IMPLICIT DIFFERENCE APPROXIMATIONS

2579, 2819, 2879, 2903, 2963, 2999, 3023, 3119, 3167, 3203, 3467, 3623, 3799, 3803, 3863, 3947, 4007, and 4079.

The Ohio State University,
Columbus, Ohio


On the Accuracy of Implicit Difference Approximations to the Equation of Heat Flow

By Wolfgang Wasow

1. Introduction. Implicit finite difference methods are gaining favor in the numerical solution of initial value problems for partial differential equations, because their computational stability is comparatively insensitive to the mesh ratios. The difference equation problems of importance in the applications are mostly too complicated to permit a theoretical analysis of their approximating qualities. It is a reasonable working hypothesis that the properties of the simplest linear difference equations with constant coefficients are indicative of the situation to be expected in many more involved problems.

The present paper is devoted to a study of the convergence, the stability and the truncation error of implicit difference approximations to the initial value problem defined by the differential equation

(1.1) $L[u] = u_t - u_{xx} = 0, \quad 0 < x < \pi, \quad 0 < t \leq T,$

where the subscripts indicate partial differentiations, and the subsidiary conditions

(1.2) $u(x, 0) = f(x), \quad u(0, t) = u(\pi, t) = 0,$

in which $f(x)$ is a given function. The approximating difference equation to be considered is

(1.3) $L_h,k[u] = k^{-1} \{ U(x, t + k) - U(x, t) \} - sh^{-2} \{ U(x + h, t + k) - 2U(x, t + k) + U(x - h, t + k) \} - (1 - s)h^{-2} \{ U(x + h, t) - 2U(x, t) + U(x - h, t) \} = 0, \quad 0 < x < \pi, \quad 0 < t \leq T.$

Received 1 May 1957. Sponsored by the United States Army under Contract No. DA-11-022-ORD-2059.
The function $U(x, t)$ must also satisfy the same subsidiary conditions as $u(x, t)$, i.e.,

$$U(x, 0) = f(x), \quad U(0, t) = U(\pi, t) = 0.$$  

(1.4)

It will be assumed throughout that $\pi/h$ is an integer. Furthermore, $h$ and $k$ will be restricted to intervals $0 < h < h_1 < 1$, $0 < k < k_1 < 1$.

The cases $s = 0$, $s = 1$ have been extensively studied in the literature (See, e.g., [1], [2], [3]). A paper in process of publication by M. L. Juncosa and D. Young [4] contains an appraisal of the truncation error for $s = \frac{1}{2}$. This appraisal is less sharp than the one contained in theorem 2 of the present paper. P. Lax and R. Richtmyer [5] have applied their general theory of finite difference approximations for initial value problems to an equation slightly more general than (1.3). Their arguments are based on the norm $\left\{ \int_0^\pi \phi^2(x, t) \, dx \right\}^{\frac{1}{2}}$. The elegant methods by means of which they study the convergence and stability can doubtless be extended to yield appraisals of the truncation error as well, but since uniform or pointwise appraisals are more satisfactory to the numerical analyst than bounds on the mean square norm, it is hoped that the results of the present paper are of some interest.

2. The Convergence of $U(x, t)$. If $f(x)$ is integrable and $\sum_{r=1}^\infty b_r \sin rx$ is its Fourier sine-series the function

$$u(x, t) = \sum_{r=1}^\infty \exp \left( -r^2 t \right) b_r \sin rx$$

(2.1)

is, for $t > 0$, a solution of the differential equation (1.1), as can be readily verified. Under rather mild additional conditions on $f(x)$ this function fulfills also the boundary conditions (1.2), at least in some weak sense [3]. Here, the stronger hypothesis

$$\sum_{r=1}^\infty | b_r | < \infty$$

(2.2)

will be made, which is, for instance, satisfied if

$$f(0) = f(\pi) = 0,$$

(2.3)

and $f(x)$ is continuous and has a piecewise continuous derivative.

The method of separation of variables leads to the formal trigonometric series

$$U(x, t) \sim \sum_{r=1}^\infty E_r E_k b_r \sin rx$$

(2.4)

for the solution of (1.3) and (1.4). Here $E_r$ is defined by

$$E_r = \frac{h^2 - 4(1 - s)k \sin^2 (rh/2)}{h^2 + 4sk \sin^2 (rh/2)}.$$  

(2.5)

The derivation of (2.4) is straightforward and will be omitted.

If

$$2(1 - 2s)k - h^2 \leq 0,$$

(2.6)
and only then, it follows that

\[(2.7) \quad |E_r| \leq 1,\]

for all r. In this case the series in (2.4) is uniformly convergent, thanks to condition (2.2), and represents therefore the solution of the difference problem.

In order to prove the convergence of \(U(x, t)\) to \(u(x, t)\) it will now be shown that

\[(2.8) \quad \lim_{h, k \to 0} \frac{E_r^h}{k} = \exp \left(-r^2t\right).\]

If \(k\) is so small that \(|kr^2(1 - s)| < 1\) and \(|kr^2s| < 1\), it follows from part a) of the Lemma in section 5 that

\[|\log E_r + A_h r^2 k \sin^2 \left(\frac{rh}{2}\right)| \leq c(4h^{-2}k \sin^2 (rh/2))^2, \quad c \text{ a constant}.\]

The right member is a bounded function of \(h\), and it can therefore be concluded from this inequality that

\[\lim_{k \to 0} \left(t/k \right) \left(\log E_r + 4h^{-2}k \sin^2 \left(\frac{rh}{2}\right)\right) = 0,\]

uniformly in \(h\). Hence,

\[(2.9) \quad \lim_{h, k \to 0} \frac{E_r^h}{k} = \exp \left\{-4h^{-2}t \sin^2 \left(\frac{rh}{2}\right)\right\},\]

and (2.8) is now obtained by letting \(h\) tend to zero.

Since, under the hypotheses made, the series (2.4) converges uniformly, for all small \(h\) and \(k\), the passage to the limit indicated by (2.8) can be performed termwise and leads to the series (2.1). Thus the following theorem has been proved.

**Theorem 1.** If the Fourier sine coefficients \(b_r\) of \(f(x)\) satisfy the condition

\[\sum_{r=1}^{\infty} |b_r| < \infty,\]

and if \(h\) and \(k\) tend to zero in such a way that \(2(1 - 2s)k - h^2 \leq 0\) then

\[\lim_{h, k \to 0} U(x, t) = u(x, t),\]

where \(U\) and \(u\) solve the problems (1.3), (1.4), and (1.1), (1.2), respectively.

For \(s < \frac{1}{2}\), the inequality (2.6) restricts the size of the ratio \(k/h^2\). For \(s \geq \frac{1}{2}\) theorem 1 proves the convergence of \(U\) to \(u\) for any form of the passage to the limit \(h \to 0, k \to 0\). This goes beyond the frequently made statement that \(k/h^2\) may have any constant value in this case. Observe that \(s\) is not restricted to the interval \(0 \leq s \leq 1\).

**Corollary.** Denote by \(u_h(x, t)\) the solution of the difference-differential equation problem obtained from (1.3) by substituting \(U\) for \(U\) for \(k^{-1}(U(x, t + k) - U(x, t))\) and by imposing the subsidiary conditions (1.4). Then

\[\lim_{k \to 0} U(x, t) = u_h(x, t),\]

provided (2.2) is true.

To prove this corollary it suffices to refer to (2.9) and to show that the uniformly convergent series obtained from (2.4) by substituting \(\exp \left\{-4h^{-2}t \sin^2 \left(\frac{rh}{2}\right)\right\}\) for
$E_r$ satisfies the difference-differential equation and the boundary conditions (1.4). This can be done by simple verification. The condition (2.6) is, of course, true for all sufficiently small $k$, if $h$ is kept fixed.

It is easily seen that an analogous corollary could be stated concerning the difference-differential equation obtained if the difference quotients with respect to $x$ in (1.3) are replaced by the corresponding derivatives. Here, however, the condition $s \geq \frac{1}{2}$ has to be imposed in order to maintain the inequality (2.6). The solution of this problem is the series obtained from (2.4) by substituting

$$\frac{1 - (1 - s)kr^2}{1 + skr^2}$$

for $E_r$.

Such difference-differential equations are sometimes used on analog machines. The arguments of this section justify these procedures mathematically.

3. The Truncation Error. In this section the order of magnitude of the truncation error $U - u$ will be determined (see theorem 2 below). The condition (2.6) is known to be necessary at least for $s = 0$, if convergence for all reasonably regular initial functions $f(x)$ is desired. Very likely it is necessary for all $0 < s < \frac{1}{2}$. Various heuristic arguments lead to the conjecture that $s = \frac{1}{2}$ is the most favorable choice as far as the order of magnitude of the truncation error is concerned. For instance, if $L_h[u]$ in (1.3) is developed in powers of $h$ and $k$, where $u$ is a solution of (1.1), the resulting series begins with terms of order $0(k + h^2)$, except if $s = \frac{1}{2}$, in which case the leading terms are $0(k^2 + h^2)$. The results of this section will confirm the expected exceptional role of the value $s = \frac{1}{2}$. Since the purpose of implicit methods is to gain more freedom in the choice of $h$ and $k$, the range $0 < s < \frac{1}{2}$, in which the condition (2.6) is necessary, is of little interest. Throughout this section the inequalities

$$(3.1) \quad \frac{1}{2} \leq s \leq 1$$

will therefore be imposed.

In view of theorem 1 the convergence of $U$ to $u$ takes place no matter how $h$ and $k$ tend to zero, if (3.1) is true. However, the order of magnitude of $U - u$ is very likely affected if $h$ is exponentially small with respect to $k$. Here the inequality

$$(3.2) \quad h/k \geq \rho > 0, \quad \rho \text{ a constant}$$

will be assumed to be true. Without too much complication the arguments apply also to the more general case that $h^\beta/k \geq \rho > 0, \beta > 0$. The restriction to (3.2) is motivated by the desire to avoid involved formulas. The condition (3.2) is considerably weaker than the hypothesis $k/h^2 = \text{const}$ frequently met in this connection. It is also milder than the assumption $k = 0(h/\log h)$ made in [4].

Finally, the conditions on $f(x)$ must be tightened in this section. It will be assumed that

$$(3.3) \quad |b_r| \leq Kr^{-1}.$$  

This is, for instance, the case if (2.3) is true and $f(x)$ possesses an integrable third
derivative in $0 \leq x \leq \pi$, for three integrations by parts show that then

$$b_r = \frac{2}{\pi} \int_0^\pi f(x) \sin rx \, dx = \frac{2}{\pi} r^{-3} [-(-1)^{j} f^\prime\prime(\pi) - f^\prime\prime(0)]$$

$$- \frac{2}{\pi} r^{-3} \int_0^\pi f^\prime\prime(x) \cos rx \, dx.$$

The following theorem will now be proved.

**Theorem 2.** Let $\frac{1}{2} \leq s \leq 1$, $h/k \geq \rho > 0$, $k \leq t \leq T$, $h \leq h_1 < p/2$. Assume that $f(x)$ possesses an absolutely integrable bounded third derivative and that $f(0) = f(\pi) = 0$. Then

$$|U(x, t) - u(x, t)| \leq c t^{-1} (h^2 + k^\alpha),$$

where $\alpha = 1$ for $s \neq \frac{1}{2}$, and $\alpha = 2$ for $s = \frac{1}{2}$. Also,

$$c = M \sup_{0 \leq s \leq r} (|f''(x)| + |f'''(x)|).$$

The constant $M$ depends on $s$, $\rho$ and $T$ only.

The truncation error will be appraised by splitting the series in

$$U(x, t) - u(x, t) = \sum_{r=1}^{\infty} (E_r t/k - \exp(-r^2 t)) b_r \sin rx$$

into four parts $V_1$, $V_2$, $V_3$, $V_4$ corresponding, respectively, to the ranges $r \leq r_1$, $r_1 < r \leq r_2$, $r_2 < r \leq r_3$, $r_3 < r$. The integers $r_j, j = 1, \ldots, 4$, are defined below.

a) *The range $r \leq r_1$.* The numerator

$$h^2 - 4(1 - s)k \sin^2 (rh/2)$$

of $E_r$ vanishes if

$$r = \frac{2}{h} \arcsin \frac{h}{2 \sqrt{(1 - s)k}}. \tag{3.6}$$

Let $r_0$ be the smallest positive determination of the right member of (3.6) if the latter is real, i.e., if $h/(2 \sqrt{(1 - s)k}) \leq 1$. Otherwise, set $r_0 = \infty$. If $r_0$ is finite it lies in the interval

$$\frac{1}{\sqrt{(1 - s)k}} \leq r_0 \leq \frac{\pi/2}{\sqrt{(1 - s)k}}, \tag{3.7}$$

because $1 \leq (\arcsin q)/q \leq \pi/2$, for $0 \leq q \leq 1$, if the principal value of $\arcsin q$ is taken. Define $r_1$ by

$$r_1 = \left\lfloor \sqrt{\theta p/k} \right\rfloor, \quad 0 < \theta < 1. \tag{3.8}$$

Here $[x]$ designates the largest integer not exceeding $x$.

The letters $c_j, j = 1, 2, \ldots,$ below denote positive constants that may depend on $s$, $\rho$, and $T$, but not on $h$, $k$, $x$, $t$, or $f(x)$.

In order to appraise $V_1$ let $r \leq r_1$, and observe that, by (3.2), (3.7) and (3.8),

$$r_1 \leq r_0/\sqrt{2}. \tag{3.9}$$
Also, if \( z_r \) is defined by

\[
z_r = 4kh^{-2} \sin^2 (rh/2),
\]

it follows from (3.8) that

\[
z_r \leq k r_i^2 \leq \theta,
\]

for \( r \leq r_1 \).

Since

\[
E_r = \frac{1 - (1 - s) z_r}{1 + sz_r},
\]

an application of the Lemma of section 5 shows that

\[
E_r^{1/k} = \exp \left[ -t(z_r - R)/k \right],
\]

where \( R \) satisfies the inequalities

\[
-c z_r^2 \leq R \leq \frac{1}{2} s^2 z_r^2,
\]

for \( r \leq r_1 \).

If \( S \) is defined by

\[
S = r^2 - z_r k^{-1} + R k^{-1},
\]

formula (3.12) can be written

\[
E_r^{1/k} = \exp (-r^2 t + St),
\]

so that

\[
E_r^{1/k} - \exp (-r^2 t) = \exp (-r^2 t) (\exp St - 1) = St \exp ((S_* - r^2)t),
\]

with \( S_* \) having a value between 0 and \( S \). The quantity \( S \) can be appraised by inserting for \( z_r \) its expression from (3.10) and by applying the inequalities (3.13) and

\[
0 \leq r^2 - \sin^2 r \leq \frac{1}{6} r^4.
\]

This leads to

\[
-c k r^4 \leq S \leq \frac{1}{24} r^4 h^2 + \frac{1}{2} s^2 k r^4
\]

and therefore, if (3.2), (3.8) and (3.11) are recalled, to

\[
S_* \leq S \leq \frac{1}{24} r^4 h^2 + \frac{1}{2} s^2 k r^4 \leq \frac{1}{24} p^2 + \frac{1}{2} s^2 r^2.
\]

Applying the inequalities (3.17) and (3.18) to (3.16) the appraisal

\[
| E_r^{1/k} - \exp (-r^2 t) | \leq c_3 (h^2 + k) r^4 \exp (-3r^2 t/4),
\]

is obtained.

The expected improvement in the order of magnitude, when \( s = \frac{1}{3} \) can be based on part c) of the Lemma in section 5, according to which

\[
-c z_r^3 \leq R < 0,
\]

for \( s = \frac{1}{3} \).
so that (3.17) can be replaced by

\[-c_5k^2r^6 \leq S \leq \frac{1}{24} r^4h^2 \leq \frac{1}{24} p^2, \quad s = \frac{1}{2}\]

Inserted into (3.16) this leads to

\[
| E_{s/k} - \exp (-r^2t) | \leq tc_4(r^4h^2 + r^6k^2) \exp (-r^2t), \quad s = \frac{1}{2}, \quad r \leq r_1.
\]

The quantity $V_1$ is now readily appraised by means of (3.3), (3.19), and (3.20). One has

\[
(3.21) \quad | V_1 | = \left| \sum_{r=1}^{r_1} (E_{s/k} - \exp (-r^2t)) b_r \sin rx \right| \leq tc_8(h^2 + k)K \sum_{r=1}^{r_1} r \exp (-3r^2t/4) \leq tc_8(h^2 + k) K \int_0^\infty r \exp (-3r^2t/4) dr = c_8K(h^2 + k),
\]

and for $s = \frac{1}{2}$, in addition, in view of (3.20),

\[
(3.22) \quad | V_1 | \leq tc_8K \left\{ h^2 \int_0^\infty r \exp (-r^2t) dr + k^2 \int_0^\infty r^2 \exp (-r^2t) dr \right\} \leq t^{-3} c_8K(h^2 + k)^{3/2}, \quad s = \frac{1}{2}.
\]

b) The range $r_1 < r \leq r_2$. The quantity $E_r$ is a decreasing function of $r$ and positive for $r < r_0$. Hence, $E_{s/k}$ is a positive decreasing function of $r$ in $r < r_0$. Let $r_2$ be defined by

\[
(3.23) \quad r_2 = \left[ r_0 \right]
\]

(so that $r_2 = \infty$, if $r_0 = \infty$), then

\[0 \leq E_{s/k} \leq E_{r_1}, \quad r_1 \leq r \leq r_2\]

or, in view of (3.19),

\[
| E_{s/k} - \exp (-r^2t) | \leq E_{r_1} + \exp (-r_1^2t) \leq | E_{r_1} - \exp (-r_1^2t) | + \exp (-r_1^2t) \leq \{ tc_8(h^2 + k)r_1^4 \} \exp (-r_1^2t/2).
\]

By virtue of formulas (3.3) and (3.5) this implies

\[
| V_2 | \leq \{ tc_8(h^2 + k)r_1^4 + 1 \} \exp (-r_1^2t/2) K \sum_{r=1}^{r_1} r^{-3}.
\]

The last summation is less than $\int_{r_1}^{\infty} r^{-3} dr = 2r_1^2$. Now we assume that $h_1 \leq \theta p/16$, so that $r_1^2 > \theta p/h - 2\sqrt{\theta p/h} + 1 \geq \theta p/2h$. Then the inequality for $| V_2 |$ becomes

\[
(3.24) \quad | V_2 | \leq \{ c_7(h^2 + k) + (8h/\theta p) \} \exp (-\theta p/4h) K \leq c_9 r_1^{-1}K(h^2 + k), \quad \text{for } r_1 \leq r \leq r_2.
\]
v) The range $r_2 < r \leq r_3$. The set $r > r_3$ is empty, if $r_0 = \infty$. In the opposite case, $E_r$ is a decreasing negative function in $r_0 < r < \pi/h$. Let

$$r_3 = \pi/h$$

Then, for $r_2 < r \leq r_3$,

$$|E_r^{t/h}| = |E_r|^{t/h} = \left(\frac{4(1-s)kh^2(rh/2)-h^2}{4skh^2(rh/2)+h^2}\right)^{t/h} \leq \left(\frac{r^2ks-1}{r^2ks+1}\right)^{t/h} \leq \left(\frac{r^2ks-1}{r^2ks+1}\right)^{t/h}.$$

Here we have made use of the monotonic dependence of $E_r$ on the variable $4k \sin 2rh/2$.

As $r^2ks > r_0ks > s/(1-s)$, part b) of the Lemma in section 5 can be applied to the last bound for $|E_r^{t/h}|$, provided $s > \frac{1}{2}$. This leads to

$$|E_r^{t/h}| \leq \left(\frac{1}{1 + 1/r^2ks}\right)^{t/h} \leq \exp(-2t/r^2k^2s), \quad r_2 < r \leq r_3.$$  

By continuity, formula (3.25) remains true for $s = \frac{1}{2}$. Hence, using (3.3) and (3.25)

$$\left|\sum_{r=r_3}^{r_2} b_r E_r^{t/h} \sin rx\right| \leq K \sum_{r=r_3}^{r_2} r^{-3} \exp(-2t/r^2k^2).$$

The function of $r$ in the last summation assumes its maximum for $r = r_m = \sqrt{4t/3k^2}$. It is an increasing function of $r$ for $r < r_m$, a decreasing one for $r > r_m$, and the value of the maximum is less than $t^{-3/2}kh^3$. Therefore, (3.26) can be replaced by

$$\left|\sum_{r=r_3}^{r_2} b_r E_r^{t/h} \sin rx\right| \leq K \int_0^\infty r^{-3} \exp(-2t/r^2k^2) dr + Kt^{-1/2}k^3 \leq K(k^2/4t + t^{-1/2}k^3) \leq c_{19}t^{-1}Kh^3, \quad \text{for } t \geq k.$$  

On the other hand,

$$\left|\sum_{r=r_3}^{r_2} b_r \exp(-rt \sin rx)\right| \leq K \int_{r_3}^{\infty} \exp(-rt)^{-3} dr$$

$$= -(1/2t) \exp(-rt^2-r^4)\bigg|_{r_3}^{\infty} - (2/t) \int_{r_3}^{\infty} \exp(-rt^2) r^4 dr \leq t^{-1}K_r t^4 \leq t^{-1}K^3,$$

because $r_3 \geq r_0 \geq ((1-s)k)^{-1}$. Combination of (3.27) and (3.28) proves that

$$|V_3| \leq c_{15}t^{-1}Kh^3, \quad \text{for } t \geq k.$$  

d) The range $r > r_3$. For $r > r_3 = \pi/h$,

$$|V_3| \leq \sum_{r=\pi/h}^{\infty} |b_r| \leq K \int_{\pi/h}^{\infty} r^{-3}dr \leq c_{15}Kh^3.$$
The desired bound for the truncation error is now obtained by adding the
inequalities (3.24), (3.29), and (3.30) to (3.21) or (3.22), respectively. This com-
pletes the proof of theorem 2.

The order of magnitude of this appraisal cannot, in general, be improved. One
consequence of this theorem is that the particular choice \( s = \frac{1}{2} \) improves the order
of magnitude of the truncation error only if \( h = o(\sqrt{k}) \).

As was mentioned before, it is possible to carry out the arguments of this section
with the milder hypothesis

\[
h^\beta/k \geq \rho > 0,
\]

taking the place of (3.2). However, higher powers of \( t^{-1} \) will then appear in the bound
for the truncation error. If \( k \) shrinks so much more slowly than \( h \) that

\[
| \log t^{-1/k} | \leq \rho < \infty,
\]

then the method of proof breaks down decisively.

4. Stability. The term computational stability refers to the extent to which
small errors in the data and in the stepwise computation of \( U(x, t) \) affect the final
result. In the case of linear homogeneous difference equations the cumulative mod-
ification of the solution by such errors is the sum of the departures caused by errors
on each line \( t = 0, k, 2k, \ldots \). Thus, a knowledge of a bound on \( U(x, t) \) in terms of
the initial function \( f(x) \) conveys information concerning the stability of the pro-
cedure. In the previous section it was decisive that \( f(x) \) had certain smoothness
properties. If the round-off errors are to be treated as uniformly bounded, but
otherwise arbitrary, quantities, no condition beyond (2.3) and boundedness may be
imposed on \( f(x) \) in the present section.

In view of the last remark it is natural to base the arguments here on the finite
trigonometric interpolation series for \( f(x) \), i.e., on

\[
f(mh) = \sum_{r=1}^{N-1} B_r \sin mrh, \quad m = 1, 2, \ldots, N - 1,
\]

where \( N = \pi/k \) and

\[
B_r = \frac{2}{N} \sum_{r=1}^{N-1} f(rh) \sin \rho h, \quad r = 1, 2, \ldots, N - 1.
\]

It can be directly verified that the expression

\[
V(mh, nk) = \sum_{r=1}^{N-1} B_r E_r^n \sin mrh
\]

solves the difference problem (1.3), (1.4) for \( U(x, t) \) at the grid points \( x = mh \)
\( t = nk, m = 1, 2, \ldots, N - 1; n = 1, 2, \ldots \). Hence,

\[
V(mh, nk) = U(mh, nk).
\]
Insertion of (4.2) into (4.3) and an interchange of summations leads to

\[ U(mh, nk) = \sum_{\rho=1}^{N-1} g_{\rho mn}(\rho h) \]

with

\[ g_{\rho mn} = \left(\frac{2}{N}\right) \sum_{r=1}^{N-1} E_r^n \sin mrh \sin \rho rh. \]

The quantity (4.5) will now be appraised under the assumption that (2.6) and hence (2.7) is true. Moreover, the restriction

\[ 0 \leq s \leq 1 \]

will be imposed. Under this hypothesis \( E_r \) is a decreasing function of \( r \), for \( r \leq N-1 \). Therefore, \( E_r^n \) is a decreasing function of \( r \) for all odd \( n \), while for even \( n \) the quantity \( E_r^n \) decreases as long as \( E_r \geq 0 \) and increases thereafter, for \( r \leq N-1 \).

In order to take advantage of this observation (4.5) is written in the form

\[ g_{\rho mn} = \left(\frac{1}{N}\right) \sum_{r=0}^{N-1} E_r^n \cos (m - \rho) rh - \left(\frac{1}{N}\right) \sum_{r=0}^{N-1} E_r^n \cos (m + \rho) rh, \]

and each of the two right hand terms is subjected to a summation by parts. Recalling the identities

\[ \sum_{r=0}^{N-1} a_r b_r = - \sum_{r=0}^{N-1} \left\{ (a_{r+1} - a_r) \sum_{\mu=0}^{r} b_{\mu} \right\} + a_N \sum_{\mu=0}^{N-1} b_{\mu} \]

and

\[ \sum_{\mu=0}^{r} \cos \mu \theta = \frac{\sin \left( \frac{r + \frac{1}{2}}{2} \right) \theta}{\sin \frac{1}{2} \theta} + \frac{1}{2} = \frac{\sin \left( (r + 1) \theta/2 \right) \cos (r \theta/2)}{\sin \frac{1}{2} \theta} \]

one finds that

\[ \left| \left(\frac{1}{N}\right) \sum_{r=0}^{N-1} E_r^n \cos (m - \rho) rh \right| \]

\[ = \left| - \left(\frac{1}{N}\right) \sum_{r=0}^{N-1} \left\{ (E_{r+1}^n - E_r^n) \sum_{\mu=0}^{r} \cos (m - \rho) \mu h \right\} \right| \]

\[ + \left(\frac{1}{N}\right) E_N^n \sum_{\mu=0}^{N-1} \cos (m - \rho) \mu h \]

\[ \leq h N \sum_{r=0}^{N-1} \left| E_r^n - E_{r+1}^n \right| + \left| E_N^n \right| \]

\[ \leq \frac{1}{\pi} \left| \sin \frac{1}{2}(m - \rho)kh \right|. \]

Now, if \( n \) is odd,

\[ \sum_{r=0}^{N-1} \left| E_r^n - E_{r+1}^n \right| = \sum_{r=0}^{N-1} (E_r^n - E_{r+1}^n) = E_0^n - E_N^n \leq 2, \quad \text{odd } n. \]

If \( E_r \) is non-negative for \( 0 \leq r_1 \leq N \) the last formula is correct for even \( n \) as well.
IMPLICIT DIFFERENCE APPROXIMATIONS

But if \( E_r \) vanishes for \( r = r_0 \) one has, instead, with \( r_s = \lceil r_0 \rceil \),

\[
\sum_{r=0}^{N-1} | E_r^n - E_{r+1}^n | = \sum_{r=0}^{[r_0]-1} (E_r^n - E_{r+1}^n) + | E_{r+1}^n - E^n_s |
\]

\[
+ \sum_{r=[r_0]+1}^{N-1} (E_r^n - E_{r+1}^n) = E_0^n - E^n_s + | E_{r+1}^n - E^n_s - E^n_s + E^n_s | \leq 3,
\]
even \( n \).

Inserting the result of the last two inequalities into (4.8) and taking account of the fact that both \( m \) and \( \rho \) range between 0 and \( N = \pi/h \) it follows that the first right hand term in (4.7) is numerically not greater than \( 4 | m - \rho |^{-1} \). The second summation in (4.7) is similarly found to be bounded, numerically, by

\[
4h/(\pi \sin \frac{1}{2} (m + \rho)h),
\]

where \( m \) and \( \rho \) range from zero to \( \pi/h \). For \( m + \rho \leq \pi/h \) that bound is clearly not greater than \( 4(m + \rho)^{-1} \), which is less than \( 4 | m - \rho |^{-1} \). For

\[
m + \rho = (2\pi/h) - \gamma, \quad 0 < \gamma < \pi/h,
\]

the bound \( 4\gamma^{-1} = ((2\pi/h) - m - \rho)^{-1} \) is obtained, instead. But this last quantity is also not greater than \( 4 | m - \rho |^{-1} \), since \((2\pi/h) - m - \rho = ((\pi/h) - m) + ((\pi/h) - \rho) \geq | ((\pi/h) - m) - ((\pi/h) - \rho) | = | \rho - m | \). Collecting these results it is seen that (4.7) implies the inequality

\[
| g_{m\rho} | \leq \frac{8}{| m - \rho |}.
\]

On the other hand,

\[
| g_{m\rho} | \leq 2,
\]

because of (2.7) and (4.5). Combination of the last two inequalities yields

(4.9)

\[
| g_{m\rho} | < \frac{10}{| m = \rho | + 1}.
\]

Let \( \tilde{f} \) be defined by

\[
\tilde{f} = \text{l.u.b.} | f(x) |.
\]

Then (4.4) and (4.9) imply

\[
| U(mh, nk) | \leq 10^{\tilde{f}} \sum_{\rho=1}^{N-1} \frac{1}{| m - \rho | + 1}
\]

\[
\leq 10^{\tilde{f}} \left\{ \int_0^m \frac{dx}{1 + m - x} + \int_m^{N-1} \frac{dx}{1 - m + x} + 1 \right\}
\]

\[
\leq 10^{\tilde{f}} (1 + 2 \log (\pi/h)) \leq C^{\tilde{f}} \log (1/h),
\]
where $C$ is a constant. This proves

**Theorem 3.** If $f(0) = f(\pi) = 0$, and $|f(x)| \leq \tilde{f}$ in $0 \leq x \leq \pi$; and if

$$2(1 - 2s)k - h^2 \leq 0, \quad 0 \leq s \leq 1,$$

then $|U(mh, nk)| \leq C\tilde{f} \log 1/h$, for $0 \leq m \leq \pi/h$, $0 \leq n$. The constant $C$ depends on $h_1$ only.

As a consequence of this theorem the departure caused at a grid point by the cumulative effect of all errors at grid points below that $t$-level is less than

$\epsilon C\tilde{f} \log (1/h)k^{-1},$

where $\epsilon$ is an upper bound for the error committed at one point. If $h_1$ is small $C$ is about equal to 10. Round-off errors of this size are considered quite manageable in numerical work.

It is an open question whether the result of theorem 3 is the best possible, i.e., if there exists a constant $C^*$ such that corresponding to arbitrarily small values of $h$ and $k$ satisfying (2.6) a function $f(x)$ can be found for which

$$|U(mh, nk)| > C^*\tilde{f} \log (1/h)$$

at some point $x = mh, t = nk$. It is well known that the factor $\log (1/h)$ can be omitted if $h, k$ and $s$ satisfy the inequality

$$s(1 - s)k - h^2 \leq 0,$$

for then the difference equation (1.3) satisfies a maximum principle. But this inequality is more restrictive than (2.6). For $s = \frac{1}{2}$, for instance, it implies the undesirable condition $k/h^2 < 1$. The boundedness statement proved in [5] is compatible with values of $U$ of the order $O(h^{-4})$ and is therefore weaker than that of theorem 3.

**5. A Lemma.**

**Lemma:** Let $\Phi_s(z) = \log \frac{1 - (1 - s)z}{1 + sz} + z$.

a) If $\min(sz, -(1 - s)z) \geq -\theta_1 > -1$, then

$$|\Phi_s(z)| \leq c_1sz^2,$$

$c_1$ a constant.

b) If $0 \leq s \leq 1, z \geq 0$ and $(1 - s)z < 1$, then

$$\Phi_s(z) \leq \frac{1}{2}s^2z^2.$$

c) If $z/2 \leq \theta_2 < 1$, then

$$-c_2z^3 \leq \Phi_s(z) \leq 0,$$

c_2 a constant.

**Proof.** a) It is readily verified that

$$\Phi_s(z) = \int_{-(1-s)z}^{\infty} \frac{\tau}{1 + \tau} \, d\tau, \text{ if } -(1 - s)z > -1, \quad sz > -1.$$

Under the assumption of part a) one has $|1 + \tau| > 1 - \theta_1$ in the interval of integration. Hence,

$$|\Phi_s(z)| \leq (1 - \theta_1)^{-1} \int_{-(1-s)z}^{\infty} \frac{|\tau| \, d\tau}{|1 + \tau|} \leq c_1sz^2.$$
b) Since \(-(1 - s)z < 0, sz > 0\) it follows from (5.1) that

\[
\Phi_s(z) \leq \int_0^s \frac{r}{1 + r} \, dr \leq \int_0^s r \, dr \leq \frac{1}{2} (sz)^2.
\]

c) This result is an immediate consequence of the identity

\[
\Phi_s(z) = -2 \int_0^s \frac{r^2}{1 - r^2} \, dr.
\]

Mathematics Research Center,
United States Army,
University of Wisconsin,
Madison, Wisconsin


An Open Formula for the Numerical Integration of First Order Differential Equations (II)

By Herbert S. Wilf

In a previous paper [1], referred to below as \(I\), a set of formulas was derived for the numerical integration of systems of first order differential equations. In what follows we will consider the questions of convergence and stability of the method and higher order formulas.

We quote here, for reference, the final results of \(I\), namely the propagation formulas:

\[
(1) \quad y_1 = y_0 + \frac{h}{12} \left[ 5f(x_0, y_0) + 8f(x_1, y_1) - f(x_2, y_2^*) \right]
\]

\[
(2) \quad y_2^* = 5y_0 - 4y_1 + 2h[f(x_0, y_0) + 2f(x_1, y_1)]
\]

for the solution of

\[
(3) \quad y' = f(x, y)
\]

where \(y_0\) is given.

1. **Convergence.** The solution of (1) and (2) for \(y_1\) proceeds by taking as an initial guess the \(y_2^*\) from the preceding point. One then applies (2) and (1) successively until consecutive values of \(y_1\) agree to sufficient accuracy.

Received 16 August 1957.