THE PROPAGATION OF ERROR IN NUMERICAL INTEGRATIONS

MARK LOTKIN

1. **Introduction.** The numerical integration of differential equations is generally performed by replacing the differential equations by approximate difference equations whose solutions are expected to approach those of the associated differential equations as the step size approaches zero. The replacement of differential by difference equations may clearly be carried out in a variety of ways; the actual choice will depend on particular circumstances, accuracy requirements, computational facilities, etc.

It is now a well known fact that whenever the order of the difference equations exceeds that of the original differential equations there are introduced certain numerical solutions that are extraneous to the original differential equations. The behavior of these extraneous solutions in general determines the usefulness of the integration method. For such a method to be effective it must be "stable" in the sense that the extraneous solutions always remain of negligible size as compared with the actual solutions.

Thus it is of interest to distinguish first between stable and unstable methods. In addition, it is also of interest to determine, in either case, the growth of error in the large, since the knowledge of this quantity permits an estimation of the accuracy obtained.

This paper, then, deals with a number of standard methods of integration, and investigates their stability and propagation of error. Round-off is considered to a certain extent, but not completely; see footnote 1. While some of the results have been obtained previously, mainly by L. H. Thomas [1] and H. Rutishauser [2], others do not seem to be as well known.

The propagation of error was already treated previously by Rademacher [3] and others. While the method of adjoint differential equations employed there seems to be capable of general application, it was used, in [3] especially, for Heun's method only.

Finally there are carried out a few illustrative examples; they show that the theoretical expressions obtained frequently lead to good estimates.

2. **Solutions of linear difference equations.** Let us assume that the

Presented to the Society, September 1, 1953; received by the editors November 11, 1953 and, in revised form, May 3, 1954.

integration of the nth order differential equation has proceeded from a starting point x_0 to a point x_k, and that the original differential equation has been replaced by a difference equation of order s:

$$\alpha_{ks}v_{k+s} + \alpha_{k,s-1}v_{k+s-1} + \cdots + \alpha_{k,0}v_k + \alpha_k = 0, \tag{2.1}$$

where the coefficients α_{kj}, α_k are known, with $\alpha_{ks} \neq 0$ for each k, and initial values $v_0, v_1, \cdots, v_{s-1}$ have been supplied. For our purposes it is now convenient to use matrix notation. Introducing, then, the column matrices

$$u_k = \begin{bmatrix} v_{k+s-1} \\ v_{k+s-2} \\ \cdot \\ \cdot \\ \cdot \\ v_k \end{bmatrix}, \qquad b_k = \begin{bmatrix} \alpha_k \\ 0 \\ \cdot \\ \cdot \\ \cdot \\ 0 \end{bmatrix}, \qquad u_0 = \begin{bmatrix} v_{s-1} \\ v_{s-2} \\ \cdot \\ \cdot \\ \cdot \\ v_0 \end{bmatrix}$$

and the square matrices of order s:

$$J_k = \begin{bmatrix} -\alpha_{ks} & 0 \cdots 0 \\ 0 & 1 \cdots 0 \\ \cdot & \cdot \\ \cdot & \cdot \\ \cdot & \cdot \\ 0 & 0 \cdots 1 \end{bmatrix}, \qquad A_k = \begin{bmatrix} \alpha_{k,s-1} & \alpha_{k,s-2} & \cdots & \alpha_{k0} \\ 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \cdot & & & \cdot \\ \cdot & & & \cdot \\ \cdot & & & \cdot \\ 0 & 0 & \cdots 1 & 0 \end{bmatrix}$$

we may express (2.1) as a system of difference equations of the first order:

$$J_k E u_k = A_k u_k + b_k, \tag{2.2}$$

with E denoting the displacement operator

$$Ev_i = v_{i+1}.$$

Since J_k is nonsingular, (2.2) may be written as

$$u_{k+1} = J_k^{-1} A_k u_k + J_k^{-1} b_k = C_k u_k + d_k, \tag{2.3}$$

with

$$C_k = J_k^{-1} A_k, \qquad d_k = J_k^{-1} b_k.$$

The general solution of (2.3) is

$$u_k = P_{k-1}\left(u_0 + \sum_{t=0}^{k-1} P_t^{-1} d_t\right), \tag{2.4}$$

where

$$P_k = C_k P_{k-1}, \qquad P_0 = C_0,$$

or

$$P_{k-1} = \prod_{t=1}^{k} C_{k-t}.$$

In particular, if C_k is actually independent of k, then $P_{k-1} = C^k$ and

$$u_k = C^k\left(u_0 + \sum_{t=0}^{k-1} C^{-t-1} d_t\right). \tag{2.5}$$

The use of Sylvester's theorem [4] now permits us to express the solution u_k in slightly different form, as follows: let $G(\lambda) = \lambda I - C$, $G_a(\lambda)$ denote the adjoint of $G(\lambda)$, $\delta(\lambda)$ the determinant of $G(\lambda)$, and $\delta'(\lambda) = d\delta/d\lambda$. If the characteristic roots λ_m, $m = 1, 2, \cdots, s$, of C are distinct, then

$$C^k = \sum_{m=1}^{s} \lambda_m^k H_m \tag{2.6}$$

with

$$H_m \equiv H(\lambda_m) = G_a(\lambda_m)/\delta'(\lambda_m).$$

Consequently, from (2.5),

$$u_k = \sum_{m=1}^{s} \lambda_m^k H_m \left(u_0 + \sum_{t=0}^{k-1} \lambda_m^{-t-1} d_t\right),$$

or, if $u_0 = 0$,

$$u_k = \sum_{m=1}^{s} H_m \sum_{t=0}^{k-1} \lambda_m^{k-t-1} d_t. \tag{2.7}$$

If the distinct roots λ_i, $i = 1, 2, \cdots, q$, of C have multiplicities μ_i, then (2.6) must be replaced by

$$C^k = \sum_i \frac{1}{(\mu_i - 1)!}\left[\frac{d^{\mu_i - 1}}{d\lambda^{\mu_i - 1}}\left(\frac{\lambda^k G_a(\lambda)}{\delta_{\mu_i}(\lambda)}\right)\right]_{\lambda=\lambda_i}, \tag{2.8}$$

$$\delta_{\mu_i}(\lambda) = (\lambda - \lambda_i)^{-\mu_i}\delta(\lambda).$$

In particular, if the only multiple root of C has the value zero, then clearly (2.8) reverts to the form (2.6) with the summation to be extended over all the nonvanishing roots. This observation will be put to use in the subsequent discussion.

3. **The variational difference equations.** The foregoing treatment

of the difference equation will now be applied to the numerical integration of the nth order differential equation

$$y^{(n)} = f_n(x, y, y', \cdots, y^{(n-1)}) \tag{3.1}$$

subject to suitable boundary conditions. The exact solution $y(x)$ of (3.1), assumed to exist uniquely, may be expressed in the form [2]

$$\begin{aligned} y(x_{k+1}) &= \sum_{j=0}^{r} a_{0j}y(x_{k-j}) + h \sum_{j=-1}^{r} a_{1j}y'(x_{k-j}) + \cdots \\ &\quad + h^N \sum_{j=-1}^{r} a_{Nj}y^{(N)}(x_{k-j}) + T_k, \\ y^{(i)}(x_{k+1}) &= \sum_{j=0}^{r} a_{ij}^{(i)} y^{(i)}(x_{k-j}) + h \sum_{-1}^{r} a_{i+1,j}^{(i)} y^{(i+1)}(x_{k-j}) + \cdots \\ &\quad + h^{N-i} \sum_{-1}^{r} a_{Nj}^{(i)} y^{(N)}(x_{k-j}) + T_{ki}, \end{aligned} \tag{3.2}$$

for $i=1, 2, \cdots, n-1$. Here $h=x_{k+1}-x_k$ is the step used in the integration, the $a_{\nu j}^{(i)}$ are constants that will in general depend on r, where r itself indicates a certain range of points x_{k-j} to the left of x_k; to each numerical method of integration there is associated a fixed value of r. The functions T_k, T_{ki} are truncation errors of orders h^{N+1}, h^{N-i+1}, respectively, with $N \geqq n$ denoting a positive integer. In case N exceeds the order n of the equation (3.1) the derivatives of orders $n+1$, $n+2, \cdots, N$ occurring in (3.2) may be obtained by $N-n$ successive differentiations of (3.1):

$$y^{(i)}(x) = f_i(x, y(x), y'(x), \cdots, y^{(i-1)}(x)), \tag{3.3}$$

$i=n+1, \cdots, N$. The coefficients $a_{\nu j}^{(i)}$ are not arbitrary but normally depend on certain conditions (4.5) derived below.

In solving the integration problem numerically (3.1) is actually replaced by

$$*y^{(n)} = *f_n(x, *y, *y', \cdots, *y^{(n-1)}),$$

where the asterisks indicate rounded values, and the method of integration actually employed may be of the form

$$*y_{k+1}^{(i)} = \sum_{j=0}^{r} *a_{ij}^{(i)} \odot *y_{k-j}^{(i)} \oplus \cdots \oplus h^{N-i} \sum_{j=-1}^{r} *a_{Nj}^{(i)} \odot *y_{k-j}^{(N)} \tag{3.2a}$$

for $i=0, 1, 2, \cdots, n-1$, and

$$*y_{k+1}^{(i)} = *f_i(x_{k+1}, *y_{k+1} \cdots *y_{k+1}^{(i-1)}), \qquad n \leqq i \leqq N. \tag{3.3a}$$

In (3.2a) the circled symbols indicate pseudo-addition and pseudo-multiplication, i.e. certain digital operations by which the corresponding arithmetical operations must be replaced whenever numerical calculations (which of necessity involve rounding) are carried out.[1]

In practice the numerical solution of the sets (3.2a) and (3.3a) is obtained iteratively in the following manner: Extrapolation, or some other means, permits the determination of a first set of values for $^*y_{k+1}^{(i)}$, $n \leqq i \leqq N$. Then (3.2a), with $i=n-1, n-2, \cdots, 0$, lead to a first set of values for $^*y_{k+1}^{(i)}$, $0 \leqq i \leqq n-1$. Next an improved set $^*y_{k+1}^{(i)}$, $n \leqq i \leqq N$, is computed by means of (3.3a), etc., this cycle being repeated until duplication occurs.

Our main interest is now the determination of the errors

$$\eta_\nu^{(i)} = {}^*y_\nu^{(i)} - y^{(i)}(x_\nu),$$

and of the associated property of "numerical stability" in the sense that all the $\eta_\nu^{(i)}$ remain small throughout the entire region of integration. By (3.2a) and (3.2),

$$\eta_{k+1}^{(i)} = \sum_{j=0}^{r} [{}^*a_{ij}^{(i)} \odot {}^*y_{k-j}^{(i)} - a_{ij}^{(i)} y^{(i)}(x_{k-j}) + \cdots] - T_{ki}.$$

However, $^*a \odot {}^*y = a^*y + ({}^*a \odot {}^*y - {}^*a^*y) + ({}^*a - a)^*y$, and $^*a \odot {}^*y - {}^*a^*y = \rho$, $^*a - a = \sigma$, with $|\rho| \leqq \mu$, $|\sigma| \leqq \mu$, $\mu = 2^{-1}\beta^{-\gamma}$ denoting the basic rounding error of a computation carried out to γ places in a number system of base β.

Consequently

$$\eta_{k+1}^{(i)} = \sum_{j=0}^{r} a_{ij}^{(i)} \eta_{k-j}^{(i)} + h \sum_{j=-1}^{r} a_{i+1,j}^{(i)} \eta_{k-j}^{(i+1)} + \cdots$$

$$+ h^{N-i} \sum_{j=-1}^{r} a_{Nj}^{(i)} \eta_{k-j}^{(N)} - T_{ki} + \tau_{ki}, \tag{3.4}$$

$$\tau_{ki} = \sum_j (\rho_{ijk} + \sigma_{ij} {}^*y_{k-j}^{(i)}).$$

Further, by (3.3) and (3.3a),

[1] Note that there is no provision in formula (3.2a) for rounding the product involving h^{N-i}. Thus the formula and later special cases of it in §4 are based on the assumption that the independent variable x and the step size h may be chosen exactly, and that multiplication by powers of h does not necessitate rounding. These assumptions could be dropped at the expense of including additional rounding items. As written, however, the conclusions of the paper may not be precisely applicable to numerical integrations in which the term in question is rounded.

$$\overset{(i)}{\eta}_{k+1} = {}^*f_i(x_{k+1}, {}^*y_{k+1}, \cdots) - f_i(x_{k+1}, y(x_{k+1}), \cdots)$$
$$\text{for } i = n, n+1, \cdots, N.$$

But

$${}^*f_i(x, {}^*y, \cdots) = f_i(x, {}^*y, \cdots) + \phi_i,$$

with ϕ_i denoting certain quantities that depend on the procedure employed in calculating f_i from its arguments x, ${}^*y$, $\cdots$. Thus, correct to terms of the first order in η,

$$\overset{(i)}{\eta}_{k+1} = \sum_{j=0}^{i-1} (\partial f_i/\partial \overset{(j)}{y})\overset{(j)}{\eta}_{k+1} + \phi_i, \qquad n \leqq i \leqq N, \tag{3.5}$$

the partial derivatives to be evaluated at x_{k+1}, ${}^*y_{k+1}$, $\cdots$, ${}^*y_{k+1}^{(i-1)}$.

The system (3.4) and (3.5) of difference equations is transformed into the form (2.2) by the introduction of the column matrix U_k, where

$$U_k^T = [\eta_k, \eta_{k-1}, \cdots, \eta_{k-r}; \eta'_k, \cdots, \eta'_{k-r}; \cdots, \overset{(N)}{\eta}_k, \cdots, \overset{(N)}{\eta}_{k-r}],$$

the superscript T denoting transposition. Then our system becomes

$$J_k E U_k = A U_k + b_k, \tag{3.6}$$

where J_k, A are square matrices of order $s = (r+1)(N+1)$, and b_k is a column matrix of the same order. These matrices are composed as follows:

(i) The elements J_{ij} of J_k are square matrices of order $r+1$,
$J_{ii} = I$ for $i = 0, 1, \cdots, N$,
$J_{ij} = 0$ for $0 \leqq i \leqq n-1$, $j < i$, and for $n \leqq i \leqq N$, $j > i$,

$$J_{ij} = \begin{bmatrix} -h^{j-i}\overset{(i)}{a}_{j,-1} & 0 \cdots 0 \\ 0 & \cdot \\ \cdot & \cdot \\ \cdot & \cdot \\ \cdot & \cdot \\ 0 & \cdots 0 \end{bmatrix} \quad \text{for } 0 \leqq i \leqq n-1,\ 1 \leqq j \leqq N,\ i < j,$$

$$J_{ij} = \begin{bmatrix} -\partial f_i/\partial y^{(j)} & 0 \cdots 0 \\ 0 & \cdot \\ \cdot & \cdot \\ \cdot & \cdot \\ \cdot & \cdot \\ 0 & \cdots 0 \end{bmatrix} \quad \text{for } n \leqq i \leqq N,\ 0 \leqq j \leqq N-1,\ i > j;$$

(ii) the elements A_{ij} of A are square matrices of order $r+1$,

$$A_{ii} = \begin{bmatrix} a_{i0}^{(i)} & a_{i1}^{(i)} & \cdots & a_{ir}^{(i)} \\ 1 & 0 & & 0 \\ 0 & & & \cdot \\ \cdot & & & \cdot \\ \cdot & & & \cdot \\ 0 & 0 & 1 & 0 \end{bmatrix} \qquad \text{for } 0 \leqq i \leqq n-1,$$

$$A_{ii} = \begin{bmatrix} 0 & 0 \cdots & 0 \\ 1 & 0 \cdots & 0 \\ 0 & & \cdot \\ \cdot & & \cdot \\ \cdot & & \cdot \\ 0 & 0 \;\; 1 & 0 \end{bmatrix} \qquad \text{for } n \leqq i \leqq N,$$

$$A_{ij} = h^{j-i} \begin{bmatrix} a_{j0}^{(i)} & a_{j1}^{(i)} & \cdots & a_{jr}^{(i)} \\ 0 & & & 0 \\ \cdot & & & \cdot \\ \cdot & & & \cdot \\ 0 & & \cdots & 0 \end{bmatrix} \quad \text{for } 0 \leqq i \leqq n-1,\ 1 \leqq j \leqq N,\ i < j,$$

$A_{ij} = 0$ for $n \leqq i \leqq N$, $j \neq i$, and also for $0 < i \leqq n-1$, $j < i$;

(iii) the column matrix b_k has the transpose

$$b_k^T = [-T_k + \tau, 0, \cdots, 0, \cdots, -T_{k,n-1} + \tau_{k,n-1}, 0, \cdots, 0; \phi_n, 0, \cdots, 0; \cdots, \phi_N, 0, \cdots, 0].$$

Clearly the determinant $D(J)$ of J is of the form

$$D(J) = 1 + C_1 h + C_2 h^2 + \cdots,$$

i.e. $D(J) \neq 0$ for sufficiently small h.

Let us now assume that the partial derivatives $(\partial f_i/\partial y^{(j)})$ have the property that there exist points α_0, $\beta_0^{(j)}$ in a suitably defined space $|x-x_0| \leqq \alpha$, $|y^{(j)}-y_0^{(j)}| \leqq \beta^{(j)}$, $j=0, 1, \cdots, N$, such that $\partial f_i/\partial y^{(i)}$ is approximately equal to a constant f_{ij} $(\alpha_0, \beta_0, \beta_0', \cdots, \beta_0^{(i-1)})$. In such a case let J_0 denote the matrix J whose elements are evaluated at α_0, $\beta_0^{(j)}$. If the characteristic roots λ of $C_0 = J_0^{-1}A$ are distinct, then by (2.7)

$$U_k = \sum_m H_m \sum_{t=0}^{k-1} \lambda_m^{k-t-1} J_0^{-1} b_t. \tag{3.7}$$

Now A has at least $N-n+1$ rows of zeros. There are thus

$(N-n+1)$ $(r+1)$ artificial characteristic roots of C_0 of value zero. There must be, further, the n roots associated with the n independent solutions of the variational equations; these roots are of the form

$$\lambda_m = 1 + \gamma_m h + \cdots, \qquad m = 1, 2, \cdots, n.$$

Thus there remain

$$(r + 1)(N + 1) - (r + 1)(N - n + 1) - n = nr$$

additional roots. These are "extraneous," introduced by the method of integration. "Extraneous" solutions of the integration problem are, consequently, solutions belonging to extraneous roots of C_0.

Now in the solution vector U_k the only components of interest are those of Ω_k, where

$$\Omega_k^T = \left[\eta_k, \eta_k', \cdots, \eta_k^{(N)}\right].$$

These may be calculated obviously from (3.7) by simply replacing b_k in (3.7) by c_k where

$$c_k^T = \left[- T + \tau, - T_{k1} + \tau_{k1}, \cdots, - T_{k,n-1} + \tau_{k,n-1}; \phi_n, \cdots, \phi_N\right],$$

accompanied by a similar contraction of the matrices $H(\lambda)$ and J_0^{-1}.

The determination of the characteristic values λ of $C_0 = J_0^{-1}A$ and the construction of the error vector may be simplified somewhat, as follows:

Let us define

$$\Gamma(\Lambda) = \Lambda I - J_a A, \tag{3.8}$$

where J, A are square matrices of orders s, and J_a denotes the adjoint of J. Similarly, let G_a denote the adjoint of $G(\lambda) = \lambda I - J^{-1}A$. Let, further,

$$\Delta(\Lambda) = \det \Gamma(\Lambda),$$

and, as before, $\delta(\lambda) = \det G(\lambda)$.

Then we have the following

LEMMA. $G_a/\delta' = \Gamma_a/\Delta'$.

PROOF. Clearly

$$\Gamma(\Lambda) = D\left[\Lambda D^{-1} - J^{-1}A\right] = DG(\Lambda/D), \tag{3.9}$$

where again $D = \det J$, $G(\lambda) = \lambda I - J^{-1}A$. Consequently,

$$\Delta(\Lambda) = D^s\delta(\Lambda/D).$$

To each root Λ of $\Delta(\Lambda)=0$ there is then associated a root

$$\lambda = \Lambda/D$$

of $\delta(\lambda)=0$. Further,

$$\delta'(\lambda) = D^{-s+1}\cdot\Delta'(\Lambda). \tag{3.10}$$

However, by (3.9),

$$\Gamma_a(\Lambda) = D^{s-1}G_a(\lambda) \tag{3.11}$$

which, together with (3.9), proves the lemma.

We have thus obtained the following general

PROPAGATION THEOREM.

$$\Omega_k = \sum_m \left[\Gamma_a(\Lambda_m)J_a/D\Delta'(\Lambda_m)\right]\sum_t \lambda_m^{k-t-1}c_t. \tag{3.12}$$

In this theorem the index m is to be extended over all distinct non-zero characteristic roots Λ_m of $\Delta(\Lambda)=0$, $\lambda_m=\Lambda_m/D(J)$, and the elements of the vector c_k are due to truncation error and rounding. The theorem shows again that for a method to be stable for sufficiently small h it is sufficient that all characteristic roots λ_m be of absolute value not exceeding unity.

The actual computation of Ω_k would then proceed in obvious fashion from the construction of J_aA and $\Delta(\Lambda)$ to the calculation of $D(J)$, Λ_m, and $\Gamma_a(\Lambda_m)\cdot J_a$, and could be carried out concurrently with the integration.

4. **The propagation of error in the case $n=N=1$.** The deductions of the previous sections will be applied now to a number of well known methods of numerical integration. We shall start by considering the general first order differential equation

$$y' = f(x, y).$$

In this case the associated homogeneous variational equation is

$$\eta' = f_y\eta;$$

it has the fundamental solution

$$\eta(x) = \exp\left(\int_{x_0}^{x} f_y dt\right).$$

Among the solutions λ of the characteristic equation $\delta(\lambda)=0$ inherent in any useful method of integration there must be one, $\lambda=\lambda_1$, that approaches this fundamental solution $\eta(x)$ as the step size h

goes to zero. It will be seen that this root λ_1 is of the form

$$\lambda_1 = 1 + hf_y + \cdots \approx \exp hf_y \equiv \exp p, \ p = hf_y;$$

it gives rise in (3.12) to the principal term of the form

$$\Omega_k(\lambda_1) = M_1\lambda_1^{k-t-1} \approx M_1 \exp\left(\int_{x_{t+1}}^{x_k} f_y dt\right).$$

In order to prevent the error in the large from increasing rapidly it is thus sufficient to carry out the integration in the direction Δx in which $f_y\Delta x$ is nonpositive.

Let us consider first the important case $n=N=1$. Here we have

$$J = \begin{bmatrix} I & -J_{01} \\ -J_{10} & I \end{bmatrix}, \quad J_{01} = \begin{bmatrix} h\, a_{1,-1} & \cdots & 0 \\ \vdots & & \vdots \\ 0 & \cdots & 0 \end{bmatrix}, \quad J_{10} = \begin{bmatrix} f_y & \cdots & 0 \\ \vdots & & \vdots \\ 0 & \cdots & 0 \end{bmatrix},$$

$$A = \begin{bmatrix} A_{00} & A_{01} \\ 0 & A_{11} \end{bmatrix}, \qquad A_{00} = \begin{bmatrix} a_{00} & a_{01} & \cdots & a_{0r} \\ 1 & 0 & \cdots & 0 \\ \vdots & & & \vdots \\ 0 & 0 & 1 & 0 \end{bmatrix},$$

$$A_{01} = \begin{bmatrix} ha_{10} & ha_{11} & \cdots & ha_{1r} \\ 0 & \cdot & \cdot & \cdot\, 0 \\ \vdots & & & \\ 0 & \cdot & \cdot & \cdot\, 0 \end{bmatrix}, \quad A_{11} = \begin{bmatrix} 0 & \cdots & 0 & 0 \\ 1 & & 0 & 0 \\ \vdots & & \vdots & \\ 0 & \cdots & 1 & 0 \end{bmatrix}.$$

It follows that

$$D = 1 - pa_1, \qquad a_1 \equiv a_{1,-1},$$

$$J_a = \begin{bmatrix} K & J_{01} \\ J_{10} & K \end{bmatrix}, \qquad K = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & D & \cdots & 0 \\ \vdots & & & \\ 0 & 0 & \cdots & D \end{bmatrix},$$

(4.1)

$$\Gamma(\Lambda) = \Lambda I - J_a A = \begin{bmatrix} \Gamma_{00} & \Gamma_{01} \\ \Gamma_{10} & \Gamma_{11} \end{bmatrix},$$

$$\Gamma_{00} = \begin{bmatrix} \Lambda - a_{00} & -a_{01} & \cdots & -a_{0,r-1} & -a_{0r} \\ -D & \Lambda & \cdots & 0 & 0 \\ \vdots & & & & \\ 0 & 0 & \cdots & -D & \Lambda \end{bmatrix},$$

$$\Gamma_{01} = \begin{bmatrix} -ha_{10} & -ha_{11} & \cdots & -ha_{1r} \\ 0 & 0 & & 0 \\ \vdots & & & \\ 0 & \cdots & \cdots & 0 \end{bmatrix},$$

$$\Gamma_{10} = -f_y \begin{bmatrix} a_{00} & a_{01} & \cdots & a_{0r} \\ 0 & & & 0 \\ \vdots & & & \\ 0 & \cdots & \cdots & 0 \end{bmatrix}, \tag{4.1}$$

$$\Gamma_{11} = \begin{bmatrix} \Lambda - pa_{10} & -pa_1 & \cdots & -pa_{1,r-1} & -pa_{1r} \\ -D & \Lambda & & 0 & 0 \\ \vdots & & & & \\ 0 & 0 & & -D & \Lambda \end{bmatrix}.$$

The characteristic equation $\Delta(\Lambda) = 0$ may be expressed in the form

$$\begin{aligned} \Delta(\Lambda) &= \Lambda^{r+1}\Delta_c(\Lambda) = 0, \\ \Delta_c(\Lambda) &= \Lambda^{r+1} - e_0\Lambda^r - e_1\Lambda^{r-1} - \cdots - e_r, \end{aligned} \tag{4.2}$$

where

$$e_\rho = D^\rho(a_{0\rho} + pa_{1\rho}), \qquad \rho = 0, 1, \cdots, r. \tag{4.3}$$

The equation for the extraneous roots Λ_m, if any, is now quite easily obtained from (4.2).

It is convenient to write the characteristic expression in the form

$$\Delta_c(\Lambda) = \Lambda^{r+1} - \Delta_{c0}(\Lambda) - p\Delta_{c1}(\Lambda),$$

where

$$\Delta_{ci}(\Lambda) = \Lambda^r a_{i0} + \Lambda^{r-1}Da_{i1} + \cdots + D^r a_{ir}, \qquad i = 0, 1. \tag{4.4}$$

Since $\Lambda_1 = D\lambda_1 = (1+p)D$ is a solution of $\Delta_c(\Lambda) = 0$, there are obtained for the coefficients a_{ij} the relationships

$$\begin{aligned} \sum_{j=0}^{r} a_{0j} &= 1, \\ a_1 + \sum_{j=0}^{r} a_{1j} - \sum_{j=0}^{r} (j)a_{0j} &= 1. \end{aligned} \tag{4.5}$$

One may show, after some lengthy calculations, that the contracted adjoint of $\Gamma(\Lambda)$ may be expressed as

$$\Gamma_a(\Lambda) = \Lambda^r \begin{bmatrix} \Delta_{c0}(\Lambda) & h\Delta_{c1}(\Lambda) \\ f_y\Delta_{c0}(\Lambda) & p\Delta_{c1}(\Lambda) \end{bmatrix},$$

or

$$\Gamma_a(\Lambda) = \Lambda^r \begin{bmatrix} 1 \\ f_y \end{bmatrix} [\Delta_{c0}, \Delta_{c1}] \begin{bmatrix} 1 & 0 \\ 0 & h \end{bmatrix}.$$

Therefore,

$$\Gamma_a(\Lambda)J_a = \Lambda^r \begin{bmatrix} 1 \\ f_y \end{bmatrix} [\Lambda^{r+1}, h(a_1\Delta_{c0}(\Lambda) + \Delta_{c1}(\Lambda))],$$

and, finally,

$$\begin{bmatrix} \eta_k \\ \eta_k' \end{bmatrix} = \frac{1}{D} \sum_m \frac{1}{\Lambda_m \Delta_c'(\Lambda_m)} [\Lambda_m^{r+1}, h(a_1\Delta_{c0} + \Delta_{c1})] \sum_t \lambda_m^{k-t-1} c_t. \tag{4.6}$$

The error η_k' may thus be obtained quite simply from η_k by multiplication with f_y.

The contribution of the root $\Lambda_1 = (1+p)D$ to the error η_k may now be written down at once; it is

$$\begin{aligned} \eta_k(\Lambda_1) &= \frac{1}{\Delta_c'(\Lambda_1)} \left[1 + p(r - (r-1)a_1), h\left(a_1 + \sum_0^r a_{1j}\right)\right] \\ &\quad \cdot \sum_t \exp\left(\int f_y dx\right) c_t. \end{aligned} \tag{4.7}$$

It is of interest to apply above deductions to some specific cases.

I. Case $r=0$. As was pointed out above no extraneous solutions arise in this case, so that the methods are stable in the direction in which $p<0$. Techniques falling into this class are due to Euler, Heun, Runge-Kutta, Milne, and others.

Since now

$$\Delta_c(\Lambda) = \Lambda - (a_{00} + pa_{10}),$$

there is obtained from (4.7) and (4.5)

$$\eta_k = [1 + pa_1, h] \sum_t \exp\left(\int_{x_{t+1}}^{x_k} f_y dx\right) c_t,$$

$$c_t = \begin{bmatrix} -T_t + \tau_t \\ \phi \end{bmatrix},$$

or

$$(4.8)\qquad \eta_k \approx \int_{x_0}^{x_k} [(-T_t + \tau_t)/h + a_1\tau_t f_y + \phi] \exp\left(\int_{x_{t+1}}^{x_k} f_y dx\right) dt.$$

I, 1. Euler's method.

$$^*y_{k+1} = {}^*y_k + h^*y'_k,$$

$$(a_{00}\ a_1\ a_{10}) = (1, 0, 1), \qquad T = (h^2/2)y''.$$

$$\eta_k \approx \int [-(h/2)y''_t + \tau_t/h + \phi] \exp\left(\int f_y dx\right) dt.$$

I, 2. Heun's (modified Euler) method.

$$^*y_{k+1} = {}^*y_k + (h/2)({}^*y'_k + {}^*y'_{k+1}),$$

$$(a_{00}\ a_1\ a_{10}) = (1, 1/2, 1/2),$$

$$T = -(h^3/12)y'''.$$

$$\eta_k \approx [-1 + p/2, h] \sum_t \exp\left(\int f_y dx\right) c_t.$$

II. Case $r=1$. There is one extraneous root Λ_2; it satisfies the equation

$$(4.9)\qquad \Delta_c(\Lambda) \equiv \Lambda^2 - \Delta_{c0} - p\Delta_{c1} = 0$$

where

$$\Delta_{c0}(\Lambda) = \Lambda a_{00} + Da_{01}, \qquad \Delta_{c1}(\Lambda) = \Lambda a_{10} + Da_{11}.$$

Since $\Lambda_1 + \Lambda_2 = a_{00} + pa_{10}$, and a_{ij} satisfy (4.5), it follows that

$$(4.10)\qquad \Lambda_2 = -a_{01} + p(a_{01} - a_{11}).$$

Further

$$\Delta'_c(\Lambda_1) = (2 - a_{00}) + p[2(1 - a_1) - a_{10}], \qquad \Delta'_c(\Lambda_2) = -\Delta'_c(\Lambda_1).$$

II, 1. Simple central difference method.

$$^*y_{k+1} = {}^*y_{k-1} + 2h^*y'_k,$$

$$(a_{00}\ a_{01}\ a_1\ a_{10}\ a_{11}) = (0, 1, 0, 2, 0),$$

$$T = (h^3/3)y'''.$$

Thus

$$D(J) = 1,$$

$$\Lambda_m = \pm 1 + p, \qquad m = 1, 2,$$

$$\Delta_c'(\Lambda_1) = 2,$$

and, consequently,

$$(4.11)\qquad \eta_k = \frac{1}{2}\left\{[1 + p, 2h]\sum_t \exp\left(\int f_y dx\right) + [-1 + p, 2h]\sum_t (-1)^{k-t} \exp\left(-\int f_y dx\right)\right\} c_t.$$

The extraneous root Λ_2 may thus give rise to an oscillating term of increasing magnitude whenever the integration is applied in a direction in which $hf_y<0$. However, it is entirely possible that the actual increase of this term is choked off by the rounding procedure itself. See footnote 1. Used for integrations in the opposite direction, over short ranges, the method may give useful results.

II, 2. Simpson's method. Here

$$^*y_{k+1} = {}^*y_{k-1} + (h/3)(^*y'_{k+1} + 4^*y'_k + {}^*y'_{k-1}),$$

whence

$$(a_{00}\ a_{01}\ a_1\ a_{10}\ a_{11}) = (0, 1, 1/3, 4/3, 1/3),$$
$$T = -(h^5/90)y^V.$$

It follows that

$$D(J) = 1 - p/3, \qquad \Lambda_m = \pm 1 + 2p/3, \qquad \Delta_c'(\Lambda_1) = 2.$$

Therefore,

$$(4.12)\qquad \eta_k = \frac{1}{2}\left\{[1 + p, 2h]\sum_t \exp\left(\int f_y dx\right) + [-1 + p/3, 2h/3]\sum_t (-1)^{k-t} \cdot\exp\left(-\int f_y dx/3\right)\right\} c_t.$$

The second root Λ_2 may thus again make this integration method unsuitable.

III. Case $r=2$. The three roots Λ_m, $m=1, 2, 3$, satisfy the characteristic equation

$$\Delta_c(\Lambda) \equiv \Lambda^3 - \Delta_{c0}(\Lambda) - p\Delta_{c1}(\Lambda) = 0,$$

with

$$\Delta_{ci}(\Lambda) = \Lambda^2 a_{i0} + \Lambda D a_{i1} + D^2 a_{i2}, \qquad i = 0, 1.$$

The 2 extraneous roots Λ_2, Λ_3 may thus be obtained from

$$(4.13)\quad \begin{aligned} \Lambda^2 - \Lambda[(a_{00} - 1) + p(a_1 + a_{10} - 1)] \\ - D[- a_{02} + p(a_{02} - a_{12})] = 0. \end{aligned}$$

III, 1. Adams method.

$$*y_{k+1} = *y_k + h[*y_k' + (1/2)\nabla *y_k' + (5/12)\nabla'' *y_k'] + \cdots$$

with

$$\nabla^{(i+1)}q_k = \nabla^{(i)}q_k - \nabla^{(i)}q_{k-1}.$$

Thus, alternately,

$$*y_{k+1} = *y_k + (h/12)[23*y_k' - 16*y_{k-1}' + 5*y_{k-2}'] + \cdots .$$

Then,

$$(a_{00}, a_{01}, a_{02}, a_1, a_{10}, a_{11}, a_{12}) = (1/12)(12, 0, 0, 0, 23, -16, 5).$$

The 2 extraneous roots Λ_2, Λ_3 satisfy

$$\Lambda^2 - (11p/12)\Lambda + (5p/12) = 0.$$

Therefore,

$$\Lambda_m = \pm (- 5p/12)^{1/2} + 11p/24, \qquad m = 2, 3.$$

For sufficiently small h this method, then, is stable, and the propagation of error depends essentially on Λ_1.

Since

$$\Delta_c'(\Lambda_1) = 1 + 3p/2,$$

it is found that

$$(4.14)\qquad \eta_k = [1 + p/2, h] \sum_t \exp\left(\int f_y dx\right) c_t.$$

III, 2. Gregory's method. Here

$$\begin{aligned} *y_{k+1} &= *y_k + h[*y_{k+1} - (1/2)\nabla' *y_{k+1} - (1/12)\nabla'' *y_{k+1}' \\ &\qquad - (1/24)\nabla''' *y_{k+1}'] + \cdots \\ &= *y_k + (h/24)[9*y_{k+1}' + 19*y_k' - 5*y_{k-1}' + *y_{k-2}'] + \cdots . \end{aligned}$$

Thus

$$(a_{00}\ a_{01}\ a_{02}\ a_1\ a_{10}\ a_{11}\ a_{12}) = (1/24)(24, 0, 0, 9, 19, -5, 1),$$
$$T = - (19/720)h^5 y^v.$$

By (4.13), then, $D=1-3p/8$,

$$\Lambda^2 - (p/6)\Lambda + p/24 = 0,$$

so that

$$\Lambda_m = \pm (-p/24)^{1/2} + p/12, \qquad m = 2, 3.$$

The method thus has the same stability properties as Adams' method.

Since again

$$\Delta_c'(\Lambda_1) = 1 + 3p/2,$$

one obtains from (4.7)

$$\eta_k = [1 + p/8, h]\sum_t \exp\left(\int f_y dx\right) c_t. \tag{4.15}$$

5. **Numerical example.** To test the propagation theorem let us integrate the differential equation

$$y' = y - 2x/y \tag{5.1}$$

by means of Simpson's method II, 2. Taking $h=0.5$ and starting at $x=0$ with values computed from the exact solution

$$y(x) = (2x + 1)^{1/2},$$

there are obtained the "solutions" shown in columns (2) and (3) of Table I. At each step a sufficient number of iterations is carried out in order to achieve agreement to five decimals (column (2)), or four decimals (column (3)). Due to the instability of the method the five-decimal "solution" diverges more and more from the four-decimal "solution."

The fourth column (4) contains the exact solution $y(x)$, and the fifth column (5) the error $\eta = {}^*y - y(x)$, the solution ${}^*y$ taken from column (3).

The growth of error may be inferred from (4.12), or, somewhat more accurately, from

$$\eta_k = \eta_k(\lambda_1) + \eta_k(\lambda_2),$$

$$\eta_k(\lambda_1) \approx \frac{1}{2}\sum_{t=0}^{k-1} [(1 + p)((h^5/90)y_t^{v} + \tau_t) + 2h\phi]\lambda_1^{k-t-1}, \tag{5.2}$$

$$\eta_k(\lambda_2) \approx -\frac{1}{2}\sum_{t=0}^{k-1} [(-1 + p/3)((h^5/90)y_t^{v} + \tau_t) + (2/3)h\phi]\lambda_2^{k-t-1}. \tag{5.3}$$

TABLE I. INTEGRATION OF $y'=y-2x/y,\ h=0.5$

(1)	(2)	(3)	(4)	(5)
x	$^*y$	$^*y$	$y(x)$	η
0	1.00000	1.0000	1.0000	0
.5	1.41421	1.4142	1.4142	0
1.0	1.73516	1.7352	1.7320	.0032
1.5	2.00529	2.0053	2.0000	.0053
2.0	2.25064	2.2507	2.2361	.0146
2.5	2.48438	2.4845	2.4495	.0350
3.0	2.73397	2.7342	2.6458	.0884
3.5	3.04775	3.0483	2.8284	.220
4.0	3.53706	3.5383	3.0000	.538
4.5	4.42054	4.4232	3.1623	1.26
5.0	6.07815	6.0834	3.3166	2.77
5.5	9.08576	9.0953	3.4641	5.63
6.0	14.31274	14.3292	3.6056	10.7
6.5	23.14506	23.1727	3.7417	19.4
7.0	37.86292	37.9089	3.8730	34.0
7.5	62.23708	62.3131	4.0000	58.3
8.0	102.49977	102.6253	4.1231	98.5
8.5	168.93852	169.1431	4.2426	165
9.0	278.52584	278.8552	4.3589	274
9.5	459.25450	459.8020	4.4721	455
10.0	757.28847	758.1877	4.5826	754

For our example,

$$k = 20, \qquad h = 1/2, \qquad p \equiv hf_y = 1 - [2(2x+1)]^{-1},$$
$$y^v = 105(2x+1)^{-9/2},$$
$$|\tau|, \quad |\phi| = c_1 10^{-5}, \qquad c_1 < 10,$$

and

$$\lambda_1 = 1 + p.$$

Now p increases from 0.5 at $x=0$ to 0.98 at $x=20$, so that $p\approx 0.7$ could be taken as average value. Furthermore, for the terms corresponding to the low values of x, which contribute most to $\eta_k(\lambda_1)$, τ and ϕ are negligible. Thus, by (5.2),

$$\eta_k(\lambda_1) \approx \frac{1}{2}\sum_{t=0}^{19}(1/2880)\overset{v}{y}_t(1.7)^{20-t}. \tag{5.4}$$

One obtains

$$\eta_k(\lambda_1) \approx 762.$$

Since $\lambda_2 = -1 + p/3$, so that

$$\eta_k(\lambda_2) \approx -\frac{1}{2}\sum_{t=0}^{19} (1/2880) y_t^v (-0.77)^{20-t},$$

and consequently negligible, we get

$$\eta_k \approx \eta_k(\lambda_1) \approx 762,$$

which compares very favorable indeed with the actual value of 754 for the total error.

In order to examine the oscillating term $\eta_k(\lambda_2)$, let us integrate again (5.1), this time starting at $x=60$, and taking $h=-1$. The "solution" is shown in column (2) of Table 2.

Now,

$$\begin{aligned} k &= 34, \qquad h = -1, \\ p &= -2 + (2x+1)^{-1} \approx -2, \\ |\tau|&, \qquad |\phi| = c_2 \cdot 10^{-6}, \qquad\qquad c_2 \approx 2, \\ \lambda_2 &= -1 + p/3 \approx -1.66. \end{aligned}$$

Since for low values of t the term $(h^5/90)y_v^t$ is less then 1.10^{-7}, in absolute value, there is obtained the expression

$$|\eta_k(\lambda_2)| \approx 10^{-6} \sum_{t=0}^{33} (-1 + p/3)^{34-t}. \tag{5.5}$$

This formula leads to

$$|\eta_k(\lambda_2)| \approx 19.0,$$

which is very close indeed to the exact error given in Table 2.

The exhibited expressions, then, lead to quite useful estimates of the error.

References

1. L. H. Thomas, *Stability of solution of partial differential equations*, Symposium on Theoretical Compressible Flow, U. S. Naval Ordnance Laboratory, 1949, pp. 83–94.

2. H. Rutishauser, *Über die Instabilität von Methoden zur Integration von Differentialgleichungen*, ZAMP vol. 3 (1952) pp. 65–74.

3. H. Rademacher, *On the accumulation of errors in processes of integration*, Proceedings, Symposium on Large-Scale Digital Calculating Machines, Harvard Computation Laboratory, 1948, pp. 176–185.

4. R. A. Frazer, W. J. Duncan, A. R. Collar, *Elementary matrices*, Cambridge University Press, 1950, p. 78.

TABLE 2. INTEGRATION OF $y'=y-2x/y$, $h=-1$

(1)	(2)	(3)	(4)
x	$^*y$	$y(x)$	$10^5\eta$
60	11.00000	11.00000	0
59	10.90871	10.90871	0
58	10.81665	10.81665	0
57	10.72381	10.72381	0
56	10.63014	10.63015	−1
55	10.53566	10.53565	1
54	10.44030	10.44031	−1
53	10.34409	10.34408	1
52	10.24694	10.24695	−1
51	10.14891	10.14889	2
50	10.04984	10.04988	−4
49	9.94994	9.94987	7
48	9.84875	9.84886	−11
47	9.74698	9.74679	19
46	9.64333	9.64365	−32
45	9.53994	9.53939	55
44	9.43304	9.43398	−94
43	9.32899	9.32738	161
42	9.21679	9.21954	−275
41	9.11515	9.11043	472
40	8.99193	9.00000	−807
39	8.90203	8.88819	1384
38	8.75130	8.77496	−2366
37	8.70087	8.66025	4062
36	8.47474	8.54400	−6926
35	8.54560	8.42615	11945
34	8.10464	8.30662	−20198
33	8.53865	8.18535	35330
32	7.47987	8.06226	−58239
31	9.00031	7.93725	106306
30	6.19044	7.81025	−161981
29	11.03789	7.68115	335674
28	3.61153	7.54983	−393830
27	19.42204	7.41620	1200584
26	−12.03925	7.28011	−1931936

RCA SERVICE COMPANY, PATRICK AIR FORCE BASE