MA 7007: Numerical Solution of Differential Equations I
Diffusion Equations and Parabolic Problems

Suh-Yuh Yang (楊肅煜)

Department of Mathematics, National Central University
Jhongli District, Taoyuan City 32001, Taiwan

E-mail: syyang@math.ncu.edu.tw
Website: http://www.math.ncu.eud.tw/~syyang/

Heat equation (or diffusion equation)

We consider the heat equation on a bounded domain $x \in (0,1)$ and for time $t > t_0 := 0$ which is the classical example of a parabolic equation:

$$u_t = \kappa u_{xx} \quad \text{for } 0 < x < 1,\ t > 0.$$

We need initial condition at time $t_0 = 0$,

$$u(x,0) = \eta(x) \quad \text{for } 0 \leq x \leq 1,$$

and also boundary conditions,

$$\begin{aligned} u(0,t) &= g_0(t) \quad \text{for } t > 0, \\ u(1,t) &= g_1(t) \quad \text{for } t > 0. \end{aligned}$$

For simplicity, we assume that $\kappa = 1$. If $\kappa < 0$, then $u_t = \kappa u_{xx}$ would be a *backward heat equation*, which is an ill-posed problem.

A natural discretization: the forward difference method

Consider a discrete grid with grid points (x_i, t_n), where

$$x_i = ih, \quad t_n = nk.$$

Here $h = \Delta x$ is the mesh spacing on the x-axis and $k = \Delta t$ is the time step. Let $U_i^n \approx u(x_i, t_n)$ represent the numerical approximation to u at (x_i, t_n).

One natural discretization of the heat equation $u_t = u_{xx}$ would be the forward difference method:

$$\frac{U_i^{n+1} - U_i^n}{k} = \frac{1}{h^2}\left(U_{i-1}^n - 2U_i^n + U_{i+1}^n\right).$$

This uses our standard centered difference in space and a forward difference in time. This is an *explicit method* since we can compute each U_i^{n+1} explicitly in terms of the previous data:

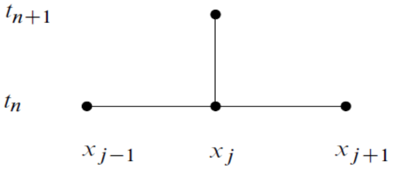
$$U_i^{n+1} = U_i^n + \frac{k}{h^2}\left(U_{i-1}^n - 2U_i^n + U_{i+1}^n\right).$$

Stencil of the forward difference method

Below figure shows the stencil of the forward difference method:

$$U_i^{n+1} = U_i^n + \frac{k}{h^2}\left(U_{i-1}^n - 2U_i^n + U_{i+1}^n\right).$$

This is a one-step method in time, which is also called a two-level method in the context of PDEs, since it involves the solution at two different time levels.



The matrix form of forward difference method

The forward difference method can be represented in the following matrix form:

$$\begin{bmatrix} U_1^{n+1} \\ U_2^{n+1} \\ U_3^{n+1} \\ \vdots \\ U_{m-1}^{n+1} \\ U_m^{n+1} \end{bmatrix} = \begin{bmatrix} (1-2\lambda) & \lambda & & & & \\ \lambda & (1-2\lambda) & \lambda & & & \\ & \lambda & (1-2\lambda) & \lambda & & \\ & & \ddots & \ddots & \ddots & \\ & & & \lambda & (1-2\lambda) & \lambda \\ & & & & \lambda & (1-2\lambda) \end{bmatrix} \begin{bmatrix} U_1^n \\ U_2^n \\ U_3^n \\ \vdots \\ U_{m-1}^n \\ U_m^n \end{bmatrix}$$

$$+ \begin{bmatrix} \lambda g_0(t_n) \\ 0 \\ 0 \\ \vdots \\ 0 \\ \lambda g_1(t_n) \end{bmatrix},$$

where $\lambda := \frac{k}{h^2}$. In compact form, we have

$$U^{n+1} = AU^n + \boldsymbol{g}^n.$$

The Crank-Nicolson method

Another one-step method is the Crank-Nicolson method,

$$\begin{aligned}\frac{U_i^{n+1}-U_i^n}{k} &= \frac{1}{2}\left(D^2U_i^n + D^2U_i^{n+1}\right)\\ &= \frac{1}{2h^2}\left(U_{i-1}^n - 2U_i^n + U_{i+1}^n + U_{i-1}^{n+1} - 2U_i^{n+1} + U_{i+1}^{n+1}\right),\end{aligned}$$

which can be rewritten as

$$U_i^{n+1} = U_i^n + \frac{k}{2h^2}\left(U_{i-1}^n - 2U_i^n + U_{i+1}^n + U_{i-1}^{n+1} - 2U_i^{n+1} + U_{i+1}^{n+1}\right)$$

or

$$-rU_{i-1}^{n+1} + (1+2r)U_i^{n+1} - rU_{i+1}^{n+1} = rU_{i-1}^n + (1-2r)U_i^n + rU_{i+1}^n,$$

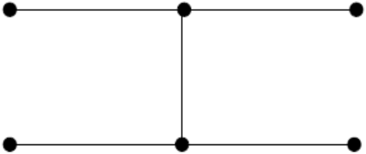
where $r := k/(2h^2)$. This is an *implicit method* and gives a tridiagonal system of equations to solve for all the value U_i^{n+1} simultaneously.

Stencil of the Crank-Nicolson method

Recall the Crank-Nicolson method,

$$-rU_{i-1}^{n+1} + (1+2r)U_i^{n+1} - rU_{i+1}^{n+1} = rU_{i-1}^{n} + (1-2r)U_i^{n} + rU_{i+1}^{n}.$$

The stencil of the Crank-Nicolson method is given below:



The matrix form of the Crank-Nicolson method

The Crank-Nicolson method can be represented in the following matrix form:

$$\begin{bmatrix} (1+2r) & -r & & & & \\ -r & (1+2r) & -r & & & \\ & -r & (1+2r) & -r & & \\ & & \ddots & \ddots & \ddots & \\ & & & -r & (1+2r) & -r \\ & & & & -r & (1+2r) \end{bmatrix} \begin{bmatrix} U_1^{n+1} \\ U_2^{n+1} \\ U_3^{n+1} \\ \vdots \\ U_{m-1}^{n+1} \\ U_m^{n+1} \end{bmatrix}$$

$$= \begin{bmatrix} r(g_0(t_n)+g_0(t_{n+1})) + (1-2r)U_1^n + rU_2^n \\ rU_1^n + (1-2r)U_2^n + rU_3^n \\ rU_2^n + (1-2r)U_3^n + rU_4^n \\ \vdots \\ rU_{m-2}^n + (1-2r)U_{m-1}^n + rU_m^n \\ rU_{m-1}^n + (1-2r)U_m^n + r(g_1(t_n)+g_1(t_{n+1})) \end{bmatrix}.$$

Note how the BCs $u(0,t) = g_0(t)$ and $u(1,t) = g_1(t)$ come into these equations. Since a *tridiagonal system of m equations* can be solved with $O(m)$ work, this implicit method is essentially as efficient per time step as an explicit method.

LTE of the forward difference method

The local truncation error $\tau_i^n := \tau_F(x_i, t_n)$ of the forward difference method,

$$U_i^{n+1} = U_i^n + \frac{k}{h^2}\left(U_{i-1}^n - 2U_i^n + U_{i+1}^n\right),$$

is based on the form

$$\frac{U_i^{n+1} - U_i^n}{k} = \frac{1}{h^2}\left(U_{i-1}^n - 2U_i^n + U_{i+1}^n\right),$$

where

$$\tau_F(x,t) := \frac{u(x,t+k) - u(x,t)}{k} - \frac{1}{h^2}\Big(u(x-h,t) - 2u(x,t) + u(x+h,t)\Big).$$

Again we should be careful to use the difference form that directly models the original differential equation.

LTE of the forward difference method (cont.)

Although we don't know $u(x,t)$, if we assume that it is sufficiently smooth and use Taylor series expansions about $u(x,t)$, we find that

$$\tau_F(x,t) = \Big(u_t + \frac{1}{2}ku_{tt} + \frac{1}{6}k^2 u_{ttt} + \cdots\Big) - \Big(u_{xx} + \frac{1}{12}h^2 u_{xxxx} + \cdots\Big).$$

Since $u_t = u_{xx}$, the $O(1)$ terms drop out. By differentiating $u_t = u_{xx}$ we find that $u_{tt} = u_{txx} = u_{xxxx}$ and so

$$\tau_F(x,t) = \Big(\frac{1}{2}k - \frac{1}{12}h^2\Big)u_{xxxx} + O(k^2 + h^4).$$

This method is said to be second order accurate in space and first order accurate in time, since the local truncation error is $O(h^2 + k)$.

LTE of the Crank-Nicolson method

A local truncation error analysis of the Crank-Nicolson method shows that it is second order accurate in both space and time,

$$\tau(x,t) = O(k^2 + h^2).$$

The local truncation error is given by $\tau_i^n := \tau(x_i, t_n)$, where

$$\begin{aligned}
&\tau(x,t) = \frac{1}{2}\Big\{\tau_F(x,t) + \tau_B(x,t)\Big\} \\
&= \frac{1}{2}\left\{\frac{u(x,t+k)-u(x,t)}{k} - \frac{1}{h^2}\Big(u(x-h,t) - 2u(x,t) + u(x+h,t)\Big)\right\} \\
&+\frac{1}{2}\left\{\frac{u(x,t+k)-u(x,t)}{k} - \frac{1}{h^2}\Big(u(x-h,t+k) - 2u(x,t+k) + u(x+h,t+k)\Big)\right\}.
\end{aligned}$$

LTE of the Crank-Nicolson method (cont.)

The LTE of the backward difference method (an implicit method) is given by

$$\begin{aligned}\tau_B(x,t) &= \frac{u(x,t+k)-u(x,t)}{k} - \frac{1}{h^2}\Big(u(x-h,t+k)-2u(x,t+k)+u(x+h,t+k)\Big)\\ &= \left(u_t - \frac{1}{2}ku_{tt} + \frac{1}{6}k^2u_{ttt} + \cdots\right)_{(x,t+k)} - \left(u_{xx} + \frac{1}{12}h^2u_{xxxx} + \cdots\right)_{(x,t+k)}\\ &= \left(-\frac{1}{2}ku_{tt} + \frac{1}{6}k^2u_{ttt} + \cdots\right)_{(x,t+k)} - \left(\frac{1}{12}h^2u_{xxxx} + \cdots\right)_{(x,t+k)}.\end{aligned}$$

Note that

$$u(x,t) = u(x,t+k) - ku_t(x,t+k) + \frac{k^2}{2!}u_{tt}(x,t+k) - \frac{k^3}{3!}u_{ttt}(x,t+k) + \cdots$$

Therefore,

$$\frac{u(x,t+k)-u(x,t)}{k} = \left(u_t - \frac{1}{2}ku_{tt} + \frac{1}{6}k^2u_{ttt} + \cdots\right)_{(x,t+k)}.$$

Also note that

$$\frac{1}{2}ku_{tt}(x,t) - \frac{1}{2}ku_{tt}(x,t+k) = \frac{1}{2}k\Big(-ku_{ttt}(x,t) + O(k^2)\Big) = O(k^2).$$

Consistency $\oplus$ stability $\Longleftrightarrow$ convergence

Definition: A method is said to be *consistent* if $\tau(x,t) \to 0$ as $k, h \to 0$.

Just as in the other cases we have studied, we expect that consistency, plus some form of stability, will be enough to prove that the method converges at each fixed point (X,T) as we refine the grid in both space and time. Moreover, we expect that for a stable method the global order of accuracy will agree with the order of the local truncation error, e.g., for the Crank-Nicolson method we expect that

$$U_i^n - u(X,T) = O(k^2 + h^2) \to 0,$$

as $k, h \to 0$ when $ih \equiv X$ and $nk \equiv T$ are fixed.

Lax equivalence theorem: For linear PDEs, the fact that "consistency" plus "some form of stability" is equivalent to "convergence."

Stability of the forward difference method

Assume that $g_0(t) = g_1(t) = 0$. Then the matrix form of the forward difference method can be written as

$$U^{n+1} = AU^n.$$

If an error E^0 is made in representing the initial data U^0, then an error AE^0 propagates in U^1, since

$$U^1 = A(U^0 + E^0) = AU^0 + AE^0.$$

Repeating this precess, we have

$$U^n = A^n U^0 + A^n E^0.$$

We say the forward difference method is stable when the errors do not grow as $n \to \infty$. That is, the method is stable if and only if for any initial error E^0, $\|A^n E^0\| \leq \|E^0\|$ for all $n \geq 1$ for some norm $\| \cdot \|$.
$(\Longleftrightarrow \|A^n\| \leq 1$ for some matrix norm $\Longrightarrow \rho(A^n) = (\rho(A))^n \leq 1)$

The forward difference method is conditionally stable

- So $\lambda = k/h^2$ should be chosen such that $\rho(A) \leq 1$. The eigenvalues of A are

  $$\lambda_j = 1 - 2\lambda(1 - \cos\theta_j),$$

  where $\theta_j = \dfrac{j\pi}{m+1}$, $1 \leq j \leq m$. For $\rho(A) \leq 1$ we require

  $$-1 \leq 1 - 2\lambda(1 - \cos\theta_j) \leq 1.$$

  This is true if and only if

  $$\lambda \leq (1 - \cos\theta_j)^{-1}.$$

- The worse case $\cos\theta_j = -1$, which does not happen since the largest $\theta_{j(=m)} = \dfrac{m\pi}{m+1}$. So we have

  $$0 < \lambda \leq \frac{1}{2} \quad \text{or} \quad \frac{k}{h^2} \leq \frac{1}{2} \Longrightarrow k \leq \frac{h^2}{2}.$$

Stability of the Crank-Nicolson method

- Taking $r := k/(2h^2)$ and recalling the CN method, we have

$$-rU_{i-1}^{n+1} + (1+2r)U_i^{n+1} - rU_{i+1}^{n+1} = rU_{i-1}^n + (1-2r)U_i^n + rU_{i+1}^n.$$

- Let $U^n = (U_1^n, U_2^n, \cdots, U_m^n)^\top$ and

$$B = \begin{bmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & & \ddots & & \\ & & & \ddots & -1 \\ & & & -1 & 2 \end{bmatrix}.$$

Then the method can be written in the matrix form

$$(I + rB)U^{n+1} = (I - rB)U^n.$$

The Crank-Nicolson method is unconditionally stable

- For stability, we need $\rho((I+rB)^{-1}(I-rB)) \leq 1$.
- Set $A := (I+rB)^{-1}(I-rB)$ with $U^{n+1} = AU^n$. If $\boldsymbol{x}_i$ is an eigenvector of B associated with the eigenvalue μ_i, then

$$
\begin{aligned}
(I-rB)\boldsymbol{x}_i &= \boldsymbol{x}_i - rB\boldsymbol{x}_i = \boldsymbol{x}_i - r\mu_i\boldsymbol{x}_i = (1-r\mu_i)\boldsymbol{x}_i,\\
(I+rB)\boldsymbol{x}_i &= \boldsymbol{x}_i + rB\boldsymbol{x}_i = \boldsymbol{x}_i + r\mu_i\boldsymbol{x}_i = (1+r\mu_i)\boldsymbol{x}_i,\\
(I+rB)^{-1}\boldsymbol{x}_i &= (1+r\mu_i)^{-1}\boldsymbol{x}_i,\\
(I+rB)^{-1}(I-rB)\boldsymbol{x}_i &= (I+rB)^{-1}(1-r\mu_i)\boldsymbol{x}_i = (1-r\mu_i)(1+r\mu_i)^{-1}\boldsymbol{x}_i.
\end{aligned}
$$

$\Longrightarrow \boldsymbol{x}_i$ is also an eigenvector of A with the eigenvalue $\dfrac{1-r\mu_i}{1+r\mu_i}$.

- To have $\rho((I+rB)^{-1}(I-rB)) \leq 1$, we get it if $|(1+r\mu_i)^{-1}(1-r\mu_i)| \leq 1$.
- Because $\mu_i = 2(1-\cos\theta_i)$, we see that $0 < \mu_i < 4$.

  Thus

$$
\left|\frac{1-r\mu_i}{1+r\mu_i}\right| = \frac{|1-r\mu_i|}{1+r\mu_i} \leq 1, \quad \forall\, r = \frac{k}{2h^2} > 0.
$$

  So, the Crank-Nicolson method is an unconditionally stable method.

Method of lines (MOL)

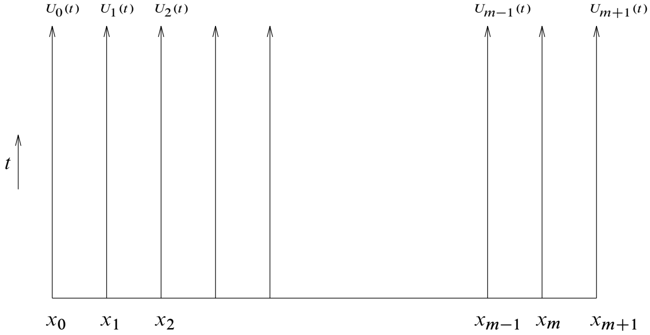
To understand how stability theory for time-dependent PDEs relates to the stability theory we have already developed for time-dependent ODEs, it is easiest to first consider the so-called *method of lines (MOL)* discretization of the PDE, which is a *semidiscrete method.*

For example, we might discretize the heat equation in space at grid point x_i by

$$U_i'(t) = \frac{1}{h^2}\Big(U_{i-1}(t) - 2U_i(t) + U_{i+1}(t)\Big) \quad \text{for} \quad i = 1, 2, \cdots, m,$$

where prime now means differentiation with respect to time. We can view this as a coupled system of m ODEs for the variables $U_i(t)$, which vary continuously in time along the lines shown in figure on next page.

$U_i(t)$ is the solution at the grid point x_i



$U_i(t)$ is the solution along the line forward in time at the grid point x_i

Initial value problem (IVP)

This system can be written as the following IVP:

$$U'(t) = AU(t) + \boldsymbol{g}(t), \qquad U_i(0) = \eta(x_i) \quad 1 \le i \le m,$$

where the tridiagonal matrix A is exactly in (2.9) and $\boldsymbol{g}(t)$ includes the terms needed for the boundary conditions, $U_0(t) \equiv g_0(t)$ and $U_{m+1}(t) \equiv g_1(t)$,

$$A = \frac{1}{h^2} \begin{bmatrix} -2 & 1 & & & & \\ 1 & -2 & 1 & & & \\ & 1 & -2 & 1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & 1 & -2 & 1 \\ & & & & 1 & -2 \end{bmatrix}, \quad \boldsymbol{g}(t) = \frac{1}{h^2} \begin{bmatrix} g_0(t) \\ 0 \\ 0 \\ \vdots \\ 0 \\ g_1(t) \end{bmatrix}.$$

Discretizations of the IVP

- The MOL approach has the advantage of being relatively easy to apply to a fairly general set of time-dependent PDEs, but the resulting method is often not as efficient as specially designed methods for the PDE.

  As a tool in understanding stability theory, however, the MOL discretization is extremely valuable.

- Applying forward Euler's method to

$$U_i'(t) = \frac{1}{h^2}\Big(U_{i-1}(t) - 2U_i(t) + U_{i+1}(t)\Big) \quad \text{for} \quad i = 1, 2, \cdots, m,$$

  we have the fully discrete method, the forward difference method:

$$U_i^{n+1} = U_i^n + \frac{k}{h^2}(U_{i-1}^n - 2U_i^n + U_{i+1}^n).$$

  Similarly, applying the trapezoidal method results in the CN method:

$$U_i^{n+1} = U_i^n + \frac{k}{2h^2}(U_{i-1}^n - 2U_i^n + U_{i+1}^n + U_{i-1}^{n+1} - 2U_i^{n+1} + U_{i+1}^{n+1}).$$

Stability theory

We can investigate the stability of schemes like

$$U_i^{n+1} = U_i^n + \frac{k}{h^2}(U_{i-1}^n - 2U_i^n + U_{i+1}^n).$$

or

$$U_i^{n+1} = U_i^n + \frac{k}{2h^2}(U_{i-1}^n - 2U_i^n + U_{i+1}^n + U_{i-1}^{n+1} - 2U_i^{n+1} + U_{i+1}^{n+1})$$

since these can be interpreted as standard ODE methods applied to the linear system of IVP,

$$U'(t) = AU(t) + \mathbf{g}(t), \qquad U_i(0) = \eta(x_i) \quad 1 \le i \le m.$$

We expect the method to be stable if $k\lambda \in S$, i.e., if the time step k multiplied by any eigenvalue λ of A lies in the *stability region* S of the ODE method (cf. Section 7.4.2).

Remarks on the eigenvalues of matrix A

We have determined the eigenvalues of A in

$$\lambda_p = \frac{2}{h^2}(\cos(p\pi h) - 1) \quad \text{for} \quad p = 1, 2, \ldots, m,$$

where again m and h are related by $h = 1/(m+1)$.

Note that there is a new wrinkle here relative to the ODEs: the eigenvalues λ_p depend on the mesh width h. As we refine the grid and $h \to 0$, the dimension of A increases, the number of eigenvalues we must consider increases, and the values of the eigenvalues change.

Absolute stability region

We consider the following simple IVP:

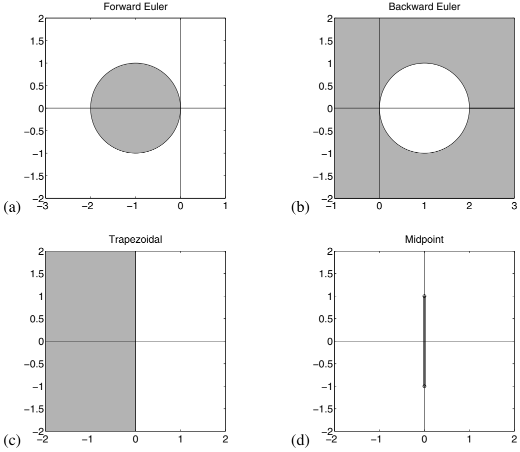
$$u'(t) = \lambda u(t), \quad u(t_0) = \eta.$$

Forward Euler's method applied to this problem gives

$$U^{n+1} = (1 + k\lambda)U^n.$$

- We say that the method is *absolutely stable* when $|1 + k\lambda| \leq 1$; otherwise it is *unstable.*
- Let $z := k\lambda$. Then the method is absolutely stable whenever $-2 \leq z \leq 0$ and we say the interval of absolute stability for Euler's method is $[-2, 0]$.
- It is more common to speak of the region of absolute stability as a region in the complex z plane, allowing the possibility that λ is complex ($k > 0$ is the time step).

Stability region of some methods for IVP



Absolute stability for linear systems

Consider the general linear system $u' = Au$ with constant $m \times m$ matrix A. For simplicity, assume that A is diagonalizable, i.e., there exists linearly independent eigenvectors r_p such that $Ar_p = \lambda_p r_p$ for $p = 1, 2, \cdots, m$.

Let $R = [r_1, r_2, \cdots, r_m]$ and $\Lambda = diag(\lambda_1, \lambda_2, \cdots, \lambda_m)$. Then we have $A = R\Lambda R^{-1}$ and $\Lambda = R^{-1}AR$.

Now, the idea lies in transforming $u' = Au$ to an equivalent system $v' = \Lambda v$ with diagonal Λ by introducing $v(t) = R^{-1}u(t)$.

$$v' = R^{-1}u' = R^{-1}Au = R^{-1}[R\Lambda R^{-1}]u = \Lambda R^{-1}u = \Lambda v.$$

This equivalent system is then decoupled into m independent scalar equations. The pth such equation is $v_p'(t) = \lambda_p v_p(t)$ and the corresponding Euler's method is $V_p^{n+1} = (1 + k\lambda_p)V_p^n$.

For the overall method to be stable, each of the scalar problems must be stable, and this clearly requires that $k\lambda_p$ be in the stability region of Euler's method for all values of p, i.e., $|1 + k\lambda_p| \le 1, \forall p$.

Example 1: forward difference method

If we use Euler's method to obtain the discretization

$$U_i^{n+1} = U_i^n + \frac{k}{h^2}(U_{i-1}^n - 2U_i^n + U_{i+1}^n),$$

then we must require $|1 + k\lambda| \leq 1$ for each eigenvalue. Since the one farthest from the origin is

$$\lambda_m = \frac{2}{h^2}(\cos(m\pi h) - 1) \approx -4/h^2,$$

we require $-2 \leq -4k/h^2 < 0$. This limits the time step allowed to

$$\frac{k}{h^2} \leq \frac{1}{2}.$$

This is a severe restriction: the time step must decrease at the rate of h^2 as we refine the grid, which is much smaller than the spatial width h when h is small.

Example 2: Crank-Nicolson method

If we apply the trapezoidal method to the IVP

$$U'(t) = AU(t) + \mathbf{g}(t), \qquad U_i(0) = \eta(x_i) \quad 1 \le i \le m,$$

we obtain the Crank-Nicolson method,

$$U_i^{n+1} = U_i^n + \frac{k}{2h^2}(U_{i-1}^n - 2U_i^n + U_{i+1}^n + U_{i-1}^{n+1} - 2U_i^{n+1} + U_{i+1}^{n+1}).$$

The trapezoidal method for the ODE is absolutely stable in the whole left half-plane and the eigenvalues

$$\lambda_p = \frac{2}{h^2}(\cos(p\pi h) - 1) \quad \text{for} \quad p = 1, 2, \ldots, m,$$

are always negative. Hence the Crank-Nicolson method is stable for any time step $k > 0$, i.e. it is unconditionally stable!

Stiffness (硬性; 頑性)

In mathematics, a stiff equation is a differential equation for which certain numerical methods for solving the equation are numerically unstable, unless the step size is taken to be extremely small. It has proven difficult to formulate a precise definition of stiffness, but the main idea is that the equation includes some terms that can lead to rapid variation in the solution. – Wikipedia

Example: Consider the IVP with $\lambda \ll -1$,

$$u'(t) = \lambda u(t), \quad u(t_0) = \eta.$$

The the IVP is a stiff problem, since the forward Euler's method applied to this problem is unstable unless the time step $k \ll 1$.

Stiffness of the heat equation

Note that the IVP for the system of ODEs,

$$U'(t) = AU(t) + \mathbf{g}(t), \qquad U_i(0) = \eta(x_i) \quad 1 \le i \le m,$$

we are solving is quite stiff, particularly for small h. The eigenvalues of A are given by

$$\lambda_p = \frac{2}{h^2}(\cos(p\pi h) - 1) \quad \text{for} \quad p = 1, 2, \ldots, m.$$

The smallest and largest eigenvalues in magnitude of matrix A are

$$\begin{aligned} \lambda_1 &\approx \frac{2}{h^2}\left(1 - \frac{(\pi h)^2}{2} + \frac{(\pi h)^4}{24} - \cdots - 1\right) \approx -\pi^2 + \frac{\pi^4}{12}h^2, \\ \lambda_m &\approx -\frac{4}{h^2}, \end{aligned}$$

as h is small. The "stiffness ratio" of the system is $\dfrac{4}{\pi^2 h^2}$, which grows rapidly as $h \to 0$. As a result the explicit Euler method is stable only for very small time steps $k \le h^2/2$.

Stiffness of the heat equation (cont.)

- The stiffness is a reflection of the very different time scales present in solutions to the physical problem modeled by the heat equation. High frequency spatial oscillations in the initial data will decay very rapidly, while smooth data decay much more slowly. See the figures on next page.

  The exact solution of the heat equation on $0 < x < 1$ with $g_0(t) = g_1(t) \equiv 0$ can be represented as a Fourier sine series:

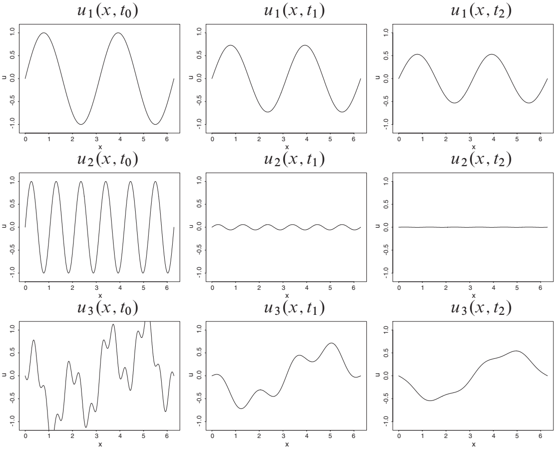
  $$u(x,t) = \sum_{j=1}^{\infty} \hat{u}_j(t) \sin(j\pi x),$$

  where $\hat{u}_j'(t) = -j^2\pi^2\hat{u}_j(t)$ for $j = 1, 2, \cdots$ and so

  $$\hat{u}_j(t) = e^{-j^2\pi^2 t}\hat{u}_j(0).$$

- A method that requires $k \approx h^2$ forces us to take a much finer temporal discretization than we should need to represent smooth solutions. If $h = 0.001$, for example, then if we must take $k = h^2$ rather than $k = h$ we would need to take 1000 time steps to cover each time interval that should be well modeled by a single time step. This is the difficulty we encountered with stiff ODEs.

Stiffness of the heat equation (cont.)



Solutions to the heat equation for $0 < x < 1$ and $g_0(t) = g_1(t) \equiv 0$ at three different times (columns) shown for three different sets of initial conditions (rows)

Convergence analysis

The methods we have studied so far can be written in the form

$$U^{n+1} = B(k)U^n + b^n(k) \qquad (*)$$

for some matrix $B(k) \in \mathbb{R}^{m \times m}$ on a grid with $h = 1/(m+1)$ and $b^n(k) \in \mathbb{R}^m$.

In general these depend on both k and h, but we will assume some fixed rule is specified relating h to k as $k \to 0$. For example, we fix $k = 0.4h^2$ for forward difference method and fix $k = h$ for the Crank-Nicolson method.

Examples

For example, applying forward Euler to the MOL system

$$U'(t) = AU(t) + \boldsymbol{g}(t),$$

gives $B(k) = I + kA$, where A is the tridiagonal matrix in

$$A = \frac{1}{h^2}\begin{bmatrix} -2 & 1 & & & & \\ 1 & -2 & 1 & & & \\ & 1 & -2 & 1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & 1 & -2 & 1 \\ & & & & 1 & -2 \end{bmatrix}, \quad g(t) = \frac{1}{h^2}\begin{bmatrix} g_0(t) \\ 0 \\ 0 \\ \vdots \\ 0 \\ g_1(t) \end{bmatrix}.$$

The Crank-Nicolson method results from applying the trapezoidal method to $U'(t) = AU(t) + g(t)$, which gives

$$B(k) = (I - \frac{k}{2}A)^{-1}(I + \frac{k}{2}A).$$

Lax Equivalence Theorem

To prove convergence we need consistency and a suitable form of stability. As usual, consistency requires that the local truncation error vanishes as $k \to 0$. The form of stability that we need is often called Lax-Richtmyer stability.

Definition. A linear method of the form $U^{n+1} = B(k)U^n + b^n(k)$ is Lax-Richtmyer stable if, for each time T, there is a constant $C_T > 0$ such that

$$\|B(k)^n\| \leq C_T$$

for all $k > 0$ and integers n for which $kn \leq T$.

Lax Equivalence Theorem: A consistent linear method of the form $U^{n+1} = B(k)U^n + b^n(k)$ is convergent if and only if it is Lax-Richtmyer stable.

Main idea of the proof ($\Leftarrow$)

If we apply the numerical method to the exact solution $u(x,t)$, we obtain

$$u^{n+1} = Bu^n + b^n + k\tau^n, \qquad (**)$$

where we suppress the dependence on k for clarity and where

$$u^n = \begin{bmatrix} u(x_1,t_n) \\ u(x_2,t_n) \\ \vdots \\ u(x_m,t_n) \end{bmatrix}, \quad \tau^n = \begin{bmatrix} \tau(x_1,t_n) \\ \tau(x_2,t_n) \\ \vdots \\ \tau(x_m,t_n) \end{bmatrix}.$$

Subtracting (**) from (*) gives the difference equation for the global error $E^n := U^n - u^n$,

$$E^{n+1} = BE^n - k\tau^n.$$

Main idea of the proof ($\Leftarrow$)

After N time steps, we have

$$E^N = B^N E^0 - k \sum_{n=1}^{N} B^{N-n} \tau^{n-1},$$

from which we obtain

$$\|E^N\| \leq \|B^N\| \|E^0\| + k \sum_{n=1}^{N} \|B^{N-n}\| \|\tau^{n-1}\|.$$

If the method is Lax-Richtmyer stable, then for $Nk \leq T$,

$$\|E^N\| \leq C_T \|E^0\| + TC_T \max_{1 \leq n \leq N} \|\tau^{n-1}\| \to 0 \quad \text{as } k \to 0 \text{ for } Nk \leq T,$$

provided the method is consistent ($\|\tau\| \to 0$) and we use appropriate initial data ($\|E^0\| \to 0$ as $k \to 0$).

The forward difference method is convergent with global error $O(h^2)$, provided $k \le h^2/2$

For the heat equation the matrix A is symmetric,

$$A = \frac{1}{h^2}\begin{bmatrix} -2 & 1 & & & & \\ 1 & -2 & 1 & & & \\ & 1 & -2 & 1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & 1 & -2 & 1 \\ & & & & 1 & -2 \end{bmatrix}$$

- The matrix $B := I + kA$ is symmetric and hence the 2-norm is equal to the spectral radius.
- Thus, $\|B\|_2 \le 1$, provided $k \le h^2/2$ is satisfied. So the forward difference method is Lax-Richtmyer stable and hence convergent under this restriction on the time step.
- The global error is $O(h^2)$.

The Crank-Nicolson method is convergent with global error $O(h^2)$, provided $k = h$:

- The matrix
$$B = \left(I - \frac{k}{2}A\right)^{-1}\left(I + \frac{k}{2}A\right)$$
is symmetric and has eigenvalues
$$\frac{1 + k\lambda_p/2}{1 - k\lambda_p/2} \quad p = 1, 2, \cdots, m,$$
and
$$\left|\frac{1 + k\lambda_p/2}{1 - k\lambda_p/2}\right| \leq 1 \quad \text{for all } p.$$
So the Crank-Nicolson method is stable in the 2-norm if $k = h$ and then the global error is $O(h^2)$.
- For the two methods considered so far we have obtained $\|B\| \leq 1$. This is called *strong stability*. If there is a constant α such that $\|B\| \leq 1 + \alpha k$ holds in some norm, then we will have Lax-Richtmyer stability in this norm, since
$$\|B^n\| \leq (1 + \alpha k)^n \leq e^{\alpha T} \text{ for } nk \leq T.$$

Von Neumann analysis

- The von Neumann approach to stability analysis is based on Fourier analysis and hence is generally limited to constant coefficient linear PDEs.
- For simplicity it is usually applied to the Cauchy problem, which is the PDE on all space with no boundaries, $-\infty < x < \infty$ in the 1-D case.
- The von Neumann analysis can also be used to study the stability of problems with periodic boundary conditions, e.g., in $0 < x < 1$ with $u(0,t) = u(1,t)$ imposed. This is generally equivalent to a Cauchy problem with periodic initial data.
- Fourier analysis allows us to obatin simple scalar recursions for each wave number (frequency) separately, rather than dealing with a system of equations that couples together all values of j.
- The underlying physical meaning is that, for a stable scheme, the energy of every frequency components will independently decay to zero as time evoles. Moreover, highly oscillatory components decay much faster than those with low wave numbers.

Basic concepts of Fourier transform

Before formally introducing the von Neumann approach to stability analysis, we briefly review some basic concepts of Fourier transform.

- If $v(x) \in L^2(\mathbb{R})$, i.e., $\|v\|_2 := (\int_{-\infty}^{\infty} |v(x)|^2\, dx)^{1/2} < \infty$. Then $v(x)$ can be expressed as a linear combination of the set of basis functions $\{e^{i\xi x}, \xi \in \mathbb{R}\}$ of $L^2(\mathbb{R})$ in the following form:

  $$v(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \hat{v}(\xi) e^{i\xi x}\, d\xi.$$

  The set of projection coefficients $\hat{v}(\xi)$ is known as the Fourier transform of $v(x)$, which can be computed as:

  $$\hat{v}(\xi) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} v(x) e^{-i\xi x}\, dx,$$

  the L^2 inner product of $v(x)$ and $e^{i\xi x}$.

- In fact, the linear transformation between $v(x)$ and $\hat{v}(\xi)$ is isometric. It can be proved that $\|v\|_2 = \|\hat{v}\|_2$, namely, the Fourier transform is norm-preserving. This is known as *Parseval's relation*.

- The Fourier transform of $v(x - \bar{x})$ equals $e^{-i\xi\bar{x}}\hat{v}(\xi)$, for any shifting point $\bar{x}$. This is known as *Translation rule of Fourier transform*.

Basic concepts of Fourier transform (cont.)

Now suppose V_j is sampled from $v(x) \in L^2(\mathbb{R})$ at grid points $x_j = jh$ for $j = 0, \pm 1, \pm 2, \cdots$, which can be considered as a discrete version of $v(x)$ and is usually called a grid function. Similar Fourier results and properties hold just as in the continuous case:

- If $V_j \in l_2(\mathbb{Z})$, i.e., $\|V\|_2 := (h \sum_{j=-\infty}^{\infty} |V_j|^2)^{1/2} < \infty$. Then V_j can be expressed as a linear combination of $\{e^{i(jh)\xi}, -\pi/h \le \xi \le \pi/h\}$:

  $$V_j = \frac{1}{\sqrt{2\pi}} \int_{-\pi/h}^{\pi/h} \hat{V}(\xi) e^{i(jh)\xi} \, d\xi,$$

  where

  $$\hat{V}(\xi) = \frac{h}{2\pi} \sum_{j=-\infty}^{\infty} V_j e^{-i(jh)\xi},$$

  the l_2 inner product of V_j and $e^{i(jh)\xi}$. Notice that the wave number (frequency) $\pi/h \to \infty$ as mesh size $h \to 0$.

- The linear transformation between V_j and $\hat{V}(\xi)$ is also norm-preserving. The following discrete version of *Parseval's relation* holds:

  $$\|V\|_2 := \left(h \sum_{j=-\infty}^{\infty} |V_j|^2 \right)^{1/2} = \|\hat{V}\|_2 := \left(\int_{\pi/h}^{\pi/h} |\hat{V}(\xi)|^2 \, d\xi \right)^{1/2}.$$

Von Neumann analysis (cont.)

- To show that a finite difference method $U^{n+1} = B(k)U^n + b^n(k) \qquad (*)$ is stable in the 2-norm, we would have to show that
$$\|U^{n+1}\|_2 \le (1+\alpha k)\|U^n\|_2 \qquad \forall n.$$
However, this can be difficult to attack directly. Alternatively one can work with the matrix B itself, but this matrix is growing as we refine the grid.

- Since $\|U\|_2 = \|\hat{U}\|_2$, it is sufficient to instead show that
$$\|\hat{U}^{n+1}\|_2 \le (1+\alpha k)\|\hat{U}^n\|_2 \qquad \forall n.$$
Furthermore, Fourier transforming both sides of $(*)$ and applying the translation rule gives
$$\hat{U}^{n+1}(\xi) = g(\xi)\hat{U}^n(\xi),$$
where the factor $g(\xi)$ depends on the method, namely, the matrix $B(k)$ itself.

- If we can show that
$$|g(\xi)| \le 1 + \alpha k,$$
where α is independent of ξ; then it follows that the method is stable, since
$$|\hat{U}_j^{n+1}(\xi)| \le (1+\alpha k)|\hat{U}_j^n(\xi)| \qquad \forall \xi.$$

The forward difference method

Consider the forward difference method,

$$U_j^{n+1} = U_j^n + \frac{k}{h^2}(U_{j-1}^n - 2U_j^n + U_{j+1}^n). \qquad (\star)$$

To apply von Neumann analysis we consider how this method works on a single wave number $\xi \neq 0$, i.e., we set $U_j^n = e^{ijh\xi}$. Then we expect that

$$U_j^{n+1} = g(\xi)e^{ijh\xi},$$

where $g(\xi)$ is the amplification factor for this wave number. Inserting these expressions into $(\star)$ gives

$$\begin{aligned} g(\xi)e^{ijh\xi} &= e^{ijh\xi} + \frac{k}{h^2}(e^{i\xi(j-1)h} - 2e^{ijh\xi} + e^{i\xi(j+1)h}) \\ &= \Big(1 + \frac{k}{h^2}(e^{-ih\xi} - 2 + e^{ih\xi})\Big)e^{ijh\xi}, \end{aligned}$$

and hence

$$g(\xi) = 1 + \frac{2k}{h^2}\Big(\cos(\xi h) - 1\Big).$$

The forward difference method (cont.)

Since $-1 < \cos(\xi h) < 1$ for any value of $\xi \neq 0$, we see that

$$1 - \frac{4k}{h^2} < g(\xi) < 1$$

for all $\xi \neq 0$. We can guarantee that $|g(\xi)| \leq 1$ for all $\xi \neq 0$ if we require

$$\frac{4k}{h^2} \leq 2.$$

This is exactly the stability restriction

$$k \leq \frac{h^2}{2}$$

we found earlier for this method.

The Crank-Nicolson method

The fact that the Crank-Nicolson method is stable for all k and h can also be shown using von Neumann analysis. Substituting $U_j^n = e^{ijh\xi}$ and $U_j^{n+1} = g(\xi)e^{ijh\xi}$ into the difference equation

$$U_j^{n+1} = U_j^n + \frac{k}{2h^2}\left(U_{j-1}^n - 2U_j^n + U_{j+1}^n + U_{j-1}^{n+1} - 2U_j^{n+1} + U_{j+1}^{n+1}\right)$$

and canceling the common factor of $e^{ijh\xi}$ gives the following relation for $g = g(\xi)$:

$$g = 1 + \frac{k}{2h^2}\left(e^{-ih\xi} - 2 + e^{ih\xi}\right)(1+g),$$

and hence $g = \dfrac{1+\frac{1}{2}z}{1-\frac{1}{2}z}$, where

$$z = \frac{k}{h^2}\left(e^{-ih\xi} - 2 + e^{ih\xi}\right) = \frac{2k}{h^2}\left(\cos(\xi h) - 1\right).$$

Since $z < 0$ for all $\xi \neq 0$, we see that $|g| \leq 1$ and the method is stable for any choice of k and h.

Multidimensional problems

In two space dimensions the heat equation takes the form

$$u_t = u_{xx} + u_{yy} \quad \text{for } (x,y) \in \Omega := (0,1) \times (0,1), t > 0$$

with ICs $u(x,y,0) = \eta(x,y)$ and BCs all along the boundary of our spatial domain Ω. We can discretize in space using the 5-point Laplacian,

$$\nabla_h^2 U_{ij} = \frac{1}{h^2}\Big(U_{i-1,j} + U_{i+1,j} + U_{i,j-1} + U_{i,j+1} - 4U_{ij}\Big), \quad 1 \le i,j \le m.$$

If we then discretize in time using the trapezoidal method, we will obtain the 2-D version of the Crank-Nicolson method,

$$U_{ij}^{n+1} = U_{ij}^n + \frac{k}{2}\Big(\nabla_h^2 U_{ij}^n + \nabla_h^2 U_{ij}^{n+1}\Big), \quad 1 \le i,j \le m.$$

We can rewrite the equations as

$$\Big(1 - \frac{k}{2}\nabla_h^2\Big)U_{ij}^{n+1} = \Big(1 + \frac{k}{2}\nabla_h^2\Big)U_{ij}^n,$$

or in matrix form,

$$\Big(I - \frac{k}{2}\nabla_h^2\Big)U^{n+1} = \Big(I + \frac{k}{2}\nabla_h^2\Big)U^n.$$

Multidimensional problems

Let $A = I - \frac{k}{2}\nabla_h^2$. The matrix A has the same pattern of nonzeros as the matrix for ∇_h^2, but the values are scaled by $k/2$ and then subtracted from the identity matrix. We find that the eigenvalues of A are ($h := 1/(m+1)$)

$$\lambda_{p,q} = 1 - \frac{k}{h^2}\Big((\cos(p\pi h) - 1) + (\cos(q\pi h) - 1)\Big) > 0 \quad \text{for } p, q = 1, 2, \cdots, m.$$

The largest and smallest eigenvalues of the matrix A are given by

$$\begin{aligned} \lambda_{m,m} &\approx 1 - \frac{k}{h^2}\Big(-2-2\Big) = 1 + \frac{4k}{h^2} = O(k/h^2), \\ \lambda_{1,1} &\approx 1 - \frac{2k}{h^2}\Big(1 - \frac{\pi^2 h^2}{2} + \frac{\pi^4 h^4}{24} + \cdots - 1\Big) = 1 + \pi^2 k + O(kh^2) = 1 + O(k). \end{aligned}$$

As a result the condition number of A is

$$\kappa_2(A) = \|A\|_2 \|A^{-1}\|_2 = \frac{\lambda_{m,m}}{\lambda_{1,1}} = O(kh^{-2}).$$

By contrast, the discrete Laplacian ∇_h^2 alone has condition number $O(h^{-2})$. The smaller condition number in the present case can be expected to lead o faster convergence of iterative methods.

Initial values for an iterative method

We have excellent starting guesses for U^{n+1} to the Crank-Nicolson method,

$$U_{ij}^{n+1} = U_{ij}^{n} + \frac{k}{2}\left(\nabla_h^2 U_{ij}^{n} + \nabla_h^2 U_{ij}^{n+1}\right), \quad 1 \le i,j \le m.$$

- Since $U_{ij}^{n+1} = U_{ij}^{n} + O(k)$, we can use U_{ij}^{n}, the values from the previous time step, as the initial values $U_{ij}^{[0]}$ for an iterative method.
- We might do even better by extrapolating forward in time, using

  $$U_{ij}^{[0]} = 2U_{ij}^{n} - U_{ij}^{n-1},$$

  or by using an explicit method (forward Euler),

  $$U_{ij}^{[0]} = (1 + k\nabla_h^2)U_{ij}^{n}.$$

The locally one-dimensional (LOD) method

Rather than solving the coupled sparse matrix equation for all the unknowns

$$U_{ij}^{n+1} = U_{ij}^n + \frac{k}{2}\left(\nabla_h^2 U_{ij}^n + \nabla_h^2 U_{ij}^{n+1}\right) \quad \text{or} \quad \left(1 - \frac{k}{2}\nabla_h^2\right)U_{ij}^{n+1} = \left(1 + \frac{k}{2}\nabla_h^2\right)U_{ij}^n,$$

for $1 \leq i, j \leq m$, an alternative approach is the *locally one-dimensional (LOD) method:*

$$\begin{aligned} U_{ij}^* &= U_{ij}^n + \frac{k}{2}\left(D_x^2 U_{ij}^n + D_x^2 U_{ij}^*\right), \quad (\text{CN method for } u_t = u_{xx}) \\ U_{ij}^{n+1} &= U_{ij}^* + \frac{k}{2}\left(D_y^2 U_{ij}^* + D_y^2 U_{ij}^{n+1}\right), \quad (\text{CN method for } u_t = u_{yy}) \end{aligned}$$

or in matrix form,

$$\begin{aligned} \left(I - \frac{k}{2}D_x^2\right)U^* &= \left(I + \frac{k}{2}D_x^2\right)U^n, \\ \left(I - \frac{k}{2}D_y^2\right)U^{n+1} &= \left(I + \frac{k}{2}D_y^2\right)U^*. \end{aligned}$$

Boundary conditions

- In solving the second set of systems

$$U_{ij}^{n+1} = U_{ij}^{*} + \frac{k}{2}\left(D_y^2 U_{ij}^{*} + D_y^2 U_{ij}^{n+1}\right), \quad 1 \le i,j \le m,$$

  we need boundary values U_{i0}^{*} and U_{i0}^{n+1} along the bottom boundary and $U_{i,m+1}^{*}$ and $U_{i,m+1}^{n+1}$ along the top boundary, for terms that go on the right-hand side of each tridiagonal system.

- To obtain the values U_{i0}^{*} and $U_{i,m+1}^{*}$, we solve the first part

$$U_{ij}^{*} = U_{ij}^{n} + \frac{k}{2}\left(D_x^2 U_{ij}^{n} + D_x^2 U_{ij}^{*}\right), \quad \text{for } j = 0, m+1.$$

Boundary conditions (continued)

- In solving the first set of systems

$$U_{ij}^{*} = U_{ij}^{n} + \frac{k}{2}\left(D_x^2 U_{ij}^{n} + D_x^2 U_{ij}^{*}\right), \quad 1 \le i, j \le m,$$

  we need boundary values U_{0j}^{*} and U_{0j}^{n} along the left boundary and values $U_{m+1,j}^{*}$ and $U_{m+1,j}^{n}$ along the right boundary.

- To determine proper values for U_{0j}^{*} and $U_{m+1,j}^{*}$, we can use

$$U_{ij}^{n+1} = U_{ij}^{*} + \frac{k}{2}(D_y^2 U_{ij}^{*} + D_y^2 U_{ij}^{n+1}), \quad \text{for } i = 0, m+1.$$

$$\Longrightarrow \left(1 + \frac{k}{2}D_y^2\right)U_{ij}^{*} = \left(1 - \frac{k}{2}D_y^2\right)U_{ij}^{n+1}, \quad \text{for } i = 0, m+1.$$

Some remarks on LOD method

- Physically the LOD method corresponds to modeling diffusion in the $x-$ and $y-$directions over time k as a decoupled process in which u is allowed to diffuse only in the $x-$direction and then only in the $y-$direction.

- For the constant coefficient diffusion problem, it can be shown that this alternating diffusion approach gives exactly the same behavior as the original two-dimensional diffusion (This is because the differential operators ∂_x^2 and ∂_y^2 commute).

- Numerically there is much less computational cost using the LOD method than the fully coupled C-N method. Taking a single time step requires solving only $2m + 2$ tridiagonal systems of size m, and thus $(2m + 2) \times O(m) = O(m^2)$ work, which is the optimal order.

- With proper treatment of the boundary conditions, it can be shown that the LOD method is second order accurate. It can also be shown that this method, like full Crank-Nicolson, is unconditionally stable for any time step k.

The alternating direction implicit (ADI) method

A modification of the LOD method is also often used, in which the two steps each involve discretization in only one spatial direction at the advanced time level but coupled with discretization in the opposite direction at the old time level. The classical method of this form is

$$\begin{aligned} U_{ij}^{*} &= U_{ij}^{n} + \frac{k}{2}\left(D_y^2 U_{ij}^{n} + D_x^2 U_{ij}^{*}\right), \\ U_{ij}^{n+1} &= U_{ij}^{*} + \frac{k}{2}\left(D_x^2 U_{ij}^{*} + D_y^2 U_{ij}^{n+1}\right). \end{aligned}$$

This is called the *alternating direction implicit (ADI) method* and was first introduced by Douglas and Rachford (Transactions of the AMS, 1956). This again gives decoupled tridiagonal systems to solve in each step:

$$\begin{aligned} \left(I - \frac{k}{2} D_x^2\right) U^{*} &= \left(I + \frac{k}{2} D_y^2\right) U^{n}, \\ \left(I - \frac{k}{2} D_y^2\right) U^{n+1} &= \left(I + \frac{k}{2} D_x^2\right) U^{*}. \end{aligned}$$