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 $\Delta x, \Delta t \rightarrow 0$.

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).

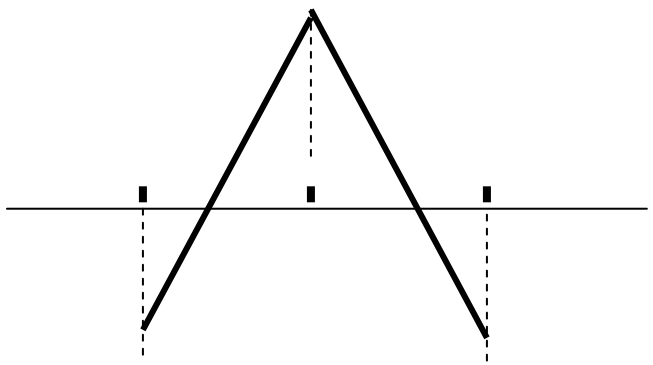
Next longest wave = 2L (k =1)

.
.
.

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 $M(k) \equiv \left[1 - 4\mu\sin^2\left(\dfrac{\pi k\Delta x}{2L}\right)\right]$, 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 $\sin^2(\,) = 1$, →

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

$\rightarrow \mu \leq 1/2.$

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

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

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

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

Be 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 <u>identical</u> 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

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

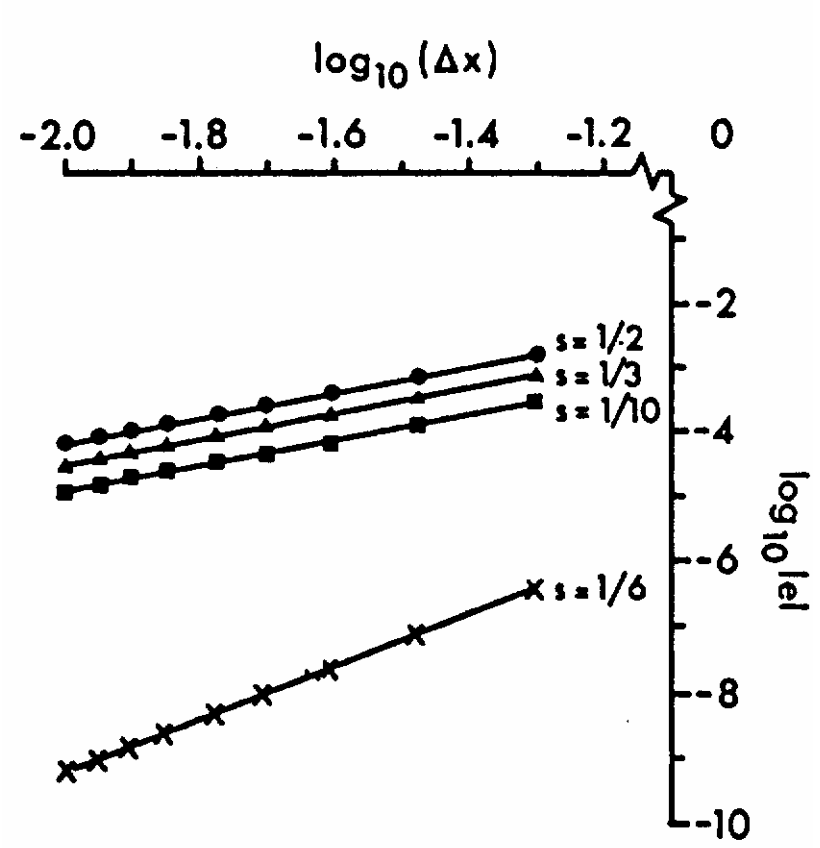
we have $\tau = \frac{K(\Delta x)^2}{2}\left(s - \frac{1}{6}\right)\frac{\partial^4 u}{\partial x^4} + O(\Delta x^4 + \Delta t^2)$.

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.

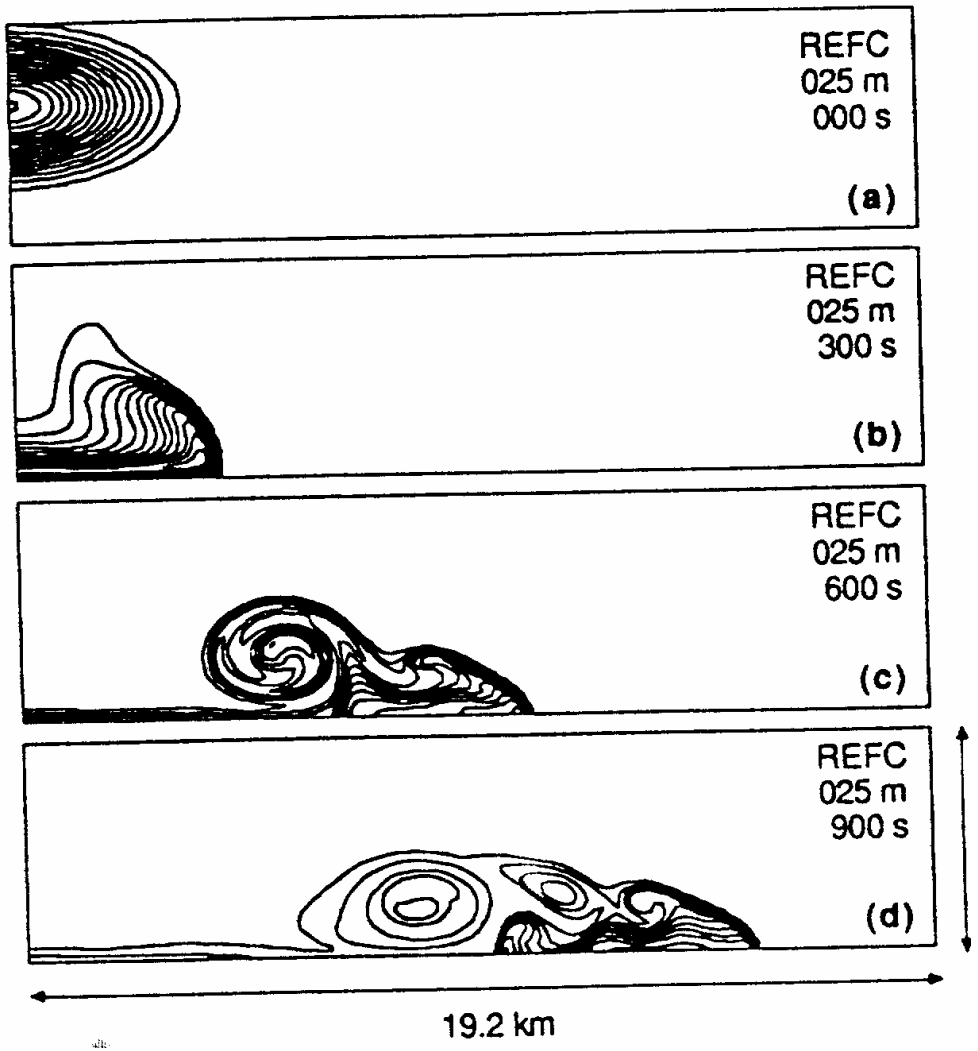
Thus the slope of the line gives the rate of convergence.

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.

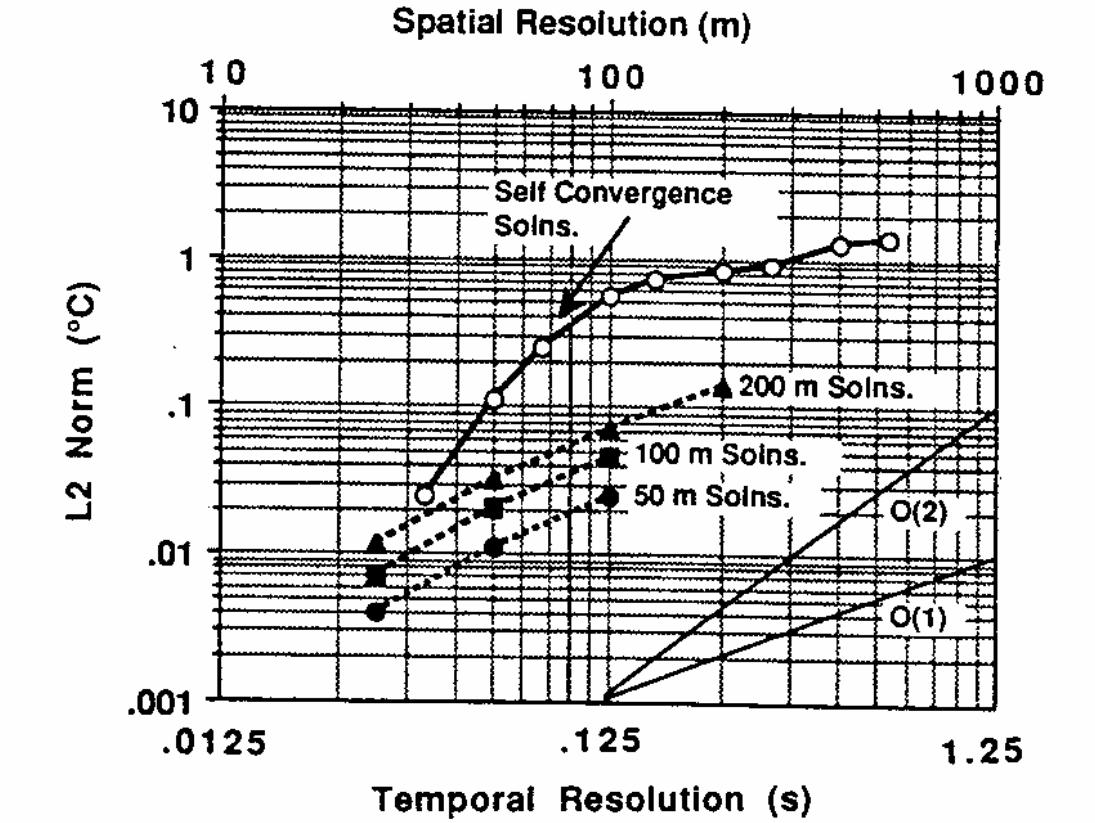


Figure (from Straka et al 1993). Graph of θ' L2 norms (°C) from self-convergence tests with the compressible reference model (REFC). The bold solid line labeled with 'self-convergence solutions' represents the L2 norms for spatial truncation errors of solutions made with Δt = constant and varying grid 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 $a + b = 1$ →

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