2.3. Quantitative Properties of Finite Difference Schemes

2.3.1. Consistency, Convergence and Stability of F.D. schemes

Reading: Tannehill et al. Sections 3.3.3 and 3.3.4.

Three important properties of F.D. schemes:

Consistency – An F.D. representation of a PDE is consistent if the difference between PDE and FDE, i.e., the truncation error, vanishes as the grid interval and time step size approach zero,

i.e., when $\lim_{\Delta \to 0}(\text{PDE} - \text{FDE}) = 0$.

Comment:

- Consistency deals with how well the FDE approximates the PDE.

Stability – For a stable numerical scheme, the errors from any source will not grow unboundedly with time.

Comments:

- A concept that is applicable only to marching (time-integration) problems.
- Generally we are much more concerned with stability than consistency.
- Some hard work is often needed to establish analytically the stability of a scheme.

Convergence – It means that the solution to a FDE approaches the true solution to the PDE as both grid interval and time step size are reduced.

Lax's Equivalence Theorem

> For a well-posed, linear initial value problem, the necessary and sufficient condition for convergence is that the FDE is stable and consistent.
>
> The theorem has been proved for initial value problems governed by linear PDE's (Richtmyer and Morton 1967).

We will discuss the three concepts one by one.

2.3.2. Consistency

Consistency means $\text{PDE} - \text{FDE} \to 0$ when $\Delta x \to 0$ and $\Delta t \to 0$.

Clearly consistency is the necessary condition for convergence.

Example:

Consider a 1-D diffusion equation:

$$\frac{\partial u}{\partial t} = K \frac{\partial^2 u}{\partial x^2} \quad (K > 0 \text{ and constant})$$

We use the forward-in-time and centered-in-space (FTCS) scheme:

$$\frac{u_i^{n+1} - u_i^n}{\Delta t} = K \frac{u_{i-1}^n - 2u_i^n + u_{i+1}^n}{\Delta x^2}$$

To show consistency, we need to determine the truncation error τ.

Using Taylor series expansion method,

$$u_i^{n+1} = u_i^n + \Delta t \frac{\partial u}{\partial t} + \frac{(\Delta t)^2}{2!} \frac{\partial^2 u}{\partial t^2} + \frac{(\Delta t)^3}{3!} \frac{\partial^3 u}{\partial t^3} + \ldots$$

$$u_{i\pm1}^n = u_i^n \pm \Delta x \frac{\partial u}{\partial x} + \frac{(\Delta x)^2}{2!} \frac{\partial^2 u}{\partial x^2} \pm \frac{(\Delta x)^3}{3!} \frac{\partial^3 u}{\partial x^3} + \ldots$$

Substituting into the FDE, we have

$$\frac{\partial u}{\partial t} + \frac{\Delta t}{2} \frac{\partial^2 u}{\partial t^2} + O(\Delta t^2) = K \left[\frac{\partial^2 u}{\partial x^2} + O(\Delta x^2) \right] + \ldots$$

therefore

$$\tau = \frac{\Delta t}{2} \frac{\partial^2 u}{\partial t^2} + O(\Delta t^2 + \Delta x^2).$$

$\tau \rightarrow 0$ when $\Delta x \rightarrow 0$ and $\Delta t \rightarrow 0$ => the scheme is consistent.

A counter example: The Dufort-Frankel method for the same diffusion equations:

$$\frac{u_i^{n+1}-u_i^{n-1}}{2\Delta t}=K\frac{(u_{i+1}^n+u_{i-1}^n)-(u_i^{n+1}+u_i^{n-1})}{\Delta x^2}.$$

It's a centered-in-time scheme that is 2nd-order accurate in both space and time.

We can find again the truncation error (do it yourself!)

$$\tau=\frac{K(\Delta x)^2}{12}\frac{\partial^4 u}{\partial x^4}-K\left(\frac{\Delta t}{\Delta x}\right)^2\frac{\partial^2 u}{\partial t^2}-\frac{(\Delta t)^2}{6}\frac{\partial^3 u}{\partial t^3}+H.R.T.$$

We can see that $\lim_{\Delta x,\Delta t\to 0}\tau\to 0$ except for the second term.

If $\lim_{\Delta x,\Delta t\to 0}\frac{\Delta t}{\Delta x}=0$ then the scheme is consistent therefore Δt must approaches zero faster then Δx.

If they approaches zero at the same rate, then $\lim_{\Delta x,\Delta t\to 0}\frac{\Delta t}{\Delta x}=\beta$, then

$$\lim_{\Delta x,\Delta t\to 0}\tau=-K^2\beta^2\frac{\partial^2 u}{\partial t^2},$$

our equation becomes

$$\frac{\partial u}{\partial t}+K^2\beta^2\frac{\partial^2 u}{\partial t^2}=K\frac{\partial^2 u}{\partial x^2}$$

Thus, we are solving the wrong equation. In fact this equation is hyperbolic instead of parabolic.

Note that if $\frac{\Delta t}{\Delta x} \sim 1$ you might see spurious waves in your solution, due to the hyperbolic nature of the "new" PDE.

2.3.3. Convergence

General Discussion

Definition is given earlier. Symbolically, it is

$$\lim_{\Delta x, \Delta t \to 0} u_i^n = u(x,t) .$$

Convergence is generally hard to prove, especially for nonlinear problems. The Lax's Theorem we presented earlier is very helpful in understanding the convergence for linear systems, and is often extended to nonlinear systems.

We will also discuss numerical convergence and methods for measuring solution accuracy later.

We will first show a convergence proof for a diffusion problem. Certain concept introduced will be useful later.

Convergence proof for a 1-D diffusion problem

Consider

$$\frac{\partial u}{\partial t} = K\frac{\partial^2 u}{\partial x^2} \quad (K > 0 \text{ and constant}) \text{ for } 0 \le x \le L$$

which has an initial condition:

$$u(x, t = 0) = \sum_{k=1}^{\infty} a_k \sin(\frac{k\pi x}{L}) = f(x).$$

The B.C. is

$$u(0,t) = u(L,t) = 0.$$

This is a well-posed, linear initial value problem (notice the I.C. satisfies the B.C. as well).

First, let's **find the analytical solution** to the PDE.

Because the problem in linear, we need only to examine the solution for a single wavenumber *k*, and can assume a solution of the form:

$$u_k(x,t) = A_k(t)\sin(\frac{k\pi x}{L})$$

Here A_k is the amplitude and sin gives the spatial structure and the final solution should be the sum of all wave components. Note that this solution satisfies the boundary conditions.

Substituting the solution into the PDE, we obtain an ODE for the amplitude A_k:

$$\frac{dA_k(t)}{dt} = -\left(\frac{\pi k}{L}\right)^2 KA_k$$

→ $$\frac{d\ln(A_k)}{dt} = -K\left(\frac{\pi k}{L}\right)^2$$

→ $$A_k(t) = A_k(0)\exp\left[-K\left(\pi k / L\right)^2 t\right].$$

It says that the amplitude of the solutions for all wave numbers decreases with time.

From the I.C., $A_k(0) = a_k$, so we have

$$u_k(x,t) = a_k \exp\left[-K\left(\pi k / L\right)^2 t\right]\sin(\frac{k\pi x}{L})$$

and

$$u(x,t) = \sum_k u_k(x,t)$$

which is the analytical solution to the original diffusion equation.

A numerical approximation to the diffusion equation should converge to this solution as Δx, $\Delta t \rightarrow 0$.

Consider the forward-in-time centered-in-space (FTCS) scheme we derived earlier.

Goal: Show that $u_i^t \rightarrow u(x,t)$ as $\Delta x, \Delta t \rightarrow 0$.

First, **find the numerical solution**.

This time, we use the FDE and substitute a discrete Fourier series into the equation.

Let the I.C. be given by

$$f(x_i) = u_i^0 = \sum_{k=0}^{J} \tilde{a}_k \sin(\frac{k\pi x_i}{L}) \qquad \text{for i= 0, 1, 2, ….., J}$$

where J+1 = total number of grid points used to represent the initial condition.

The coefficient $\tilde{a}_k$ is given by a discrete Fourier transform:

$$\tilde{a}_k = \frac{2}{J} \sum_{i=0}^{J} f(x_i) \sin(\frac{k\pi x_i}{L}) \qquad \text{for } k\text{=0, 1, 2, …., } J.$$

Note 1: $L = J\Delta x$. As $\Delta x \rightarrow 0$, J$\rightarrow \infty$, the discrete Fourier series becomes continuous and $\tilde{a}_k \rightarrow a_k$.

Note 2: The number of harmonics or Fourier wave components that can be represented is a function of the number of grid points (J+1), which is the number of degrees of freedom. Spectral methods represent fields in terms of spectral components, whose amplitudes are solved for.

The wavenumber for the wave components is $\frac{k\pi}{L}$ in the above equations.

Recall that wavelength

$$\lambda = \frac{2\pi}{w.n.} = \frac{2\pi}{(\pi k / L)} = \frac{2L}{k}$$

where *k* is the number (index) of wave components.

Longest wavelength = ∞ , corresponding to wavenumber zero (k=0).

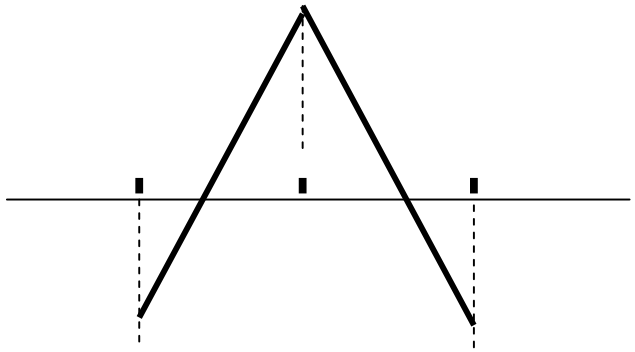
Next longest wave = 2L (k =1)

.
.
.

Shortest wave = 2L/J = 2JΔx/J = 2Δx.

Comments:

A 2Δx wave is the shortest wave that can be resolved on any grid and it takes at least 3 points to represent a wave.



2Δx waves often have some special properties. They are also represented most poorly by numerical methods – recall that smooth fields are more accurately represented by a finite number of grid points.

As in the continuous case, we examine only one wavenumber k (the solution is the sum of all waves), so for our discrete problem, assume a solution of the form

$$u_i^n = A_k(n)\sin\left(\frac{k\pi x_i}{L}\right)$$

(It satisfies B.C.) and *n* is the time level.

We also have from I.C.

$$A_k(0) = \tilde{a}_k.$$

Substituting this into the FDE

$$\frac{u_i^{n+1} - u_i^n}{\Delta t} = K\frac{u_{i-1}^n - 2u_i^n + u_{i+1}^n}{\Delta x^2},$$

and letting

$$S_i = \sin\left(\frac{k\pi x_i}{L}\right),$$

we have

$$\frac{A_k(n+1)S_i - A_k(n)S_i}{\Delta t} = \frac{KA_k(n)}{(\Delta x)^2}[S_{i+1} - 2S_i + S_{i-1}].$$

Since $x_i = i\,\Delta x$, $x_{i+1} = (i+1)\,\Delta x$ therefore

$$S_{i+1} = \sin\left(\frac{k\pi(x+\Delta x)}{L}\right).$$

Using standard trigonometric identities, we can write the above in the form of a recursion relation:

$$A_k(n+1) = A_k(n)\left[1 - 4\mu\sin^2\left(\frac{\pi k\Delta x}{2L}\right)\right]$$

where $\mu = \dfrac{K\Delta t}{(\Delta x)^2}$.

If we let $\mathrm{M(k)} \equiv \left[1 - 4\mu\sin^2\left(\frac{\pi k\Delta x}{2L}\right)\right]$, then we have

$$A_k(n+1) = M(k)A_k(n).$$

Write it out for n =0, 1, 2, …., n:

$$A_k(1) = M(k)A_k(0) = M(k)\tilde{a}_k$$
$$A_k(2) = M(k)A_k(1) = [M(k)]^2\tilde{a}_k$$
$$\vdots$$
$$A_k(n) = M(k)A_k(n-1) = [M(k)]^n\tilde{a}_k$$

we therefore have the solution of u for wave mode k:

$$(u_i^n)_k = A_k(n)S_i = \tilde{a}_k[M(k)]^n \sin\left(\frac{k\pi x_i}{L}\right).$$

Definition: $M(k)$ is known as the amplification factor, and if $|M(k)| \le 1$, the solution will not grow in time as $n \to \infty$. This had better to be the case because the amplitude of the analytical solution is supposed to always decrease with time.

For our problem we can see that $|M(\mathrm{k})| \le 1$ means

$$\left|1 - 4\mu \sin^2\left(\frac{\pi k \Delta x}{2L}\right)\right| \le 1.$$

If we take the maximum possible value of $\sin^2(\,) = 1$, →

$$-1 \le 1 - 4\mu \le 1$$

→ $\mu \le 1/2$.

This condition needs to be met for all *k* to prevent solution growth. Based on the definition of μ, the condition becomes

$$\Delta t \le \frac{(\Delta x)^2}{2K}.$$

This imposes an upper bound on the *Δt* that can be used for a given value of *Δx*, and such a condition is known as the Stability Constraint.

Now we have our solution, let's check convergence for single mode *k*.

By definition of convergence, we take

$$\lim_{\Delta x,\Delta t\to 0}(u_i^n)_k = \lim_{\Delta x,\Delta t\to 0}\tilde{a}_k\left[1-\Delta t\frac{4K}{(\Delta x)^2}\sin^2\left(\frac{\pi k\Delta x}{2L}\right)\right]^n \sin\left(\frac{k\pi x_i}{L}\right)$$

or if we let $f(\Delta x) \equiv \frac{4K}{(\Delta x)^2}\sin^2\left(\frac{\pi k\Delta x}{2L}\right)$,

$$\lim_{\Delta x,\Delta t\to 0}(u_i^n)_k = \lim_{\Delta x,\Delta t\to 0}\tilde{a}_k\left[1-\Delta t\, f(\Delta x)\right]^n \sin\left(\frac{k\pi x_i}{L}\right).$$

It can be shown that if $f(x)$ is a complex-valued function of a real argument, say Δx, such that

$\lim_{\Delta x \to 0} f(\Delta x) = a$, then,

$\lim_{\Delta x, \Delta t \to 0} [1 \pm \Delta t \, f(\Delta x)]^n = e^{\pm at}$ (we will show this later).

We know that $\lim_{y \to 0} \frac{\sin(y)}{y} = 1$, therefore

$$\lim_{\Delta x \to 0} f(\Delta x) = \lim_{\Delta x \to 0} \frac{K(\pi k)^2}{L^2} \left[\frac{\sin\left(\frac{\pi k \Delta x}{2L} \right)}{\left(\frac{\pi k \Delta x}{2L} \right)} \right]^2 = K \left(\frac{\pi k}{L} \right)^2 = a$$

Also $\lim_{\Delta x \to 0} \tilde{a}_k = a_k$, therefore

$$\lim_{\Delta x, \Delta t \to 0} (u_i^n)_k = a_k \exp\left[-K \left(\frac{\pi k}{L} \right)^2 t \right] \sin\left(\frac{\pi k x_j}{2L} \right).$$

Interestingly, this solution is identical to our analytical solution derived earlier! Therefore the numerical solution converges to the PDE solution when Δx, $\Delta t \rightarrow 0$.

Finally, we show here (noting that $n\Delta t = t$)

$$\lim_{\Delta x \to 0}(1 \pm f\Delta t)^n = 1 \pm fn\Delta t + \frac{n(n-1)}{2!}(f\Delta t)^2 + \frac{n(n-1)(n-2)}{3!}(f\Delta t)^3 + \ldots$$

$$= 1 \pm at + \frac{a^2}{2!}\left(1 - \frac{1}{n}\right)(n\Delta t)^2 \pm \frac{a^3}{3!}\left(1 - \frac{1}{n}\right)\left(1 - \frac{2}{n}\right)(n\Delta t)^3 + \ldots$$

$$= 1 \pm at + \frac{(at)^2}{2!} \pm \frac{(at)^3}{3!} + \ldots$$

$$= e^{\pm at}$$

2.3.4. Numerical Convergence

Reading: Fletcher (handout), Sections 4.1.2, 4.2.1, 4.4.1.

Numerical Convergence

Convergence is often hard to demonstrate theoretically.
True analytical solution is hard or even impossible to find is one of the reasons.

We can, however, find out the convergence of a given scheme numerically.

We compute solutions at successively higher resolutions and see how the error changes with the resolution.

Does $\tau \rightarrow 0$ when $\Delta x \rightarrow 0$?

And how fast τ decreases?

The procedure can be very expensive (remember the cost factor increase as Δ doubles).

A typical measure of error is the L2 norm or RMS error:

$$L2 = \sqrt{\sum \frac{[u_{i\Delta x} - u_i]^2}{n}}$$

where *u* is a true solution or a 'converged' numerical solution when exact solution is not available.

Example: 1-D diffusion equation using FTCS scheme:

$$\tau = \frac{\Delta t}{2}\frac{\partial^2 u}{\partial t^2} - K\frac{(\Delta x)^2}{12}\frac{\partial^4 u}{\partial x^4} + ...$$

Making use of

$$\frac{\partial u}{\partial t} = K\frac{\partial^2 u}{\partial x^2} \Rightarrow \frac{\partial^2 u}{\partial t^2} = K\frac{\partial}{\partial t}\left(\frac{\partial^2 u}{\partial x^2}\right) = K\frac{\partial^2}{\partial x^2}\left(\frac{\partial u}{\partial t}\right) = K^2\frac{\partial^4 u}{\partial x^4}.$$

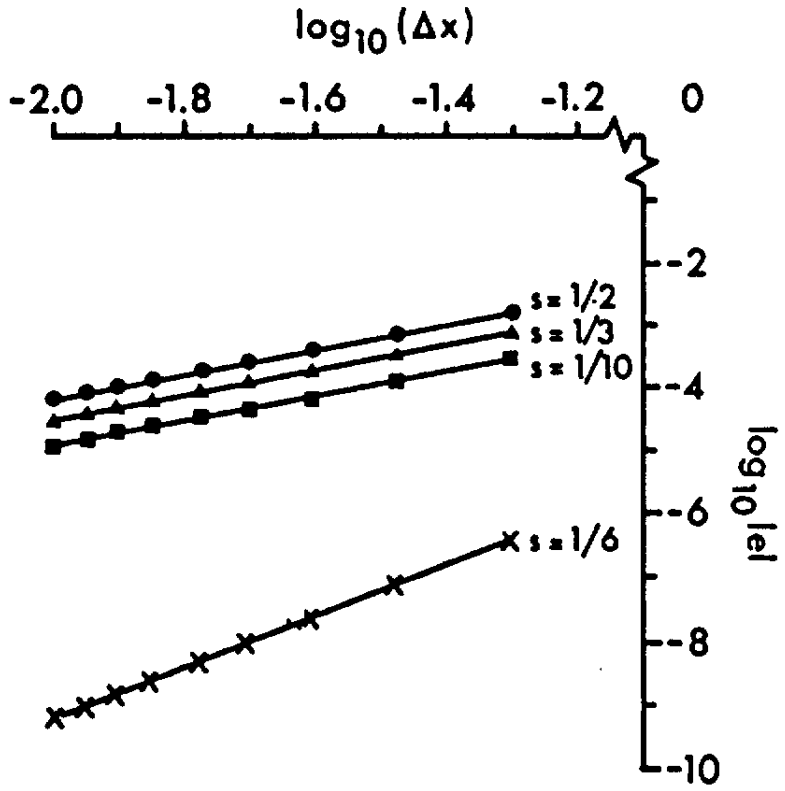
we have $\tau = \frac{K(\Delta x)^2}{2}\left(s - \frac{1}{6}\right)\frac{\partial^4 u}{\partial x^4} + O(\Delta x^4 + \Delta t^2).$

where $s = \frac{K\Delta t}{(\Delta x)^2}$

Table 4.1 (from Fletcher) shows the error reduction with Δx for two values of s.

Table 4.1. Solution error (rms) reduction with grid refinement

	rms error			
$s=\alpha\Delta t/\Delta x^2$	$\Delta x=0.2$	$\Delta x=0.1$	$\Delta x=0.05$	$\Delta x=0.025$
0.50	1.658	0.492	0.121	0.030
0.30	0.590	0.187	0.048	0.012



The above figure shows plots of $\log_{10}$(err) as a function of $\log_{10}(\Delta x)$.

Recall that

$$\tau = A\,(\Delta x)^n \quad \rightarrow$$

$$\log\,\tau = \log A + n \log \Delta x.$$

This is a straight line in log-log diagram with a slope of *n* and intercept *A*.

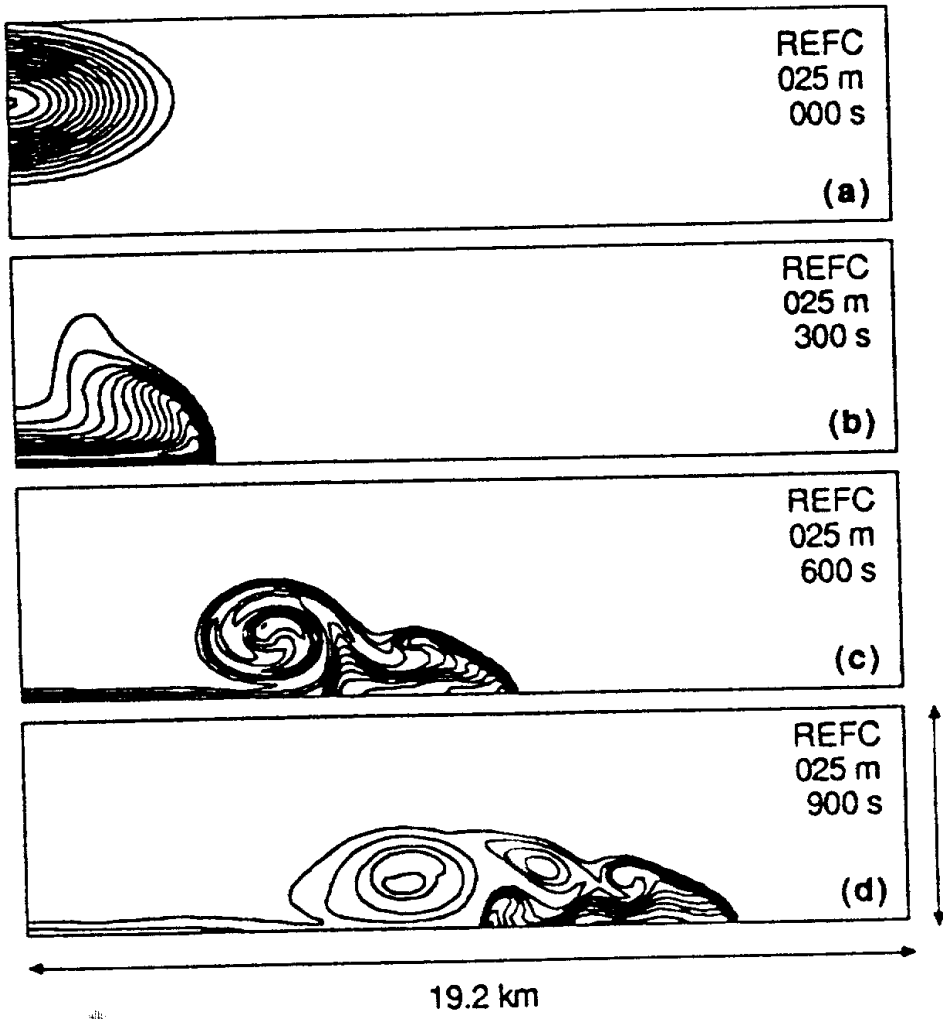
Thus the slope of the line gives the rate of convergence.

In the above figure, we see that when s = 1/6, the error line has a steeper slope and the error is smaller for all Δx. This is because for this value of s, the first term in τ drops out and the scheme because 4th-order accurate. The scheme is second-order accurate for all other values of s.

Note that you can choose Δx and Δt such that s=1/6 only when K is constant in the entire domain.

In cases where no exact solutions is available, a so-called 'grid-convergence' or reference solution is often sought and this solution can be used in the place of true solution in the estimating the solution error.

An example from Straka et al (1993).



It shows a reference solution obtained at Δx=25 m, for a density current resulting from a dropping cold bubble.

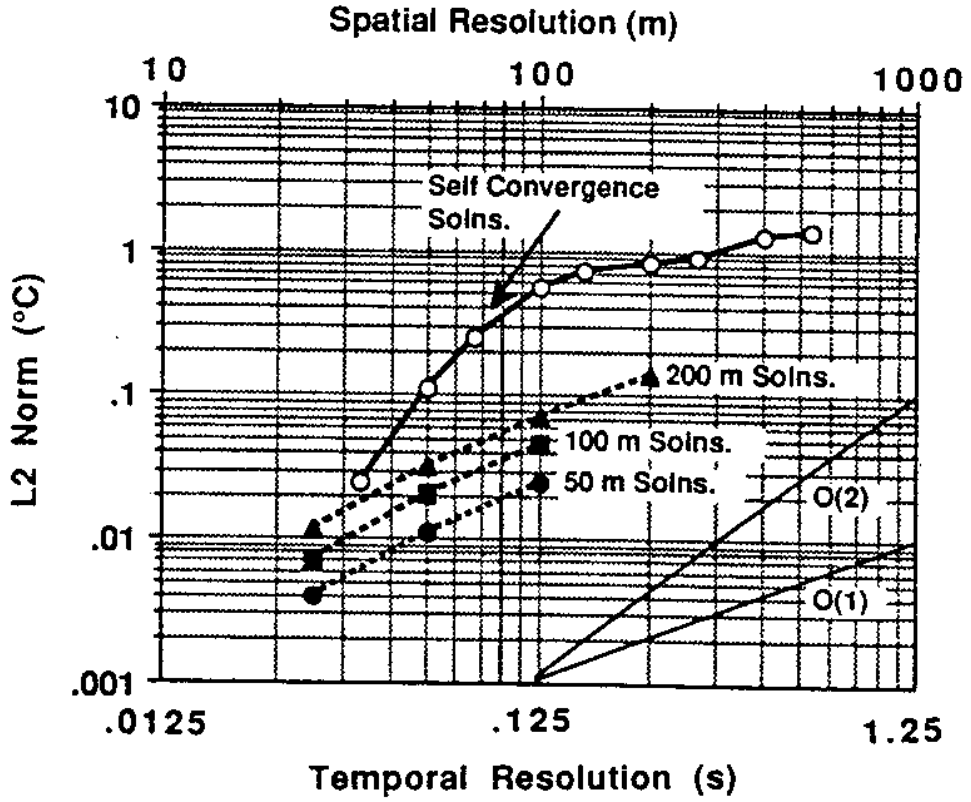


Figure (from Straka et al 1993). Graph of θ' L2 norms (°C) from self-convergence tests with the compressible reference model (REFC). The bold solid line labeled with 'self-convergence solutions' represents the L2 norms for spatial truncation errors of solutions made with Δt = constant and varying grid spacings. The L2 norms were computed against a 25.0 m reference solution. The bold dashed lines labeled with, for example, '200.0 m solutions' represent L2 norms for temporal truncation errors of solutions made with Δx =constant (e.g. 200.0 m) and varying time steps. The reference solutions for these computations were made using a time step consistent with Δt = 12.5 s times a constant in each of the cases. The solid fines labeled O(l) and O(2) represent first- and second-order convergence, respectively.

Richardson Extrapolation

As the grid becomes very fine, the error behaves much like that predicted from the leading terms in τ. Further refinements are expensive, so we use another technique to improve the solution – Richardson Extrapolation.

Consider two numerical solutions obtained at Δx_a and Δx_b.

With FTCS scheme (assuming $s \neq 1/6$),

$$\tau_a = \frac{K(\Delta x_a)^2}{2}\left(s - \frac{1}{6}\right)\frac{\partial^4 u}{\partial x^4} + O(\Delta x_a{}^4 + \Delta t^2)$$

$$\tau_b = \frac{K(\Delta x_b)^2}{2}\left(s - \frac{1}{6}\right)\frac{\partial^4 u}{\partial x^4} + O(\Delta x_b{}^4 + \Delta t^2)$$

Find a linear combination of the two solutions, u_a and u_b

$$u_c = a\, u_a + b\, u_b$$

where $a + b = 1$ and a and b are chosen so that the leading terms in τ_a and τ_b cancel and the scheme becomes 4th-order accurate (Of course this assumes that $\partial^4 u / \partial x^4$ is the same in both case, which is reasonably assumption only at relatively high resolutions when the solution is well resolved).

If $\Delta x_b = \Delta x_a/2$, then $4a + b = 0$

with $a + b = 1 \rightarrow a = -1/3$, $b = 4/3$ and $u_c = -1/3\, u_a + 4/3\, u_b$.

2.3.5. Stability Analysis

Reading: Tannehill et al. Section 3.6.

Stability – For a stable numerical scheme, the errors in the initial condition will not grow unboundedly with time.

In this section, we discuss the methods for determining the stability of F.D. schemes. This is very important when designing a F.D. scheme and for understanding its behavior.

There are several methods:

- Energy method
- von Neumann method
- Matrix method (for systems of equations)
- Discrete perturbation method (will not discuss)

Note: Stability refers to the F.D.
Does not involve B.C. or I.C.
Refers to time-matching problems only

The energy method

Read Durran Section 2.2 (handout, see web link)

This method is used much less often than the von Neumann method.

It's attractive because it works for nonlinear problem and problems without periodic B.C.

The key is to show that a positive definite quantity like $\sum_i (u_i^n)^2$ is bounded for all *n*.

We illustrate this method using the upstream-forward scheme for wave (advection) equation $\frac{\partial u}{\partial t} + c\frac{\partial u}{\partial x} = 0$:

$$\frac{u_i^{n+1} - u_i^n}{\Delta t} + c\frac{u_i^n - u_{i-1}^n}{\Delta x} = 0$$

Let $\mu = c\Delta t/\Delta x$ →

$$u_i^{n+1} = (1-\mu)u_i^n + \mu u_{i-1}^n$$

Squaring both sides and summing over all grid points:

$$\sum_i (u_i^{n+1})^2 = \sum_i [(1-\mu)^2 (u_i^n)^2 + 2(1-\mu)\mu u_i^n u_{i-1}^n + \mu^2 (u_{i-1}^n)^2] \qquad (1)$$

Assuming periodic B.C. →

$$\sum_i (u_i^n)^2 = \sum_i (u_{i-1}^n)^2$$

and using the Schwarz inequality (which says that for two vectors *U* and *V*, $|U \cdot V| \le |U| \cdot |V|$),

$$\sum_i u_i^n u_{i-1}^n \le \sqrt{\sum_i (u_i^n)^2}\ \sqrt{\sum_i (u_{i-1}^n)^2} = \sum_i (u_i^n)^2 .$$

If $\mu(1-\mu) \ge 0$, all coefficients of RHS terms in (1) are positive and we have

$$\sum_i (u_i^{n+1})^2 \le [(1-\mu)^2 + 2(1-\mu)\mu + \mu^2] \sum_i (u_i^n)^2 = \sum_i (u_i^n)^2 ,$$

i.e., the L2 norm at n+1 is no greater than that at n, therefore the scheme is stable!

The condition $\mu(1-\mu) \ge 0$ gives

$$\mu = c\ \Delta t/\Delta x\ \le 1$$

which is the stability condition for this scheme. As we discussed earlier, it says that waves cannot propagate more than one grid interval during one Δt in order to maintain stability.

von Neumann method

Read Tannehill et al, Section 3.6.1.

In a sense, we have already used this method to find stability of the 1-D diffusion equation – when we were proving the convergence of the solution using FTCS scheme.

We found then

$$u^{n+1} = [\ M(t)\ ]^{n+1}\ u^0$$

where $M(t)$ = amplification factor.

If $M(t) \leq 1$ by some measure (M can be a matrix or a complex number), then $u^{n+1} \leq u^n$ and the solution cannot grow in time – the scheme is stable.

Essentially, von Neumann method expands the F.D.E. in a Fourier series, finds the amplification factor and determines under what condition the factor is less than or equal to 1 for stability.

Assumptions:

1. The equation has to be Linear with constant coefficients.
2. It is assumed that the solution is periodic.

With this method, the dependent variable is decomposed into a complex (or a real) Fourier series:

$$u(x,y,z,t) = \sum_{k,l,m} U(k,l,m) \exp[i(kx + ly + mz - \omega t)] \qquad (2)$$

where U is the complex amplitude and $\omega = \omega_R + i\ \omega_I$ is the complex frequency.

In fact ω_R gives the wave propagation speed and ω_I gives the growth and decaying rate.

$$e^{-i\omega t} = e^{-i[\omega_R + i\omega_I]t} = e^{-i\omega_R t} e^{\omega_I t}$$

$e^{-i\omega_R t}$ - phase function of Fourier components
$e^{\omega_I t}$ - growth or decay rate

If $\omega_I > 0$, the solution will grow exponentially in time.

Example 1: 1-D diffusion equation with FTCS scheme.

$$u_i^{n+1} - u_i^n = \mu(u_{i-1}^n - 2u_i^n + u_{i+1}^n) \qquad (3)$$

where $\mu = \dfrac{K\Delta t}{(\Delta x)^2}$.

Let's examine a single wave *k*:

$$u_i^n = U_k e^{i(kx-\omega t)}$$
$$u_i^{n+1} = U_k e^{i(kx-\omega t)} e^{-i\omega\Delta t}$$
$$u_{i\pm1}^n = U_k e^{i(kx-\omega t)} e^{\pm ik\Delta x}$$

Substitute the above into (3) →

$$U_k e^{i(kx-\omega t)}(e^{-i\omega\Delta t} - 1) = \mu U_k e^{i(kx-\omega t)}[e^{ik\Delta x} - 2 + e^{-ik\Delta x}] \rightarrow$$

$$U_k e^{i(kx-\omega t)}[e^{-i\omega\Delta t} - 1 - 2\mu(\cos k\Delta x - 1)] = 0 \rightarrow$$

$$U_k e^{i(kx-\omega t)}[e^{-i\omega\Delta t} - 1 + 4\mu\sin^2(k\Delta x/2)] = 0$$

For non-trivial solution, we require

$$e^{-i\omega\Delta t} - 1 + 4\mu \sin^2(k\Delta x/2) = 0 \rightarrow$$

$$e^{-i\omega\Delta t} = 1 - 4\mu \sin^2(k\Delta x/2)$$

Here, $e^{-i\omega\Delta t}$ is actually the amplification factor, the same as the *M* discussed earlier.

$\lambda \equiv e^{-i\omega\Delta t} = u^{n+1}/u^n$ - the amplification factor.

For stability we require

$|\lambda| \le 1 \rightarrow$

$-1 \le 1 - 4\mu \sin^2(k\Delta x/2) \le 1 \rightarrow$

$\mu \le 1/2$ as before!

In practice, when μ=1/2, the solution (amplification factor) switches between –1 and +1 for $2\Delta x$ waves ($\sin^2(k\Delta x/2) = 1$) very other step, which is unrealistic. The standard requirement is therefore

$\mu \le 1/4$.

Therefore, do not naively think diffusion terms in a numerical model does not cause numerical instability! When integrated stably, the diffusion term in CFD models tends to stabilize the solution by killing off/damping small scale waves, but when stability condition is not met, it tem itself will cause problem!

Read Pielke (1984) section 10.1.2 (handout, see web link).

Shorthand notations for discrete/finite difference operators and discretization identities

Notations:

$$\overline{A}^{nx} = \frac{A_{j+n/2} + A_{j-n/2}}{2}$$

$$\delta_{nx} x = \frac{A_{j+n/2} - A_{j-n/2}}{n\Delta x}$$

$$\delta_{+x} x = \frac{A_{j+1} - A_j}{\Delta x}$$

$$\delta_{-x} x = \frac{A_j - A_{j-1}}{\Delta x}$$

Identities:

$$\delta_{2x} A \equiv \delta_x \overline{A}^x \equiv \overline{\delta_x A}^x$$

$$A\delta_x B \equiv \delta_x (\overline{A}^x B) - \overline{B\delta_x A}^x$$

$$\overline{A}^x \delta_x A \equiv \delta_x (A^2/2)$$

2.3.6. Implicit Methods

Read Tannehill et al, second part of section 3.4.1.

So far, we have dealt with only explicit schemes which have the form of

$$u^{n+1} = f(u^n, u^{n-1}, \ldots).$$

With these schemes, the future state at each grid point is only dependent on the current and past time levels, therefore the solution can be obtained directly or explicitly. c.f., explicit functions such as $y = x^2$.

Implicit scheme involves variables of the future time level at more than one grid point (often resulting from finite difference of variable(s) at the future time level). Mathematically it can be expressed as

$$u^{n+1} = f(u^{n+1}, u^n, u^{n-1}, \ldots).$$

This is analogous to implicit functions such as $x = \sin(x)$. As one can imagine, implicit schemes are more difficult to solve. Usually matrix inversion is involved.

We will first look at the stability property of an implicit scheme.

Example. Consider the 1-D diffusion equation $u_t = K\, u_{xx}$ again.

It is approximated by the following F.D. scheme:

$$\delta_{+t} u_i = K[\alpha \delta_{xx} u_i^{n+1} + (1-\alpha)\delta_{xx} u_i^n] \qquad (4)$$

(Note: shorthand notations for F.D. are used. See Appendix. e.g., $\delta_{+t}u = (u^{n+1} - u^n)/\Delta t$).

- When α = 0, the scheme is explicit, and is the FTCS scheme discussed earlier.
- When α = 1/2, it is implicit and is called Crank-Nicolson scheme.
- For other values of α, it is a general implicit scheme.

We can show that

$$\tau = K\frac{\partial^4 u}{\partial x^4}\left[K\Delta t\left(\frac{1}{2}-\alpha\right)-\frac{(\Delta x)^2}{12}\right]+O(\Delta x^4+\Delta t^2)$$

(show if for yourself!).

We can see that when α = 1/2, it is 2nd-order accurate in time and space. Otherwise, it's first-order in time – which is expected for un-centered time-differencing scheme (when α=1/2, the right hand side is an averaged between the current and future time levels valid at n+1/2. Relative to this RHS, the LHS time difference becomes centered in time. We know that the simplest centered difference scheme is second-order accurate).

When [ ] = 0, the scheme becomes fourth-order in space.

Let's perform stability analysis on (4) using von Neumann method.

$$u_j^n = U_k e^{i(kx-\omega t)} = U_k e^{-i\omega n\Delta t} e^{ikj\Delta x} \equiv U_k \lambda^n e^{ikj\Delta x} \qquad (5)$$

Here $\lambda \equiv e^{-i\omega\Delta t}$. Note that we are now using j as the grid point index.

Substitute (5) into (4) →

$$U_k e^{ikj\Delta x}(\lambda^{n+1}-\lambda^n)=\mu U_k e^{ikj\Delta x}\left[\alpha\lambda^{n+1}+(1-\alpha)\lambda^n\right](e^{ik\Delta x}-2+e^{-ik\Delta x})\rightarrow$$

Dividing $U_k e^{ikj\Delta x}\lambda^n$ on both sides, and rearranging →

$$\lambda-1=-4\mu\sin^2(k\Delta x/2)[\alpha\lambda+(1-\alpha)]$$

$$\lambda=\frac{1-4(1-\alpha)\mu\sin^2(k\Delta x/2)}{1+4\alpha\mu\sin^2(k\Delta x/2)}$$

Look at several cases:

Case I: $\alpha = 0,\ \lambda=1-4\mu\sin^2(k\Delta x/2)$ → $\mu \le 1/2$ as before. The scheme is <u>conditionally</u> stable.

Case II: $\alpha = 1/2$ (Crank-Nicolson)

$$|\lambda|=\left|\frac{1-2\mu\sin^2(k\Delta x/2)}{1+2\mu\sin^2(k\Delta x/2)}\right|\le 1 \quad \text{for all values of } \mu,$$

therefore the scheme is <u>absolutely or unconditionally</u> stable.

Case III: α =1, the time difference is backward, relative to the RHS terms.

$$|\lambda| = \left| \frac{1}{1 + 4\mu \sin^2(k\Delta x/2)} \right| \le 1$$, again for all values of μ,

therefore the scheme is also absolutely stable. However, this scheme is only first-order accurate in time, as discussed earlier (consistent with the time difference scheme being un-centered).

In general, when $0 \le \alpha < 1/2$, it is required that $\mu \le 1/(2 - 4\alpha)$, therefore the scheme is conditionally stable. When $1/2 \le \alpha \le 1$, the scheme is unconditionally stable (it is sometimes referred to as the forward-biased scheme).

In the ARPS, the implicit diffusion scheme is an option for treating the vertical turbulent mixing terms. This treatment is necessary in order to remove the severe stability constraint from these terms when vertical mixing is strong inside the planetary boundary layer (PBL). The latter occurs when the PBL is convectively unstably and the non-local PBL mixing is invoked with the Sun and Chang (1986) parameterization. Parameter *alfcoef* in arps.input corresponds to 1-α here (see handout).

Finally, we note that for multi-time level schemes, there is usually multiple solutions for the amplification factor λ. some of them might represent spurious computational modes due to the use of extra (artificial) initial conditions.

The expression of λ can be too complicated so that a graphic plotting is needed to understand its dependency on wave number k. $|\lambda|$ has to be no greater than 1 for all possible waves. The shortest wave resolvable on a grid has a wavelength of 2 Δ, and the longest is 2L, where L is the domain width.

Tridiagonal Solver

1-D implicit method often leads to tridiagonal systems of linear algebraic equations.

(In the ARPS, this appears twice – once when sound waves are treated implicitly in the vertical direction and once when the vertical turbulence mixing is treated implicitly).

For example, Eq. (4) can be rewritten as

$$\frac{u_i^{n+1}-u_i^n}{\Delta t}=\alpha K\frac{u_{i-1}^{n+1}-2u_i^{n+1}+u_{i+1}^{n+1}}{\Delta x^2}+(1-\alpha)K\frac{u_{i-1}^{n}-2u_i^{n}+u_{i+1}^{n}}{\Delta x^2}\rightarrow \qquad (6)$$

It can be rearranged into

$$u_i^{n+1}-\alpha\frac{\Delta tK}{\Delta x^2}(u_{i-1}^{n+1}-2u_i^{n+1}+u_{i+1}^{n+1})=(1-\alpha)\frac{\Delta tK}{\Delta x^2}(u_{i-1}^{n}-2u_i^{n}+u_{i+1}^{n})+u_i^n\rightarrow$$

$$-\alpha\frac{\Delta tK}{\Delta x^2}u_{i-1}^{n+1}+(1+2\alpha\frac{\Delta tK}{\Delta x^2}u_i^{n+1})-\alpha\frac{\Delta tK}{\Delta x^2}u_{i+1}^{n+1}=d_i^n.$$

where $d_i^n=\Delta t(1-\alpha)K\frac{u_{i-1}^{n}-2u_i^{n}+u_{i+1}^{n}}{\Delta x^2}+u_i^n$.

Let $A_i=C_i=-\frac{\alpha\Delta tK}{\Delta x^2}$, $B_i=(1+\frac{2\alpha\Delta tK}{\Delta x^2})$, we then have

$$A_iu_{i-1}^{n+1}+B_iu_i^{n+1}+C_iu_{i+1}^{n+1}=d_i^n, \qquad (7)$$

for i = 1, 2, …, N-1, assuming the boundaries are at i =0 and N.

If we have Dirichlet boundary conditions, i.e., u at i=0 and N are known, then for i =1, the equation becomes

$$B_1 u_1^{n+1} + C_1 u_2^{n+1} = d_1^n - A_1 u_0^{n+1} \tag{8}$$

and for $i = N$, the equation is

$$A_{N-1} u_{N-2}^{n+1} + B_{N-1} u_{N-1}^{n+1} = d_{N-1}^n - C_{N-1} u_N^{n+1}. \tag{9}$$

For i = 2, 3, …., N-2, the equation remains of the form in Eq.(7).

If we write the equations (7-9) in a matrix form, we have

$$\begin{bmatrix} B_1 & C_1 & & & & & \\ A_2 & B_2 & C_2 & & & & \\ & . & . & . & & & \\ & & A_i & B_i & C_i & & \\ & & & . & . & . & \\ & & & & A_{N-2} & B_{N-1} & C_{N-2} \\ & & & & & A_{N-1} & B_{N-1} \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ . \\ u_i \\ . \\ u_{N-2} \\ u_{N-1} \end{bmatrix} = \begin{bmatrix} D_1 \\ D_2 \\ . \\ D_i \\ . \\ D_{N-2} \\ D_{N-1} \end{bmatrix} \tag{10}$$

where

$D_i = d_i^n$ for i=2, …, N-2,

$$D_1 = d_1^n - A_1 u_0^{n+1},$$
$$D_{N-1} = d_{N-1}^n - C_{N-1} u_N^{n+1}.$$

If we have Neumann boundary conditions, i.e., we know the gradient of u at the boundaries which in discretized form are $u_1 - u_0 = L$ and $u_N - u_{N-1} = R$. Plug these relations into Eq.(7) for *i*=1 and *i*=*N*-1, we obtain equations similar to (8) and (9):

$$(A_1 + B_1)u_1^{n+1} + C_1 u_2^{n+1} = d_1 + A_1 L \qquad (11)$$
$$A_{N-1} u_{N-2}^{n+1} + (B_{N-1} + C_{N-1})u_{N-1}^{n+1} = d_{N-1} - C_{N-1} R. \qquad (12)$$

In this case, the final coefficients in (10) are different for the first and last equation.

Since in each except for the first and last row of the coefficient matrix, only three elements are non-zero and the non-zero elements of the matrix are aligned along the diagonal axis, this system is called tridiagonal system of equations. It can be solved efficiently using Thomas Algorithm.

The procedure consists of two parts. First, Eq. (10) is manipulated into the following form:

$$\begin{bmatrix} 1 & C'_1 & & & & & \\ & 1 & C'_2 & & & & \\ & . & . & . & & & \\ & & & 1 & C'_i & & \\ & & & . & . & . & \\ & & & & & 1 & C'_{N-2} \\ & & & & & & 1 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ . \\ u_i \\ . \\ u_{N-2} \\ u_{N-1} \end{bmatrix} = \begin{bmatrix} D'_1 \\ D'_2 \\ . \\ D'_i \\ . \\ D'_{N-2} \\ D'_{N-1} \end{bmatrix} \qquad (13)$$

in which the sub-diagonal coefficients A are eliminated and the diagonal coefficients are normalized. For the first equation

$$C'_1 = \frac{C_1}{B_1}, \qquad D'_1 = \frac{D_1}{B_1}. \tag{14a}$$

For the general equations:

$$C'_i = \frac{C_i}{B_i - A_i C'_{i-1}}, \qquad D'_i = \frac{D_i - A_i D'_{i-1}}{B_i - A_i C'_{i-1}} \quad \text{for i=2, …, } N\text{-1.} \tag{14b}$$

Equations in (14) represent a forward sweep step (see figure below). It is followed by a backward substitution step that finds solution u_i from (13). The solution is:

$$\begin{aligned} u_{N-1} &= D'_{N-1} \\ u_i &= D'_i - u_{i+1} C'_i \quad \text{for } i \text{ from } N-2 \text{ to } 1. \end{aligned} \tag{15}$$

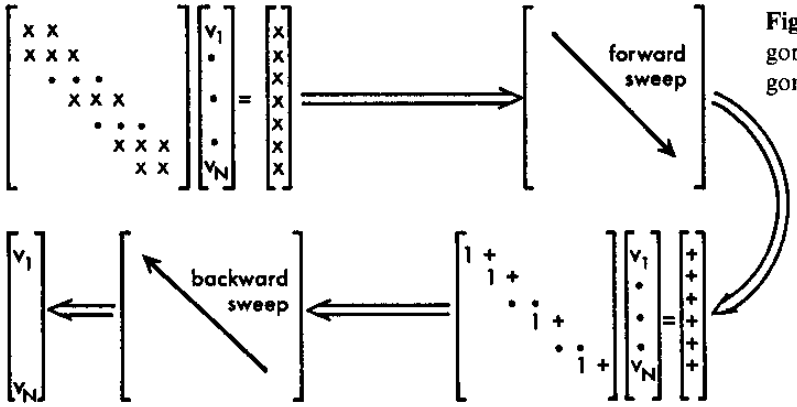


Fig. 6.17. The Thoma
gorithm for solving a t
gonal system of equati

Note both (14) and (15) involve reduction, the algorithm is inherently non-parallelizable. Fortunately, for multi-dimensional problems, multiple systems of equations often need to be solved, and one can exploit parallelism along other dimensions (e.g., *j* instead of *i* direction).

Read Section 4.3.3 of Tannehill et al and Appendix A.

2.3.7. Stability Analysis for Systems of Equations

When we are dealing with a system of equations, we can also apply the von Neumann method to find the stability property of a given F.D. scheme.

As with single equations, von Neumann can only be used for linear systems of equations. For nonlinear systems, linearization has to be performed first.

Without going into details, we point out that a system of linear equations can be expressed in a matrix form like

$$\frac{\partial \mathbf{u}}{\partial t} + \mathbf{A}\frac{\partial \mathbf{u}}{\partial x} = 0 \qquad (16)$$

The equation is first discretized using certain F.D. scheme, **u** can be written in terms of a discrete Fourier series and the wave component is then substituted into the discrete equation to obtain something like:

$$\mathbf{U}_k^{n+1} = \mathbf{M}(\Delta t, \Delta x)\,\mathbf{U}_k^n \qquad (17)$$

where $\mathbf{U}_k^n$ is the amplitude vector for wave *k* at time level *n*, and **M** is called the amplification matrix.

The scheme is stable when the maximum absolute eigenvalue of **M** is no greater than 1.

Why the maximum absolute eigenvalue?

Because as you saw earlier (in Chapter 1) that a system of equation like (17) can be transformed into a system of decoupled equations, and the eigenvalues of **M** become the amplification factors for each of the new dependent variables, v_i (the element of vector **V**), i.e., we can obtain from (17)

$$\mathbf{V}_k^{n+1} = \mathbf{N}\,\mathbf{V}_k^n ,$$

where **N** is a diagonal matrix with the eigenvalues of **M** as its diagonal elements.

$$\mathbf{T}^{-1}\,\mathbf{M}\,\mathbf{T} = \mathbf{N}$$

$$\mathbf{T}^{-1}\mathbf{U}_k^{n+1} = \mathbf{T}^{-1}\mathbf{M}\,\mathbf{T}\mathbf{T}^{-1}\mathbf{U}_k^n$$

Let $\mathbf{V}_k^n = \mathbf{T}^{-1}\,\mathbf{U}_k^n$, we have

$$\mathbf{V}_k^{n+1} = \mathbf{N}\mathbf{V}_k^n \ .$$

Therefore $(v_i)_k^{n+1} = \lambda_i\,(v_i)_k^n$ where λ_i are the eigenvalues.

Since the system is stable only when all dependent variables remain bounded, the absolute value of the maximum eigenvalue has to be no greater than 1.

Read section 3.6.2 of Tannehill et al.