LA--1287

c. 3

LOS ALAMOS SCIENTIFIC LABORATORY

OF THE

UNIVERSITY OF CALIFORNIA

CONTRACT W-7405-ENG. 36 WITH

U. S. ATOMIC ENERGY COMMISSION

C.13

LOS ALAMOS SCIENTIFIC LABORATORY

of

THE UNIVERSITY OF CALIFORNIA

August 22, 1951

LA-1287

CRITERIA OF STABILITY FOR THE NUMERICAL SOLUTION OF PARTIAL DIFFERENTIAL EQUATIONS

Work done by:

Herbert A. Forrester

Report written by:

Herbert A. Forrester

PHYSICS AND MATHEMATICS

PHYSICS AND MATHEMATICS

LA-1287

OCT 2 1951

Los Alamos Document Room	20
J R. Oppenheimer	1

STANDARD DISTRIBUTION

American Cyanamid Company	1
Argonne National Laboratory	8
Armed Forces Special Weapons Project	1
Atomic Energy Commission - Washington	6
Battelle Memorial Institute	1
Brush Beryllium Company	1
Brookhaven National Laboratory	4
Bureau of Medicine and Surgery	1
Bureau of Ships	1
Carbide and Carbon Chemicals Company (C-31), Paducah	2
Carbide and Carbon Chemicals Company (K-25)	4
Carbide and Carbon Chemicals Company (Y-12)	4
Columbia University (Failla)	1
Columbia University (Havens)	1
Du Pont de Nemours and Company	5
Eldorado Mining and Refining Ltd.	2
General Electric Company, Richland	3
Idaho Operations Office	4
Iowa State College	2
Kansas City Operations Branch	1
Kellex Corporation	2
Kirtland Air Force Base	2
Knolls Atomic Power Laboratory	4
Mallinckrodt Chemical Works	1
Massachusetts Institute of Technology (Kaufmann)	1
Mound Laboratory	3
National Advisory Committee for Aeronautics	1
National Bureau of Standards (Taylor)	1
Naval Medical Research Institute	1
Naval Radiological Defense Laboratory	2
New Brunswick Laboratory	1
New York Operations Office	3
North American Aviation, Inc.	1
Oak Ridge National Laboratory (X-10)	8
Patent Branch, Washington	1
Rand Corporation	1

Sandia Corporation	1
Savannah River Operations Office	1
Technical Information Service, Oak Ridge	75
USAF, Major James L. Steele	6
U. S. Geological Survey	2
U. S. Public Health Service	2
University of California at Los Angeles	1
University of California Radiation Laboratory	5
University of Rochester	2
University of Washington	1
Western Reserve University	1
Westinghouse Electric Corporation	4

SUPPLEMENTARY DISTRIBUTION

Atomic Energy Project, Chalk River (The Library)	4
Dr. Gregory Breit (Sloane Physics Laboratory)	1
Carnegie Institute of Technology (Dr. E. Creutz)	1
Chief of Naval Research	1
H. K. Ferguson Company (Miss Dorothy M. Lasky)	1
Harshaw Chemical Company (Mr. K. E. Long)	1
Isotopes Division (Mr. J. A. McCormick)	1
Library of Congress (Alton H. Keller)	2
National Bureau of Standards (The Library)	1
National Research Council, Ottawa (Mr. J. M. Manson)	1
Naval Research Laboratory (Code 2028)	15
Nevis Cyclotron Laboratories (Mr. M. W. Johnson)	2
Oak Ridge Institute of Nuclear Studies (Mr. R. A. Schlueter)	1
United Kingdom Scientific Mission	10
USAF, Central Air Documents Office (CADO-E)	5
USAF, Director of Research & Development (Research Division)	1
USAF, Wright-Patterson Air Force Base (Colonel Leo V. Harman)	1
U. S. Army, Army Field Forces (CWO Edwin H. Hoffman)	1
U. S. Army, Army Medical Service Graduate School (Col. W. S. Stone)	1
U. S. Army, Chemical and Radiological Laboratories (Miss Amoss)	3
U. S. Army, Director of Operations Research Office (Dr. E. Johnson)	1
U. S. Army, Office, Chief of Ordnance (Col. A. R. Del Campo)	1
U. S. Army, Office of the Chief Signal Officer (Mr. N. Stulman thru Lt. Col. G. C. Hunt)	2
U. S. Army, Special Weapons Branch (Lt. Col. A.W. Betts)	1
UT-AEC Agricultural Research Program (C. L. Comar)	1

TABLE OF CONTENTS

CRITERIA OF STABILITY FOR THE NUMERICAL SOLUTION OF PARTIAL DIFFERENTIAL EQUATIONS

by
Herbert A. Forrester

1. INTRODUCTION.

As an introduction to the problem of stability in the numerical solution of partial differential equations, we can do no better than quote the opening paragraphs of [OHK 1]*:

"One of the most common and useful methods employed in the numerical integration of partial differential equations involves the replacement of the differential equation by an equivalent difference equation. This technique has become particularly important in recent years because of the development of modern high-speed computing machines.

"In the present paper we shall show that the accuracy of a finite difference solution to a partial differential problem is conveniently discussed in terms of the 'convergence' and 'stability' of the difference scheme. Courant, Friedrichs and Lewy [(CFL 1)] discussed the convergence of difference solutions for the basic types of linear partial differential equations; for equations of parabolic or hyperbolic character, they found the important result that the 'mesh ratio' must satisfy certain inequalities. J. von Neumann obtained the same inequalities from a study of error growth (stability of the difference scheme). The partly heuristic technique of stability analysis developed by von Neumann was applied by him to a wide variety of difference and differential equation problems during World War II.

*Numbers in brackets refer to the bibliographs at the end of Section 3).

"We begin with terminology and definitions. Let D represent the *exact solution* of the partial *differential* equation, Δ represent the *exact solution* of the partial *difference* equation, and N represent the *numerical solution* of the partial *difference* equation. We call $(D-\Delta)$ the truncation error, it arises because of the finite distance between points of the difference mesh. To find the conditions under which $\Delta \rightarrow D$ is the problem of *convergence*. We call $(\Delta-N)$ the numerical error. If a faultless computer working to an infinite number of decimal places were employed, the numerical error would be zero. Although $(\Delta-N)$ may consist of several kinds of error, we usually consider it limited to round-off errors. To find the conditions under which $(\Delta-N)$ is small throughout the entire region of integration is the problem of *stability*.

"Whether a given finite-difference scheme satisfies the criteria for convergence and stability (we say, for short, that the difference-scheme is convergent/divergent and stable/unstable) depends on the form of the Δ-equation and upon the initial and boundary conditions. If the Δ-equation is linear, stability (and usually convergence also) will not depend on the initial and boundary conditions. Now for most problems, D and Δ are either unavailable or can only be obtained with much greater labor than is involved in finding N. The principal problem in the numerical solution of partial differential equations is to determine N such that (D-N) is smaller than some preassigned allowable error throughout the whole region considered. We can assert that

$$(D-N) \equiv (D-\Delta) + (\Delta-N)$$

is small for a numerical calculation over a fine mesh using a stable, convergent difference scheme. Sometimes, for convenience or from necessity, a convergent but unstable difference scheme is used; then provision must be made for controlling the error-growth (See Ref.[H 1], where the numerical solution of elliptic problems is discussed; here the governing difference equations are inherently unstable).

"In this paper we shall be interested, for the most part, in partial differential equations of parabolic or hyperbolic type, for which the data is naturally given on an open curve (or surface) from which the solution is stepped-off. Many remarks will apply, however, to elliptic problems. We shall mostly discuss equations of the second order in two independent variables, but extensions to more than two variables will be obvious, though algebraically more complex. In the first part of this paper, we shall give a method for determining the stability of partial differential equations and shall discuss implicit difference schemes. In the second part, we shall work with a simple parabolic problem and investigate directly the magnitudes of the truncation error and the numerical error for various methods of numerical solution; in particular, when (D-N) is large, we shall ask whether _lack of convergence or lack of stability is chiefly responsible for the discrepancy_. We shall find that very often in such cases, the truncation error overshadows the numerical error, contrary to what is generally thought.

"In studying the effect of round-off errors fed into the calculation (the problem of stability), we may ask:

a) Does the over-all error due to all round-offs

$$\begin{pmatrix}\text{Grow}\\ \text{Now grow}\end{pmatrix}\text{? This we term }\underline{\text{strong}}\begin{pmatrix}\text{Instability}\\ \text{Stability}\end{pmatrix}\text{ ?}$$

b) Does a single, general, round-off error

$$\begin{pmatrix}\text{Grow}\\ \text{Not grow}\end{pmatrix}\text{? }\underline{\text{This we term weak}}\begin{pmatrix}\text{Instability}\\ \text{Stability}\end{pmatrix}\text{ ?}$$

We mean "growth" during the uninterrupted stepping-ahead of the solution, where no use is made of special devices applied from time to time to limit the error growth (See end of second paragraph above). What we need to know in our numerical work is whether a given difference equation is strongly stable or strongly unstable. It is much easier, however, to demonstrate weak stability or instability. The gap between the two types of stability is closed by the following

Assumption: Weak $\begin{pmatrix}\text{Stability}\\ \text{Instability}\end{pmatrix}$ implies strong $\begin{pmatrix}\text{Stability}\\ \text{Instability}\end{pmatrix}$.

In the following text, then, whenever we refer to the stability of instability of a difference scheme we shall mean the weak form. We intend to examine rather closely in another paper the validity of the Assumption; for the present we note that it is true for all those calculations we have seen where care was taken that the round-off errors should be random. (As pointed out by Huskey and Hartree, in the Journal of Research of the National Bureau of Standards, vol. 42, pp. 57-62, round-off errors may be non-random in certain regions of integration.

They observe that randomness may be regained by carrying extra figures in calculating these regions. For general purposes, the assumption of random round-off errors is probably a good one).

"It is important to note that the overall error may be considered as the sum of the individual errors fed in (modified from step to step by the numerical process) because the variational equation which governs error propagation is always linear and solutions may be superposed. For studying weak stability, we may adopt either of two procedures:

1) Consider a unit error introduced at an arbitrary mesh point and follow its progress.
2) Make a Fourier expansion of a line of errors and follow the progress of the general term of the expansion.

The first procedure occurs occasionally in the literature but, to our knowledge, has not been developed in any systematic way; such a development has now been completed by R. P. Eddy (Ref. [E 1]). The second procedure was developed and used by J. von Neumann during World War II, but has never been published by him*. With his permission we present below some of Professor von Neumann's results."

The paragraphs quoted from [OHK 1] provide an outline of the fundamentals of the problems we will consider. In outline: We will first consider existence, uniqueness and stability equations for the solutions of linear partial differential equations particular in-so-far as they are relevant to the stability of numerical solutions; secondly, we will develop the stability criteria of von Neumann and R. P.Eddy; and thirdly,

*But see reference [NR 1] .

we will consider applications of the general theories to the special cases of certain parabolic and hyperbolic equations.

Bibliography is given at the end of Sections 2, 3, and 4; numbered references are to the bibliography at the end of Article 4.

2. EXISTENCE QUESTIONS.

We consider the equation

$$L[u] = \sum a_{i_1 i_2 \ldots i_n} \frac{\partial^{i_1+\ldots+i_n} u}{\partial x_1^{i_1}\, \partial x_2^{i_2} \ldots \partial x_n^{i_n}} = B$$

where $a_{i_1, i_2, \ldots i_n}$ and B are analytic functions; the relevant results will carry over to systems of such equations.

Generally, $L[u] = B$ has an infinity of solutions; the essential problem is to find additional conditions which will specify a unique solution u. Since the equation often represents a physical problem, it is essential that small (observationally unavoidable) "errors" in the $a_{i_1, i_2, \ldots, i_n}$, B, and the determining data should appear in the solution u as small errors. These criteria (existence, uniqueness, and stability) cause a subdivision of problems concerning differential equations into hyperbolic, parabolic and elliptic differential equations and problems; while this classification is not complete in case of equations of order higher than two it will serve for our purposes.

The form in which the additional data is given is usually the specification of u and some of its derivatives or normal derivatives along a surface S, together with the requirement (especially in the elliptic case) that the solution exist throughout some specified region. The relevant classification ties certain types of operators L with certain forms in which data is specified, and classifies surfaces S with respect to the operator L; this classification is based on the consideration of certain characteristic surfaces (or hypersurfaces) connected with L.

Let L be of order m, i.e., $a_{i_1 i_2 \ldots i_n} = 0$ for $i_1+i_2+\ldots+i_n > m$, while for some $i_1+i_2+\ldots+i_n = m$, $a_{i_1 i_2 \ldots i_n} \neq 0$. Let $Q(\xi_1, \ldots, \xi_n)$ be a form defined in an (n-1)-dimensional projective space for a point $x^* = (x_1^*, x_2^*, \ldots, x_n^*)$ by

$$Q(\xi_1, \ _2, \ldots, \xi_n) = \sum_{i_1+\ldots+i_n=m} a_{i_1 i_2 \ldots i_n}(x_1^*, x_2^*, \ldots, x_n^*)\, \xi_1^{i_1} \xi_2^{i_2} \ldots \xi_n^{i_n}.$$

A normal to a surface S at x* is of a singular nature with respect to L if its components $(y_1, \ldots, y_n)$ satisfy the equation

$$Q(y_1, y_2, \ldots, y_n) = 0.$$

If this equation *does not* hold at x*, we say that S is *free* at x* with respect to L. If S is free at every point, it is called non-characteristic for L. Generally, data for L can be specified only on non-characteristic surfaces.

The normal derivatives of a function $u(x_1, \ldots, x_n)$ at a point $x^* = (x_1^*, \ldots, x_n^*)$ of a surface S, whose normal at x* is $\xi^* = (\xi_1^*, \ldots, \xi_n^*)$ are defined as the functions

$$\left(\sum_{i=1}^{n} \xi_i^* \frac{\partial}{\partial x_i} \right)^m u$$

for m = 0, 1, 2, ...

Two problems, and corresponding types of equations, can now be described.

I) The Elliptic Case. If $Q(\xi_1, \ldots, \xi_n) = 0$ has no real points in projective space, we say that $L[u] = B$ is an elliptic equation. In this case, any real surface S is non-characteristic. The problem associated with this equation is the boundary value problem: The surface S is to be a closed surface, the values of the solution $u(x_1, \ldots, x_n)$ are specified on S, and the solution is required to exist throughout the interior of S.

II) The Hyperbolic (And Parabolic) Case. The equation Q = 0 has real solutions in projective space. The problem associated with this case is the initial value or Cauchy problem: The surface S is required to be open and the values of u and its first (m-1) normal derivatives are specified on S, where m is the order of L.

In either case S is assumed to be given by an equation

$$f(x_1, x_2, \ldots, x_n) = 0$$

where f is an analytic function; and the functions specifying u or its normal derivatives on S are to be analytic on S.

The solution in either case is unique, and is stable in the sense that small variations in the functions specifying u and/or its normal derivatives on S, in the function f specifying S, in the coefficients

$a_{i_1 i_2 \ldots i_n}$ of L, and in the function B in L[u] = B are reflected in small variations in the solution u throughout some neighborhood of S, which neighborhood will grow to the full domain of existence as the magnitude of the variations becomes small.

In the elliptic case the function u exists throughout the region bounded by S, except possibly at isolated points.

In either case, u is also an analytic function of $x_1, \ldots, x_n$.

The single difficulty, in the hyperbolic case, lies in the region of existence of u. Here u exists in regions cut out by "characteristic surfaces" through the points of S; here a characteristic surface, in ordinary space, through the point $x^* = (x_1^*, \ldots, x_n^*)$ is given by

$$Q(x_1 - x_1^*, x_2 - x_2^*, \ldots, x_n - x_n^*) = 0.$$

Thus the influence of the initial conditions spread like a wave front throughout space, under conditions of propagation controlled by the operator L.

Further, u, is determined at a point x** only by the portion of S, and data thereon, which is cut out by the characteristic surface through x**; the importance of this fact for numerical computation lies in this: Any method of computing u numerically which makes u at x** depend on substantially more of S and the initial data than is cut out by the characteristic surface through x** will be highly unstable, while a processing depending on substantially less of S will generally be highly in error.

This highly incomplete presentation will not be needed explicitly in what follows. For reference purposes a short bibliography on existence theorems is appended. For further reference, see D. Bernstein's book.

REFERENCES

D. L. Bernstein — "Existence Theorems in Partial Differential Equations". Princeton, 1950

F. Johns — "General Properties of Solutions of Linear, Elliptic Partial Differential Equations". The Proceedings of the Symposium on Spectral Theory and Differential Problems. Oklahoma, 1951.

"On Linear Partial Differential Equations with Analytic Data". Communications on Pure & Applied Mathematics, vol. 2 (1949), pp. 209-254.

"The Fundamental Solution of Linear Elliptic Differential Equations with Analytic Coefficients". ibid, vol. 3 (1950), pp. 273-304.

(An unpublished paper on Parabolic Equations).

R. Courant & D. Hilbert — "Methoden der Mathematischen Physic" vol. 2 . New York.

E. Kamke — "Differentialgleichungen Reeller Funktionen". New York, 1947

3. ON CONVERGENCE PROBLEMS.

The basic approximation procedure which is considered in this paper is the replacement of derivatives by partial difference quotients. For example, we could replace $\frac{\partial u}{\partial x_1}$ by

$$\frac{u(x_1+h_1,x_2,\ldots,x_n) - u(x_1,x_2,\ldots,x_n)}{h_1} ,$$

or by

$$\frac{u(x_1+h_1,x_2,\ldots,x_n) - u(x_1-h_1,x_2,\ldots,x_n}{2h_1} ;$$

let $\Delta_{i_1 i_2 \ldots i_n}$ be the operator which replaces $\frac{\partial^{i_1+\ldots+i_n}}{\partial x_1^{i_1} \partial x_2^{i_2} \ldots \partial x_n^{i_n}}$

and let L_Δ be the difference operator corresponding to L, i.e.,

$$L_\Delta [u] = \sum a_{i_1 i_2 \ldots i_n} \Delta_{i_1 \ldots i_n} (u)$$

This is defined on a mesh whose steps are $h_1, h_2, \ldots, h_n$. We must replace S by an approximation which passes through mesh points. Let then S_Δ be the approximation to S, together with the appropriate data for u and the difference expressions of the normal derivatives of u. This is the difference problem corresponding to the original differential problem. Let u_Δ be the solution of the difference equation, u_o the solution of the differential equation. The convergence problem is the problem of when

$$\lim_{h_1 \to 0, h_2 \to 0, \ldots} u_\Delta = u_o .$$

For elliptic problems the answer is that this always happens. For hyperbolic problems the condition is that certain ratios of the $h_1,h_2,\ldots,h_n$ must satisfy inequalities determined by L; without making these precise, let us only say that these inequalities coincide, in form and substance, with the conditions of stability which will be derived later.

As for a precise estimate of the truncation error $(=u_o-u_\Delta)$, no such exists; nor is there any rigorous test for convergence. In practice, one may solve the difference equations for several progressively smaller meshes; if the solutions coincide to many decimal places, one can assume that the truncation error occurs beyond those places. Obviously, this procedure has its dangers; in slowly converging cases it is bound to fail.

The fundamental paper, in which was first published the discovery of the need for inequalities on the mesh ratios, is:

R. Courant, K. Friederichs, and H. Lewy - "Uber die Partielle
Differenzengleichungen der Mathematischen Physik".
Mathematishce Annalen, vol. 100 (1928), pp. 32-74.

4. STABILITY: GENERALITIES.

Given the equations $L[u_o] = B$, $L_\Delta\left[u_\Delta\right] = B$, their analytic solution in explicit form may be entirely beyond our means; we are forced to turn to the numerical solution of the difference problem, generally, with the aid of machines.

In numerical work, numbers must be rounded off; this means that we derive a numerical approximation u_N to u_Δ . Even if u_Δ is a good approximation of u_o, u_N may be a bad one; for the errors introduced, by rounding off, into u_N may grow large, or may not grow small and thereby accumulate. We need, therefore, criteria that insure that the round off error will grow small, i.e., tend to zero (See sections quoted in Section 1).

In the case of elliptic problems, no criterion is needed; $u_\Delta - u_N$ remains small, just as $u_o - u_\Delta$ becomes small. Thus no further discussion of the elliptic case is necessary. With reference to this case; see:

J. D. Tamarkin & W. Feller - "Partial Differential Equations".

Brown University Lecture Notes, 1941, Chap. V, pp. 160-196.

The rest of this paper is devoted to a discussion of the stability criteria for the hyperbolic and parabolic cases. The basic references upon which this discussion rests are given on the following page. The papers in this list will be referred to by the bracketed numbers preceding them.

REFERENCES

[E 1] R.P. Eddy "Stability in the Numerical Solution of Initial Value Problems in Partial Differential Equations". NOLM 10232

[OHK 1] G. O'Brien, M. Hymen & S. Kaplan "A Study of the Numerical Solution of Partial Differential Equations". NOLM 10433. (Both the above are in the Journal of Mathematics and Physics, Vol. 29 (1951), pp. 223-251.

[T 1] L.H. Thomas "Stability of Solution of Partial Differential Equations". In the Symposium on Theoretical Compressible Flow, NOLR 1132, pp. 83-94.

[T 2] L.H. Thomas "Numerical Solution of Partial Differential Equations of Parabolic Type". Proceedings of a Seminar on Scientific Computation, Nov. 1949, pp. 71-78.

[NR 1] J. von Neumann & R.D. Richtmyer "On the Numerical Solution of Partial Differential Equations of Parabolic Type". LA-657.

[H 1] M. Hyman "On the Solution of Boundary-Value Problems as Initial Value Problems". Abstract 324, Bulleting of the American Mathematical Society, vol. 56 (1950), p. 346. (To appear under the title "On the Non-iterative Numerical Solution of Boundary-Value Problems".

[L 1] W. Leutert "On the Convergence of Approximate Solutions of the Heat Equation to the Exact Solution". Proceedings of the American Mathematical Society. Vol. 2 (1951), pp. 433-439.

[R 1] L.F. Richardson "The Approximate Arithmetical Solution by Finite Differences of Physical Problems Involving Differential Equations". Philosophical Transactions of the Royal Society of London, Series A, Vol. 210 (1919), pp. 307-357.

5. STABILITY: von NEUMANNS' CRITERIA.

The references for this and the next section are [OHK 1], [NR 1] and [L 1] .

The criteria developed here are sufficient for the stability of the difference-scheme solutions of partial differential equations with constant coefficients; the criteria can be applied to equations with variable coefficients by dividing the region of integration into sufficiently small subregions in which the coefficients vary slowly, and applying the criteria to each subregion. This procedure must be used with care, and will not apply if, in particular, the coefficients are discontinuous.

The paper of W. Leutert [L 1] shows, by example, that von Neumann's criteria are not necessary.

We will make some simplifications in our differential and difference problems; we assume first that the independent variables are t and $x_1, x_2, \ldots, x_n$ (this amounts merely to renaming the variables). The reason for this assumption is the next one, namely that that surface S on which the initial data are give is defined by

$$S: \ t = 0.$$

The particular difference operator L_Δ which replaces the differential operator L is assumed to be a polynomial in the operators Δ, Δ_1, $\Delta_2, \ldots, \Delta_n$ and Δ^{-1}, $\Delta_1{}^{-1}, \ldots, \Delta_n{}^{-1}$, where

$$\Delta u = u(t+k, x_1, \ldots, x_n) - u(t, x_1, \ldots, x_n)$$

$$\Delta_n^{-1} u = u(t, x_1, \ldots, x_n) - u(t-k, x_1, \ldots, x_n)$$

and

$$\Delta_i u = u(t,x_1,\dots,x_2 + h_i,\dots,x_n) - u(t,x_1,\dots,x_n)$$

$$\Delta_i^{-1} u = u(t,x_1,\dots,x_i,\dots,x_n) - u(t,x_1,\dots,x_i - h_i,\dots,x_n),$$

and where k, h_1, $h_2,\dots,h_n$ are positive quantities which determine the mesh structure.

Now (See paragraphs quoted on page 5) we will assume that at t = 0 an error occurs, say $e(x_1,x_2,\dots,x_n)$. This error affects the solution of

$$L_\Delta \; [u] = B$$

and is therefore propagated in the form of an error $f(t,x_1,\dots,x_n)$ whose law of propagation is given by

$$L_\Delta \; \Big[f(t;x_1,\dots,x_n)\Big] = 0$$

$$f(0;x_1,\dots,x_n) = e(x_1,\dots,x_n).$$

Since L_Δ is a linear homogeneous operator, it is appropriate to make a Fourier analysis of $f(t;x_1,\dots,x_n)$; the consequence of this analysis will be an inequality to be satisfied for all $\beta_1,\dots,\ \beta_n$ by a certain function of the new variables $\beta_1,\dots,\ \beta_n$ and of the $k,h_1,\dots,h_n$. In practise this will result in certain inequalities to be satisfied by certain ratios of the $k,h_1,\dots,h_n$.

Let us expand $e(x_1,\dots,x_n)$ (the initial error at t = 0) in a Fourier series

$$e(x_1,\dots,x_n) = \sum_{\beta_1\dots\beta_n} A_{\beta_1\dots\beta_n} \, e^{(\beta_1 x_1+\dots+\beta_n x_n)i\pi} \quad ,$$

or else in a Fourier integral

$$e(x_1,\dots,x_n) = \int\dots\int A(\beta_1,\dots,\beta_n)\, e^{(\beta_1 x_1+\dots+\beta_n x_n)i\pi} \, d\beta_1\dots d\beta_n,$$

according as the initial data is given over a finite or infinite portion of the surface S: $t = 0$.

The principal fact to notice is that $e(x_1,\dots,x_n)$ is given by linear superposition of functions of the form $e^{(\beta_1 x_1+\dots+\beta_n x_n)i\pi}$; an analysis of the propagation of error which is initially given by such a function will lead to an analysis of the propagation of $e(x_1,\dots,x_n)$, simply again by superposition. Therefore our problem becomes this: To find $f(t;x_1,\dots,x_n)$ where

$$L_\Delta \left[f(t;x_1,\dots,x_n)\right] = 0$$

$$f(0;x_1,\dots,x_n) = e^{i\ (\beta_1 x_1+\dots+\beta_n x_n)}$$

The solution of this problem will be built up linearly from a finite number of functions of the form

$$F = e^{\alpha t}\ e^{(\beta_1 x_1+\dots+\beta_n x_n)i\pi}$$

where $\alpha = \alpha(\beta_1,\dots,\beta_n)$ is a (generally) complex number, depending on $\beta_1,\dots,\beta_n$, and the whole expression F is itself a solution of

$$L_\Delta [F] = 0 \qquad F(0;x_1,\dots,x_n) = e^{i\pi\ (\Sigma\ \beta_i x_i)} .$$

Now the requirement that the error remain small can be expressed by the inequality

$$\left|e^{\alpha\Delta t}\right| = \left|e^{\alpha k}\right| \leqq 1;$$

the requirement that the error die out (See the discussion in Section 4, first paragraph) is that

$$\left|e^{\alpha k}\right| < 1.$$

Now let us consider the result of substituting $F = e^{\alpha t} e^{i\pi\ (\beta_1 x_1+\dots\ \beta_n x_n)}$

into $L_\Delta[F] = 0$. Since L_Δ is a linear homogeneous function, we have

$$L_\Delta[F] = e^{\alpha t} e^{i\pi(\beta_1 x_1 + \ldots + \beta_n x_n)} L_*(e^{\alpha k}; k, h_1, \ldots, h_{n_j} \ \beta_1, \ldots, \beta_n$$

where L_* is a function of $e^{\alpha k}$, of $h_1, \ldots, h_n$, and of $\beta_1, \ldots, \beta_n$. Solving $L_* = 0$ for $e^{\alpha k}$ gives

$$e^{\alpha k} = G(k, h_1, \ldots, h_n; \beta_1, \ldots, \beta_n)$$

and the inequality $|e^{\alpha k}| < 1$ results in the inequality

$$|G(k, h_1, \ldots, h_n; \beta_1, \ldots, \beta_n)| < 1$$

which must hold for all $\beta_1, \ldots, \beta_n$ (See [OHK 1] , footnote on page 227, for special circumstances under which the inequality $|G| < 1$ need not hold for all β ; the inequality must nevertheless hold if the difference scheme is to be true for arbitrarily small $k, h_1, \ldots, h_n$).

We have thus found a sufficient condition for stability. In the next section we will develop similar criteria of stability according to a method developed by R. P. Eddy [E 1] .

6. STABILITY: R. P. EDDY'S CRITERIA.

The reference for this section is [E 1] . As in Section 5, we assume that the variables are $t, x_1, \ldots, x_n$, and that the mesh steps are $k, h_1, \ldots, h_n$.

Here we use, instead of the difference operators Δ, $\Delta_1, \ldots, \Delta_n$, the translation operators E, $E_1, \ldots, E_n$ whose powers are defined by

$$(E^{\pm\nu})f = f(t \pm \nu k, x_1, \ldots, x_n)$$

$$(E_i^{\pm\nu})f = f(t, x_1, \ldots, x_{i-1} \pm \nu h_i, x_{i+1}, \ldots, x_n).$$

Thus $\Delta = E-1$, $\Delta^{-1} = 1 - E$ and $\Delta_i = E_i-1$, $\Delta_i^{-1} = 1 - E_i$.

If now we replace the partial derivatives in the linear differential operator L by partial difference quotients, we obtain an operator L_E which is a polynomial in E, E^{-1}, $\ldots, E_n, E_n^{-1}$. We write

$$L_E = L_E(E;\ E_1,\ldots,E_n;\ k,h_1,\ldots,h_n).$$

If now the solution u of

$$L_E u = B$$

is disturbed by an error $e(m_1,\ldots,m_n)$ at $t = 0$, $x_1 = x_1 + m_1\ h$, $x_2 = x_2 + m_2\ h_2,\ldots,x_n = x_{no}+m_h\ h_n$, where $m_1,\ldots,m_n$ are integers, then the error is propagated according to the equation

$$L_E u = 0.$$

Let the error at $t = s\ k$, $s_1 = x_{10} + m_1\ h_1,\ldots\ x_n = x_{no} + m_n h_n$ be denoted by

$$f(s;m_1,\ldots,m_n)$$

so that

$$L_E\ f = 0$$

$$f(0;m_1,m_2,\ldots,m_n) = e(m_1,\ldots,m_n).$$

The operators E, $E_1,\ldots,E_n$ operate on f according to

$$(E^{\pm\nu})f = f(n \pm \nu k;\ m_1,\ldots,m_n)$$

$$(E_i{}^{\pm\nu})f = f(n;\ m_1,\ldots,m_2 \pm \nu h_i;\ \ldots,m_n).$$

Now let us replace the operators $E_1,\ldots,E_n$ in $L_E = L_E(E,E_1,\ldots,E_n;$ $k,\ h_1,\ldots,h_n)$ by exponentials

$e^{-i\theta 1}$, $e^{-i\theta 2},\ldots,e^{-i\theta n}$, so that $E_j{}^{\pm\nu}$ is replaced by $e^{\pm i \nu \theta j}$, and E_j^{-1} by $e^{i \theta j}$, and let us consider the solution $\varphi(s;\ \theta_1, \ldots,\ \theta_n; k, h_1, \ldots, h_n) = \varphi(s)$ of the equation

$$L_E(E; e^{-i\theta 1}, \ldots, e^{-i\theta n};\ k, h_1, \ldots, h_n)\ \varphi(s) = 0.$$

Then the error $f(s; m_1, \ldots, m_n)$ is given by

$$f(s; m_1, \ldots, m_n) = \varphi(s; E_1^{-1}, E_2^{-1}, \ldots, E_n^{-1}; k, h_1, \ldots, h_n)\ e(m_1, \ldots, m_n).$$

Thus the nature of the error is entirely contained in the function $\varphi(s;\ \theta_1,\ \theta_2, \ldots,\ \theta_n; k, h_1, \ldots, h_n)$; in order to have stability we must satisfy the condition

$$\lim_{s \to \infty} \varphi(s) = 0.$$

Let us write $L_E \varphi = 0$ in the form

$$\varphi(s) = \beta_1 \varphi(s-1) + \beta_2 \varphi(s-2) + \ldots + \beta_p \varphi(s-p),$$

where p is an integer determined by L_E and β_i $(i = 1, \ldots, p)$ are functions of the $\theta_1, \ldots,\ \theta_n, k, h_1, \ldots, h_n$.

Now $\varphi(s)$ can be given explicitly. Let $\rho_1, \ldots,\ \rho_q$ be the roots of

$$\lambda^p - \beta_1\ \lambda^{p-1} - \ldots - \beta_{p-1}\ \lambda - \beta_p = 0$$

and let $a_1, \ldots, a_q$ be the multiplicities of these roots. Then $\varphi(s)$ can be given as a combination

$$\varphi(s) = \sum_{i=1}^{q} \left(\sum_{j=0}^{a_j - 1} A_{ji}\ s^j\ \rho_i^{\,s} \right)$$

where A_{ji} are functions of $\theta_1, \ldots,\ \theta_n, k, h_1, \ldots, h_n$. In order that $\lim_{s \to \infty} \varphi(s) = 0$, a necessary and sufficient condition is that

$$\lim_{s \to \infty} s^j \rho_i^s = 0 \quad i = 1,\ldots,q, \quad 0 \leqq j \leqq a_i - 1.$$

But this is satisfied if $|\rho_i| < 1$. Thus the condition for stability is that every root ρ of

$$\lambda^p - \beta_1 \lambda^{p-1} - \ldots - \lambda \beta_{p-1} - \beta_p = 0$$

satisfy

$$|\rho| < 1.$$

Further, this must hold for every value of the variable $\theta_1, \ldots, \theta_n$; the end result is therefore a set of conditions on k, $h_1, \ldots, h_n$.

According to Eddy [E 1; p. 3] , in all cases which have so far been tested the methods of this section and of the last section yield identical results.

For the cases p = 1 or 2, we can develop immediate criteria for the roots to satisfy $|\rho| < 1$.

When p = 1, the only root of $\lambda - \beta_1 = 0$ is $\rho_1 = \beta_1$. Therefore the condition of stability is

$$-1 < \beta_1 < 1.$$

When p = 2, the usual expression for the roots of a quadratic can be applied to get the desired conditions on β_1 and β_2; when β_1 and β_2 are real, these conditions are particularly simple. There are two cases to be considered: Real roots, and complex conjugate roots.

First case: Complex conjugate roots, i.e.,

$$\left(\frac{\beta_1}{2}\right)^2 + \beta_2 < 0.$$

Then the roots are $\rho_1 = \frac{\beta_1}{2} + \sqrt{\left(\frac{\beta_1}{2}\right)^2 + \beta_2}$ and $\rho_2 = \frac{\beta_1}{2} \sqrt{\left(\frac{\beta_1}{2}\right)^2 + \beta_2}$; the conditions of stability are $|\rho_1| = |\rho_2| < 1$. But

$|\rho_1| = \left(\frac{\beta_1}{2}\right)^2 - \left(\left(\frac{\beta_1}{2}\right)^2 + \beta_2\right)$ or $|\rho_1| = |\rho_2| = - \beta_2$. Thus we must have

$$- 1 < \beta_2 < 0.$$

which combined with $\left(\frac{\beta_1}{2}\right)^2 + \beta_2 < 0$ gives

$$|\beta_1| < 2 \sqrt{- \beta_2}$$

Second case: Real roots, i.e.,

$$\left(\frac{\beta_1}{2}\right)^2 + \beta_2 > 0.$$

Now the roots $\rho_1 = \beta_2 + \sqrt{\left(\frac{\beta_1}{2}\right)^2 + \beta_2}$ and $\rho_2 = \beta_1{}^2 - \left(\frac{\beta_1}{2}\right)^2 + \beta_2$

must satisfy $-1 < \rho_1 < 1$ $- 1 < \rho_2 < 1$. These conditions lead to

$$\beta_2 + \beta_1 < 1$$

$$\beta_2 - \beta_1 < 1$$

for

$$1 > \frac{\beta_1}{2} + \sqrt{\left(\frac{\beta_1}{2}\right)^2 + \beta_2} \text{ yields}$$

$$\left(1 - \frac{\beta_1}{2}\right)^2 > \left(\frac{\beta_1}{2}\right)^2 + \beta_2$$

or $1 > \beta_1 + \beta_2$,

and likewise $\frac{\beta_1}{2} - \sqrt{\left(\frac{\beta_1}{2}\right)^2 + \beta_2} > -1$

yields $$\left(1+\frac{\beta_1}{2}\right)^2 > \left(\frac{\beta_1}{2}\right)^2 + \beta_2$$

or $$1 > \beta_2 - \beta_1 \quad .$$

Thus for p = 1,2 the criteria can be written

I. $p = 1$ $\qquad - 1 < \beta_1 < 1$

II. $p = 2$, β_1, β_2 real;

for complex roots $\qquad -1 < \beta_2 < 0, \quad |\beta_1| \;\; 2\sqrt{-\beta_2}\,,$

for real roots $\qquad \beta_2 + \beta_1 < 1, \; \beta_2 - \beta_1 < 1.$

7. THE HEAT EQUATION AS AN EXAMPLE.

We will, in this section, discuss the heat equation

$$\frac{\partial u}{\partial t} = a \frac{\partial^2 u}{\partial x^2}$$

as an example and illustration of the methods of the preceding sections. The basic references are [E 1] , [OHK 1] , [T 1] and [NR 1] . In particular, the numerical examples, tables and graphs are copied directly from [OHK 1] ; they form the most enlightening feature of all these discussions of numerical methods.

We will consider several possible choices for the difference scheme to be used in the numerical solution of the heat equation, and we will repeat the derivation of the stability conditions for these special cases in order to illustrate the general proceedures of the preceding section.

We will consider first Richardson's difference scheme (See [R 1]).

$$\frac{u(x,t+\Delta t) - u(x,t-\Delta t)}{2\,\Delta t} = a\,\frac{u(x+\Delta x\ t) - 2\,u(x,t) + u(x-\Delta x\ t)}{\Delta x^2}$$

(in this simple case, we will use Δt, Δx rather than k, h_1).

This was used by L. F. Richardson to integrate the heat equation over a short range; as we shall see, this scheme is ordinarily unstable, both by the stability conditions and experimentally. Nevertheless, W. Leutert has shown [L 1] , that under appropriate conditions, Richardson's scheme can be stable without satisfying either the von Neumann or Eddy criteria.

The von Neumann criteria proceeds by a Fourier analysis of error; by linearity we need only consider solutions of Richardson's scheme of the form $e^{\alpha t} e^{i\beta x}$. The requirement that $\left| e^{\alpha \Delta t} \right| < 1$ gives conditions which Δt, Δx, must satisfy for all β. Introducing $e^{\alpha t}\, e^{i\beta x}$ in place of u(x,t) in Richardson;s scheme, we get

$$\frac{e^{\alpha t} e^{i\beta x}(e^{\alpha \Delta t} - e^{-\alpha \Delta t})}{2\,\Delta t} = a\, e^{\alpha t} e^{i\beta x} \frac{(e^{i\beta \Delta x} - 2 + e^{-i\beta \Delta x})}{\Delta x^2}$$

or setting $\xi = e^{\alpha \Delta t}$

$$\xi - \frac{1}{\xi} = \left(\frac{2\,a\,\Delta t}{\Delta x^2}\right) 2(\cos\,\beta\,\Delta x - 1).$$

$$= -\,8\left(\frac{a\,\Delta t}{\Delta x^2}\right) \sin^2\,\left(\frac{\beta\,\Delta x}{2}\right)$$

or setting $r = \dfrac{a\,\Delta t}{\Delta x^2}$

$$\xi - \frac{1}{\xi} = -8\ r\ \sin^2\,\left(\frac{\beta\,\Delta x}{2}\right).$$

Now the condition that $\left| e^{\alpha \Delta t} \right| < 1$ for every for which $e^{\alpha t} e^{i \beta x}$ is a solution of Richardson's scheme gives us that we must have

$$\left| \xi \right| < 1 \qquad \left| \frac{1}{\xi} \right| < 1.$$

But if $\beta \neq 0$, $\xi - \frac{1}{\xi} < 0$ by the equation for $\xi - \frac{1}{\xi}$; and this is possible only if either $\xi < -1$ or $- \frac{1}{\xi} < -1$. Thus Richardson's scheme is, by von Neumann's test, always unstable.

Now in order to apply Eddy's method, we use the operators E. E_x where

$$Eu(x,t) = u(x,t+\Delta t), \; E_x \, u(x,t) = u(x+ \Delta x,t)$$

Richardson's scheme is then

$$\frac{E - E^{-1}}{2 \Delta t} u = a \left(\frac{E_x - 2 + E_x^{-1}}{\Delta x^2} \right) u$$

or replacing $u(x,t+s \Delta t)$ by $\varphi(s)$ and E_x by $e^{-i\theta}$, we have

$$\varphi(s+1) - \varphi(s-1) = \left(\frac{2a \Delta t}{\Delta x^2} \right) (e^{-i\theta} -2 + e^{i\theta}) \, \varphi(s)$$

or translating to the left by 1, i.e., replacing s by (s-1) and setting $r = \frac{a \Delta t}{\Delta x^2}$

$$\varphi(s) - 4 \, r(\cos \theta -1) \, \varphi(s-1) - \varphi(s-2) = 0$$

The "characteristic equation"

$$\lambda^2 - 4 \, r(\cos \theta -1) \, \lambda - 1 = 0.$$

The conditions at the end of Section 6 for the case p = 2 apply here; since $-1 < 0$, the roots are real. Thus $\beta_2 = 1$ and $\beta_1 = 4r \, (\cos \theta -1)$ must satisfy

$$\beta_1 + \beta_2 < 1 \qquad \beta_2 - \beta_1 < 1$$

or

$$1 + 4t\ (\cos\theta - 1) < 1 \qquad 1 - 4r(\cos\theta - 1) < 1.$$

In particular, for $\cos\theta = -1$, these give

$$1 - 8r < 1 \qquad 1 + 8r < 1$$

or

$$8r > 0 \text{ and } \qquad 8r < 0$$

which is obviously impossible.

Thus by Eddy's criteria the scheme of Richardson is unstable. We will later compare Richardson's scheme with a stable scheme numerically, and see that it is in fact unstable; the results of W. Leutert, which show that Richardson's scheme can be made stable do not apply to the particular numerical methods used.

Now let us turn our attention to the difference scheme

$$\frac{u(x,t+\Delta t) - u(x,t)}{\Delta t} = a\,\frac{u(x+\Delta x,t) - 2u(x,t) + u(x-\Delta x,t)}{\Delta x^2}$$

This scheme has a larger truncation error than Richardson's scheme, but on the other hand, it is a stable scheme.

Using von Neumann's method of analysis, we substitute $e^{\alpha t}e^{i\beta x}$ for $u(x,t)$, obtaining

$$\frac{e^{\alpha t}e^{i\beta x}(e^{\alpha\Delta t}-1)}{\Delta t} = a\,e^{\alpha t}e^{i\beta x}\frac{(e^{i\beta\Delta x} - 2 + e^{-i\beta\Delta x})}{\Delta x^2}$$

or setting $\xi = e^{\alpha\Delta t}$, $r = \dfrac{a\Delta t}{\Delta x^2}$

$$\xi - 1 = 2r\ (\cos \beta \Delta x - 1)$$

$$= - 4r \sin^2 \frac{\beta \Delta x}{2}$$

or

$$\xi = 1 - 4r \sin^2 \frac{\beta \Delta x}{2}$$

The condition $|\xi| < 1$ gives

$$-1 < 1 - 4r \sin^2 \frac{\beta \Delta x}{2} < 1.$$

The right inequality is satisfied trivially. The left inequality gives

$$-1 < 1 - 4r \sin^2 \frac{\beta \Delta x}{2}$$

or

$$r \sin^2 \frac{\beta \Delta x}{2} < \frac{1}{2} \cdot$$

since this inequality must hold for all β , we find

$$r < \frac{1}{2}$$

as the condition for stability.

Turning to Eddy's criterion, and using E, E_x as the translation operators, we have as the difference scheme

$$\frac{(E-1)u}{\Delta t} = a \frac{(E_x - 2 + E_x - 1)u}{\Delta x^2}$$

and the Eddy equation becomes

$$\varphi(s+1) - \varphi(s) = \frac{a \Delta t}{\Delta x^2} \left(e^{-i\theta} - 2 + e^{i\theta}\right) \varphi(s)$$

or replacing s by s-1, and $\frac{a \Delta t}{\Delta x^2}$ by r

$$\varphi(s) - (1 + 2r\ \overline{\cos \theta - 1})\ \varphi(s-1) = 0.$$

By the conditions at the end of Section 6, the stability condition for p = 1 is

$$-1 < 1 + 2r\ (\cos\theta - 1) < 1$$

or

$$-2 < 2r\ (\cos\theta - 1) < 0$$

or

$$0 < r\ (1 - \cos\theta) < 1.$$

The left inequality always holds, while the right inequality requires that we have

$$r < \frac{1}{2}$$

which is thus the condition for stability; this conclusion agrees with that drawn from von Neumann's method.

Before we turn to a numerical consideration of the above two schemes, let us consider a strong objection to them; in order to have stability, we must have

$$r < \frac{1}{2}$$

i.e.,

$$\Delta t < \frac{a}{2}\ \Delta x^2.$$

Thus if Δx is very small, Δt must be very, very small; in consequence, a certain degree of accuracy in x calls for much more accuracy in t. Methods have been developed by von Naumann (the "implicit methods") which are stable for all values of r, i.e., all mesh ratios; see [NR 1] ,[E 1], [OHK 1] . For numerical work, such schemes are obviously preferable; we concentrate here on an unstable scheme and one which is stable for $r < \frac{1}{2}$ because they show the various phenomena which may be expected in a much clearer manner than an absolutely stable scheme can.

We will now consider the result of numerical integration applied to a special case of the heat equation. The differential problem we will now consider will be

$$\frac{\partial u(x,t)}{\partial t} = \frac{\partial^2 u(x,t)}{\partial x}$$

$$u(x,0) = 1 \qquad 0<x<1$$

$$u(0,t) = u(1,t) = 0 \qquad t \geq 0.$$

The exact solution of this differential problem is given by

$$u(x,t) = \frac{4}{\pi} \sum_{m=1,3,5,\ldots}^{\infty} \frac{1}{m} e^{-m^2 \pi^2 t} \sin(m\pi x).$$

We will calculate both the exact and numerical solutions of the difference scheme (A) and (B)

$$(A)\ldots \frac{u(x,t+\Delta t) - u(x,t)}{\Delta t} = \frac{u(x+\Delta x,t) - 2\,u(x,t) + u(x-\Delta x,t)}{\Delta x^2}$$

$$(B)\ldots \frac{u(x,t+\Delta t) - u(x,t-\Delta t)}{2\,\Delta t} = \frac{u(x+\Delta x,t) - 2u(x,t) + u(x-\Delta x,t)}{\Delta x^2}$$

the latter scheme being, of course, Richardson's.

We will refer to the exact solution as D, the exact and numerical solutions of (A) as Δ and N, and the exact and numerical solutions of (B) as Δ_R and N_R, both in this discussion and in the tables and graphs.

The tables and graphs are taken directly from [OHK 1; pp. 240-251]; they are given at the end of this section.

Plate I compares D and N_R, using t = 0.001 (or 1 millisecond) and

$\Delta x = 0.1$. The data from which Plate I was drawn is given in Table II, columns two and three. It shows that for $t > 0.005$, N_R begins to diverge from D and shortly goes into strong oscillations.

The remaining plates compare D, Δ and N (the solutions of scheme (A)) for various values of Δx and various values of the mesh ratio

$$r = \frac{\Delta t}{\Delta x^2} ,$$

Scheme (A) being stable for $r < \frac{1}{2}$. The data from which the remaining plates are drawn is given in Table II, columns other than number three, and in Tables III, IV, and V.

We can make the following observations on the graphs. For $r < \frac{1}{2}$, all of the solutions are quite close tb the true solution D, $r = 0.45$ being very nearly as good as $r = 0.1$, in spite of the fact that this causes Δt to be 4.5 times larger in the first case as in the second. The fact that the curve for $r = 0.45$ is better over a portion of the graph than $r = 0.1$, and better over the entire graph than $r = 0.3$ is without significance since the total time interval is only one tenth of a second. We get decidedly poorer results using $r = 0.5$.

In consequence, in numerical work it behooves us to use a mesh ratio r close to but less than 1/2.

In Plate 3, we compare the exact solution D with N for $r = 0.45$, $r = 0.5$, $r = 0.55$, and $r = 0.7$. The rate at which the situation degenerates is tremendous; for while N for $r = 0.45$ is a fairly good approximation to D, N for $r = 0.5$ is a poor approximation, and N for $r = 0.55$ is no approximation at all. The further increases to $r = 0.7$ causes unreasonably large oscillations.

In Plates 4 and 5 we compare D and N for stable mesh ratios and various values of Δx; as Δx becomes smaller the solutions N become better approximations to D in a regular manner, i.e., the solutions N converge to D (thus bearing out experimentally a statement made on the equivalence of convergence and stability criteria made in Section 3).

If we compare the numerical solution N for $r = 0.1$ $\Delta x = 0.1$ (Table II, column 4) and the exact difference solution Δ for $r = 0.1$, $\Delta x = 0.1$, (Table V, column 2) we see a remarkable agreement; which indicates that round-off errors are damped out too rapidly to accumulate; this substantiates experimentally the assumption made in Section 1 (in the quoted paragraphs) that weak stability implies strong stability.

Finally, if we compare the numerical and exact solutions N and Δ of the difference scheme (A), we find that they are nearly equal, even in the unstable case, as the following brief tabulation (taken from Tables II, III, V) shows at a glance:

r = 0.5			r = 0.55		
t(ms)	Δ	N	t(ms)	Δ	N
20	0.9375	0.9375	22	0.9083	0.9085
30	0.8594	0.8594	38.5	0.8824	0.8824
45	0.78126	0.78125	55	0.5756	0.5753
60	0.6409	0.6409	71.5	0.7609	0.7609
85	0.5245	0.5245	88	0.2287	0.2287
100	0.4292	0.4292	104.5	0.8218	0.8218

We can therefore conclude that the error in a numerical integration is due, not to the round-off error, but in fact, is due principally to the truncation error; which conclusion is contrary to the opinion generally held at the present time.

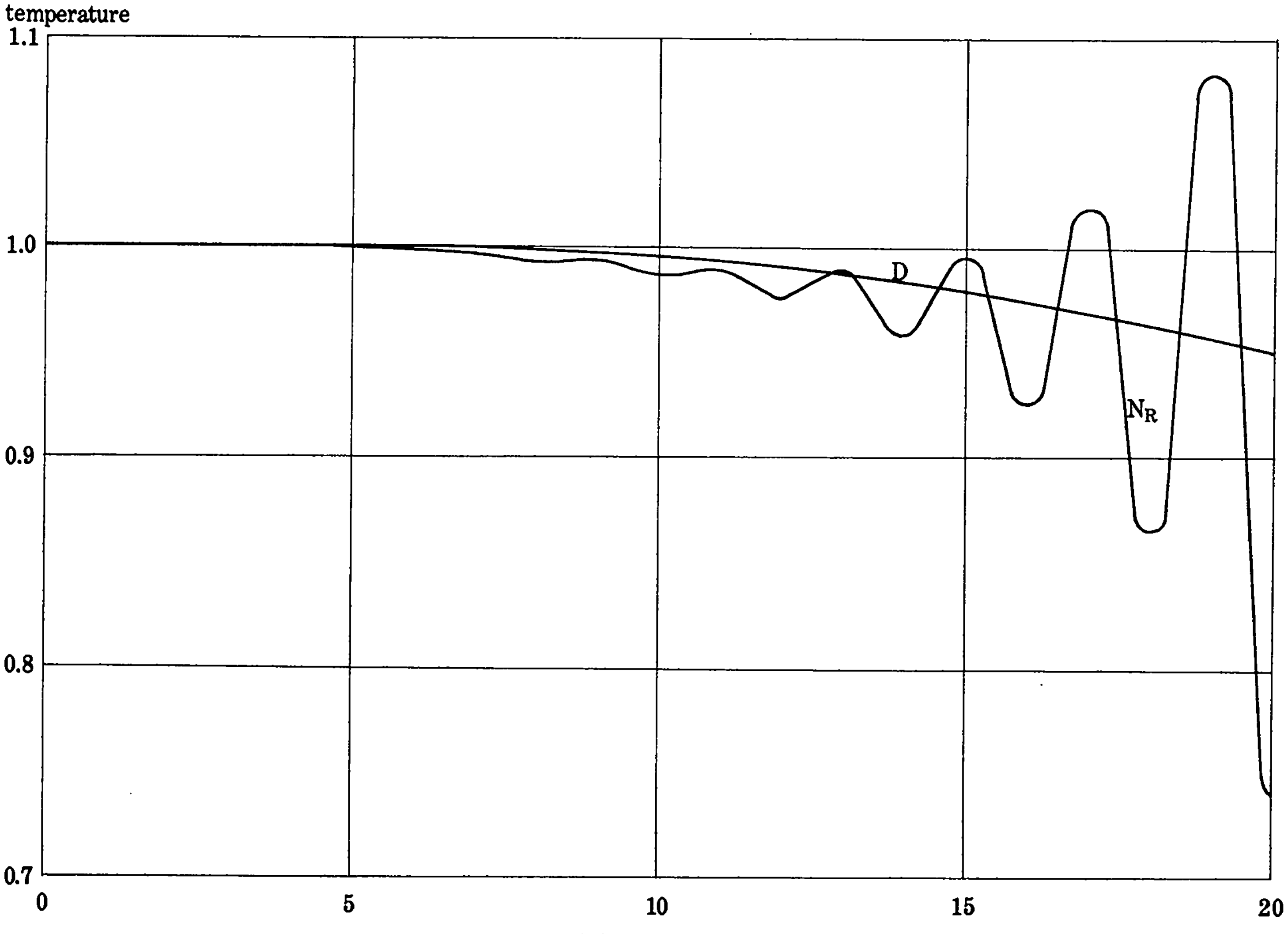

Plate 1. N_R COMPARED WITH D.

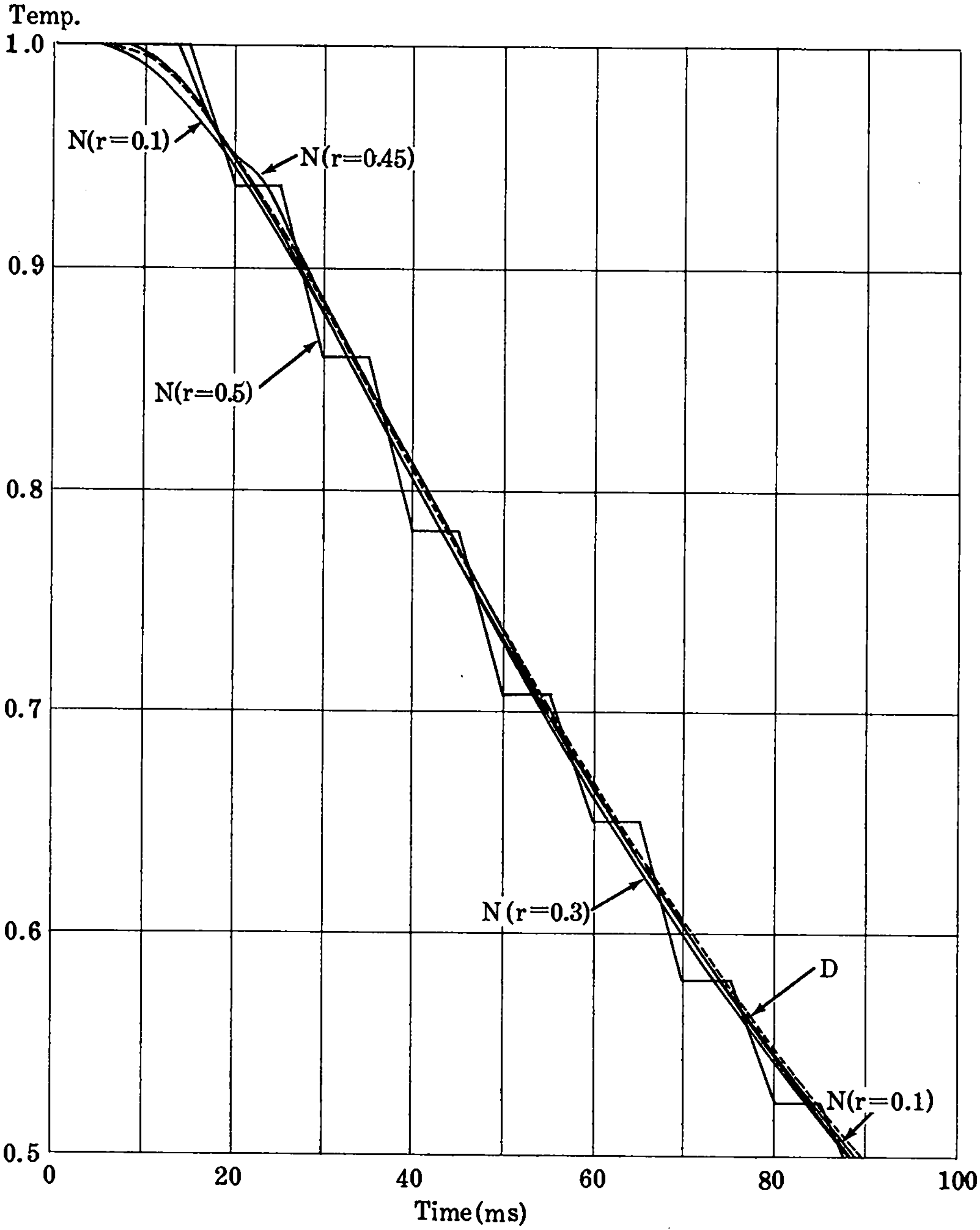

Plate 2. Stable Numerical Solutions. ($\Delta x=0.1$)

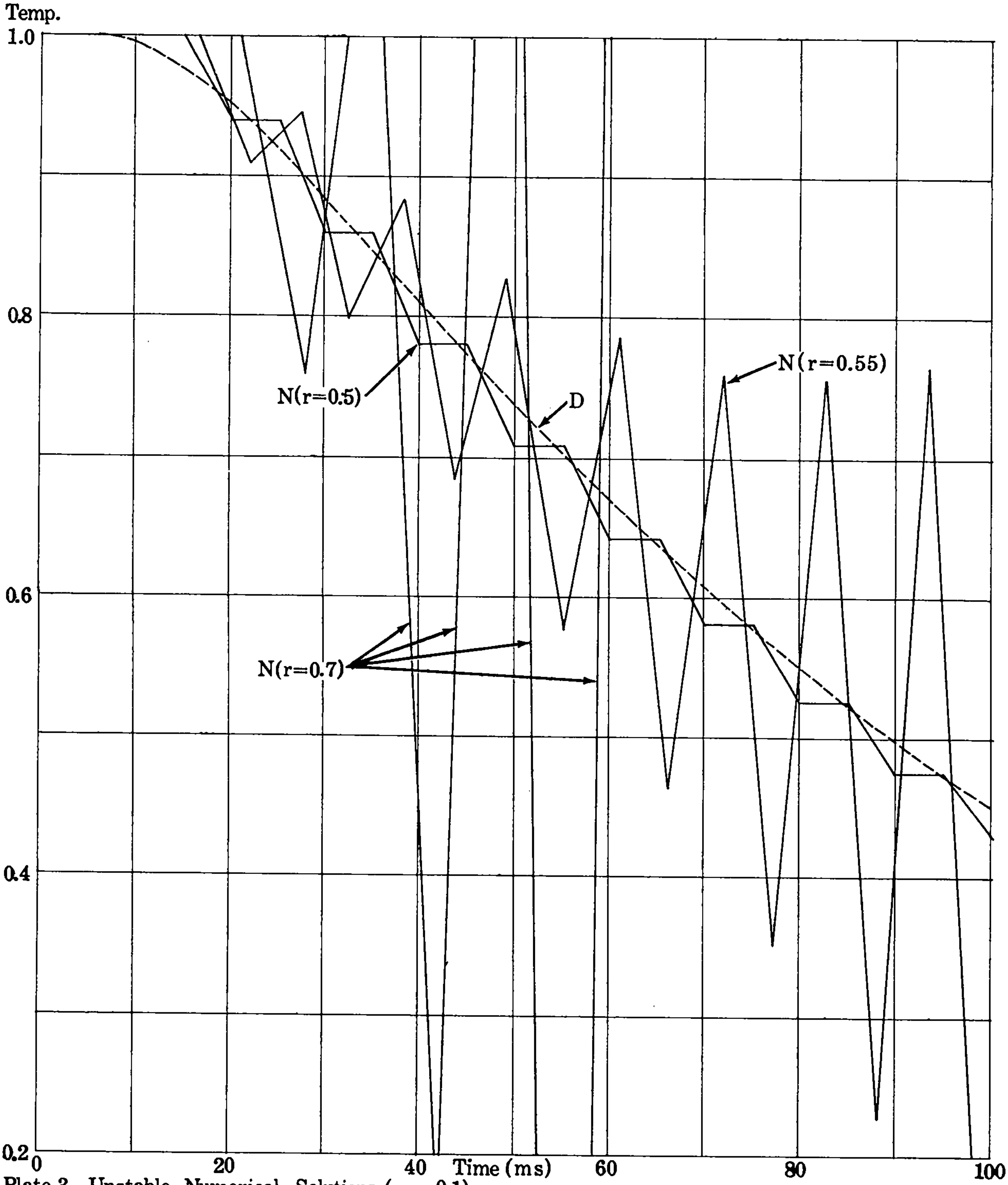

Plate 3. Unstable Numerical Solutions. ($\Delta x = 0.1$)

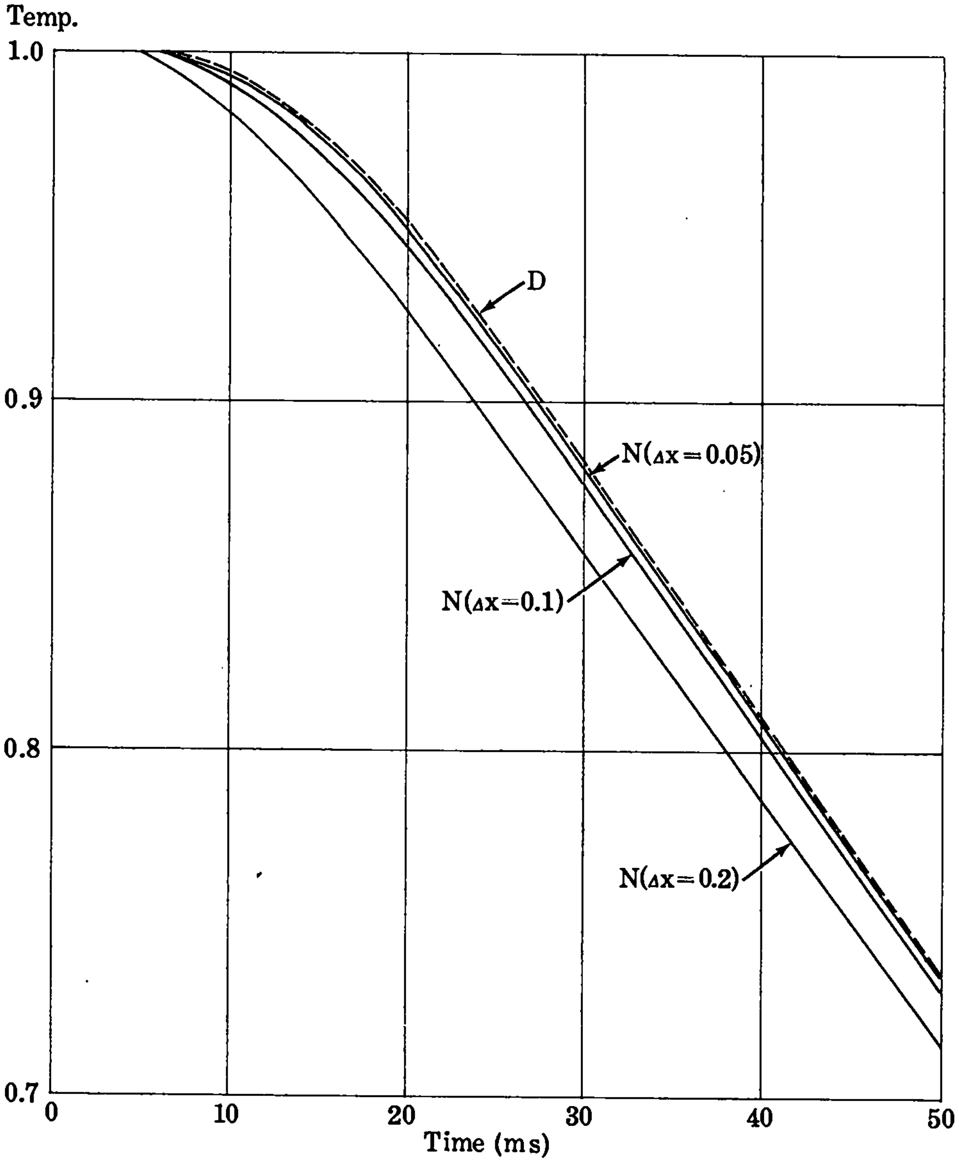

Plate 4. Converging Numerical Solutions. (r=0.1)

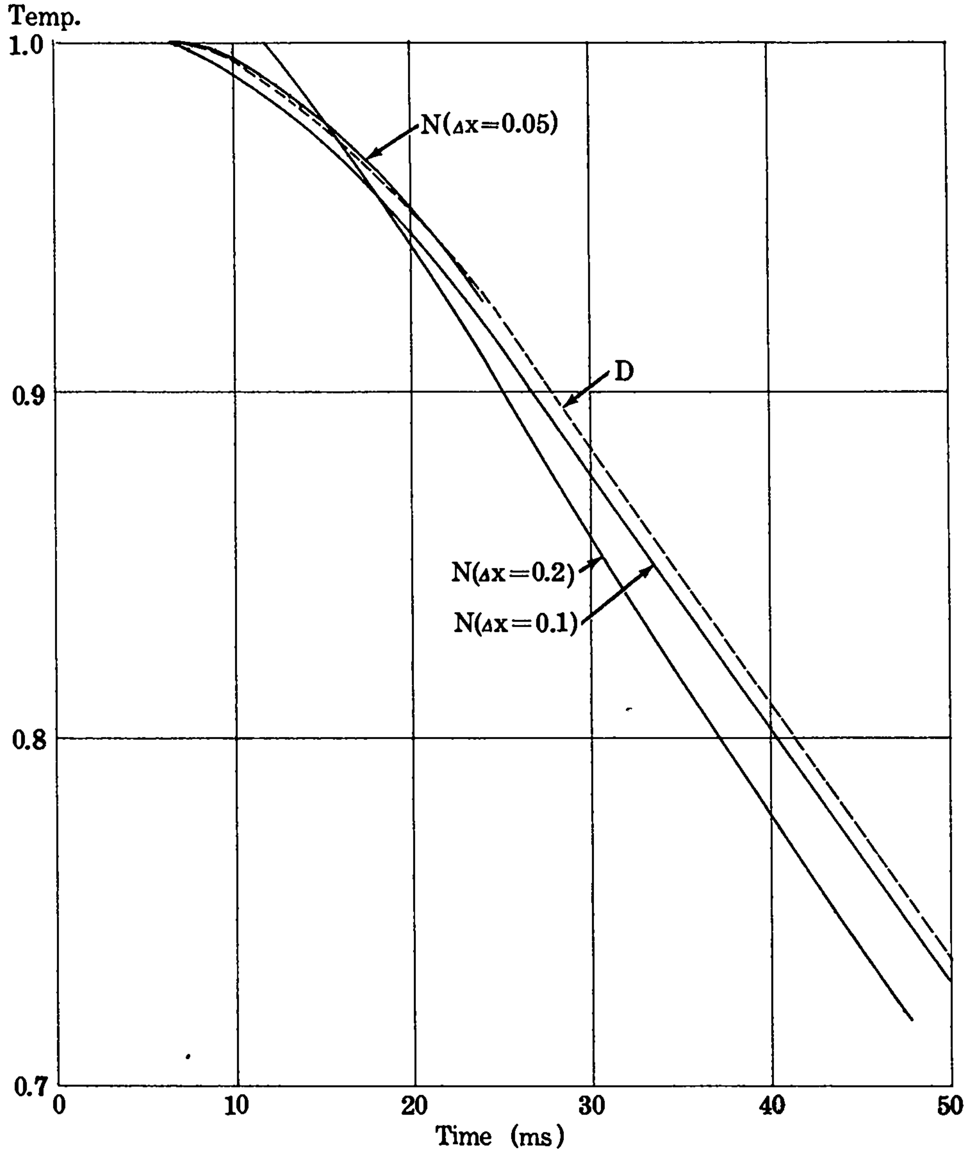

Plate 5. Converging Numerical Solutions. (r=0.3)

t(ms)	D	Δ_R	N_R	$\Delta = N$	$D-L_R$	$\Delta_R - N_R$	$D-\Delta$
5	1.0000	.9990	.9990	.9996	.0010	.0000	.0004
10	.9953	.9872	.9870	.9919	.0081	.0002	.0034
15	.9785	.9941	.9957	.9729	-.0156	-.0016	.0056
20	.9518	.7491	.7390	.9452	.2027	.0101	.0066
25	.9192	2.3771	2.5504	.9123	-1.4579	-.1733	.0069
30	.8832	-9.7547	-10.8768	.8766	10.6379	1.1221	.0066

TABLE I

Here we use $\Delta t = 0.001$, $\Delta x = 0.1$, which have the stable mesh ratio

$$r = \frac{\Delta t}{\Delta x^2} = 0.1.$$

In this and succeding tables (as in the graphs) we use the notation:

D = exact solution of $\dfrac{\partial u}{\partial t} = \dfrac{\partial^2 u}{\partial x^2}$

Δ = exact solution of

$$\frac{u(x,t+\Delta t) - u(x,t)}{\Delta t} = \frac{u(x+\Delta x,t) - 2u(x,t) + u(x-\Delta x,t)}{\Delta x^2}$$

N = numerical solution of the preceding difference system.

Δ_R = exact solution of Richardson's scheme

$$\frac{u(x,t+\Delta t) - u(x,t-\Delta t)}{2\Delta t} = \frac{u(x+\Delta x,t) - 2u(x,t) + u(x\ \Delta x,t)}{\Delta x^2}$$

N_R = Numerical solution of Richardson's scheme.

Throughout all tables and graphs, $x = 0.4$.

TABLE II (Δ x=0.1, r = $\frac{\Delta t}{\Delta x^2}$)

t(ms)	D	N_R	N(r=0.1)	N(r=0.3)	N(r=0.5)	N(r=0.7)
0	1.00000	1.0000	1.0000	1.0000	1.0000	1.0000
1		1.0000	1.0000			
2		1.0000	1.0000			
3		0.9999	1.0000	1.0000		
4		0.9994	0.9999			
5	1.0000	0.9990	0.9996		1.0000	
6		0.9975	0.9989	1.0000		
7		0.9970	0.9979			1.0000
8		0.9936	0.9964			
9		0.9940	0.9944	1.0000		
10	0.9953	0.9870	0.9919		1.0000	
11		0.9909	0.9890			
12		0.9763	0.9856	0.9919		
13		0.9898	0.9818			
14		0.9585	0.9775			1.0000
15	0.9785	0.9957	0.9729	0.9789	1.0000	
16		0.9267	0.9679			
17		1.0193	0.9627			
18		0.8652	0.9571	0.9623		
19		1.0845	0.9513			
20	0.9518	0.7390	0.9452		0.9375	
21		1.2439	0.9389	0.9432		1.0000
25	0.9192	2.5504	0.9123		0.9375	
27		4.5251	0.8982	0.9005		
28		-4.4469	0.8911			0.7599
30	0.8832	-10.8768	0.8766	0.8779	0.8594	
35	0.8461		0.8398		0.8594	1.1441
36			0.8324	0.8322		
40	0.8088		0.8029		0.7812	
42			0.7883	0.7869		0.1717

TABLE II (Cont'd.)

t(ms)	D	N_R	N(r=0.1)	N(r=0.3)	N(r=0.5)	N(r=0.7)
45	0.7721		0.7666	0.7647	0.7812	
48			0.7453	0.7430		
49			0.7382			1.8908
50	0.7363		0.7312		0.7080	
54			0.7038	0.7009		
55	0.7018		0.6971		0.7080	
56			0.6904			-1.4538
60	0.6686		0.6642	0.6608	0.6409	
63			0.6452	0.6416		3.4096
65	0.6368		0.6328		0.6409	
66			0.6266	0.6229		
70	0.6063		0.6027		0.5798	
72			0.5910	0.5870		
75	0.5773		0.5739	0.5698	0.5798	
78			0.5573	0.5531		
80	0.5496		0.5465		0.5245	
84			0.5254	0.5212		
85	0.5232		0.5203		0.5245	
90	0.4981		0.4954	0.4911	0.4745	
95	0.4741		0.4716		0.4745	
96			0.4670	0.4627		
100	0.4513		0.4490		0.4292	
102				0.4359		

TABLE III ($\Delta x = 0.1, \quad r = \frac{\Delta t}{\Delta x^2}$)

t(ms)	N(r=0.45	N(r=0.55)
0	1.0000	1.0000
4.5	1.0000	
5.5		1.0000
9.0	1.0000	
11.0		1.0000
13.5	1.0000	
16.5		1.0000
18.0	0.9590	
22.0		0.9085
22.5	0.9426	
27.0	0.9047	
27.5		0.9451
31.5	0.8743	
33.0		0.7976
36.0	0.8373	
38.5		0.8824
40.5	0.8043	
44.0		0.6856
45.0	0.7692	
49.5	0.7371	0.8273
54.0	0.7046	
55.0		0.5753
58.5	0.6744	
60.5		0.7857
63.0	0.6446	
66.0		0.4649
67.5	0.6166	
71.5		0.7609
72.0	0.5893	
76.5	0.5636	

TABLE III (Cont'd)

t(ms)	N(r=0.45)	N(r=0.55)
77.0		0.3509
81.0	0.5386	
82.5		0.7560
85.5	0.5151	
88.0		0.2287
90.0	0.4923	
93.5		0.7746
94.5	0.4707	
99.0	0.4499	0.0928
103.5	0.4301	
104.5		0.8218
110.0		-0.0639
115.0		0.9042
121.0		-0.2502

TABLE IV $(r = \frac{\Delta t}{\Delta x^2})$

	r = 0.1		r = 0.3	
t(ms)	N(Δ x=0.05)	N(Δ x=0.2)	N(Δ x=0.05)	N(Δ x=0.2)
0	1.000000	1.000000	1.000000	1.000000
1.5	1.000000			
3.0	0.999997			
4.0	0.999970	1.000000		
4.5	0.999928			
6.0	0.999540		0.999934	
7.5	0.998469		0.999299	
8.0	0.997905	0.990000		
9.0	0.996420		0.997613	
10.5	0.993243		0.994666	
12.0	0.988919	0.973000	0.990450	1.000000
13.5	0.983508		0.985059	
15.0	0.977110		0.978621	
16.0	0.972354	0.951200		
16.5	0.969842		0.971275	
18.0	0.961820		0.963153	
19.5	0.953154		0.954373	
20.0	0.950140	0.926210		
21.0	0.943944		0.945041	
22.5	0.934278		0.935251	
24.0	0.924235	0.899205	0.925084	0.910000
28.0	0.896113	0.871039		
32.0	0.866812	0.842331		
36.0	0.837052	0.813525		0.811000
40.0	0.807323	0.784938		
44.0	0.777957	0.756792		
48.0	0.749174	0.729240		0.719200
50.0	0.735048			
52.0		0.702386		

TABLE V (Cont'd)

t(ms)	Δ(r=0.1)	Δ(r=0.5)	Δ(r=0.55)
71.5			0.7609
75.0		0.5798	
77.0			0.3509
80.0	0.5464	0.5245	
82.5			0.7560
85.0		0.5245	
88.0			0.2287
90.0	0.5954	0.4745	
93.5			0.7746
95.0		0.4745	
99.0			0.0928
100.0	0.4490	0.4292	
104.5			0.8218
105.0		0.4292	
110.0			-0.0640

TABLE V ($\Delta x = 0.1$, $r = \frac{\Delta t}{\Delta x^2}$)

t(ms)	Δ(r=0.1)	Δ(r=0.5)	Δ(r=0.55)
10.0	0.9919	1.0000	
11.0			1.0000
15.0		1.0000	
16.5			1.0000
20.0	0.9452	0.9375	
22.0			0.9083
25.0		0.9375	
27.5			0.9451
30.0	0.8766	0.8594	
33.0			0.7975
35.0		0.8594	
38.5			0.8824
40.0	0.8029	0.7812	
44.0			0.6856
45.0		0.7812	
49.5			0.8273
50.0	0.7312	0.7080	
55.0		0.7080	0.5756
60.0	0.6642	0.6409	
60.5			0.7857
65.0		0.6409	
66.0			0.4650
70.0	0.6026	0.5798	

8. THE WAVE EQUATION.

We will consider briefly the application of von Neumann's and Eddy's methods to the wave equation

$$\frac{\partial^2 u}{\partial t^2} = a^2 \frac{\partial^2 u}{\partial x^2}$$

as an example of the hyperbolic case.

We will consider only the difference scheme

$$\frac{u(x,t+\Delta t) - 2u(x,t) + u(x,t-\Delta t)}{\Delta t^2} = a^2 \frac{u(x+\Delta x,t)- 2u(x,t) +u(x-\Delta x,t)}{\Delta x^2}$$

The results will be expressed in terms of the mesh ratio

$$r = \frac{a\Delta t}{\Delta x}.$$

To apply von Neumann's method, we substitute $e^{\alpha t} e^{i\beta x}$ for $u(x,t)$ and obtain, upon cancelling common factors and writing $\xi = e^{\alpha \Delta t}$

$$\xi + \frac{1}{\xi} = 2 + 2 r^2 (\cos \beta \Delta x - 1)$$

$$= 2 - 4 r^2 \sin^2 (\frac{\beta \Delta x}{2})$$

or setting

$$A = 1 - 2r^2 \sin^2 (\frac{\beta \Delta x}{2})$$

we have

$$\xi^2 - 2 A \xi + 1 = 0.$$

The roots of this quadratic are

$$\xi_1 = A + \sqrt{A^2 - 1} \quad \text{and} \quad \xi_2 = A - \sqrt{A^2 - 1}.$$

From the equation

$$\xi + \frac{1}{\xi} = 2 A$$

we see that $\xi_1 = 1/\xi_2$. If $A > 1$, $|\xi_1| > 1$; if $A < -1$, then $|\xi_2| > 1$; if $|A| \leqq 1$, then $|\xi_1| = |\xi_2| = 1$. Thus we can not choose r so that $|\xi| < 1$, i.e., we can never insure that the error tends to zero, but only that if it is initially small, it will remain small (See Section 4); this is known as semi-stability. The condition for semistability is then

$$-1 \leqq 1 - 2 r^2 \sin^2 \left(\frac{\beta \Delta x}{2}\right) \leqq 1$$

The right side is trivial; the left is satisfied if and only if

$$r \leqq 1.$$

Now we turn to Eddy's criteria. The difference scheme becomes

$$\frac{E - 2 + E^{-1}}{\Delta t^2} u = a^2 \frac{E_x - 2 + E_x^{-1}}{\Delta x^2} u$$

Replacing u by $\varphi(s)$, E_x by $e^{-i\theta}$ we have

$$\varphi(s+1) - 2\varphi(s) + \varphi(s-1) = 2 r^2 (\cos\theta - 1) \varphi(s)$$

or replacing s by s-1 and $(\cos\theta - 1)$ by $-2 \sin^2 \frac{\theta}{2}$,

$$\varphi(s) - 2(1 - 2r^2 \sin^2\theta) \varphi(s-1) + \varphi(s-2) = 0.$$

The characteristic equation is then

$$\lambda^2 - 2 A \lambda + 1 = 0$$

where

$$A = 1 - 2 r^2 \sin^2 \theta .$$

This then leads to the same results as in von Neumann's method, i.e., the roots ρ_1, ρ_2 of the equation satisfy

$$|\rho_1| = |\rho_2| = 1 \qquad \text{if } r \leqq 1$$
$$|\rho_1| > 1 \text{ or } |\rho_2| > 1 \qquad \text{if } r > 1.$$

Thus we can never ensure that the error will die out, but only that the error will not grow.

9. ESTIMATES FOR THE ERROR: ORDINARY DIFFERENTIAL EQUATIONS.

Methods developed by H. Rademacher permit the evaluation of both the truncation and round-off errors in the step-by-step numerical integration of systems of first order differential equations; these methods depend on the evaluation of two constants by an initial rough numerical integration, and therefore are of quantitative importance only in the more complicated cases which occur. The qualitative information supplied by these methods is of general utility. These methods have been extended to partial differential equations by L. H. Thomas.

In what follows we will consider a single ordinary equation,

$$y' = \frac{dy}{dx} = f(x,y)$$

with the initial condition

$$y(x_o) = y_o.$$

All our considerations will apply to systems of equations; a direct translation can be made if y and f are replaced by vectors, $\frac{\partial f}{\partial y}$ by the Jacobian matrix, and products by matricial products (the Y_j, λ, R, Φ and u which occur later must also be vectors).

We assume that we have a procedure of numerical integration, i.e., a function $\bar{y}(x,h)$ of $x,y(x),y'(x),h$, etc., such that

$$y(x+h) - \bar{y}(x,h) = h^{\beta+1}\Phi(x) + h^{\beta+2}\Psi(x,h)$$

$$\sim h^{\beta+1}\Phi(x)$$

where $\Phi \neq 0$ is an expression in $x,y(x)$ and its derivatives, and $f(x,y)$ and its derivatives which determines the coefficient of the approximate error; $\Psi(x,h)$ is some bounded coefficient, so that $h^{\beta+2}\Psi$ can be ignored in comparison with $h^{\beta+1}\Phi$ when h is small enough.

We will use this process of integration to compute $y(x_j)$ at $x_j = x_o + j\,h$, and we will endeavor to evaluate the approximate error involved in this numerical computation. Initially, we will assume that all numerical computation is carried to an infinite number of decimal places; the error resulting from rounding-off will be considered later.

We assume that the result of the numerical integration is a set of values y_j - an approximation to $y(x_j)$. The "truncation error" is denoted by

$$u_j = y(x_j) - y_j.$$

We introduce functions $Y_j(x)$ such that

$$Y_j(x_{j-1}) = y_{j-1} \qquad \frac{dY_j}{d\,x} = f(x,Y_j) \quad .$$

Then the extent by which $Y_j(x_j)$ and y_j differ is a measure of the growth of the truncation error in the interval from x_{j-1} to $x_j = x_{j-1} + h$. We set

$$R_j = Y_j(x_j) - y_j.$$

Then, approximately, $R_j \sim h^{\beta+1}\,\Phi(x_j)$.

Now let us consider the rate of change of (Y_j-y), i.e.,

$$\frac{d}{dx}(Y_j-y) = f(x,Y_j) - f(x,y)$$

$$\sim f_y(x,y)\,(Y_j-y)$$

neglecting higher powers of (Y-y). We form the "adjoint equation"

$$\frac{d}{dx} \lambda = - f_y(x,y(x))\lambda \;;$$

we will not yet specify initial values for λ .

Now $\frac{\lambda d}{dx} (Y_j-y) = \lambda f_y(x,y) (Y_j-y)$

$$= \left(\frac{-d}{d\ x} \lambda\right) (Y_j-y)$$

or

$$\frac{d}{dx}\left\{\lambda (Y_j-y)\right\} = 0.$$

If we integrate from x_{j-1} to x_j we obtain

$$\lambda (Y_j-y) \Big|_{x_{j-1}}^{x_j} = 0$$

$$= \lambda (x_j)\Big(Y_j(x_j) - y(x_j)\Big) - \lambda(x_{j-1})\Big(Y_j(x_{j-1}) - y(x_{j-1})\Big).$$

Then $Y_j(x_j) = y_j + R_j$, and $y(x_j) - y_j = u_j$.

Setting $\lambda(x_j) = \lambda_j$ we have

$$\lambda_j R_j = \lambda_j u_j - \lambda_{j-1} u_{j-1}.$$

Since $y(x_o) = y_o$, $u_o = 0$. Summing we have then

$$\sum_{j=1}^{n} \lambda_j R_j = \lambda_n u_n.$$

If we replace R_j by $h^{\beta+1} \Phi(x_j)$ we have

$$\lambda_n u_n \sim h^{\beta} \epsilon (\lambda_j \Phi(x_j)h$$

$$\sim h \int_{x_o}^{x_n} \lambda (x) \Phi(x) \, dx$$

or replacing x_n by X, u_n by u(X) and λ_n by λ(X)

$$\lambda(X)\; u(X) \sim h^{\beta} \int_{x_0}^{X} \lambda(x)\;\; \Phi(x)\; dx.$$

If we choose λ to satisfy the initial condition $\lambda(X) = 1$, we have an expression for the truncation error.

Now let us assume that in our computation we round-off in the (k+1)-st place, i.e., that we carry k places. We denote the rounded-off quantities by a dash over the symbols for exact quantities; thus, the $\bar{y}_j$ are the results of the numerical integration under rounding-off. We make two further simplifying assumptions: That we compute $f(\bar{x}_j,\bar{y}_j)$ so accurately that all doubtful figures in $hf(\bar{x}_j,\bar{y}_j)$ are shifted beyond the round-off point, i.e.,

$$\overline{hf(\bar{x}_j,\bar{y}_j)} = \overline{hf(\overline{\bar{x}_j,\bar{y}_j})}$$

and also we assume that $\overline{x_j} = x_j$.

Because of the magnitude of the round-off error (this is a "forward prediction") we will assume that h is so small that

$$\bar{y}_j = \bar{y}_{j-1} + h\; f(x_{j-1},\bar{y}_{j-1}) + \;\epsilon_j\; 10^{-k}$$

WHERE $0 \leqq \epsilon_j \leqq +1$ (this is the crudest representation of $\bar{y}_j$ in terms of $\bar{y}_{j-1}$, h that we can use).

The round-off error is given by

$$\bar{u}_j = y_j - \bar{y}_j\; .$$

Then

$$\bar{u}_j = \bar{u}_{j-1} + h(f(x_{j-1},y_{j-1}) - f(x_{j-1},\bar{y}_{j-1}) - \;\epsilon_j\; 10^{-k}$$

or

$$\bar{u}_j - \bar{u}_{j-1} = h\, \frac{\partial f}{\partial y}\; \bar{u}_{j-1} - \epsilon_j\; 10^{-k}$$

upon ignoring higher powers of $\bar{u}_{j-1}$. We introduce the solution of the difference equation

$$\bar{\lambda}_j - \bar{\lambda}_{j-1} = - h \frac{\partial f}{\partial y} \bar{\lambda}_j .$$

Then approximately

$$\lambda_j \sim \lambda(x_j).$$

Now we obtain

$$\bar{\lambda}_j(\bar{u}_j - \bar{u}_{j-1}) = h \; \bar{\lambda}_j \frac{\partial f}{\partial y} \; \bar{u}_{j-1} - \epsilon_j \, 10^{-k} \; \bar{\lambda}_j$$

$$= - (\bar{\lambda}_j - \bar{\lambda}_{j-1}) \, \bar{u}_{j-1} - \epsilon_j \, 10^{-k} \; \bar{\lambda}_j$$

or $$(\bar{\lambda}_j \, \bar{u}_j - \bar{\lambda}_{j-1}\bar{u}_{j-1}) = - \epsilon_j \; \bar{\lambda}_j \; 10^{-k} .$$

Thus setting $x_n = X$ and summing from $j = 1$ to n, and using the fact that $\bar{u}_{j-1} = 0$, we have

$$\lambda(X)\bar{u}(X) = - (\Sigma \, \epsilon_j \, \lambda_j) \, 10^{-k}$$

or $$\left| \lambda(X)u(X) \right| \leqq (\Sigma \; |\lambda_j|) \; 10^{-k} \sim \frac{10^{-k}}{h} \int_{x_o}^{X} |\, \lambda \,| \, dx \quad .$$

Thus the round-off error is, in general, of the order of the number of steps $(\frac{1}{h})$ of integration. This estimate can be much improved if we can assume that the rounded-off numbers were distributed at random, i.e., that ϵ_j is a random number. For then the dispersion of $\lambda(x)\bar{u}(x)$ will be given by

$$\sigma^2\Big(\lambda(X)\bar{u}(X)\Big) = (\Sigma \; \lambda_j^2 \int_0^1 \epsilon^2 \, d\, \epsilon) \; 10^{-2k}$$

$$\sim \frac{10^{-2k}}{3 \; h} \int_{x_o}^{X} \lambda^2 \, dx$$

or

$$\sigma\Big(\lambda(X)\bar{u}(X)\Big)\sim\frac{10^{-k}}{\sqrt{3}\,h}\sqrt{\int_{x_o}^{X}\lambda^2\,dx}.$$

The probable round-off error is then 0.674 times this.

Numerical examples indicate that the estimate of the truncation error will be accurate to about 1/2%; and if enough places are carried so that the truncation error lies in the last four places, the estimate for the dispersion of the round-off error will be greater than the round-off error and generally within 20% of the round-off error. If the number of places carried is such that the round-off and truncation errors are approximately equal (as suggested by Rademacher) the error estimates may be off by a factor of 200. Therefore we should have h and k so chosen that, setting

$$D=\frac{0.674}{\sqrt{3}}\sqrt{\int_{x_o}^{X}\lambda^2\,dx}\qquad\qquad E=\left|\int_{x_o}^{X}\Phi\,\lambda\,dx\right|$$

$$1\leqq n\leqq 3$$

we should have

$$h^{\beta+\frac{1}{2}}\sim 10^{-k+n}\;\frac{D}{E}.$$

Among many assumptions, some somewhat justified, the assumption that ϵ_j is randomly distributed stands out as unjustified; indeed Huskey and Hartree have shown that in a simple case it is quite strongly unsatisfied. A criterion suggested by Rademacher is shown in the paper of Huskey and Hartree to be inadequate; the following is the criterion developed by Hartree (for references see the end of this section).

If ϵ_j is not randomly distributed we have systematic errors, i.e., the first integer digits dropped are the same, i.e.,

the last integer digit of $\Delta(10^{k+1}y) = 0$

or

the last integer digit of $10^{h+1}y' \ h = 0$

or

$$10n-0.5 \leqq 10^{k+1} \ y'h \leqq 10n + 0.5$$

where n is an integer. This will occur over a range of values of $\Delta y'$, and we have

$$\Delta y' = \frac{1}{10^{k+1h}}$$

The number N of intervals necessary to cover this range satisfies

$$N \ h \ y'' \sim \Delta y'$$

or

$$N = \frac{1}{10^{k+1}h^2 \ |y''|} \quad .$$

The number of errors will be serious in practice only if N is greater than 3 or 4; thus let us require $N<4$, or

$$4 \ 10^{k+1} \ h^2 \ |y''| > 1 \quad .$$

This gives us the condition

$$h > \frac{1}{\sqrt[2]{4 \cdot 10^{k+1} \ |y''|}}$$

We will now state explicitly the results obtained for the case of a system of equations. Let

$$\frac{dy_i}{dx} = y'_i = f_i(x,y_1,\ldots,y_n) \qquad (i=1,2,\ldots,n)$$

and $\qquad y_i(x_o) = y_{10} \qquad (i=1,2,\dots,n)$.

Let us have a process of integration expressed as $y_i(x,h)$ such that

$$y_i(x+h) - y_i(x,h) = h^{\beta+1}\Phi_i(x) + h^{\beta+2}\Psi_i(x,h)$$

$$\sim h^{\beta+1}\Phi_i(x) \qquad (i=1,\dots,n)$$

where β is now assumed to be independent of i. Introduce the solutions of the "adjoint equations".

$$\frac{d\lambda_j}{d\,x} = -\sum_{i=1} \frac{\partial f_i}{\partial y_j}(x,y_1(x),\dots,y_n(x) \cdot \lambda_i \ .$$

If the exact numerical solution is y_{2j} at $x_j = x \ + j\,h$, and $u_{ij} = y_i(x_j) - y_{ij}$, then setting $u_i(x_j) = u_{ij}$ we find that the truncation error is given by

$$\lambda_1(X)u_1(X) + \dots + \lambda_n(X)u_n(X) \sim h^{\beta} \int_{x_o}^{X} (\lambda_1\Phi_1 + \dots + \lambda_n\Phi_n)\,dx$$

If we round off after k places then the round-off error $\bar{u}_i$, given by the difference between the exact and approximate numerical solution, is bounded as

$$\left|\lambda_1(X)\bar{u}_1(X) + \dots + \lambda_n(X)\bar{u}_n(X)\right| \leq \frac{10^{-k}}{h} \int_{x_o}^{X} \Big(|\lambda_1| + |\lambda_2| + \dots + |\lambda_n|\Big)\,dx$$

If, however, the rounded-off term is distributed at random, then the standard deviation is

$$\sigma(\lambda_1\bar{u}_1 + \dots + \lambda_n\bar{u}_n) \sim \frac{10^{-k}}{\sqrt{3}\,h^{\frac{1}{2}}} \sqrt{\int_{x_o}^{X} (\lambda_1^2 + \dots + \lambda_n^2)\,dx}$$

The probable error is then 0.674 times this σ. The condition that the rounded-off terms be random is

$$h \gtreqqless \max_{j} \left(\frac{1}{\sqrt{4 \cdot 10^{k+1} \quad y_j''}} \right)$$

The optimum choice of h is then such that

$$h \left| \int_{x_o}^{X} (\epsilon \lambda_i \Phi_i) dx \right| \sim \frac{10^{-k+n}}{h^{\frac{1}{2}}} \quad \frac{0.674}{\sqrt{3}} \quad \sqrt{\int_{x_o}^{X} (\epsilon \ \lambda^2) dx}$$

where $1 \leqq n \leqq 3$.

Set

$$D = \frac{0.674}{\sqrt{3}} \quad \sqrt{\int_{x_o}^{X} (\epsilon \ {\lambda^2}_i) \, dx} \qquad E = \left| \int_{x_o}^{X} (\epsilon \ \lambda_i \ \Phi_i) \, dx \right| .$$

Then we should have

$$h^{\beta + \frac{1}{2}} \sim 10^{-k+n} \ \frac{D}{E} \quad .$$

The lower inequality on h then gives

$$\left(10^{-k+n} \quad \frac{D}{E} \right)^{\frac{1}{\beta + \frac{1}{2}}} \geqq \max \ \frac{1}{\sqrt[2]{4 \cdot 10^{k+1} \ \cdot \quad y_j''}}$$

or

$$\left(10^{-k+n} \ \frac{D}{E} \right)^{\frac{2}{\beta + 1}} \cdot 4 \cdot 10^{k+1} \geqq \max \ \frac{1}{y''}$$

which gives the condition

$$10^{-2\frac{4 \, k}{\beta + 1} + k + \frac{2n}{2 \beta + 1}} \ \frac{40 \ D^{4/2 \beta + 1}}{E^{4/2 \beta + 1}} \gtreqqless \max \ \frac{1}{|y''|} \quad .$$

Among the conclusions we can draw from this formula is that a method for which $\beta < 2$ cannot be the most efficient, i.e., if we choose k large, h must not satisfy the condition which makes it an optimum - since $10^{-\frac{4k}{2\beta+1}+k}$ grows small as k grows large.

The references for this section are as follows:

D. Brouwer — Astronomical Journal, vol. 46 (1937) pp. 146-

H. Rademacher — "On the Accumulation of Errors in Processes of Integration on High-Speed Calculating Machines".

Proceedings of a Symposium on Large-Scale Digital Computing Machines. Cambridge, 1948. pp. 176-187.

Huskey & Hartree — "On the Precision of a Certain Procedure of Numerical Integration".

Journal of Research, N.B.S., vol. 42 (1949) pp. 57-62.

(In particular, this last reference gives a numerical example, using an incorrect condition of Rademacher for the round-off errors to be randomly distributed. The agreement between predicted and actual values of the errors is still good).

10. ESTIMATES FOR THE ERROR: PARTIAL DIFFERENTIAL EQUATIONS.

We consider in this section L. H. Thomas extension of the methods of the preceding section to partial differential equations; the reference is to [T 2] .

Any partial differential equation or system of such equations can be put in the form of a system of parabolic equations.

$$\frac{\partial u_i}{\partial t} = f_i(t,x_1,\dots,x_n;u_1,\dots,u_m; \frac{\partial u_1}{\partial x_1},\dots,\frac{\partial^2 u_1}{\partial x^2_1}, \frac{\partial^2 u_1}{\partial x_1 \partial x_2},\dots\dots)$$

$$(i = 1,\dots,m).$$

For example, the equation

$$\frac{\partial^2 u}{\partial t^2} + \frac{\partial^2 u}{\partial x^2} = 0$$

can be reduced to this form by setting $u_2 = \frac{\partial u}{\partial t}$; the system is then

$$\frac{\partial u}{\partial t} = u_2$$

$$\frac{\partial u_2}{\partial t} = - \frac{\partial^2 u}{\partial x^2} .$$

We will consider the round-off and truncation error in a single (parabolic) equation of the form

$$\frac{\partial u}{\partial t} = f(t,x; u,\frac{\partial u}{\partial x} , \frac{\partial^2 u}{\partial x^2}$$

or using the abbreviations

$$p = \frac{\partial u}{\partial x} , q = \frac{\partial^2 u}{\partial x^2}$$

$$\frac{\partial u}{\partial t} = f(t,x; u, p, q)$$

(We will use p and q to give us a convenient notation for the partial derivatives of $f(t,x;u,p,q)$, i.e., as in f_q, f_p). The methods used on this equation can be extended to other equations and systems of equations.

In the given differential equation we replace the differentiation operators $\frac{\partial}{\partial t}$, $\frac{\partial}{\partial x}$, $\frac{\partial^2}{\partial x^2}$ by difference operators L, L_p, L_q such that

$$L\,v = \frac{\partial v}{\partial t} + \Phi(\Delta t, \Delta x;\; u;\; \frac{\partial u}{\partial t};\ldots)$$

$$= \frac{\partial v}{\partial t} + \Phi$$

$$L_p v = \frac{\partial v}{\partial x} + \Phi_p(\Delta t, \Delta x;\; u;\; \frac{\partial u}{\partial t};\; \ldots)$$

$$= \frac{\partial v}{\partial x} + \Phi_p$$

$$L_q v = \frac{\partial^2 v}{\partial x^2} + \Phi_q$$

where Φ, Φ_p, Φ_q are terms of higher order in the mesh steps Δt, Δx. Then we have a difference equation

$$Lv = f(t,x;v,L_p v,L_q v)$$

The truncation error

$$e = u - v$$

then satisfies a linear partial differential equation which is found by subtracting the equations for u and v from one another

$$\frac{\partial u}{\partial t} - Lu = \frac{\partial(u-v)}{\partial t} - \Phi = \frac{\partial e}{\partial t} - \Phi$$

$$= f(x,t;u,\frac{\partial y}{\partial x}, \frac{\partial^2 u}{\partial x^2}) - f(x,t;v,L_p v,L_q v)$$

$$= - f_u(v-u) - f_p\left(L_p v - \frac{\partial u}{\partial x}\right) - f_q\left(L_q v - \frac{\partial^2 u}{\partial x^2}\right)$$

$$= - f_u e - f_p\left(\frac{\partial(u-v)}{\partial x} - \Phi_p\right) - f_q\left(\frac{\partial^2(u-v)}{\partial x^2} - \Phi_q\right)$$

$$= - f_u e - f_p \frac{\partial e}{\partial x} - f_q \frac{\partial^2 e}{\partial x^2} + f_p \Phi_p + f_q \Phi_q$$

or

$$\frac{\partial e}{\partial t} + f_u e + f_p \frac{\partial e}{\partial x} + f_q \frac{\partial^2 e}{\partial x^2} = \Phi + f_p \Phi_p + f_q \Phi_q .$$

(The f, Φ's are here functions of v).

A numerical integration of this equation then gives an approximate value for the truncation error.

We will now consider the rounding-off error. As in the preceding section, it is necessary that the mesh intervals be small in order that the rounding-off errors of each step shall be random. Let us round off after k places, so that the rounding-off error is $\epsilon(x,t)10^{-k}$, where

$$0 \leqq \epsilon(x,t) \leqq 1.$$

We will denote by E(x,t) the accumulated rounding-off error.

By an analysis directly analogous to that of the preceeding sectiom we could integrate the error $\epsilon(x,t)$ of a single step in t to obtain the error E(x,t). In the stable case, we already know that the greater part of the error is truncation error (See the end of Section 7); the analysis of the round-off error will therefore be carried no further here.

As an example, consider the equation

$$\frac{\partial u}{\partial t} = a \frac{\partial^2 u}{\partial x^2}$$

The equation, for the truncation error then becomes

$$\frac{\partial e}{\partial t} = a \frac{\partial^2 e}{\partial x^2} + \frac{1}{2} \Delta t \frac{\partial^2 v}{\partial t^2} \left(1 - \frac{1}{6} \frac{\Delta x^2}{a \Delta t}\right)$$

$$+ \frac{1}{15} \Delta t^2 \frac{\partial^3 v}{\partial +3}$$

(where v is the approximate solution).

If $r = \frac{a \Delta t}{\Delta x^2} \neq \frac{1}{6}$, the equation is

$$\frac{\partial e}{\partial t} = a \frac{\partial^2 e}{\partial x^2} + \frac{1}{2} \Delta t \frac{\partial^2 v}{\partial t^2} \left(1 - \frac{1}{6 r}\right)$$

whose approximate solution is

$$e \sim \frac{1}{2} t \Delta t \frac{\partial^2 v}{\partial t^2} \left(1 - \frac{1}{6 r}\right)$$

while if $r = \frac{1}{6}$, the error becomes of the next higher order in Δt, i.e.,

$$e \sim \frac{1}{15} t \Delta t^2 \frac{\partial^3 v}{\partial t^3}$$