JOURNAL OF APPROXIMATION THEORY 1, 236–242 (1968)

Non-Linear Difference Schemes

D. P. SQUIER

Colorado State University, Fort Collins, Colorado 80521

1. The purpose of this paper is to give a brief theory of non-linear difference schemes for solving the initial value problem

$$\frac{dy}{dt} = f(t, y) \tag{1}$$

$$y(0) = y_0. \tag{2}$$

Non-linear here means non-linear in f. Timlake [1] and Timlake and Kendall [2] have indicated how such non-linear schemes arise from linear multistep schemes by averaging and by using predictor–corrector methods. A brief account of the latter will be given here.

2. A general k-step difference equation may be given by

$$\phi_n = F(\phi_n, \phi_{n-1}, \ldots, \phi_{n-k}; n) \qquad n \geqslant k, \tag{3}$$

where we assume F is defined for all values of its arguments and

$$|F(\alpha, x_{k-1}, x_{k-2}, \ldots, x_0) - F(\beta, x_{k-1}, x_{k-2}, \ldots, x_0)| \leqslant R|\alpha - \beta|$$

with $R < 1$. It is therefore possible to solve for ϕ_n uniquely in (3) in terms of $\phi_{n-j}, j = 1, 2, \ldots k$ to obtain

$$\phi_n = G(\phi_{n-1}, \phi_{n-2}, \ldots, \phi_{n-k}; n) \qquad n \geqslant k. \tag{4}$$

DEFINITION 1. The equation (4) (or 3) is called stable if and only if there is a number $K \geqslant 1$ such that if $\{u_n\}$ and $\{v_n\}$ are any two sequences satisfying (4) for $n \geqslant N_0$, then

$$|u_n - v_n| \leqslant K \sum_{r=0}^{k-1} |u_{j+r} - v_{j+r}|, \tag{5}$$

for all n, j for which $n \geqslant j \geqslant N_0 - k$. K is independent of the sequences, n, N_0, and j; K is called a stability constant for (4).

It follows that G satisfies a Lipschitz condition in its first k arguments and that K is not less than the Lipschitz constant.

If $\rho(x) = \sum_{j=0}^{k} \alpha_j x^j$, $\alpha_k \neq 0$, and $\rho(z) = 0$ implies $|z| \leqslant 1$ and $\rho'(z) \neq 0$ if $|z| = 1$, then

$$\sum_{j=0}^{k} \alpha_j u_{n+j} = 0 \tag{6}$$

is a stable k-step equation [3]. Such an equation is called a ρ-equation.

For convenience in the sequel it will be assumed that $k = 3$, though the reasoning will apply for any k.

LEMMA 1. *If* (i) $\phi_n = G(\phi_{n-1}, \phi_{n-2}, \phi_{n-3}; n) + \omega_n$, *where* $\{\omega_n\}$ *is a given sequence, is stable;* (ii) $\{u_n\}$ *is a sequence satisfying the equation in* (i) *and* $\{v_n\}$ *is a sequence satisfying* $v_n = G(v_{n-1}, v_{n-2}, v_{n-3}; n) + \lambda_n$, *then*

$$|u_n - v_n| \leqslant K \sum_{r=j+3}^{n-1} |\omega_r - \lambda_r| + K \sum_{r=j}^{j+2} |u_r - v_r| + |\omega_n - \lambda_n| \tag{7}$$

for all $n \geqslant j \geqslant 0$. *Here* K *is a stability constant for* (i).

Proof. The notation $u_n(\alpha, \beta, \gamma; s)$ denotes the nth member of the sequence satisfying (i) for $n \geqslant s + 1$ with initial values $u_s = \alpha$, $u_{s-1} = \beta$, $u_{s-2} = \gamma$. If $Z_n^s = u_n(v_s, v_{s-1}, v_{s-2}; s)$, then

$$u_n - v_n = u_n(u_{j+2}, u_{j+1}, u_j, j+2) - u_n(v_{j+2}, v_{j+1}, v_j, j+2)$$

$$+ \sum_{s=j+2}^{n-2} (Z_n^s - Z_n^{s+1}) + Z_n^{n-1} - v_n. \tag{8}$$

Now

$$\begin{aligned} Z_n^{n-1} &= u_n(v_{n-1}, v_{n-2}, v_{n-3}, n-1) \\ &= G(v_{n-1}, v_{n-2}, v_{n-3}, n) + \omega_n \\ &= v_n - \lambda_n + \omega_n. \end{aligned}$$

Also,

$$Z_n^s = u_n(Z_{s+1}^s, v_s, v_{s-1}; s+1).$$

Thus,

$$\begin{aligned} |Z_n^s - Z_n^{s+1}| &\leqslant |u_n(Z_{s+1}^s, v_s, v_{s-1}; s+1) - u_n(v_{s+1}, v_s, v_{s-1}; s+1)| \\ &\leqslant K|Z_{s+1}^s - v_{s+1}| \\ &= K|G(v_s, v_{s-1}, v_{s-2}; s+1) + \omega_{s+1} - v_{s+1}| \\ &= K|\omega_{s+1} - \lambda_{s+1}|. \end{aligned}$$

Combining these estimates in (8) produces (7), the second sum on the right in (7) coming from the first two terms on the right in (8) and the stability of (i).

3. In solving (1) by finite differences we assume f is continuous on $I \times R$ where $t \in I = [0, T]$ and $y \in R$, the totality of reals. Further,

$$|f(t, y_1) - f(t, y_2)| \leqslant L|y_1 - y_2|$$

for all $t \in I, y_1, y_2 \in R$. For $h > 0$ we form the set of points $\{t_m\}$ where $0 \leqslant t_m \leqslant T$ and $t_m = mh$.

DEFINITION 2. A difference equation

$$\phi_n = G(\phi_{n-1}, \phi_{n-2}, \phi_{n-3}, n, h) \tag{9}$$

is said to be consistent with the differential equation (1) if, for any y a solution of (1), there is a constant h_0 such that

$$y(t_n) - G(y(t_{n-1}), y(t_{n-2}), y(t_{n-3}), n, h) = \lambda_n$$

for all n for which $t_n \leqslant T$, $0 < h \leqslant h_0$, and $|\lambda_n| \leqslant \epsilon_1(y,h)h$, where $\epsilon_1(y,h) \to 0$ as $h \to 0$. λ_n is called the discretization error of (9). (9) is said to be of order p if there is a constant c_1 such that $|\lambda_n| \leqslant c_1 h^{p+1}$ but no c_2 such that $|\lambda_n| \leqslant c_2 h^{p+2}$. c_1 may depend on y. Equation (9) is called h-stable if the inequality in (5) holds for all n for which $nh \leqslant T$ and all sufficiently small h, say $0 < h \leqslant h_1$. K may depend on T, but not on h. A sequence $\{u_n\}$ satisfying (9) is said to converge to a solution y of (1) if, for fixed $nh \leqslant T$, $|u_n - y(nh)| \to 0$ as $h \to 0$. A set of numbers $s_0(h)$, $s_1(h)$, $s_2(h)$ is called a compatible set of starting values for the initial value problem (1)–(2) if

$$|s_0 - y(0)| + |s_1 - y(h)| + |s_2 - y(2h)| \to 0 \text{ as } h \to 0$$

if y is the solution of (1)–(2).

THEOREM 1. *If equation* (9) *is h-stable and consistent with* (1) *and* s_0, s_1, s_2 *is a compatible set of starting values for* (1)–(2), *then the sequence* $\{u_n\}$ *satisfying* (9) *with* $u_j = s_j, j = 0, 1, 2$, *converges to the solution of* (1)–(2).

PROOF. With $v_n = y(t_n)$, y the solution of (1)–(2), it follows that $\{v_n\}$ satisfies

$$v_n = G(v_{n-1}, v_{n-2}, v_{n-3}, n, h) + \lambda_n, \qquad n \geqslant 3,$$

with $|\lambda_n| \leqslant \epsilon_1(y,h)h$. From Lemma 1, with $\omega_n = 0$ all n,

$$\begin{aligned} |u_n - v_n| &\leqslant K \sum_{r=3}^{n} |\lambda_r| + K \sum_{r=0}^{2} |s_r - v_r| \\ &\leqslant K\epsilon_1(y, h)\, nh + K\epsilon_2(h) \\ &\leqslant KT\epsilon_1(y, h) + K\epsilon_2(h) \end{aligned}$$

and $\epsilon_1, \epsilon_2 \to 0$ as $h \to 0$.

4. We show here how h-stable equations may be generated from stable or h-stable equations. An application to the work [2] will be made.

LEMMA 2. *If Q, P, r_m are non-negative for all m, $m = 0, 1, 2, \ldots$ and*

$$r_n \leqslant Q \sum_{m=0}^{n-1} r_m + P$$

for all n, then $r_n \leqslant (1+Q)^{n-1}(Qr_0 + P)$.

Proof. $r_1 \leqslant Qr_0 + P$. If $Q \sum_{m=1}^{n-1} r_m + P \leqslant B_{n-1}$, then

$$\begin{aligned} r_{n+1} &\leqslant Qr_n + Q \sum_{m=0}^{n-1} r_m + P \\ &\leqslant Qr_n + B_{n-1} \\ &\leqslant QB_{n-1} + B_{n-1} = (1+Q)\, B_{n-1}. \end{aligned}$$

Thus if B_k is defined recursively by $B_m = (1+Q)B_{m-1}$, $B_0 = (Qr_0 + P)$, the result follows.

LEMMA 3. *If $\phi_n = G(\phi_{n-1}, \phi_{n-2}, \phi_{n-3}, n, h) + \omega_n$ is stable or h-stable for every sequence $\{\omega_n\}$ with stability constant K independent of $\{\omega_n\}$ and h for $h \leqslant h_0$, then*

$$\phi_n = G(\phi_{n-1}, \phi_{n-2}, \phi_{n-3}, n, h) + hS(\phi_{n-1}, \phi_{n-2}, \phi_{n-3}, n, h) \tag{10}$$

is h-stable if S is Lipschitz continuous in its first three arguments, the Lipschitz constant L being independent of n and h. A stability constant for (10) is $K\exp(3KLT)$; for a k-step method the 3 in the exponent is replaced by k.

Proof. Let $\{u_n\}$ & $\{v_n\}$ be two sequences satisfying (10). Then with

$$\begin{aligned} \omega_n &= hS(u_{n-1}, u_{n-2}, u_{n-3}, n, h) \\ \lambda_n &= hS(v_{n-1}, v_{n-2}, v_{n-3}, n, h) \end{aligned}$$

it follows from Lemma 1 that

$$\begin{aligned} |u_n - v_n| &\leqslant K \sum_{r=j+3}^{n} |\omega_r - \lambda_r| + K \sum_{r=j}^{j+2} |u_r - v_r| \\ &\leqslant KhL \sum_{r=j+3}^{n} \left(\sum_{p=0}^{2} |u_{r-1-p} - v_{r-1-p}| \right) + K \sum_{r=j}^{j+2} |u_r - v_r| \\ &\leqslant 3KLh \sum_{r=j}^{n-1} |u_r - v_r| + K \sum_{r=j}^{j+2} |u_r - v_r|. \end{aligned}$$

From Lemma 2, with $Q = 3KLh$,

$$r_m = u_{j+m} - |v_{j+m}|, \qquad m = 0, 1, 2, \ldots$$

and

$$P = K \sum_{r=j}^{j+2} |u_r - v_r|,$$

it follows that

$$|u_n - v_n| \leqslant (1 + 3KLh)^{n-j-1}(3KLh|u_j - v_j| + P)$$

$$\leqslant \left(1 + \frac{3KLT}{n}\right)^n \frac{K(3Lh+1)}{(3KLh+1)^{j+1}} \sum_{r=j}^{j+2} |u_r - v_r|$$

$$\leqslant (\exp 3KLT)\, K \sum_{r=j}^{j+2} |u_r - v_r|.$$

To obtain the last inequality, $K \geqslant 1$ has been used.

The $\rho - \sigma$ equations of [3] are defined by

$$\sum_{r=0}^{k} \alpha_r u_{n+r} = h \sum_{r=0}^{k} \beta_r f(t_{n+r}, u_{n+r}) \tag{11}$$

where the α's and β's are constants. If $\alpha_k \neq 0$, $\beta_k = 0$, the h-stability of the $\rho - \sigma$ equation follows from the stability of the ρ-equation by Lemma 3 under the assumed Lipschitz continuity of $f(t, y)$. If $\beta_k \neq 0$, a slight modification of the proof in Lemma 3 shows the stability constant given there should be multiplied by

$$\left(1 - \frac{\beta_k}{\alpha_k} KLh\right)^{-1}, \text{ assuming } \left|\frac{\beta_k}{\alpha_k} KLh\right| < 1.$$

In predictor–corrector methods for solving (1)–(2) by finite differences one uses "predictors"

$$\begin{aligned} \phi_n &= H(\phi_{n-1}, \phi_{n-2}, \phi_{n-3}, n, h) \\ \phi_{n+1} &= J(\phi_{n-1}, \phi_{n-2}, \phi_{n-3}, n, h) \end{aligned} \tag{12}$$

in conjunction with a "corrector"

$$\phi_n = G(\phi_{n-1}, \phi_{n-2}, \phi_{n-3}, n, h) + hS(\phi_{n+1}, \phi_n, \phi_{n-1}, \phi_{n-2}, \phi_{n-3}, n, h). \tag{13}$$

Here H, J, S are Lipschitz continuous in their ϕ arguments, the Lipschitz constant L being independent of n and h. (12) and (13) are consistent with (1) but not necessarily h-stable, though it is assumed that

$$\phi_n = G(\phi_{n-1}, \phi_{n-2}, \phi_{n-3}, n, h) + \omega_n \tag{14}$$

is h-stable for every sequence $\{\omega_n\}$ as in Lemma 3. Values of ϕ_n and ϕ_{n+1} are first computed from (12) and then placed in the right side of (13) to obtain the final value of ϕ_n. Thus S becomes a function S^* of $\phi_{n-1}, \phi_{n-2}, \phi_{n-3}$. From the Lipschitz continuity of S follows the Lipschitz continuity of S^* with Lipschitz constant $L(2L + 1)$. By Lemma 3, (12)–(13) is h-stable since (14) is. It is a simple matter to show that if $\lambda_n^{(1)}$, $\lambda_n^{(2)}$, $\lambda_n^{(3)}$ are the truncation errors in (12)

and (13), respectively, then the truncation error λ_n in the combined (12)–(13) satisfies

$$|\lambda_n| \leqslant |\lambda_n^{(3)}| + Lh(|\lambda_n^{(1)}| + |\lambda_n^{(2)}|).$$

Thus, if (13) is of order p and (12) is of order $p-1$, then (12)–(13) is of order at least p. In particular, (12)–(13) is consistent with (1) if that is true of (12) and of (13). By Theorem 1, (12)–(13) is then a convergent scheme for (1)–(2).

It is known [3] that a k-step $\rho - \sigma$ equation (11) cannot be of order greater than $k+2$ and still be stable. Thus a k-step scheme of order $>k+2$ must be unstable or non-linear. In [2] k-step schemes of order $>k+2$ are constructed from linear predictors and correctors. The predictors are of the form

$$\phi_{n+s} = \sum_{r=-k}^{-1} \alpha_r^{(s)} \phi_{n+r} + h \sum_{r=-k}^{-1} \beta_r^{(s)} f(t_{n+r}, \phi_{n+r}) \tag{15}$$

$$s = 0, 1, 2, \ldots q; q \leqslant k-1,$$

and a corrector of the form

$$\phi_n - \sum_{r=-k}^{-1} \alpha_{k+r} \phi_{n+r} = h \sum_{r=-k}^{q} \beta_{k+r} f(t_{n+r}, \phi_{n+r}). \tag{16}$$

The α's and β's in (15) can be chosen so that each equation in (15) is consistent with (1) of order $2k-1$. The α's in (16) are chosen so that

$$\phi_n - \sum_{r=-k}^{-1} \alpha_{k+r} \phi_{n+r} = 0$$

is stable and then the β's in (16) are determined so that the order of (16) is at least $k+q+1$ [3]. If (15) is substituted into the right side of (16) a (non-linear) k-step equation develops. It is consistent of order $k+q+1$ and h-stable by our general theory. By Theorem 1 the combined scheme (15)–(16) is convergent. We note that in general none of the equations in (15) or (16) is stable by itself.

5. The sequence $\{v_n\}$ actually computed from (4) or (9) may be quite different from the theoretical sequence $\{u_n\}$. Though one desires to obtain u_k from

$$u_k = G(u_{k-1}, u_{k-2}, u_{k-3}, k, h)$$

one actually computes v_k from

$$v_k = G(v_{k-1}, v_{k-2}, v_{k-3}, k, h) + \epsilon_k.$$

ϵ_k is the error introduced by rounding, by using approximations of G instead of G itself, by using wrong numbers for v_{k-1}, v_{k-2}, or v_{k-3}, or by computer fault. The question is how $\{u_n\}$ is related to $\{v_n\}$. If (9) is stable or h-stable, the answer is given by Lemma 1, with $\omega_n = 0$ all n and $\lambda_n = \epsilon_n$.

$$|u_n - v_n| \leqslant K \sum_{r=3}^{n-1} |\epsilon_r| + K \sum_{r=0}^{2} |u_r - v_r| + |\epsilon_n|.$$

If $|\epsilon_r|$ is small relative to h, then no serious error results. This illustrates the need for not taking too small a step in numerical computing.

References

1. W. P. Timlake, On an algorithm of Milne and Reynolds. *BIT*, **5** (1965), 276–281.
2. R. P. Kendall, and W. P. Timlake, A stable k Step Method of Order Greater Than $k+2$. I.B.M. publ. 37.022 (Houston, Tex.), April, 1967.
3. P. Henrici, "Discrete Variable Methods in Ordinary Differential Equations." Wiley, New York, 1962.