Numerical Analysis - Part II

Anders C. Hansen

Lecture 5

Partial differential equations of evolution

Solving the diffusion equation

We consider the solution of the *diffusion equation*

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2}\,, \qquad 0 \le x \le 1, \quad t \ge 0,$$

with *initial conditions* $u(x,0) = u_0(x)$ for $t = 0$ and Dirichlet *boundary conditions* $u(0,t) = \phi_0(t)$ at $x = 0$ and $u(1,t) = \phi_1(t)$ at $x = 1$. By Taylor's expansion

$$\begin{aligned} \tfrac{\partial u(x,t)}{\partial t} &= \tfrac{1}{k}\big[u(x,t+k) - u(x,t)\big] + \mathcal{O}(k), & k = \Delta t\,, \\ \tfrac{\partial^2 u(x,t)}{\partial x^2} &= \tfrac{1}{h^2}\big[u(x-h,t) - 2u(x,t) + u(x+h,t)\big] + \mathcal{O}(h^2), & h = \Delta x\,, \end{aligned}$$

so that, for the true solution, we obtain

$$u(x,t+k) = u(x,t) + \tfrac{k}{h^2}\big[u(x-h,t) - 2u(x,t) + u(x+h,t)\big] + \mathcal{O}(k^2 + kh^2)\,. \tag{1}$$

Numerical scheme for the diffusion equation

That motivates the numerical scheme for approximation $u_m^n \approx u(x_m, t_n)$ on the rectangular mesh $(x_m, t_n) = (mh, nk)$:

$$u_m^{n+1} = u_m^n + \mu \left(u_{m-1}^n - 2u_m^n + u_{m+1}^n \right), \qquad m = 1...M\,. \tag{2}$$

Here $h = \frac{1}{M+1}$ and $\mu = \frac{k}{h^2} = \frac{\Delta t}{(\Delta x)^2}$ is the so-called *Courant number.* With μ being fixed, we have $k = \mu h^2$, so that the local truncation error of the scheme is $\mathcal{O}(h^4)$. Substituting whenever necessary initial conditions u_m^0 and boundary conditions u_0^n and u_{M+1}^n, we possess enough information to advance in (2) from $\boldsymbol{u}^n := [u_1^n, \dots, u_M^n]$ to $\boldsymbol{u}^{n+1} := [u_1^{n+1}, \dots, u_M^{n+1}]$.

Convergence

Similarly to ODEs or Poisson equation, we say that the method is *convergent* if, for a fixed μ, and for every $T > 0$, we have

$$\lim_{h \to 0} |u_m^n - u(x_m, t_n)| = 0 \quad \text{uniformly for} \quad (x_m, t_n) \in [0,1] \times [0,T] \,.$$

In the present case, however, a method has an extra parameter μ, and it is entirely possible for a method to converge for some choice of μ and diverge otherwise.

Proving convergence

Theorem 1

If $\mu \le \frac{1}{2}$, then method (2) converges.

Proof. Let $e_m^n := u_m^n - u(mh, nk)$ be the error of approximation, and let $\boldsymbol{e}^n = [e_1^n, \dots, e_M^n]$ with $\|\boldsymbol{e}^n\| := \max_m |e_m^n|$. Convergence is equivalent to

$$\lim_{h \to 0} \max_{1 \le n \le T/k} \|\boldsymbol{e}^n\| = 0$$

for every constant $T > 0$. Subtracting (1) from (2), we obtain

$$\begin{aligned} e_m^{n+1} &= e_m^n + \mu(e_{m-1}^n - 2e_m^n + e_{m+1}^n) + \mathcal{O}(h^4) \\ &= \mu e_{m-1}^n + (1-2\mu)e_m^n + \mu e_{m+1}^n + \mathcal{O}(h^4). \end{aligned}$$

Then

$$\|\boldsymbol{e}^{n+1}\| = \max_m |e_m^{n+1}| \le (2\mu + |1-2\mu|)\,\|\boldsymbol{e}^n\| + ch^4 = \|e^n\| + ch^4,$$

by virtue of $\mu \le \frac{1}{2}$. Since $\|\boldsymbol{e}^0\| = 0$, induction yields

$$\|\boldsymbol{e}^n\| \le cnh^4 \le \frac{cT}{k}\,h^4 = \frac{cT}{\mu}\,h^2 \to 0 \qquad (h \to 0) \qquad \square$$

Stability, consistency and the Lax equivalence theorem

Suppose that a numerical method for a partial differential equation of evolution can be written in the form[1]

$$\boldsymbol{u}^{n+1} = A_h \boldsymbol{u}^n,$$

where $\boldsymbol{u}^n \in \mathbb{R}^M$, $A_h \in \mathbb{R}^{M\times M}$ is a matrix, and $h = \frac{1}{M+1}$. Fix a norm $\|\cdot\|$ on $\mathbb{R}^M$, and let $\|A_h\| = \sup \frac{\|A_h \boldsymbol{x}\|}{\|\boldsymbol{x}\|}$ be the corresponding induced matrix norm. If we define *stability* as preserving the boundedness of $\boldsymbol{u}^n$ with respect to the norm $\|\cdot\|$, then since

$$\|\boldsymbol{u}^n\| \leq \|A_h^n \boldsymbol{u}^0\| \leq \|A_h\|^n \|\boldsymbol{u}^0\|,$$

we get:

$$\|A_h\| \leq 1 \text{ as } h \to 0 \quad \Rightarrow \quad \text{the method is stable.}$$

[1] Assuming zero boundary conditions

Stability, consistency and the Lax equivalence theorem

If we denote the exact solution of the PDE by $u(x,t)$ and let $\widehat{\boldsymbol{u}}^n = (u(mk,nt))_{1\leq m\leq M}$, then we have $\widehat{\boldsymbol{u}}^{n+1} = A_h\widehat{\boldsymbol{u}}^n + \boldsymbol{\eta}^n$ where $\boldsymbol{\eta}^n$ is the local truncation error. The error vector $\boldsymbol{e}^n = \widehat{\boldsymbol{u}}^n - \boldsymbol{u}^n$ satisfies

$$\boldsymbol{e}^{n+1} = A_h\boldsymbol{e}^n + \boldsymbol{\eta}^n.$$

Using $\|A_h\| \leq 1$ and assuming $\|\boldsymbol{e}^0\| = 0$, we get

$$\|\boldsymbol{e}^n\| \leq \|\boldsymbol{\eta}^{n-1}\| + \cdots + \|\boldsymbol{\eta}^0\|.$$

If *consistency* holds, i.e., $\|\boldsymbol{\eta}^n\| = O(k^2)$, then we see that $\|\boldsymbol{e}^n\| \leq nck^2$ for some constant $c > 0$. Since $n \leq T/k$ we end up with $\|\boldsymbol{e}^n\| \leq cTk$, and so $\|\boldsymbol{e}^n\| \to 0$ as $k \to 0$ uniformly in $n \in [1, T/k]$. This shows convergence.

Stability, consistency and the Lax equivalence theorem

We have thus arrived at the *Lax equivalence theorem*:

Theorem 2

"consistency $+$ *stability* $=$ *convergence"*

(more precisely what we have proved here is the implication $\implies$)

Norms

The discussion above involves a choice of norm on $\mathbb{R}^M$. There are two standard choices of norms:

- *Sup-norm.* Here, we choose

$$\|\boldsymbol{u}\| = \|\boldsymbol{u}\|_\infty = \max_{i=1,\dots,M} |u_i|.$$

It can be easily shown that the corresponding induced norm for a matrix $A \in \mathbb{R}^{M \times M}$ is given by:

$$\|A\|_{\infty\to\infty} := \sup_{\boldsymbol{x}} \frac{\|A\boldsymbol{x}\|_\infty}{\|\boldsymbol{x}\|_\infty} = \max_{i=1,\dots,M} \sum_{j=1}^{M} |A_{ij}|.$$

This the choice of norm we implicitly used in the convergence proof of Theorem 1. The matrix in this case was

$$A_h = \begin{bmatrix} 1-2\mu & \mu & & & \\ & & & & \\ \mu & & \ddots & \ddots & \\ & & & & \\ & & \ddots & \ddots & \mu \\ & & & & \\ & & & \mu & 1-2\mu \end{bmatrix},$$

for which we get $\|A_h\|_{\infty\to\infty} = |1-2\mu| + 2\mu \leq 1$ if $\mu \leq 1/2$.

Stability, consistency and the Lax equivalence theorem

- *Normalized Euclidean norm.* Another common of choice of norm is the normalized Euclidean length, namely,

  $$\|\boldsymbol{u}\| := \sqrt{\frac{1}{M}\sum\nolimits_{i=1}^{M} |u_i|^2}.$$

  The reason for the factor $\frac{1}{M}$ is to ensure that, because of the convergence of Riemann sums, we obtain

  $$\|\boldsymbol{u}\| := \left[\tfrac{1}{M}\sum\nolimits_{i=1}^{M} |u_i|^2\right]^{1/2} \to \left[\int_0^1 |u(x)|^2 \mathrm{d}x\right]^{1/2} =: \|u\|_{L_2} \quad (h = 1/(M+1) \to 0),$$

  The induced matrix norm in this case is the *spectral norm* (or the *operator norm*) and is denoted $\|A\|_2$

  $$\|A\|_2 := \sup_{\boldsymbol{x}} \frac{\|A\boldsymbol{x}\|_2}{\|\boldsymbol{x}\|_2}.$$

  The spectral norm of A is equal to the largest singular value of A. Equivalently, we can write $\|A\|_2 = [\rho(AA^T)]^{1/2}$ where ρ is the spectral radius:

  $$\rho(M) := \max\left\{|\lambda| : \lambda \text{ eigenvalue of } M\right\}.$$

Proving stability directly

Although we can deduce from the theorem that $\mu \leq \frac{1}{2}$ implies stability, we will prove directly that stability $\Leftrightarrow \mu \leq \frac{1}{2}$. Let $\boldsymbol{u}^n = [u_1^n, \ldots, u_M^n]^T$. We can express the recurrence (2)

$$u_m^{n+1} = u_m^n + \mu \left(u_{m-1}^n - 2u_m^n + u_{m+1}^n \right), \qquad m = 1 \ldots M\,,$$

in the matrix form

$$\boldsymbol{u}_h^{n+1} = A_h \boldsymbol{u}_h^n, \qquad A_h = I + \mu A_*, \qquad A_* = \begin{bmatrix} -2 & 1 & & \\ 1 & \ddots & \ddots & \\ & \ddots & \ddots & 1 \\ & & 1 & -2 \end{bmatrix}_{M \times M} .$$

Proving stability directly

Here A_* is TST, with $\lambda_\ell(A_*) = -4\sin^2\frac{\pi\ell h}{2}$, hence $\lambda_\ell(A_h) = 1 - 4\mu\sin^2\frac{\pi\ell h}{2}$, so that its spectrum lies within the interval $[\lambda_M, \lambda_1] = [1 - 4\mu\cos^2\frac{\pi h}{2}, 1 - 4\mu\sin^2\frac{\pi h}{2}]$. Since A_h is symmetric, we have

$$\|A_h\|_2 = \rho(A_h) = \begin{cases} |1 - 4\mu\sin^2\frac{\pi h}{2}| \le 1\,, & \mu \le \frac{1}{2}, \\ |1 - 4\mu\cos^2\frac{\pi h}{2}| > 1, & \mu > \frac{1}{2} \quad (h \le h_\mu)\,. \end{cases}$$

Proving stability directly

We distinguish between two cases.

1) $\mu \leq \frac{1}{2}$: $\|\boldsymbol{u}^n\| \leq \|A\| \cdot \|\boldsymbol{u}^{n-1}\| \leq \cdots \leq \|A\|^n \|\boldsymbol{u}^0\| \leq \|\boldsymbol{u}^0\|$ as $n \to \infty$, for every $\boldsymbol{u}^0$.

2) $\mu > \frac{1}{2}$: Choose $\boldsymbol{u}^0$ as the eigenvector corresponding to the largest (in modulus) eigenvalue, $|\lambda| > 1$. Then $\boldsymbol{u}^n = \lambda^n \boldsymbol{u}^0$, becoming unbounded as $n \to \infty$.

Recall Euler's method

Suppose that we want to solve the differential equation

$$y' = f(t, y), \qquad y(t_0) = y_0.$$

Euler's method is given by

$$y_{n+1} = y_n + kf(t_n, y_n),$$

where $k = t_{n+1} - t_n$ is the step size.

Semidiscretization

Let $u_m(t) = u(mh, t)$, $m = 1...M$, $t \geq 0$. Approximating $\partial^2/\partial x^2$ as before, we deduce from the PDE that the *semidiscretization*

$$\frac{du_m}{dt} = \frac{1}{h^2}(u_{m-1} - 2u_m + u_{m+1}), \qquad m = 1...M \tag{3}$$

carries an error of $\mathcal{O}(h^2)$. This is an ODE system, and we can solve it by any ODE solver. Thus, Euler's method yields (2), while backward Euler results in

$$u_m^{n+1} - \mu(u_{m-1}^{n+1} - 2u_m^{n+1} + u_{m+1}^{n+1}) = u_m^n.$$

Semidiscretization

This approach is commonly known as *the method of lines.* Much (although not all!) of the theory of finite-difference methods for PDEs of evolution can be presented as a two-stage task: first semidiscretize, getting rid of space variables, then use an ODE solver.

Typically, each stage is conceptually easier than the process of discretizing in unison in both time and in space (so-called *full discretization*).

Recall the trapezoidal rule

Suppose that we want to solve the differential equation

$$y' = f(t, y), \qquad y(t_0) = y_0.$$

The trapezoidal rule is given by the formula

$$y_{n+1} = y_n + \tfrac{1}{2}k\Big(f(t_n, y_n) + f(t_{n+1}, y_{n+1})\Big),$$

where $k = t_{n+1} - t_n$ is the step size.

The Crank–Nicolson scheme

Discretizing the ODE (3) with the trapezoidal rule, we obtain

$$u_m^{n+1} - \tfrac{1}{2}\mu(u_{m-1}^{n+1} - 2u_m^{n+1} + u_{m+1}^{n+1}) = u_m^n + \tfrac{1}{2}\mu(u_{m-1}^n - 2u_m^n + u_{m+1}^n), \tag{4}$$

where $m = 1...M$. Thus, each step requires the solution of an $M \times M$ TST system. The error of the scheme is $\mathcal{O}(k^3 + kh^2)$, so basically the same as with Euler's method. However, as we will see, Crank–Nicolson enjoys superior stability features, as compared with the method (2).

The Crank–Nicolson scheme

Note further that (4) is an *implicit* method: advancing each time step requires to solve a linear algebraic system. However, the matrix of the system is TST and its solution by sparse Cholesky factorization can be done in $\mathcal{O}(M)$ operations.