QUANTITATIVE ERROR ANALYSIS OF NUMERICAL METHODS FOR PARTIAL DIFFERENTIAL EQUATIONS

By

R. Vichnevetsky

K. W. Tu

J. A. Steen

DCS-TR-#28

Reprinted from Proceedings
of the Eighth Annual
Princeton Conference on
Information Science and
Systems, Princeton University,
March 1974

Department of Computer Science
Rutgers University
The State University of New Jersey
New Brunswick, New Jersey 08903

Proceedings of the Eighth Annual Princeton Conference on Information Science and Systems, Princeton University, March 1974

QUANTITATIVE ERROR ANALYSIS OF NUMERICAL METHODS FOR PARTIAL DIFFERENTIAL EQUATIONS

R. Vichnevetsky, K. W. Tu, J. A. Steen

Department of Computer Science
Rutgers University, New Brunswick, N. J.

1. Introduction

The classical theory of numerical methods for partial differential equations is concerned to a large extent with problems of consistency, stability and convergence of algorithms. While these aspects of the theory are useful in establishing the validity of difference approximations, they fail to provide quantitative measures of errors which may be used in practice to characterize the effect of truncation. Nor do they provide relevant criteria whereby a specified accuracy requirement in a numerical solution may be converted into the choice of algorithms and/or the choice of grid sizes. When one attempts to develop such criteria, one readily finds out that they are very much dependent upon the initial and boundary conditions of the problem: by contrast, the classical properties of consistency, stability and convergence depend upon those initial/boundary conditions in a much weaker sense. It also becomes readily apparent that something other than the conventional tools must be used to perform such quantitative error analyses. One approach to this problem, which has proved to be useful in several case studies, is outlined in the remainder of this paper.

2. Models of computational processes

In attempting to predict the magnitude of errors resulting from the use of difference methods for partial differential equations, it is justified to use "models" which are only approximations of these methods, inasmuch as (a) these models are analytically more tractable than the actual computation and (b) the predicted errors are reasonable approximations of the actual errors.

As a case in point, consider the simple advection equation

$$\frac{\partial u}{\partial t} + V \frac{\partial u}{\partial x} = 0 \tag{1}$$

and its six-point implicit approximation

$$\frac{u_n^{j+1} - u_n^j}{\Delta t} + V\left[\theta \frac{u_{n+1}^{j+1} - u_{n-1}^{j+1}}{2\Delta x} + (1-\theta)\frac{u_{n+1}^{j} - u_{n-1}^{j}}{2\Delta x}\right] = 0 \tag{2}$$

$u_n^j \cong u(x_n, t^j)$; $x_n = n\Delta x$; $t^j = j\Delta t$; $0 \le \theta \le 1$

where truncation errors are due to non-zero spatial increments Δx and non-zero time increments Δt. We use this equation here strictly as an example, suggesting that similar procedures are applicable to more complex cases. As a prelude to analysing these errors one may postulate the following principle of independence: when errors are "small" one may analyse separately the effect of each discretization (i.e. space into Δx, and time into Δt) and add the individual contributions to obtain an estimate of the global error. For lack of space, we shall not elaborate upon the validity of this principle, except to the extent of indicating that it is based on exactly the same premises as those used in the classical error theory of experimental physics [1]. As a consequence of this viewpoint, we may derive two "models" of (2)

1) Model to be used in the analysis of errors due to $\Delta x \neq 0$

This model is obtained by letting $\Delta t \to 0$ in (2);

$$\frac{du_n}{dt} + V \frac{u_{n+1} - u_{n-1}}{2\Delta x} = 0 \tag{3}$$

$u_n(t) \cong u(x_n, t)$

2) Model to be used in the analysis of errors due to $\Delta t \neq 0$

This model is obtained by letting $\Delta x \to 0$ in (2);

$$\frac{u^{j+1} - u^j}{\Delta t} + V\left[\theta \frac{du^{j+1}}{dx} + (1-\theta)\frac{du^j}{dx}\right] = 0 \tag{4}$$

$u^j(x) \cong u(x, t^j)$

Note that these models ((3)-(4) above or other) are not intended to yield information about the numerical stability of the algorithms analysed. The assumption of smallness of the errors is equivalent to restricting the analysis to algorithms which are known (by other means) to be numerically stable.

3. Analysis by the use of a frequency method

One of the ways in which the error due to replacement of (1) by either (3) or (4) may be analyzed is by assuming a sinusoidal boundary condition

$$u(o,t) = A \cdot \sin \Omega t \tag{5}$$

from which the exact solution of (1) is known to be:

$$u(x,t) = A \cdot \sin \Omega (t - x/V) \tag{6}$$

Solutions of the models (3) and (4) in response to this boundary condition are expressible also in sinusoidal form, but with an amplitude A* and phase velocity V* which are in general not equal to A and V. The quantitative aspect of the frequency approach stems from the fact that non-sinusoidal solutions may be synthesized by the superposition of sinusoidal solutions. It should be noted in passing that there are other ways in which the models (3) and (4) may be used: step changes in the boundary condition u(o,t) lead to solutions analytically expressible in terms of Bessel functions for (3) and in terms of Laguerre functions for (4).

Frequency Analysis of Δx-related errors

Let $E^*(\delta)$ be the numerical step transfer function of the model (3), defined as:

$$E^*(\delta) \triangleq \frac{\mathcal{L}_t(u_{n+1}(t))}{\mathcal{L}_t(u_n(t))} \tag{7}$$

$\mathcal{L}_t(\cdot)$ being the classical Laplace Transform of $(\cdot)$ with $\delta = \frac{d}{dt}$.

To obtain an expression for $E^*(\delta)$, we take the Laplace Transform of (3), thus obtaining:

$$E^*(\delta)^{-1}[E^*(\delta)^2 + \frac{2\cdot\Delta x}{V}\cdot\delta\cdot E^*(\delta) - 1]\cdot\mathcal{L}_t(u_n(t)) = 0 \tag{8}$$

By solving the characteristic equation in (8) (and by eliminating the unstable root which would correspond to $u(\infty,t) = \infty$), we find:

$$E^*(\delta) = -\frac{\delta\cdot\Delta x}{V} + \sqrt{(\frac{\delta\cdot\Delta x}{V})^2 + 1} \tag{9}$$

It may also be shown (see e.g. [2]) that an equivalent expression to (9) is:

$$E^*(\delta) = e^{-\mathrm{Arg\ Sh}(\frac{\delta\cdot\Delta x}{V})} \tag{10}$$

By contrast, the exact step transfer function $E(\delta)$ defined as the counterpart of (7) for the exact equation is:

$$E(\delta) \triangleq \frac{\mathcal{L}_t(u(x_{n+1},t))}{\mathcal{L}_t(u(x_n,t))} = e^{-\frac{\delta\cdot\Delta x}{V}} \tag{11}$$

We may rewrite (6) as:

$$u(x_n,t) = A_n\cdot\sin(\Omega\cdot t - \varphi_n) \tag{12}$$

where the amplitude A_n and phase φ_n are, respectively:

$$A_n = A\cdot|E(i\Omega)|^n = A \tag{13}$$
$$\varphi_n = -n\cdot\angle E(i\Omega) = n\cdot\Omega\cdot\Delta x/V = \Omega x_n/V$$

By constrast, the solution of the model (3) of the numerical approximation is:

$$u_n(t) = A^*_n\cdot\sin(\Omega\cdot t - \varphi^*_n) \tag{14}$$

where the amplitude and phase are now:

$$A^*_n = A\cdot|E^*(i\Omega)|^n = A\cdot|\frac{i\Omega\Delta x}{V} + \sqrt{1 - (\frac{\Omega\Delta x}{V})^2}\,|^n \tag{15}$$

$$\varphi^*_n = -n\cdot\angle E^*(i\Omega) = -n\cdot\angle(\frac{i\Omega\Delta x}{V} + \sqrt{1 - (\frac{\Omega\Delta x}{V})^2}) \tag{16}$$

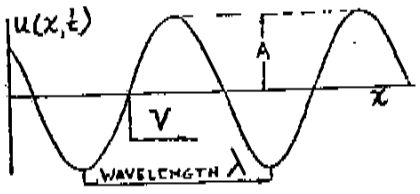

Fig. 1
Exact solution of Equation (1)

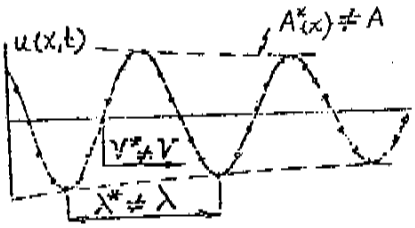

Fig. 2
Numerical solution of Equation (1)

It is also useful to define the phase velocity V^* of numerical solutions by the identity:

$$u_n(t) = A^*_n\sin\Omega(t - x_n/V^*)$$

which by comparison with (14) and (15) yields:

$$V^* = -\frac{\Omega\cdot\Delta x}{\angle E^*(i\Omega)} \tag{17}$$

Using the expression (10) for $E^*(\delta)$ we find the analytical expression (valid for $\Omega\cdot\Delta x/V \leq 1$):

$$V^* = V^*(\Omega) = \Omega\cdot\Delta x/\arcsin(\Omega\cdot\Delta x/V) \tag{18}$$

As we may observe, the velocity V^* is (by contrast with V) a function of the frequency Ω.

Differences between E^* and E on the one hand, and V^* and V on the other hand describe the amplitude and phase velocity errors due to $\Delta x \neq 0$ in the numerical approximation. These are illustrated graphically in Figures 3 and 4. About these figures we may make the following observations:

a) The amplitude error in numerical sinusoidal solutions of (1) described by the model (3) is equal to zero when $\Omega\cdot\Delta x/V \leq 1$. For larger values of $\Omega\cdot\Delta x/V$, the amplitude of sinusoidal solutions decays rapidly as one moves away from the boundary $x = 0$ where the boundary condition $u(o,t) = A\cdot\sin(\Omega t)$ is applied.

b) Under the same conditions V^*/V is close to 1 for sufficient small $\Omega\Delta x/V$. The phase velocity error $V^*(\Omega) - V$ becomes non-negligible when $\Omega\Delta x/V$ becomes larger than about .5. The phase velocity error for $\Omega\Delta x/V < \pi/2$ is negative. Numerical sinusoidal components in that range travel slower than the exact velocity V. Non-constancy of the phase velocity V^* is responsible for a phenomenon of spurious dispersion in the solution, and appears most clearly when sharp variations in u are present.

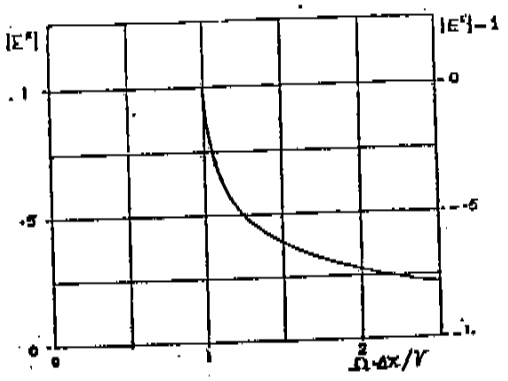

Fig. 3

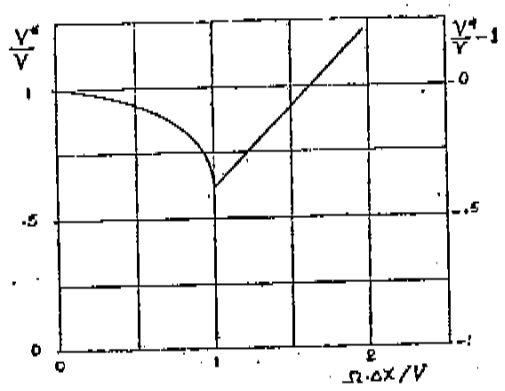

Fig. 4

Analysis of Δt-related errors

We may likewise express the exact solution (6) at the discrete instants $t^j = j \cdot \Delta t$

$$u(x,t^j) = A \cdot \sin\Omega(j \cdot \Delta t - x/V) \tag{19}$$

To find the response of the model (4) to the boundary condition (5) we seek by identification a solution of the form:

$$u^j(x) = A^*(x)\ \sin \cdot \Omega(t^j - B \cdot x) \tag{20}$$

It is reasonably straightforward to substitute (20) in (4). One obtains an expression containing terms in $\sin\Omega(t^j - Bx)$ and in $\cos\Omega(t^j - Bx)$. Equating these two groups of terms to zero provides us with two equations

$$A^*[\cos\Omega\Delta t - 1] + V\Delta t \frac{dA^*}{dx}[\theta \cos\Omega\Delta t + (1-\theta)] - V B\Omega\Delta t\ A^* \sin\Omega\Delta t = 0 \tag{21}$$

$$A^* \sin\Omega\Delta t + V\Delta t \frac{dA^*}{dx} \sin\Omega\Delta t + V B\Omega\Delta t\ A^*[\theta \cos\Omega\Delta t + (1-\theta)] = 0 \tag{22}$$

which may be solved into

$$\left(\frac{1}{A^*} \cdot \frac{dA^*}{dx}\right) = \frac{(1 - \cos\Omega\Delta t)(1-2\theta)}{[1+2(\theta-\theta^2)(\cos\Omega\Delta t - 1)] \cdot V\Delta t} \tag{23}$$

$$B = \frac{\sin\Omega\Delta t}{[1 + 2(\theta-\theta^2)(\cos\Omega\Delta t - 1)]\ \Omega V\Delta t} \tag{24}$$

The function $\left(\frac{1}{A^*} \frac{dA^*}{dx}\right)$ represents a spurious variation of the amplitude with x due to $\Delta t \neq 0$. This term would be equal to zero if there were no amplitude error, whilst B contains a measure of the phase error of numerical solutions due to the same factor.

To obtain a more useful expression, we observe that rewriting (20) in the form:

$$u^j(x) = A^*(x)\ \sin\ \Omega(t^j - x/V^*) \tag{25}$$

defines implicitly the numerical phase velocity $V^*(\Omega)$ as:

$$V^*(\Omega) = \frac{1}{B} = V \cdot \frac{[1 + 2(\theta-\theta^2)(\cos(\Omega\Delta t)-1)]\Omega\Delta t}{\sin(\Omega\Delta t)} \tag{26}$$

Amplitude and velocity errors corresponding to (23) - (26) are illustrated in Fig. (5).

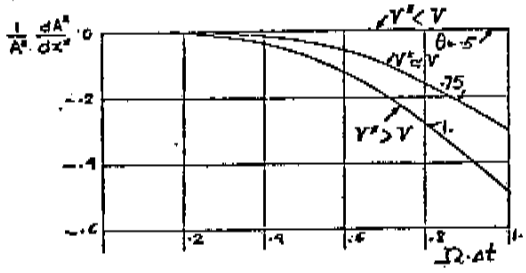

Fig. 5

As we may observe, errors becomes non-negligible in this case when $\Omega\Delta t > .5$

4. An example and numerical verification

An example of the preceding theories was provided in the process of developing a computer program to compute tidal flows in the Delaware estuary. The equations under consideration are well known (see e.g. ref. [4,5]).

Conservation of Mass: $$\frac{\partial h}{\partial t} + \frac{\partial}{\partial x}(h \cdot v) = 0 \tag{27.a}$$

Conservation of Momentum:

$$\frac{\partial v}{\partial t} + \frac{\partial}{\partial x}(gh + \frac{v^2}{2}) = g(s - \frac{k}{h^{m_1}}|v|^{m_2} \operatorname{sgn} v) = G(h,v) \tag{27.b}$$

where h = height
v = velocity
s = slope
g = acceleration of gravity
k, m_1, m_2 = positive constants describing the friction force

The problem at hand consisted in having to choose the largest values of Δx and Δt such that numerical solutions with an accuracy of a few percent (i.e. consistent with the kind of accuracy one deals with in this kind of work) would be obtained.

The system of equation (27) is quasi-linear and hyperbolic. What plays an identical role to the velocity V in the preceding sections of this paper is not the velocity v(x,t) of the water, but the velocities v_1 and v_2 of the two wave equations

which are obtained by transforming (27) into its characteristic form

$$\left.\begin{aligned} &\frac{\partial w_1}{\partial t} + v_1 \frac{\partial w_1}{\partial x} = G \quad ; w_1 = v + 2\sqrt{g \cdot h}\ ; v_1 = v + \sqrt{g \cdot h} \\ &\frac{\partial w_2}{\partial t} + v_2 \frac{\partial w_2}{\partial x} = G \quad ; w_2 = v - 2\sqrt{g \cdot h}\ ; v_2 = v - \sqrt{g \cdot h} \end{aligned}\right\} \tag{28}$$

v_1 and v_2 are the eigenvalues of the matrix coefficients of $\frac{\partial v}{\partial x}$ and $\frac{\partial h}{\partial x}$ in (27)). Tidal flows in an estuary are the response of (27) to a sinusoidal imposed boundary condition

$$h(x_n,t) = A\cdot\sin \Omega t \qquad (29)$$

where $\Omega = 2\pi/12.5$ hours $= .5$ rad/hour (12.5 hours in the tidal period.) The second boundary in our case was $v(o,t) = 0$ (corresponding to the presence of a dam at Trenton, N. J.).

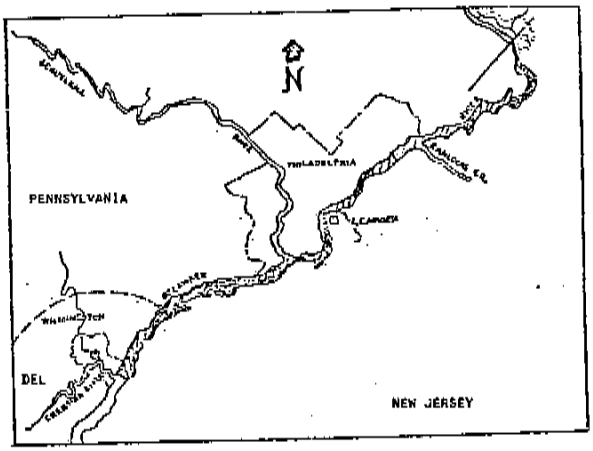

Fig. 6

Flow conditions which are typical of the Delaware River south of Trenton are

$$7 < v_1 < 18 \quad \text{and} \quad -11.5 < v_2 < -3.5 \text{ ft/second} \qquad (30)$$

A value of Δt was first chosen to satisfy $\Omega\cdot\Delta t < .5$ (i.e. $\Delta t < 1$ hour; we chose $\Delta t = .5$ hours). Variations of the numerical solution as a function of varying Δx were then observed. The worst case occurs where either $|v_1|$ or $|v_2|$ reaches a minimum. This minimum is seen from (30) to be $|v|_{min} = 3.5$ ft/sec $\simeq 2.5$ mph and is presumed to occur at least for some length of time during each tidal period. According to the theory of section 3, numerical results should therefore be relatively insensitive to changes in Δx (i.e. remain sufficiently accurate) as long as the condition

$$\left.\begin{array}{l} \Omega\Delta x/|v_{min}| < .4 \\ \text{or } \Delta x < 2 \text{ miles} \end{array}\right\} \qquad (31)$$

is not being violated.

Results shown in Fig. 7 bear this prediction out.

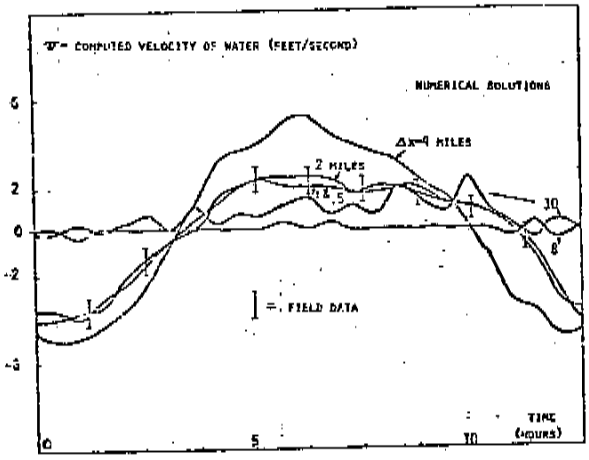

Fig. 7

With $\Delta x = 1$ mile, several numerical simulations were made with increasing values for Δt. The theoretically predicted evidence of significant errors when $\Delta t > 1$ hour was clearly apparent. Likewise, numerical solutions with $\Delta x > 2$ miles and $\Delta t > 1$ hour were observed to be significantly in error.

5. Conclusions

The use of models of computational processes, augmented by the principle of independence formulated in Section 1 has been shown to provide, at least in some instances, useful ways to predict quantitatively the errors which occur in numerical solutions of partial differential equations and to obtain insight into the process of their generation.

What this suggests is that there is plenty of room for investigation in the direction of a search for tractable models of computational processes, as opposed to the too frequent use of more straightforward, classical but hard to handle tools which leave much to be desired in terms of results which are useful in practice.

6. Acknowledgements

It is our pleasure to acknowledge the assistance of several graduate students at Rutgers University who have participated in the discussions and have programmed several examples (including the one shown in section 4 of this paper) in support of the theory. Particular credit goes to F. H. Aydin and Y. S. Shieh in this respect. Credit is also due to Professors R. Ahlert and B. Davidson for the interest which they have shown in these studies.

This work was supported in part by Grant RR 643 of the Biotechnology Resources Branch of the NIH.

7. References

[1] Beers, Y. (1957) "Introduction to the theory of Errors", Addison Wesley Publishing Co., Inc.

[2] Vichnevetsky, R. and Shieh, Y. S. (1972). "On the numerical method of lines for one-dimensional water quality equations". DCS Report #20. Dept. of Computer Science, Rutgers University, New Brunswick, New Jersey.

[3] Vichnevetsky, R. (1973). "Physical criteria in computer methods for partial differential equations". Proceedings, 7th International AICA Congress. Reprinted in Proceedings of AICA, Vol. 16, No. 1, January 1974. European Academic Press, Brussels.

[4] Eagleson, P. S. (1970). "Dynamic Hydrology". McGraw-Hill, New York.

[5] Vichnevetsky, R. and Aydin, F. M. (1973). "The accuracy of estuarine tidal flow computations". NAM 57, Department of Computer Science, Rutgers University, New Brunswick, N. J.

[6] Stone, H. L. and Brian, P.L.T. (1963). "Numerical solution of connective transport problems". A.I. Ch. E Journal, Vol. 9, No. 5.