Numerical Methods II

Lecture 1: Introduction, numerical solution of the heat equation.

(1) $$\partial_t u = D \cdot \partial_x^2 u$$

$D > 0$ = diffusion coefficient

Domain: $0 \le x \le L$, $t \ge 0$

Boundary condition: $u(0, t) = u(L, t) = 0$
"Dirichlet"

Initial condition: $u(x, 0) = u_0(x)$.

Initial / Boundary value problem:

given u_0, boundary conditions, calculate $u(x, t)$ for $t > 0$, x in the "domain" $0 < x < L$.

Discretization:

time step Δt

(2)

space step Δx

grid points: $x_j = j \cdot \Delta x$

$t_k = k \cdot \Delta t$

grid = "computational" grid, (mesh)

$= \{ (x_j, t_k) \quad k \geq 0, \ 0 < x_j < L \}$

N = # of "interior" grid points "in space" = # of unknowns per "time level".

$x_0 = 0, \quad x_{N+1} = L, \ \{x_1, \dots, x_N\}$ = interior points.

Finite difference approximation

$U_{j,k} \cong u(x_j, t_k)$

Discrete boundary condition

$U_{0,k} = U_{N+1,k}$ (from the problem)

Discrete initial condition

$U_{j,0} = u_0(x_j) \quad j = 1, 2, \dots, N.$

③

Marching ≡ time stepping:

have U_{jk} for $j=1,\dots,N$

= solution at time t_k

compute $U_{j,k+1}$ for $j=1,\dots,N$

= solution at the next time.

Marching formula derived from finite difference approximations of derivatives

$$\partial_x^2 u(x,t) \cong \frac{u(x+\Delta x,t)-2u(x,t)+u(x-\Delta x,t)}{\Delta x^2}$$

$$\partial_x^2 u(x_j,t_k) \cong \frac{U_{j+1,k}-2U_{j,k}+U_{j+1,k}}{\Delta x^2}$$

$$\partial_t u(x,t) \cong \frac{u(x,t+\Delta t)-u(x,t)}{\Delta t}$$

$$\partial_t u(x_j,t_k) \cong \frac{U_{j,k+1}-U_{jk}}{\Delta t}$$

Approximate the PDE (1) by

$$\frac{U_{j,k+1}-U_{jk}}{\Delta t} = D\cdot\frac{U_{j+1,k}-2U_{jk}+U_{j-1,k}}{\Delta x^2}$$

(4)

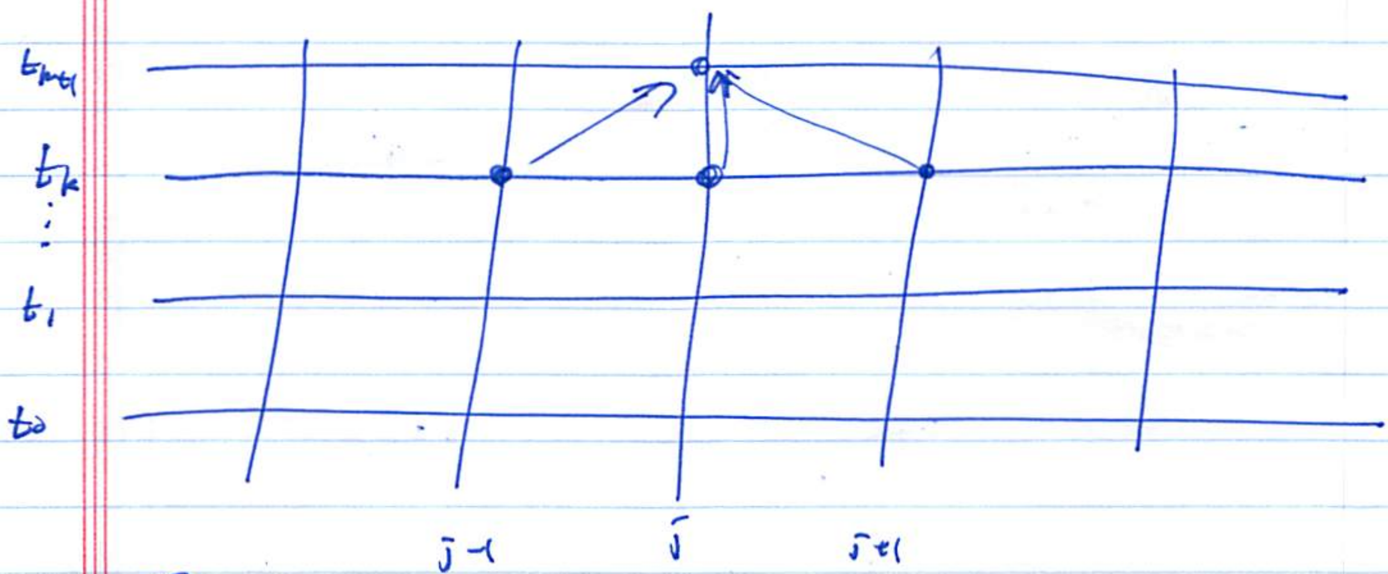
This may be re-written as a formula for $U_{j,k+1}$ in terms of values at time t_k:

(2) $$U_{j,k+1} = a_1\, U_{j+1,k} + a_0\, U_{j,k} + a_{-1}\, U_{j-1,k}$$

$$a_1 = D\frac{\Delta t}{\Delta x^2}$$

$$a_{-1} = D\frac{\Delta t}{\Delta x^2}$$

$$a_0 = 1 - 2D\frac{\Delta t}{\Delta x^2}$$



The stencil of the finite difference formula (2).

Goal: find U_{jk} close enough to $u(x_j, t_k)$ for fixed $T = t_k$, so $k \to \infty$, $\Delta t \to 0$, $\Delta x \to 0$ with $k \Delta t = T$ fixed.

Work = waiting time, depends on the number of space points and time steps

Accuracy depends on $\Delta x, \Delta t$.

Theory: ① how does the error depend on Δx and Δt as $\Delta x \to 0$, $\Delta t \to 0$?

Asymptotic error bound, approximation.

② Time step constraint/strategy

here $\Delta t \leq 2 \frac{\Delta x^2}{D}$, or it doesn't work

Other methods have different constraints.

Here, N space points $\Rightarrow O(N^2)$ time

⑥

steps $\Rightarrow O(N^3)$ work.

Even simple 1-D computations, which should be trivial, can be too slow by this method

Professional scientific computing experts have better methods.

Quantitative analysis of (2): Lax strategy

Consistency / accuracy analysis

+ Stability (uniform bounds as $\Delta x \to 0$, $\Delta t \to 0$)

$\Rightarrow$ error bounds.

R = residual = amount by which the exact solution of (1) fails to satisfy the difference equation (2).

7

E = error = ~~difference~~ $U - u$

R_{jk}: $u(x_j, t_{k+1})$

$$= a_1 u(x_{j+1}, t_k) + a_0 u(x_j, t_k) + a_{-1} u(x_{j-1}, t_k) + \Delta t \, R_{jk}$$

factor of Δt here to make the resulting formulas simpler.

E_{jk}: $E_{j,k} = U_{j,k} - u(x_j, t_k)$

Consistency: $|R| \le C(\Delta x^p + \Delta t^q)$

if $\Delta t = c \cdot \Delta x^2$, just

$$|R| \le C \cdot \Delta x^p$$

p = ~~order~~ "formal" order of accuracy

= 2 for method (2)

Stability: $|E| \le C \cdot |R|$

8

$$\Rightarrow |E| \le C \cdot \Delta x^p$$

p = actual order of accuracy.

To say this correctly, you need to replace vague $|R|$ with specific carefully chosen norms that depend on Δx and δt in specific ways.

Consistency + order of accuracy + error expansion for finite difference approximation.

Review of Taylor series as asymptotic expansion.

Asymptotic expansion: h is a small param that goes to zero.

$$A(h) \approx A_0 + hA_1 + h^2A_2 + \cdots$$

9

means that for any p there is an H_p and C_p so that if $|h| \le H_p$ then

~~$A(h) - (A_0 + \cdots + A_p h_p$~~

$$\left| A(h) - \left(A_0 + hA_1 + \cdots + h^{p-1} A_{p-1}\right) \right| \le C_p h^p .$$

more succinctly: ~~for every $p \geq$~~ every positive integer p,

$$A(h) - \left(\sum_{k<p} h^k A_k \right) = O(h^p).$$

Taylor series as an example of asymptotic series:

if $f(y)$ is C^∞ for $y \in (a,b)$ then $f(y+h) \sim f(y) + hf'(y) + h^2 \frac{1}{2} f''(y) \cdots$

is an asymptotic expansion.

Note well: we don't claim the asymptotic expansion converges to $A(h)$ as $p \to \infty$

$$\times \quad A(h) = \sum_{k=0}^{\infty} h^k A_k \quad \times$$

we don't say this, we don't know whether it's true in specific applications. Sometimes it is true, sometimes not. For our applications today it doesn't matter.

This makes our life simpler. We don't ask a question (convergence of the asymptotic expansion) that is ① hard to answer, and ② irrelevant.

Finite difference, an example:

11

$$\frac{f(y+h)-f(y)}{h} \sim f'(y) + h \cdot \frac{1}{2} f''(y) + h^2 \frac{1}{6} f'''(y) + \cdots$$

This implies:

(a) $$\frac{f(y+h)-f(y)}{h} \to f'(y) \quad \text{as } h \to 0$$

qualitative

(b) $$\frac{f(y+h)-f(y)}{h} = f'(y) + O(h)$$

quantitative error bound

$$\left| \frac{f(y+h)-f(y)}{h} - f'(y) \right| \le C_1 h \quad \text{if } |h| \le H_1.$$

(c) $$\frac{f(y+h)-f(y)}{h} - f'(y) \sim h \frac{1}{2} f''(y) + h^2 \frac{1}{6} f'''(y) + \cdots$$

- asymptotic error expansion.

The one sided two point approximation to the first derivative is first order accurate.

(12)

Centered differences:

if $A(-h) = A(h)$ then

$$A(h) = A_0 + h^2 A_2 + h^4 A_4 + \cdots$$

the odd terms are all zero.

This implies an extra order of accuracy

$$A(h) - A_0 = O(h^2) \quad (\text{not } O(h) \text{ as before})$$

e.g. $$\frac{f(y+h) - f(y-h)}{2h} = D_0(h, y, f)$$

has

$$D_0(-h, y, f) = D_0(h, y, f)$$

Therefore $D_0(h, y, f) = f'(y) + O(h^2)$.

Explicit check:

$$\frac{f(y+h) - f(y-h)}{2h} \sim f'(y) + h^2 \cdot \frac{1}{3} f'''(y) + h^4 \cdot \frac{1}{60} f^{(5)}(y) + \cdots$$

(13)

Second derivative: (centered 3 pt. formula)

$$\frac{f(y+h) - 2f(y) + f(y-h)}{h^2}$$

$$\sim f''(y) + h^2 \tfrac{1}{12} f^{(4)}(y) + \cdots$$

second order.

Residual in (2):

$$u(x_j, t_{k+1}) = u(x_j, t_k) + D \cdot \frac{u(x_{j+1}, t_k) - 2u(x_j, t_k) \cdots}{\Delta x^2}$$

$$+ \Delta t \, R_{jk}$$

$$R_{jk} = \frac{u(x_j, t_k + \Delta t) - u(x_j, t_k)}{\Delta t}$$

$$- D \cdot \frac{u(x_j + \Delta x, t_k) - 2u(x_j, t_k) + u(x_j - \Delta x, t_k)}{\Delta x^2}$$

$$= O(\Delta t) + O(\Delta x^2)$$

$$\sim \tfrac{1}{2} \partial_t^2 u(x_j, t_k) \cdot \Delta t$$

$$+ D \cdot \tfrac{1}{12} \cdot \partial_x^4 u(x_j, t_k) \, \Delta x^2 + \cdots$$

(14)

if $\Delta t = \mu \cdot \Delta x^2$, then

$R_{jk} = O(\Delta x^2)$

The method (2) is formally second order accurate.

__Norms__: The error at time t_k is an N-component vector $E_k = \begin{pmatrix} E_{1,k} \\ \vdots \\ E_{N,k} \end{pmatrix} \in \mathbb{R}^N$

Suppose V is a vector space. A __norm__ on V is a function

$v \rightsquigarrow \|v\| \qquad V \to \mathbb{R} \qquad$ with

$\|v\| \geq 0$ for all $v \in V$

$\|v\| = 0$ only if $v = 0$

$\|a \cdot v\| = |a| \cdot \|v\|$ if $a \in \mathbb{R}$, $v \in V$

$\|v_1 + v_2\| \leq \|v_1\| + \|v_2\|$ if $v_1, v_2 \in V$.

triangle inequality

Examples,

$V = \mathbb{R}^N$, $\quad \|v\| = \max_j |v_j| \qquad L^\infty$ norm

$$\|v\| = \sum |v_j| \qquad L^1 \text{ norm}$$

$$\|v\| = \left(\sum v_j^2\right)^{\frac{1}{2}} \qquad L^2$$

$V =$ "functions" on $[0, L]$.

$$\|v\|_{L^\infty} = \sup_{0<x<L} |v(x)|$$

$$\|v\|_{L^1} = \int_0^L |v(x)|\, dx$$

$$\|v\|_{L^2} = \left(\int_0^L v(x)^2\, dx\right)^{\frac{1}{2}}.$$

$$\|v\|_{H_1} = \left(\int_0^L v(x)^2\, dx + \int v'(x)^2\, dx\right)^{\frac{1}{2}}$$

Sobolev space.

$V =$ functions of x, t

$$\|v\|_{L^\infty_t L^2_x} = \sup_{0\le t\le T} \left(\int_0^L v(x,t)^2\, dx\right)^{\frac{1}{2}}$$

$$\|v\|_{L^2_t L^1_x} = \left(\int_0^T \left(\int_0^L |v(x,t)|\, dx\right)^2 dt\right)^{\frac{1}{2}}$$

16

For us now: define "discrete" norms on $\mathbb{R}^N$ as $N \to \infty$ to be "consistent" with continuous norms of functions in the sense that:

if $v(x)$ is defined for $0 < x < L$

$V^{(N)} \in \mathbb{R}^N$ has components $V_j^{(N)} = v(x_j)$
$= v(j\Delta x)$

Then $\|V^{(N)}\| \to \|v\|$ as $N \to \infty$, $\Delta x \to 0$.

<u>Examples</u> ① $\|v\| = \|v\|_{L^\infty} = \max_x |v(x)|$

(suppose v is continuous)

$$\|V^{(N)}\| = \max_{j=1,\dots,N} |V_j^{(N)}|.$$

The discrete version of the max norm is the max norm.

② $\|v\| = \|v\|_{L^1} = \int_0^L |v(x)|\,dx$

$$\|V^{(N)}\| = \sum_{j=1}^{N} |v(x_j)| \cdot \Delta x = \Delta x \cdot \sum_{j=1}^{N} |V_j^N|$$

The "right" discrete version of the L^1 norm has a Δx prefactor.

(3) $\|v\| = \|v\|_{L^2} = \int_0^L v(x)^2 \, dx$

$$\|V^{(N)}\| = \left(\Delta x \sum_{j=1}^{N} V_j^2 \right)^{\frac{1}{2}}$$

$$= (\Delta x)^{\frac{1}{2}} \left(\sum_{j=1}^{N} V_j^2 \right)^{\frac{1}{2}}$$

The "right" discrete L^2 norm is the "natural" discrete L^2 norm scaled by $\Delta x^{1/2}$.

error at time t_k

~~Th~~ Our goal: $\|E_k^{(N)}\| \leq c \cdot \Delta x^2$

(second order accuracy)

We do this in discrete L^1, L^∞, L^2 norms, ~~the~~ but only if properly scaled with the correct powers of Δx as prefactors.

The main idea of the Day:

$A = N \times N$ matrix ("discrete operator") that advances one time step

$$U_{k+1} = A U_k .$$

The exact soln on the mesh is

$$u_k = \begin{pmatrix} u(x_1, t_k) \\ \vdots \\ u(x_N, t_k) \end{pmatrix}.$$

This satisfies

$$u_{k+1} = A u_k + \Delta t R_k$$

where R_k is the residual at time k

$$R_k = \begin{pmatrix} R_{1,k} \\ \vdots \\ R_{N,k} \end{pmatrix}$$

The error is $u_k - U_k = E_k$.

This satisfies

(19)

(3) $$E_{k+1} = A E_k + \Delta t R_k, \qquad E_0 = 0$$

The operator A is a <u>contraction</u> in the norm $\|\cdot\|$ if $\|Av\| \le \|v\|$ for all $v \in V$. It really should be called a non-expansion, since I (the identity) is a "contraction" in this sense.

We will see that:

① If $\Delta t \le \frac{\Delta x^2}{2D}$ then A (defined by (2)) is a contraction in L^1, L^2, L^∞.

② If the exact solution has
$$|\partial_x^4 u(x,t)| \le C \quad \text{for all } x,\ t \le T,$$
then $\|R_k\| \le C\Delta x^2$ for $t_k \le T$.

<u>Lax "equivalence" theorem</u> (the practical part, adapted to this case):

20

① and ② imply that

$$\|E_k\| \le C\, t_k\, \Delta x^2$$

Proof: Note: if A is a contraction then A^p is a contraction for any positive integer p.

$$\|A^p v\| = \|A \cdot A^{p-1} v\| \le \|A^{p-1} v\| \text{ etc.}$$

From ③,

$$E_1 = \Delta t\, R_0$$

$$E_2 = \Delta t (A R_0 + R_1)$$

$$\vdots$$

$$E_k = \Delta t \cdot \sum_{j=0}^{k-1} A^{k-j-1} R_j$$

From the triangle inequality:

$$\|E_k\| \le \Delta t \cdot \sum_{j=0}^{k-1} \|A^{k-j-1} R_j\|$$

$$\le \Delta t \sum_{j=0}^{k-1} \|R_j\|$$

$$\le \Delta t \cdot C \cdot k \, \Delta x^2 \quad \text{from } (1)$$

$$\le C \, t_k \qquad (t_k = k \, \Delta t)$$

Technical core: stability.

often very technical,

A is a contraction is the same as

$\|A\| \le 1$.

~~If $\|A\| \le 1$ and~~

Matrix/operator norms satisfy

$\|a \cdot A\| = |a| \cdot \|A\|$ if a is a number

$\|A + B\| \le \|A\| + \|B\|$

— the triangle inequality for matrices/operators

We will prove the method (2) is stable if $\frac{\Delta t}{2D \, \Delta x^2} \le 1$. stability limit time step constraint, CFL

(22)

In this case:

$a_1 \geq 0, \; a_0 \geq 0, \; a_{-1} \geq 0$ (only if CFL)

$a_1 + a_0 + a_{-1} = 1$ (even if CFL is violated)

Define left shift, S_L, and right shift S_R by "shift in" a zero

$S_L U = V$ means $V_j = U_{j+1}, \; V_N = 0$

$S_R U = V$ means $V_j = U_{j-1}, \; V_1 = 0$

The scheme (2) may be written as $AU = V$ if

$V_j = a_1 U_{j+1} + a_0 U_j + a_{-1} U_{j-1}$

$V = a_1 S_L + a_0 I + a_{-1} S_R.$

for the vector norms $\|U\|_{L^\infty}, \|U\|_{L^1}, \|U\|_{L^2}$ it is "easy" to see that S_L and S_R are contractions $\|S_L U\| \leq \|U\|$ etc.

There are other natural norms for which S_L and S_R are not contraction (e.g., discrete versions of Sobolev norms).

But with our norms

$$\|AU\| = \|a_1 S_L U + a_0 U + a_{-1} S_R U\|$$
$$\le \|a_1 S_L U\| + \|a_0 U\| + \|a_{-1} S_R U\|$$
$$= |a_1| \|S_L U\| + |a_0| \|U\| + |a_{-1}| \|S_R U\|$$
$$\le |a_1| \|U\| + |a_0| \|U\| + |a_{-1}| \|U\|$$
$$= (|a_1| + |a_0| + |a_{-1}|) \|U\|.$$

(If $a_1 \ge 0$, $a_0 \ge 0$, $a_{-1} \ge 0$, & $a_1 + a_0 + a_{-1} = 1$)

$\|AU\| \le \|U\|$.

In terms of matrices

$$A = a_1 S_L + a_0 I + a_{-1} S_R$$

so
$$\|A\| \le |a_1| \|S_L\| + |a_0| \cdot \|I\| + |a_{-1}| \|S_R\|$$
$$\le |a_1| + |a_0| + |a_{-1}| \quad \text{as before.}$$