# Numerical Analysis - Part II

Anders C. Hansen

Lecture 6

*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 $(x_m, t_n) = (mh, nk)$:

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

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 $\mu$ 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 $\boldsymbol{u}^n := [u_1^n, \ldots, u_M^n]$ to $\boldsymbol{u}^{n+1} := [u_1^{n+1}, \ldots, u_M^{n+1}]$.

# Convergence

Similarly to ODEs or Poisson equation, we say that the method is *convergent* if, for a fixed $\mu$, 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 $\mu$, and it is entirely possible for a method to converge for some choice of $\mu$ 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 $\boldsymbol{e}^n = [e_1^n, \ldots, e_M^n]$ 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 $\widehat{\boldsymbol{u}}^n = (u(mk, nt))_{1 \le m \le M}$, 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 $\|A_h\| \le 1$ and assuming $\|\boldsymbol{e}^0\| = 0$, we get

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

If *consistency* holds, i.e., $\|\boldsymbol{\eta}^n\| = O(k^2)$, then we see that $\|\boldsymbol{e}^n\| \le nck^2$ for some constant $c > 0$. Since $n \le T/k$ we end up with $\|\boldsymbol{e}^n\| \le cTk$, 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 $\Longrightarrow$ )

# 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 $\|A_h\|_{\infty\to\infty} = |1-2\mu| + 2\mu \le 1$ if $\mu \le 1/2$.

# Stability, consistency and the Lax equivalence theorem

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

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

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

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

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 $\rho$ is the spectral radius:

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

# Norm inequalities

Recall the basic norm inequalities:

$$\|x\|_2 \leq \|x\|_1 \leq \sqrt{n}\,\|x\|_2,$$

$$\|x\|_\infty \leq \|x\|_2 \leq \sqrt{n}\,\|x\|_\infty,$$

$$\|x\|_\infty \leq \|x\|_1 \leq n\,\|x\|_\infty,$$

where $x \in \mathbb{C}^n$.

# 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) $\mu \le \frac{1}{2}$: $\|\boldsymbol{u}^n\| \le \|A\| \cdot \|\boldsymbol{u}^{n-1}\| \le \cdots \le \|A\|^n \|\boldsymbol{u}^0\| \le \|\boldsymbol{u}^0\|$ 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

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

where $m = 1 \ldots M$. 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).

**Definition 3 (Normal matrices)**

We say that a matrix $A$ is *normal* if $A = QD\bar{Q}^T = QDQ^*$, where $D$ is a (complex) diagonal matrix and $Q$ is a unitary matrix (such that $Q\bar{Q}^T = I$, where the bar in $\bar{Q}$ means complex conjugation). In other words, a matrix is normal if it has a complete set of orthonormal eigenvectors.

Examples of the real normal matrices, besides the familiar symmetric matrices ($A = A^T$), inlclude also the matrices which are skew-symmetric ($A = -A^T$), and more generally the matrices with skew-symmetric off-diagonal part.

### Proposition 4

*If $A$ is normal, then $\|A\| = \rho(A)$.*

**Proof.** Let $\boldsymbol{u}$ be any vector (complex-valued as well). We can expand it in the basis of the orthonormal eigenvectors $\boldsymbol{u} = \sum_{i=1}^{n} a_i \boldsymbol{q}_i$ . Then $A\boldsymbol{u} = \sum_{i=1}^{n} \lambda_i a_i \boldsymbol{q}_i$, and since $\boldsymbol{q}_i$ are orthonormal, we obtain

$$\|A\|_2 := \sup_{\boldsymbol{u}} \frac{\|A\boldsymbol{u}\|_2}{\|\boldsymbol{u}\|_2} = \sup_{a_i} \frac{\{\sum_{i=1}^{n} |\lambda_i a_i|^2\}^{1/2}}{\{\sum_{i=1}^{n} |a_i|^2\}^{1/2}} = |\lambda_{\max}| \,.$$

### Remark 5

More generally, one can prove that, for any matrix $A$, we have $\|A\|_2 = [\rho(A\bar{A}^T)]^{1/2}$, and the previous result for normal matrices can be deduced from that formula.

# Crank–Nicolson method for diffusion equation

Let

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

where $m = 1...M$. Then $B\boldsymbol{u}^{n+1} = C\boldsymbol{u}^n$, where the matrices $B$ and $C$ are Toeplitz symmetric tridiagonal (TST),

$$\boldsymbol{u}^{n+1} = B^{-1}C\boldsymbol{u}^n, \qquad \begin{array}{l} B = I - \tfrac{1}{2}\mu A_*\,, \\[1em] C = I + \tfrac{1}{2}\mu A_*\,, \end{array} \qquad A_* = \begin{bmatrix} -2 & 1 & & \\ 1 & \ddots & \ddots & \\ & \ddots & \ddots & 1 \\ & & 1 & -2 \end{bmatrix}_{M\times M} .$$

# Crank–Nicolson method for diffusion equation

All $M \times M$ TST matrices share the same eigenvectors, hence so does $B^{-1}C$. Moreover, these eigenvectors are orthogonal. Therefore, also $A = B^{-1}C$ is normal and its eigenvalues are

$$\lambda_k(A) = \frac{\lambda_k(C)}{\lambda_k(B)} = \frac{1 - 2\mu \sin^2 \frac{1}{2}\pi kh}{1 + 2\mu \sin^2 \frac{1}{2}\pi kh} \quad \Rightarrow \quad |\lambda_k(A)| \leq 1, \qquad k = 1 \ldots M.$$

Consequently Crank–Nicolson is stable for all $\mu > 0$.

**Matlab demo:** Download the Matlab GUI for *Stability of 1D PDEs* from `http://www.damtp.cam.ac.uk/user/hf323/M21-II-NA/demos/pde_stability/pde_stability.html` and solve the diffusion equation in the interval $[0, 1]$ with the Euler method and with Crank–Nicolson. See the effect of unconditional stability!

# Convergence of the Crank-Nicolson method for diffusion equation

It is not difficult to verify that the local error of the Crank-Nicolson scheme is $\eta_m^n = \mathcal{O}(k^3 + kh^2)$, where $\mathcal{O}(k^3)$ is inherited from the trapezoidal rule (compared to $\mathcal{O}(k^2)$ for the Euler method). We also have

$$\|\boldsymbol{\eta}^n\| = \{h \sum_{m=1}^{M} |\eta_m^n|^2\}^{1/2} = \mathcal{O}(k^3 + kh^2)\,.$$

Hence, for the error vectors $\boldsymbol{e}^n$ we have

$$B\boldsymbol{e}^{n+1} = C\boldsymbol{e}^n + \boldsymbol{\eta}^n \quad \Rightarrow \quad \|\boldsymbol{e}^{n+1}\| \le \|B^{-1}C\| \cdot \|\boldsymbol{e}^n\| + \|B^{-1}\| \cdot \|\boldsymbol{\eta}^n\|\,.$$

We have just proved that $\|B^{-1}C\| \le 1$, and we also have $\|B^{-1}\| \le 1$, because all the eigenvalues of $B$ are greater than 1 (by Gershgorin's theorem). Therefore, $\|\boldsymbol{e}^{n+1}\| \le \|\boldsymbol{e}^n\| + \|\boldsymbol{\eta}^n\|$, and

$$\|\boldsymbol{e}^n\| \le \|\boldsymbol{e}^0\| + n\|\boldsymbol{\eta}\| = n\|\boldsymbol{\eta}\| \le \tfrac{cT}{k}(k^3 + kh^2) = cT(k^2 + h^2).$$

Thus, taking $k = \alpha h$ will result in $\mathcal{O}(h^2)$ error of approximation.

# The advection equation

We consider the solution of the *advection equation*

$$\frac{\partial u}{\partial t} = \frac{\partial u}{\partial x}, \qquad 0 \leq x \leq 1, \quad t \geq 0,$$

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 $x = 1$.

If we discretize the right-hand side by
$\frac{\partial u}{\partial x} = \frac{1}{2h}(u(x+h,t) - u(x-h,t)) + \mathcal{O}(h^2)$ we end up with the ODE

$$\frac{du_m}{dt} = \frac{1}{2h}(u_{m+1} - u_{m-1}).$$

# Crank–Nicolson for advection equation

Let

$$u_m^{n+1} - u_m^n = \tfrac{1}{4}\mu(u_{m+1}^{n+1} - u_{m-1}^{n+1}) + \tfrac{1}{4}\mu(u_{m+1}^n - u_{m-1}^n), \qquad m = 1 \ldots M\,.$$

(This is the trapezoidal rule applied to the semidiscretization of advection equation $\frac{\partial u}{\partial t} = \frac{\partial u}{\partial x}$). In this case, $\boldsymbol{u}^{n+1} = B^{-1} C \boldsymbol{u}^n$, where the matrices $B$ and $C$ are Toeplitz antisymmetric tridiagonal,

$$B = \begin{bmatrix} 1 & -\frac{1}{4}\mu & & \\ \frac{1}{4}\mu & 1 & \ddots & \\ & \ddots & \ddots & -\frac{1}{4}\mu \\ & & \frac{1}{4}\mu & 1 \end{bmatrix}, \qquad C = \begin{bmatrix} 1 & \frac{1}{4}\mu & & \\ -\frac{1}{4}\mu & 1 & \ddots & \\ & \ddots & \ddots & \frac{1}{4}\mu \\ & & -\frac{1}{4}\mu & 1 \end{bmatrix}.$$

# Crank–Nicolson for advection equation

Similarly to Exercise 4, the eigenvalues and eigenvectors of the matrix

$$S = \begin{bmatrix} \alpha & \beta & & \\ -\beta & \alpha & \ddots & \\ & \ddots & \ddots & \beta \\ & & -\beta & \alpha \end{bmatrix},$$

are given by $\lambda_k = \alpha + 2\,\mathrm{i}\,\beta\cos kx$, and $\boldsymbol{w}_k = (\mathrm{i}^m \sin kmx)_{m=1}^{M}$, where $x = \pi h = \frac{\pi}{M+1}$. So, all such $S$ are normal and share the same eigenvectors, hence so does $A = B^{-1}C$, hence $A$ is normal and

$$\lambda_k(A) = \frac{\lambda_k(C)}{\lambda_k(B)} = \frac{1 + \frac{1}{2}\,\mathrm{i}\,\mu\cos kx}{1 - \frac{1}{2}\,\mathrm{i}\,\mu\cos kx} \quad \Rightarrow \quad |\lambda_k(A)| = 1, \qquad k = 1...M.$$

So, Crank–Nicolson is again stable for all $\mu > 0$.

# Euler for advection equation

Finally, consider the Euler method for advection equation

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

We have $\boldsymbol{u}^{n+1} = A\boldsymbol{u}^n$, where

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

but $A$ is *not* normal, and although its eigenvalues are bounded by 1 for $\mu \leq 2$ (note $1-\mu$ is the only eigenvalue of $A$), it is the matrix induced norm of $A$ that matters. For this example, it is easier to work with $\|A\|_{\infty\to\infty}$ which we see is given by $|1-\mu| + \mu$ (by the formula in Lecture 5), and this is smaller than 1 precisely when $\mu \leq 1$.