Numerical PDEs

1 Elements of function spaces

An n-tuple $\alpha = (\alpha_1, \dots, \alpha_n) \in \mathbb{N}^n$ is called a *multi-index*.

$|\alpha| := \alpha_1 + \cdots + \alpha_n$ is called the length of the multi-index $\alpha = (\alpha_1, \dots, \alpha_n)$.

$$D^\alpha := \left(\frac{\partial}{\partial x_1}\right)^{\alpha_1} \cdots \left(\frac{\partial}{\partial x_n}\right)^{\alpha_n} = \frac{\partial^{|\alpha|}}{\partial x_1^{\alpha_1} \cdots \partial x_n^{\alpha_n}}.$$

$$\|u\|_{C^k(\overline{\Omega})} := \sum_{|\alpha| \le k} \sup_{x \in \Omega} |D^\alpha u(x)|,$$

when $k = 0$, we shall write $C(\overline{\Omega})$ instead of $C^0(\overline{\Omega})$; $\|u\|_{C(\overline{\Omega})} = \sup_{x \in \Omega} |u(x)| = \max_{x \in \overline{\Omega}} |u(x)|$

1.2 Spaces of integrable functions

Next we define a class of spaces that consist of (Lebesgue) integrable functions. Let p be a real number, $p \ge 1$; we denote by $L_p(\Omega)$ the set of all real-valued functions defined on an open set $\Omega \subset \mathbb{R}^n$ such that

$$\int_\Omega |u(x)|^p \, \mathrm{d}x < \infty. \qquad \text{with norm} \qquad \|u\|_{L_p(\Omega)} := \left(\int_\Omega |u(x)|^p \, \mathrm{d}x\right)^{1/p}.$$

$$(u, v) := \int_\Omega u(x) v(x) \, \mathrm{d}x.$$

Inner product for L2

Lemma 1 *(The Cauchy–Schwarz inequality). Let $u, v \in L_2(\Omega)$; $|(u, v)| \le \|u\|_{L_2(\Omega)} \|v\|_{L_2(\Omega)}$.*

Corollary 1 *(The triangle inequality) Let u, v belong to $L_2(\Omega)$; then $u + v \in L_2(\Omega)$, and*

$$\|u + v\|_{L_2(\Omega)} \le \|u\|_{L_2(\Omega)} + \|v\|_{L_2(\Omega)}.$$

1.3 Sobolev spaces

consist of functions $u \in L_2(\Omega)$ whose weak derivatives $D^\alpha u$ are also elements of $L_2(\Omega)$. To give a precise definition of a Sobolev space, we shall first explain the meaning of *weak derivative*.

Suppose that u is a smooth function, say $u \in C^k(\Omega)$, and let $v \in C_0^\infty(\Omega)$; then we have the following integration-by-parts formula:

$$\int_\Omega D^\alpha u(x) \, v(x) \, \mathrm{d}x = (-1)^{|\alpha|} \int_\Omega u(x) \, D^\alpha v(x) \, \mathrm{d}x \qquad \forall \alpha : |\alpha| \le k, \quad \forall v \in C_0^\infty(\Omega).$$

We note here that all integrals on $\partial\Omega$ that arise in the course of partial integration, based on the divergence theorem,[1] have vanished because $v \in C_0^\infty(\Omega)$. However, in the theory of partial differential equations one often has to consider functions u that do not possess the smoothness hypothesized above, yet they have to be differentiated (in some sense). It is for this purpose that we introduce the idea of a *weak derivative*.

Suppose that u is locally integrable on Ω (i.e., $u \in L_1(\omega)$ for each bounded open set ω, with $\overline{\omega} \subset \Omega$). Suppose also that there exists a function w_α, locally integrable on Ω, and such that

$$\int_\Omega w_\alpha(x) \, v(x) \, \mathrm{d}x = (-1)^{|\alpha|} \int_\Omega u(x) \, D^\alpha v(x) \, \mathrm{d}x \quad \forall v \in C_0^\infty(\Omega).$$

Then we say that w_α is the *weak derivative* of u (of order $|\alpha| = \alpha_1 + \cdots + \alpha_n$) and write $w_\alpha = D^\alpha u$.

(coincides with regular derivative when it exists)

Now we are ready to give a precise definition of a Sobolev space. Let k be a nonnegative integer. We define (with D^α denoting a weak derivative of order $|\alpha|$)

$$H^k(\Omega) := \{u \in L_2(\Omega) : D^\alpha u \in L_2(\Omega), \quad |\alpha| \le k\}.$$

$H^k(\Omega)$ is called a Sobolev space of order k; it is equipped with the (Sobolev) norm

$$\|u\|_{H^k(\Omega)} := \left(\sum_{|\alpha| \le k} \|D^\alpha u\|^2_{L_2(\Omega)} \right)^{1/2}$$

and inner product

$$(u, v)_{H^k(\Omega)} := \sum_{|\alpha| \le k} (D^\alpha u, D^\alpha v).$$

With this inner product, $H^k(\Omega)$ is a Hilbert space (for the definition of Hilbert space, see the remark in Section 1.2). Letting

$$|u|_{H^k(\Omega)} := \left(\sum_{|\alpha| = k} \|D^\alpha u\|^2_{L_2(\Omega)} \right)^{1/2} \quad \Rightarrow \quad \|u\|_{H^k(\Omega)} = \left(\sum_{j=0}^{k} |u|^2_{H^j(\Omega)} \right)^{1/2}.$$

Throughout these notes we shall frequently use $H^1(\Omega)$ and $H^2(\Omega)$.

Finally, we define a special Sobolev space, $H^1_0(\Omega) := \{u \in H^1(\Omega) : u = 0 \text{ on } \partial\Omega\}$,

Lemma 2 *(Poincaré–Friedrichs inequality). Suppose that Ω is a bounded open set in $\mathbb{R}^n$ (with a sufficiently smooth boundary $\partial\Omega$) and let $u \in H^1_0(\Omega)$; then, there exists a positive constant $c_\star(\Omega)$, independent of u, such that*

$$\int_\Omega u^2(x)\,\mathrm{d}x \le c_\star \sum_{i=1}^{n} \int_\Omega |\partial_{x_i} u(x)|^2\,\mathrm{d}x. \tag{1}$$

Write u(x,y) as integral fro a to x + u(a,y), then apply cauchy schwarz. Do same for y side

2: Elliptic boundary-value problems

Laplace equation: $\Delta u = 0$, Poisson's equation $-\Delta u = f$.

More generally, let Ω be a bounded open set in $\mathbb{R}^n$, and consider the (linear) second-order partial differential equation

$$-\sum_{i,j=1}^{n} \frac{\partial}{\partial x_j}\left(a_{i,j}(x) \frac{\partial u}{\partial x_i} \right) + \sum_{i=1}^{n} b_i(x) \frac{\partial u}{\partial x_i} + c(x)u = f(x), \quad x \in \Omega, \tag{2}$$

where the coefficients $a_{i,j}$, b_i, c and f satisfy the following conditions:

$$\begin{aligned} a_{i,j} &\in C^1(\overline{\Omega}), && i,j = 1,\dots,n; \\ b_i &\in C(\overline{\Omega}), && i = 1,\dots,n; \\ c &\in C(\overline{\Omega}), && f \in C(\overline{\Omega}), \quad \text{and} \\ \sum_{i,j=1}^{n} a_{i,j}(x)\xi_i\xi_j &\ge \tilde{c} \sum_{i=1}^{n} \xi_i^2, && \forall\, \xi = (\xi_1,\dots,\xi_n) \in \mathbb{R}^n, \quad \forall\, x \in \overline{\Omega}; \end{aligned} \tag{3}$$

here $\tilde{c}$ is a positive constant independent of x and ξ.

The equation (2) is supplemented with one of the following boundary conditions:

(a) $u = g$ on $\partial\Omega$ (*Dirichlet boundary condition*);

(b) $\frac{\partial u}{\partial \nu} = g$ on $\partial\Omega$, where ν denotes the unit outward normal vector to the boundary $\partial\Omega$ of Ω, and where the derivative in the direction of ν is defined by $\frac{\partial u}{\partial \nu} := \nabla u \cdot \nu$ (*Neumann boundary condition*);

(c) $\frac{\partial u}{\partial \nu} + \sigma u = g$ on $\partial\Omega$, where $\sigma(x) \ge 0$ on $\partial\Omega$ (*Robin boundary condition*);

(d) A more general version of the boundary conditions (b) and (c) is

$$\sum_{i,j=1}^{n} a_{i,j} \frac{\partial u}{\partial x_i} \cos\alpha_j + \sigma(x)u = g \quad \text{on } \partial\Omega,$$

We begin by considering the homogeneous Dirichlet boundary-value problem

$$-\sum_{i,j=1}^{n}\frac{\partial}{\partial x_j}\left(a_{i,j}(x)\frac{\partial u}{\partial x_i}\right)+\sum_{i=1}^{n}b_i(x)\frac{\partial u}{\partial x_i}+c(x)u=f(x) \qquad \text{for } x\in\Omega, \tag{4}$$

$$u=0 \qquad \text{on } \partial\Omega, \tag{5}$$

where $a_{i,j}$, b_i, c and f are as in (3).

A function $u\in C^2(\Omega)\cap C(\overline{\Omega})$ satisfying (4) and (5) is called a *classical solution* of this problem. The theory of partial differential equations tells us that (4), (5) has a unique classical solution, provided that $a_{i,j}$, b_i, c, f and $\partial\Omega$ are sufficiently smooth. However, in many applications one has to consider boundary-value problems where these smoothness requirements are violated, and for such problems the classical theory of partial differential equations is inappropriate. Take, for example, Poisson's equation on the cube $\Omega=(-1,1)^n$ in $\mathbb{R}^n$, subject to a zero Dirichlet boundary condition:

$$\left.\begin{aligned} -\Delta u &= \operatorname{sgn}\left(\tfrac{1}{2}-|x|\right), & x\in\Omega,\\ u &= 0, & x\in\partial\Omega. \end{aligned}\right\} \tag{$*$}$$

This problem does not have a classical solution, $u\in C^2(\Omega)\cap C(\overline{\Omega})$, for otherwise Δu would be a continuous function on Ω, which is not possible because $\operatorname{sgn}(1/2-|x|)$ is not a continuous function on Ω.

3 Introduction to the theory of finite difference schemes

Let Ω be a bounded open set in $\mathbb{R}^n$ $\mathcal{L}u=f$ in Ω, $\mathcal{B}u=g$ on $\Gamma:=\partial\Omega$,

where $\mathcal{L}$ is a linear partial differential operator, and $\mathcal{B}$ is a linear operator which specifies the boundary condition. For example,

$$\mathcal{L}u\equiv-\sum_{i,j=1}^{n}\frac{\partial}{\partial x_j}\left(a_{i,j}(x)\frac{\partial u}{\partial x_i}\right)+\sum_{i=1}^{n}b_i\frac{\partial u}{\partial x_i}+cu,$$

e.g

$\mathcal{B}u\equiv u$ (Dirichlet boundary condition), $\mathcal{B}u\equiv\dfrac{\partial u}{\partial\nu}$ (Neumann boundary condition),

$$\mathcal{B}u\equiv\sum_{i,j=1}^{n}a_{i,j}(x)\cos\alpha_j+\sigma(x)u \quad \text{(oblique derivative boundary condition)},$$

In general, impossible to solve in closed form. So want to approximate via *finite difference method*

'approximated' $\overline{\Omega}=\Omega\cup\Gamma$ via $\overline{\Omega}_h=\Omega_h\cup\Gamma_h$, finite set of points $\Omega_h\subset\Omega$ and $\Gamma_h\subset\Gamma$;

$\overline{\Omega}_h$ is called a *mesh*, Ω_h is the *set of interior mesh-points* and Γ_h the *set of boundary mesh-points*.

The parameter $h=(h_1,\ldots,h_n)$ measures the 'fineness' of the mesh (size of mesh in Oxi direction) smaller $\max_{1<i<n}h_i$ is, the finer the mesh.

Having constructed the mesh, we proceed by replacing the derivatives in $\mathcal{L}$ by divided differences, and we approximate the boundary condition in a similar fashion. This yields the finite difference scheme

$$\mathcal{L}_hU(x)=f_h(x), \qquad x\in\Omega_h, \quad \mathcal{B}_hU(x)=g_h(x), \qquad x\in\Gamma_h,$$

where f_h and g_h are suitable approximations of f and g, respectively. Now (19) is a system of linear algebraic equations involving the values of U at the mesh-points, and can be solved by Gaussian elimination or an iterative method, provided, of course, that it has a unique solution. The sequence $\{U(x):x\in\overline{\Omega}_h\}$ is an approximation to $\{u(x):x\in\overline{\Omega}_h\}$, the values of the exact solution at the mesh-points.

There are two classes of problems associated with finite difference schemes:

(1) the first, and more fundamental, is the problem of approximation, that is, whether (19) approximates the boundary-value problem (18) in some sense, and whether its solution $\{U(x):x\in\overline{\Omega}_h\}$ approximates $\{u(x):x\in\overline{\Omega}_h\}$, the values of the exact solution at the mesh-points.

(2) the second problem concerns the effective solution of the discrete problem (19) using techniques from Numerical Linear Algebra.

3.1 Finite difference approximation of a two-point boundary-value problem

$$-u'' + c(x)u = f(x), \quad x \in (0,1),$$
$$u(0) = 0, \quad u(1) = 0,$$

where f and c are real-valued functions, which are defined and continuous on the interval $[0,1]$ and $c(x) \geq 0$ for all $x \in [0,1]$.

The first step in the construction of a finite difference scheme for this boundary-value problem is to define the mesh. Let N be an integer, $N \geq 2$, and let $h := 1/N$ be the mesh-size; the mesh-points are $x_i := ih$, $i = 0, \ldots, N$. Formally, $\Omega_h := \{x_i : i = 1, \ldots, N-1\}$ is the set of interior mesh-points, $\Gamma_h := \{x_0, x_N\}$ the set of boundary mesh-points and $\overline{\Omega}_h := \Omega_h \cup \Gamma_h$ the set of all mesh-points. Suppose that u is sufficiently smooth (e.g. $u \in C^4([0,1])$). Then, by Taylor series expansion,

$$D_x^+ u(x_i) := \frac{u(x_{i+1}) - u(x_i)}{h} = u'(x_i) + \mathcal{O}(h), \quad D_x^- u(x_i) := \frac{u(x_i) - u(x_{i-1})}{h} = u'(x_i) + \mathcal{O}(h),$$

$$D_x^+ D_x^- u(x_i) = D_x^- D_x^+ u(x_i) \quad = \frac{u(x_{i+1}) - 2u(x_i) + u(x_{i-1})}{h^2} = u''(x_i) + \mathcal{O}(h^2).$$

$$D_x^0 u(x_i) := \frac{1}{2}\left(D_x^+ u(x_i) + D_x^- u(x_i)\right) = \frac{u(x_{i+1}) - u(x_{i-1})}{2h} \quad (= u'(x_i) + \mathcal{O}(h^2)).$$

Thus we replace the second derivative u'' in the differential equation by the second divided difference $D_x^+ D_x^- u(x_i)$; hence,

$$-D_x^+ D_x^- u(x_i) + c(x_i)u(x_i) \approx f(x_i), \quad i = 1, \ldots, N-1, \quad u(x_0) = 0, \quad u(x_N) = 0.$$

This is, in fact, a system of $N-1$ linear algebraic equations for the $N-1$ unknowns, U_i, $i = 1, \ldots, N-1$.

$AU = F$, where A is the symmetric tridiagonal $(N-1) \times (N-1)$

U and F are column vectors of size $N-1$, corresponding to the $N-1$ 'interior' mesh-points

3.2 Existence and uniqueness of solutions, stability, consistency, and convergence

We begin the analysis of the finite difference scheme (22) by showing that it has a unique solution. It suffices to show that the matrix A is nonsingular (i.e., $\det A \neq 0$), and therefore invertible. We shall do so by developing a technique which we shall, in subsequent sections, extend to the finite difference approximation of partial differential equations. The purpose of this section is to introduce the key ideas through the finite difference approximation (21) of the simple two-point boundary-value problem (20).

For this purpose, we introduce, for two functions V and W defined at the interior mesh-points x_i, $i = 1, \ldots, N-1$, the inner product

$$(V, W)_h := \sum_{i=1}^{N-1} h V_i W_i,$$

resembles L2, inner product, aim to mimic integration-by-parts

Lemma 3 *Suppose that V is a function defined at the mesh-points x_i, $i = 0, \ldots, N$, and let $V_0 = V_N = 0$;*

$$(-D_x^+ D_x^- V, V)_h = \sum_{i=1}^{N} h \left|D_x^- V_i\right|^2.$$

(proof: write in sigma notation, expands Ds, shift one of the sigmas, combine)

Returning to the finite difference scheme (22), let V be as in the above lemma and note that as, by hypothesis, $c(x) \geq 0$ for all $x \in [0,1]$, we have that

$$(AV, V)_h = (-D_x^+ D_x^- V + cV, V)_h \; = (-D_x^+ D_x^- V, V)_h + (cV, V)_h \quad \geq \sum_{i=1}^{N} h \left|D_x^- V_i\right|^2 \tag{25}$$

Thus, if $AV = 0$ for some V, then $D_x^- V_i = 0$, $i = 1, \ldots, N$; because $V_0 = V_N = 0$, this implies that $V_i = 0$, $i = 0, \ldots, N$. Hence $AV = 0$ if and only if $V = 0$. It therefore follows that A is a nonsingular

Theorem 3 *Suppose that c and f are continuous real-valued functions defined on the interval $[0,1]$, and $c(x) \geq 0$ for all $x \in [0,1]$; then, the finite difference scheme (22) possesses a unique solution U.*

Next, we investigate the approximation properties of the finite difference scheme

discrete L_2-norm $$\|U\|_h := (U,U)_h^{1/2} = \left(\sum_{i=1}^{N-1} h|U_i|^2 \right)^{1/2},$$

discrete Sobolev norm $$\|U\|_{1,h} := (\|U\|_h^2 + \|D_x^- U]\|_h^2)^{1/2},$$ where $$\|V]\|_h^2 := \sum_{i=1}^{N} h\,|V_i|^2$$

$$(V,W]_h := \sum_{i=1}^{N} h\,V_i\,W_i.$$

(25) can be rewritten as follows:

$$(AV,V)_h \geq \|D_x^- V]\|_h^2.$$

Lemma 4 *(Discrete Poincaré–Friedrichs inequality.) Let V be a function defined on the finite difference mesh $\{x_i := ih : i = 0,\ldots,N\}$, where $h := 1/N$ and $N \geq 2$, and such that $V_0 = V_N = 0$; then, there exists a positive constant $c_\star$, independent of V and h, such that*

$$\|V\|_h^2 \leq c_\star \|D_x^- V]\|_h^2 \tag{27}$$

for all such V.

(prove something similar for each Vi via C-S, then sum all Vi together)

Theorem 4 *The scheme (22) is stable in the sense that* $$\|U\|_{1,h} \leq \frac{1}{c_0}\|f\|_h.$$

(use lemma 4 on rewriting of (25))

. We define the *global error*, e, by $e_i := u(x_i) - U_i, \quad i = 0,\ldots,N.$

$$Ae_i = Au(x_i) - AU_i = Au(x_i) - f(x_i) = -D_x^+ D_x^- u(x_i) + c(x_i)u(x_i) - f(x_i)$$

$$= u''(x_i) - D_x^+ D_x^- u(x_i), \qquad i = 1,\ldots,N-1.$$

$$Ae_i = \varphi_i, \qquad i = 1,\ldots,N-1, \quad e_0 = 0, \quad e_N = 0,$$

where $\varphi_i := Au(x_i) - f(x_i) = u''(x_i) - D_x^+ D_x^- u(x_i)$ is the *consistency error*

. By applying the inequality (30) to the finite difference scheme (31), we obtain

$$\|u - U\|_{1,h} = \|e\|_{1,h} \leq \frac{1}{c_0}\|\varphi\|_h.$$

bound trunc error by Taylor expansion

$$\varphi_i = u''(x_i) - D_x^+ D_x^- u(x_i) = \mathcal{O}(h^2), \qquad \|\varphi\|_h = \left(\sum_{i=1}^{N-1} h\,|\varphi_i|^2 \right)^{1/2} \leq Ch^2.$$

Combine $$\|u - U\|_{1,h} \leq \frac{C}{c_0} h^2.$$

Theorem 5 *Let $f \in C([0,1])$, $c \in C([0,1])$, with $c(x) \geq 0$ for all $x \in [0,1]$, and suppose that the corresponding (weak) solution of the boundary-value problem (20) belongs to $C^4([0,1])$; then*

$$\|u - U\|_{1,h} \leq \frac{1}{8} h^2 \|u^{IV}\|_{C([0,1])}. \tag{35}$$

(1) The first step is to prove the stability of the scheme in an appropriate mesh-dependent norm (c.f. inequality (30), for example). A typical stability result for the general finite difference scheme (19) is

$$|||U|||_{\Omega_h} \leq C_1 \left(\|f_h\|_{\Omega_h} + \|g_h\|_{\Gamma_h}\right), \tag{36}$$

where $|||\cdot|||_{\Omega_h}$, $\|\cdot\|_{\Omega_h}$ and $\|\cdot\|_{\Gamma_h}$ are mesh-dependent norms involving mesh-points of Ω_h (or $\overline{\Omega}_h$) and Γ_h, respectively, and C_1 is a positive constant, independent of h.

(2) The second step is to estimate the size of the *consistency error*,

$$\begin{aligned} \varphi_{\Omega_h} &:= \mathcal{L}_h u - f_h, \qquad \text{in } \Omega_h, \\ \varphi_{\Gamma_h} &:= \mathcal{B}_h u - g_h, \qquad \text{on } \Gamma_h. \end{aligned}$$

(in the case of the finite difference scheme (20) $\varphi_{\Gamma_h} = 0$, and therefore φ_{Γ_h} did not appear explicitly in our error analysis). If

$$\|\varphi_{\Omega_h}\|_{\Omega_h} + \|\varphi_{\Gamma_h}\|_{\Gamma_h} \to 0 \quad \text{as} \quad h \to 0,$$

for a sufficiently smooth solution u of the boundary-value problem (18), we say that the scheme (19) is *consistent*. If p is the largest positive integer such that

$$\|\varphi_{\Omega_h}\|_{\Omega_h} + \|\varphi_{\Gamma_h}\|_{\Gamma_h} \leq C_2 h^p \quad \text{as} \quad h \to 0,$$

(where C_2 is a positive constant independent of h) for all sufficiently smooth u, the scheme is said to have *order of accuracy* (or *order of consistency*) p.

The finite difference scheme (19) is said to provide a *convergent* approximation to the solution u of the boundary-value problem (18) in the norm $|||\cdot|||_{\Omega_h}$, if

$$|||u - U|||_{\Omega_h} \to 0 \quad \text{as } h \to 0.$$

If q is the largest positive integer such that

$$|||u - U|||_{\Omega_h} \leq C h^q \quad \text{as } h \to 0$$

(where C is a positive constant independent of the mesh-size h), then the scheme is said to have *order of convergence* q.

From these definitions we deduce the following fundamental theorem.

Theorem 6 *Suppose that the finite difference scheme (19), involving linear finite difference operators $\mathcal{L}_h$ and $\mathcal{B}_h$, is stable (i.e., the inequality (36) holds for all f_h and g_h) and that the scheme is a consistent approximation of the boundary-value problem (18); then the finite difference scheme (19) is a convergent approximation of the boundary-value problem (18), and the order of convergence q is not smaller then the*

Thus, paraphrasing Theorem 3.6, *stability* and *consistency* imply *convergence*.

PROOF. We define the *global error* $e := u - U$. Then, thanks to the assumed linearity of $\mathcal{L}_h$, we have

$$\mathcal{L}_h e = \mathcal{L}_h(u - U) = \mathcal{L}_h u - \mathcal{L}_h U = \mathcal{L}_h u - f_h. \quad \Rightarrow \quad \mathcal{L}_h e = \varphi_{\Omega_h}. \quad + \quad \mathcal{B}_h e = \varphi_{\Gamma_h}. \quad \Rightarrow$$

$$|||u - U|||_{\Omega_h} = |||e|||_{\Omega_h} \leq C_1(\|\varphi_{\Omega_h}\|_{\Omega_h} + \|\varphi_{\Gamma_h}\|_{\Gamma_h}),$$

4 Finite difference approximation of elliptic boundary-value problems

$$-\Delta u + c(x,y)u = f(x,y) \qquad \text{in } \Omega, \quad u = 0 \qquad \text{on } \partial\Omega, \tag{37}$$

where $\Omega = (0,1) \times (0,1)$, c is a continuous function on $\overline{\Omega}$ and $c(x,y) \geq 0$. As far as the smoothness of the function f is concerned, we shall consider two separate cases:

(a) First we shall assume that f is a continuous function on $\overline{\Omega}$. In this case, the error analysis will proceed along the same lines as in Section 3.

(b) We shall then consider the case when f is only in $L_2(\Omega)$. As f need not be continuous on Ω, the boundary-value problem (37) need not have a classical solution – only a weak solution exists. This gives rise to technical difficulties: in particular, we cannot use a Taylor series expansion to estimate the size of the consistency error. We shall bypass the problem by employing a different technique.

(a) $(f \in C(\overline{\Omega}))$ mesh: $i,j = 0,\ldots,N$, where $x_i := ih$, $y_j := jh$.

$\overline{\Omega}_h := \{(x_i,y_j) \in \overline{\Omega} : i,j = 0,\ldots,N\}$ $\Omega_h := \{(x_i,y_j) \in \Omega : i,j = 1,...,N-1\}$ $\Gamma_h := \overline{\Omega}_h \setminus \Omega_h$.

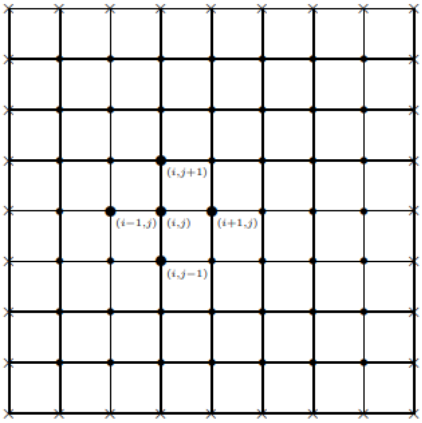
just like before scheme is: $U = 0$ on Γ_h

$$-(D_x^+ D_x^- U_{i,j} + D_y^+ D_y^- U_{i,j}) + c(x_i,y_j)U_{i,j} = f(x_i,y_j) \qquad \text{for } (x_i,y_j) \in \Omega_h,$$

(all we're doing is swapping derivatives of u with their corresponding approximations)

It is again possible to write (39), (40) as a system of linear algebraic equations

$$U = (U_{11}, U_{12}, \ldots, U_{1,N-1}, U_{21}, U_{22}, \ldots, U_{2,N-1}, \ldots, U_{i,N-1}, \ldots, U_{N-1,1}, U_{N-1,2}, \ldots, U_{N-1,N-1})^{\mathrm{T}}$$

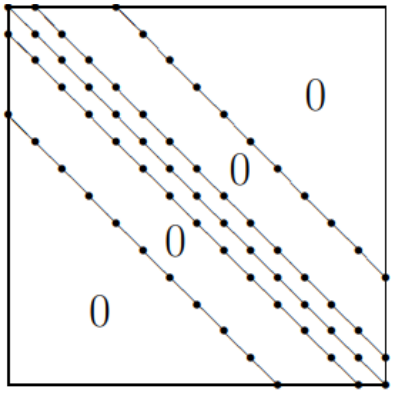


$$F = (F_{11}, F_{12}, \ldots, F_{1,N-1}, F_{21}, F_{22}, \ldots, F_{2,N-1}, \ldots, F_{i1}, F_{i2}, \ldots, F_{i,N-1}, \ldots, F_{N-1,1}, F_{N-1,2}, \ldots, F_{N-1,N-1})^{\mathrm{T}},$$

A is an $(N-1)^2 \times (N-1)^2$ sparse matrix of banded structure

(banded means non-zero elements are confined to diagonals)

$A =$



$$(V,W)_h := \sum_{i=1}^{N-1}\sum_{j=1}^{N-1} h^2 V_{i,j} W_{i,j},$$

(Resembles L2 inner product)

4.1 Existence and uniqueness of a solution, stability, consistency, and convergence

Lemma 5 *Suppose that V is a function defined on $\overline{\Omega}_h$ and that $V = 0$ on Γ_h; then,*

$$(-D_x^+ D_x^- V, V)_h + (-D_y^+ D_y^- V, V)_h = \sum_{i=1}^{N}\sum_{j=1}^{N-1} h^2 |D_x^- V_{i,j}|^2 + \sum_{i=1}^{N-1}\sum_{j=1}^{N} h^2 |D_y^- V_{i,j}|^2.$$

Directly from lemma 3

now proceed in much the same way as in the univariate case considered in the previous section

$$(AV,V)_h = (-D_x^+ D_x^- V - D_y^+ D_y^- V + cV, V)_h$$

$$= (-D_x^+ D_x^- V, V)_h + (-D_y^+ D_y^- V, V)_h + (cV, V)_h \geq \sum_{i=1}^{N}\sum_{j=1}^{N-1} h^2 |D_x^- V_{i,j}|^2 + \sum_{i=1}^{N-1}\sum_{j=1}^{N} h^2 |D_y^- V_{i,j}|^2,$$

for any V defined on $\overline{\Omega}_h$ such that $V = 0$ on Γ_h. Now this implies, just as in the one-dimensional analysis presented in Section 3, that A is a nonsingular matrix. Indeed if $AV = 0$, then (43) yields:

$$D_x^- V_{i,j} = \frac{V_{i,j} - V_{i-1,j}}{h} = 0, \quad \begin{matrix} i = 1, \ldots, N, \\ j = 1, \ldots, N-1; \end{matrix} \qquad D_y^- V_{i,j} = \frac{V_{i,j} - V_{i,j-1}}{h} = 0, \quad \begin{matrix} i = 1, \ldots, N-1 \\ j = 1, \ldots, N. \end{matrix}$$

Since $V = 0$ on Γ_h, these imply that $V \equiv 0$.

In order to prove the stability of the finite difference scheme (38), we introduce (similarly as in the univariate case) the mesh–dependent norms

$$\|U\|_h := (U, U)_h^{1/2}, \quad \text{and} \quad \|U\|_{1,h} := (\|U\|_h^2 + \|D_x^- U]\|_x^2 + \|D_y^- U]\|_y^2)^{1/2},$$

$$\|D_x^- U]\|_x := \left(\sum_{i=1}^{N}\sum_{j=1}^{N-1} h^2 |D_x^- U_{i,j}|^2 \right)^{1/2} \qquad \|D_y^- U]\|_y := \left(\sum_{i=1}^{N-1}\sum_{j=1}^{N} h^2 |D_y^- U_{i,j}|^2 \right)^{1/2}.$$

$$\|u\|_{H^1(\Omega)} := \left(\|u\|_{L_2(\Omega)}^2 + \left\| \frac{\partial u}{\partial x} \right\|_{L_2(\Omega)}^2 + \left\| \frac{\partial u}{\partial y} \right\|_{L_2(\Omega)}^2 \right)^{1/2}$$ (discrete Sobolev Norm)

(43) can be rewritten in the following compact form: $(AV, V)_h \geq \|D_x^- V]\|_x^2 + \|D_y^- V]\|_y^2$.

Lemma 6 *(Discrete Poincaré–Friedrichs inequality.) Suppose that V is a function defined on $\overline{\Omega}_h$ and such that $V = 0$ on Γ_h; then, there exists a constant c_*, independent of V and h, such that*

$$\|V\|_h^2 \leq c_* \left(\|D_x^- V]\|_x^2 + \|D_y^- V]\|_y^2 \right) \quad \textit{for all such } V. \quad (45)$$ (same as univariate proof)

Theorem 7 *The finite difference scheme (38) is stable in the sense that* $\|U\|_{1,h} \leq \frac{1}{c_0} \|f\|_h$.

(same as univariate proof)

4.1.1 Convergence in the class of classical solutions

$$e_{i,j} := u(x_i, y_j) - U_{i,j}, \quad 0 \leq i, j \leq N. \quad \varphi_{i,j} := Au(x_i, y_j) - f_{i,j}.$$

assuming that $u \in C^4(\overline{\Omega})$, + taylor expansion $\Rightarrow$ $Ae_{i,j} = Au(x_i, y_j) - AU_{i,j} = Au(x_i.y_j) - f_{i,j}$

$$= \varphi_{i,j} = -\frac{h^2}{12} \frac{\partial^4 u}{\partial x^4}(\xi_i, y_j) - \frac{h^2}{12} \frac{\partial^4 u}{\partial y^4}(x_i, \eta_j), \qquad 1 \leq i, j \leq N-1,$$

$$Ae_{i,j} = \varphi_{i,j}, \quad 1 \leq i, j \leq N-1, \quad e = 0 \qquad \text{on } \Gamma_h. \implies \text{(stability result)}$$

$$\|u - U\|_{1,h} = \|e\|_{1,h} \leq \frac{1}{c_0} \|\varphi\|_h.$$

now must bound phi

$$|\varphi_{i,j}| \leq \frac{h^2}{12} \left(\left\| \frac{\partial^4 u}{\partial x^4} \right\|_{C(\overline{\Omega})} + \left\| \frac{\partial^4 u}{\partial y^4} \right\|_{C(\overline{\Omega})} \right) \implies \|\varphi\|_h \leq \frac{h^2}{12} \left(\left\| \frac{\partial^4 u}{\partial x^4} \right\|_{C(\overline{\Omega})} + \left\| \frac{\partial^4 u}{\partial y^4} \right\|_{C(\overline{\Omega})} \right)$$

Theorem 8 *Let $f \in C(\overline{\Omega})$, $c \in C(\overline{\Omega})$, with $c(x, y) \geq 0$, $(x, y) \in \overline{\Omega}$, and suppose that the corresponding weak solution of the boundary-value problem (37) belongs to $C^4(\overline{\Omega})$; then,*

$$\|u - U\|_{1,h} \leq \frac{5h^2}{48} \left(\left\| \frac{\partial^4 u}{\partial x^4} \right\|_{C(\overline{\Omega})} + \left\| \frac{\partial^4 u}{\partial y^4} \right\|_{C(\overline{\Omega})} \right). \quad (52)$$

(b) ($f \in L_2(\Omega)$). We retain the same finite difference mesh as in case (a), but we shall modify the right-hand side in the finite difference scheme (39) to cater for the fact that f is no longer assumed to be a continuous function on $\overline{\Omega}$.

The idea is to replace $f(x_i, y_j)$ in (39) by a 'cell-average' of f, $Tf_{i,j} := \dfrac{1}{h^2}\displaystyle\int_{K_{i,j}} f(x,y)\,\mathrm{d}x\,\mathrm{d}y,$

$$K_{i,j} = \left[x_i - \frac{h}{2}, x_i + \frac{h}{2}\right] \times \left[y_j - \frac{h}{2}, y_j + \frac{h}{2}\right]$$ (justification in notes, uses divergence theorem)

Clearly, $Tf_{i,j}$ is well defined for f in $L_2(\Omega)$ (in fact, $Tf_{i,j}$ is well defined even if $f \in L_1(\Omega)$ only)

$$|Tf_{i,j}| = \frac{1}{h^2}\left|\int_{K_{i,j}} f(x,y)\,\mathrm{d}x\,\mathrm{d}y\right| \le \frac{1}{h^2}\left(\int_{K_{i,j}} 1^2\,\mathrm{d}x\,\mathrm{d}y\right)^{1/2}\left(\int_{K_{i,j}} |f(x,y)|^2\,\mathrm{d}x\,\mathrm{d}y\right)^{1/2}$$

$$= \frac{1}{h}\|f\|_{L_2(K_{i,j})} \le \frac{1}{h}\|f\|_{L_2(\Omega)}.$$

Thus we define our finite difference (or, more precisely, finite volume) approximation of (37) by

$$-(D_x^+ D_x^- U_{i,j} + D_y^+ D_y^- U_{i,j}) + c(x_i, y_j)U_{i,j} = Tf_{i,j}, \quad \text{for } (x_i, y_j) \in \Omega_h, \qquad U = 0, \quad \text{on } \Gamma_h.$$

As fin diff operator is unchanged, our old argument still works

Theorem 9 *The scheme (54) is stable in the sense that* $\|U\|_{1,h} \le \dfrac{1}{c_0}\|Tf\|_h \left(\le \dfrac{1}{c_0}\|f\|_{L_2(\Omega)}\right)$

Global error: $Ae_{i,j} = Au(x_i, y_j) - AU_{i,j} = Au(x_i, y_j) - Tf_{i,j}$

$$= -(D_x^+ D_x^- u(x_i, y_j) + D_y^+ D_y^- u(x_i, y_j)) + c(x_i, y_j)u(x_i, y_j)$$

$$+ \left(T\left(\frac{\partial^2 u}{\partial x^2}\right)(x_i, y_j) + T\left(\frac{\partial^2 u}{\partial y^2}\right)(x_i, y_j) - T(cu)(x_i, y_j)\right)$$ By noting that

$$T\left(\frac{\partial^2 u}{\partial x^2}\right)(x_i, y_j) = \frac{1}{h}\int_{y_j - h/2}^{y_j + h/2} \frac{\frac{\partial u}{\partial x}(x_i + h/2, y) - \frac{\partial u}{\partial x}(x_i - h/2, y)}{h}\,\mathrm{d}y$$

$$= \frac{1}{h}\int_{y_j - h/2}^{y_j + h/2} D_x^+ \frac{\partial u}{\partial x}(x_i - h/2, y)\,\mathrm{d}y = D_x^+\left[\frac{1}{h}\int_{y_j - h/2}^{y_j + h/2} \frac{\partial u}{\partial x}(x_i - h/2, y)\,\mathrm{d}y\right]$$ and

$$T\left(\frac{\partial^2 u}{\partial y^2}\right)(x_i, y_j) = D_y^+\left[\frac{1}{h}\int_{x_i - h/2}^{x_i + h/2} \frac{\partial u}{\partial y}(x, y_j - h/2)\,\mathrm{d}x\right]$$ conclude

$$\varphi_1(x_i, y_j) := \frac{1}{h}\int_{y_j - h/2}^{y_j + h/2} \frac{\partial u}{\partial x}(x_i - h/2, y)\,\mathrm{d}y - D_x^- u(x_i, y_j),$$

$$\varphi_2(x_i, y_j) := \frac{1}{h}\int_{x_i - h/2}^{x_i + h/2} \frac{\partial u}{\partial y}(x, y_j - h/2)\,\mathrm{d}x - D_y^- u(x_i, y_j),$$

$$Ae = D_x^+\varphi_1 + D_y^+\varphi_2 + \psi, \qquad \psi(x_i, y_j) := (cu)(x_i, y_j) - T(cu)(x_i, y_j).$$

$$Ae = D_x^+\varphi_1 + D_y^+\varphi_2 + \psi \quad \text{in } \Omega_h, \qquad e = 0 \quad \text{on } \Gamma_h.$$ stability of scheme won't give us a good bound $\|e\|_{1,h} \le \dfrac{1}{c_0}\|D_x^+\varphi_1 + D_y^+\varphi_2 + \psi\|_h,$ (doesn't use $\varphi := D_x^+\varphi_1 + D_y^+\varphi_2 + \psi,$)

Use different approach $c_0\|e\|_{1,h}^2 \le (Ae,e)_h = (D_x^+\varphi_1,e)_h + (D_y^+\varphi_2,e)_h + (\psi,e)_h$ (58)

$e = 0$ on Γ_h, we then have that $(D_x^+\varphi_1,e)_h = \sum_{j=1}^{N-1} h\left(\sum_{i=1}^{N-1} h\frac{\varphi_1(x_{i+1},y_j)-\varphi_1(x_i,y_j)}{h}e_{i,j}\right)$

$$= -\sum_{j=1}^{N-1} h\left(\sum_{i=1}^{N} h\varphi_1(x_i,y_j)\frac{e_{i,j}-e_{i-1,j}}{h}\right) = -\sum_{i=1}^{N}\sum_{j=1}^{N-1} h^2\varphi_1(x_i,y_j)D_x^- e_{i,j}$$

$$\le \left(\sum_{i=1}^{N}\sum_{j=1}^{N-1} h^2|\varphi_1(x_i,y_j)|^2\right)^{1/2}\left(\sum_{i=1}^{N}\sum_{j=1}^{N-1} h^2|D_x^- e_{i,j}|^2\right)^{1/2} = \|\varphi_1]|_x\|D_x^- e]|_x. \quad \text{(C-S)}$$

Thus, $(D_x^+\varphi_1,e)_h \le \|\varphi_1]|_x\|D_x^- e]|_x$. $(D_y^+\varphi_2,e)_h \le \|\varphi_2]|_y\|D_y^- e]|_y$ $(\psi,e)_h \le \|\psi\|_h\|e\|_h$. (C-S)

Subbing these into (58) : $c_0\|e\|_{1,h}^2 \le \left(\|\varphi_1]|_x^2 + \|\varphi_2]|_y^2 + \|\psi\|_h^2\right)^{1/2}\|e\|_{1,h}$.

Lemma 7 *The global error, e, of the finite difference scheme (54) satisfies the inequality*

$$\|e\|_{1,h} \le \frac{1}{c_0}(\|\varphi_1]|_x^2 + \|\varphi_2]|_y^2 + \|\psi\|_h^2)^{1/2}, \tag{62}$$

where φ_1, φ_2, and ψ are defined by

$$\varphi_1(x_i,y_j) := \frac{1}{h}\int_{y_j-h/2}^{y_j+h/2}\frac{\partial u}{\partial x}(x_i-h/2,y)\,\mathrm{d}y - D_x^- u(x_i,y_j), \tag{63}$$

for $i = 1,\ldots,N$, $j = 1,\ldots,N-1$;

$$\varphi_2(x_i,y_j) := \frac{1}{h}\int_{x_i-h/2}^{x_i+h/2}\frac{\partial u}{\partial y}(x,y_j-h/2)\,\mathrm{d}x - D_y^- u(x_i,y_j), \tag{64}$$

for $i = 1,\ldots,N-1$, $j = 1,\ldots,N$; and

$$\psi(x_i,y_j) := (cu)(x_i,y_j) - \frac{1}{h^2}\int_{x_i-h/2}^{x_i+h/2}\int_{y_j-h/2}^{y_j+h/2}(cu)(x,y)\,\mathrm{d}x\,\mathrm{d}y, \tag{65}$$

for $i,j = 1,\ldots,N-1$.

To complete the error analysis, it remains to bound φ_1, φ_2 and ψ. Using Taylor series expansions it is easily seen that

$$|\varphi_1(x_i,y_j)| \le \frac{h^2}{24}\left(\left\|\frac{\partial^3 u}{\partial x\partial y^2}\right\|_{C(\overline{\Omega})} + \left\|\frac{\partial^3 u}{\partial x^3}\right\|_{C(\overline{\Omega})}\right) \quad |\varphi_2(x_i,y_j)| \le \frac{h^2}{24}\left(\left\|\frac{\partial^3 u}{\partial x^2\partial y}\right\|_{C(\overline{\Omega})} + \left\|\frac{\partial^3 u}{\partial y^3}\right\|_{C(\overline{\Omega})}\right)$$

$$|\psi(x_i,y_j)| \le \frac{h^2}{24}\left(\left\|\frac{\partial^2 (cu)}{\partial x^2}\right\|_{C(\overline{\Omega})} + \left\|\frac{\partial^2 (cu)}{\partial y^2}\right\|_{C(\overline{\Omega})}\right)$$

Theorem 10 *Let $f \in L_2(\Omega)$, $c \in C^2(\overline{\Omega})$ with $c(x,y) \ge 0$, $(x,y) \in \overline{\Omega}$, and suppose that the corresponding weak solution, u, of the boundary-value problem (37) belongs to $C^3(\overline{\Omega})$; then,*

$$\|u - U\|_{1,h} \le \frac{5}{96}h^2 M_3, \tag{69}$$

where

$$M_3 = \left\{\left(\left\|\frac{\partial^3 u}{\partial x\partial y^2}\right\|_{C(\overline{\Omega})} + \left\|\frac{\partial^3 u}{\partial x^3}\right\|_{C(\overline{\Omega})}\right)^2 + \left(\left\|\frac{\partial^3 u}{\partial x^2 y}\right\|_{C(\overline{\Omega})} + \left\|\frac{\partial^3 u}{\partial y^3}\right\|_{C(\overline{\Omega})}\right)^2 + \left(\left\|\frac{\partial^2 (cu)}{\partial x^2}\right\|_{C(\overline{\Omega})} + \left\|\frac{\partial^2 (cu)}{\partial y^2}\right\|_{C(\overline{\Omega})}\right)^2\right\}^{1/2}.$$

4.1.2 Convergence in the class of weak solutions that belong to H3 (Ω)

Comparing (69) with (52), we see that while the smoothness requirement on the solution has been relaxed from $u \in C^4(\overline{\Omega})$ to $u \in C^3(\overline{\Omega})$, second-order convergence has been retained.

The hypothesis $u \in C^3(\overline{\Omega})$ can be further relaxed by using integral representations of φ_1, φ_2 and ψ instead of Taylor series expansions. We show how this is done for φ_1; φ_2 and ψ are handled analogously. The key idea is to use the Newton–Leibniz formula (also known as the *fundamental theorem of calculus*):

Very very long calculations, will have to revise it properly and replace this section with notes for it

Theorem 11 *Let $f \in L_2(\Omega)$, $c \in C^2(\overline{\Omega})$, with $c(x,y) \geq 0$, $(x,y) \in \overline{\Omega}$, and suppose that the corresponding weak solution of the boundary-value problem (37) belongs to $H^3(\Omega)$; then,*

$$\|u - U\|_{1,h} \leq Ch^2 \|u\|_{H^3(\Omega)}, \tag{73}$$

where C is a positive constant (computable from (70)–(72)).

4.2 Nonaxiparallel domains and nonuniform meshes

So far done stuff only for a simple fin dif scheme in square domain. Similar for other elliptic equations too. E.g

$$-\left[\frac{\partial}{\partial x}\left(a_1(x,y)\frac{\partial u}{\partial x}\right) + \frac{\partial}{\partial y}\left(a_2(x,y)\frac{\partial u}{\partial y}\right)\right] + b_1(x,y)\frac{\partial u}{\partial x} + b_2(x,y)\frac{\partial u}{\partial y} + c(x,y)u = f(x,y)$$

$$\begin{aligned}
&-\frac{1}{h}\left[a_1(x_{i+1/2}, y_j)\frac{U_{i+1,j} - U_{i,j}}{h} - a_1(x_{i-1/2}, y_j)\frac{U_{i,j} - U_{i-1,j}}{h}\right] \\
&-\frac{1}{h}\left[a_2(x_i, y_{j+1/2})\frac{U_{i,j+1} - U_{i,j}}{h} - a_2(x_i, y_{j-1/2})\frac{U_{i,j} - U_{i,j-1}}{h}\right] \\
&+ b_1(x_i, y_j)\frac{U_{i+1,j} - U_{i-1,j}}{2h} + b_2(x_i, y_j)\frac{U_{i,j+1} - U_{i,j-1}}{2h} \\
&+ c(x_i, y_j)U_{i,j} = \frac{1}{h^2}\int_{x_{i-1/2}}^{x_{i+1/2}}\int_{y_{j-1/2}}^{y_{j+1/2}} f(x,y)\,dx\,dy.
\end{aligned}$$

gives scheme

When Ω has a curved boundary, a nonuniform mesh has to be used near ∂Ω to avoid a loss of accuracy. To be more precise, let us introduce the following notation: let $h_{i+1} := x_{i+1} - x_i$, $h_i := x_i - x_{i-1}$, and let

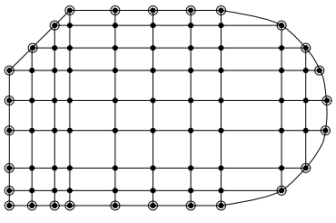
$$\hbar_i := \frac{1}{2}(h_{i+1} + h_i). D_x^+ U_i := \frac{U_{i+1} - U_i}{\hbar_i}, \quad D_x^- U_i := \frac{U_i - U_{i-1}}{h_i},$$

$$D_x^+ D_x^- U_i := \frac{1}{\hbar_i}\left(\frac{U_{i+1} - U_i}{h_{i+1}} - \frac{U_i - U_{i-1}}{h_i}\right)$$

$$k_{j+1} := y_{j+1} - y_j,\ k_j := y_j - y_{j-1}, \qquad \hbar_i := \frac{1}{2}(k_{j+1} + k_j).$$

$$D_y^+ U_j := \frac{U_{j+1} - U_j}{\hbar_j}, \quad D_y^- U_j := \frac{U_j - U_{j-1}}{k_j}, \quad D_y^+ D_y^- U_j := \frac{1}{\hbar_j}\left(\frac{U_{j+1} - U_j}{k_{j+1}} - \frac{U_j - U_{j-1}}{k_j}\right)$$

gives non-uniform mesh $\overline{\Omega}_h := \{(x_i, y_j) \in \overline{\Omega} : x_{i+1} - x_i = h_{i+1},\ y_{j+1} - y_j = k_{j+1}\},$



• Ω_h; ⊙ Γ_h, $\overline{\Omega}_h = \Omega_h \cup \Gamma_h$.
Figure 4: Nonuniform mesh $\overline{\Omega}_h$.

$$\begin{aligned}
-(D_x^+ D_x^- U_{i,j} + D_y^+ D_y^- U_{i,j}) &= f(x_i, y_j) \qquad \text{in } \Omega_h, \\
U_{i,j} &= 0 \qquad \text{on } \Gamma_h.
\end{aligned}$$

Proceed just as before

4.3 The discrete maximum principle (Regular max principle I know very well, see notes for more)

The Laplace operator, Δ, is approximated by $D_x^+D_x^- + D_y^+D_y^-$, with the difference operators $D_x^+D_x^-$, $D_y^+D_y^-$ defined as in Section 4.2. The finite difference approximation of the Dirichlet problem

$$-(D_x^+D_x^-U_{i,j} + D_y^+D_y^-U_{i,j}) = f(x_i, y_j) \qquad \text{in } \Omega_h, \quad U_{i,j} = g(x_i, y_j) \qquad \text{on } \Gamma_h. \quad (78)$$

Suppose that $f(x_i, y_j) < 0$ for all $(x_i, y_j) \in \Omega_h$ and that the maximum value of U is attained at a point $(x_{i_0}, y_{j_0}) \in \Omega_h$. Clearly,

$$\left(\frac{1}{\hbar_i}\left(\frac{1}{h_{i+1}} + \frac{1}{h_i}\right) + \frac{1}{\bar{\kappa}_j}\left(\frac{1}{k_{j+1}} + \frac{1}{k_j}\right)\right) U_{i,j} = \frac{U_{i+1,j}}{\hbar_i\, h_{i+1}} + \frac{U_{i-1,j}}{\hbar_i\, h_i} + \frac{U_{i,j+1}}{\bar{\kappa}_j\, k_{j+1}} + \frac{U_{i,j-1}}{\bar{\kappa}_j\, k_j} + f(x_i, y_j)$$

for any $(x_i, y_j) \in \Omega_h$. Therefore, because $U_{i_0\pm1,j_0} \le U_{i_0,j_0}$ and $U_{i_0,j_0\pm1} \le U_{i_0,j_0}$, and $f(x_{i_0}, y_{j_0}) < 0$, it follows that

$$\left(\frac{1}{\hbar_{i_0}}\left(\frac{1}{h_{i_0+1}} + \frac{1}{h_{i_0}}\right) + \frac{1}{\bar{\kappa}_{j_0}}\left(\frac{1}{k_{j_0+1}} + \frac{1}{k_{j_0}}\right)\right) U_{i_0j_0} < \frac{U_{i_0j_0}}{\hbar_{i_0}\, h_{i_0+1}} + \frac{U_{i_0j_0}}{\hbar_{i_0}\, h_{i_0}} + \frac{U_{i_0j_0}}{\bar{\kappa}_{j_0}\, k_{j_0+1}} + \frac{U_{i_0j_0}}{\bar{\kappa}_{j_0}\, k_{j_0}}.$$

Both sides here are equal ⇒ # for strict f < 0. For not strict define

$$V_{i,j} := U_{i,j} + \frac{\varepsilon}{4}(x_i^2 + y_j^2) \qquad \text{for } (x_i, y_j) \in \overline{\Omega}_h.$$

and sandwich like in traditional method

$$\boxed{\max_{(x_i,y_j)\in\Gamma_h} U_{i,j} = \max_{(x_i,y_j)\in\overline{\Omega}_h} U_{i,j}.} \qquad \boxed{\min_{(x_i,y_j)\in\Gamma_h} U_{i,j} = \min_{(x_i,y_j)\in\overline{\Omega}_h} U_{i,j}.}$$

(for f <= and f>= respectively)

4.4 Stability in the discrete maximum norm

Lemma 10 *The finite difference scheme (78) has a unique solution.*

assume 2, take difference (noting f = 0 for the new scheme)

We are now ready to embark on the analysis of the stability of the scheme (78) with respect to perturbations in the boundary data.

Consider the mesh functions $U^{(1)}$ and $U^{(2)}$, which satisfy, respectively:

$$-(D_x^+D_x^-U_{i,j}^{(1)} + D_y^+D_y^-U_{i,j}^{(1)}) = f(x_i, y_j) \qquad \text{in } \Omega_h, \quad U_{i,j}^{(1)} = g^{(1)}(x_i, y_j) \qquad \text{on } \Gamma_h \quad (80)$$

$$-(D_x^+D_x^-U_{i,j}^{(2)} + D_y^+D_y^-U_{i,j}^{(2)}) = f(x_i, y_j) \qquad \text{in } \Omega_h\, U_{i,j}^{(2)} = g^{(2)}(x_i, y_j) \qquad \text{on } \Gamma_h \quad (81)$$

Let $U := U^{(1)} - U^{(2)}$ and $g := g^{(1)} - g^{(2)}$. ⇒

$$-(D_x^+D_x^-U_{i,j} + D_y^+D_y^-U_{i,j}) = 0 \qquad \text{in } \Omega_h\, U_{i,j} = g(x_i, y_j) \qquad \text{on } \Gamma_h. \quad (82) \Rightarrow$$

for all $(x_i, y_j) \in \overline{\Omega}_h$, $U_{i,j} \le \max_{(x_i,y_j)\in\Gamma_h} |g(x_i, y_j)|$.

(discrete maximum principle) and same with min

$$|U_{i,j}| \le \max_{(x_i,y_j)\in\Gamma_h} |g(x_i, y_j)| \Rightarrow \max_{(x_i,y_j)\in\overline{\Omega}_h} |U_{i,j}| \le \max_{(x_i,y_j)\in\Gamma_h} |g(x_i, y_j)|. \Rightarrow$$

$$\max_{(x_i,y_j)\in\overline{\Omega}_h} |U_{i,j}^{(1)} - U_{i,j}^{(2)}| \le \max_{(x_i,y_j)\in\Gamma_h} |g^{(1)}(x_i, y_j) - g^{(2)}(x_i, y_j)|.$$

Continuous dependence/ stability

4.5 Iterative solution of linear systems: linear stationary iterative methods

Before embarking on our discussion of the main topic of this section, we require a few technical tools. Let us start by considering the finite difference approximation of the eigenvalue problem:

$$-u''(x) + c\,u(x) = \lambda u(x), \qquad x \in (0,1), \quad u(0) = 0, \quad u(1) = 0,$$ c >= 0 real number

A nontrivial solution $u(x) \not\equiv 0$ of this boundary-value problem is called an *eigenfunction*, corresponding lambda is eigenvalue. Here $u^k(x) := \sin(k\pi x)$ and $\lambda_k := c + k^2\pi^2, \quad k = 1, 2, \ldots$

The finite difference approximation of this eigenvalue problem on the mesh $\{x_i := ih : i = 0, \ldots, N\}$ of uniform spacing $h := 1/N$, with $N \geq 2$, is given by

$$-\frac{U_{i+1} - 2U_i + U_{i-1}}{h^2} + cU_i = \Lambda U_i, \quad i = 1, \ldots, N-1, \quad U_0 = 0, \quad U_N = 0.$$

Again, we seek nontrivial solutions, and a simple calculation yields that $U_i := U^k(x_i)$, where

$$U^k(x) := \sin(k\pi x), \quad x \in \{x_0, x_1, \ldots, x_N\} \quad \text{and} \quad \Lambda_k := c + \frac{4}{h^2}\sin^2\frac{k\pi h}{2}, \qquad k = 1, 2, \ldots, N-1.$$

Using matrix notation the finite difference approximation of the eigenvalue problem becomes

$$\begin{bmatrix} \frac{2}{h^2}+c & -\frac{1}{h^2} & & & 0 \\ -\frac{1}{h^2} & \frac{2}{h^2}+c & -\frac{1}{h^2} & & \\ & \ddots & \ddots & \ddots & \\ & & -\frac{1}{h^2} & \frac{2}{h^2}+c & -\frac{1}{h^2} \\ 0 & & & -\frac{1}{h^2} & \frac{2}{h^2}+c \end{bmatrix} \begin{bmatrix} U_1 \\ U_2 \\ \vdots \\ U_{N-2} \\ U_{N-1} \end{bmatrix} = \Lambda \begin{bmatrix} U_1 \\ U_2 \\ \vdots \\ U_{N-2} \\ U_{N-1} \end{bmatrix},$$

$$\begin{aligned} -\Delta u + cu &= \lambda u \quad \text{in } \Omega, \\ u &= 0 \quad \text{on } \Gamma := \partial\Omega, \end{aligned} \qquad \begin{aligned} -\Delta u + cu &= f(x,y) \quad \text{in } \Omega, \\ u &= 0 \quad \text{on } \Gamma := \partial\Omega, \end{aligned}$$ open square

Motivated by these examples, we shall be interested in developing a simple iterative method for the approximate solution of systems of linear algebraic equations of the form $AU = F$, where $A \in \mathbb{R}^{M\times M}$ is a symmetric matrix with positive eigenvalues, which are contained in a nonempty closed interval $[\alpha, \beta]$, with $0 < \alpha < \beta$, $U \in \mathbb{R}^M$ is the vector of unknowns and $F \in \mathbb{R}^M$ is a given vector. To this end, we consider the following iteration for the approximate solution of the linear system $AU = F$.

$$U^{(j+1)} := U^{(j)} - \tau(AU^{(j)} - F), \qquad j = 0, 1, \ldots, \tag{90}$$

Aim: find tau to converge fastest

where $U^{(0)} \in \mathbb{R}^M$ is a given initial guess, and $\tau > 0$ is a parameter to be chosen so as to ensure that the sequence of iterates $\{U^{(j)}\}_{j=0}^{\infty} \subset \mathbb{R}^M$ converges to $U \in \mathbb{R}^M$ as $j \to \infty$. We begin by observing that $U = U - \tau(AU - F)$. Therefore, upon subtraction of (90) from this equality we find that

$$U - U^{(j+1)} = U - U^{(j)} - \tau A(U - U^{(j)}) = (I - \tau A)(U - U^{(j)}), \qquad j = 0, 1, \ldots, \tag{91}$$

where $I \in \mathbb{R}^{M\times M}$ is the identity matrix. Consequently,

$$U - U^{(j)} = (I - \tau A)^j (U - U^{(0)}), \qquad j = 1, 2, \ldots.$$

induced matrix norm $\|B\| := \sup_{V\in\mathbb{R}^M\setminus\{0\}} \frac{\|BV\|}{\|V\|}$ $\|BV\| \leq \|B\|\|V\|$ for all $V \in \mathbb{R}^M$.

hence, by induction $\|B^jV\| \leq \|B\|^j\|V\| \Rightarrow$

$\|U - U^{(j)}\| = \|(I - \tau A)^j(U - U^{(0)})\| \leq \|I - \tau A\|^j\|U - U^{(0)}\|$

In order to continue, we need to bound $\|I - \tau A\|$, and to this end we need a few tools from linear algebra; we shall therefore make a brief detour. Our first observation is that $\mathbb{R}^M$ is a finite-dimensional linear space, and in a finite-dimensional linear spaces all norms are equivalent.[6] Therefore, if the sequence $\{U^{(j)}\}_{j=0}^{\infty}$ converges to U in one particular norm on $\mathbb{R}^M$, it will also converge to U in any other norm on $\mathbb{R}^M$. For the sake of simplicity of the exposition we shall therefore assume that the norm $\|\cdot\|$ on $\mathbb{R}^M$ appearing in the inequality above is the Euclidean norm:

$$\|V\| := \left(\sum_{i=1}^{M} V_i^2\right)^{1/2}, \qquad V = (V_1, \ldots, V_M)^{\mathrm{T}} \in \mathbb{R}^M.$$

A symmetric matrix $B \in \mathbb{R}^{M\times M}$ has real eigenvalues, and the associated set of orthonormal eigenvectors spans the whole of $\mathbb{R}^M$. Denoting by $\{e_i\}_{i=1}^{M}$ the (orthonormal) eigenvectors of B and by λ_i, $i = 1, \ldots, M$, the corresponding eigenvalues, for any vector $V = \alpha_1 e_1 + \cdots + \alpha_M e_M$, expanded in terms of the eigenvectors of B, thanks to orthonormality the Euclidean norms of V and BV can be expressed, respectively, as follows:

$$\|V\| = \left(\sum_{i=1}^{M} \alpha_i^2\right)^{1/2} \quad \text{and} \quad \|BV\| = \left(\sum_{i=1}^{M} \alpha_i^2 \lambda_i^2\right)^{1/2}.$$

Clearly, $\|BV\| \le \max_{i=1,\ldots,M} |\lambda_i| \, \|V\|$ for all $V \in \mathbb{R}^M$, and the inequality becomes an equality if V happens to be the eigenvector of B associated with the largest in absolute value eigenvalue of B. Therefore, $\|B\| = \max_{i=1,\ldots,M} |\lambda_i|$, where now $\|\cdot\|$ is the matrix norm induced by the Euclidean norm.

We are now ready to return to (92) to find that $\|I - \tau A\|$ appearing on the right-hand side of (92), where again $\|\cdot\|$ denotes the matrix norm induced by the Euclidean norm, is equal to the largest in absolute value eigenvalue of the symmetric matrix $I - \tau A$. As the eigenvalues of A are assumed to belong to the interval $[\alpha, \beta]$, where $0 < \alpha < \beta$, and the parameter τ is by assumption positive, the eigenvalues of $I - \tau A$ are contained in the interval $[1 - \tau\beta, 1 - \tau\alpha]$, whereby $\|I - \tau A\| \le \max\{|1 - \tau\beta|, |1 - \tau\alpha|\}$. As $\tau > 0$ is a free parameter, to be suitably chosen, we would like to select it so that the iterative method (90) converges as fast as possible, and to this end we see from (92) that it is desirable to choose τ so that $\|I - \tau A\|$ is as small as possible, and less than 1. We shall therefore seek $\tau > 0$ so as to ensure that

$$\min_{\tau>0} \max\{|1 - \tau\beta|, |1 - \tau\alpha|\} < 1.$$

By plotting the nonnegative piecewise linear functions $\tau \mapsto |1 - \tau\beta|$ and $\tau \mapsto |1 - \tau\alpha|$ for $\tau \in [0, \infty)$, we see that they vanish at $\tau = 1/\beta$ and $\tau = 1/\alpha$, respectively; their graphs intersect at $\tau = 0$ and at $\tau = \frac{2}{\alpha+\beta}$. As $0 < \alpha < \beta$, clearly $0 < 1/\beta < 1/\alpha$. Next, by plotting the continuous piecewise linear function $\tau \mapsto \max\{|1 - \tau\beta|, |1 - \tau\alpha|\}$ for $\tau \in [0, \infty)$, we observe that it attains its minimum at $\tau = \frac{2}{\alpha+\beta}$ where $1 - \tau\beta = \tau\alpha - 1$. Thus,

$$\min_{\tau>0} \max\{|1 - \tau\beta|, |1 - \tau\alpha|\} = \max\{|1 - \tau\beta|, |1 - \tau\alpha|\}|_{\tau=\frac{2}{\alpha+\beta}} = \frac{\beta - \alpha}{\beta + \alpha} < 1.$$

In summary then, the iterative method proposed for the approximate solution of the linear system $AU = F$ is the one stated in (90), with $\tau := \frac{2}{\beta+\alpha}$, and $[\alpha, \beta]$ being a closed subinterval of $(0, \infty)$ that contains all eigenvalues of the symmetric matrix $A \in \mathbb{R}^{M\times M}$.

5 Finite difference approximation of parabolic equations

(heat equation) in one space dimension: $$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2}, \quad (96)$$

which we shall consider for $x \in (-\infty, \infty)$ and $t \ge 0$, st $u(x, 0) = u_0(x), \qquad x \in (-\infty, \infty),$

where u_0 is a given function

Fourier transform of a function v: $$\hat{v}(\xi) = F[v](\xi) := \int_{-\infty}^{\infty} v(x)\, \mathrm{e}^{-\mathrm{i}x\xi}\, \mathrm{d}x.$$

We shall assume henceforth that the functions under consideration are sufficiently smooth and that they decay to 0 as $x \to \pm\infty$ sufficiently quickly in order to ensure that our formal manipulations make sense.

By Fourier-transforming the partial differential equation (96) we obtain

$$\int_{-\infty}^{\infty} \frac{\partial u}{\partial t}(x,t)\,\mathrm{e}^{-\imath x\xi}\,\mathrm{d}x = \int_{-\infty}^{\infty} \frac{\partial^2 u}{\partial x^2}(x,t)\,\mathrm{e}^{-\imath x\xi}\,\mathrm{d}x.$$

After (formal) integration by parts on the right-hand side and ignoring 'boundary terms' at $\pm\infty$, we obtain

$$\frac{\partial}{\partial t}\hat{u}(\xi,t) = (\imath\xi)^2\hat{u}(\xi,t), \quad \text{whereby} \quad \hat{u}(\xi,t) = \mathrm{e}^{-t\xi^2}\hat{u}(\xi,0), \quad \text{and so} \quad u(x,t) = F^{-1}\left(\mathrm{e}^{-t\xi^2}\hat{u}_0\right)$$

$$v(x) = F^{-1}[\hat{v}](x) := \frac{1}{2\pi}\int_{-\infty}^{\infty} \hat{v}(\xi)\mathrm{e}^{\imath x\xi}\,\mathrm{d}\xi.$$ (inverse Fourier transform) $\Rightarrow$ (after long calculations)

$$u(x,t) = F^{-1}\left(\mathrm{e}^{-t\xi^2}\hat{u}_0(\xi)\right) = \int_{-\infty}^{\infty} w(x-y,t)u_0(y)\,\mathrm{d}y, \quad w(x,t) := \frac{1}{\sqrt{4\pi t}}\mathrm{e}^{-x^2/(4t)},$$ *(heat kernel)*

So finally

$$u(x,t) = \frac{1}{\sqrt{4\pi t}}\int_{-\infty}^{\infty} \mathrm{e}^{-(x-y)^2/(4t)}u_0(y)\,\mathrm{d}y. \qquad (97)$$

This formula gives an explicit expression for the solution of the heat equation (96) in terms of the initial datum u_0. Because $w(x,t) > 0$ for all $x \in (-\infty,\infty)$ and all $t > 0$, and

$$\int_{-\infty}^{\infty} w(y,t)\,\mathrm{d}y = 1 \quad \text{for all } t > 0, \quad \sup_{x\in(-\infty,+\infty)}|u(x,t)| \le \sup_{x\in(-\infty,\infty)}|u_0(x)|, \quad t > 0.$$

(deduce from (97) that if u_0 is a bounded continuous function,)

In other words, the 'largest' and 'smallest' values of $u(\cdot,t)$ at $t > 0$ cannot exceed those of $u_0(\cdot)$

Lemma 11 (Parseval's identity) *Let $L_2(-\infty,\infty)$ denote the set of all complex-valued square-integrable functions defined on the real line. Suppose that $u \in L_2(-\infty,\infty)$. Then, $\hat{u} \in L_2(-\infty,\infty)$, and the following equality holds:*

$$\|u\|_{L_2(-\infty,\infty)} := \frac{1}{\sqrt{2\pi}}\|\hat{u}\|_{L_2(-\infty,\infty)}, \quad \text{where} \quad \|u\|_{L_2(-\infty,\infty)} = \left(\int_{-\infty}^{\infty}|u(x)|^2\,\mathrm{d}x\right)^{1/2}.$$

Proof $$\int_{-\infty}^{\infty}\hat{u}(\xi)\,v(\xi)\,\mathrm{d}\xi = \int_{-\infty}^{\infty}\left(\int_{-\infty}^{\infty}u(x)\,\mathrm{e}^{-\imath x\xi}\,\mathrm{d}x\right)v(\xi)\,\mathrm{d}\xi$$

$$= \int_{-\infty}^{\infty}\left(\int_{-\infty}^{\infty}v(\xi)\,\mathrm{e}^{-\imath x\xi}\,\mathrm{d}\xi\right)u(x)\,\mathrm{d}x = \int_{-\infty}^{\infty}u(x)\,\hat{v}(x)\,\mathrm{d}x. \quad \text{take} \quad v(\xi) = \overline{\hat{u}(\xi)} = 2\pi F^{-1}[\overline{u}](\xi)$$

(96), we thus have by Parseval's identity that $\|u(\cdot,t)\|_{L_2(-\infty,\infty)} = \frac{1}{\sqrt{2\pi}}\|\hat{u}(\cdot,t)\|_{L_2(-\infty,\infty)}, \quad t > 0$

and therefore $\|u(\cdot,t)\|_{L_2(-\infty,\infty)} = \frac{1}{\sqrt{2\pi}}\|\mathrm{e}^{-t\xi^2}\hat{u}_0(\cdot)\|_{L_2(-\infty,\infty)} \le \frac{1}{\sqrt{2\pi}}\|\hat{u}_0\|_{L_2(-\infty,\infty)}$

$$= \|u_0\|_{L_2(-\infty,\infty)} \Rightarrow \|u(\cdot,t)\|_{L_2(-\infty,\infty)} \le \|u_0\|_{L_2(-\infty,\infty)} \quad \text{for all } t > 0 \quad (99)$$

This is a useful result as it can be used to deduce stability (wrt perturbations of datum)

Suppose perturbate … $$\|u(\cdot,t) - \tilde{u}(\cdot,t)\|_{L_2(-\infty,\infty)} \le \|u_0 - \tilde{u}_0\|_{L_2(-\infty,\infty)}$$

5.1 Finite difference approximation of the heat equation

We take our computational domain to be $\{(x,t) \in (-\infty,\infty)\times[0,T]\}$

$T > 0$ is a given final time., spacing $\Delta x > 0$, spacing $\Delta t := T/M$ in the t-direction.

$x_j := j\Delta x \quad t_m := m\Delta t,$ approximate derivatives of u accordingly,

$$\frac{U_j^{m+1} - U_j^m}{\Delta t} = \frac{U_{j+1}^m - 2U_j^m + U_{j-1}^m}{(\Delta x)^2}, \quad j = 0, \pm 1, \pm 2, \ldots, \quad m = 0, \ldots, M-1,$$

$U_j^0 := u_0(x_j), \qquad j = 0, \pm 1, \pm 2, \ldots$ Equivalently, we can write this as

$$U_j^{m+1} = U_j^m + \mu(U_{j+1}^m - 2U_j^m + U_{j-1}^m), \quad j = 0, \pm 1, \pm 2, \ldots, \quad m = 0, \ldots, M-1,$$

$$U_j^0 := u_0(x_j), \qquad j = 0, \pm 1, \pm 2, \ldots, \quad \text{where } \mu = \tfrac{\Delta t}{(\Delta x)^2}.$$

Alternatively, if instead of time level m the expression on the right-hand side of the explicit Euler scheme is evaluated on the time level $m+1$, we arrive at the *implicit Euler scheme*:

$$\frac{U_j^{m+1} - U_j^m}{\Delta t} = \frac{U_{j+1}^{m+1} - 2U_j^{m+1} + U_{j-1}^{m+1}}{(\Delta x)^2}, \quad j = 0, \pm 1, \pm 2, \ldots, \quad m = 0, \ldots, M-1,$$

$$U_j^0 := u_0(x_j), \qquad j = 0, \pm 1, \pm 2, \ldots .$$

The explicit and implicit Euler schemes are special cases of a more general one-parameter family of numerical methods for the heat equation, called the *θ-method*, which is a convex combination of the two Euler schemes, with a parameter $\theta \in [0, 1]$. The θ-method is defined as follows:

$$\frac{U_j^{m+1} - U_j^m}{\Delta t} = (1-\theta)\frac{U_{j+1}^m - 2U_j^m + U_{j-1}^m}{(\Delta x)^2} + \theta\frac{U_{j+1}^{m+1} - 2U_j^{m+1} + U_{j-1}^{m+1}}{(\Delta x)^2}, \quad \begin{cases} j = 0, \pm 1, \pm 2, \ldots, \\ m = 0, \ldots, M-1, \end{cases}$$

$$U_j^0 := u_0(x_j), \qquad j = 0, \pm 1, \pm 2, \ldots,$$

where $\theta \in [0, 1]$ is a parameter. For $\theta = 0$ the θ-scheme coincides with the explicit Euler scheme, for $\theta = 1$ it is the implicit Euler scheme, and for $\theta = 1/2$ it is the arithmetic average of the two Euler schemes, and is called the *Crank–Nicolson scheme*.

Numerical methods of this kind are called *fully-discrete approximations*. An alternative approach is to approximate the spatial partial derivative only in the heat equation, resulting in the following initial-value problem for a system of ordinary differential equations:

$$\frac{\mathrm{d}U_j(t)}{\mathrm{d}t} = \frac{U_{j+1}(t) - 2U_j(t) + U_{j-1}(t)}{(\Delta x)^2}, \qquad j = 0, \pm 1, \pm 2, \ldots,$$
$$U_j(0) := u_0(x_j), \qquad j = 0, \pm 1, \pm 2, \ldots .$$

This is called a spatially semi-discrete approximation, because no discretization with respect to the

5.1.1 Accuracy of the θ-method

$$T_j^m := \frac{u_j^{m+1} - u_j^m}{\Delta t} - (1-\theta)\frac{u_{j+1}^m - 2u_j^m + u_{j-1}^m}{(\Delta x)^2} - \theta\frac{u_{j+1}^{m+1} - 2u_j^{m+1} + u_{j-1}^{m+1}}{(\Delta x)^2}, \quad \begin{cases} j = 0, \pm 1, \pm 2, \ldots, \\ m = 0, \ldots, M-1, \end{cases}$$

$u_j^m := u(x_j, t_m).$ Taylor expand…

$$T_j^m = \begin{cases} \mathcal{O}\left((\Delta x)^2 + (\Delta t)^2\right) & \text{for } \theta = 1/2, \\ \mathcal{O}\left((\Delta x)^2 + \Delta t\right) & \text{for } \theta \neq 1/2. \end{cases}$$

5.2 Stability of finite difference schemes

We shall say that a finite difference scheme for the unsteady heat equation is

(practically) stable in the ℓ_2 norm, if $\|U^m\|_{\ell_2} \leq \|U^0\|_{\ell_2}, \qquad m = 1, \ldots, M,$

$$\|U^m\|_{\ell_2} := \left(\Delta x \sum_{j=-\infty}^{\infty} |U_j^m|^2\right)^{1/2}$$

Definition 2 *The semidiscrete Fourier transform of a function U defined on the infinite mesh with mesh-points $x_j = j\Delta x$, $j = 0, \pm 1, \pm 2, \ldots$, is:*

$$\hat{U}(k) := \Delta x \sum_{j=-\infty}^{\infty} U_j \, \mathrm{e}^{-\mathrm{i}kx_j}, \qquad k \in [-\pi/\Delta x, \pi/\Delta x].$$

Lemma 12 (Discrete Parseval's identity) *Let*

$$\|U\|_{\ell_2} := \left(\Delta x \sum_{j=-\infty}^{\infty} |U_j|^2 \right)^{1/2} \quad \text{and} \quad \|\hat{U}\|_{L_2} := \left(\int_{-\pi/\Delta x}^{\pi/\Delta x} |\hat{U}(k)|^2 \, \mathrm{d}k \right)^{1/2}$$

If $\|U\|_{\ell_2}$ *is finite, then so is* $\|\hat{U}\|_{L_2}$, *and* $\|U\|_{\ell_2} = \frac{1}{\sqrt{2\pi}} \|\hat{U}\|_{L_2}$. (similar proof to lemma 11)

5.2.1 Stability analysis of the explicit Euler scheme

By inserting $U_j^m = \frac{1}{2\pi} \int_{-\pi/\Delta x}^{\pi/\Delta x} \mathrm{e}^{\imath k j \Delta x} \hat{U}^m(k) \, \mathrm{d}k$ into the explicit Euler scheme $\Rightarrow$

$$\frac{1}{2\pi} \int_{-\pi/\Delta x}^{\pi/\Delta x} \mathrm{e}^{\imath k j \Delta x} \frac{\hat{U}^{m+1}(k) - \hat{U}^m(k)}{\Delta t} \, \mathrm{d}k = \frac{1}{2\pi} \int_{-\pi/\Delta x}^{\pi/\Delta x} \frac{\mathrm{e}^{\imath k (j+1) \Delta x} - 2\mathrm{e}^{\imath k j \Delta x} + \mathrm{e}^{\imath k (j-1) \Delta x}}{(\Delta x)^2} \hat{U}^m(k) \, \mathrm{d}k$$

$$\Rightarrow \quad \frac{1}{2\pi} \int_{-\pi/\Delta x}^{\pi/\Delta x} \mathrm{e}^{\imath k j \Delta x} \frac{\hat{U}^{m+1}(k) - \hat{U}^m(k)}{\Delta t} \, \mathrm{d}k = \frac{1}{2\pi} \int_{-\pi/\Delta x}^{\pi/\Delta x} \mathrm{e}^{\imath k j \Delta x} \frac{\mathrm{e}^{\imath k \Delta x} - 2 + \mathrm{e}^{-\imath k \Delta x}}{(\Delta x)^2} \hat{U}^m(k) \, \mathrm{d}k.$$

deduce that the two integrands are identicall $\Rightarrow$

$\hat{U}^{m+1}(k) = \hat{U}^m(k) + \mu(\mathrm{e}^{\imath k \Delta x} - 2 + \mathrm{e}^{-\imath k \Delta x}) \hat{U}^m(k) \Rightarrow \quad \hat{U}^{m+1}(k) = \lambda(k) \hat{U}^m(k),$

$\lambda(k) := 1 + \mu(\mathrm{e}^{\imath k \Delta x} - 2 + \mathrm{e}^{-\imath k \Delta x})$ is the *amplification factor* and $\mu := \frac{\Delta t}{(\Delta x)^2}$ *CFL number*

By the discrete Parseval identity $\|U^{m+1}\|_{\ell_2} = \frac{1}{\sqrt{2\pi}} \|\hat{U}^{m+1}\|_{L_2}$ ... $\leq \frac{1}{\sqrt{2\pi}} \max_k |\lambda(k)| \, \|\hat{U}^m\|_{L_2}$

$= \max_k |\lambda(k)| \, \|U^m\|_{\ell_2}$. so we demand $\max_k |\lambda(k)| \leq 1,$ because

In order to mimic the bound (99) we would like to ensure that $\|U^{m+1}\|_{\ell_2} \leq \|U^m\|_{\ell_2}$,

i.e., that $\max_k |1 + \mu(\mathrm{e}^{\imath k \Delta x} - 2 + \mathrm{e}^{-\imath k \Delta x})| \leq 1.$

...

$-1 \leq 1 - 4\mu \sin^2\left(\frac{k \Delta x}{2}\right) \leq 1 \quad \forall k \in [-\pi/\Delta x, \pi/\Delta x].$ holds if, and only if, $\mu = \frac{\Delta t}{(\Delta x)^2} \leq \frac{1}{2}$

Theorem 12 *Suppose that* U_j^m *is the solution of the explicit Euler scheme*

$$\frac{U_j^{m+1} - U_j^m}{\Delta t} = \frac{U_{j+1}^m - 2U_j^m + U_{j-1}^m}{(\Delta x)^2}, \quad j = 0, \pm 1, \pm 2, \ldots, \quad m = 0, \ldots, M-1,$$

$U_j^0 := u_0(x_j), \quad j = 0, \pm 1, \pm 2, \ldots,$ *and* $\mu = \frac{\Delta t}{(\Delta x)^2} \leq \frac{1}{2}$. *Then,*

$\|U^m\|_{\ell_2} \leq \|U^0\|_{\ell_2}, \quad m = 1, 2, \ldots, M.$ (101)

In other words the explicit Euler scheme is *conditionally practically stable*,

One can also show that if μ > 1/2, then (101) will fail.

5.2.2 Stability analysis of the implicit Euler scheme

We shall now perform a similar analysis for the *implicit Euler scheme* for the heat equation (96), which is defined as follows:

$$\frac{U_j^{m+1} - U_j^m}{\Delta t} = \frac{U_{j+1}^{m+1} - 2U_j^{m+1} + U_{j-1}^{m+1}}{(\Delta x)^2}, \quad j = 0, \pm 1, \pm 2, \ldots, \quad m = 0, \ldots, M-1,$$

(equally $U_j^{m+1} - \mu(U_{j+1}^{m+1} - 2U_j^{m+1} + U_{j-1}^{m+1}) = U_j^m$,) $U_j^0 := u_0(x_j), \qquad j = 0, \pm 1, \pm 2, \ldots$)

$\mu := \dfrac{\Delta t}{(\Delta x)^2}$. Using an identical argument as for the explicit Euler scheme,

$\lambda(k) := \dfrac{1}{1 + 4\mu \sin^2\left(\frac{k\Delta x}{2}\right)}$ $\Rightarrow$ $\max_k |\lambda(k)| \le 1$ $\Rightarrow$

Theorem 13 *Suppose that U_j^m is the solution of the implicit Euler scheme*

$$\frac{U_j^{m+1} - U_j^m}{\Delta t} = \frac{U_{j+1}^{m+1} - 2U_j^{m+1} + U_{j-1}^{m+1}}{(\Delta x)^2}, \quad j = 0, \pm 1, \pm 2, \ldots, \quad m = 0, \ldots, M-1,$$

$$U_j^0 := u_0(x_j), \qquad j = 0, \pm 1, \pm 2, \ldots.$$

Then, for all $\Delta t > 0$ and $\Delta x > 0$,

$$\|U^m\|_{\ell_2} \le \|U^0\|_{\ell_2}, \qquad m = 1, 2, \ldots, M. \tag{102}$$

In other words, the implicit Euler scheme is *unconditionally practically stable*, meaning that the bound (102) holds without any restrictions on Δx and Δt.

5.3 Von Neumann stability

Definition 4 *We shall say that a finite difference scheme for the unsteady heat equation on the time interval $[0, T]$ is von Neumann stable in the ℓ_2 norm, if there exists a positive constant $C = C(T)$ such that*

$$\|U^m\|_{\ell_2} \le C\|U^0\|_{\ell_2}, \qquad m = 1, \ldots, M = \frac{T}{\Delta t} \quad \textit{where} \quad \|U^m\|_{\ell_2} = \left(\Delta x \sum_{j=-\infty}^{\infty} |U_j^m|^2\right)^{1/2}$$

Clearly, practical stability implies von Neumann stability, with stability constant $C = 1$.

As C dependent on T, it only makes sense for finite time intervals

Lemma 13 *Suppose that the semidiscrete Fourier transform of the solution $\{U_j^m\}_{j=-\infty}^{\infty}$, $m = 0, 1, \ldots, \frac{T}{\Delta t}$, of a finite difference scheme for the heat equation satisfies*

$$\hat{U}^{m+1}(k) = \lambda(k)\hat{U}^m(k)$$

and there exists a nonnegative constant C_0 such that

$$|\lambda(k)| \le 1 + C_0 \Delta t \qquad \forall\, k \in [-\pi/\Delta x, \pi/\Delta x].$$

Then the scheme is von Neumann stable. In particular, if $C_0 = 0$ then the scheme is practically stable.

PROOF: By Parseval's identity for the semidiscrete Fourier transform we have that

$$\|U^{m+1}\|_{\ell_2} = \frac{1}{\sqrt{2\pi}}\|\hat{U}^{m+1}\|_{L_2} = \frac{1}{\sqrt{2\pi}}\|\lambda\hat{U}^m\|_{L_2} \le \frac{1}{\sqrt{2\pi}} \max_k |\lambda(k)|\ \|\hat{U}^m\|_{L_2} = \max_k |\lambda(k)|\ \|U^m\|_{\ell_2}$$

$\Rightarrow$ $\|U^{m+1}\|_{\ell_2} \le (1 + C_0\Delta t)\|U^m\|_{\ell_2}$, $\Rightarrow$ $\|U^m\|_{\ell_2} \le (1 + C_0\Delta t)^m\|U^0\|_{\ell_2}$ $\Rightarrow$ $\|U^m\|_{\ell_2} \le \mathrm{e}^{C_0 T}\|U^0\|_{\ell_2}$,

meaning that von Neumann stability holds, with stability constant $C = \mathrm{e}^{C_0 T}$. $C_0 = 0 \Rightarrow C = 1$

5.4 Stability of the θ-scheme

To analyze the practical stability of the θ-scheme in the ℓ_2 norm, we shall use Lemma 13 with $C_0 = 0$.

Suppose that $U_j^m = [\lambda(k)]^m \mathrm{e}^{\mathrm{i}kx_j}$.

Substitution of this 'Fourier mode' into the θ-scheme gives the equality

$$\lambda(k) - 1 = -4(1-\theta)\,\mu \sin^2\left(\frac{k\Delta x}{2}\right) - 4\theta\,\mu\,\lambda(k)\sin^2\left(\frac{k\Delta x}{2}\right) \Rightarrow$$

$$\lambda(k) = \frac{1 - 4(1-\theta)\mu \sin^2\left(\frac{k\Delta x}{2}\right)}{1 + 4\theta\mu \sin^2\left(\frac{k\Delta x}{2}\right)}.$$ For practical stability, we demand that $|\lambda(k)| \leq 1$ which holds if, and only if, $2(1-2\theta)\mu \leq 1$. Thus we have shown that:

- For $\theta \in [1/2, 1]$ the θ-scheme is *unconditionally practically stable*;
- For $\theta \in [0, 1/2)$ the θ-scheme is *conditionally practically stable*, the stability condition being that

$$\mu \leq \frac{1}{2(1-2\theta)}.$$

5.5 Boundary-value problems for parabolic problems

When a parabolic partial differential equation is considered on a bounded spatial domain, one needs to impose boundary conditions at the boundary of the domain. Here we shall concentrate on the simplest

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2}, \qquad a < x < b, \quad 0 < t \leq T, \quad u(x,0) = u_0(x), \qquad x \in [a,b],$$

$$u(a,t) = A(t), \quad u(b,t) = B(t), \qquad t \in (0,T].$$

5.5.1 θ-scheme for the Dirichlet initial-boundary-value problem

We approximate the Dirichlet initial-boundary-value problem with the following θ-scheme:

$$\frac{U_j^{m+1} - U_j^m}{\Delta t} = (1-\theta)\frac{U_{j+1}^m - 2U_j^m + U_{j-1}^m}{(\Delta x)^2} + \theta\frac{U_{j+1}^{m+1} - 2U_j^{m+1} + U_{j-1}^{m+1}}{(\Delta x)^2},$$

for $j = 1, \ldots, J-1$, $m = 0, \ldots, M-1$,

$$U_j^0 := u_0(x_j), \qquad j = 0, \ldots, J, \quad U_0^{m+1} := A(t_{m+1}), \quad U_J^{m+1} := B(t_{m+1}), \quad m = 0, \ldots, M-1.$$

In order to implement this scheme it is helpful to rewrite it as a system of linear algebraic equations to compute the values of the approximate solution on time-level $m+1$ from those on time-level m. We have that

$$\begin{aligned} [1 - \theta\mu\delta^2]U_j^{m+1} &= [1 + (1-\theta)\mu\delta^2]U_j^m, \quad j = 1, \ldots, J-1, \quad m = 0, \ldots, M-1, \\ U_j^0 &:= u_0(x_j), \qquad j = 0, \ldots, J, \\ U_0^{m+1} &:= A(t_{m+1}), \quad U_J^{m+1} := B(t_{m+1}), \qquad m = 0, \ldots, M-1, \end{aligned}$$

$\delta^2 U_j := U_{j+1} - 2U_j + U_{j-1}$.

θ-scheme can be written as

$$(\mathcal{I} - \theta\mu\mathcal{A})\mathbf{U}^{m+1} = (\mathcal{I} + (1-\theta)\mu\mathcal{A})\mathbf{U}^m + \theta\mu\mathbf{F}^{m+1} + (1-\theta)\mu\mathbf{F}^m$$

where

$$\mathcal{A} := \begin{pmatrix} -2 & 1 & 0 & 0 & 0 & \ldots & 0 & 0 & 0 \\ 1 & -2 & 1 & 0 & 0 & \ldots & 0 & 0 & 0 \\ 0 & 1 & -2 & 1 & 0 & \ldots & 0 & 0 & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & 0 & 0 & 0 & \ldots & 1 & -2 & 1 \\ 0 & 0 & 0 & 0 & 0 & \ldots & 0 & 1 & -2 \end{pmatrix}$$

$$\mathbf{U}^m := (U_1^m, U_2^m, \ldots, U_{J-2}^m, U_{J-1}^m)^{\mathrm{T}} \qquad \mathbf{F}^m := (A(t_m), 0, \ldots, 0, B(t_m))^{\mathrm{T}}.$$

and

Thus, for each $m = 0, \ldots, M-1$, we are required to solve a system of linear algebraic equations with (the same) tridiagonal matrix $\mathcal{I} - \theta\mu\mathcal{A}$ in order to compute $\mathbf{U}^{m+1}$ from $\mathbf{U}^m$.

5.5.2 The discrete maximum principle

Theorem 14 (Discrete maximum principle for the θ-scheme)
The θ-scheme for the Dirichlet initial-boundary-value problem for the heat equation, with $0 \leq \theta \leq 1$ and $\mu(1-\theta) \leq \frac{1}{2}$, yields a sequence of numerical approximations $\{U_j^m\}_{j=0,\ldots,J;\ m=0,\ldots,M}$ satisfying

$$U_{\min} \leq U_j^m \leq U_{\max}$$

where

$$U_{\min} := \min\left\{\min\{U_0^m\}_{m=0}^M,\ \min\{U_j^0\}_{j=0}^J,\ \min\{U_J^m\}_{m=0}^M\right\}$$

and

$$U_{\max} := \max\left\{\max\{U_0^m\}_{m=0}^M,\ \max\{U_j^0\}_{j=0}^J,\ \max\{U_J^m\}_{m=0}^M\right\}.$$

PROOF: We rewrite the θ-scheme as

$$(1+2\theta\mu)\,U_j^{m+1} = \theta\mu\left(U_{j+1}^{m+1}+U_{j-1}^{m+1}\right) + (1-\theta)\mu\left(U_{j+1}^{m}+U_{j-1}^{m}\right) + [1-2(1-\theta)\mu]\,U_j^m, \qquad (103)$$

and recall that, by hypothesis, $\theta\mu \geq 0 \qquad (1-\theta)\mu \geq 0, \qquad 1-2(1-\theta)\mu \geq 0.$

Suppose that U attains its maximum value at an interior mesh-point U_j^{m+1}, $1 \leq j \leq J-1,\ 0 \leq m \leq M-1$. If this is not the case, the proof is complete. We define

$$U^\star = \max\{U_{j+1}^{m+1}, U_{j-1}^{m+1}, U_{j+1}^{m}, U_{j-1}^{m}, U_j^m\}.$$

Upper bound our equation, will reach an equality, implying all neighbours are also maxima.
Induction… eventually hit boundary

In summary then, for

$$\mu(1-\theta) \leq \frac{1}{2} \quad \text{with } \theta \in [0,1]$$

the θ-scheme satisfies the discrete maximum principle. Clearly, this condition is more demanding than the ℓ_2-stability condition:

$$\mu(1-2\theta) \leq \frac{1}{2} \qquad \text{for} \qquad 0 \leq \theta \leq \tfrac{1}{2}.$$

E.g Crank Nicholson scheme

5.5.3 Convergence analysis of the θ-scheme in the maximum norm

We close our discussion of finite difference schemes for the heat equation (96) in one space-dimension with the convergence analysis of the θ-scheme for the Dirichlet initial-boundary-value problem. We begin by rewriting the scheme as follows:

$$(1+2\theta\mu)\,U_j^{m+1} = \theta\mu\left(U_{j+1}^{m+1}+U_{j-1}^{m+1}\right) + (1-\theta)\mu\left(U_{j+1}^{m}+U_{j-1}^{m}\right) + [1-2(1-\theta)\mu]\,U_j^m,$$

Define consistency and global error as expected, notice

$$(1+2\theta\mu)\,e_j^{m+1} = \theta\mu\left(e_{j+1}^{m+1}+e_{j-1}^{m+1}\right) + (1-\theta)\mu\left(e_{j+1}^{m}+e_{j-1}^{m}\right) + [1-2(1-\theta)\mu]\,e_j^m + \Delta t\,T_j^m.$$

$$E^m := \max_{0\leq j\leq J}|e_j^m| \quad \text{and} \quad T^m := \max_{0\leq j\leq J}|T_j^m|. \qquad \theta\mu \geq 0, \quad (1-\theta)\mu \geq 0, \quad 1-2(1-\theta)\mu \geq 0,$$

$$\Rightarrow (1+2\theta\mu)E^{m+1} \leq 2\theta\mu E^{m+1} + E^m + \Delta t T^m, \qquad m = 0,\ldots,M-1. \Rightarrow$$

$$E^{m+1} \leq E^m + \Delta t\,T^m, \text{ As } E^0 = 0, \text{ upon summation, } E^m \leq \Delta t \sum_{n=0}^{m-1} T^n \leq m\Delta t \max_{0\leq n\leq m-1} T^n$$

$$\leq T \max_{0\leq m\leq M-1}\ \max_{1\leq j\leq J-1}|T_j^m|, \qquad m = 1,\ldots,M, \Rightarrow$$

$$\max_{0\leq m\leq M}\ \max_{0\leq j\leq J}|u(x_j,t_m) - U_j^m| \leq T \max_{0\leq m\leq M-1}\ \max_{1\leq j\leq J-1}|T_j^m|.$$ (know order from scheme type)

5.6 Finite difference approximation of parabolic equations in two space-dime

Consider the heat equation $\dfrac{\partial u}{\partial t} = \dfrac{\partial^2 u}{\partial x^2} + \dfrac{\partial^2 u}{\partial y^2}, \qquad (x,y) \in \Omega := (a,b)\times(c,d),\ t \in (0,T],$

subject to the initial condition $u(x,y,0) = u_0(x,y), \qquad (x,y) \in [a,b]\times[c,d],$

and the Dirichlet boundary condition $u|_{\partial\Omega} = B(x,y,t), \qquad (x,y) \in \partial\Omega,\ t \in (0,T],$

5.6.1 The explicit Euler scheme

$\delta_x^2 U_{i,j} := U_{i+1,j} - 2U_{i,j} + U_{i-1,j}$,

$\delta_y^2 U_{i,j} := U_{i,j+1} - 2U_{i,j} + U_{i,j-1}$. $x_i := a + i\Delta x$, $y_j := c + j\Delta y$, $t_m := m\Delta t$,

$\Delta x := (b-a)/J_x$, $\Delta y := (d-c)/J_y$, $\Delta t := T/M$ $\Rightarrow$ $$\frac{U_{i,j}^{m+1} - U_{i,j}^m}{\Delta t} = \frac{\delta_x^2 U_{i,j}^m}{(\Delta x)^2} + \frac{\delta_y^2 U_{i,j}^m}{(\Delta y)^2},$$

5.6.2 The implicit Euler scheme

$$\frac{U_{i,j}^{m+1} - U_{i,j}^m}{\Delta t} = \frac{\delta_x^2 U_{i,j}^{m+1}}{(\Delta x)^2} + \frac{\delta_y^2 U_{i,j}^{m+1}}{(\Delta y)^2},$$

5.6.3 The θ-scheme

$$\frac{U_{i,j}^{m+1} - U_{i,j}^m}{\Delta t} = (1-\theta)\left(\frac{\delta_x^2 U_{i,j}^m}{(\Delta x)^2} + \frac{\delta_y^2 U_{i,j}^m}{(\Delta y)^2}\right) + \theta\left(\frac{\delta_x^2 U_{i,j}^{m+1}}{(\Delta x)^2} + \frac{\delta_y^2 U_{i,j}^{m+1}}{(\Delta y)^2}\right)$$

$U_{i,j}^m = [\lambda(k_x,k_y)]^m \, \mathrm{e}^{\imath(k_x x_i + k_y y_j)}$

$$\lambda - 1 = -4(1-\theta)\left[\mu_x \sin^2\left(\frac{k_x\Delta x}{2}\right) + \mu_y \sin^2\left(\frac{k_y\Delta y}{2}\right)\right] - 4\theta\lambda\left[\mu_x \sin^2\left(\frac{k_x\Delta x}{2}\right) + \mu_y \sin^2\left(\frac{k_y\Delta y}{2}\right)\right]$$

$$\mu_x := \frac{\Delta t}{(\Delta x)^2}, \qquad \mu_y := \frac{\Delta t}{(\Delta y)^2}. \qquad \lambda = \lambda(k_x,k_y) = \frac{1 - 4(1-\theta)\left[\mu_x \sin^2\left(\frac{k_x\Delta x}{2}\right) + \mu_y \sin^2\left(\frac{k_y\Delta y}{2}\right)\right]}{1 + 4\theta\left[\mu_x \sin^2\left(\frac{k_x\Delta x}{2}\right) + \mu_y \sin^2\left(\frac{k_y\Delta y}{2}\right)\right]}$$

For practical stability in the ℓ_2 norm, $|\lambda(k_x,k_y)| \le 1$ so need

$$-1 \le \frac{1 - 4(1-\theta)\left[\mu_x + \mu_y\right]}{1 + 4\theta\left[\mu_x + \mu_y\right]} \le 1, \qquad 2(1-2\theta)(\mu_x + \mu_y) \le 1.$$

i.e

Under a suitable condition the θ-scheme for the initial-boundary-value problem also satisfies a discrete maximum principle. To see this, we rewrite the θ-scheme as

$$(1 + 2\theta(\mu_x+\mu_y))U_{i,j}^{m+1} = (1 - 2(1-\theta)(\mu_x+\mu_y))U_{i,j}^m + (1-\theta)\mu_x(U_{i+1,j}^m + U_{i-1,j}^m) + (1-\theta)\mu_y(U_{i,j+1}^m + U_{i,j-1}^m)$$
$$+ \theta\mu_x(U_{i+1,j}^{m+1} + U_{i-1,j}^{m+1}) + \theta\mu_y(U_{i,j+1}^{m+1} + U_{i,j-1}^{m+1}),$$

Theorem 15 *Suppose that* $(\mu_x + \mu_y)(1-\theta) \le \frac{1}{2}, \qquad \theta \in [0,1]$

Then, the θ-scheme satisfies the following discrete maximum principle: $U_{\min} \le U_{i,j}^m \le U_{\max}$,

where $U_{\min} := \min\left\{\min\{U_{i,j}^0\}_{i,j=0}^{J_x,J_y},\ \min\{U_{i,j}^m\}_{m=0}^{M}|_{(x_i,y_j)\in\partial\Omega}\right\}$ and

$U_{\max} := \max\left\{\max\{U_{i,j}^0\}_{i,j=0}^{J_x,J_y},\ \max\{U_{i,j}^m\}_{m=0}^{M}|_{(x_i,y_j)\in\partial\Omega}\right\}$

PROOF: The proof proceeds by an obvious modification of the proof of the discrete maximum principle for the θ-scheme in one space-dimension. □

In summary, then, for

$$(\mu_x + \mu_y)(1-\theta) \le \frac{1}{2}$$

the θ-scheme satisfies the discrete maximum principle. This condition is more demanding than the one for the practical stability of the scheme in the ℓ_2 norm, which requires that

$$(\mu_x + \mu_y)(1-2\theta) \le \frac{1}{2} \qquad \text{for} \qquad 0 \le \theta \le \tfrac{1}{2}.$$

We close our discussion by returning to the θ-scheme for the initial-boundary-value problem, and discussing its error analysis. The starting point is to rewrite the scheme as follows:

$$(1+2\theta(\mu_x+\mu_y))U_{i,j}^{m+1} = (1-2(1-\theta)(\mu_x+\mu_y))U_{i,j}^{m}$$

$$+(1-\theta)\mu_x(U_{i+1,j}^{m}+U_{i-1,j}^{m})+(1-\theta)\mu_y(U_{i,j+1}^{m}+U_{i,j-1}^{m}) \quad +\theta\mu_x(U_{i+1,j}^{m+1}+U_{i-1,j}^{m+1})+\theta\mu_y(U_{i,j+1}^{m+1}+U_{i,j-1}^{m+1})$$

$U_{i,j}^{m} := B(x_i, y_j, t_m)$, at the boundary mesh-points, for $m = 1, \ldots, M$.

By performing some elementary but tedious Taylor series expansions, one can deduce that

$$T_{i,j}^{m} = \begin{cases} \mathcal{O}\left((\Delta x)^2+(\Delta y)^2+(\Delta t)^2\right) & \theta = 1/2, \\ \mathcal{O}\left((\Delta x)^2+(\Delta y)^2+\Delta t\right) & \theta \neq 1/2. \end{cases}$$

$$(1+2\theta(\mu_x+\mu_y))E^{m+1} \leq 2\theta(\mu_x+\mu_y)E^{m+1}+E^{m}+\Delta t T^{m} \Rightarrow \quad E^{m+1} \leq E^{m}+\Delta t\, T^{m}$$

As $E^{0} = 0$, upon summation we deduce that $E^{m} \leq \Delta t \sum_{}^{m-1} T^{n}$

$$\leq T \max_{0\leq m\leq M} \max_{1\leq i\leq J_x-1,\, 1\leq j\leq J_y-1} |T_{i,j}^{m}|, \Rightarrow$$

$$\max_{0\leq i\leq J_x, 0\leq j\leq J_y} \max_{0\leq m\leq M} |u(x_i,y_j,t_m)-U_{i,j}^{m}| \leq T \max_{1\leq i\leq J_x-1,\, 1\leq j\leq J_y-1} \max_{0\leq m\leq M} |T_{i,j}^{m}|.$$

6 Finite difference approximation of hyperbolic equations

In this section we shall be concerned with the finite difference approximation of the simplest example of a second-order linear hyperbolic equation, the linear wave equation

$$\frac{\partial^2 u}{\partial t^2} - c^2\frac{\partial^2 u}{\partial x^2} = f(x,t),$$

$c > 0$ is the wave speed and f is a given source term

In the special case when f is identically zero and the equation is considered on the whole real line, $-\infty < x < \infty$, by supplying two initial conditions

$$u(x,0) = u_0(x) \quad \text{for } x \in \mathbb{R}, \quad \frac{\partial u}{\partial t}(x,0) = u_1(x) \quad \text{for } x \in \mathbb{R},$$

where u_0 and u_1 are defined on $\mathbb{R}$, u_0 is twice continuously differentiable and u_1 is once continuously differentiable on $\mathbb{R}$, the solution is given by d'Alembert's formula

$$u(x,t) = \frac{1}{2}\left[u_0(x-ct)+u_0(x+ct)\right] + \frac{1}{2c}\int_{x-ct}^{x+ct} u_1(\xi)\,\mathrm{d}\xi.$$

More generally, if f is a continuous function on $\mathbb{R}\times[0,\infty)$ such that $\frac{\partial f}{\partial x}$ is a continuous function on $\mathbb{R}\times[0,\infty)$, then there is still an explicit formula for the solution:

$$u(x,t) = \frac{1}{2}\left[u_0(x-ct)+u_0(x+ct)\right] + \frac{1}{2c}\int_{x-ct}^{x+ct} u_1(\xi)\,\mathrm{d}\xi + \frac{1}{2c}\int_0^t \int_{x-c(t-\tau)}^{x+c(t-\tau)} f(s,\tau)\,\mathrm{d}s\,\mathrm{d}\tau.$$

We focus on closed bounded spatial interval [a,b] and finite time T (so need to give data at a and b)

6.1 Second-order hyperbolic equations: initial-boundary-value problem and energy estimate

$$\frac{\partial^2 u}{\partial t^2} - c^2\frac{\partial^2 u}{\partial x^2} = f(x,t) \quad \text{for } (x,t) \in (a,b)\times(0,T] \qquad u(x,0) = u_0(x) \quad \text{for } x \in [a,b],$$

$$\frac{\partial u}{\partial t}(x,0) = u_1(x) \quad \text{for } x \in [a,b] \qquad u(a,t) = 0 \quad \text{and} \quad u(b,t) = 0 \quad \text{for } t \in [0,T] \quad (109)$$

f, u_0 and u_1 are continuous and real in relevant areas ((a,b)x[0,T] for f and [a,b] for the u's). Assume compatibility key analytical tools will be 'discrete energy inequalities' used for stability and convergence

We begin by multiplying the partial differential equation $(109)_1$ by the time derivative of u, and we then integrate the resulting expression over the interval $[a,b]$; thus,

$$\int_a^b \frac{\partial^2 u}{\partial t^2}(x,t)\,\frac{\partial u}{\partial t}(x,t)\,\mathrm{d}x - c^2\int_a^b \frac{\partial^2 u}{\partial x^2}(x,t)\,\frac{\partial u}{\partial t}(x,t)\,\mathrm{d}x = \int_a^b f(x,t)\frac{\partial u}{\partial t}(x,t)\,\mathrm{d}x. \qquad (110)$$

$$u(a,t)=0 \text{ and } u(b,t)=0 \text{ for all } t\in[0,T] \Rightarrow \frac{\partial u}{\partial t}(a,t)=0 \quad \text{and} \quad \frac{\partial u}{\partial t}(b,t)=0 \quad \text{for all } t\in[0,T]$$

Thus, by performing partial integration with respect to x in the second term on the left-hand side of (110), we arrive at the following equality:

$$\int_a^b \frac{\partial}{\partial t}\left(\frac{\partial u}{\partial t}(x,t)\right)\frac{\partial u}{\partial t}(x,t)\,\mathrm{d}x + c^2\int_a^b \frac{\partial u}{\partial x}(x,t)\,\frac{\partial}{\partial t}\left(\frac{\partial u}{\partial x}(x,t)\right)\mathrm{d}x = \int_a^b f(x,t)\frac{\partial u}{\partial t}(x,t)\,\mathrm{d}x. \qquad (111)$$

$$\frac{\partial}{\partial t}\left(\frac{\partial u}{\partial t}\right)\frac{\partial u}{\partial t} = \frac{1}{2}\frac{\partial}{\partial t}\left(\frac{\partial u}{\partial t}\right)^2 \quad \text{and} \quad \frac{\partial u}{\partial x}\frac{\partial}{\partial t}\left(\frac{\partial u}{\partial x}\right) = \frac{1}{2}\frac{\partial}{\partial t}\left(\frac{\partial u}{\partial x}\right)^2, \Rightarrow \text{(swapping order of}$$

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\int_a^b \left(\frac{\partial u}{\partial t}\right)^2(x,t)\,\mathrm{d}x + \frac{c^2}{2}\frac{\mathrm{d}}{\mathrm{d}t}\int_a^b \left(\frac{\partial u}{\partial x}\right)^2(x,t)\,\mathrm{d}x = \int_a^b f(x,t)\frac{\partial u}{\partial t}(x,t)\,\mathrm{d}x.$$

operations)

In the special case when f is identically zero, the right-hand side of (112) vanishes, and after integrating the resulting expression from 0 to t, for any $t\in(0,T]$, we deduce that

$$\frac{1}{2}\int_a^b \left(\frac{\partial u}{\partial t}\right)^2(x,t)\,\mathrm{d}x + \frac{c^2}{2}\int_a^b \left(\frac{\partial u}{\partial x}\right)^2(x,t)\,\mathrm{d}x = \frac{1}{2}\int_a^b \left(\frac{\partial u}{\partial t}\right)^2(x,0)\,\mathrm{d}x + \frac{c^2}{2}\int_a^b \left(\frac{\partial u}{\partial x}\right)^2(x,0)\,\mathrm{d}x.$$

(113) View LHS as *total energy* and RHS and *initial total energy* equality says conservation (without

$$\mathcal{L}^2(u(\cdot,t)) := \int_a^b \left(\frac{\partial u}{\partial t}\right)^2(x,t)\,\mathrm{d}x + c^2\int_a^b \left(\frac{\partial u}{\partial x}\right)^2(x,t)\,\mathrm{d}x$$

source term) ⇒ (113) says

$$\mathcal{L}^2(u(\cdot,t)) = \mathcal{L}^2(u(\cdot,0)) \qquad \text{for all } t\in[0,T].$$

More generally, if f is not identically zero, then (112) implies that

$$\mathcal{L}^2(u(\cdot,t)) = \mathcal{L}^2(u(\cdot,0)) + 2\int_0^t\int_a^b f(x,t)\frac{\partial u}{\partial t}(x,\tau)\,\mathrm{d}x\,\mathrm{d}\tau.$$

As $2\alpha\beta \le \alpha^2+\beta^2$ for all $\alpha,\beta\in\mathbb{R}$

$$\mathcal{L}^2(u(\cdot,t)) \le \mathcal{L}^2(u(\cdot,0)) + \int_0^t\int_a^b f^2(x,\tau)\,\mathrm{d}x\,\mathrm{d}\tau + \int_0^t\int_a^b \left(\frac{\partial u}{\partial t}\right)^2(x,\tau)\,\mathrm{d}x\,\mathrm{d}\tau$$

$$\le \mathcal{L}^2(u(\cdot,0)) + \int_0^t\int_a^b f^2(x,\tau)\,\mathrm{d}x\,\mathrm{d}\tau + \int_0^t \mathcal{L}^2(u(\cdot,\tau))\,\mathrm{d}\tau. \qquad (114)$$

now need Gronwall's Lemma

Lemma 14 *(Gronwall's Lemma) Suppose that A and B are continuous real-valued nonnegative functions defined on $[0,T]$, and B is a nondecreasing function of its argument. Suppose further that*

$$A(t) \le B(t) + \int_0^t A(s)\,\mathrm{d}s \qquad \textit{for all } t\in[0,T]\textit{; then} \qquad A(t) \le \mathrm{e}^t\,B(t) \qquad \textit{for all } t\in[0,T].$$

proof starts with $\mathrm{e}^{-t}A(t) - \mathrm{e}^{-t}\int_0^t A(s)\,\mathrm{d}s \le \mathrm{e}^{-t}B(t), \;\Rightarrow\; \frac{\mathrm{d}}{\mathrm{d}t}\left[\mathrm{e}^{-t}\int_0^t A(s)\,\mathrm{d}s\right] \le \mathrm{e}^{-t}B(t)$, then integrates (LHS vanishes at t=0), then upperbounds RHS swapping B(s) for B(t), put into start

$$A(t) := \mathcal{L}^2(u(\cdot,t)) \quad \text{and} \quad B(t) := \mathcal{L}^2(u(\cdot,0)) + \int_0^t\int_a^b f^2(x,\tau)\,\mathrm{d}x\,\mathrm{d}\tau \Rightarrow$$

$$\mathcal{L}^2(u(\cdot,t)) \le \mathrm{e}^t\left(\mathcal{L}^2(u(\cdot,0)) + \int_0^t\int_a^b f^2(x,\tau)\,\mathrm{d}x\,\mathrm{d}\tau\right) \quad \mathcal{L}^2(u(\cdot,t)) := \int_a^b\left(\frac{\partial u}{\partial t}\right)^2(x,t)\,\mathrm{d}x + c^2\int_a^b\left(\frac{\partial u}{\partial x}\right)^2(x,t)\,\mathrm{d}x$$

$$\mathcal{L}^2(u(\cdot,0)) := \int_a^b\left(\frac{\partial u}{\partial t}\right)^2(x,0)\,\mathrm{d}x + c^2\int_a^b\left(\frac{\partial u}{\partial x}\right)^2(x,0)\,\mathrm{d}x = \|u_1\|^2_{L_2((a,b))} + c^2\,|u_0|^2_{H^1((a,b))}.$$

6.2 The implicit scheme: stability, consistency and convergence

$$\frac{U_j^{m+1} - 2U_j^m + U_j^{m-1}}{(\Delta t)^2} - c^2\frac{U_{j+1}^{m+1} - 2U_j^{m+1} + U_{j-1}^{m+1}}{(\Delta x)^2} = f(x_j, t_{m+1}) \qquad \text{for} \begin{cases} j = 1,\dots,J-1, \\ m = 1,\dots,M-1, \end{cases}$$

$$U_j^0 := u_0(x_j)\; U_j^1 := U_j^0 + \Delta t\,u_1(x_j)\; U_0^m := 0 \quad \text{and} \quad U_J^m := 0 \text{ (all with obvious limits for j,m)} \quad (115)$$

The second numerical initial condition, featuring in equation $(115)_3$, stems from the observation that if

$\frac{\partial^2 u}{\partial t^2} \in C([a,b]\times[0,T])$ then $\frac{u(x_j,\Delta t) - U_j^0}{\Delta t} = \frac{u(x_j,\Delta t) - u(x_j,0)}{\Delta t} = \frac{\partial u}{\partial t}(x_j,0) + \mathcal{O}(\Delta t) = u_1(x_j) + \mathcal{O}(\Delta t)$;

thus, by ignoring the $\mathcal{O}(\Delta t)$ term and replacing $u(x_j,\Delta t)$ by its numerical approximation U_j^1 we arrive at the numerical initial condition $(115)_3$.

implicit as after having U(m-1) and U(m) we need to solve a system of equations to get our U(m+1)s

Stability

$$(U,V) := \sum_{j=1}^{J-1}\Delta x\,U_j\,V_j, \quad (U,V] := \sum_{j=1}^{J}\Delta x\,U_j\,V_j, \quad \|U\| := (U,U)^{\frac{1}{2}} \text{ and } \|U]| := (U,U]^{\frac{1}{2}}.$$

$$(A-B,A) = \frac{1}{2}(\|A\|^2 - \|B\|^2) + \frac{1}{2}\|A-B\|^2 \Rightarrow \text{taking } A = U^{m+1} - U^m \text{ and } B = U^m - U^{m-1}$$

$$(U^{m+1} - 2U^m + U^{m-1}, U^{m+1} - U^m) = \frac{1}{2}(\|U^{m+1} - U^m\|^2 - \|U^m - U^{m-1}\|^2) + \frac{1}{2}\|U^{m+1} - 2U^m + U^{m-1}\|^2.$$

$$(A-B,A] = \frac{1}{2}(\|A]|^2 - \|B]|^2) + \frac{1}{2}\|A-B]|^2 \Rightarrow \text{taking } A = D_x^- U^{m+1} \text{ and } B = D_x^- U^m$$

$$\left(-D_x^+ D_x^- U^{m+1}, U^{m+1} - U^m\right) = (D_x^- U^{m+1}, D_x^-(U^{m+1} - U^m)] = (D_x^- U^{m+1} - D_x^- U^m, D_x^- U^{m+1}]$$

$$= \frac{1}{2}(\|D_x^- U^{m+1}]|^2 - \|D_x^- U^m]|^2) + \frac{1}{2}\|D_x^-(U^{m+1} - U^m)]|^2$$

By taking the $(\cdot,\cdot)$ inner product of $(115)_1$ with $U^{m+1} - U^m$ and using the identities stated above

$$\frac{1}{2}\left(\left\|\frac{U^{m+1} - U^m}{\Delta t}\right\|^2 - \left\|\frac{U^m - U^{m-1}}{\Delta t}\right\|^2\right) + \frac{1}{2}(\Delta t)^2\left\|\frac{U^{m+1} - 2U^m + U^{m-1}}{(\Delta t)^2}\right\|^2 \quad (116)$$

$$+\frac{1}{2}c^2(\|D_x^- U^{m+1}]|^2 - \|D_x^- U^m]|^2) + \frac{1}{2}c^2\,(\Delta t)^2 \left\| D_x^- \left(\frac{U^{m+1}-U^m}{\Delta t} \right) \right]\Big\|^2 = (f(\cdot,t_{m+1}), U^{m+1}-U^m).$$

In the special case when f is identically zero the equality (116) implies that

$$\left\| \frac{U^{m+1}-U^m}{\Delta t} \right\|^2 + c^2 \|D_x^- U^{m+1}]|^2 \le \left\| \frac{U^m - U^{m-1}}{\Delta t} \right\|^2 + c^2 \|D_x^- U^m]|^2. \tag{117}$$

$$\mathcal{M}^2(U^m) := \left\| \frac{U^{m+1}-U^m}{\Delta t} \right\|^2 + c^2 \|D_x^- U^{m+1}]|^2. \quad \Rightarrow \quad \mathcal{M}^2(U^m) \le \mathcal{M}^2(U^{m-1}), \qquad \text{for all } m = 1, \ldots, M-1,$$

$$\Rightarrow \quad \mathcal{M}^2(U^m) \le \mathcal{M}^2(U^0), \qquad \text{for all } m = 1, \ldots, M-1.$$

One can verify that the mapping $U \mapsto \max_{m\in\{0,\ldots,M-1\}} [\mathcal{M}^2(U^m)]^{1/2}$ is a norm on the linear space of mesh functions U defined on the space-time mesh $\{(x_j, t_m) : j = 0, 1, \ldots, J,\ m = 0, 1, \ldots, M\}$ such that $U_0^m = U_J^m = 0$ for all $m = 0, 1, \ldots, M$. Thus we have shown that when f is identically zero the implicit scheme (115) is (unconditionally) stable in this norm.

We now return to the general case when f is not identically zero. Our starting point is the equality (116) and we focus our attention on the term on its right-hand side. By the Cauchy–Schwarz inequality,

$$(f(\cdot,t_{m+1}), U^{m+1}-U^m) \le \|f(\cdot,t_{m+1})\|\,\|U^{m+1}-U^m\| \tag{118}$$

$$= \sqrt{\Delta t\, T}\,\|f(\cdot,t_{m+1})\| \sqrt{\frac{\Delta t}{T}} \left\| \frac{U^{m+1}-U^m}{\Delta t} \right\| \le \frac{\Delta t\, T}{2}\|f(\cdot,t_{m+1})\|^2 + \frac{\Delta t}{2T} \left\| \frac{U^{m+1}-U^m}{\Delta t} \right\|^2,$$

(using $\alpha\beta \le \frac{1}{2}\alpha^2 + \frac{1}{2}\beta^2, \qquad \text{for } \alpha, \beta \in \mathbb{R}.$) substituting (118) into (116) (119)

$$\left(1 - \frac{\Delta t}{T}\right) \left(\left\| \frac{U^{m+1}-U^m}{\Delta t} \right\|^2 + c^2 \|D_x^- U^{m+1}]|^2 \right) \le \left\| \frac{U^m - U^{m-1}}{\Delta t} \right\|^2 + c^2 \|D_x^- U^m]|^2 + \Delta t\, T\, \|f(\cdot,t_{m+1})\|^2$$

definition of $\mathcal{M}^2(U^m)$ we can rewrite (119) in the following compact form:

$$\left(1 - \frac{\Delta t}{T}\right) \mathcal{M}^2(U^m) \le \mathcal{M}^2(U^{m-1}) + \Delta t\, T\, \|f(\cdot,t_{m+1})\|^2.$$

$$1 - x \ge \frac{1}{1+2x} \qquad \forall\, x \in \left[0, \tfrac{1}{2}\right],$$

and M >= 2 ⇒

$$\mathcal{M}^2(U^m) \le \left(1 + \frac{2\,\Delta t}{T}\right) \mathcal{M}^2(U^{m-1}) + \Delta t\, T \left(1 + \frac{2\,\Delta t}{T}\right) \|f(\cdot,t_{m+1})\|^2$$

$$\le \left(1 + \frac{2\,\Delta t}{T}\right) \mathcal{M}^2(U^{m-1}) + 2\,\Delta t\, T\, \|f(\cdot,t_{m+1})\|^2.$$

Next lemma needed, easily proved by induction

Lemma 15 *Suppose that $M \ge 2$ is an integer, $\{a_m\}_{m=0}^{M-1}$ and $\{b_m\}_{m=1}^{M-1}$ are nonnegative real numbers, $\alpha > 0$, and*

$$a_m \le \alpha\, a_{m-1} + b_m \qquad \text{for } m = 1, 2, \ldots, M-1.$$

Then,

$$a_m \le \alpha^m a_0 + \sum^{m} \alpha^{m-k} b_k \qquad \text{for } m = 1, 2, \ldots, M-1.$$

apply Lemma 15 with $a_m = \mathcal{M}^2(U^m), \quad b_m = 2\,\Delta t\,T\,\|f(\cdot,t_{m+1})\|^2, \quad \alpha = 1 + \dfrac{2\,\Delta t}{T}$

$$\mathcal{M}^2(U^m) \le \left(1+\frac{2\,\Delta t}{T}\right)^m \mathcal{M}^2(U^0) + 2\,\Delta t\,T \sum_{k=1}^{m} \left(1+\frac{2\,\Delta t}{T}\right)^{m-k} \|f(\cdot,t_{k+1})\|^2 \quad \text{for } m = 1,2,\ldots,M-1.$$

$$\left(1+\frac{2\,\Delta t}{T}\right)^m \le \left(1+\frac{2\,\Delta t}{T}\right)^M = \left(1+\frac{2\,\Delta t}{T}\right)^{\frac{T}{\Delta t}} \le \mathrm{e}^2, \quad \text{as} \quad (1+2x)^{\frac{1}{x}} \le \mathrm{e}^2 \quad \forall\, x \in \left(0, \tfrac{1}{2}\right],$$

with $x = \Delta t/T$, which, in turn, follows by noting that $1 + 2x \le \mathrm{e}^{2x}$ for all $x \ge 0$. Thus we deduce the following stability result for the implicit scheme (115).

Theorem 16 *The implicit finite difference approximation* (115) *of the initial-boundary-value problem* (109), *on a finite difference mesh of spacing* $\Delta x := (b-a)/J$ *with* $J \ge 2$ *in the* x*-direction and* $\Delta t := T/M$ *with* $M \ge 2$ *in the* t*-direction, is (unconditionally) stable in the sense that*

$$\mathcal{M}^2(U^m) \le \mathrm{e}^2\,\mathcal{M}^2(U^0) + 2\,\mathrm{e}^2\,T \sum_{k=1}^{m} \Delta t\,\|f(\cdot,t_{k+1})\|^2, \quad \textit{for } m = 1,\ldots,M-1,$$

independently of the choice of Δx *and* Δt.

Consistency of the implicit scheme define consistency error as normal and $T_j^1 := \dfrac{u_j^1 - u_j^0}{\Delta t} - u_1(x_j)$

$$|T_j^{m+1}| \le \frac{1}{12}c^2(\Delta x)^2 M_{4x} + \frac{5}{3}\Delta t M_{3t}, \qquad \begin{cases} j = 1,\ldots,J-1, \\ m = 1,\ldots,M-1, \end{cases} \qquad (120)$$

…

$$M_{4x} := \max_{(x,t)\in[a,b]\times[0,T]} \left|\frac{\partial^4 u}{\partial x^4}(x,t)\right| \quad \text{and} \quad M_{3t} := \max_{(x,t)\in[a,b]\times[0,T]} \left|\frac{\partial^3 u}{\partial t^3}(x,t)\right|. \quad \text{It remains to bound } T_j^1.$$

$$T_j^1 = \frac{1}{\Delta t}\int_0^{\Delta t} (\Delta t - t)\,\frac{\partial^2 u}{\partial t^2}(x_j,t)\,\mathrm{d}t, \quad \Rightarrow \quad |T_j^1| \le \frac{1}{2}\Delta t\,M_{2t}, \quad j = 1,\ldots,J-1, \quad \Rightarrow$$

$$M_{2t} := \max_{(x,t)\in[a,b]\times[0,T]} \left|\frac{\partial^2 u}{\partial t^2}(x,t)\right|.$$

Convergence of the implicit scheme: Same as before, note global error e satisfies an identical finite difference scheme as U, but with $f(x_j,t_{m+1})$ replaced by T_j^{m+1}, $U_j^0 = u_0(x_j)$ replaced by $e_j^0 = 0$, and $u_1(x_j)$ replaced by T_j^1. It therefore follows from Theorem 16 with U^m replaced by e^m, U^0 replaced by e^0 and $f(x_j,t_{k+1})$ replaced by T_j^{k+1} for $j = 1,\ldots,J-1$ and $k = 1,\ldots,M-1$, that

$$\mathcal{M}^2(e^m) \le \mathrm{e}^2\,\mathcal{M}^2(e^0) + 2\,\mathrm{e}^2\,T \sum_{k=1}^{m} \Delta t \left\|T^{k+1}\right\|^2, \quad \text{for } m = 1,\ldots,M-1.$$

$$\max_{1\le k\le m} \left\|T^{k+1}\right\|^2 = \max_{1\le k\le m} \sum_{i=1}^{J-1} \Delta x\,|T_j^{k+1}|^2 \le (b-a)\left[\frac{1}{12}c^2(\Delta x)^2 M_{4x} + \frac{5}{3}\Delta t M_{3t}\right]^2$$

$$\mathcal{M}^2(e^0) = \left\|\frac{e^1 - e^0}{\Delta t}\right\|^2 + c^2\|D_x^- e^1]|^2 = \|T^1\|^2 + c^2\|D_x^- e^1]|^2 \le (b-a)\left[\frac{1}{2}\Delta t\,M_{2t}\right]^2 + c^2\|D_x^- e^1]|^2.$$

$$D_x^- e_j^1 = D_x^- e_j^0 + \Delta t\,D_x^- T_j^1 = \Delta t\,D_x^- T_j^1 = \int_0^{\Delta t} (\Delta t - t)\,D_x^- \frac{\partial^2 u}{\partial t^2}(x_j,t)\,\mathrm{d}t \;= \frac{1}{\Delta x}\int_0^{\Delta t} (\Delta t - t) \int_{x_{j-1}}^{x_j} \frac{\partial^3 u}{\partial x\,\partial t^2}(x,t)\,\mathrm{d}x\,\mathrm{d}t, \quad \Rightarrow$$

$$\|D_x^- e^1]\|^2 \le (b-a)\left[\frac{1}{2}(\Delta t)^2 M_{1x2t}\right]^2. \quad \Rightarrow \quad \mathcal{M}^2(e^0) \le (b-a)\left[\frac{1}{2}\Delta t\, M_{2t}\right]^2 + c^2(b-a)\left[\frac{1}{2}(\Delta t)^2 M_{1x2t}\right]^2.$$

$$\mathcal{M}^2(e^m) \le \mathrm{e}^2(b-a)\left[\frac{1}{2}\Delta t\, M_{2t}\right]^2 + c^2\mathrm{e}^2(b-a)\left[\frac{1}{2}(\Delta t)^2 M_{1x2t}\right]^2 + 2\mathrm{e}^2\, T^2(b-a)\left[\frac{1}{12}c^2(\Delta x)^2 M_{4x} + \frac{5}{3}\Delta t M_{3t}\right]^2$$

for $m = 1, \ldots, M-1$. Thus, provided that M_{2t}, M_{1x2t}, M_{4x} and M_{3t} are all finite, we have that

$$\max_{m\in\{1,\ldots,M-1\}} [\mathcal{M}^2(u^m - U^m)]^{\frac{1}{2}} = \mathcal{O}((\Delta x)^2 + \Delta t),$$

meaning that the implicit scheme exhibits second order convergence with respect to the spatial discretization step Δx and first-order convergence with respect to the temporal discretization step Δt in the norm $\max_{m\in\{1,\ldots,M-1\}}[\mathcal{M}^2(\cdot)]^{\frac{1}{2}}$. Thanks to the unconditional stability of the implicit scheme, its convergence is also *unconditional* in the sense that there is no limitation on the size of the time step Δt in terms of the spatial mesh-size Δx for convergence of the sequence of numerical approximations to the solution of the wave equation to occur as Δx and Δt tend to 0.

6.3 The explicit scheme: stability, consistency and convergence

$$\frac{U_j^{m+1} - 2U_j^m + U_j^{m-1}}{(\Delta t)^2} - c^2\frac{U_{j+1}^m - 2U_j^m + U_{j-1}^m}{(\Delta x)^2} = f(x_j, t_m) \quad \text{(rest as before) could also do} \quad (122)$$

$$U_j^1 := U_j^0 + \Delta t\, u_1(x_j) + \frac{1}{2}(\Delta t)^2\left(c^2 D_x^+ D_x^- u_1(x_j) + f(x_j, 0)\right). \quad \text{as f cont (one more accurate)} \quad (123)$$

Stability of the explicit scheme

$$\begin{aligned} U_j^{m+1} - U_j^{m-1} &= (U_j^{m+1} - U_j^m) + (U_j^m - U_j^{m-1}) = (U_j^{m+1} + U_j^m) - (U_j^m + U_j^{m-1}), \\ U_j^{m+1} - 2U_j^m + U_j^{m-1} &= (U_j^{m+1} - U_j^m) - (U_j^m - U_j^{m-1}), \\ U_j^{m+1} + 2U_j^m + U_j^{m-1} &= (U_j^{m+1} + U_j^m) + (U_j^m + U_j^{m-1}). \end{aligned} \quad (124)$$

$$\begin{aligned} &\frac{U_j^{m+1} - 2U_j^m + U_j^{m-1}}{(\Delta t)^2} - c^2 D_x^+ D_x^- U_j^m \\ &= \frac{U_j^{m+1} - 2U_j^m + U_j^{m-1}}{(\Delta t)^2} + \frac{c^2(\Delta t)^2}{4} D_x^+ D_x^- \frac{U_j^{m+1} - 2U_j^m + U_j^{m-1}}{(\Delta t)^2} - c^2 D_x^+ D_x^- \frac{U_j^{m+1} + 2U_j^m + U_j^{m-1}}{4} \end{aligned}$$

$\Rightarrow$

$$\left(I + \frac{1}{4}c^2(\Delta t)^2 D_x^+ D_x^-\right)\frac{U_j^{m+1} - 2U_j^m + U_j^{m-1}}{(\Delta t)^2} = c^2 D_x^+ D_x^- \frac{U_j^{m+1} + 2U_j^m + U_j^{m-1}}{4} + f(x_j, t_m) \quad (125)$$

$$(U, V) := \sum_{j=1}^{J-1} \Delta x\, U_j\, V_j, \quad (U, V] := \sum_{j=1}^{J} \Delta x\, U_j\, V_j,$$

$$(\mathcal{D}(A-B), A+B) = (\mathcal{D}A, A) - (\mathcal{D}B, B), \ (\mathcal{D}(A+B), A-B) = (\mathcal{D}A, A) - (\mathcal{D}B, B) \text{ using}$$

$$\mathcal{D} = I + \frac{1}{4}c^2(\Delta t)^2 D_x^+ D_x^- \quad \text{and} \quad \mathcal{D} = c^2 D_x^+ D_x^-$$

$$\left(\left(I + \frac{1}{4}c^2(\Delta t)^2 D_x^+ D_x^-\right)\frac{U^{m+1} - U^m}{\Delta t}, \frac{U^{m+1} - U^m}{\Delta t}\right) - \left(\left(I + \frac{1}{4}c^2(\Delta t)^2 D_x^+ D_x^-\right)\frac{U^m - U^{m-1}}{\Delta t}, \frac{U^m - U^{m-1}}{\Delta t}\right)$$

$$= -c^2\left(-D_x^+ D_x^- \frac{U^{m+1} + U^m}{2}, \frac{U^{m+1} + U^m}{2}\right) + c^2\left(-D_x^+ D_x^- \frac{U^m + U^{m-1}}{2}, \frac{U^m + U^{m-1}}{2}\right) + (f(\cdot, t_m), U^{m+1} - U^{m-1})$$

$(-D_x^+ D_x^- V, V) = (D_x^- V, D_x^- V] = \|D_x^- V]\|^2$ $V = \frac{1}{2}(U^{m+1} + U^m)$ and $V = \frac{1}{2}(U^m + U^{m-1})$, $\Rightarrow$

$$\left(\left(I+\frac{1}{4}c^2(\Delta t)^2 D_x^+ D_x^-\right)\frac{U^{m+1}-U^m}{\Delta t},\frac{U^{m+1}-U^m}{\Delta t}\right)+c^2\left\|D_x^-\frac{U^{m+1}+U^m}{2}\right\|^2$$

$$=\left(\left(I+\frac{1}{4}c^2(\Delta t)^2 D_x^+ D_x^-\right)\frac{U^{m}-U^{m-1}}{\Delta t},\frac{U^{m}-U^{m-1}}{\Delta t}\right)+c^2\left\|D_x^-\frac{U^{m}+U^{m-1}}{2}\right\|^2 \quad (126)$$

$$+(f(\cdot,t_m),U^{m+1}-U^{m-1}) \qquad V_j^m:=\frac{U_j^{m+1}-U_j^m}{\Delta t} \text{ noting that } V_0^m=V_J^m=0,$$

$$\left(\left(I+\frac{1}{4}c^2(\Delta t)^2 D_x^+ D_x^-\right)V^m,V^m\right)=\|V^m\|^2+\frac{1}{4}c^2(\Delta t)^2(D_x^+D_x^-V^m,V^m)=\|V^m\|^2-\frac{1}{4}c^2(\Delta t)^2\|D_x^-V^m]\|^2$$

$$\|V^m\|^2-\frac{1}{4}c^2(\Delta t)^2\|D_x^-V^m]\|^2\geq 0. \text{ noting } (\alpha-\beta)^2\leq 2\alpha^2+2\beta^2,$$

$$\|D_x^-V^m]\|^2=\sum_{j=1}^{J}\Delta x\,|D_x^-V_j^m|^2=(\Delta x)^{-1}\sum_{j=1}^{J}(V_j^m-V_{j-1}^m)^2$$

$$\leq 2(\Delta x)^{-1}\sum_{j=1}^{J}(V_j^m)^2+(V_{j-1}^m)^2=4(\Delta x)^{-1}\sum_{j=1}^{J-1}(V_j^m)^2=4(\Delta x)^{-2}\sum_{j=1}^{J-1}\Delta x\,(V_j^m)^2=\left(\frac{2}{\Delta x}\right)^2\|V\|^2.$$

$$\Rightarrow \left(\left(I+\frac{1}{4}c^2(\Delta t)^2 D_x^+ D_x^-\right)V^m,V^m\right)\geq\left(1-\left(\frac{c\,\Delta t}{\Delta x}\right)^2\right)\|V^m\|^2. \text{ suppose } \frac{c\,\Delta t}{\Delta x}\leq c_0<1. \quad (128)$$

$$\left(\left(I+\frac{1}{4}c^2(\Delta t)^2 D_x^+ D_x^-\right)\frac{U^{m+1}-U^m}{\Delta t},\frac{U^{m+1}-U^m}{\Delta t}\right)\geq(1-c_0^2)\left\|\frac{U^{m+1}-U^m}{\Delta t}\right\|^2.$$

$$\mathcal{N}^2(U^m):=\left(\left(I+\frac{1}{4}c^2(\Delta t)^2 D_x^+ D_x^-\right)\frac{U^{m+1}-U^m}{\Delta t},\frac{U^{m+1}-U^m}{\Delta t}\right)+c^2\left\|D_x^-\frac{U^{m+1}+U^m}{2}\right\|^2$$

$$\mathcal{N}^2(U^m)=\mathcal{N}^2(U^{m-1})+(f(\cdot,t_m),U^{m+1}-U^{m-1}). \text{ so } \mathcal{N}^2(U^m)=\mathcal{N}^2(U^0), \text{ for f = 0}$$

that $U_0^m=U_J^m=0$ for all $m=0,1,\ldots,M$. Thus we have shown that, provided that the CFL condition (128) holds and f is identically zero, the explicit scheme (122) is (conditionally) stable in this norm.

6.4 First-order hyperbolic equations: initial-boundary-value problem and energy estimate

$$\frac{\partial u}{\partial t}+\sum_{i=1}^{n}b_i(x)\frac{\partial u}{\partial x_i}+c(x,t)u=f(x,t),\quad x\in\Omega,\quad t\in(0,T], \qquad u(x,t)=0,\quad x\in\Gamma_-,\quad t\in[0,T],$$

$u(x,0)=u_0(x)\quad x\in\Omega$. Ω be a bounded open set in $\mathbb{R}^n$ $Q=\Omega\times(0,T]$, $\Gamma_-=\{x\in\Gamma: b(x)\cdot\nu(x)<0\}$

$b=(b_1,\ldots,b_n)$ and $\nu(x)$ denotes the unit outward normal to Γ at $x\in\Gamma$. Γ_- will be called the *inflow boundary*. Its complement, $\Gamma_+=\Gamma\backslash\Gamma_-$, will be referred to as the *outflow boundary*. It is important to

Note boundary condition only given to inflow boundary, else may have no solution/ continuous dependence

assume that $b_i\in C^1(\overline{\Omega}),\quad i=1,\ldots,n,\quad c\in C(\overline{Q}),\quad f\in L_2(Q),\quad u_0\in L_2(\Omega)$. (143 - 145)

$$c(x,t)-\frac{1}{2}\sum_{i=1}^{n}\frac{\partial b_i}{\partial x_i}(x)\geq 0,\quad x\in\bar{\Omega},\quad t\in[0,T].$$

Assume unique solutions and assume (146)

$$\left(\frac{\partial u}{\partial t}(\cdot,t),u(\cdot,t)\right)+\left(c(\cdot,t)-\frac{1}{2}\sum_{i=1}^{n}\frac{\partial b_i}{\partial x_i}(\cdot),u^2(\cdot,t)\right)+\frac{1}{2}\int_{\Gamma_+}\left[\sum_{i=1}^{n}b_i(x)\nu_i(x)\right]u^2(x,t)\,ds(x)=(f(\cdot,t),u(\cdot,t)),$$

(147)

(by inner product in $L_2(\Omega)$ of the equation (140) with $u(\cdot,t)$ + BC + partial integration)

where $\nu(x) = (\nu_1(x), \dots, \nu_n(x))$ is the unit outward normal vector to Γ at $x \in \Gamma$. noting

$$\left(\frac{\partial u}{\partial t}, u\right) = \int_\Omega \frac{\partial u}{\partial t}(x,t)\, u(x,t)\,\mathrm{d}x = \int_\Omega \frac{1}{2}\frac{\partial}{\partial t}u^2(x,t)\,\mathrm{d}x = \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\int_\Omega u^2(x,t)\,\mathrm{d}x = \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|u(\cdot,t)\|^2, \quad \Rightarrow$$

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|u(\cdot,t)\|^2 + \frac{1}{2}\int_{\Gamma_+}\left[\sum_{i=1}^n b_i(x)\nu_i(x)\right] u^2(x,t)\,\mathrm{d}s(x) \le (f(\cdot,t), u(\cdot,t)).$$ C-S $\Rightarrow$

$$(f(\cdot,t), u(\cdot,t)) \le \|f(\cdot,t)\|\,\|u(\cdot,t)\| \le \frac{1}{2}\|f(\cdot,t)\|^2 + \frac{1}{2}\|u(\cdot,t)\|^2 \quad \Rightarrow$$

$$\frac{\mathrm{d}}{\mathrm{d}t}\|u(\cdot,t)\|^2 + \int_{\Gamma_+}\left[\sum_{i=1}^n b_i(x)\nu_i(x)\right] u^2(x,t)\,\mathrm{d}s(x) - \|u(\cdot,t)\|^2 \le \|f(\cdot,t)\|^2, \quad t \in [0,T].$$

rewritten

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\mathrm{e}^{-t}\|u(\cdot,t)\|^2\right) + \mathrm{e}^{-t}\int_{\Gamma_+}\left[\sum_{i=1}^n b_i(x)\nu_i(x)\right] u^2(x,t)\,\mathrm{d}s \le \mathrm{e}^{-t}\|f(\cdot,t)\|^2, \quad t \in [0,T].$$

integrating wrt t

$$\mathrm{e}^{-t}\|u(\cdot,t)\|^2 + \int_0^t \mathrm{e}^{-\tau}\int_{\Gamma_+}\left[\sum_{i=1}^n b_i(x)\nu_i(x)\right] u^2(x,\tau)\,\mathrm{d}s(x)\,\mathrm{d}\tau \le \|u_0\|^2 + \int_0^t \mathrm{e}^{-\tau}\|f(\cdot,\tau)\|^2\,\mathrm{d}\tau, \quad t \in [0,T]. \Rightarrow$$

$$\|u(\cdot,t)\|^2 + \int_0^t \mathrm{e}^{t-\tau}\int_{\Gamma_+}\left[\sum_{i=1}^n b_i(x)\nu_i(x)\right] u^2(x,\tau)\,\mathrm{d}s(x)\,\mathrm{d}\tau \le \mathrm{e}^t\|u_0\|^2 + \int_0^t \mathrm{e}^{t-\tau}\|f(\cdot,\tau)\|^2\,\mathrm{d}\tau, \quad t \in [0,T]$$

This, so called, energy inequality expresses the continuous dependence of the solution to (140)–(142)

Important case $c \equiv 0, \quad f \equiv 0, \quad \text{and} \quad \operatorname{div} b = \sum_{i=1}^n \frac{\partial b_i}{\partial x_i} \equiv 0, \quad \Rightarrow$

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|u(\cdot,t)\|^2 + \frac{1}{2}\int_{\Gamma_+}[b(x)\cdot\nu(x)]\,u^2(x,t)\,\mathrm{d}s(x) = 0, \quad \Rightarrow \quad \|u(\cdot,t)\|^2 + \int_0^t\int_{\Gamma_+}[b(x)\cdot\nu(x)]\,u^2(x,\tau)\,\mathrm{d}s(x)\,\mathrm{d}\tau = \|u_0\|^2$$

6.5 Explicit finite difference approximation

$$\frac{\partial u}{\partial t} + b\frac{\partial u}{\partial x} = f(x,t), \quad x \in (0,1), \quad t \in (0,T] \qquad u(x,t) = 0, \quad x \in \Gamma_-, \quad t \in [0,T], \quad u(x,0) = u_0(x), \quad x \in [0,1]$$

If $b > 0$ then $\Gamma_- = \{0\}$, and if $b < 0$ then $\Gamma_- = \{1\}$. Let us assume, for example, that $b > 0$. Then the appropriate boundary condition is

$$u(0,t) = 0, \quad t \in [0,T].$$

$$\frac{U_j^{m+1} - U_j^m}{\Delta t} + b\,D_x^- U_j^m = f(x_j, t_m), \qquad j = 1, \dots, J,$$
$$m = 0, \dots, M-1, \quad \text{or equally}$$

$$U_j^{m+1} = (1-\mu)U_j^m + \mu U_{j-1}^m + \Delta t\, f(x_j, t_m), \qquad \begin{cases} j = 1, \dots, J, \\ m = 0, \dots, M-1, \end{cases}$$

$$U_0^m := 0, \quad m = 0, \dots, M, \qquad U_j^0 := u_0(x_j), \quad j = 0, \dots, J, \qquad \mu := \frac{b\Delta t}{\Delta x};$$ *CFL number*

The explicit finite difference scheme (153) is frequently called the *first-order upwind scheme.*

Suppose that $0 \le \mu \le 1$; $\left|U_j^{m+1}\right| \le (1-\mu)\left|U_j^m\right| + \mu\left|U_{j-1}^m\right| + \Delta t\,|f(x_j,t_m)|$

$$\le (1-\mu)\max_{0\le j\le J}\left|U_j^m\right| + \mu\max_{1\le j\le J+1}\left|U_{j-1}^m\right| + \Delta t\max_{0\le j\le J}|f(x_j,t_m)| = \max_{0\le j\le J}\left|U_j^m\right| + \Delta t\max_{0\le j\le J}|f(x_j,t_m)|$$

$$\|U\|_\infty := \max_{0\le j\le J}|U_j|; \Rightarrow \|U^{m+1}\|_\infty \le \|U^m\|_\infty + \Delta t\|f(\cdot,t_m)\|_\infty, \Rightarrow$$

$$\max_{1\le k\le M}\|U^k\|_\infty \le \|U^0\|_\infty + \sum_{m=0}^{M-1}\Delta t\|f(\cdot,t_m)\|_\infty,$$ expresses stability (under $0 \le \mu = \frac{b\Delta t}{\Delta x} \le 1.$)

It is possible to show that the scheme (153)–(155) is also stable in the mesh-dependent L_2-norm, $\|\cdot]\|$,

$$\|V]\|^2 = \sum_{i=1}^J \Delta x\, V_i^2. \quad (V,W] := \sum_{i=1}^J \Delta x\, V_i W_i.$$ using $U_j^m = \frac{U_j^m + U_{j-1}^m}{2} + \frac{U_j^m - U_{j-1}^m}{2}, \Rightarrow$

$$\left(\frac{U^{m+1}-U^m}{\Delta t}, U^m\right] = \frac{1}{2\Delta t}\left(\|U^{m+1}]\|^2 - \|U^m]\|^2\right) - \frac{\Delta t}{2}\left\|\frac{U^{m+1}-U^m}{\Delta t}\right]\right\|^2$$

taking Inner Product

$$\|U^{m+1}]\|^2 + \Delta t\, b(U_J^m)^2 + b\,\Delta x\,\Delta t\,\|D_x^- U^m]\|^2 - \|U^m]\|^2 - (\Delta t)^2\left\|\frac{U^{m+1}-U^m}{\Delta t}\right]\right\|^2 = 2\Delta t\,(f^m, U^m],$$

For f = 0, substitute original equation into the above to get rid of U(m+1) - U(m), then sum through m for stability with $0 \le \mu = \frac{b\,\Delta t}{\Delta x} \le 1.$ for f not 0

$$\left\|\frac{U^{m+1}-U^m}{\Delta t}\right]\right\|^2 = \|f^m - bD_x^- U^m]\|^2 \le \{\|f^m]\| + b\|D_x^- U^m]\|\}^2$$

$$\le \left(1+\frac{1}{\epsilon'}\right)\|f^m]\|^2 + (1+\epsilon')b^2\|D_x^- U^m]\|^2, \quad \epsilon' > 0, \ (f^m, U^m] \le \|f^m]\|\,\|U^m]\| \le \tfrac{1}{2}\|f^m]\|^2 + \tfrac{1}{2}\|U^m]\|^2,$$

$$\Rightarrow \|U^{m+1}]\|^2 + \Delta t\, b\,|U_n^m|^2 + b\,\Delta x\,\Delta t\left[1-(1+\epsilon')\frac{b\Delta t}{\Delta x}\right]\|D_x^- U^m]\|^2$$

$$\le \Delta t\left[\left(1+\frac{1}{\epsilon'}\right)\Delta t + 1\right]\|f^m]\|^2 + (1+\Delta t)\|U^m]\|^2.$$ Letting $\epsilon = 1 - 1/(1+\epsilon') \in (0,1)$ and assuming that

$$0 \le \mu = \frac{b\,\Delta t}{\Delta x} \le 1-\epsilon, \Rightarrow \|U^{m+1}]\|^2 + \Delta t\, b\,|U_J^m|^2 \le \|U^m]\|^2 + \Delta t\left(1+\frac{\Delta t}{\epsilon}\right)\|f^m]\|^2 + \Delta t\|U^m]\|^2.$$

Sum through m, then need lemma

Lemma 16 *Let (a_k), (b_k), (c_k) and (d_k) be four sequences of nonnegative real numbers such that the sequence (c_k) is nondecreasing and*

$$a_k + b_k \le c_k + \sum_{m=0}^{k-1} d_m a_m, \quad k \ge 1; \ a_0 + b_0 \le c_0.$$ then $$a_k + b_k \le c_k \exp\left(\sum_{m=0}^{k-1} d_m\right), \quad k \ge 1.$$

Apply with $a_k := \|U^k]\|^2, \quad k \ge 0, \quad b_k := \sum_{m=0}^{k-1}\Delta t\, b\,|U_J^m|^2, \quad k \ge 1; \ b_0 = 0,$

$$c_k := \|U^0]\|^2 + \left(1+\frac{\Delta t}{\epsilon}\right)\sum_{m=0}^{k-1}\Delta t\,\|f^m]\|^2, \quad k\geq 1;\ \ c_0 = \|U^0]\|^2, \qquad d_k := \Delta t, \quad k=1,2,\ldots,M,$$ gives stability rule

Global error By virtue of the stability inequality $\max_{\leq m\leq M}\|e^m\|_\infty \leq \sum_{k=0}^{M-1}\Delta t\|T^m\|_\infty.$

taylor expansion for T gives $T_j^m = \frac{1}{2}\Delta t\frac{\partial^2 u}{\partial t^2}(x_j,\tau^m) + \frac{1}{2}b\,\Delta x\frac{\partial^2 u}{\partial x^2}(\xi_j,t_m)$ $M_{kxlt} := \max_{(x,t)\in\overline{Q}}\left|\frac{\partial^{k+l}}{\partial x^k\partial t^l}(x,t)\right|$

$\mathcal{M} := \max(M_{2t},M_{2x}),$ $\Rightarrow$ $|T_j^m| \leq \frac{1}{2}\mathcal{M}(\Delta t + b\,\Delta x) \quad (=\mathcal{O}(\Delta x+\Delta t))$

$\Rightarrow \max_{1\leq m\leq M}\|u^m - U^m\|_\infty \leq \frac{1}{2}T\mathcal{M}(\Delta t + b\,\Delta x)$ $c_\epsilon^\star = \frac{1}{2}e^{T/2}(1+T/\epsilon)^{1/2}T^{1/2}\mathcal{M}.$

6.6 Finite difference approximation of scalar nonlinear hyperbolic conservation laws

$\frac{\partial u}{\partial t} + \frac{\partial}{\partial x}f(u) = 0 \quad \text{for } (x,t)\in\mathbb{R}\times(0,\infty).$ $u_0\in C^1(\mathbb{R})$; has compact support

suppose that $f(0) = f'(0) = 0,$ and $f''(s)\geq 0$ for all $s\in\mathbb{R}.$ $\Rightarrow$ $f'(s)\geq 0$ for all $s\geq 0.$

assume further that $|f'(s)|\leq f'(|s|)$ for all $s\in\mathbb{R}.$

Assuming that there is a $T>0$ such that a solution $u\in C^1(\mathbb{R}\times[0,T])$ to the initial-value problem exists, then thanks to the chain rule the equation (166) can be rewritten as

$$\frac{\partial u}{\partial t} + f'(u)\frac{\partial u}{\partial x} = 0 \quad \text{for } (x,t)\in\mathbb{R}\times(0,T]. \tag{167}$$

Motivated by construction of first order upwind scheme previously, decompose f' into $[f'(u)]_+ + [f'(u)]_-$

$\Rightarrow$ $\frac{\partial u}{\partial t} + [f'(u)]_+\frac{\partial u}{\partial x} + [f'(u)]_-\frac{\partial u}{\partial x} = 0 \quad \text{for } (x,t)\in\mathbb{R}\times(0,T]$ (167)

$$\frac{U_j^{m+1}-U_j^m}{\Delta t} + [f'(U_j^m)]_+ D_x^- U_j^m + [f'(U_j^m)]_- D_x^+ U_j^m = 0, \qquad j\in\mathbb{Z}, \quad m=0,\ldots,M-1, \tag{169}$$
$$U_j^0 := u_0(x_j), \quad j\in\mathbb{Z},$$

$$\begin{aligned} U_j^{m+1} &= U_j^m - \frac{[f'(U_j^m)]_+\Delta t}{\Delta x}(U_j^m - U_{j-1}^m) - \frac{[f'(U_j^m)]_-\Delta t}{\Delta x}(U_{j+1}^m - U_j^m) \\ &= \left(1-\frac{\Delta t}{\Delta x}\left([f'(U_j^m)]_+ - [f'(U_j^m)]_-\right)\right)U_j^m + \frac{[f'(U_j^m)]_+\Delta t}{\Delta x}U_{j-1}^m + \frac{-[f'(U_j^m)]_-\Delta t}{\Delta x}U_{j+1}^m \qquad (171) \\ &= \left(1-\frac{|f'(U_j^m)|\,\Delta t}{\Delta x}\right)U_j^m + \frac{[f'(U_j^m)]_+\Delta t}{\Delta x}U_{j-1}^m + \frac{-[f'(U_j^m)]_-\Delta t}{\Delta x}U_{j+1}^m \end{aligned}$$

Suppose that the following CFL condition holds: $\frac{f'(\|U^0\|_\infty)\,\Delta t}{\Delta x}\leq 1.$

$\frac{f'(\|U^k\|_\infty)\,\Delta t}{\Delta x}\leq 1$ for all $k=0,\ldots,m.$ (for inductive hypothesis)

$$\Rightarrow \quad \frac{|f'(U_j^m)|\,\Delta t}{\Delta x} \le 1 \qquad \text{for all } j \in \mathbb{Z}, \qquad \Rightarrow$$

$$|U_j^{m+1}| \le \left(1 - \frac{|f'(U_j^m)|\,\Delta t}{\Delta x}\right)|U_j^m| + \frac{[f'(U_j^m)]_+\,\Delta t}{\Delta x}|U_{j-1}^m| + \frac{-[f'(U_j^m)]_-\Delta t}{\Delta x}|U_{j+1}^m|$$

$$\le \left(1 - \frac{|f'(U_j^m)|\,\Delta t}{\Delta x}\right)\|U^m\|_\infty + \frac{[f'(U_j^m)]_+\,\Delta t}{\Delta x}\|U^m\|_\infty + \frac{-[f'(U_j^m)]_-\,\Delta t}{\Delta x}\|U^m\|_\infty$$

$$= \left(1 - \frac{|f'(U_j^m)|\,\Delta t}{\Delta x}\right)\|U^m\|_\infty + \frac{|f'(U_j^m)|\,\Delta t}{\Delta x}\|U^m\|_\infty = \|U^m\|_\infty$$

To complete the inductive step it remains to show that (173) holds with m replaced by $m+1$. By (174) and the fact that f' is nondecreasing imply that

$$\frac{f'(\|U^{m+1}\|_\infty)\,\Delta t}{\Delta x} \le \frac{f'(\|U^m\|_\infty)\,\Delta t}{\Delta x} \le 1. \tag{175}$$

The inequality (175) shows that (173) holds with m replaced by $m+1$, which then completes the inductive step. Thus we have shown that, under the CFL condition (172),

$$\|U^{m+1}\|_\infty \le \|U^m\|_\infty \le \cdots \le \|U^0\|_\infty \qquad \text{for all } m = 0, 1, \ldots, M-1, \tag{176}$$

which completes the proof of the assertion that the sequence $\{U_j^m\}_{j\in\mathbb{Z},\, 0\le m\le M}$ of finite difference approximations generated by the scheme is bounded: in particular (170) has been shown to hold.

Assuming that u has continuous and bounded second partial derivatives with respect to x and t defined on $\mathbb{R} \times [0, T]$, it can be shown that

$$\max_{1\le m\le M} \|u^m - U^m\|_\infty = \mathcal{O}(\Delta x + \Delta t),$$