Journal of Research of the National Bureau of Standards Vol. 50, No. 6, June 1953 Research Paper 2424

On the Numerical Solution of Parabolic Partial Differential Equations[1]

Gertrude Blanch

The numerical results presented here relate to a two-dimensional parabolic partial differential equation containing a nonlinear term. Denoting the independent variables by t and x, a lattice is introduced, with intervals k and h in the t-and x-directions, respectively. Much attention has been devoted recently to the study of the conditions on the mesh ratio, k/h^2, under which an approximation by a difference equation converges to the solution of the differential equation for sufficiently small h. Some known results are summarized in sections 1 and 2, and three approximation formulas are given, one of order two, and two of order four. The feasibility of using approximation formulas of order higher than the differential equation is studied in later sections. The primary objective of this paper is to seek the *most economical* mesh ratio for a given approximation formula, that is, of all mesh ratios that will lead to a preassigned upper bound of error in the approximation, to choose that mesh ratio that will lead to the least amount of work. It is shown in section 3 that the largest admissible mesh ratio is not necessarily the most economical one and that a great deal depends on the form of the differential system and the boundary conditions.

In section 4 a generalization is given of the method of Hartree and Womersley (1937) for improving a solution from two difference approximations. The method is shown to be very effective for suitable boundary conditions. Five numerical examples are presented and analyzed in section 5. An appendix, with detailed derivations of the formulas used, is given for the benefit of those who may want to apply the formulas to specific studies.

1. Definitions; Basic formulas

Once the existence of a unique solution to a differential equation has been established, and an approximating function has been found that converges to the solution under suitable conditions, there remains the problem of providing an effective numerical treatment of the approximation. Our study concerns itself with one phase of this problem, for the case when the approximating function is expressed by a difference equation. We shall further limit the discussion to a specific type of differential equation, namely,

$$\frac{\partial u}{\partial t}=\frac{\partial^2 u}{\partial x^2}+f(x,t,u) \quad 0\leq x\leq x_a; \quad 0\leq t\leq t_N \qquad (1)$$

with given initial and boundary conditions. Let us introduce a lattice covering the region, at intervals h in the x-direction and k in the t-direction. Let $\lambda=k/h^2$ be the *mesh ratio*. It is a trivial restriction to assume that $x_a=sh$, where s is an integer.

NOTATION. For the sake of brevity, we shall write $u(x,t)=u_{m,n}$, if (x,t) is a lattice point.

$$\left[\frac{\partial^p g}{\partial x^s \partial t^{p-s}}\right]_{\xi,n}=\frac{\partial^p g_{\xi,\eta}}{\partial x^s \partial t^{p-s}},$$

where g is any function under consideration. Furthermore, following the usual convention for even central differences, we define,

$$\delta_x^{2p} u_{m,n}=\sum_{w=-p}^{w=p}(-1)^{p+w}\binom{2p}{p+w}u_{m+w,n}. \qquad (2)$$

The point in the x,t-plane with coordinates $x=mh$, $t=nk$ will be denoted by (m,n).

Let u be regular at (m,n). Then for k sufficiently small there is a Taylor series in t around (m,n)

$$u_{m,n+1}=u_{m,n}+\sum_{p=1}^{\infty}\frac{k^p}{p!}\frac{\partial^p u_{m,n}}{\partial t^p}. \qquad (3)$$

If in (3) terms involving $p\geq 2$ are dropped, and if $\partial u/\partial t$ is replaced by the right-hand side of (1), with $\partial^2 u/\partial x^2$ approximated by central differences, we get the well-known approximation to u:

$$\begin{aligned} v_{m,n+1}&=v_{m,n}+\lambda\delta^2_{m,n}+kf_{m,n}\\ &=(1-2\lambda)v_{m,n}+\lambda(v_{m-1,n}+v_{m+1,n})+kf_{m,n}. \qquad (4)\end{aligned}$$

Similarly, if terms through $p=2$ are retained, we obtain,

$$\begin{aligned} v_{m,n+1}&=v_{m,n}+\left[\lambda\delta^2+\left(\frac{\lambda^2}{2}-\frac{\lambda}{12}\right)\delta^4\right]v_{m,n}+k\varphi_{m,n}\\ &=\left(1-\frac{5}{2}\lambda+3\lambda^2\right)v_{m,n}+\left(\frac{4}{3}\lambda-2\lambda^2\right)(v_{m-1,n}+v_{m+1,n})\\ &\quad+\left(\frac{\lambda^2}{2}-\frac{\lambda}{12}\right)(v_{m-2,n}+v_{m+2,n})+k\varphi_{m,n}, \qquad (5)\end{aligned}$$

[1] The preparation of this paper was sponsored (in part) by the Office of Naval Research, USN.

where

$$\varphi_{m,n}=kf_{m,n}+\tfrac{1}{2}k^2\left(\frac{df_{m,n}}{dt}+\frac{d^2f_{m,n}}{dx^2}\right).$$

Formulas (4) and (5) need modification at the boundary to satisfy given initial and boundary conditions to a required accuracy. This modification can, in general, be made. In the present study we shall assume that all required derivatives exist and are continuous. For parabolic equations this is not a serious restriction, for let us assume that we have generated values of v for a given t. From (4) and (5) it is clear that these serve as boundary values in a subdomain for generating the set of values for the next t in the lattice. In imposing the continuity restrictions, we therefore merely imply that regions close to the given boundary, which may have "corners" or other discontinuities, will be treated separately. This in fact must usually be done in practice, either by choosing a lattice that is much finer than that required over the major portion of the region or by special approximations that are appropriate for the particular problem. Our main concern here is with the choice of the mesh ratio, λ, for the major portion of the domain where the function is presumed to be regular. The manner in which an error (or variation) in the boundary conditions is propagated over the rest of the domain must, of course, be examined; but this problem is no different at the boundary of the given domain than at any of the subdomains (that is, at the successive values of t in the lattice). This problem will not be considered here.

Truncation terms. Let $\sigma_{m,n}=u_{m,n}-v_{m,n}$. It can be verified that corresponding to (4),

$$\sigma_{m,n+1}=\sigma_{m,n}+\lambda\delta^2\sigma_{m,n}+k\,\frac{\partial\overline{f}_{m,n}}{\partial u}\,\sigma_{m,n}+kT_{2,m,n},\quad(6)$$

where

$$T_{2,m,n}=h^2\left[\left(\frac{\lambda}{2}-\frac{1}{12}\right)\frac{\partial^4u_{m,n}}{\partial x^4}+\theta_1\right]+O(h^3).\quad(7\text{a})$$

Another form for $T_{2,m,n}$ is given below:

$$T_{2,m,n}=h^2\left[\frac{\lambda}{2}\,\frac{\partial^2u_{m,n}}{\partial t^2}-\frac{1}{12}\,\frac{\partial^4u_{m,n}}{\partial x^4}\right]+O(h^3).\quad(7\text{b})$$

Corresponding to (5),

$$\sigma_{m,n+1}\cong\sigma_{m,n}+\lambda\delta^2+\left(\frac{\lambda^2}{2}-\frac{\lambda}{12}\right)\delta^4\sigma_{m,n}+\left[k\,\frac{\partial\overline{\varphi}}{\partial u}\right]_{m,n}\sigma_{m,n}+kT_{4,m,n},\quad(8)$$

where two expressions for $T_{4,m,n}$ are given below.

$$T_{4,m,n}=h^4\left[M(\lambda)\,\frac{\partial^6u_{m,n}}{\partial x^6}+\theta_2\right]+O(h^5),\quad(9\text{a})$$

$$T_{4,m,n}=h^4\left[\frac{\lambda^2}{6}\,\frac{\partial^3u_{m,n}}{\partial t^3}+\left(\frac{1}{90}-\frac{\lambda}{12}\right)\frac{\partial^6u_{m,n}}{\partial x^6}\right]+O(h^5)\quad(9\text{b})$$

where

$$M(\lambda)=\frac{\lambda^2}{6}-\frac{\lambda}{12}+\frac{1}{90}.\quad(10)$$

In the above θ_1 and θ_2 depend on f and its derivatives; if $f=0$, then $\theta_1=\theta_2=0$. Further, $\partial\overline{f}/\partial u$ and $\partial\overline{\phi}/\partial u$ imply evaluation of these functions at (m,n) corresponding to a value of $\overline{u}$ intermediate between $u_{m,n}$ and $v_{m,n}$. Again, (6), (7), (8), and (9) need modification in the immediate vicinity of the boundary.

Definition. $T_{r,m,n}$ will be said to be of order r if it contains h^r as a factor, but not h^{r+1}.

2. Stability Considerations; Bounds for the Errors in the Approximations

Definition. The solution $v(x,t)$ of an approximating difference equation will be defined as *stable* if it is bounded for all finite t, independently of h.

Consider the special case when $f(x,t,u)=0$, and $u(x,0)=A(x)$ is defined and bounded for $-\infty<x<\infty$. It will be convenient to refer to the differential eq (1) under these special conditions as the *basic homogeneous* equation. Let this basic homogeneous equation be approximated by a difference equation of order $2p$, so that we can write,

$$v_{m,n+1}=\sum_{w=0}^{p}b_w\delta^{2w}v_{m,n}\equiv\sum_{w=0}^{p}a_wv_{m+w,n},\quad(11)$$

where the coefficients b_w and a_w are suitable constants. It is easy to show that $\Sigma a_w=1$; in general, the coefficients a_w will be functions of λ. Let q be an upper bound of $|A(x)|$. If it is possible to choose λ so that all the coefficients a_w are nonnegative, we shall have, from (11),

$$|v_{m,1}|\leq\Sigma a_w|v_{m+w,0}|\leq q\Sigma a_w=q,$$

and by induction on n we establish that $|v_{m,n}|\leq q$ for all m,n; hence $v(x,t)$ is stable. The condition that all a_w be nonnegative is therefore a sufficient, though not a necessary condition [2] for the stability of $v(x,t)$ in the case of the basic, homogeneous

[2] The above definition of stability and the observation that (11) is stable when all the coefficients a_w are nonnegative were mentioned by F. John in seminar talks at the Institute for Numerical Analysis.

equation. We shall refer to the range of λ for which the basic, homogeneous equation is stable as the "admissible" range of λ. In what follows λ will always be chosen to lie within the admissible range; this may not guarantee that the solution $v(x,t)$ will be stable for arbitrary boundary conditions and functions $f(x,t,u)$. However, the choice of λ within this range usually simplifies the error analysis for any set of boundary conditions, even when (1) is not homogeneous.

Turning now to the first approximation formula defined by (4), it is clear that the admissible range of λ is $0<\lambda\leq\frac{1}{2}$. The corresponding equations for $\sigma_{m,n}$ are given in (6). Let τ be an upper bound of $|T_{2,m,n}|$ and let ω be an upper bound of $|\partial f/\partial u|$, this upper bound to be independent of h. If $\sigma_{m,0}=0$, eq (6) yields, for a wide range of initial and boundary conditions, $|\sigma_{m,1}|<k\tau$. Since λ is in the admissible range, we have further

$$|\sigma_{m,n+1}|\leq|\sigma_{m,n}|(1+k\omega)|+k\tau.$$

Now $(1+\text{k}\omega)^{N}\leq e^{t\omega}$, where $t=Nk$; hence at time t,

$$|\sigma_{m,N}|\leq t\tau e^{t\omega}. \qquad (12)$$

Boundary conditions may impose modification of (12). If the conditions are such that no negative power of h is added as a factor to the right-hand side of the inequality (12), then it can be shown from the form of $T_{2,m,n}$ that an upper bound of τ can be found that has h^2 as a factor. It follows then from (12) that $v_{m,n}\rightarrow u_{m,n}$ as $h\rightarrow 0$.

Let us now consider the second approximation formula, defined by (5). The coefficients of $v_{m+w,n}[\pm w=0,1,2]$ in (5) will all be positive if $0<\lambda\leq\frac{2}{3}$. The approximation $v(x,t)$ to the basic homogeneous equation will therefore be stable for this range of λ, and by the same analysis as before, we can show that $v(x,t)$ approaches $u(x,t)$ as h approaches zero, for a wide range of boundary conditions.

3. Criteria for the Choice of a Suitable Mesh Ratio

The error $\sigma_{m,n}$ is a function of h,λ, and of the boundary conditions associated with the differential equation. Given an upper bound of error that can be tolerated in the solution, the problem involves choosing h and λ (the latter within the admissible range) so as to meet requirements with the least amount of work. We shall assume that for a given approximation scheme, the work is proportional to the number of lattice points at which $u_{m,n}$ must be evaluated. This is not strictly true. For let us define a *profile* as a set of values of $v_{m,n}$ for a fixed n, and all $m\leq s$. If, for example, successive profiles are generated from preceding ones on an IBM machine such as the card programmed calculator, then merging operations may be required at the end of a profile which may consume some time. Thus a grid of 10 points in the x-direction and 100 points in the t-direction may take more time to generate than a grid of 20 points in the x-direction and 50 points in the t-direction. Nevertheless, the assumption that the work is proportional to the number of lattice points is close enough to reality to be useful. Of course, the complexity of the programming must be considered; thus an approximation formula of order four may take more machine cycles (hence more time) than one of order two. However, for the same approximation scheme, the choice of λ does not change the amount of work radically. Let $X=sh$ be the range of x and $t_1=Nk$ the range of t. The number of lattice points in the region is $Ns=(Xt_1/hk)=Xt_1/\lambda h^3$. As Xt_1 is fixed, the work required for a given approximation scheme is therefore inversely proportional to λh.[3]

If the exact solution for $\sigma_{m,n}$ were known, it would be theoretically possible to study the magnitude of the error for various choices of λ and h corresponding to a given approximation scheme. The precise solution $\sigma_{m,n}$ is not easy to find. However, from (12), it is clear that an upper bound which can be approximated has $|T_{r,m,n}|$ as a factor. We shall, therefore, aim to choose λ and h in such a manner as to make $|T_{r,m,n}|$ small. Moreover, for both approximation formulas (4) and (5), the successive terms of $h^2T_{r,m,n}$ involve $h^{r+p}\partial^{r+p}u/\partial x^{r+p}$ and $h^{r+p}\partial^{r+p}f/\partial x^{r+p}$. We shall require that for all choices of λ, the interval h be sufficiently small so as to satisfy

$$|\delta^{r+p}u|>c|\delta^{r+p+1}u|, \quad p\geq 0, \quad 0<c<1, \qquad (13)$$

almost everywhere. From the known relations connecting derivatives with differences, it is clear that (13) implies that successive terms of $T_{r,m,n}$ will be numerically smaller than preceding ones. The phrase "almost everywhere" for the condition (13) needs explanation. It may happen that in a region where $\delta^{r+p}u$ changes sign, a few entries of $\delta^{r+p}u$ may be numerically smaller than corresponding entries in the higher differences. Such a case may also arise near critical points. In particular the condition (13) shall be satisfied by the initially given values and $f(x,0,u)$. In practice one often requires that c be $\frac{1}{10}$ or $\frac{1}{5}$. For one powerful check on the accuracy of computed values is obtained from the pattern of successive differences of the entries. Hence, even if the criterion (13) were unnecessary from the viewpoint of estimating an upper bound of error in the solution, it would still be a desirable condition to impose, in order to insure that the computed values difference with reasonable ease. We shall further require that the term of $T_{r,m,n}$ involving the lowest power of h shall approximate the magnitude of $T_{r,m,n}$ to within a factor of two. The fact that a restriction is thereby imposed on h must be clearly kept in mind. In general a value of h small enough to satisfy (13) will not necessarily make $|T_{r,m,n}|$ small enough to meet requirements for a given upper bound of error in

[3] It is no serious restriction to consider the range t_1 as an integral multiple of h.

the solution. In the instance when it does not, the problem posed is to choose λ and h judiciously, so as to bring the truncating error within permissible bounds with the least amount of work. But if the required accuracy is rather low, it may well happen that after h is chosen small enough to satisfy (13), the error measured by $T_{r,m,n}$ may already be small enough to meet all requirements. In that case the largest λ within the admissible range will, of course, lead to the least amount of work.

With these observations we shall now attempt to study the dependance of the truncation term on λ and on h. Ideally, it would be desirable to classify differential equations into several types, according to the value of λ which is appropriate for equations belonging to the type in question. A complete classification is difficult to set down, but an attempt in this direction is made by considering two types:

Type I. This type is characterized by the following conditions:

$$\frac{df}{dt}+\frac{d^2f}{dx^2}=\frac{d^4f}{dx^4}=0. \tag{14}$$

When (14) holds, it follows that

$$\frac{\partial^2 u}{\partial t^2}=\frac{\partial^4 u}{\partial x^4}; \quad \frac{\partial^3 u}{\partial t^3}=\frac{\partial^6 u}{\partial x^6}.$$

It can be shown that for systems belonging to type I, the terms of $T_{r,m,n}$, $r\leq 4$, do not involve $f(x,t,u)$ or its derivatives. The basic homogenous equation belongs to this type.

Type II. This type is characterized by the condition that successive derivatives of $u(x,t)$ with respect to t are known to be much smaller than corresponding derivatives with respect to x (usually of twice the order) with which they are associated in truncation terms such as (7b) and (9b). It is quite easy to find systems belonging to this type; the numerical illustration given in section 5 belongs to type II.

Consider differential systems of type I, and let us start with the case when formula (4) is used. The corresponding truncation term, $T_{2,m,n}$, is given by (7a). It has been observed by Milne, as well as by Thomas and others, that if λ is chosen as $\frac{1}{6}$, then the leading term of $T_{2,m,n}$ vanishes, so that the order of $T_{2,m,n}$ becomes four. With $\lambda=\frac{1}{6}$ and h small enough to satisfy the conditions for an upper bound of $|T_{2,m,n}|$—enough to meet the requirements for a given upper bound of error in the solution—a certain amount of work will be done, which we shall measure in units of $Z=1/\lambda h^3$, as it has been shown that for a rectangular lattice the work is proportional to Z, approximately. If another λ and h were chosen, then we shall ask whether, for the *same amount of work,* and using the same formula (4), greater accuracy can be obtained in the solution, assuming that accuracy to be measured by the magnitude of the truncation term. In other words, we shall compare various choices of λ and h, for a constant Z. It can be readily shown that for equations belonging to type I, no other choice of h and λ, keeping Z constant, will be as good. In this sense it is correct to state that $\lambda=\frac{1}{6}$ is the best choice for formula (4). However, one important point has been overlooked. We have seen that h is not completely free, for h must be small enough to meet the minimum conditions imposed by (13). It may happen that after h has been taken small enough to insure that (13) is satisfied, the conditions for an upper bound of error in the solution can be satisfied with some range of $\lambda>\frac{1}{6}$. In that case, we would certainly get a more *accurate* solution by choosing $\lambda=\frac{1}{6}$, but that would be more accuracy that required, and work could be saved by choosing a larger λ. We conclude that even for equations of type I, the choice of $\lambda=\frac{1}{6}$ will be best only if a relatively high accuracy is required in the solution. Moreover, in practice the differences of $u(x,t)$ can usually be judged only from the initially given profile, at the time when λ and h are chosen, so that a *safe* interval h (rather than the maximum permissible for the given profile) is often chosen, and it may happen that the value of h considered necessary may, as stated, bring the error within the tolerance limits for all choices of $\boldsymbol{\lambda}$.

Still considering systems that belong to type I, let us now examine approximations of order four. The expression for the truncation term $T_{4,m,n}$ is given in (9a) and the term of order four is

$$h^4A_4(x,t,\lambda)=h^4\left(\frac{\lambda^2}{6}-\frac{\lambda}{12}+\frac{1}{90}\right)\frac{\partial^6 u_{m,n}}{\partial x^6}=h^4M(\lambda)\ \frac{\partial^6 u_{m,n}}{\partial x^6}.$$

It can be readily verified that $M(\lambda)$ is positive for all choices of λ, hence the leading term of $T_{4,m,n}$ cannot be eliminated completely, as in the case of the simpler approximation of order two. However, we may seek that value of λ which will make the unit of work, $Z=1/\lambda h^3$, a minimum, subject to a given permissible upper bound of $|T_{4,m,n}|$. By hypothesis, h is small enough so that $h^4A_4(x,t,\lambda)$ approximates the magnitude of $|T_{4,m,n}|$; hence if C is the permissible tolerance of $|T_{4,m,n}|$, we wish to satisfy

$$h^4|A_4(x,t,\lambda)|=h^4M(\lambda)\left|\frac{\partial^6 u}{\partial x^6}\right|\leq C, \tag{15}$$

Since $|\partial^6u/\partial x^6|$ is independent of the choice of parameters h and λ, the inequality expressed in (15) can be satisfied by taking

$$h^4M(\lambda)=C_1; \quad C_1\leq C/\left|\frac{\partial^6 u}{\partial x^6}\right|. \tag{16}$$

Thus we seek to determine h and λ, which minimize Z, subject to the condition

$$h^4 M(\lambda) = C_1. \quad (17$$

From (17), $h = C_1/[M(\lambda)]^{\frac{1}{4}}$. Substituting this value of h into Z and differentiating the resulting expression of Z with respect to λ, we obtain the following condition for a minimum of Z:

$$F = 3\lambda M'(\lambda) - 4M(\lambda) = 0,$$

or

$$60\lambda^2 + 15\lambda - 8 = 0. \quad (18)$$

The positive root of (18) is $\lambda_1 = 0.26095\ .\ .\ .$, and it can be verified that $F'(\lambda_1)$ is positive, so that Z is indeed a minimum for this value of λ. Since λ lies within the admissible range, it can be used for the mesh ratio, in conjunction with a suitable value of h which satisfies (18). In practice, $\lambda = 0.25$ will normally be used, because an irrational value of λ is inconvenient.

It will be instructive to examine the following schedule, which gives $M(\lambda)$, h, and the work unit Z *for a constant value of* $h^4M(\lambda)$, and various values of λ within the admissible range, in units of the corresponding quantities when $\lambda = \frac{1}{2}$.

λ	$M(\lambda) = \left(\frac{\lambda^2}{6} - \frac{\lambda}{12} + \frac{1}{90}\right)$	h	$Z = 1/(\lambda h^3)$
1/6	0. 001851	1. 565h_1	0. 782z_1
1/5	. 001111	1. 778h_1	. 444z_1
1/4	. 000694	2. 000h_1	. 25z_1
3/10	. 001111	1. 778h_1	. 297z_1
1/3	. 001851	1. 568h_1	. 391z_1
3/8	. 003299	1. 354h_1	. 536z_1
1/2	. 011111	h_1	z_1
2/3	. 029630	0. 783h_1	1. 565z_1

Compared with a unit of work Z_1 for the case $\lambda = \frac{1}{2}$, only $\frac{1}{4}Z_1$ is required when $\lambda = \frac{1}{4}$. The largest admissible value of λ is the poorest of all, from the viewpoint of the amount of work required. Again, a word of caution is required. For the same magnitude of the error in the leading term of $T_{4,m,r}$ and different choices of λ, the above schedule shows that with $\lambda = \frac{1}{4}$, h can be chosen twice as large as that required when $\lambda = \frac{1}{2}$. If h is made twice as large, the eighth difference in the x-direction is multiplied by about 2^8, and the differences must be reexamined to see whether this larger value of the eighth difference still satisfies the fundamental conditions imposed by (13) namely that successive differences in the x-direction beyond the fourth be numerically smaller than preceding ones. Moreover, it has been pointed out before that the maximum h corresponding to which (13) is satisfied may already be such that $T_{4,m,n}$ is within the required tolerance for all admissible values of λ. In that case it is of course best to choose $\lambda = \frac{2}{3}$ or one close to it.

The following question arises: since the simpler approximation (4) has a truncating error of order four when $\lambda = \frac{1}{6}$, and $T_{4,m,n}$ is also of order four, is there any gain in using the approximation formula of higher order? The answer to this question is complicated by the fact that the time required to generate a profile corresponding to the higher approximation may be considerably longer than that for the simpler approximation. Much will depend not only on the computing instrument which will be used, but also on the complexity of the boundary conditions. If the IBM card programmed calculator is to be used, and the boundary conditions are not too complex, the simpler approximation (4) can perhaps be generated in only two-thirds of the time per profile, compared with the more elaborate fourth order approximation given in (5). There are, however, compensating factors which make the higher approximation worth considering. Let us assume we are dealing with a case where the "best" values of λ are used in the approximation of order two and the one of order four. For the same h used in both cases, the number of lattice points is inversely proportional to λ. Hence, we shall use only two-thirds the number of lattice points when the higher approximation is used. This would about compensate for the longer time it may take to generate each profile. There is, however, a gain in accuracy when the higher approximation is used. For the coefficient of $h^4(\partial^6u/\partial x^6)$ in (9a), corresponding to the simpler approximation, is 1/540 when $\lambda = \frac{1}{6}$. On the other hand, the coefficient of the corresponding term in the higher approximation with $\lambda = \frac{1}{4}$ is only 0.000694; or less than three-eighths that of the simpler approximation. Furthermore, in cases where the maximum permissible h is such that $T_{4,m,n}$ is already within the required tolerance limit for all values of λ, we may be able to take $\lambda = \frac{2}{3}$, or close to it, when the higher approximation is used. Such a choice of λ would cut the number of lattice points to $\frac{1}{4}$ that required for the simpler approximation, with $\lambda = \frac{1}{6}$, and that might more than compensate for the greater difficulty in generating $v(x,t)$ by the higher approximation.

Let us now consider differential systems that belong to type II. Since, by hypothesis, the derivatives in the t-direction are negligible compared with those in the x-direction, (7b) shows that the leading term in $T_{2,m,n}$ is practically independent of λ, and this term cannot be eliminated by any choice of λ. Hence, it is reasonable to take the maximum admissible λ, or one close to it. Similarly, the leading term of $T_{4,m,n}$ is complex in structure, and there is no optimum λ that stands out as suitable for all functions falling under this type. For a given problem, it may be possible to compute estimates of the various terms that contribute to the truncation error. Or if the problem involves a family of parameters, in the boundary condition, then available results for some members of the family may lead to an optimum choice of λ and h for the remaining ones.

For equations falling under type II the higher approximation formula usually has distinct advantages over the simpler one of order two. For in this case the truncation term is necessarily of a lower order of magnitude in the higher approximation formula than in the simpler one; moreover, a larger mesh ratio, namely, $\frac{2}{3}$, can be taken. Whenever the boundary conditions are such that the coding problem is manageable, the higher approximation formula is to be recommended.

It might be well to remark that in some cases the solution may be an oscillating function of x but not of t. Such functions may fall under type II, and a Fourier approximation may actually be better than a finite difference approximation. However, there is no reason to expect that all, or even a good portion of functions belonging to type II, can be most simply treated by Fourier approximations. A finite difference approximation such as (4) or (5) is often preferred, because of its simplicity, to other types of approximations (by Fourier series or perhaps "implicit" difference approximations). Our concern here, as stated before, is to study the dependence of the solution on the choice of λ, after either (4) or (5) has been selected as the approximation formula.

4. Method of Improving the Solution from Two Difference Approximations

The method to be explained below has been presented in [1] [4] by Hartree and Womersley, who ascribe the idea to L. F. Richardson [5]. In [1] the method is applied to a somewhat different type of approximation—a mixed difference—differential scheme, suitable for computation by a differential analyzer. The results will here be extended to difference approximations of any other.

Let us suppose that values of $v_{m,n}$ have been generated by an approximation formula, corresponding to a true solution $u_{m,n}$, and let us assume that it is possible to write

$$\begin{aligned}\sigma_{m,n}&=u_{m,n}-v_{m,n}\\&=h^rC_r(x,t,\lambda)+h^{r+1}C_{r+1}(x,t,\lambda)+\ldots\quad,\end{aligned}\tag{19}$$

where the functions $C_j(x,t,\lambda)$ are independent of h. It can be shown that corresponding to a wide range of boundary conditions the expression for $\sigma_{m,n}$ does assume the form (19) for both approximation formulas (4) and (5).

Consider now the case when $v_{m,n}$ has been generated by the use of an interval h_1 in the x-direction and a mesh-ratio λ. Let these values of $v_{m,n}$ be designated by $v_{m,n}(h_1)$. Now let another computation be made, based on the same mesh ratio λ, but at an interval h_2 in x, where $h_2=\rho h_1$, $0<\rho<1$. These values will be designated [5] by $v_{m,n}(h_2)$. It is presumed that the same approximation formula will be used in both cases. Clearly,

$$v_{m,0}(h_1)=v_{m,0}(h_2)=u_{m,0};\quad v_{s,n}(h_1)=v_{s,n}(h_2)=u_{s,n}.$$

By hypothesis, we have from (19)

$$\begin{aligned}u_{m,n}-v_{m,n}(h_1)&=h_1^rC_r+h_1^{r+1}C_{r+1}\\&\quad+h_1^{r+2}C_{r+2}+\ldots\quad.\end{aligned}\tag{20}$$

Similarly, remembering that $h_2=\rho h_1$ and that C_j is independent of h,

$$\begin{aligned}u_{m,n}-v_{m,n}(h_2)&=\rho^rh_1^rC_r+\rho^{r+1}h_1^{r+1}C_{r+1}\\&\quad+\rho^{r+2}h_1^{r+2}C_{r+2}+\ldots\quad.\end{aligned}\tag{21}$$

Multiply (20) by ρ^r and subtract from (21). This gives, after transposing some terms and dividing by $(1-\rho^r)$,

$$\begin{aligned}u_{m,n}&=v_{m,n}(h_1,h_2)-\frac{\rho^r(1-\rho)}{1-\rho^r}h^{r+1}C_{r+1}\\&\quad-\rho^r\frac{(1-\rho^2)}{1-\rho^r}h^{r+2}C_{r+2}+\ldots\quad,\end{aligned}\tag{22}$$

where

$$v_{m,n}(h_1,h_2)=\frac{v_{m,n}(h_2)-\rho^rv_{m,n}(h_1)}{1-\rho^r}.\tag{23}$$

We shall refer to the process defined in (23) as the "ρ^r" correction. In [1] it is recommended that ρ be taken as $\frac{1}{2}$. This is a desirable choice for numerical work, since every other value in the x-direction at the smaller interval is available at the larger interval. Similarly, every fourth value in the t-direction will be available at both intervals. There are many ways of applying corrections of this type. One way, for example, is to generate four values in the t-direction at the finer interval, then generate the corresponding values at the larger interval; apply the ρ^r correction to the last profile, and use these corrected values as initial data for generating the next profile. Such a use of the correction scheme would, of course, require interpolating for values of $v(h_1,h_2)$ for every other value of x, since we can correct only for those points that are available at both intervals. The coding of this method would be complicated. The simplest way of using the correction scheme is to actually make two completely separate computations for all required values of n, and then to apply the correction process only to those functional values that are required. Often the last profile generated is of most interest, and in that case it may be enough to correct the values on the last computed profile only.

The process furnishes a powerful check on the convergence rate of the approximations, and since the work done at the larger interval is one-eighth that done at the smaller interval if $\rho=\frac{1}{2}$, the added labor is not too costly, when the two computations are carried separately. Moreover, the coding is the same

[4] Figures in brackets indicate the literature references at the end of this paper.
[5] In the subsequent discussions we shall write $v(h_2)$, or $v(h_1,h_2)$ to indicate $v_{m,n}(h_2)$ or $v_{m,n}(h_1,h_2)$, when no ambiguity is likely to arise.

for both approximation schemes. For the special case when $r=2$, (22) and (23) reduce to

$$u_{m,n}=v_{m,n}(h_1,h_2)-\frac{\rho^2h^3C_3}{1+\rho}-\rho^2h^4C_4+\cdots, \tag{24}$$

where

$$v_{m,n}(h_1,h_2)=\frac{v_{m,n}(h_2)-\rho^2v_{m,n}(h_1)}{1-\rho^2}. \tag{25}$$

The "ρ^r" correction has been applied in the numerical examples given in section 5 with highly satisfactory results. In [1] Hartree and Womersley give *sufficient* conditions on the nature of the boundary for the method to be valid; the method can probably be used over a wider class of functions than those specified in [1].

If h_1 is sufficiently small, $v(h_1,h_2)$ will always be an improvement over $u(h_2)$, since the truncation term is of lower order of magnitude in $v(h_1,h_2)$. The question arises: how small must h be to insure that $v(h_1,h_2)$ shall be an improvement over $u(h_2)$? Let the truncation term, $T_{r,m,n}$, now be identified by $T_{r,m,n}(h)$, when associated with an interval h in the x directions, and let

$$\sigma(h_1,h_2)=u-v(h_1,h_2);\quad \sigma(h)=u=v(h);$$

$$T_{r,m,n}(h_1,h_2)=\frac{T_{r,m,n}(h_2)-\rho^rT_{r,m,n}(h_1)}{1-\rho^r}. \tag{26}$$

It is more difficult to get a good estimate of σ than of $T_{r,m,n}$. Just as in section 3, we shall, therefore, inquire under what conditions the inequality

$$|T_{r,m,n}(h_1,h_2)|<|T_{r,m,n}(h_2)| \tag{27}$$

will be satisfied. A numerically smaller truncation term will usually be associated with a smaller total error. For the approximation formulas considered here the truncation term is of the form

$$T_{r,m,n}(h_2)=\sum_{p=r}^{\infty}\rho^ph^pg_p(x,t,u).$$

Moreover, the condition following eq (13) guarantees that

$$T_{r,m,n}(h_2)\cong\rho^rh^rg_r(x,t,u). \tag{28}$$

But no similar statement was made about $T_{r,m,n}(h_1)$. Now from

$$T_{r,m,n}(h_1,h_2)=\frac{-\sum_{p=1}^{\infty}\rho^r(1-\rho^p)h^{r+p}g_{r+p}(x,t,u)}{(1-\rho^r)}.$$

If we can satisfy

$$|hg_{r+p}(x,t,u)|\leq|\rho^cg_{r+p-1}(x,t,u)|, \tag{29}$$

where

$$\rho^c=1-\rho^r, \tag{30}$$

then from (27) and (30) it can be shown by summing the absolute values of the terms of $T_{r,m,n}(h_1,h_2)$ that (27) is satisfied. When $\rho=\frac{1}{2}$ and $r=2$, the condition (29) implies that $\rho^c=\frac{3}{4}$. If $\rho=\frac{1}{2}$ and $r=4$, then $\rho^c=\frac{15}{16}$, or what is equivalent, successive terms of $T_{r,m,n}$ at the larger interval are required to be just smaller numerically than preceding terms. Under such conditions $v(h_1,h_2)$ will always be an improvement over $u(h_2)$.

In the foregoing, it has been assumed that $v_{m,n}$ can be computed exactly by the prescribed formula, and that all initial and boundary conditions are exact. This can seldom be realized in practice, and there will be rounding errors committed at every step of the computation, due to carrying a fixed number of decimals or significant figures in the computations. The cumulative effect of such errors can perhaps be studied statistically, or upper bounds for errors of this type [2] can be found. In the numerical examples given in section 5, the cumulative round-off error was very small, after 100 steps in t.

5. Numerical Examples

The problem selected for analysis was the following one:

$$\frac{\partial u}{\partial t}=\frac{\partial^2u}{\partial x^2}+f(x,u). \tag{31}$$

$$\text{At } t=0,\quad u=(10x^2+x^3+0.64)^{1/2}=A(x);$$

$$\text{at } x=1,\quad u=(11.64+t)^{1/2}=B(t).$$

$$\left[\frac{\partial u}{\partial x}\right]_{x-0}=0;\quad f(x,u)=\frac{-(3x+9.5)}{u}+\frac{(10x+1.5x^2)^2}{u^3}.$$

This particular form was chosen because it represents a case in which derivatives with respect to t are much smaller numerically than corresponding derivatives with respect to x, with which they are associated in truncation terms such as those of (7b). The differential system belongs to type II of section 3. It is known that the exact solution to the problem is

$$u=(10x^2+x^3+0.64+t)^{1/2}.$$

The choice of a nonlinear form of the differential equation was deliberate. It was made in order to study the cumulative error in cases where a considerable number of operations, that involve the approximate values of u, have to be performed at each step.

The following five sets of solutions were generated:

Example 1.

Formula:

$$v_{m,n+1}=v_{m,n}+\lambda\delta^2 v_{m,n}+kf_{m,n};\quad v_{m,o}=A(x);$$

$$v_{s,n}=B(t);\quad sh=1.$$

$$v_{-1,n}=v_{1,n}+\frac{h^3}{3}\left[\frac{\partial f}{\partial x}\right]_{x=0}=v_{1,n}-\frac{h^3}{v_{0,n}}.$$

Parameters: $h=0.05=h_2$; $k=0.00125$; $\lambda=\frac{1}{2}$.
Range of t: $0\leq t\leq 0.125$.
Number of lattice points: $20\times 100=2{,}000$.
Initial values and the computed values of $v_{n,n}$ for the last profile are given in table 1.

Table 1. *Solution of the difference equation*

$$v_{m,n+1}=v_{m,n}+\lambda\delta^2 v_{m,n}+kf(x,u);\quad v_{-1,n}=v_{1,n}-\frac{h^3}{v_{0,n}}$$

$$At\ t=0,\quad v(x,0)=(10x^2+x^3+0.64)^{\frac{1}{2}};\quad At\ x=1,\quad v(1,t)=(11.64+t)^{\frac{1}{2}}$$

$$f(x,v)=\frac{-(3x+9.5)}{v}+\frac{(10x+1.5x^2)^2}{v^3};\quad \lambda=\tfrac{1}{2},\quad h_2=0.05;\quad k=.00125$$

$$u=\sqrt{10x^2+x^3+t+0.64}$$

x	$v(x,0)$	At $t=0.125$		
		$v(x,t)$	$u(x,t)$	$u(x,t)-v(x,t)$
0. 00	0. 8000000	0. 8720400	0. 8746427	+. 0026027
. 05	. 8155520	. 8864322	. 8888897	+. 0024575
. 10	. 8608136	. 9285124	. 9305912	+. 0020788
. 15	. 9318664	. 9950813	. 9966819	+. 0016006
. 20	1. 0237187	1. 0819105	1. 0830512	+. 0011407
. 25	1. 1316470	1. 1848335	1. 1855905	+. 0007570
. 30	1. 2517987	1. 3003062	1. 3007690	+. 0004628
. 35	1. 3812585	1. 4255420	1. 4257892	+. 0002472
. 40	1. 5178933	1. 5584295	1. 5585249	+. 0000954
. 45	1. 6601581	1. 6973964	1. 6973876	−. 0000088
. 50	1. 8069311	1. 8412718	1. 8411952	−. 0000766
. 55	1. 9573898	1. 9891819	1. 9890638	−. 0001181
. 60	2. 1109240	2. 1404665	2. 1403270	−. 0001395
. 65	2. 2670741	2. 2946235	2. 2944770	−. 0001465
. 70	2. 4254896	2. 4512642	2. 4511221	−. 0001421
. 75	2. 5858993	2. 6100866	2. 6099568	−. 0001298
. 80	2. 7480902	2. 7708507	2. 7707399	−. 0001108
. 85	2. 9118937	2. 9333664	2. 9332788	−. 0000876
. 90	3. 0771740	3. 0974789	3. 0974182	−. 0000607
. 95	3. 2438210	3. 2630627	3. 2630315	−. 0000312
1. 00	3. 4117444	3. 4300145	3. 4300145	0

Example 2.
Formula: The same as in example 1.
Parameters: $h=0.1=h_1$; $k=0.005$; $\lambda=\frac{1}{2}$.
Range of t: The same as in example 1.
Number of lattice points: $10\times 25=250$.
Initial values and the computed values of $v_{m,n}$ for the last profile are given in table 2.

TABLE 2. *Solution of the difference equation described in table 1 and values of $v(h_1,h_2)$*

Parameters: $h_1=0.1$; $k=0.005$, $\lambda=\frac{1}{2}$.

$v(x,t,h_1,h_2)\equiv v(h_1,h_2)=[v(h_2)-\rho^2 v(h_1)]/(1-\rho^2)$; $\rho=\frac{1}{2}$.

$v(h_2)$ given in table 1.

x	$v(x,0)$	At $t=0.125$				
		$v(x,t)$	$u(x,t)$	$u(x,t)-v(x,t)$	$v(h_1,v_2)$	$u(x,t)-v(h_1,h_2)$
0. 0	0. 8000000	0. 8638878	0. 8746427	+. 0107549	0. 8747574	−. 0001147
. 1	. 8608136	. 9221349	. 9305912	+. 0084563	. 9306382	−. 0000470
. 2	1. 0237187	1. 0784844	1. 0830512	+. 0045668	1. 0830525	−. 0000013
. 3	1. 2517987	1. 2989434	1. 3007690	+. 0018256	1. 3007605	+. 0000085
. 4	1. 5178933	1. 5581507	1. 5585249	+. 0003742	1. 5585224	+. 0000025
. 5	1. 8069311	1. 8415191	1. 8411952	−. 0003239	1. 8411893	+. 0000059
. 6	2. 1109240	2. 1408939	2. 1403270	−. 0005669	2. 1403240	+. 0000030
. 7	2. 4254896	2. 4517066	2. 4511221	−. 0005845	2. 4511167	+. 0000054
. 8	2. 7480902	2. 7711921	2. 7707399	−. 0004522	2. 7701369	+. 0000030
. 9	3. 0771740	3. 0976676	3. 0974182	−. 0002494	3. 0974160	+. 0000022
1. 0	3. 4117444	3. 4300145	3. 4300145	0		0

EXAMPLE 3.

Formula: The same as in example 1.
Parameters: $h=1/14$; $k=1/1176$; $\lambda=\frac{1}{6}$.
Range of t: The same as in example 1.
Number of lattice points: 14×147=2058.
Initial values and the computed values of $v_{m,n}$ for the last profile are given in table 3.

The formula for examples 1 to 3 comes from (4).

TABLE 3. *Solution of the difference equation described in table 1*

Parameters: $h=1/14$, $k=1/1176$; $\lambda=1/6$, $x=mh$.

m	$v(x,0)$	At $t=0.125$		
		$v(x,t)$	$u(x,t)$	$u(x,t)-v(x,t)$
0	0. 8000000	0. 8692554	0. 8746428	+. 0053874
1	. 8314955	. 8987549	. 9035402	+. 0047853
2	. 9203245	. 9824630	. 9858991	+. 0034361
3	1. 0531018	1. 1087660	1. 1108660	+. 0021000
4	1. 2164087	1. 2656385	1. 2667479	+. 0011094
5	1. 4003800	1. 4438567	1. 4443213	+. 0004646
6	1. 5985781	1. 6371338	1. 6372086	+. 0000748
7	1. 8069311	1. 8413398	1. 8411953	−. 0001445
8	2. 0228433	2. 0537601	2. 0535080	−. 0002521
9	2. 2446210	2. 2725815	2. 2722948	−. 0002867
10	2, 4711277	2, 4965649	2, 4962917	−. 0002732
11	2. 7015788	2, 7248430	2, 7246151	−. 0002279
12	2. 9354176	2, 9567941	2, 9566326	−. 0001615
13	3. 1722397	3. 1919624	3. 1918811	−. 0000813
14	3. 4117444	3. 4300145	3. 4300145	0

EXAMPLE 4.

Formula:

$$v_{m,n+1}=v_{m,n}+\lambda\left(\delta^2-\tfrac{1}{12}\delta^4\right)v_{m,n}+kf_{m,n} \qquad (32)$$

$$v_{-m,n}=v_{m,n}+\frac{m^3h^3}{3}\left[\frac{\partial f}{\partial x}\right]_{x=0}=v_{m,n}-\frac{m^3h^3}{v_{0,n}}, \quad m=1,2.$$

It will be convenient to rewrite (32) in the following form:

$$v_{m,n+1}=\sum_{q=-2}^{q=2} a_q v_{m+q,n}+kf_{m,n}, \qquad (33)$$

where

$$a_{-2}=a_2=-\tfrac{1}{12}\lambda; \quad a_{-1}=a_1=\tfrac{4}{3}\lambda; \quad a_0=1-2.5\lambda.$$

Parameters: $h=0.0625$; $k=0.0014$; $\lambda=0.3584$.
Range of t: $0\leq t\leq 0.1288$.
Number of lattice points: $16\times 92=1472$.

Formula (32) is a modification of (5). It is obtained by dropping the terms of (3) involving $p\geq 2$. However, the second derivative with respect to x is approximated by differences, including the fourth order. The resulting formula is almost as accurate as (5) for this problem, since derivatives with respect to t are very small numerically. In view of the fact that (5) was modified, it is necessary to examine (33) to determine the admissible range of λ. It is clear from (33) that all the coefficients a_q cannot be made positive by any choice of λ; hence the stability criteria given earlier do not apply. However, it has been shown in [2] that if there exists a positive number M, independent of x and t, such that the coefficients a_q of (33) satisfy

$$\left|\sum_{p=-2}^{p=2} a_q \exp(iqy)\right| \leq \exp(-My^2), \quad \text{for } |y|\leq\pi, \qquad (34)$$

then the basic homogeneous differential equation,[6] is stable. From that result we may then deduce the stability of (32) or (33).

[6] In [2] the theorem applies to a more general case. Similar results relating to stability are given in [4].

TABLE 4. *Solution of the difference equation*

$$v_{m,n+1}=v_{m,n}+\left(\lambda\delta^2-\frac{\lambda\delta^4}{12}\right)v_{m,n}+kf_{m,n}$$

$$v_{-m,n}=v_{m,n}+\frac{m^3h^3}{3}\left[\frac{\partial f}{\partial x}\right]_{x=0}=v_{m,n}-\frac{m^3h^3}{v_{0,n}}$$

$v_{m,o}=u_{m,o}$; $v_{s,n}=u_{s,n}$; $\delta^4 v_{s-1,n}=\delta^4 v_{s-2,n}$; $sh=1$

$f(x,v)$ defined in table 1.

Parameters: $h_2=0.0625$; $k=0.0014$; $\lambda=0.3584$.

x	$v(x,0)$	At $t=0.1288$		
		$v(x,t)$	$u(x,t)$	$u(x,t)-v(x,t)$
0. 0000	0. 8000000	0. 8767367	0. 8768124	+. 0000757
. 0625	. 8242006	. 8988926	. 8989475	+. 0000545
. 1250	. 8934221	. 9627974	. 9628100	+. 0000125
. 1875	. 9990767	1. 0615954	1. 0615804	−. 0000150
. 2500	1. 1316470	1. 1872134	1. 1871920	−. 0000214
. 3125	1. 2833862	1. 3326403	1. 3326211	−. 0000192
. 3750	1. 4487872	1. 4925905	1. 4925762	−. 0000143
. 4375	1. 6241314	1. 6633207	1. 6633105	−. 0000102
. 5000	1. 8069311	1. 8422337	1. 8422269	−. 0000068
. 5625	1. 9955052	2. 0275254	2. 0275206	−. 0000048
. 6250	2. 1886961	2. 2179277	2. 2179247	−. 0000030
. 6875	2. 3856893	2. 4125343	2. 4125324	−. 0000019
. 7500	2. 5858993	2. 6106854	2. 6106848	−. 0000006
. 8125	2. 7888957	2. 8118924	2. 8118925	+. 0000001
. 8750	2. 9943567	3. 0157864	3. 0157871	+. 0000007
. 9375	3. 2020364	3. 2220847	3. 2220858	+. 0000011
1. 0000	3. 4117444	3. 4305684	3. 4305684	0

EXAMPLE 5.

Formula: The same as in example 4.
Parameters: $h_1=0.125$; $k=0.0056$; $\lambda=0.3584$.
Range of t: The same as in example 4. Results are given in table 5.

TABLE 5. *Solution of the difference equation described in table 4*

Parameters: $h_1=0.125$; $k=0.0056$; $h=0.3584$

$v(x, t, h_1, h_2)=v(h_1, h_2)=[v(h_2)-\rho^4(v(h_1))]/[1-\rho^4]$; $\rho=\frac{1}{2}$

$v(h_2)$ given in table 4.

x	$v(x,0)$	At $t=0.1288$				
		$v(x,t)$	$u(x,t)$	$u(x,t)-v(x,t)$	$v(h_1,h_2)$	$u(x,t)-v(h_1,h_2)$
0	0. 8000000	0. 8753628	0. 8768124	+. 0014496	0. 8768282	−. 0000158
0. 125	. 8934221	. 9623470	. 9628099	+. 0004629	. 9628274	−. 0000175
. 250	1. 1316470	1. 1872934	1. 1871920	−. 0001014	1. 1872081	−. 0000161
. 375	1. 4487872	1. 4926595	1. 4925762	−. 0000833	1. 4925859	−. 0000097
. 500	1. 8069311	1. 8422296	1. 8422269	−. 0000027	1. 8422340	−. 0000071
. 625	2. 1886961	2. 2178929	2. 2179247	+. 0000318	2. 2179300	−. 0000053
. 750	2. 5858993	2. 6106245	2. 6106847	+. 0000602	2. 6106895	−. 0000048
. 875	2. 9943567	3. 0157128	3. 0157869	+. 0000741	3. 0157913	−. 0000044
1. 000	3. 4117444	3. 4305684	3. 4305684	0		0

In all the above examples, $v_{m,n}$, $u_{m,n}$, and $u_{m,n}-v_{m,n}$ were generated. Fourth differences were generated in all the examples, even where the computing formula did not call for them. These fourth differences were jotted down by the operator of the card programmed IBM calculator, and any lack of continuity in the differences was a warning that the machine was not functioning properly. The operations were carried out with a board wired to perform eight-place multiplication. The calculations were carried to the fullest possible accuracy, that is, to seven decimals in $v_{m,n}$ and to eight decimals in some of the subsidiary computations for $(v_{m,n+1}-v_{m,n})$. This accuracy is far in excess of the truncation error, for small values of x.

A "ρ^2" correction was applied to the values of the last available profile, as explained in section 4, based on the entries in examples 1 and 2. The results are given in example 2. Similarly, a ρ^4 correction was applied to the last profile in examples 4 and 5; the results are given in example 5.

5.1. Observations

(a) The ρ^r Correction Process

It is to be noted that in spite of the fact that in the last profile $v_{0,n}$ differs from the true values by 0.0026 in example 1 and by more than 0.01 in example 2, the values of $v(h_1,h_2)$ resulting from the ρ^2 correction are correct to within 0.00011 or better. It can be verified from the tabulated entries that

$$[u-v(h_1)]/[u-v(h_2)]\simeq h_1^2/h_2^2.$$

It follows that to gain an accuracy comparable to that of $v(h_1,h_2)$ without the "ρ^2" correction, an interval of $h=0.01$ would be needed, and hence the amount of work would be 125 times that used in example 1.

The improvement due to the "ρ^4" correction in examples 4 and 5 is not quite so striking. However, even here there is considerable improvement for small values of x, and it must be remembered that the formula used does include an h^2 term in the truncation error, which is not eliminated by the ρ^4 correction, although the error from this term is somewhat lessened. However, for x larger than $\frac{1}{2}$, $v(h_2)$ in example 4 is closer to the true value than $v(h_1,h_2)$. An explanation for this may come from the following considerations: Assume

$$u-v(h_2)=\sum_{p=0}^{\infty}\rho^{p+2}h^{p+2}C_{p+2}. \tag{35}$$

then it follows that

$$u-v(h_1,h_2)=\frac{\rho^2h^2C_2}{1+\rho^2}+\frac{\rho^3(1-\rho)}{1+\rho^4}\,h^3C_3 -\frac{\rho^4(1-\rho)}{1-\rho^4}\,h^5C_5-\ldots. \tag{36}$$

It is to be noticed that the terms involving h^2 and h^3 are numerically smaller in (36) than in (35). However, the term involving h^5 is somewhat larger, if $\rho=\frac{1}{2}$, and all subsequent terms will be larger. In fact, it can be readily verified that if a "ρ^r" correction is applied, then compared with a term involving $\rho^{p+r}h^{p+r}C_{p+r}$ in (35), there is the corresponding term $[-\rho^r(1-\rho^p)/(1-\rho^r)h^{p+r}C_{p+r}$ in $[u-v(h_1,h_2)]$, which is, of course, numerically larger; but usually the leading term after the elimination is smaller than the term eliminated, namely, $h^r\rho^rC_r$. The "ρ^4" approximation, as applied in examples 4 and 5, is rather unusual in that a very small term of order two is left. If all the terms of (35) are small numerically, it may happen that the combination of the leading terms in $|u-v(h_1,h_2)|$ may be somewhat larger than in $|u-v(h_2)|$. In such cases, however, the difference between $v(h_1,h_2)$ and $v(h_2)$ will itself be small; hence, although it may not be known which is the better answer, it is to be expected that the order of magnitude of the error in $v(h_1,h_2)$ is not greater than $|v(h_2)-v(h_1,h_2)|$.

(b) The Rounding Error

The last place of $u(x,t)$ is not guaranteed, hence we can judge the rounding error only to the extent that the sixth decimal place is affected. It is to be noticed that in example 4, which seems to be the most accurate, the difference $u_{m,n}-v_{m,n}$ is systematic as far as sign is concerned. This is evidence of the fact that no large rounding-off error accumulated, after 100 steps in t, even though a nonlinear differential equation was used, and a considerable number of arithmetic operations were performed at each step.

(c) The Effect of Varying λ

Let us compare the error pattern in example 1, where $\lambda=\frac{1}{2}$, with that of example 3, where $\lambda=\frac{1}{6}$. In spite of the fact that somewhat more work was performed in generating values in example 3, the results are not as good—although the error in both examples is in the same decimal place; but the error in example 3 is about twice as large. This, in fact, is precisely what was to be expected. For as observed in section 3, in cases coming under type II, where derivatives with respect to t are relatively small, the error is not appreciably affected by λ; hence, it is almost proportional to h^2. The ratio of the two values of h^2 is $(20/14)^2\cong 2.04$. The numerical illustration verifies the observation that $\lambda=\frac{1}{6}$ is *not* necessarily the best mesh ratio for the formula given in (4). The boundary conditions and the form of $f(x,t,u)$ determine the type that the differential system belongs to, and it is only after the problem is studied from this viewpoint that a suitable choice of λ can be made.

In connection with example 3 corresponding to $\lambda=\frac{1}{6}$, it was not desirable to generate values at double the interval in the x-direction, for purposes of applying a "ρ^2" correction. For if the formula given in (4) were used to compute $v_{-1,1}$, we would get a *negative* value of $(v_{0,1}-v_{0,0})$. Even if the true answer were not known, a knowledge of the more accurate values of $v_{m,n}$ from the smaller value of h would warn us that a negative value of $(v_{0,1}-v_{0,0})$ is incorrect. It will be instructive to write down the differences of $v_{m,0}$ (the known initial profile) at double the interval used in example 3, and to use a value $v_{-1,0}$ computed from formula (4). The values of $v_{m,0}$ and some of the differences are given below:

$h=\frac{1}{7};\quad t=0.$

m	$v_{m,0}$	$\delta^2 v$	$\delta^3 v$	$\delta^4 v$	$\delta^5 v$
−1	0. 9166802				
0	. 8000000	. 2370047	−. 0612450		
1	. 9203245	. 1757597	−. 0896745	−. 0284295	+. 0741146
2	1. 2164087	. 0860852	−. 0439894	+. 0456851	−. 0197723
3	1. 5985781	. 0420958	−. 0180766	+. 0259128	
4	2. 0228433	. 0240192			
5	2. 4711277				

The fourth differences are not numerically smaller than the third differences, and the first entry in the fifth-difference column is very much larger numerically than the second one in the same column. Such a pattern is a warning that the interval is too large.

(d) The Effect of Using a Higher-Order Approximation Formula

Let us compare the results in examples 1 and 4. The coding for example 4 is somewhat more complicated than that for example 1, but since only 1,472 lattice points were used in example 4 compared with 2,000 in example 1, the over-all amount of work is about the same in both cases. The result after 100 steps shows that the higher approximation formula gives very much better results. To secure a maximum error of 0.00008 in $v_{m,n}$ with the approximation used in example 1, it would have been necessary to use an interval h of about 0.009; hence 170 times the amount of work would have been necessary. However, if the ρ^2 correction were applied, the comparative results would not be quite so unfavorable to the simpler approximation. Assuming that a ρ^2 correction were applied to two computations by the simpler approximation, and a ρ^4 correction to results of the modified fourth-order approximation, the latter would still give significantly better results—one additional decimal place, in fact. To secure comparable accuracy by the simpler approximation, it would be necessary to multiply the amount of work by the factor[7] 10. The conclusion is inescapable that the higher approximation is worth while, in cases where the boundary conditions do not introduce singularities in the higher derivatives, and when the coding problem is manageable.

The author acknowledges gratefully the many constructive suggestions which were given by Dr. Fritz John during the progress of the study.

6. References

1] D. R. Hartree and J. R. Womersley, A method for the numerical or mechanical solution of certain types of partial differential equations, Proc. Roy. Soc., London [A] **161,** 353 to 366 (1937).

[2] Fritz John, On integration of parabolic equations by difference methods, Communications on Pure and Applied Math. **5,** 155–211 (1952).

[3] L. M. Milne-Thomson, The calculus of finite differences (Macmillan Co., London, 1933).

[4] George O'Brien, Morton A. Hyman and Sidney Kaplan, A study of the numerical solution of partial differential equation, J. Math. Phys. **29,** No. 4, 223 to 251 (Jan. 1951).

[5] L. F. Richardson and J. Arthur Gaunt, The deferred approach to the limit. Part I: Single lattices, Richardson. Part II: Interpenetrating lattices, Gaunt, Phil. Trans. Roy. Soc., London [A] **226,** 299 to 361 (July 1927).

7. Appendix

DEFINITION. We define $z\ (m,M;n,N)$ to be a region in the x,t-plane, which contains the region

$$(m-M)h \leq x \leq (m+M)h; \quad nk \leq t \leq (n+N)k.$$

7.1. Expressions for Derivatives[8] in Terms of Differences

$$\frac{\partial^2 u}{\partial x^2}=\frac{1}{h^2}\left[\left(\delta^2-\frac{1}{12}\delta^4+\frac{1}{90}\delta^6\right)u_{m,n}\right]-\frac{h^6}{560}\frac{\partial^8 u_{x_1,nk}}{\partial x^8}, \qquad (37)$$

$$\frac{\partial^4 u_{m,n}}{\partial x^4}=\frac{1}{h^4}\left[\left(\delta^4-\frac{1}{6}\delta^6\right)u_{m,n}\right]+h^4\left[\frac{7}{240}\frac{\partial^8 u_{x_2,n}}{\partial x^8}-\frac{h^2}{4200}\frac{\partial^{10} u_{x_3,n}}{\partial x^{10}}\right], \qquad (38)$$

$$\frac{\partial^6 u_{m,n}}{\partial x^6}=\frac{1}{h^6}\delta^6 u_{m,n}-h^2\left[\frac{1}{4}\frac{\partial^8 u_{x_4,n}}{\partial x^8}-\frac{7h^2}{720}\frac{\partial^{10} u_{x_5,n}}{\partial x^{10}}+\frac{h^4\partial^{12} u_{x_6,n}}{18480\partial x^{12}}\right], \qquad (39)$$

where the points (x_j,nk) are in $Z(m,3;n,0)$, $j=1,2,. . .,6$.

If h is sufficiently small, successive terms of (37), (38), and (39) are of a lower order of magnitude than preceding ones, and if terms involving h in the numerator are dropped, then the truncation error is of the order of magnitude of the first term neglected.

7.2. Relations Between Derivatives in the t- and x-Directions

By differentiating (1) we can establish:

$$\frac{\partial^2 u}{\partial t^2}=\frac{\partial^2}{\partial x^2}\left(\frac{\partial u}{\partial t}\right)+\frac{df}{dt}=\frac{\partial^4 u}{\partial x^4}+\frac{d^2 f}{dx^2}+\frac{df}{dt}. \qquad (40)$$

$$\frac{\partial^2 u}{\partial x\partial t}=\frac{\partial^3 u}{\partial x^3}+\frac{df}{dx} \qquad (41)$$

$$\frac{\partial^3 u}{\partial x\partial t^2}=\frac{\partial^5 u}{\partial x^5}+P_{31}; \quad P_{31}=\frac{d^3 f}{dx^3}+\frac{d^2 f}{dxdt}. \qquad (42)$$

$$\frac{\partial^3 u}{\partial t^3}=\frac{\partial^6 u}{\partial x^6}+P_{32}; \quad P_{32}=\frac{d^4 f}{dx^4}+\frac{d^2 f}{dt^2}+\frac{d^3 f}{dtdx^2}. \qquad (43)$$

In the above, P_{31} and P_{32} are functions of x and t. Where not otherwise specified, they will be understood to correspond to $x=mh$, $t=nk$. If $f(x,t,u)$ involves u, the derivatives of f may involve various derivatives of u with respect to x. It will be convenient in this section to change the notation introduced previously, and to define $f_x=\partial f/\partial x$, with t,u held fixed; $f_t=\partial f/\partial t$, with x,u held fixed; $f_u=\partial f/\partial u$, with x,t held fixed with similar definitions for $f_{x,t}, f_{x,u}$, etc. These derivatives will usually be required at $x=0$ and $t=nk$; hence evaluation of the derivatives at $(0,nk)$ will be implied unless otherwise specified, or readily understood from the context of the analysis.

7.3. The Boundary Conditions at $x=0$ for the Examples in Section 5

Since $u_{-m,0}$ has not been defined by the differential system, the expression (47) must be modified when $m=0$, or what is equivalent to it, $u_{-1,n}$ must be defined. Similarly $u_{-1,n}$ and $u_{-2,n}$ must be defined for (5) and (32). Consider the condition $(\partial u_{0,n}/\partial x)=0$. It implies

$$\frac{\partial^p u_{0,n}}{\partial t^{p-1}\partial x}=0, \quad p\geq 2. \qquad (44)$$

Now from (41) and (44)

$$0=\frac{\partial^2 u_{0,n}}{\partial t\partial x}=\frac{\partial^3 u_{0,n}}{\partial x^3}+f_x.$$

Hence

$$\frac{\partial^3 u_{0,n}}{\partial x^3}=-f_x. \qquad (45)$$

Similarly, from (6.23) and (6.30)

$$0=\frac{\partial^3 u_{0,n}}{\partial t^2\partial x}=\frac{\partial^5 u_{0,n}}{\partial x^5}+P_{31},$$

where

$$P_{3,1}=\left[\frac{d^3 f}{dx^3}+\frac{d^2 f}{dxdt}\right]_{0,t}.$$

If f involves u, P_{31} will involve $\partial^2 u/\partial x^2$. An explicit expression for P_{31} in terms of partial derivatives will therefore be useful. It is given below.

$$P_{31}=f_{xxx}+4f_{x,u}\frac{\partial^2 u}{\partial x^2}+f\cdot f_{x,u}-f_x\cdot f_u+f_{x,t}$$
$$=4f_{x,u}\frac{\partial^2 u_{0.n}}{\partial x^2}+G_n=\frac{-\partial^5 u_{0,n}}{\partial x^5}, \qquad (46)$$

[7] This is on the assumption that after the ρ^2 correction, the leading term of $\sigma_{m,n}$ has the factor h^3. Since the work is inversely proportional to h^3 for the same λ, improving the entries by a factor of M means doing M times the amount of work.

[8] See chapter VII, reference [3].

where

$$G_n = f_{xxx} + f\cdot f_{x,u} - f_x\cdot f_u + f_{x,t}. \qquad (47)$$

Using (45) the Maclaurin series around (0, nk) yields

$$u_{m,n} = u_{0,n} + \frac{m^2h^2}{2}\frac{\partial^2 u_{0,n}}{\partial x^2} - \frac{m^3h^3}{6} f_x + \sum_{p=4}^{\infty} \frac{m^p h^p}{p!}\frac{\partial^p u_{0,n}}{\partial x^p}. \qquad (48)$$

If we set $m=1$, 2 and use (46), we can obtain $\partial^{2p}u_{0,n}/\partial x^{2p}$ to an accuracy comparable to approximations used over the rest of the region. The same results are obtained if we use (48), to solve for $u_{-1,n}$ and $u_{-2,n}$ (if the latter is needed). This artificial extension of the region to negative values of m is convenient for numerical treatment, and is justifiable wherever (48) exists and the required derivatives are bounded. Thus we can write

$$u_{-m,n} = u_{m,n} + \frac{m^3h^3}{3} f_x - \frac{m^5h^5}{60}\frac{\partial^5 u_{0,n}}{\partial x^5} - \ldots . \qquad (49)$$

If h is sufficiently small, and terms involving powers of $h^{5+p}, p\geq 0$ are neglected in (49), the truncation error for $u_{-m,n}$ is of order h^5. Thus $T_{2,0,n}$, defined in (7a), has a term in h^3.

The third term on the right-hand side of (49) was also dropped in example 4, since its magnitude would have affected only the fifth decimal place at any point. This term could have been obtained, if desired, by using (46) and (37), and then setting $m=1$ and $m=2$ in (49), to obtain three linear equations for the unknowns u_{-1}, u_{-2}, and $h^5\partial^5 u/\partial x^5$.

7.4. The Boundary Condition at the Terminal Points (x_a, t)

The difference equation defined in (4) needs no special treatment for the boundary conditions at x_a, where $a=sm$, since $(s-1,n)$ is the last lattice point at which $v_{m,n}$ is generated, and $\delta^2 v_{s-1,n}$ is fully defined. However, when the fourth order approximation is used, as in (5), an expression is required for $v_{s+1,n}$, or for $\delta^2 u_{s,n}$. Since $u(x_a,t)=B(t)$ and all its derivatives in the t-direction are assumed to be known, the differential eq (1) can be used to obtain the second partial derivative in the x-direction, in terms of $\partial u/\partial t$ and $f(x,u)$. The relations between derivatives and differences can then be used to obtain $\delta^2 u_{s,n}$. In examples 4 and 5, the fourth difference in the x-direction was essentially zero at x_a; hence the last fourth difference was replaced by $\delta^4 u_{s-1,n}$.

Los Angeles, California, September 18, 1951.