High accuracy difference formulae for the numerical solution of the heat conduction equation

By A. R. Mitchell and R. P. Pearce

A method is given for generating systematically difference replacements of the heat conduction equation of any desired order of truncation error. The method is then used to obtain explicit and implicit formulae of high accuracy.

1. Introduction

In recent years many finite-difference replacements have been proposed for solving the heat conduction equation

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} \tag{1}$$

where x, t are the distance and time co-ordinates respectively. Most of these, together with conditions for their stability, can be found in works such as Richtmyer (1957), Forsythe and Wasow (1960), Collatz (1960), and Saul'ev (1962).

In the main, the formulae in common use tend to be simple formulae with stability conditions which permit relatively large time steps. There are, however, many problems involving equations of the heat conduction type where very high accuracy is required over a small range of the time co-ordinate. Thus formulae of high accuracy are required which need only be stable for small values of the mesh ratio. In view of the high-speed computing facilities now available these formulae can also, of course, be used to give high accuracy results over any range of the time co-ordinate.

2. Method of deriving High Accuracy Formulae

A formula of any desired truncation error can be obtained at the node (r, s) in Fig. 1, depending on how many additional nodes one is prepared to consider. Once the nodes have been decided upon, the value of u at each node is obtained as a Taylor expansion in terms of u and its derivatives at the point (r, s). All the derivatives with respect to t can be replaced by higher derivatives with respect to x, if the relations $\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2}$, $\frac{\partial^2 u}{\partial t^2} = \frac{\partial^4 u}{\partial x^4}$, etc., from (1) are used. In addition, if h and k are the mesh lengths in the x and t directions respectively, k can be eliminated by introducing the ratio p through the relationship $k = ph^2$. For example, the values of u at nodes in the vicinity of (r, s) are given by

$$u_{r,s\pm1} = u \pm pB + \frac{1}{2}p^2D \pm \frac{1}{6}p^3F + \frac{1}{24}p^4H \pm \frac{1}{120}p^5J + \dots$$

$$u_{r+1,s} + u_{r-1,s} = 2u + B + \frac{1}{12}D + \frac{1}{360}F + \frac{1}{20{,}160}H + \frac{1}{1{,}814{,}400}J + \dots$$

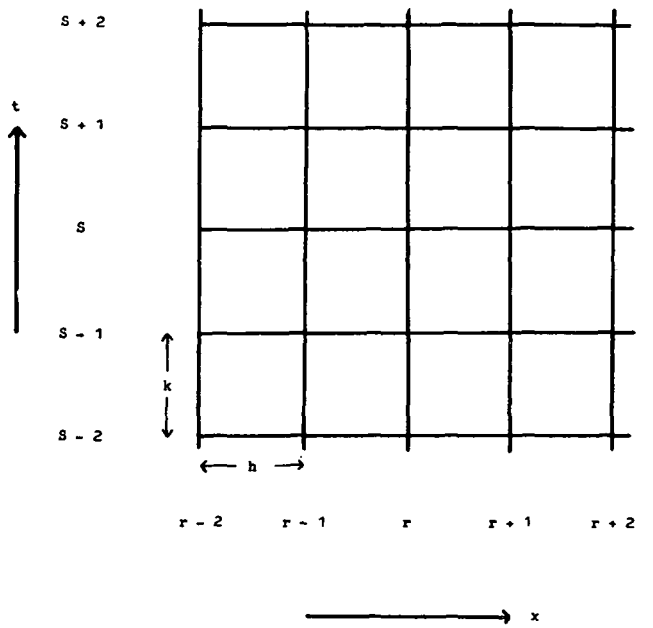

Fig. 1

$$u_{r+2,s} + u_{r-2,s} = 2u + 4B + \frac{4}{3}D + \frac{8}{45}F + \frac{4}{315}H + \frac{8}{14{,}175}J + \dots$$

$$u_{r+1,s\pm1} + u_{r-1,s\pm1} = 2u + (1 \pm 2p)B + \left(\frac{1}{12} \pm p + p^2\right)D + \left(\frac{1}{360} \pm \frac{1}{12}p + \frac{1}{2}p^2 \pm \frac{1}{3}p^3\right)F + \left(\frac{1}{20{,}160} \pm \frac{1}{360}p + \frac{1}{24}p^2 \pm \frac{1}{6}p^3 + \frac{1}{12}p^4\right)H + \left(\frac{1}{1{,}814{,}400} \pm \frac{1}{20{,}160}p + \frac{1}{720}p^2 \pm \frac{1}{72}p^3 + \frac{1}{24}p^4 \pm \frac{1}{60}p^5\right)J + \dots \tag{2}$$

where $u, B, D, \dots J, \dots$ are the values of

$$u, h^2\frac{\partial^2 u}{\partial x^2}, h^4\frac{\partial^4 u}{\partial x^4}, \dots h^{10}\frac{\partial^{10} u}{\partial x^{10}}, \dots$$

at the node (r, s). Finally, a linear relation connecting

the values of u at the nodes considered is chosen in order to eliminate the maximum possible number of $B, D, \ldots J, \ldots$ The coefficients in this linear relation depend on p only. The formula obtained has the smallest possible truncation error for the original choice of points. In the subsequent work, the truncation error quoted for each formula is based on the formula written with $u_{r,s+1}$ only on the left-hand side of the equation.

3. Explicit Formulae

Some explicit formulae in common use are optimum formulae in the sense of the previous section. A typical example is the four-point formula

$$u_{r,s+1} = (1 - 2p)u_{r;s} + p(u_{r-1,s} + u_{r+1,s}), \quad (3)$$

with a truncation error $\frac{1}{2}p(p - \frac{1}{6})D$. If $p = \frac{1}{6}$, the truncation error reduces to $\frac{1}{3,240}F$. This formula is stable if $p \leqslant \frac{1}{2}$. Recently Herman and Radok (1960) constructed a family of optimum two-level formulae, giving $u_{r,s+1}$ in terms of u at nodes on column s. The first member of the family is formula (3). The second member is

$$u_{r,s+1} = \tfrac{1}{2}(6p^2 - 5p + 2)u_{r,s} + \tfrac{2}{3}p(2 - 3p)(u_{r+1,s} + u_{r-1,s}) - \tfrac{1}{12}p(1 - 6p)(u_{r+2,s} + u_{r-2,s}), \quad (4)$$

which has a principal truncation error

$$\tfrac{1}{180}p\,(30p^2 - 15p + 2)F,$$

and is stable for $p \leqslant \frac{2}{3}$. The third member introduces the values of u at the additional points $(r \pm 3, s)$, and so on.

Radok's formulae can be improved from the point of view of truncation error by incorporating the node $(r, s - 1)$. Formula (3) is then replaced by

$$(1 + 6p)u_{r,s+1} = 2(1 - 12p^2)u_{r,s} + 12p^2(u_{r+1,s} + u_{r-1,s}) - (1 - 6p)u_{r,s-1}. \quad (5)$$

This formula, which has a truncation error $\frac{2p^2}{1 + 6p}\left(p^2 - \frac{1}{60}\right)F$, and is stable for $p \leqslant \frac{1}{2\sqrt{3}}$, is given by Saul'ev (1962). If $p = \frac{1}{2\sqrt{15}}$, the truncation error is $\frac{5 - \sqrt{15}}{151,200}H$. Similarly, in place of (4) we have

$$(2 + 15p + 30p^2)u_{r,s+1} = (4 - 3p^2 + 180p^4)u_{r,s} + 8p^2(4 - 15p^2)(u_{r+1,s} + u_{r-1,s}) - p^2(\tfrac{1}{2} - 30p^2)(u_{r+2,s} + u_{r-2,s}) - (2 - 15p + 30p^2)u_{r,s-1}, \quad (6)$$

a formula with a truncation error

$$\frac{p^2\left(5p^4 - \frac{5}{12}p^2 + \frac{1}{105}\right)}{2(30p^2 + 15p + 2)}H,$$

which is stable if $p \leqslant \frac{1}{2\sqrt{15}}$.

Other explicit schemes were considered which made use of more complicated arrangements of nodes, but it was felt that the complexity of the resulting formulae did not justify the gain in truncation error.

4. Implicit Formulae

Most implicit formulae are two-level formulae connecting values of u at nodes on levels s and $s + 1$. The six-point implicit scheme in common use is the Crank–Nicolson scheme

$$2(p + 1)u_{r,s+1} - p(u_{r+1,s+1} + u_{r-1,s+1}) = 2(1 - p)u_{r,s} + p(u_{r+1,s} + u_{r-1,s}). \quad (7a)$$

This scheme is stable for all values of p, but is not an optimum scheme in the sense of Section 2. The six-point implicit scheme with minimum truncation error is in fact

$$(5 + 6p)u_{r,s+1} + (\tfrac{1}{2} - 3p)(u_{r+1,s+1} + u_{r-1,s+1}) = (5 - 6p)u_{r,s} + (\tfrac{1}{2} + 3p)(u_{r+1,s} + u_{r-1,s}). \quad (7b)$$

This scheme is stable for all p and has a truncation error $-\frac{1}{40}\frac{p(20p^2 - 1)}{5 + 6p}F$. If $p = \frac{1}{2\sqrt{5}}$, the truncation error reduces to $-\frac{5\sqrt{5} - 3}{233,856}H$. If three additional nodes on level $s - 1$ are considered, the highly accurate nine-point formula

$$au_{r,s+1} + b(u_{r+1,s+1} + u_{r-1,s+1}) = cu_{r,s} + d(u_{r+1,s} + u_{r-1,s}) + eu_{r,s-1} + f(u_{r+1,s-1} + u_{r-1,s-1}) \quad (8)$$

is obtained using the method of Section 2, where

$$a, e = \pm 4p^4 + 5p^3 \mp \frac{1}{10}p^2 - \frac{23}{84}p \mp \frac{313}{12,600}$$

$$b, f = \mp 2p^4 + \frac{1}{2}p^3 \pm \frac{1}{20}p^2 - \frac{11}{840}p \pm \frac{13}{25,200}$$

$$c = -16p^4 + p^2 - \frac{313}{6,300}$$

$$d = 8p^4 - \frac{1}{2}p^2 + \frac{13}{12,600}.$$

This formula is stable for $p \leqslant \frac{1}{2\sqrt{5}}$, and has a truncation error

$$\frac{p^2\left(\frac{1}{15}p^6 - \frac{3}{400}p^4 + \frac{1}{7,200}p^2 - \frac{59}{127,008,000}\right)}{4p^4 + 5p^3 - \frac{1}{10}p^2 - \frac{23}{84}p - \frac{313}{12,600}}J.$$

In fact, if $p = 0{\cdot}13384$, the leading term in the truncation error contains $h^{12}\frac{\partial^{12}u}{\partial x^{12}}$.

5. Sub-optimum Formulae

The method described in Section 2 can be modified to derive formulae with larger truncation errors, but with less stringent stability conditions than the optimum formulae obtained in Sections 3 and 4.

For example, consider the node (r, s) together with the four surrounding nodes $(r+1, s)$, $(r-1, s)$, $(r, s+1)$, $(r, s-1)$. Using the appropriate expansions in (2), the linear relation

$$\begin{aligned}(1+2p)u_{r,s+1} &= (1+2p)bu_{r,s} \\ &\quad + p(2-b)(u_{r+1,s}+u_{r-1,s}) \\ &\quad + (1-2p-b)u_{r,s-1} \end{aligned} \tag{9}$$

where b is an arbitrary parameter, is obtained. This relation eliminates B and has a truncation error $\dfrac{p[(1+6p)b-2(1-12p^2)]}{12(1+2p)}D$. If $b=\dfrac{2(1-12p^2)}{1+6p}$, formula (5), which has the minimum possible truncation error for the above five points, is obtained. If, however, $b=0$, equation (9) becomes

$$(1+2p)u_{r,s+1} = 2p(u_{r+1,s}+u_{r-1,s}) + (1-2p)u_{r,s-1} \tag{10}$$

This is the Dufort and Frankel scheme which has a truncation error $\dfrac{p(12p^2-1)}{6(1+2p)}D$ and is stable for all p. This is a distinct improvement over the stability range of $p \leqslant \dfrac{1}{2\sqrt{3}}$ for the optimum formula (5).

It is worth pointing out, however, the considerable loss in accuracy resulting from the use of a sub-optimum formula like (10) with a less stringent stability requirement. For example, if $h = 0{\cdot}1$, the truncation error is $\dfrac{11}{48} \times 10^{-8}\dfrac{\partial^6 u}{\partial x^6}$ for the optimum formula (5) with $p=\dfrac{1}{4}$, and $\dfrac{47}{150} \times 10^{-3}\dfrac{\partial^4 u}{\partial x^4}$ for the sub-optimum formula (10) with $p=2$. There is no doubt that if a premium is placed on accuracy, it is much preferable to use formula (5), even although eight times the number of steps is required to reach a given time.

6. Theoretical Solutions of Difference Equations

In comparing the accuracies of the various difference formulae, both explicit and implicit, theoretical rather than numerical solutions of the difference equations are considered. This eliminates consideration of numerical errors, which although of great importance, vary considerably, depending on the method used to solve the particular difference equation.

For purposes of comparison, the problem considered is the solution of (1) together with the boundary conditions

$$u = \sin x \ (0 \leqslant x \leqslant \pi) \text{ at } t = 0,$$
$$u = 0 \text{ at } x = 0, \pi \text{ for } t \geqslant 0.$$

The theoretical solution is

$$u = e^{-t}\sin x, \tag{11}$$

and this is used to test the accuracy of the various difference schemes. In all the calculations using difference equations, $h = \pi/20$, and a comparison with (11) is made when $x = \pi/2$ $(r = 10)$.

The most accurate formula in common use at present is the six-point implicit scheme (7*b*). The solution of (7*b*) for the present problem is

$$u_{r,s} = \lambda^s \sin rh, \tag{12}$$

where $$\lambda = \frac{1-2\left(p+\frac{1}{6}\right)\sin^2\frac{h}{2}}{1+2\left(p-\frac{1}{6}\right)\sin^2\frac{h}{2}}.$$

Calculations are carried out for $p=\dfrac{1}{2\sqrt{5}}$, the value of p which makes the truncation error as small as possible.

The nine-point implicit scheme (8) proposed in the present paper has the solution

$$u_{r,s} = \left[\frac{\lambda_1^s-\lambda_2^s}{\lambda_1-\lambda_2}\psi(k) - \frac{\lambda_1^{s-1}-\lambda_2^{s-1}}{\lambda_1-\lambda_2}\lambda_1\lambda_2\right]\sin rh, \tag{13}$$

where $$\lambda_1, \lambda_2 = \frac{B \pm \sqrt{(B^2+4AC)}}{2A},$$

and

$$A = 2b\cos h + a$$
$$B = 2d\cos h + c$$
$$C = 2f\cos h + e.$$

The function $\psi(k)$ depends on the values of u used at nodes on $t = k$. In the present paper, where a difference equation of higher order than the differential equation is used, the additional boundary values required in order to solve the difference equation are taken from (11), the theoretical solution of the differential equation, and so in this case $\psi(k) = e^{-k}$. In a more general application, of course, the theoretical solution of the differential equation is not likely to be known, and the values of u at nodes on $t = k$ must be obtained by an independent procedure.

It should be pointed out that in (13), λ_2 is the approximation to the fundamental root which is present in any difference replacement of the problem, whereas λ_1 is the extra root introduced because the nine-point formula is an order higher in t than the original differential equation. It is interesting to calculate λ_1 for various values of p and h. For all h, $\lambda_1 = 1$ at $p = 0$, $\dfrac{1}{2\sqrt{5}}$. If $p > \dfrac{1}{2\sqrt{5}}$, $\lambda_1 > 1$ for all h, which is of course to be expected, since (8) is stable only if $p \leqslant \dfrac{1}{2\sqrt{5}}$.

In order to evaluate (13) for $h = \pi/20$ at $r = 10$ it is convenient to put $\psi(k) = \lambda_2 + \delta$, and so

$$u_{10,s} = \lambda_2^s + \frac{\lambda_1^s-\lambda_2^s}{\lambda_1-\lambda_2}\delta.$$

Table 1

NUMBER OF TIME STEPS		SOLUTION OF DIFFERENTIAL EQUATION	FORMULA (8) *9 POINT*	FORMULA (6) *7 POINT*	FORMULA (7*b*) *6 POINT*	FORMULA (5) *5 POINT*	FORMULA (3) *4 POINT*	FORMULA (7*a*) *CRANK–NICOLSON*
$p = \frac{1}{\sqrt{20}}$	$p = \frac{1}{2\sqrt{20}}$		$p = \frac{1}{2\sqrt{20}}$	$p = \frac{1}{2\sqrt{20}}$	$p = \frac{1}{\sqrt{20}}$	$p = \frac{1}{2\sqrt{20}}$	$p = \frac{1}{\sqrt{20}}$	$p = \frac{1}{\sqrt{20}}$
1	2	994,497,915,630	0	−7	−26	−916	−4,000,000	11,000,000
2	4	989,026,104,192	0	−16	−51	−3,138	−8,000,000	22,000,000
4	8	978,172,634,773	0	−33	−101	−7,614	−15,000,000	40,000,000
8	16	956,821,703,419	0	−67	−198	−16,274	−30,000,000	79,000,000
16	32	915,507,772,134	0	−125	−379	−32,464	−58,000,000	151,000,000
80	160	643,146,895,793	−1	−445	−1,331	−117,739	−200,000,000	531,000,000
160	20	413,637,929,568	−1	−573	−1,712	−152,043	−257,000,000	683,000,000
320	640	171,096,336,778	−2	−476	−1,417	−126,028	−212,000,000	565,000,000
640	1,280	029,273,956,459	−1	−163	−485	−43,169	−73,000,000	194,000,000
800	1,600	012,108,818,740	0	−82	−251	−22,325	−37,000,000	100,000,000
Stability			$p \leqslant \frac{1}{2\sqrt{5}}$	$p \leqslant \frac{1}{2\sqrt{15}}$	All p.	$p \leqslant \frac{1}{2\sqrt{3}}$	$p \leqslant \frac{1}{2}$	All p.
Truncation error			$f(p)J$	$g(p)H$	$-\frac{1}{40}\frac{p(20p^2-1)}{5+6p}F$ $\left(p \neq \frac{1}{2\sqrt{5}}\right)$; $-\frac{5\sqrt{5}-3}{233,856}H$ $\left(p = \frac{1}{2\sqrt{5}}\right)$.	$\frac{2p^2}{1+6p}\left(p^2-\frac{1}{60}\right)F$ $\left(p \neq \frac{1}{2\sqrt{15}}\right)$; $\frac{5-\sqrt{15}}{151,200}H$ $\left(p = \frac{1}{2\sqrt{15}}\right)$.	$\frac{1}{2}p\left(p-\frac{1}{6}\right)D$ $\left(p \neq \frac{1}{6}\right)$; $\frac{1}{3,240}F\left(p = \frac{1}{6}\right)$.	$\frac{1}{6}pD$

All results are multiplied by 10^{12}. In above, $f(p) = \dfrac{p^2\left(\frac{1}{15}p^6 - \frac{3}{400}p^4 + \frac{1}{7,200}p^2 - \frac{59}{127,008,000}\right)}{4p^4 + 5p^3 - \frac{1}{10}p^2 - \frac{23}{84}p - \frac{313}{12,600}}$, $g(p) = \dfrac{p^2\left(5p^4 - \frac{5}{12}p^2 + \frac{1}{105}\right)}{2(30p^2 + 15p + 2)}$.

In order to compare results directly with the six-point implicit formula, solutions are carried out for $p = \frac{1}{4\sqrt{5}}$. (The nine-point formula is only marginally stable for $p = \frac{1}{2\sqrt{5}}$). As this value of p does not coincide with the value of p for which the truncation error of the nine-point formula is a minimum, it seems likely that greater accuracy can be obtained with the nine-point formula than is demonstrated in the present example.

The explicit scheme which is most likely to compare with the nine-point implicit scheme from the point of view of accuracy is formula (6). This has solution (13), where this time λ_1, λ_2 are the roots of the equation

$$(2 + 15p + 30p^2)\lambda^2 - [(4 - 3p^2 + 180p^4) + 16p^4(4 - 15p^2)\cosh - p^2(1 - 60p^2)\cos 2h]\lambda + (2 - 15p + 30p^2) = 0.$$

A solution is obtained for $p = \frac{1}{4\sqrt{5}}$, by a method similar to that used to solve the nine-point implicit formula. Once again λ_2 is the approximation to the fundamental root and λ_1 is the extra root introduced because the difference formula is an order higher in t than the original differential equation.

A less complicated explicit formula is given by (5). This also has a solution of the form (13), where this time λ_1, λ_2 are the roots of the equation

$$(1 + 6p)\lambda^2 - [2 - 24p^2(1 - \cos h)]\lambda + (1 - 6p) = 0.$$

A solution is obtained for $p = \frac{1}{2\sqrt{20}}$.

7. Numerical Results

The numerical values obtained by solving the difference equations discussed in the present paper are shown in Table 1, together with a summary of the stability requirements and truncation errors of the formulae. As stated previously, the calculations are carried out for $h = \pi/20$, and proceed as far as $t = \pi^2/\sqrt{5}$. This is equivalent to 800 time steps when $p = \frac{1}{\sqrt{20}}$, and to 1,600 time steps when $p = \frac{1}{2\sqrt{20}}$. The results are accurate to twelve places of decimals, and are presented in the form of the solution of the differential equation, together with the discrepancies between the theoretical solutions of the difference equations and the solution of the differential equation. In the case of the four-point explicit scheme and the Crank–Nicolson formula, it was sufficient to work correct to six places of decimals, and so the zeros in decimal places seven to twelve have no significance. In the case of the nine-point implicit formula, fifteen places of decimals were retained, and the discrepancy between the theoretical solution and the theoretical solution of the differential equation was only one in the fourteenth place after one time step. In each calculation, the maximum discrepancy occurred in the region of the 160th time step if $p = \frac{1}{\sqrt{20}}$, and the 320th step if $p = \frac{1}{2\sqrt{20}}$.

References

COLLATZ, L. (1960). *The Numerical Treatment of Differential Equations.* Berlin: English Translation, Springer.

FORSYTHE, G. E., and WASOW, W. R. (1960). *Finite Difference Methods for Partial Differential Equations*, New York: Wiley.

HERMAN, R., and RADOK, J. R. M. (1960). "High Order Correct Difference Schemes for Anisotropic Parabolic Equations," Polytechnic Institute of Brooklyn Report No. 581.

RICHTMYER, R. D. (1957). *Difference Methods for Initial-Value Problems*, New York: Interscience.

SAUL'EV, V. K. (1962). *The Integration of Parabolic Equations by the Method of Nets*, London: English Translation, Pergamon (in press).