J. Sci. Hiroshima Univ. Ser. A-I
25 (1961), 3-62

Theory of Errors in Numerical Integration of Ordinary Differential Equations [1)]

Minoru Urabe

(Received March 3, 1961)

Preface

Concerning errors in numerical integration of ordinary differential equations, there are three problems.

The first one is the problem of stable convergence, namely the problem of determining necessary and sufficient conditions that, for sufficiently small length of divided intervals, approximate solutions can be actually obtained by numerical integration in any finite interval where the true solution exists; and moreover, as the length of divided intervals tends to zero, these approximate solutions converge to the true solution in that interval provided all round-off errors including the errors of starting values tend to zero in a suitable manner. In this paper, we say that an integration formula is *stable* if it satisfies the above conditions. To the problem of stable convergence, so far as the author knows, an almost complete answer has been given first by G. Dahlquist [3, 4][2)] for general multi-step integration formulas. Of course, before him, the problem has been studied by many scholars, for instance, by J. Todd [14], H. Rutishauser [13], and F. B. Hildebrand [8]. But, by all of these, it has been assumed that the initial differential equations, given in the canonical form, are linear in the unknown functions with constant coefficients and moreover, even for such equations, the treatment of the problem has been illustrative rather than demonstrative. Dahlquist, on the contrary, has derived necessary conditions for general differential equations that a general multi-step integration formula may be stable and, after that, he has proved that, for any multi-step integration formula satisfying the necessary conditions derived, there actually exist numerical solutions satisfying that multi-step integration formula with any prescribed accuracy and that the numerical solutions obtained actually converge to the true solution as the length of divided intervals and the sum of round-off errors in all steps tend to zero. But he has not proved that, by means of the multi-step integration formulas satisfying his necessary conditions, for sufficiently small length of divided intervals, the numerical approximate solutions can be actually constructed in

1) Mainly sponsored by the United States Army under Contract No. DA-11-022-ORD-2059, Mathematics Research Center, United States Army, Madison, Wisconsin.

2) The numbers in brackets refer to the references listed at the end of the paper.

a finite interval where the true solution exists. Indeed this fact does not follow from the proof of Dahlquist, because his proof has been irrelevant to actual construction of numerical solutions. The proof of Dahlquist, afterwards, has been simplified and applied to proof of existence theorems of solutions for differential equations by T. E. Hull and W. A. J. Luxemburg [9]. But, in their paper, the domain of definition of differential equations has been assumed to be so broad that the numerical solutions can be always actually constructed. This defect, however, can be easily removed by extending the initial domain of definition to the broader one by the method used often in the theory of differential equations.

In the present paper, after removing the above defect by the method mentioned, the problem of stable convergence is completely solved by a method quite different from those of Dahlquist or Hull and Luxemburg, and, moreover, it is done in a unified form for three representative integration formulas including not only the usual multi-step formulas but also compound multi-step formulas[1] and general Runge-Kutta formulas [6, 5].

The second of the problems is that of *propagation of errors*, namely behavior of growth of errors as the steps advance. While, in the problem of stable convergence, the length of divided intervals is considered as a variable tending to zero, in the problem of propagation of errors it is considered as a fixed quantity. This problem has been studied separately in particular cases by many scholars, for instance, by J. Todd [14], H. Rutishauser [13], M. Lotkin [10], L. Collatz [1], F. B. Hildebrand [8], G. Dahlquist [4], R. W. Hamming [7], W. E. Milne and R. R. Reynolds [11, 12], H. S. Wilf [19], etc.

In the present paper, in the same manner as in the first problem, the problem is studied for general differential equations in a unified form for the three integration formulas mentioned above. By the results of the present paper, the behavior of growth of errors is made clear rigorously though most of them are already known heuristically.

The last problem is the problem of *estimation of errors*. Estimates of errors in terms of Lipschitz constants [1, 8] are well known and they are usually derived from the difference equations which are satisfied by the errors of numerical solutions. But, as is well known, these estimates of errors are too crude for practical use. Hence, it has been long necessary to get better estimates of errors. Recently W. Uhlmann [15, 16] proposed a new method, namely to obtain estimates of errors from the differential equations which are satisfied by the errors of the continuously differentiable approximate solutions obtained from discrete numerical solutions by interpolation. By this method, he could obtain better estimates of errors. But, as Dahlquist stated [4], Uhlmann's method seems not to be suited to general integration formulas, for example, to general multi-step formulas though it is very satisfactory for

1) For compound multi-step formulas, see **1.2.**

Adams' formulas or Runge-Kutta formulas, etc. Afterwards, Dahlquist [4] derived a new estimation formula for a particular multi-step integration formula, but his estimation formula is confined to a very particular formula and, in addition, is very complicated. On the other hand, for Runge-Kutta formulas including the general one, new estimation formulas have been obtained by J. W. Carr III [2], B. A. Galler and D. P. Rozenberg [5]. But, of course, these are confined only to Runge-Kutta formulas.

In the present paper, a new estimate of errors is derived also in a unified form for the three integration formulas mentioned in the beginning. And, for estimation of errors, there are used quantities like

$$\mu_P^{\pm}[Q]=\lim_{h\to\pm 0}\frac{|P+hQ|-|P|}{h},$$

where P and Q are matrices and $|\cdots|$ denotes the norm of the matrix. The above quantities are generalizations of the quantities introduced by Dahlquist [4]. The estimate of errors which is obtained has properties analogous to those obtained by the above people and is fairly better than the classical estimates of errors in terms of Lipschitz constants.

In the present paper, there are first derived the difference equations which are satisfied by the errors of the numerical solutions obtained by the three integration formulas stated in the beginning. These equations are then rewritten as a simultaneous system of equations of first order. In the sequel, by analysis of this system by means of the theory of matrices, the three problems concerning errors are discussed.

Chapter I. Difference equations for errors

1.1 Difference equations for errors of usual multi-step integration formulas

A usual multi-step integration formula can be written as

$$x_{n+k}=\alpha_1 x_{n+k-1}+\alpha_2 x_{n+k-2}+\cdots+\alpha_k x_n+h(\beta_0 \dot{x}_{n+k}+\beta_1 \dot{x}_{n+k-1}+\cdots+\beta_k \dot{x}_n)\,, \tag{1.1}$$

where h is a length of the divided intervals and $\dot{x}$ is the derivative of x with respect to the independent variable t.

Let us derive the difference equations which the errors satisfy when the above formula is applied to the N-dimensional differential system

$$\frac{dx}{dt}=f(x,\ t). \tag{1.2}$$

For this equation, it is assumed that this equation has a solution $x=x(t)$ in the interval $[t_0-L,\ t_0+L]$ and that, in the domain

$$D:\ |t-t_0|\leqq L,\quad |x-x(t)|\leqq r\ ^{1)},$$

$f(x,\ t)$ is continuous in $(x,\ t)$ and is continuously differentiable in x.

Put

$$x_i=x(t_i)=x(t_0+i\,h)$$

and let $\tilde{x}_i$ be the approximate values of $x_i=x(t_i)$ computed by (1.1) in the domain D. Then evidently

$$\begin{aligned} x_{n+k}=&\alpha_1\ x_{n+k-1}+\alpha_2\ x_{n+k-2}+\cdots+\alpha_k\ x_n\\ &+h\{\beta_0\ f(x_{n+k},\ t_{n+k})+\beta_1\ f(x_{n+k-1},\ t_{n+k-1})+\cdots+\beta_k\ f(x_n,\ t_n)\}+T_n, \end{aligned} \tag{1.3}$$

$$\begin{aligned} \tilde{x}_{n+k}=&\alpha_1\ \tilde{x}_{n+k-1}+\alpha_2\ \tilde{x}_{n+k-2}+\cdots+\alpha_k\ \tilde{x}_n\\ &+h\{\beta_0\ f(\tilde{x}_{n+k},\ t_{n+k})+\beta_1\ f(\tilde{x}_{n+k-1},\ t_{n+k-1})+\cdots+\beta_k\ f(\tilde{x}_n\ t_n)\}+R_n, \end{aligned} \tag{1.4}$$

where T_n and R_n are respectively the truncation and round-off errors. Put

$$\tilde{x}_i-x_i=e_i \qquad (i=0,\ 1,\ 2,\cdots). \tag{1.5}$$

Then evidently the e_i express the errors of the approximate solution computed by the multi-step formula (1.1).

Subtracting (1.3) from (1.4), there is obtained

---

1) In the present paper, for norms of vectors and matrices, the following definitions are adopted:

$$|v|=\max_{\nu}|v^{\nu}|,\quad |A|=\max_{\nu}\sum_{\mu}|a_{\mu}^{\nu}|,$$

where $v=(v^{\nu})$ or $A=(a_{\mu}^{\nu})$ is an arbitrary vector or matrix respectively.

$$(1.6)\qquad \begin{aligned} e_{n+k}=&\alpha_1\, e_{n+k-1}+\alpha_2\, e_{n+k-2}+\cdots+\alpha_k\, e_n\\ &+h\big[\beta_0\{f(\tilde{x}_{n+k},\, t_{n+k})-f(x_{n+k},\, t_{n+k})\}\\ &\quad+\cdots+\beta_k\{f(\tilde{x}_n,\, t_n)-f(x_n,\, t_n)\}\big]+R_n-T_n. \end{aligned}$$

But, from the continuous differentiability of $f(x,\, t)$ and continuity of the solution $x=x(t)$, the quantities

$$f(\tilde{x}_i,\, t_i)-f(x_i\; t_i)\qquad (i=n,\, n+1,\cdots,\, n+k)$$

can be written as follows:

$$f(\tilde{x}_i,\, t_i)-f(x_i,\, t_i)=F(x_n,\, t_n)\, e_i+\Phi_{n,\,i-n}e_i\qquad (i=n,\, n+1,\cdots,\, n+k),$$

where $F(x,\, t)=(F_\mu^\nu)$ is the Jacobian matrix of $f(x,\, t)$ with respect to x. Since

$$(1.7)\qquad \Phi_{n,\,i-n}=\int_0^1 F(x_i+\theta e_i,\, t_i)\,d\theta-F(x_n,\, t_n),$$

it is evident that

$$(1.8)\qquad \Phi_{n,\,i-n}=o(1)\quad \text{uniformly as } |h|,\ |e_i|\to 0.$$

Then (1.6) can be rewritten as follows:

$$(1.9)\qquad \begin{aligned} e_{n+k}=&\alpha_1\, e_{n+k-1}+\alpha_2\, e_{n+k-2}+\cdots+\alpha_k\, e_n\\ &+h(\beta_0\, F_n\, e_{n+k}+\beta_1\, F_n\, e_{n+k-1}+\cdots+\beta_k\, F_n\, e_n)\\ &+h(\beta_0\, \Phi_{n,k}\, e_{n+k}+\beta_1\, \Phi_{n,k-1}\, e_{n+k-1}+\cdots+\beta_k\, \Phi_{n,0}\, e_n)\\ &+R_n-T_n, \end{aligned}$$

where $F_n=F(x_n,\, t_n)=(F_{n\,\mu}^{\ \nu})$. Since $|F_n|$ is bounded in D, evidently

$$\det\,(I-h\beta_0\, F_n)\neq 0\ ^{1)}$$

for sufficiently small $|h|$. Then (1.9) can be rewritten further as follows:

$$(1.10)\qquad \begin{aligned} e_{n+k}=&\{\alpha_1+h(\beta_0\alpha_1+\beta_1)F_n\}e_{n+k-1}+\{\alpha_2+h(\beta_0\alpha_2+\beta_2)F_n\}e_{n+k-2}+\cdots\\ &+\{\alpha_k+h(\beta_0\alpha_k+\beta_k)F_n\}e_n\\ &+h(\Psi_{n,k}\, e_{n+k}+\Psi_{n,k-1}\, e_{n+k-1}+\cdots+\Psi_{n,0}\, e_n)\\ &+S_n, \end{aligned}$$

where

$$(1.11)\qquad \begin{cases} \Psi_{n.\,k}=(I-h\,\beta_0\, F_n)^{-1}\ \beta_0\, \Phi_{n,k}, & \\ \Psi_{n,k-l}=(I-h\,\beta_0\, F_n)^{-1}\ \beta_l\, \Phi_{n,k-l} & \\ \qquad\qquad +\dfrac{1}{h}\big[(I-h\beta_0\, F_n)^{-1}(\alpha_l+h\beta_l\, F_n)-\{\alpha_l+h(\beta_0\,\alpha_l+\beta_l)F_n\}\big] & \\ \qquad =(I-h\beta_0\, F_n)^{-1}\beta_l\, \Phi_{n,k-l}+O(|h|)\qquad (l=1,\, 2,\cdots,\, k), & \\ S_n=(I-h\beta_0\, F_n)^{-1}(R_n-T_n). & \end{cases}$$

1) I denotes the unit matrix.

(1.10) is a system of difference equations of order k. For simplicity of handling, let us transform (1.10) into a system of equations of order one.

For this purpose, we put

$$e_n^\nu = e_n^{\nu 1},\ e_{n+1}^\nu = e_n^{\nu 2},\cdots,\ e_{n+k-1}^\nu = e_n^{\nu k} \qquad (\nu = 1, 2,\cdots, N) \tag{1.12}$$

and consider the k-dimensional vector

$$\boldsymbol{e}_n^\nu = \begin{pmatrix} e_n^{\nu 1} \\ e_n^{\nu 2} \\ \vdots \\ e_n^{\nu k} \end{pmatrix} \tag{1.13}$$

and the kN-dimensional vector

$$\boldsymbol{e}_n = \begin{pmatrix} \boldsymbol{e}_n^1 \\ \boldsymbol{e}_n^2 \\ \vdots \\ \boldsymbol{e}_n^N \end{pmatrix}. \tag{1.14}$$

Further let us introduce the notations

$$\underset{n}{A_{\mu l}^\nu}(h) = \delta_\mu^\nu\, \alpha_l + h(\beta_0\, \alpha_l + \beta_l)\underset{n}{F_\mu^\nu}\ ^{1)}, \tag{1.15}$$

and consider the matrices

$$A_n(h) = \begin{pmatrix} \underset{n}{B_1^1} & \underset{n}{B_2^1} & \cdots & \underset{n}{B_N^1} \\ \vdots & \vdots & & \vdots \\ \underset{n}{B_1^N} & \underset{n}{B_2^N} & \cdots & \underset{n}{B_N^N} \end{pmatrix}, \tag{1.16}$$

where $\underset{n}{B_\mu^\nu}$ $(\nu, \mu = 1, 2,\cdots, N)$ are matrices of order k such that

$$\underset{n}{B_\mu^\nu} = \begin{pmatrix} 0 & 0 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 0 \\ \underset{n}{A_{\mu k}^\nu}(h) & \underset{n}{A_{\mu k-1}^\nu}(h) & \cdots & \underset{n}{A_{\mu 1}^\nu}(h) \end{pmatrix} \quad \text{for } \nu \neq \mu \tag{1.17}$$

and

$$\underset{n}{B_\nu^\nu} = \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & & \vdots \\ \vdots & \vdots & \ddots\ddots & \ddots\ddots & 0 \\ 0 & 0 & \cdots & 0 & 1 \\ \underset{n}{A_{\nu k}^\nu}(h) & \underset{n}{A_{\nu k-1}^\nu}(h) & \cdots & \underset{n}{A_{\nu 2}^\nu}(h) & \underset{n}{A_{\nu 1}^\nu}(h) \end{pmatrix} \tag{1.18}$$

---

1) δ_μ^ν is the Kronecker delta.

Then (1.10) can be expressed in a simple form as follows:

$$\boldsymbol{e}_{n+1}=A_n(h)\,\boldsymbol{e}_n+h\boldsymbol{\varphi}_n(\boldsymbol{e}_{n+1},\,\boldsymbol{e}_n)+\boldsymbol{s}_n, \tag{1.19}$$

where

$$\boldsymbol{\varphi}_n(\boldsymbol{e}_{n+1},\,\boldsymbol{e}_n)=\begin{pmatrix}\boldsymbol{\varphi}_n^1\\ \boldsymbol{\varphi}_n^2\\ \vdots\\ \boldsymbol{\varphi}_n^N\end{pmatrix},\qquad \boldsymbol{s}_n=\begin{pmatrix}\boldsymbol{s}_n^1\\ \boldsymbol{s}_n^2\\ \vdots\\ \boldsymbol{s}_n^N\end{pmatrix}. \tag{1.20}$$

Here $\boldsymbol{\varphi}_n^\nu$ and $\boldsymbol{s}_n^\nu$ $(\nu=1, 2,\cdots, N)$ are the k-dimensional vectors of which the first $(k-1)$ components are all zero and the k-th components are respectively the ν-th components of the quantities

$$\Psi_{n,k}e_{n+k}+\Psi_{n,k-1}e_{n+k-1}+\cdots+\Psi_{n,0}e_n \text{ and } S_n. \tag{1.21}$$

From (1.11), it is evident that

$$\boldsymbol{\varphi}_n(\boldsymbol{e}_{n+1},\,\boldsymbol{e}_n)=o(|\boldsymbol{e}_{n+1}|+|\boldsymbol{e}_n|)\qquad \text{uniformly as } |\boldsymbol{e}_{n+1}|,\ |\boldsymbol{e}_n|,\ |h|\to 0. \tag{1.22}$$

Let us assume that rounding is done always so that

$$R_n=o(|h|)\qquad \text{uniformly as } |h|\to 0. \tag{1.23}$$

Then, since it can be always assumed that

$$T_n=o(|h|)\qquad \text{uniformly as } |h|\to 0, \tag{1.24}$$

we see that

$$|\boldsymbol{s}_n|=o(|h|)\qquad \text{uniformly as } |h|\to 0.$$

Then (1.19) can be rewritten as follows:

$$\boldsymbol{e}_{n+1}=A_n(h)\,\boldsymbol{e}_n+h\,\boldsymbol{\varphi}_n(\boldsymbol{e}_{n+1},\,\boldsymbol{e}_n)+h\,\boldsymbol{r}_n, \tag{1.25}$$

where $\boldsymbol{r}_n$ are the quantities such that

$$|\boldsymbol{r}_n|=o(1)\qquad \text{uniformly as } |h|\to 0. \tag{1.26}$$

(1.25) is the equation of the desired form which the errors satisfy when the multi-step integration formula (1.1) is applied to the N-dimensional differential system (1.2).

1.2 Difference equations for errors of compound multi-step integration formulas

Previously, in order to obtain the periodic solutions of van der Pol's equation as accurately as possible, the author devised a new method of nu-

merical integration by combining the integrated Stirling's interpolation formula with the ordinary Adams' extrapolation formula [18]. Though the computation is a little more complicated than usual multi-step formulas, as is shown in the computation of solutions of van der Pol's equation in the paper [18], this new method is also brought into consideration even though it is not yet popular.

The formula in question can be written in its general form as follows:

$$(1.27)\quad \begin{cases} x_{n+k}=\alpha_1 x_{n+k-1}+\alpha_2 x_{n+k-2}+\cdots+\alpha_{k-1}x_{n+1} \\ \qquad +h(\beta_{-1}\hat{\dot{x}}_{n+k+1}+\beta_0\dot{x}_{n+k}+\beta_1\dot{x}_{n+k-1}+\cdots+\beta_{k-1}\dot{x}_{n+1}), \\ \hat{x}_{n+k+1}=\hat{\alpha}_0\, x_{n+k}+\hat{\alpha}_1\, x_{n+k-1}+\cdots+\hat{\alpha}_{k-1}\, x_{n+1} \\ \qquad +h(\hat{\beta}_0\,\dot{x}_{n+k}+\hat{\beta}_1\,\dot{x}_{n+k-1}+\cdots+\hat{\beta}_k\,\dot{x}_n), \end{cases}$$

where $\hat{\dot{x}}_{n+k+1}$ and $\hat{x}_{n+k+1}$ aré respectively the subsidiary approximate values of $\dot{x}$ and x for $t=t_{n+k+1}$.

As in the preceding paragraph, let us apply the above formula to the equation (1.2) and let $\tilde{x}_i$ be the obtained approximate values of x_i—the true values of the solution of (1.2) for $t=t_i=t_0+i\,h$ [1]. Then, as in (1.3) and (1.4),

$$(1.28)\quad \begin{cases} x_{n+k}=\alpha_1\, x_{n+k-1}+\alpha_2\, x_{n+k-2}+\cdots+\alpha_{k-1}\, x_{n+1} \\ \qquad +h\{\beta_{-1}f(x_{n+k+1},\, t_{n+k+1})+\beta_0 f(x_{n+k},\, t_{n+k})+\cdots+\beta_{k-1}f(x_{n+1},\, t_{n+1})\} \\ \qquad +T_n, \\ \hat{x}_{n+k+1}=\hat{\alpha}_0\, x_{n+k}+\hat{\alpha}_1\, x_{n+k-1}+\cdots+\hat{\alpha}_{k-1}\, x_{n+1} \\ \qquad +h\{\hat{\beta}_0 f(x_{n+k},\, t_{n+k})+\hat{\beta}_1 f(x_{n+k-1},\, t_{n+k-1})+\cdots+\hat{\beta}_k f(x_n,\, t_n)\} \\ \qquad +\hat{T}_n; \end{cases}$$

$$(1.29)\quad \begin{cases} \tilde{x}_{n+k}=\alpha_1\,\tilde{x}_{n+k-1}+\alpha_2\,\tilde{x}_{n+k-2}+\cdots+\alpha_{k-1}\,\tilde{x}_{n+1} \\ \qquad +h\{\beta_{-1}f(\hat{x}_{n+k+1},\, t_{n+k+1})+\beta_0 f(\tilde{x}_{n+k},\, t_{n+k})+\cdots+\beta_{k-1}f(\tilde{x}_{n+1},\, t_{n+1})\} \\ \qquad +R_n, \\ \hat{x}_{n+k+1}=\hat{\alpha}_0\,\tilde{x}_{n+k}+\hat{\alpha}_1\tilde{x}_{n+k-1}+\cdots+\hat{\alpha}_{k-1}\,\tilde{x}_{n+1} \\ \qquad +h\{\hat{\beta}_0 f(\tilde{x}_{n+k},\, t_{n+k})+\hat{\beta}_1 f(\tilde{x}_{n+k-1},\, t_{n+k-1})+\cdots+\hat{\beta}_k f(\tilde{x}_n,\, t_n)\} \\ \qquad +\hat{R}_n\ ^{2)}, \end{cases}$$

where $(T_n,\ \hat{T}_n)$ and $(R_n,\ \hat{R}_n)$ are respectively the truncation and round-off errors. Put

$$(1.30)\qquad \tilde{x}_i-x_i=e_i,\quad \hat{x}_i-x_i=\hat{e}_i \qquad (i=0,\ 1,\ 2,\cdots).$$

Then evidently the e_i are the errors of the approximate solution computed by

1) Here, of course, $(\tilde{x},\ t_i)$ are assumed to lie in the domain D.
2) Here, of course, $(\hat{x}_i,\ t_i)$ are assumed to lie in the domain D.

the formula (1.27) and the $\hat{e}_i$ are the errors of the subsidiary approximate values $\hat{x}_i$.

Subtracting (1.28) from (1.29), there are obtained

$$
(1.31)\quad \begin{cases} e_{n+k}=\alpha_1 e_{n+k-1}+\alpha_2 e_{n+k-2}+\cdots+\alpha_{k-1} e_{n+1} \\ \qquad +h[\beta_{-1}\{f(\hat{x}_{n+k+1}, t_{n+k+1})-f(x_{n+k+1}, t_{n+k+1})\} \\ \qquad\quad +\beta_0\{f(\tilde{x}_{n+k}, t_{n+k})-f(x_{n+k}, t_{n+k})\}+\cdots \\ \qquad\qquad +\beta_{k-1}\{f(\tilde{x}_{n+1}, t_{n+1})-f(x_{n+1}, t_{n+1})\}]+R_n-T_n, \\ \hat{e}_{n+k+1}=\hat{\alpha}_0 e_{n+k}+\hat{\alpha}_1 e_{n+k-1}+\cdots+\hat{\alpha}_{k-1} e_{n+1} \\ \qquad +h[\hat{\beta}_0\{f(\tilde{x}_{n+k}, t_{n+k})-f(x_{n+k}, t_{n+k})\} \\ \qquad\quad +\hat{\beta}_1\{f(\tilde{x}_{n+k-1}, t_{n+k-1})-f(x_{n+k-1}, t_{n+k-1})\}+\cdots \\ \qquad\qquad +\hat{\beta}_k\{f(\tilde{x}_n, t_n)-f(x_n, t_n)\}]+\hat{R}_n-\hat{T}_n. \end{cases}
$$

In the same way as in the preceding paragraph, these can be written as follows:

$$
(1.32)\quad \begin{cases} e_{n+k}=\alpha_1 e_{n+k-1}+\alpha_2 e_{n+k-2}+\cdots+\alpha_{k-1} e_{n+1} \\ \qquad +h(\beta_{-1} F_n \hat{e}_{n+k+1}+\beta_0 F_n e_{n+k}+\cdots+\beta_{k-1} F_n e_{n+1}) \\ \qquad +h(\beta_{-1} \hat{\Phi}_{n,k+1} \hat{e}_{n+k+1}+\beta_0 \Phi_{n,k} e_{n+k}+\cdots+\beta_{k-1} \Phi_{n,1} e_{n+1}) \\ \qquad +R_n-T_n, \\ \hat{e}_{n+k+1}=\hat{\alpha}_0 e_{n+k}+\hat{\alpha}_1 e_{n+k-1}+\cdots+\hat{\alpha}_{k-1} e_{n+1} \\ \qquad +h(\hat{\beta}_0 F_n e_{n+k}+\hat{\beta}_1 F_n e_{n+k-1}+\cdots+\hat{\beta}_k F_n e_n) \\ \qquad +h(\hat{\beta}_0 \Phi_{n,k} e_{n+k}+\hat{\beta}_1 \Phi_{n,k-1} e_{n+k-1}+\cdots+\hat{\beta}_k \Phi_{n,0} e_n) \\ \qquad +\hat{R}_n-\hat{T}_n, \end{cases}
$$

where $F_n=F(x_n, t_n)$ and

$$
(1.33)\quad \begin{cases} \hat{\Phi}_{n,k+1}=o(1) \qquad \text{uniformly as } |h|, |\hat{e}_{n+k+1}|\to 0, \\ \Phi_{n,i}=o(1) \qquad \text{uniformly as } |h|, |e_{n+i}|\to 0 \qquad (i=0, 1, \cdots, k). \end{cases}
$$

As in the preceding paragraph, for R_n and T_n, we assume that

$$
(1.34)\qquad R_n, T_n=o(|h|) \qquad \text{uniformly as } |h|\to 0.
$$

But, for $\hat{R}_n$ and $\hat{T}_n$, we assume the weaker condition that

$$
(1.35)\qquad \hat{R}_n, \hat{T}_n=o(1) \qquad \text{uniformly as } |h|\to 0.
$$

From the second of these conditions, it is evident that

$$
(1.36)\qquad |\hat{e}_{n+k+1}|\to 0 \text{ uniformly as } |h|, |e_{n+i}|\to 0 \qquad (i=0, 1, \cdots, k).
$$

Consequently, for $\hat{\Phi}_{n,k+1}$ in which $\hat{e}_{n+k+1}$ is replaced by the second expression of (1.32), we have

$$\hat{\Phi}_{n,k+1}=o(1) \text{ uniformly as } |h|,\ |e_{n+i}|\to 0 \qquad (i=0,1,\cdots,k). \tag{1.37}$$

Then, substituting the second equation of (1.32) into the first equation of (1.32), we have

$$\begin{aligned} e_{n+k}=&\alpha_1 e_{n+k-1}+\alpha_2 e_{n+k-2}+\cdots+\alpha_{k-1}e_{n+1}\\ &+hF_n\{(\beta_{-1}\hat{\alpha}_0+\beta_0)e_{n+k}+(\beta_{-1}\hat{\alpha}_1+\beta_1)e_{n+k-1}+\cdots+(\beta_{-1}\hat{\alpha}_{k-1}+\beta_{k-1})e_{n+1}\}\\ &+h(\Psi'_{n,k}e_{n+k}+\Psi'_{n,k-1}e_{n+k-1}+\cdots+\Psi'_{n,0}e_n)+S'_n, \end{aligned} \tag{1.38}$$

where

$$\begin{aligned} \Psi'_{n,k-l}=&h\beta_{-1}\hat{\beta}_l F_n(F_n+\Phi_{n,k-l})+\beta_{-1}\hat{\alpha}_l\hat{\Phi}_{n,k+1}\\ &+h\beta_{-1}\hat{\beta}_l\hat{\Phi}_{n,k+1}(F_n+\Phi_{n,k-l})+\beta_l\Phi_{n,k-l}\\ =&o(1) \qquad (l=0,1,\cdots,k)\\ &\text{uniformly as } |h|,\ |e_{n+i}|\to 0 \qquad (i=0,1,\cdots,k), \end{aligned} \tag{1.39}$$

and

$$\begin{aligned} S'_n=&R_n-T_n+h\beta_{-1}(F_n+\hat{\Phi}_{n,k+1})(\hat{R}_n-\hat{T}_n)\\ =&o(|h|) \qquad \text{uniformly as } |h|\to 0. \end{aligned} \tag{1.40}$$

Since $|F_n|$ is bounded in D, evidently $\det\{I-h(\beta_{-1}\hat{\alpha}_0+\beta_0)F_n\}\neq 0$ for sufficiently small $|h|$. Consequently (1.38) can be rewritten as follows:

$$\begin{aligned} e_{n+k}=&[\alpha_1+h\{\beta_{-1}(\hat{\alpha}_0\alpha_1+\hat{\alpha}_1)+(\beta_0\alpha_1+\beta_1)\}F_n]e_{n+k-1}\\ &+[\alpha_2+h\{\beta_{-1}(\hat{\alpha}_0\alpha_2+\hat{\alpha}_2)+(\beta_0\alpha_2+\beta_2)\}F_n]e_{n+k-2}\\ &+\cdots\cdots\\ &+[\alpha_{k-1}+h\{\beta_{-1}(\hat{\alpha}_0\alpha_{k-1}+\hat{\alpha}_{k-1})+(\beta_0\alpha_{k-1}+\beta_{k-1})\}F_n]e_{n+1}\\ &+h(\Psi_{n,k}e_{n+k}+\cdots+\Psi_{n,0}e_n)+S_n, \end{aligned} \tag{1.41}$$

where

$$\begin{cases} \Psi_{n,k-l}=\{I-h(\beta_{-1}\hat{\alpha}_0+\beta_0)F_n\}^{-1}\Psi'_{n,k-l} \qquad (l=0,k),\\ \Psi_{n,k-l}=\{I-h(\beta_{-1}\hat{\alpha}_0+\beta_0)F_n\}^{-1}\Psi'_{n,k-l}\\ \qquad +\dfrac{1}{h}[\{I-h(\beta_{-1}\hat{\alpha}_0+\beta_0)F_n\}^{-1}\{\alpha_l+h(\beta_{-1}\hat{\alpha}_l+\beta_l)F_n\}\\ \qquad\qquad -\{\alpha_l+h(\beta_{-1}(\hat{\alpha}_0\alpha_l+\hat{\alpha}_l)+(\beta_0\alpha_l+\beta_l))F_n\}]\\ \qquad =\{I-h(\beta_{-1}\hat{\alpha}_0+\beta_0)F_n\}^{-1}\Psi'_{n,k-l}+O(|h|) \qquad (l=1,2,\cdots,k-1),\\ S_n=\{I-h(\beta_{-1}\hat{\alpha}_0+\beta_0)F_n\}^{-1}S'_n. \end{cases} \tag{1.42}$$

From (1.39) and (1.40), it is evident that

$$\begin{cases} \Psi_{n,k-l}=o(1) \qquad (l=0,1,\cdots,k)\\ \qquad\qquad \text{uniformly as } |h|,\ |e_{n+i}|\to 0 \qquad (i=0,1,\cdots,k),\\ S_n=o(|h|) \text{ uniformly as } |h|\to 0. \end{cases} \tag{1.43}$$

The equation (1.41) is of the same form as (1.10). But, in (1.41), the order

of the approximate linear difference equation[1] is $k-1$ while it is k in (1.10). This difference is important.

As in the preceding paragraph, let us introduce the notations

$$(1.44)\quad \begin{cases} \underset{n}{\hat{A}}{}^{\nu}_{\mu l}(h)=\delta^{\nu}_{\mu}\,\alpha_l+h\{\beta_{-1}(\hat{\alpha}_0\,\alpha_l+\hat{\alpha}_l)+(\beta_0\,\alpha_l+\beta_l)\}\underset{n}{F}{}^{\nu}_{\mu} & (l=1,\,2,\cdots,\,k-1), \\ \underset{n}{\hat{A}}{}^{\nu}_{\mu k}(h)=0. \end{cases}$$

Then, similarly to (1.10), the equation (1.41) can be rewritten in the same form as (1.25) as follows:

$$(1.45)\qquad \boldsymbol{e}_{n+1}=\hat{A}_n(h)\,\boldsymbol{e}_n+h\,\hat{\boldsymbol{\varphi}}_n(\boldsymbol{e}_{n+1},\,\boldsymbol{e}_n)+h\,\hat{\boldsymbol{r}}_n,$$

where $\hat{A}_n(h)$, $\hat{\boldsymbol{\varphi}}_n$ and $\hat{\boldsymbol{r}}_n$ are of the same form as $A_n(h)$, $\boldsymbol{\varphi}_n$ and $\boldsymbol{r}_n$ except that $\underset{n}{A}{}^{\nu}_{\mu l}$ in $\underset{n}{B}{}^{\nu}_{\mu}$ are replaced by $\underset{n}{\hat{A}}{}^{\nu}_{\mu l}$.

Thus we have the conclusions:

The difference equation which the errors of a compound multi-step integration formula satisfy is of the same form as that which the errors of a usual multi-step integration formula satisfy. One of the important differences is that the order of the approximate linear difference equation of the former is less by one than that of the latter having the same value of k.

1.3 Difference equations for errors of general Runge-Kutta formulas

By [6], the general Runge-Kutta formula reads as follows:

$$(1.46)\qquad x_{n+1}=x_n+h(a\,k_{n1}+b\,k_{n2}+c\,k_{n3}+d\,k_{n4}),$$

where

$$(1.47)\quad \begin{cases} k_{n1}=f(x_n,\,t_n), \\ k_{n2}=f(x_n+mhk_{n1},\,t_n+mh), \\ k_{n3}=f\{x_n+(p-r)hk_{n1}+rhk_{n2},\,t_n+ph\}, \\ k_{n4}=f\{x_n+(q-s-u)hk_{n1}+shk_{n2}+uhk_{n3},\,t_n+qh\} \end{cases}$$

and $(a,\,b,\,c,\,d;\,m,\,p,\,q,\,r,\,s,\,u)$ are the constants satisfying the equations

$$(1.48)\quad \begin{cases} a+b+c+d=1, & cmr+d(pu+ms)=1/6, \\ bm+cp+dq=1/2, & cmpr+d(pu+ms)q=1/8, \\ bm^2+cp^2+dq^2=1/3, & cm^2r+d(p^2u+m^2s)=1/12, \\ bm^3+cp^3+dq^3=1/4, & dmru=1/24. \end{cases}$$

1) This means the equation obtained by neglecting the terms $h\Psi_{n,k-l}\,e_{n+k-l}$ $(l=0,\,1,\cdots,\,k)$ and S_n.

Let $\tilde{x}_i$ be the approximate values obtained for x_i[1]—the true values of the solution of (1.2) for $t=t_i=x_0+i\,h$. Then evidently

$$x_{n+1}=x_n+h(a\,k_{n1}+b\,k_{n2}+c\,k_{n3}+d\,k_{n4})+T_n \tag{1.49}$$

for

$$\begin{cases} k_{n1}=f(x_n,\ t_n), \\ k_{n2}=f(x_n+mhk_{n1},\ t_n+mh), \\ k_{n3}=f\{x_n+(p-r)hk_{n1}+rhk_{n2},\ t_n+ph\}, \\ k_{n4}=f\{x_n+(q-s-u)hk_{n1}+shk_{n2}+uhk_{n3},\ t_n+qh\} \end{cases} \tag{1.50}$$

and

$$\tilde{x}_{n+1}=\tilde{x}_n+h(a\,\tilde{k}_{n1}+b\,\tilde{k}_{n2}+c\,\tilde{k}_{n3}+d\,\tilde{k}_{n4})+R_n \tag{1.51}$$

for

$$\begin{cases} \tilde{k}_{n1}=f(\tilde{x}_n,\ t_n)+r'_{n1}, \\ \tilde{k}_{n2}=f(\tilde{x}_n+mh\tilde{k}_{n1},\ t_n+mh)+r'_{n2}, \\ \tilde{k}_{n3}=f\{\tilde{x}_n+(p-r)h\tilde{k}_{n1}+rh\tilde{k}_{n2},\ t_n+ph\}+r'_{n3}, \\ \tilde{k}_{n4}=f\{\tilde{x}_n+(q-s-u)h\tilde{k}_{n1}+sh\tilde{k}_{n2}+uh\tilde{k}_{n3},\ t_n+ph\}+r'_{n4}\text{[2]}. \end{cases} \tag{1.52}$$

Here T_n is a truncation error and R_n, r'_{n1}, r'_{n2}, r'_{n3}, r'_{n4} are round-off errors.

Put

$$\tilde{x}_i-x_i=e_i \qquad (i=0,\ 1,\ 2,\cdots). \tag{1.53}$$

Then evidently the e_i express the errors of the approximate solution computed by the general Runge-Kutta formula (1.46).

Let us consider the differences $\tilde{k}_{ni}-k_{ni}$ $(i=1,\ 2,\ 3,\ 4)$ successively. From the first equations of (1.50) and (1.52), follows readily

$$\tilde{k}_{n1}-k_{n1}=f(\tilde{x}_n,\ t_n)-f(x_n,\ t_n)+r'_{n1}=F_n\,e_n+\Phi_{n1}\,e_n+r_{n1}, \tag{1.54}$$

where

$$r_{n1}=r'_{n1} \tag{1.55}$$

and

---

1) It is assumed that $(\tilde{x}_i,\ t_i)\in D$.

2) It is assumed that $(\tilde{x}_n+mh\tilde{k}_{n1},\ t_n+mh)$, $\{\tilde{x}_n+(p-r)h\tilde{k}_{n1}+rh\tilde{k}_{n2},\ t_n+ph\}$, $\{\tilde{x}_n+(q-s-u)h\tilde{k}_{n1}+sh\tilde{k}_{n2}+uh\tilde{k}_{n3},\ t_n+qh\}\in D$.

$$\Phi_{n1}=\int_0^1 F(x_n+\theta e_n,\ t_n)\,d\theta-F(x_n,\ t_n)=o(1) \qquad \text{uniformly as } |e_n|\to 0. \tag{1.56}$$

Then, from the second equations of (1.50) and (1.52) follows

$$\tilde{k}_{n2}-k_{n2}=F_n\,e_n+\Phi_{n2}\,e_n+r_{n2}\,. \tag{1.57}$$

Here

$$\Phi_{n2}=(J_{n1}-F_n)+mhJ_{n1}(F_n+\Phi_{n1}), \tag{1.58}$$

$$r_{n2}=mhJ_{n1}\,r_{n1}+r'_{n2}, \tag{1.59}$$

where

$$J_{n1}=\int_0^1 F[x_n+mhk_{n1}+\theta\{e_n+mh(F_n+\Phi_{n1})e_n+mhr_{n1}\},\ t_n+mh]\,d\theta. \tag{1.60}$$

Since

$$J_{n1}=F_n+o(1) \qquad \text{uniformly as } |h|,\ |e_n|\to 0, \tag{1.61}$$

it is evident that

$$\Phi_{n2}=o(1) \qquad \text{uniformly as } |h|,\ |e_n|\to 0. \tag{1.62}$$

Likewise, from the third and fourth equations of (1.50) and (1.52), it follows successively that

$$\tilde{k}_{n3}-k_{n3}=F_n\,e_n+\Phi_{n3}\,e_n+r_{n3}, \tag{1.63}$$

where

$$\begin{aligned} J_{n2}=\int_0^1 F[&x_n+(p-r)hk_{n1}+rhk_{n2}\\ &+\theta\{e_n+(p-r)h((F_n+\Phi_{n1})e_n+r_{n1})+rh((F_n+\Phi_{n2})e_n+r_{n2})\},\ t_n+ph]\,d\theta, \end{aligned} \tag{1.64}$$

$$\begin{aligned} \Phi_{n3}&=(J_{n2}-F_n)+(p-r)hJ_{n2}(F_n+\Phi_{n1})+rhJ_{n2}(F_n+\Phi_{n2})\\ &=o(1) \qquad \text{uniformly as } |h|,\ |e_n|\to 0, \end{aligned} \tag{1.65}$$

$$r_{n3}=(p-r)hJ_{n2}\,r_{n1}+rhJ_{n2}\,r_{n2}+r'_{n3}, \tag{1.66}$$

and

$$\tilde{k}_{n4}-k_{n4}=F_n\,e_n+\Phi_{n4}\,e_n+r_{n4}, \tag{1.67}$$

where

$$\begin{aligned} J_{n3}=\int_0^1 F[&x_n+(q-s-u)hk_{n1}+shk_{n2}+uhk_{n3}\\ &+\theta\{e_n+(q-s-u)h((F_n+\Phi_{n1})e_n+r_{n1})+sh((F_n+\Phi_{n2})e_n+r_{n2})\\ &\quad+uh((F_n+\Phi_{n3})e_n+r_{n3})\},\ t_n+qh]\,d\theta, \end{aligned} \tag{1.68}$$

$$\Phi_{n4}=(J_{n3}-F_n)+(q-s-u)hJ_{n3}(F_n+\Phi_{n1})+shJ_{n3}(F_n+\Phi_{n2})+uhJ_{n3}(F_n+\Phi_{n3}) \tag{1.69}$$
$$=o(1) \quad \text{uniformly as } |h|,\ |e_n|\to 0,$$

$$r_{n4}=(q-s-u)hJ_{n3}\,r_{n1}+shJ_{n3}\,r_{n2}+uhJ_{n3}\,r_{n3}+r'_{n4}. \tag{1.70}$$

Now, by (1.49) and (1.51),

$$e_{n+1}=e_n+h\{a(\tilde{k}_{n1}-k_{n1})+b(\tilde{k}_{n2}-k_{n2}) \tag{1.71}$$
$$+c(\tilde{k}_{n3}-k_{n3})+d(\tilde{k}_{n4}-k_{n4})\}+R_n-T_n.$$

Consequently, subsituting (1.54), (1.57), (1.63) and (1.67) into the above equation and using the first equation of (1.48), we have

$$e_{n+1}=A_n(h)\,e_n+h\,\varphi_n(e_n)+s_n, \tag{1.72}$$

where

$$\begin{cases} A_n(h)=I+hF_n, \\ \varphi_n(e_n)=(a\,\Phi_{n1}+b\,\Phi_{n2}+c\,\Phi_{n3}+d\,\Phi_{n4})\,e_n, \\ s_n=h(a\,r_{n1}+b\,r_{n2}+c\,r_{n3}+d\,r_{n4})+R_n-T_n. \end{cases} \tag{1.73}$$

From (1.56), (1.62), (1.65) and (1.69), it is evident that

$$\varphi_n(e_n)=o(|e_n|) \quad \text{uniformly as } |h|,\ |e_n|\to 0. \tag{1.74}$$

Now, if the round-off errors $r'_{n1}, r'_{n2}, r'_{n3}, r'_{n4}$ satisfy the condition that

$$r'_{n1}, r'_{n2}, r'_{n3}, r'_{n4}=o(1) \quad \text{uniformly as } |h|\to 0, \tag{1.75}$$

then, from (1.55), (1.59), (1.66) and (1.70), it is evident that

$$r_{n1}, r_{n2}, r_{n3}, r_{n4}=o(1) \quad \text{uniformly as } |h|\to 0.$$

Then, in addition, if R_n satisfies (1.23), from the last equation of (1.73), it follows that

$$s_n=o(|h|) \quad \text{uniformly as } |h|\to 0, \tag{1.76}$$

because (1.24) is evidently valid for the present T_n for any continuous differential system of the form (1.2)[1]. The property (1.76) says that the equation (1.72) can be rewritten as

$$e_{n+1}=A_n(h)\,e_n+h\,\varphi_n(e_n)+h\,r_n \tag{1.77}$$

and that

$$r_n=o(1) \quad \text{uniformly as } |h|\to 0. \tag{1.78}$$

1) This means the differential system $dx/dt=f(x,t)$ for which $f(x,t)$ is continuous with respect to x and t.

The equation (1.77) is of the same form as (1.25) and moreover $\varphi_n(e_n)$ and r_n satisfy respectively the conditions (1.74) and (1.78) corresponding to (1.22) and (1.26).

The important point is that, *for a general Runge-Kutta formula, the approximate linear difference equation is of order one; consequently, for a general Runge-Kutta formula, there is no need of the substitution of the form* (1.12) which is needed for usual multi-step formulas in order to transform the approximate linear difference equation to a system of equations of order one.

Chapter II. Stable convergence of integration formulas

2.1 Consistency conditions for multi-step integration formulas[1)]

In (1.24) and (1.34), we have assumed that the truncation errors are of order higher than one in h. From these conditions, it is evident that (1.1) and the first formula of (1.27) are true for any poynomial $x(t)$ of degree at most 1. Also, in (1.35), we have assumed that the truncation errors of the second formula of (1.27) are infinitesimal as $|h|\to 0$. From this condition, it is also evident that the second formula of (1.27) is true for any constant x.

Then, after simple calculations, we find that the coefficients of integration formulas must satisfy the following relations:

for a simple multi-step formula,

$$\begin{cases} \alpha_1+\alpha_2+\cdots+\alpha_k=1, \\ \beta_0+\beta_1+\cdots+\beta_k=\alpha_1+2\alpha_2+\cdots+k\alpha_k; \end{cases} \tag{2.1}$$

for a compound multi-step formula,

$$\begin{cases} \alpha_1+\alpha_2+\cdots+\alpha_{k-1}=1, \\ \beta_{-1}+\beta_0+\beta_1+\cdots+\beta_{k-1}=\alpha_1+2\alpha_2+\cdots+(k-1)\alpha_{k-1}, \\ \hat{\alpha}_0+\hat{\alpha}_1+\cdots+\hat{\alpha}_{k-1}=1. \end{cases} \tag{2.2}$$

Conversely, when the coefficients of integration formulas satisfy the above conditions, from

$$x(t_{n+i})=x(t_n)+\int_{t_n}^{t_{n+i}} \dot{x}(t)\,dt,$$

it readily follows that the truncation errors of the integration formulas really satisfy (1.24) or (1.34) and (1.35) for any continuous differential system of the form (1.2).

Thus *the conditions* (2.1) *or* (2.2) *are respectively the necessary and sufficient conditions that the truncation errors of multi-step integration formulas may satisfy* (1.24) *or* (1.34) *and* (1.35) *for any continuous differential system of the form* (1.2).

According to the nomenclature of Hull and Luxemburg [9], let us call the conditions (2.1) and (2.2) the *consistency conditions.*

---

1) For brevity, in the sequel, we call both of the usual multi-step integration formulas and the compound multi-step formulas of **1.2** simply the multi-step formulas. And, if necessary, we call the former the simple multi-step formulas and the latter the compound multi-step formulas.

2.2 Necessary conditions for stability for multi-step formulas

We assume the consistency conditions for multi-step formuas.

Let us consider the case where the multi-step formulas are applied without any rounding to the one dimensional differential equation

$$\frac{dx}{dt}=0. \tag{2.3}$$

Then, since the solutions of (2.3) are constants, no truncation error appears by the consistency conditions. Also, by the assumption, no round-off error appears. In addition, in the present case,

$$F(x, t)=0,$$

consequently

$$\Phi_{n,k-l}=\hat{\Phi}_{n,k+1}=\Psi_{n,k-l}=0 \qquad (l=0, 1, \cdots, k).$$

Thus, for errors, by (1.25) and (1.45), we have the equation of the form as follows:

$$\boldsymbol{e}_{n+1}=A_0\,\boldsymbol{e}_n, \tag{2.4}$$

where

$$A_0=\begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & & \vdots \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 & 1 \\ \alpha_k & \alpha_{k-1} & \cdots & \alpha_2 & \alpha_1 \end{pmatrix} \tag{2.5}$$

Here $\alpha_k=0$ for a compound multi-step formula.

From (2.4) readily follows

$$\boldsymbol{e}_n=A_0^n\,\boldsymbol{e}_0 \qquad (n=0, 1, 2, \cdots). \tag{2.6}$$

Now, as is readily seen, the eigenvalues of A_0 are the roots of the equation

$$\rho(\lambda)\overset{\text{def}}{=}\lambda^k-(\alpha_1\lambda^{k-1}+\alpha_2\lambda^{k-2}+\cdots+\alpha_{k-1}\lambda+\alpha_k)=0 \tag{2.7}$$

and to each eigenvalue of A_0 corresponds only one eigenvector. Therefore, if we denote by λ_i $(i=1, 2, \cdots)$ the distinct roots of the equation (2.7) and by m_i $(i=1, 2, \cdots)$ their multiplicities, then the Jordan canonical form of A_0 becomes the direct sum of the matrices

$$(2.8)\qquad \begin{pmatrix} \lambda_i & 0 & \cdots & 0 \\ \delta & \lambda_i & & \vdots \\ 0 & \delta & \lambda_i & \\ \vdots & & \ddots & 0 \\ 0 & \cdots & \delta & \lambda_i \end{pmatrix} \qquad (i=1, 2,\cdots)$$

of the order m_i, where δ is an arbitrary positive number.

Let A_0 be the matrix of this Jordan canonical form and let

$$(2.9)\qquad T_0^{-1} A_0 T_0 = A_0 .$$

Then, by substitution

$$(2.10)\qquad \boldsymbol{e}_n = T_0 \boldsymbol{e}'_n ,$$

(2.6) can be rewritten as

$$(2.11)\qquad \boldsymbol{e}'_n = A_0^n \boldsymbol{e}'_0 .$$

Here A_0^n is the direct sum of the matrices

$$(2.12)\qquad \begin{pmatrix} \lambda_i & 0 & \cdots & 0 \\ \delta & \lambda_i & & \vdots \\ 0 & \delta & \ddots & \\ \vdots & & \ddots & 0 \\ 0 & \cdots & \delta & \lambda_i \end{pmatrix}^n = \begin{pmatrix} \lambda_i^n & 0 & \cdots & 0 \\ \binom{n}{1}\lambda_i^{n-1}\delta & \lambda_i^n & & \vdots \\ \binom{n}{2}\lambda_i^{n-2}\delta^2 & \binom{n}{1}\lambda_i^{n-1}\delta & \ddots & \\ \vdots & & \ddots & 0 \\ & \cdots & \binom{n}{1}\lambda_i^{n-1}\delta & \lambda_i^n \end{pmatrix}$$

$$(i=1, 2,\cdots),$$

because A_0 is the direct sum of the matrices (2.8).

Now, in order that the integration formulas be stable, $\boldsymbol{e}_n$ must be bounded for any $\boldsymbol{e}_0$ provided $|\boldsymbol{e}_0|$ is sufficiently small. This implies that, for any $\boldsymbol{e}'_0$ such that its norm is sufficiently small and $T\boldsymbol{e}'_0$ is real, $\boldsymbol{e}'_n$ determined by (2.11) must be bounded. Now, for any $t \neq t_0$,

$$n = \frac{t-t_0}{h} \to \infty \qquad \text{as } |h| \to 0.$$

Therefore, from (2.12), we see that, for stability of the integration formulas, the following two conditions are necessary:

$$(2.13)\qquad \begin{cases} 1^\circ \textit{ the roots of the equation } (2.7) \textit{ are all at most one in absolute value}; \\ 2^\circ \textit{ the roots whose absolute values are one are simple.} \end{cases}$$

These are the desired necessary conditions. In the sequel, these two conditions are called the *stability conditions.*

Remark. The roots of the equation (2.7) are the eigenvalues of the matrix

$A_n(0)$ in the case where an integration formula is applied to any one dimensional differential equation. On the other hand, for a general Runge-Kutta formula, evidently $A_n(0)=I$ as is seen from (1.73). Consequently, when a general Runge-Kutta formula is applied to any one dimensional differential equation, $A_n(0)=1$, in other words, the eigenvalue of $A_n(0)$ is merely 1. This says that we may suppose that *the stability conditions are always satisfied by general Runge-Kutta formulas.*

2.3 Stability of integration formulas

In this paragraph, we prove in a unified form stability of the general Runge-Kutta formulas and the multi-step formulas which fulfill both the consistency conditions and the stability conditions.

As in proof of Perron's existence theorem, first, we extend the function $f(x, t)$ to $\tilde{f}(x, t)$ such that

$$\tilde{f}(x, t)=f(x, t) \quad \text{in} \quad D$$

and

$$\tilde{f}(x, t)=f(\tilde{x}, t) \quad \text{for} \quad |t-t_0|\leqq L,\ |x-x(t)|\geqq r,$$

where $\tilde{x}=(\tilde{x}^\nu)$ is a point connected with $x=(x^\nu)$ as follows:

$$\tilde{x}^\nu=x^\nu \qquad \text{for } \nu \text{ such that } |x^\nu-x^\nu(t)|\leqq r,$$

$$\tilde{x}^\nu=x^\nu(t)+r \quad \text{for } \nu \text{ such that } x^\nu-x^\nu(t)>r,$$

$$\tilde{x}^\nu=x^\nu(t)-r \quad \text{for } \nu \text{ such that } x^\nu-x^\nu(t)<-r.$$

Then it is evident that, in the domain

$$\tilde{D}: |t-t_0|\leqq L,\ |x-x(t)|<\infty,$$

the function $\tilde{f}(x, t)$ is continuous and satisfies a Lipschitz condition with respect to x, because $f(x, t)$ is continuously differentiable with respect to x in the closed bounded domain D.

Corresponding to the initial differential system (1.2), let us consider the differential system

$$\frac{dx}{dt}=\tilde{f}(x, t). \tag{2.14}$$

Then, evidently, the solution $x=x(t)$ of (1.2) is also a solution of (2.14).

For (2.14), the approximate numerical solution can be actually constructed in the interval $|t-t_0|\leqq L$ by any one of the integration formulas under consideration—multi-step formulas and general Runge-Kutta formulas, provided

$|h|$ is sufficiently small in the case of multi-step formulas[1].

In constructing a numerical solution, we suppose that rounding is done so that the round-off errors satisfy (1.23), (1.34) and (1.35), or (1.75) and (1.23) in accordance with the formulas used.

For the approximate solution obtained, on replacing $f(x, t)$ by $\tilde{f}(x, t)$, we have (1.6), (1.31) or (1.71) in accordance with the formulas used. Let us denote here (1.6), (1.31) and (1.71) respectively by (1.6′), (1.31′) and (1.71′).

Since (1.23) and (1.24) hold, (1.6′) can be written in terms of the notations (1.14) as follows:

$$\boldsymbol{e}_{n+1}=A\,\boldsymbol{e}_n+h\,\boldsymbol{\psi}_n(\boldsymbol{e}_{n+1},\,\boldsymbol{e}_n)+h\,\boldsymbol{\rho}_n, \tag{2.15}$$

where

$$A=\begin{pmatrix} A_0 & & & \\ & A_0 & & \\ & & \ddots & \\ & & & A_0 \end{pmatrix} \quad \text{(direct sum of } N \text{ } A_0\text{'s)} \tag{2.16}$$

and

$$\boldsymbol{\rho}_n=o(1) \qquad \text{uniformly as } |h|\to 0. \tag{2.17}$$

Since $\tilde{f}(x, t)$ satisfies a Lipschitz condition, for certain positive constants K_1 and K_2,

$$|\boldsymbol{\psi}_n(\boldsymbol{e}_{n+1},\,\boldsymbol{e}_n)|\leqq K_1|\boldsymbol{e}_{n+1}|+K_2|\boldsymbol{e}_n|. \tag{2.18}$$

Likewise the first equation of (1.31′) can be written in the same form as (2.15) replacing $\hat{x}_{n+k+1}-x_{n+k+1}=\hat{e}_{n+k+1}$ by the second equation of (1.31′). But, in this case, (2.18) is replaced by

$$|\boldsymbol{\psi}_n(\boldsymbol{e}_{n+1},\,\boldsymbol{e}_n)|\leqq K_1|\boldsymbol{e}_{n+1}|+K_2|\boldsymbol{e}_n|+\sigma_n, \tag{2.19}$$

where

$$\sigma_n=o(1) \qquad \text{uniformly as } |h|\to 0. \tag{2.20}$$

For the general Runge-Kutta formula (1.46), by the Lipschitz boundedness of $\tilde{f}(x, t)$, we have

$$|\tilde{k}_{ni}-k_{ni}|\leqq K_{(i)}|\boldsymbol{e}_n|+r_{ni} \qquad (i=1,\,2,\,3,\,4) \tag{2.21}$$

for certain positive constants $K_{(i)}$ $(i=1, 2, 3, 4)$. Here r_{ni} $(i=1, 2, 3, 4)$ are quantities such that

$$r_{ni}=o(1) \qquad (i=1,\,2,\,3,\,4) \qquad \text{uniformly as } |h|\to 0.$$

1) This means that the points $(\bar{x}_{n+k}, t_{n+k})$ can be actually computed successively by solving the integration formulas with respect to $\bar{x}_{n+k}$ or $(\bar{x}_{n+k}, \hat{x}_{n+k+1})$ by means of the iteration method. For details of this fact, see the author's paper [17].

Consequently we see that (1.71′) can be written in the form

$$e_{n+1}=e_n+h\psi_n(e_n)+h\rho_n, \tag{2.22}$$

where

$$\rho_n=o(1) \qquad \text{uniformly as} \quad |h|\to 0$$

and

$$|\psi_n(e_n)|\leqq K_2|e_n|+\sigma_n \tag{2.23}$$

for a certain positive constant K_2. Here the σ_n are the quantities such that

$$\sigma_n=o(1) \qquad \text{uniformly as} \quad |h|\to 0.$$

(2.22) is a special case of (2.15), and (2.18) and (2.23) are special cases of (2.19). Thus, for study of (1.6′), (1.31′) and (1.71′), we need only to consider the equations (2.15) for which (2.19) hold.

Now, by the stability conditions (2.13), we can choose $\delta>0$ so small that

$$\delta+|\lambda_i|<1 \tag{2.24}$$

for any multiple eigenvalue of A_0. On the other hand, one of the eigenvalues of A_0 is 1 by the consistency conditions or by the remark at the end of **2.2** and moreover such an eigenvalue is simple by the stability conditions. Thus, by the stability conditions, we see that

$$|\Lambda_0|=1. \tag{2.25}$$

Then, if we put

$$T=\begin{pmatrix} T_0 & & & \\ & T_0 & & \\ & & \ddots & \\ & & & T_0 \end{pmatrix} \quad \text{(direct sum of } N \ \ T_0\text{'s)} \tag{2.26}$$

and

$$T^{-1}A\,T=\Lambda, \tag{2.27}$$

we see from (2.9) that

$$\Lambda=\begin{pmatrix} \Lambda_0 & & & \\ & \Lambda_0 & & \\ & & \ddots & \\ & & & \Lambda_0 \end{pmatrix} \quad \text{(direct sum of } N \ \ \Lambda_0\text{'s)}, \tag{2.28}$$

from which, by (2.25), follows

$$|\Lambda|=1. \tag{2.29}$$

Let us put

$$\boldsymbol{e}_n=T\boldsymbol{e}'_n,\quad \boldsymbol{\rho}_n=T\boldsymbol{\rho}'_n \tag{2.30}$$

and

$$T^{-1}\,\boldsymbol{\psi}_n(T\boldsymbol{e}'_{n+1},\ T\boldsymbol{e}'_n)=\boldsymbol{\psi}'_n(\boldsymbol{e}'_{n+1},\ \boldsymbol{e}'_n). \tag{2.31}$$

Then, from (2.19) and (2.17), follows readily

$$|\boldsymbol{\psi}'_n(\boldsymbol{e}'_{n+1},\ \boldsymbol{e}'_n)|\leqq K'_1|\boldsymbol{e}'_{n+1}|+K'_2|\boldsymbol{e}'_n|+\sigma'_n \tag{2.32}$$

and

$$|\boldsymbol{\rho}'_n|=o(1)\qquad \text{uniformly as}\quad |h|\to 0, \tag{2.33}$$

where

$$K'_1=|T^{-1}|\,|T|K_1,\quad K'_2=|T^{-1}|\,|T|K_2 \tag{2.34}$$

and

$$\sigma'_n=|T^{-1}|\sigma_n=o(1)\qquad \text{uniformly as}\quad |h|\to 0. \tag{2.35}$$

Let us take a positive number h_0 such that

$$h_0\,K'_1<1^{1)}. \tag{2.36}$$

Evidently, by (2.33) and (2.35), for any positive number ε', we can take a positive number h_1 ($\leqq h_0$) so that

$$|\sigma'_n|+|\boldsymbol{\rho}'_n|\leqq\varepsilon'\ \text{ for any } h \text{ such that }\ |h|\leqq h_1. \tag{2.37}$$

Now, by the substitution (2.27), (2.30) and (2.31), the equation (2.15) is written as follows:

$$\boldsymbol{e}'_{n+1}=A\,\boldsymbol{e}'_n+h\,\boldsymbol{\psi}'_n(\boldsymbol{e}'_{n+1},\ \boldsymbol{e}'_n)+h\,\boldsymbol{\rho}'_n. \tag{2.38}$$

Then, for h such that $|h|\leqq h_1$, by (2.29), (2.32) and (2.37), we have:

$$|\boldsymbol{e}'_{n+1}|\leqq|\boldsymbol{e}'_n|+|h|(K'_1|\boldsymbol{e}'_{n+1}|+K'_2|\boldsymbol{e}'_n|)+|h|\varepsilon',$$

which, due to (2.36), can be written as follows:

$$|\boldsymbol{e}'_{n+1}|\leqq\frac{1+|h|K'_2}{1-|h|K'_1}|\boldsymbol{e}'_n|+\frac{|h|}{1-|h|K'_1}\varepsilon'. \tag{2.39}$$

But, as is verified easily, for h such that $|h|\leqq h_0$,

$$\frac{1+|h|K'_2}{1-|h|K'_1}\leqq 1+\frac{K'_1+K'_2}{1-h_0\,K'_1}|h|$$

---

1) In the case of multi-step formulas, for h such that $|h|\leqq h_0$, the integration formulas can be actually solved with respect to $\tilde{x}_{n+k}$ or $(\tilde{x}_{n+k},\ \hat{x}_{n+k+1})$ by the iteration method [17]. So, for h such that $|h|\leqq h_0$, an approximate numerical solution can be always actually constructed in the interval $|t-t_0|\leqq L$ by any one of the integration formulas under consideration.

and

$$\frac{|h|}{1-|h|K_1'} \leqq \frac{|h|}{1-h_0 K_1'}.$$

Therefore, from (2.39), follows

$$|e_{n+1}'| \leqq \left(1+\frac{K_1'+K_2'}{1-h_0 K_1'}|h|\right)|e_n'| + \frac{|h|}{1-h_0 K_1'}\varepsilon'. \tag{2.40}$$

Then, by induction, we have:

$$\begin{aligned}|e_n'| \leqq & \left(1+\frac{K_1'+K_2'}{1-h_0 K_1'}|h|\right)^n |e_0'| \\ & + \frac{1}{K_1'+K_2'}\left\{\left(1+\frac{K_1'+K_2'}{1-h_0 K_1'}|h|\right)^n - 1\right\}\varepsilon' \\ & (n=0, 1, 2, \cdots),\end{aligned} \tag{2.41}$$

from which follows

$$\begin{aligned}|e_n'| \leqq & \left(\exp\frac{K_1'+K_2'}{1-h_0 K_1'}L\right)|e_0'| \\ & + \frac{1}{K_1'+K_2'}\left\{\left(\exp\frac{K_1'+K_2'}{1-h_0 K_1'}L\right) - 1\right\}\varepsilon' \\ & (n=0, 1, 2, \cdots)\end{aligned} \tag{2.42}$$

because $n|h| = |t_n - t_0| \leqq L$.

Due to (2.37), the inequalities (2.42) imply that $|e_n'| \to 0$ uniformly as $|e_0'|$ and $|h|$ tend to zero. By (2.30), this implies that $|e_n| \to 0$ uniformly as $|e_0|$ and $|h|$ tend to zero.

Then, evidently, for sufficiently small $|e_0|$ and $|h|$, all the points

$$\begin{aligned}&(\tilde{x}_n, t_n), (\hat{x}_n, t_n),\\ &(\tilde{x}_n + mh\tilde{k}_{n1}, t_n + mh),\\ &\{\tilde{x}_n + (p-r)h\tilde{k}_{n1} + rh\tilde{k}_{n2}, t_n + ph\},\\ &\{\tilde{x}_n + (q-s-u)h\tilde{k}_{n1} + sh\tilde{k}_{n2} + uh\tilde{k}_{n3}, t_n + qh\}\end{aligned}$$

computed for the differential system (2.14) lie in the domain D. This means that all the above points are nothing but the points computed for the initial differential system (1.2), because, in the domain D, $\tilde{f}(x, t)$ coincides with $f(x, t)$. Then this says that the approximate numerical solution of the initial differential system can be actually obtained in the domain D by any of the integration formulas under consideration.

For such approximate numerical solution, as is mentioned above, $|e_n| \to 0$ uniformly as $|e_0|$ and $|h|$ tend to zero, namely the numerical solution obtained tends to the true solution uniformly in the interval $|t-t_0| \leqq L$ as the starting

values tend to the true values and at the same time the length of divided intervals tends to zero.

The above two facts prove that the integration formulas under consideration—the general Runge-Kutta formulas and the multi-step formulas which fulfill both the consistency conditions and the stability conditions—are stable.

2.4 Conclusions

The results obtained in the present chapter are summarized as follows:

A necessary and sufficient condition that a multi-step formula satisfying the consistency conditions be stable is that the stability conditions are valid for it, provided computation is rounded so that round-off errors may satisfy either (1.23) *for a simple multi-step formula or* (1.34) *and* (1.35) *for a compound multi-step formula.*

A general Runge-Kutta formula is always stable, provided computation is rounded so that round-off errors may satisfy (1.75) *and* (1.23).

Chapter III. Propagation of errors

In this and the next chapter, we suppose that the multi-step formulas under consideration always fulfill both the consistency conditions and the stability conditions, namely that they are always stable.

As has been shown in the preceding chapter, by means of such multi-step formulas or general Runge-Kutta formulas, we can actually construct an approximate numerical solution lying in D by taking sufficiently accurate starting values and sufficiently small $|h|$ if we do the computation sufficiently minutely—namely if we round the computation so that round-off errors satisfy the conditions (1.23) or (1.34) and (1.35) or (1.75) and (1.23) in accordance with the formulas used.

In this and the next chapter, the errors of the approximate numerical solutions obtained in the above way are discussed.

3.1 Local approximate error formulas

Let us divide the given interval $[t_0-L,\ t_0+L]$ into subintervals so that $F\{x(t), t\}$ varies but little in each subinterval and let us consider the error formulas inside such subintervals.

By the way, the error formulas (1.45) for compound multi-step formulas and (1.77) for general Runge-Kutta formulas are of the same form as (1.25) for simple multi-step formulas. So, in the sequel, we shall represent all the error formulas by (1.25).

Let L_0 be the length of any such subinterval I_0 and $A(h)$ be a common approximate value of $A_n(h)$ corresponding to a certain common approximate value of $F(x_n, t_n)$'s in I_0. Then, by any one of (1.15), (1.44) and (1.73),

$$|A_n(h)-A(h)|/|h|$$

are always small in I_0, consequently, shifting the term

$$\{A_n(h)-A(h)\}\boldsymbol{e}_n$$

into the term $h\boldsymbol{\varphi}_n(\boldsymbol{e}_{n+1}, \boldsymbol{e}_n)$, in I_0, we can write (1.25) as follows:

$$\boldsymbol{e}_{n+1}=A(h)\,\boldsymbol{e}_n+h\,\boldsymbol{\varphi}_n(\boldsymbol{e}_{n+1}, \boldsymbol{e}_n)+h\,\boldsymbol{r}_n. \tag{3.1}$$

From this, it readily follows that

$$\begin{aligned}\boldsymbol{e}_{n+n_0}=&A^n(h)\,\boldsymbol{e}_{n_0}\\&+h\{A^{n-1}(h)\,\boldsymbol{r}_{n_0}+A^{n-2}(h)\,\boldsymbol{r}_{n_0+1}+\cdots+A(h)\,\boldsymbol{r}_{n_0+n-2}+\boldsymbol{r}_{n_0+n-1}\}\\&+h(A^{n-1}(h)\,\boldsymbol{\varphi}_{n_0}+A^{n-2}(h)\,\boldsymbol{\varphi}_{n_0+1}+\cdots+A(h)\,\boldsymbol{\varphi}_{n_0+n-2}+\boldsymbol{\varphi}_{n_0+n-1}\}\end{aligned} \tag{3.2}$$

for $t_{n_0}, t_{n_0+1}, \cdots, t_{n_0+n}, \cdots \in I_0$.

Now, as is seen from any one of (1.15), (1.44) and (1.73), $A(h)$ is always of the form

$$A(h)=A+hG_0, \tag{3.3}$$

where A is a matrix given by (2.16). Then, from (2.27) and (2.29), follows

$$|A(h)|=|T(A+hG)T^{-1}|\leqq|T||T^{-1}|(1+|h||G|), \tag{3.4}$$

where

$$G=T^{-1}G_0T. \tag{3.5}$$

On the other hand, by the definition of present φ_n and (1.22), for any positive number ε, there exist postive numbers γ, h_2 and l_0 such that

$$|\varphi_n(e_{n+1}, e_n)|\leqq\varepsilon(|e_{n+1}|+|e_n|) \tag{3.6}$$

whenever $|e_{n+1}|$, $|e_n|\leqq\gamma$, $|h|\leqq h_2$ and $L_0\leqq l_0$. Also, by stable convergence of integration formulas, if $|e_0|$ and $|h|$ are sufficiently small, it holds always that

$$|e_n|\leqq\gamma. \tag{3.7}$$

Therefore, by (3.4) and (3.6), we have:

$$\begin{aligned}
&|h\{A^{n-1}(h)\varphi_{n_0}+A^{n-2}(h)\varphi_{n_0+1}+\cdots+A(h)\varphi_{n_0+n-2}+\varphi_{n_0+n-1}\}|\\
&\leqq|h||T||T^{-1}|\{(1+|h||G|)^{n-1}+(1+|h||G|)^{n-2}+\cdots+1\}\varepsilon\cdot 2\gamma\\
&=\frac{2\varepsilon\gamma|T||T^{-1}|}{|G|}\{(1+|h||G|)^n-1\}\\
&\leqq\frac{2\varepsilon\gamma|T||T^{-1}|}{|G|}(e^{|G|L_0}-1).
\end{aligned}$$

This says that, in the right-hand side of (3.2), the sum

$$h\{A^{n-1}(h)\varphi_{n_0}+A^{n-2}(h)\varphi_{n_0+1}+\cdots+A(h)\varphi_{n_0+n-2}+\varphi_{n_0+n-1}\}$$

is small compared with the magnitudes of $|e_n|$[1]. So, neglecting this sum, we can write (3.2) approximately as follows:

$$\begin{aligned}
e_{n+n_0}=&A^n(h)e_{n_0}\\
&+h\{A^{n-1}(h)r_{n_0}+A^{n-2}(h)r_{n_0+1}+\cdots+A(h)r_{n_0+n-2}+r_{n_0+n-1}\}.
\end{aligned} \tag{3.8}$$

In the sequel, we shall call this formula the *local approximate error formula.*

3.2 Natural additional conditions for multi-step formulas

For multi-step formulas (1.1) and (1.27), let us consider the polynomials

1) Hitherto, this fact seems to have been assumed without strict proof.

$$(3.9)\qquad \begin{cases} \rho(\lambda)=\lambda^k-\sum_{l=1}^{k}\alpha_l\lambda^{k-l}, \\ \sigma(\lambda)=\sum_{l=0}^{k}\beta_l\lambda^{k-l}, \\ \tau(\lambda)=\sum_{l=0}^{k}\hat{\alpha}_l\lambda^{k-l} \qquad (\hat{\alpha}_k=0). \end{cases}$$

The first of these polynomials is $\rho(\lambda)$ as defined in (2.7).

For a simple multi-step formula, as has been remarked by Dahlquist [3, 4], we may assume that $\rho(\lambda)$ and $\sigma(\lambda)$ are relatively prime. For, if $\rho(\lambda)$ and $\sigma(\lambda)$ have a common factor $d(\lambda)\neq$const. so that

$$\rho(\lambda)=d(\lambda)\rho_1(\lambda) \quad \text{and} \quad \sigma(\lambda)=d(\lambda)\sigma_1(\lambda),$$

then, by means of the operators E and D such that

$$(3.10)\qquad Ex_n=x_{n+1} \quad \text{and} \quad Dx_n=\dot{x}_n,$$

the integration formula (1.1) can be written as follows:

$$\{\rho(E)-h\,\sigma(E)D\}x_n=d(E)\{\rho_1(E)-h\,\sigma_1(E)D\}x_n=0,$$

which says that the initial integration formula (1.1) can be reduced to the simpler one

$$\{\rho_1(E)-h\,\sigma_1(E)D\}x_n=0$$

of the same form but with smaller k.

For a compound multi-step formula, $\alpha_k=\beta_k=\hat{\alpha}_k=0$, consequently the polynomials of (3.9) are written as

$$(3.11)\qquad \begin{cases} \rho(\lambda)=\lambda\rho_1(\lambda), \\ \sigma(\lambda)=\lambda\sigma_1(\lambda), \\ \tau(\lambda)=\lambda\tau_1(\lambda), \end{cases}$$

where $\rho_1(\lambda)$, $\sigma_1(\lambda)$ and $\tau_1(\lambda)$ are polynomials of degree at most $k-1$. And, by means of the operators of (3.10), the integration formula (1.27) is written as follows:

$$(3.12)\qquad \begin{cases} [\rho(E)-h\{\beta_{-1}E^{k+1}+\sigma(E)\}D]x_n=0, \\ E^{k+1}x_n=\tau(E)x_n+h\sum_{l=0}^{k}\hat{\beta}_l E^{k-l}Dx_n. \end{cases}$$

Consequently, by eliminating $E^{k+1}Dx_n$, there is obtained

$$(3.13)\qquad [E\{\rho_1(E)-h(\beta_{-1}\tau_1(E)+\sigma_1(E))D\}-h^2\beta_{-1}\sum_{l=0}^{k}\hat{\beta}_l E^{k-l}D^2]x_n=0.$$

But, as is seen from (1.44), the term $h^2\beta_{-1}\sum_{l=0}^{k}\hat{\beta}_l E^{k-l}D^2x_n$ affects the error only

in the term $h\,\hat{\boldsymbol{\varphi}}_n(\boldsymbol{e}_{n+1}, \boldsymbol{e}_n)$ of (1.45). However, as is shown in like manner as in the preceding paragraph, the term $h\,\hat{\boldsymbol{\varphi}}_n(\boldsymbol{e}_{n+1}, \boldsymbol{e}_n)$ has little effect on the errors themselves, because such effect is very small compared with the magnitudes of the errors[1]. Therefore, neglecting the term $h^2\beta_{-1}\sum_{l=0}^{k}\hat{\beta}_l E^{k-l}D^2 x_n$, we may write (3.13) approximately as follows:

$$[\rho_1(E)-h\{\beta_{-1}\tau_1(E)+\sigma_1(E)\}D]x_n=0. \tag{3.14}$$

Then, as in the case of simple multi-step formulas, we may assume that $\rho_1(\lambda)$ and $\beta_{-1}\tau_1(\lambda)+\sigma_1(\lambda)$ are relatively prime.

Thus, in the sequel, for multi-step formulas, we shall assume the conditions mentioned above, namely the conditions that

$$\begin{cases} 1^\circ\ \textit{for simple multi-step formulas, } \rho(\lambda) \textit{ and } \sigma(\lambda) \textit{ are relatively prime;} \\ 2^\circ\ \textit{for compound multi-step formulas, } \rho_1(\lambda) \textit{ and } \beta_{-1}\tau_1(\lambda)+\sigma_1(\lambda) \textit{ are relatively prime.} \end{cases} \tag{3.15}$$

As is seen from the consistency conditions, these conditions are fulfilled for Adams' formula—namely simple multi-step formulas such that $\alpha_1=1$, $\alpha_2=\cdots=\alpha_k=0$; $\beta_k\neq 0$, and for compound multi-step formulas such that $\alpha_1=1$, $\alpha_2=\cdots=\alpha_{k-1}=0$, $\beta_{k-1}\neq 0$, $\hat{\alpha}_0=1$, $\hat{\alpha}_1=\cdots=\hat{\alpha}_{k-1}=0$[2].

Note: A proof that the term $h\,\hat{\boldsymbol{\varphi}}_n(\boldsymbol{e}_{n+1}, \boldsymbol{e}_n)$ has little effect on the errors themselves is sketched below.

From (1.45), follows

$$\begin{aligned} \boldsymbol{e}_n=&\hat{A}_{n-1}\hat{A}_{n-2}\cdots\hat{A}_0\,\boldsymbol{e}_0 \\ &+h(\hat{A}_{n-1}\cdots\hat{A}_1\,\hat{\boldsymbol{r}}_0+\hat{A}_{n-1}\cdots\hat{A}_2\,\hat{\boldsymbol{r}}_1+\cdots+\hat{A}_{n-1}\,\hat{\boldsymbol{r}}_{n-2}+\hat{\boldsymbol{r}}_{n-1}) \\ &+h(\hat{A}_{n-1}\cdots\hat{A}_1\,\hat{\boldsymbol{\varphi}}_0+\hat{A}_{n-1}\cdots\hat{A}_2\,\hat{\boldsymbol{\varphi}}_1+\cdots+\hat{A}_{n-1}\,\hat{\boldsymbol{\varphi}}_{n-2}+\hat{\boldsymbol{\varphi}}_{n-1}). \end{aligned}$$

If we put

$$\hat{A}_n(h)=A+h\,G_n \quad\text{and}\quad \max_n|T^{-1}G_nT|=G,$$

then, in the same way as in the preceding paragraph, assuming

$$|\hat{\boldsymbol{\varphi}}_n(\boldsymbol{e}_{n+1}, \boldsymbol{e}_n)|\leqq\varepsilon(|\boldsymbol{e}_{n+1}|+|\boldsymbol{e}_n|) \quad\text{and}\quad |\boldsymbol{e}_n|\leqq\gamma,$$

we have:

$$\begin{aligned} &|h(\hat{A}_{n-1}\cdots\hat{A}_1\,\hat{\boldsymbol{\varphi}}_0+\hat{A}_{n-1}\cdots\hat{A}_2\,\hat{\boldsymbol{\varphi}}_1+\cdots+\hat{A}_{n-1}\,\hat{\boldsymbol{\varphi}}_{n-2}+\hat{\boldsymbol{\varphi}}_{n-1})| \\ &\leqq 2\varepsilon\gamma|T||T^{-1}|\cdot\frac{1}{G}(e^{GL}-1). \end{aligned}$$

Since ε is a small arbitrary number, this proves the desired fact.

---

1) See the note added at the end of this paragraph.

2) The formula applied to van der Pol's equation in the paper [18] is of this type with $k=4$.

3.3 Eigenvalues of $A(h)$

Let us consider the characteristic equation of $A(h)$:

$$\{A(h)-\lambda I\}\boldsymbol{c}=0. \tag{3.16}$$

Here $\boldsymbol{c}$ is a kN-dimensional eigenvector such that

$$\boldsymbol{c}=\begin{pmatrix}\boldsymbol{c}_1\\ \boldsymbol{c}_2\\ \vdots\\ \boldsymbol{c}_N\end{pmatrix}, \tag{3.17}$$

where $\boldsymbol{c}_\nu$ ($\nu=1, 2,\cdots, N$) are k-dimensional vectors such that

$$\boldsymbol{c}_\nu=\begin{pmatrix}c_\nu^1\\ c_\nu^2\\ \vdots\\ c_\nu^k\end{pmatrix}. \tag{3.18}$$

Then, by (1.16), from (3.16), it follows that

$$\begin{cases} -\lambda c_\nu^1+c_\nu^2=0,\\ -\lambda c_\nu^2+c_\nu^3=0,\\ \quad\vdots\\ -\lambda c_\nu^{k-1}+c_\nu^k=0,\\ -\lambda c_\nu^k+\sum\limits_{\mu=1}^{N}\sum\limits_{l=1}^{k}A_{\mu l}^{\nu}(h)\,c_\mu^{k+1-l}=0 \end{cases} \tag{3.19}$$

$$(\nu=1, 2,\cdots, N).$$

Since from the first $(k-1)$ equations follows

$$c_\nu^2=c_\nu^1\lambda,\ c_\nu^3=c_\nu^1\lambda^2,\cdots,\ c_\nu^k=c_\nu^1\lambda^{k-1},$$

(3.18) can be rewritten as follows:

$$\boldsymbol{c}_\nu=c_\nu\begin{pmatrix}1\\ \lambda\\ \lambda^2\\ \vdots\\ \lambda^{k-1}\end{pmatrix}. \tag{3.20}$$

Then the last equation of (3.19) can be written as follows:

$$-c_\nu\lambda^k+\sum_{\mu=1}^{N}\sum_{l=1}^{k}A_{\mu l}^{\nu}(h)\,c_\mu\,\lambda^{k-l}=0 \qquad (\nu=1, 2,\cdots, N). \tag{3.21}$$

This equation can be rewritten in accordance with the respective definitions of $A^{\nu}_{\mu l}(h)$—(1.15), (1.44) and (1.73)—as follows:

$$h\{\sigma(\lambda)-\beta_0\rho(\lambda)\}\sum_{\mu=1}^{N}F^{\nu}_{\mu}c_{\mu}=\rho(\lambda)c_{\nu} \qquad (\nu=1, 2,\cdots, N) \tag{3.22}$$

for simple multi-step formulas;

$$h[\beta_{-1}\tau(\lambda)+\sigma(\lambda)-\{\beta_{-1}\hat{\alpha}_0+\beta_0\}\rho(\lambda)]\sum_{\mu=1}^{N}F^{\nu}_{\mu}c_{\mu}=\rho(\lambda)c_{\nu} \qquad (\nu=1, 2,\cdots, N) \tag{3.23}$$

for compound multi-step formulas;

$$h\sum_{\mu=1}^{N}F^{\nu}_{\mu}c_{\mu}=(\lambda-1)c_{\nu} \qquad (\nu=1, 2,\cdots, N) \tag{3.24}$$

for the general Runge-Kutta formulas,
where $F=(F^{\nu}_{\mu})$ is the common approximate value adopted for the $F(x_n, t_n)$'s in I_0.

Now, when $|h|$ is small, as is seen from the forms of $A(h)$, the eigenvalues of $A(h)$ lie near the eigenvalues of $A(0)$, namely of A. But, from (2.16) and (2.7), the eigenvalues of A are the roots of the equation

$$\{\rho(\lambda)\}^N=0. \tag{3.25}$$

Hence, in order to seek the eigenvalues of $A(h)$ for small $|h|$, it suffices to consider the equations (3.22)—(3.24) only in the neighborhood of the roots of (3.25).

First, let us consider the equation (3.22). Let λ_0 be any root of (3.25). Then evidently λ_0 is a root of (2.7). Let m be its multiplicity as a root of (2.7). Then, in the neighborhood of λ_0, $\rho(\lambda)$ can be expressed as follows:

$$\rho(\lambda)=\frac{\rho^{(m)}(\lambda_0)}{m!}(\lambda-\lambda_0)^m+\cdots^{1)} \qquad (\rho^{(m)}(\lambda_0)\neq 0). \tag{3.26}$$

Consequently, in the neighborhood of λ_0, the equation (3.22) can be written as

$$h\{\sigma(\lambda_0)+\sigma'(\lambda_0)(\lambda-\lambda_0)+\cdots-\beta_0\frac{\rho^{(m)}(\lambda_0)}{m!}(\lambda-\lambda_0)^m-\cdots\}\sum_{\mu=1}^{N}F^{\nu}_{\mu}c_{\mu}$$
$$=\left\{\frac{\rho^{(m)}(\lambda_0)}{m!}(\lambda-\lambda_0)^m+\cdots\right\}c_{\nu} \qquad (\nu=1, 2,\cdots, N). \tag{3.27}$$

But, by 1° of (3.15), $\sigma(\lambda_0)\neq 0$ and, by (3.20), at least one of c_{ν}'s $(\nu=1, 2,\cdots, N)$ is not zero. Consequently, from (3.27), we have:

---

1) The dots denote a sum of terms of higher order than those written explicitly. This convention is used without notice in the sequel.

$$\frac{\rho^{(m)}(\lambda_0)}{m!}(\lambda-\lambda_0)^m+\cdots \tag{3.28}$$

$$=\kappa h\{\sigma(\lambda_0)+\sigma'(\lambda_0)(\lambda-\lambda_0)+\cdots-\beta_0\frac{\rho^{(m)}(\lambda_0)}{m!}(\lambda-\lambda_0)^m-\cdots\},$$

where κ is an eigenvalue of F. From this it readily follows that

$$\lambda=\lambda_0+h^{1/m}\left(\frac{m!\,\sigma(\lambda_0)}{\rho^{(m)}(\lambda_0)}\kappa\right)^{1/m}+\cdots. \tag{3.29}$$

This is an expansion formula for an eigenvalue of $A(h)$. Now F has N eigenvalues κ_ν $(\nu=1, 2,\cdots, N)$, consequently, by (3.29), in the neighborhood of λ_0, we obtain mN eigenvalues of $A(h)$ as follows:

$$\lambda=\lambda_0+h^{1/m}\left(\frac{m!\,\sigma(\lambda_0)}{\rho^{(m)}(\lambda_0)}\kappa_\nu\right)^{1/m}+\cdots \qquad (\nu=1, 2,\cdots, N). \tag{3.30}$$

Here, of course, the m-th root represents any one of m values of the m-th root. Since the number of roots λ_0 (multiplicity being taken into consideration) is k, the total number of the eigenvalues given by (3.30) is just kN. This says that all the eigenvalues of $A(h)$ are given by (3.30) as should be so.

Secondly, let us consider the equation (3.23). As in the former case, let λ_0 be any root of (3.25), namely any root of (2.7), and let m be its multiplicity as a root of (2.7). Then, since (3.11) holds for compound multi-step formulas, in the present case, λ_0 becomes a root of multiplicity m or $m-1$ of the equation

$$\rho_1(\lambda)=0 \tag{3.31}$$

according as $\lambda_0\neq 0$ or $=0$ and further the equation (3.23) becomes

$$\begin{aligned}&\lambda=0,\\ &h\left[\beta_{-1}\tau_1(\lambda)+\sigma_1(\lambda)-\{\beta_{-1}\hat{\alpha}_0+\beta_0\}\rho_1(\lambda)\right]\sum_{\mu=1}^{N}F_\mu^\nu c_\mu=\rho_1(\lambda)c_\nu\\ &\qquad\qquad (\nu=1, 2,\cdots, N).\end{aligned} \tag{3.32}$$

The first equation of (3.32) expresses the fact that (3.23) holds for arbitrary values of c_ν for $\lambda=0$, in other words that $A(h)$ *has N linearly independent eigenvectors for a zero eigenvalue.* This says that $A(h)$ *has at least N zero eigenvalues.*

The second set of equations of (3.32) is of the same form as (3.22) because $\rho_1(\lambda)$ and $\beta_{-1}\tau_1(\lambda)+\sigma_1(\lambda)$ are relatively prime by 2° of (3.15). Therefore, as for (3.22), in the neighborhood of the root λ_0 of (3.31), there are obtained mN eigenvalues

$$\lambda=\lambda_0+h^{1/m}\left[\frac{m!\{\beta_{-1}\tau_1(\lambda_0)+\sigma_1(\lambda_0)\}}{\rho_1{}^{(m)}(\lambda_0)}\kappa_\nu\right]^{1/m}+\cdots \tag{3.33}$$

$$(\nu=1, 2,\cdots, N)$$

for $\lambda_0 \neq 0$ and $(m-1)N$ eigenvalues

$$\lambda = h^{1/(m-1)}\left[\frac{(m-1)!\{\beta_{-1}\tau_1(0)+\sigma_1(0)\}}{\rho_1^{(m-1)}(0)}\kappa_\nu\right]^{1/(m-1)} + \cdots \quad (\nu=1, 2, \cdots, N) \tag{3.34}$$

for $\lambda_0 = 0$.

As is readily seen, the total number of the eigen values obtained above is just kN. Then, by letting F_μ^ν vary by a small amount if necessary, we see that all the eigenvalues of $A(h)$ are given by (3.33), (3.34) and N zeros[1].

Lastly, let us consider the equation (3.24). This equation expresses the fact that $(\lambda-1)/h$ is an eigenvalue of F. Therefore we see that, in this case, all the eigenvalues of $A(h)$ are given by

$$\lambda = 1 + h\kappa_\nu \qquad (\nu=1, 2, \cdots, N), \tag{3.35}$$

where κ_ν $(\nu=1, 2, \cdots, N)$ are eigenvalues of F.

The above results are summarized as follows:

The eigenvalues of $A(h)$ are given by the formulas of the forms:

1° (3.30) *for simple multi-step formulas*;

2° (3.33), (3.34) *and N zeros for compound multi-step formulas*;

3° (3.35) *for the general Runge-Kutta formulas,*

where λ_0 is an m-ple root of (2.7) *and κ_ν $(\nu=1, 2, \cdots, N)$ are the eigenvalues of F.*

Here, without loss of generality, we may assume that the eigenvalues κ_ν $(\nu=1, 2, \cdots, N)$ are all distinct and differ from zero, because, by our assumption, F can be varied arbitrarily by a small amount. Under this assumption, the eigenvalues of $A(h)$ become all distinct except for N zeros which appear for compound multi-step formulas. But, for such N zero eigenvalues, there exist N linearly independent eigenvectors as was shown already. Thus it follows that, under the present assumption, *the Jordan canonical form of $A(h)$ is diagonal.*

3.4 Formulas for propagation of errors

In this paragraph, using the results of the preceding paragraph, we rewrite the local approximate error formula (3.8) so that the behavior of the propagation of errors may appear more explicitly.

First, let us note the following fact which follows readily from the forms

1) When some of the values given by (3.34) are zero, we let F_μ^ν vary by a small amount so that the values given by (3.34) may all differ from zero. Then, for such $A(h)$, our assertion evidently holds. Then, by letting such $A(h)$ vary back to the initial $A(h)$, we see our assertion holds also for the initial $A(h)$ on account of continuity of eigenvalues.

of $A(h)$ and $\boldsymbol{r}$ (cf. (1.16) and (1.20)):

When $A^p(h)\,\boldsymbol{r}$ $(p=0, 1, 2, \cdots)$ are expressed as

$$A^p(h)\,\boldsymbol{r}=\begin{pmatrix} \boldsymbol{v}_p^1 \\ \boldsymbol{v}_p^2 \\ \vdots \\ \boldsymbol{v}_p^N \end{pmatrix},$$

where $\boldsymbol{v}_p^\nu$ $(\nu=1, 2, \cdots, N)$ are k dimensional vectors, the first $(k-p-1)$ components of $\boldsymbol{v}_p^\nu$ are all zero.

From this fact it readily follows that the first components of $\boldsymbol{v}_p^\nu$ are all zero for $p=0, 1, \cdots, k-2$.

Let $\mathring{A}(h)$ be the Jordan form of $A(h)$ such that

$$U^{-1}A(h)\,U=\mathring{A}(h).$$

Then, as is mentioned in the preceding paragraph, $\mathring{A}(h)$ is a diagonal matrix whose diagonal elements are the eigenvalues $\lambda_{\nu i}$ $(\nu=1, 2, \cdots, N;\ i=1, 2, \cdots, k)$ of $A(h)$ determined in the preceding paragraph. Consequently, if we put

$$U=(u_{\mu j}^{\nu i}),\quad U^{-1}=(U_{\mu j}^{\nu i})$$
$$(\nu, \mu=1, 2, \cdots N;\ i, j=1, 2, \cdots, k),$$

from the local approximate error formula (3.8), we have

$$e_{n_0+n}^\nu=\sum_{j,l=1}^{k}\sum_{\mu,\omega=1}^{N} u_{\mu j}^{\nu 1}\,\lambda_{\mu j}^n\,U_{\omega l}^{\mu j}\,e_{n_0+l-1}^\omega + \sum_{p=k-1}^{n-1}\sum_{j=1}^{k}\sum_{\mu,\omega=1}^{N} u_{\mu j}^{\nu 1}\,\lambda_{\mu j}^p\,U_{\omega k}^{\mu j}\,S_{n_0-1+n-p}^\omega \qquad (3.36)$$
$$(\nu=1, 2, \cdots, N),$$

because

$$\sum_{j=1}^{k}\sum_{\mu,\omega=1}^{N} u_{\mu j}^{\nu 1}\,\lambda_{\mu j}^p\,U_{\omega k}^{\mu j}\,S_{n_0-1+n-p}^\omega=0 \quad \text{for} \quad p=0, 1, \cdots, k-2$$

as is remarked above.

But, since $|h|$ and $|\boldsymbol{e}_n|$ are small and, in actual computation, all bounds of round-off errors are of the same order, by (1.11), (1.40), (1.42) and (1.73) respectively, it holds excluding the small quantities of higher order that

$$S_n=R_n-T_n \qquad \text{for simple multi-step formulas,} \qquad (3.37)$$

$$S_n=R_n-(T_n+h\beta_{-1}\,F\,\hat{T}_n) \qquad \text{for compound multi-step formulas,} \qquad (3.38)$$

$$S_n=R_n-T_n \qquad \text{for the general Runge-Kutta formulas.} \qquad (3.39)$$

Then, by the same reason as that by which (3.2) has been written approximately as (3.8), the formula (3.36) can be written approximately as follows:

$$
\begin{aligned}
(3.40)\qquad e^{\nu}_{n_0+n} = &\sum_{j,l=1}^{k}\sum_{\mu,\omega=1}^{N} u^{\nu 1}_{\mu j}\,\lambda^{n}_{\mu j}\,U^{\mu j}_{\omega l}\,e^{\omega}_{n_0+l-1} \\
&+\sum_{p=k-1}^{n-1}\sum_{j=1}^{k}\sum_{\mu,\omega=1}^{N} u^{\nu 1}_{\mu j}\,\lambda^{p}_{\mu j}\,U^{\mu j}_{\omega k}\,R^{\omega}_{n_0-1+n-p} \\
&-\sum_{p=k-1}^{n-1}\sum_{j=1}^{k}\sum_{\mu,\omega=1}^{N} u^{\nu 1}_{\mu j}\,\lambda^{p}_{\mu j}\,U^{\mu j}_{\omega k}\,\hat{T}^{\omega}_{n_0-1+n-p} \\
&\qquad\qquad (\nu=1, 2,\cdots, N),
\end{aligned}
$$

where

$$
(3.41)\qquad \hat{T}_n = \begin{cases} T_n & \text{for simple multi-step and the general Runge-Kutta formulas,} \\ T_n + h\beta_{-1} F\hat{T}_n & \text{for compound multi-step formulas.} \end{cases}
$$

The formula (3.40) is the desired formula for propagation of errors. Indeed it expresses explicitly the behavior of growth of errors, or, in other words, the behavior of propagation of errors.

3.5 Analysis of propagation of errors

As is seen from (3.40), the rates of growth of errors are the eigenvalues $\lambda_{\nu i}$ $(\nu=1, 2,\cdots, N;\ i=1, 2,\cdots, k)$ of $A(h)$ determined in **3.3**. Hence, for integration formulas, it is desirable that these $\lambda_{\nu i}$ are all small as possible in absolute value. But, by the consistency conditions (2.1) and (2.2), one of the λ_0 is 1 and, by the stability conditions (2.13), $|\lambda_0|\leqq 1$ for all λ_0's. Therefore, from the present point of view, among the various integration formulas under consideration, the best will be the general Runge-Kutta formula or the multi-step formula for which

$$
(3.42)\qquad \alpha_1=1,\quad \alpha_2=\alpha_3=\cdots=\alpha_k=0,
$$

and the worst is the formula for which all λ_0's are 1 in absolute value.

The simple multi-step formulas for which (3.42) is valid are exactly Adams' formulas. For Adams' formulas, due to (2.1), from (3.30), the values of $\lambda_{\nu i}$ are found as follows:

$$
\begin{aligned}
(3.43)\qquad &\lambda_{\nu 1}=1+h\kappa_\nu+\cdots, \\
&\lambda_{\nu i}=h^{1/(k-1)}(-\beta_k\kappa_\nu)^{1/(k-1)}+\cdots \\
&\qquad\qquad (i=2, 3, \cdots, k;\ \nu=1, 2,\cdots, N).
\end{aligned}
$$

For the compound multi-step formulas for which (3.42) is valid, due to (2.2), from (3.33) and (3.34), the values of $\lambda_{\nu i}$ are found as follows:

$$
\begin{cases}
\lambda_{\nu 1}=1+h\,\kappa_\nu+\cdots, & \\
\lambda_{\nu i}=h^{1/(k-2)}\{-(\beta_{-1}\hat{\alpha}_{k-1}+\beta_{k-1})\kappa_\nu\}+\cdots & \\
\qquad\qquad (i=2,\,3,\cdots,\,k-1), & \\
\lambda_{\nu k}=0 \qquad\qquad (\nu=1,\,2,\cdots,\,N). &
\end{cases} \tag{3.44}
$$

For the general Runge-Kutta formulas, from (3.35), the values of λ_ν are evidently

$$
\lambda_\nu=1+h\,\kappa_\nu \qquad (\nu=1,\,2,\cdots,\,N). \tag{3.45}
$$

Comparison of (3.43), (3.44) and (3.45) yields the conclusion:

Among the multi-step formulas and the general Runge-Kutta formulas, the best three in the sense that growth of errors is least are as follows:

1*st: the general Runge-Kutta formula,*

2*nd: the compound multi-step formula for which* (3.42) *is valid,*

3*rd: Adams' formula.*

But, for the integration formulas, as is seen from (3.40), it is desirable that, besides the rates of growth of errors, the truncation errors are also small as possible. As is well known, the truncation errors of the general Runge-Kutta formulas and the multi-step formulas for which (3.42) is valid are respectively $O(|h|^5)$ and $O(|h|^{k+2})$. However, as has been shown by Dahlquist [3, 4], among the simple multi-step formulas, there are formulas such that their truncation errors are $O(|h|^{k+3})$. But such formulas can exist only for even k and moreover, for such formulas, the roots λ_0 of the equation (2.7) are all 1 in absolute value [3, 4]. As is stated in the beginning of this paragraph, this means that such formulas are the worst ones in the sense that growth of errors is largest.

As is seen from the second of (3.41) and (3.44), for a compound multi-step formula satisfying (3.42), the effect of the second formula on both growth of errors and truncation errors is small. Consequently, without any serious change of effects, we may take any integration formula as the second formula. For this reason and for simplicity, let us take, for the present, Adams' extrapolation formula as the second formula.

formulas / k	Adams' formulas		compound multi-step formulas
	extrapolation ($\beta_0=0$)	interpolation ($\beta_0\neq 0$)	
2	$\frac{5}{12}h^3x^{\mathrm{III}}$	$-\frac{1}{24}h^4x^{\mathrm{IV}}$	$\frac{1}{24}h^4x^{\mathrm{IV}}-\frac{1}{32}h^5Fx^{\mathrm{IV}}$
3	$\frac{3}{8}h^4x^{\mathrm{IV}}$	$-\frac{19}{720}h^5x^{\mathrm{V}}$	$\frac{11}{720}h^5x^{\mathrm{V}}-\frac{251}{17280}h^6Fx^{\mathrm{V}}$
4	$\frac{251}{720}h^5x^{\mathrm{V}}$	$-\frac{3}{160}h^6x^{\mathrm{VI}}$	$\frac{11}{1440}h^6x^{\mathrm{VI}}-\frac{361}{41472}h^7Fx^{\mathrm{VI}}$
5	$\frac{95}{288}h^6x^{\mathrm{VI}}$		

The above table shows the values of $\tilde{T}_n$ of multi-step formulas satisfying (3.42), namely of the Adams' formulas and of the compound multi-step formulas mentioned just above. From this table, it is readily seen that the values of $\tilde{T}_n$ of the compound multi-step formulas under consideration are less than those of the Adams' formulas. Combined with the first results about growth of errors, this says that *the compound multi-step formulas mentioned above are always preferable to the Adams' formulas with respect to both growth of errors and truncation errors.*

But, *for* $k=3$, *since the truncation errors of the multi-step formulas are of the same order as those of the general Runge-Kutta formulas, the latter will be preferable to the former on account of their superiority with respect to growth of errors.*

The simple multi-step formulas given by Dahlquist, namely those of which the truncation errors are $O(|h|^{k+3})$ will not be preferable when the formulas are applied in many steps, because the growth of errors is very large as is mentioned formerly.

Chapter IV. Estimation of errors

The formula (2.41) or (2.42) is, of course, a kind of estimates of errors. But, as is seen from its derivation, it is too crude. So, in this chapter, assuming Lipschitz conditions upon $F(x, t)$, we shall derive a more precise estimate of errors by improving the above estimate.

In this chapter, we are concerned with estimation of errors of the approximate numerical solution obtained in the domain D by means of the multistep formulas satisfying both the consistency conditions and the stability conditions or by means of the general Runge-Kutta formulas. Since these formulas are stable as is shown in **2.4**, it is needless to say that, by means of these formulas, there is actually constructed a numerical solution in the domain D.

4.1 Lemmas on matrices

For estimation of errors, Dahlquist [4] introduced the quantities like

$$\lim_{h\to+0}\frac{|I+hQ|-1}{h}=\mu[Q],$$

where I is a unit matrix and Q is an arbitrary matrix. In the present paper, generalizing the above quantities, we introduce the quantities like

$$\lim_{h\to\pm0}\frac{|P+hQ|-|P|}{h}=\mu_P^{\pm}[Q], \tag{4.1}$$

where P and Q are arbitrary matrices.

In this paragraph, about such quantities, some lemmas which will be necessary for estimation of errors are stated.

Lemma. 1. *For arbitrary matrices P and Q, there exist always the limits in the left-hand side of* (4.1).

This lemma follows readily from convexity of the function of x: $|P+xQ|$.

Lemma 2. *When norms of matrices are defined as in Chapter I, for $P=(p_\mu^\nu)$ and $Q=(q_\mu^\nu)$ $(\nu, \mu=1, 2, \cdots, N)$, $\mu_P^{\pm}[Q]$ are given by*

$$\mu_P^{\pm}[Q]=\left(\begin{matrix}\max\\ \min\\ \alpha\end{matrix}\right)\left\{\frac{1}{2}\sum_\nu{}'\frac{p_\nu^\alpha\overline{q_\nu^\alpha}+\overline{p_\nu^\alpha}q_\nu^\alpha}{|p_\nu^\alpha|}\pm\sum_\mu{}''|q_\mu^\alpha|\right\}^{1)}, \tag{4.2}$$

1) This means

$$\mu_P^{+}[Q]=\max_\alpha\left\{\frac{1}{2}\sum_\nu{}'\frac{p_\nu^\alpha\overline{q_\nu^\alpha}+\overline{p_\nu^\alpha}q_\nu^\alpha}{|p_\nu^\alpha|}+\sum_\mu{}''|q_\mu^\alpha|\right\},$$

$$\mu_P^{-}[Q]=\min_\alpha\left\{\frac{1}{2}\sum_\nu{}'\frac{p_\nu^\alpha\overline{q_\nu^\alpha}+\overline{p_\nu^\alpha}q_\nu^\alpha}{|p_\nu^\alpha|}-\sum_\mu{}''|q_\mu^\alpha|\right\}.$$

where

the bars over the letters denote conjugate imaginaries;

α *are the indices such that* $|P| = \sum_{\nu=1}^{N} |p_\nu^\alpha|$;

$\sum_\nu'$ *is a sum over the indices* ν *such that* $|p_\nu^\alpha| \neq 0$;

$\sum_\mu''$ *is a sum over the indices* μ *such that* $|p_\mu^\alpha| = 0$.

Proof. Let κ be a number such that

$$\kappa = \begin{cases} +1 & \text{for } h>0, \\ -1 & \text{for } h<0. \end{cases}$$

From the definition of α, for sufficiently small $|h|$,

$$|P+hQ| = \max_\alpha \sum_{\nu=1}^{N} |p_\nu^\alpha + h q_\nu^\alpha|. \tag{4.3}$$

Since

$$|p_\nu^\alpha + h q_\nu^\alpha| = \begin{cases} |p_\nu^\alpha| + \dfrac{h}{2} \dfrac{p_\nu^\alpha \bar{q}_\nu^\alpha + \bar{p}_\nu^\alpha q_\nu^\alpha}{|p_\nu^\alpha|} + o(|h|) & \text{for } |p_\nu^\alpha| \neq 0, \\ \kappa h |q_\nu^\alpha| & \text{for } |p_\nu^\alpha| = 0, \end{cases}$$

from (4.3), it follows that

$$|P+hQ| = |P| + \max_\alpha \left\{ \frac{h}{2} \sum_\nu' \frac{p_\nu^\alpha \bar{q}_\nu^\alpha + \bar{p}_\nu^\alpha q_\nu^\alpha}{|p_\nu^\alpha|} + \kappa h \sum_\mu'' |q_\mu^\alpha| + o(|h|) \right\},$$

from which (4.2) readily follows.

Corollary.

$$\mu_T^{\pm}[Q] = \begin{pmatrix} \max \\ \min \end{pmatrix}_\nu \left\{ \Re(q_\nu^\nu) \pm \sum_{\mu \neq \nu} |q_\mu^\nu| \right\}^{1)},$$

and, for Λ *of* (2.28),

$$\mu_\Lambda^{\pm}[Q] = \begin{pmatrix} \max \\ \min \end{pmatrix}_\alpha \left\{ \frac{1}{2} \left(e^{i\theta_\alpha} \bar{q}_\alpha^\alpha + e^{-i\theta_\alpha} q_\alpha^\alpha \right) \pm \sum_{\mu \neq \alpha} |q_\mu^\alpha| \right\}$$

where α *is a number of the row in* Λ *where* λ_i *such that* $|\lambda_i| = 1$ *is located and* $e^{i\theta_\alpha}$ *is a value of such* λ_i.

Lemma 3.

$$|\mu_P^{\pm}[Q]| \leqq |Q|. \tag{4.4}$$

This lemma follows readily from the inequalities

1) $\Re(q_\nu^\nu)$ means the real part of q_ν^ν.

$$|P|-|h|\,|Q|\leqq|P+hQ|\leqq|P|+|h|\,|Q|.$$

Lemma 4.

$$\mu_P^{\pm}[\lambda Q]=\begin{cases}\lambda\mu_P^{\pm}[Q] & \text{for } \lambda>0,\\ \lambda\mu_P^{\mp}[Q] & \text{for } \lambda<0,\end{cases} \tag{4.5}$$

$$\begin{cases}\mu_P^{+}[Q_1+Q_2]\leqq\mu_P^{+}[Q_1]+\mu_P^{+}[Q_2],\\ \mu_P^{-}[Q_1+Q_2]\geqq\mu_P^{-}[Q_1]+\mu_P^{-}[Q_2].\end{cases} \tag{4.6}$$

Proof. (4.5) follows readily from

$$\frac{|P+h\lambda Q|-|P|}{h}=\frac{|P+h\lambda\,Q|-|P|}{h\lambda}\cdot\lambda.$$

(4.6) follows from the inequalities:

$$\frac{|P+h\,(Q_1+Q_2)|-|P|}{h}=\frac{|(P+2h\,Q_1)+(P+2h\,Q_2)|-2|P|}{2h}$$

$$\leqq\text{ or }\geqq\frac{|P+2h\,Q_1|-|P|}{2h}+\frac{|P+2h\,Q_2|-|P|}{2h}$$

according as $h>0$ or $h<0$.

Lemma 5.

$$|\mu_P^{\pm}[Q_1]-\mu_P^{\pm}[Q_2]|\leqq|Q_1-Q_2|. \tag{4.7}$$

Proof. By (4.6) and (4.4), for arbitrary Q_1 and Q_2, it holds that

$$\mu_P^{+}[Q_1]-\mu_P^{+}[Q_2]\leqq\mu_P^{+}[Q_1-Q_2]\leqq|Q_1-Q_2| \tag{4.8}$$

and

$$\mu_P^{-}[Q_1]-\mu_P^{-}[Q_2]\geqq\mu_P^{-}[Q_1-Q_2]\geqq-|Q_1-Q_2|,$$

namely

$$\mu_P^{-}[Q_2]-\mu_P^{-}[Q_1]\leqq|Q_1-Q_2|. \tag{4.9}$$

Since Q_1 and Q_2 are arbitrary, (4.8) and (4.9) hold also when Q_1 and Q_2 are interchanged with each other. These inequalities combined with the initial ones prove (4.7).

Corollary. *$\mu_P^{\pm}[Q]$ is continuous with respect to Q.*

Lemma 6. *In the definition of $\mu_P^{\pm}[Q]$, the convergence is locally uniform with respect to Q.*

Proof. In the definition (4.1) of $\mu_P^{\pm}[Q]$, the quantity

$$\frac{|P+h\,Q|-|P|}{h}$$

converges to $\mu_P^{\pm}[Q]$ monotonically as $h\to\pm 0$ due to convexity of the function of $x:|P+x\,Q|$. And evidently the above quantity is continuous with respect to Q for fixed h. Besides, the limits $\mu_P^{\pm}[Q]$ are also continuous with respect to Q by Corollary of Lemma 5. Thus the lemma is valid by the theorem of Dini.

4.2 Preliminary estimation of some quantities

As is mentioned in the beginning of this chapter, we assume Lipschitz conditions upon $F(x, t)$ as follows:

$$(4.10)^{1)}\quad \begin{cases} |F(x', t)-F(x'', t)|\leqq L_1|x'-x''| & \text{for } (x', t),\ (x'', t)\in D, \\ |F\{x(t'), t'\}-F\{x(t''), t''\}|\leqq L_2|t'-t''| & \text{for } t', t''\in[t_0-L,\ t_0+L]. \end{cases}$$

Let us suppose that, in the domain D,

$$|f(x, t)|\leqq M_1,\quad |F(x, t)|\leqq M, \tag{4.11}$$

$$\begin{cases} |R_n|\leqq|h|\xi_n\leqq|h|\xi,\ |\hat{R}_n|\leqq\hat{\xi}_n\leqq\hat{\xi}, \\ |r'_{n1}|,\ |r'_{n2}|,\ |r'_{n3}|,\ |r'_{n4}|\leqq\eta_n\leqq\eta, \end{cases} \tag{4.12}$$

$$|T_n|\leqq|h|\zeta_n\leqq|h|\zeta,\ |\hat{T}_n|\leqq\hat{\zeta}_n\leqq\hat{\zeta}. \tag{4.13}$$

In (4.12) and (4.13), by the assumptions ((1.23), (1.24)), ((1.34), (1.35)) and, ((1.75), (1.23)), we may assume that

$$\xi,\ \zeta,\ \hat{\xi},\ \hat{\zeta},\ \eta\to 0 \quad\text{as}\quad |h|\to 0. \tag{4.14}$$

First, let us seek a rough estimate of $\max|\boldsymbol{e}_n|$ by means of the formula

$$(2.42')\qquad |\boldsymbol{e}'_n|\leqq\left(\exp\frac{K'_1+K'_2}{1-|h|K'_1}L\right)|\boldsymbol{e}'_0| + \frac{1}{K'_1+K'_2}\left\{\left(\exp\frac{K'_1+K'_2}{1-|h|K'_1}L\right)-1\right\}\varepsilon' \qquad (n=0, 1, 2, \cdots),$$

which is a slight modification of (2.42) and whose validity is evident from the derivation of (2.42).

In order to find K'_1, K'_2 and ε' in the above formula, we have only to know K_1, K_2, σ_n and $\boldsymbol{\rho}_n$ because of (2.34) and ((2.30), (2.35), (2.37)). Now, due to the second inequality of (4.11), we have a Lipschitz condition:

1) These are valid provided $F(x, t)\in C^1_{x,y}[D]$.

$$|f(x', t)-f(x'', t)| \leqq M|x'-x''| \qquad \text{for } (x', t), (x'', t) \in D. \tag{4.15}$$

Hence, after elementary calculations, we find:

for a simple multi-step formula,

$$\begin{cases} K_1=M|\beta_0|, \quad K_2=M\sum_{l=1}^{k}|\beta_l|, \\ \sigma_n=0, \quad |\rho_n| \leqq \xi_n+\zeta_n; \end{cases} \tag{4.16}$$

for a compound multi-step formula,

$$\begin{cases} K_1=M\{|\beta_0|+|\beta_{-1}|(|\hat{\alpha}_0|+|h|M|\hat{\beta}_0|)\}, \\ K_2=M\left\{\sum_{l=1}^{k-1}|\beta_l|+|\beta_{-1}|\left(\sum_{l=1}^{k-1}|\hat{\alpha}_l|+|h|M\sum_{l=1}^{k}|\hat{\beta}_l|\right)\right\}, \\ \sigma_n=M|\beta_{-1}|(\hat{\xi}_n+\hat{\zeta}_n), \quad |\rho_n| \leqq \xi_n+\zeta_n; \end{cases} \tag{4.17}$$

for the general Runge-Kutta formula,

$$\begin{aligned} &K_1=0, \quad K_2=MW, \\ &\sigma_n=W\eta_n, \quad |\rho_n| \leqq \xi_n+\zeta_n \end{aligned} \tag{4.18}$$

where

$$\begin{cases} M_2=M|m|, \\ M_3=M\{|p-r|+|r|(1+|h|M_2)\}, \\ M_4=M\{|q-s-u|+|s|(1+|h|M_2)+|u|(1+|h|M_3)\}, \\ W=|a|+|b|(1+|h|M_2)+|c|(1+|h|M_3)+|d|(1+|h|M_4). \end{cases} \tag{4.19}$$

From these, by (2.34) and ((2.30), (2.35), (2.37)), K_1', K_2' and ε' are found as follows:

$$K_1'=|T^{-1}||T|K_1, \quad K_2'=|T^{-1}||T|K_2; \tag{4.20}$$

$$\varepsilon'=\begin{cases} |T^{-1}|(\xi+\zeta) & \text{for a simple multi-step formula,} \\ |T^{-1}|\{(\xi+\zeta)+(\hat{\xi}+\hat{\zeta})M|\beta_{-1}|\} & \text{for a compound multi-step formula,} \\ (\xi+\zeta+\eta W) & \text{for the general Runge-Kutta formula.} \end{cases} \tag{4.21}$$

Hence, by substitution of these values for K_1', K_2' and ε' in (2.42'), a following rough estimate γ of $\max|e_n|$ is obtained:

$$\max|\boldsymbol{e}_n| \leqq \gamma = |T|\,|T^{-1}|\,|\boldsymbol{e}_0| \exp\left[\frac{K_1'+K_2'}{1-|h|K_1'}L\right] + \frac{|T|}{K_1'+K_2'}\left\{\left(\exp\frac{K_1'+K_2'}{1-|h|K_1'}L\right)-1\right\}\varepsilon' . \tag{4.22}$$

Next, let us seek estimates of the second and third terms in the right-hand sides of the error formulas (1.25), (1.45) and (1.77). These estimates are obtained after elementary calculations as follows:

for a simple multi-step formula,

$$|\boldsymbol{\varphi}_n| \leqq \varepsilon_1|\boldsymbol{e}_{n+1}| + \varepsilon_2|\boldsymbol{e}_n|, \quad |\boldsymbol{r}_n| \leqq \omega_n, \tag{4.23}$$

where

$$\begin{cases} \varepsilon_1 = \gamma \cdot \dfrac{L_1|\beta_0|}{1-|h|M|\beta_0|} + |h| \cdot \dfrac{kL_2|\beta_0|}{1-|h|M|\beta_0|}, \\ \varepsilon_2 = \gamma \cdot \dfrac{L_1\sum_{l=1}^{k}|\beta_l|}{1-|h|M|\beta_0|} + |h| \cdot \dfrac{\sum_{l=1}^{k}\left[M^2|\beta_0|^2|\alpha_l| + \{M^2|\beta_0| + (k-l)L_2\}|\beta_l|\right]}{1-|h|M|\beta_0|}, \\ \omega_n = \dfrac{\xi_n+\zeta_n}{1-|h|M|\beta_0|}, \end{cases} \tag{4.24}$$

and

$$|h| < \frac{1}{M|\beta_0|}; \tag{4.25}$$

for a compound multi-step formula,

$$|\hat{\boldsymbol{\varphi}}_n| \leqq \varepsilon_1|\boldsymbol{e}_{n+1}| + \varepsilon_2|\boldsymbol{e}_n|, \quad |\hat{\boldsymbol{r}}_n| \leqq \omega_n, \tag{4.26}$$

where

$$\begin{cases} L_3 = (k+1)L_2 + \gamma L_1\left\{(M+\gamma L_1)\sum_{l=0}^{k}|\hat{\beta}_l| + |h|L_2\sum_{l=0}^{k}(k-l)|\hat{\beta}_l|\right\}, \\ L_4 = M + \gamma L_1\sum_{l=0}^{k-1}|\hat{\alpha}_l| + |h|L_3 + (\hat{\xi}+\hat{\zeta})L_1, \end{cases} \tag{4.27}$$

$$\begin{aligned} \varepsilon_1 = \Big[&\gamma L_1\Big(|\beta_{-1}|\,|\hat{\alpha}_0|\sum_{l=0}^{k-1}|\hat{\alpha}_l| + |\beta_0|\Big) \\ &+ |h|\{L_3|\beta_{-1}|\,|\hat{\alpha}_0| + kL_2|\beta_0| + (M+\gamma L_1 + |h|kL_2)L_4|\beta_{-1}|\,|\hat{\beta}_0|\} \\ &+ (\hat{\xi}+\hat{\zeta})L_1|\beta_{-1}|\,|\hat{\alpha}_0|\Big] \Big/ \{1-|h|M(|\beta_{-1}|\,|\hat{\alpha}_0| + |\beta_0|)\}, \end{aligned}$$

$$
(4.28)\quad \begin{cases} \varepsilon_2 = \Big[\gamma L_1 \sum_{l=1}^{k-1}\Big(|\beta_{-1}|\,|\hat{\alpha}_l| \sum_{i=0}^{k-1} |\hat{\alpha}_i| + |\beta_l|\Big) + |h| \sum_{l=1}^{k} \{L_3 |\beta_{-1}|\,|\hat{\alpha}_l| + (k-l)L_2 |\beta_l| \\ \qquad + (M+\gamma L_1 + |h|(k-l)L_2)L_4 |\beta_{-1}|\,|\hat{\beta}_l| \\ \qquad + M^2(|\beta_{-1}|\,|\hat{\alpha}_0| + |\beta_0|)\,((|\beta_{-1}|\,|\hat{\alpha}_0| + |\beta_0|)|\alpha_l| + |\beta_{-1}|\,|\hat{\alpha}_l| + |\beta_l|)\} \\ \qquad + (\hat{\xi}+\zeta)L_1 |\beta_{-1}| \sum_{l=1}^{k-1} |\hat{\alpha}_l| \Big] \Big/ \{1 - |h| M(|\beta_{-1}|\,|\hat{\alpha}_0| + |\beta_0|)\}, \\ \omega_n = \dfrac{\xi_n + \zeta_n + (\hat{\xi}_n + \zeta_n) L_4 |\beta_{-1}|}{1 - |h| M(|\beta_{-1}|\,|\hat{\alpha}_0| + |\beta_0|)}, \end{cases}
$$

and

$$
(4.29)\quad |h| < \frac{1}{M(|\beta_{-1}|\,|\hat{\alpha}_0| + |\beta_0|)};
$$

for the general Runge-Kutta formula,

$$
(4.30)\quad |\varphi_n| \leqq \varepsilon_2 |e_n|, \quad |r_n| \leqq \omega_n,
$$

where

$$
(4.31)\quad \begin{cases} L_5 = |m|\,\{L_2 + 2L_1 M_1 + (M+\gamma L_1)^2 + \eta L_1\}, \\ L_6 = |p-r|\,\{L_1 M_1 + (M+\gamma L_1)^2 + \eta L_1\} \\ \qquad + |r|\,\{L_1 M_1 + (M+\gamma L_1)(M+\gamma L_1 + |h| L_5) \\ \qquad\quad + \eta L_1(1 + |h| M_2)\} + |p|(L_2 + L_1 M_1), \\ L_7 = |q-s-u|\,\{L_1 M_1 + (M+\gamma L_1)^2 + \eta L_1\} \\ \qquad + |s|\,\{L_1 M_1 + (M+\gamma L_1)(M+\gamma L_1 + |h| L_5) + \eta L_1(1 + |h| M_2)\} \\ \qquad + |u|\,\{L_1 M_1 + (M+\gamma L_1)(M+\gamma L_1 + |h| L_6) + \eta L_1(1 + |h| M_3)\} \\ \qquad + |q|(L_2 + L_1 M_1), \end{cases}
$$

$$
(4.32)\quad \begin{cases} \varepsilon_2 = \gamma L_1(|a| + |b| + |c| + |d|) + |h|(L_5 |b| + L_6 |c| + L_7 |d|), \\ \omega_n = \xi_n + \zeta_n + \eta_n W. \end{cases}
$$

Since we are concerned only with the case where $|h|$ is small, we may suppose that the conditions (4.25) and (4.29) are always fulfilled. Then, as the estimates of the second and third terms in the right-hand sides of the error formulas, we can really use (4.23) or (4.26) or (4.30) in accordance with the formulas used.

4.3 Estimation of errors

The error formulas (1.45) and (1.77) are of the same form as (1.25). So, in this paragraph, by (1.25), we shall represent all the error formulas under consideration, namely those given by (1.25), (1.45) and (1.77).

As is seen from (1.15), (1.44) and (1.73), $A_n(h)$ is of the form

$$A_n(h)=A+h\,G(t_n) \tag{4.33}$$

where A is a matrix given by (2.16). Consequently, from (2.27),

$$T^{-1}A_n(h)\,T=A+h\,G'_n, \tag{4.34}$$

where

$$G'_n=G'(t_n)=T^{-1}\,G(t_n)\,T. \tag{4.35}$$

Therefore, in like manner as (2.30) and (2.31), let us put

$$\begin{cases} \boldsymbol{e}_n=T\,\boldsymbol{e}'_n, \quad \boldsymbol{r}_n=T\,\boldsymbol{r}'_n, \\ T^{-1}\,\boldsymbol{\varphi}_n(T\,\boldsymbol{e}'_{n+1},\ T\boldsymbol{e}'_n)=\boldsymbol{\varphi}'_n(\boldsymbol{e}'_{n+1},\ \boldsymbol{e}'_n). \end{cases} \tag{4.36}$$

Then, by this substitution, the error formula (1.25) is rewritten as follows:

$$\boldsymbol{e}'_{n+1}=(A+h\,G'_n)\,\boldsymbol{e}'_n+h\,\boldsymbol{\varphi}'_n(\boldsymbol{e}'_{n+1},\ \boldsymbol{e}'_n)+h\,\boldsymbol{r}'_n\,. \tag{4.37}$$

Now, by (4.23), (4.26) and (4.30),

$$\begin{cases} |\boldsymbol{\varphi}_n(\boldsymbol{e}_{n+1},\ \boldsymbol{e}_n)|\leqq\varepsilon_1|\boldsymbol{e}_{n+1}|+\varepsilon_2|\boldsymbol{e}_n|, \\ |\boldsymbol{r}_n|\leqq\omega_n\,. \end{cases}$$

Consequently, if we put

$$\begin{cases} \varepsilon'_1=|T^{-1}|\,|T|\,\varepsilon_1,\ \varepsilon'_2=|T^{-1}|\,|T|\,\varepsilon_2, \\ \omega'_n=|T^{-1}|\,\omega_n\,, \end{cases} \tag{4.38}$$

by (4.36), it holds that

$$\begin{cases} |\boldsymbol{\varphi}'_n(\boldsymbol{e}_{n+1},\ \boldsymbol{e}_n)|\leqq\varepsilon'_1|\boldsymbol{e}'_{n+1}|+\varepsilon'_2|\boldsymbol{e}'_n|\,, \\ |\boldsymbol{r}'_n|\leqq\omega'_n\,. \end{cases} \tag{4.39}$$

Then, from (4.37), it follows that

$$|\boldsymbol{e}'_{n+1}|\leqq|A+hG'_n|\,|\boldsymbol{e}'_n|+|h|(\varepsilon'_1|\boldsymbol{e}'_{n+1}|+\varepsilon'_2|\boldsymbol{e}'_n|)+|h|\omega'_n,$$

which can be written as

$$|\boldsymbol{e}'_{n+1}|\leqq\frac{|A+h\,G'_n|+|h|\,\varepsilon'_2}{1-|h|\,\varepsilon'_1}|\boldsymbol{e}'_n|+\frac{|h|}{1-|h|\,\varepsilon'_1}\,\omega'_n, \tag{4.40}$$

because $|h|$ is small. Corresponding to (4.40), let us consider the linear difference equation

$$E'_{n+1}=\frac{|A+h\,G'_n|+|h|\,\varepsilon'_2}{1-|h|\,\varepsilon'_1}E'_n+\frac{|h|}{1-|h|\,\varepsilon'_1}\,\omega'_n \tag{4.41}$$

and let E'_n be a solution of this equation such that

$$E'_0 = |\boldsymbol{e}'_0| . \tag{4.42}$$

Then, by induction, it is readily seen that

$$|\boldsymbol{e}'_n| \leqq E'_n \qquad (n=0, 1, 2, \cdots),$$

which, by (4.36), implies

$$|\boldsymbol{e}_n| \leqq |T| E'_n \qquad (n=0, 1, 2, \cdots). \tag{4.43}$$

Consequently, in the sequel, we shall seek an estimate of the solution E'_n of (4.41) satisfying the initial condition (4.42).

First, let us seek bounds of E'_n.

As is seen from (4.33) and (4.35),

$$|G'(t)| \leqq M', \tag{4.44}$$

where

$$M' = \begin{cases} |T^{-1}|\,|T|\,M \sum_{l=1}^{k} (|\beta_0|\,|\alpha_l| + |\beta_l|) \\ \qquad\qquad \text{for a simple multi-step formula,} \\ |T^{-1}|\,|T|\,M \sum_{l=1}^{k-1} \{(|\beta_{-1}|\,|\hat{\alpha}_0| + |\beta_0|)|\alpha_l| + |\beta_{-1}|\,|\hat{\alpha}_l| + |\beta_l|\} \\ \qquad\qquad \text{for a compound multi-step formula,} \\ M \qquad\qquad \text{for the general Runge-Kutta formula.} \end{cases} \tag{4.45}$$

Also, as is seen from (4.39),

$$\omega'_n \leqq \omega', \tag{4.46}$$

where

$$\omega' = \begin{cases} |T^{-1}| \cdot \dfrac{\xi+\zeta}{1-|h|M|\beta_0|} \qquad \text{for a simple multi-step formula,} \\ |T^{-1}| \cdot \dfrac{\xi+\zeta+(\hat{\xi}+\hat{\zeta})L_4|\beta_{-1}|}{1-|h|M(|\beta_{-1}|\,|\hat{\alpha}_0| + |\beta_0|)} \\ \qquad\qquad \text{for a compound multi-step formula,} \\ (\xi+\zeta+\eta W) \qquad \text{for the general Runge-Kutta formula.} \end{cases} \tag{4.47}$$

Hence, due to (2.29), from (4.41), it follows that

$$E'_{n+1} \leqq \frac{1+|h|(M'+\varepsilon'_2)}{1-|h|\varepsilon'_1} E'_n + \frac{|h|}{1-|h|\varepsilon'_1} \omega'$$

$$=\left\{1+\frac{M'+\varepsilon_1'+\varepsilon_2'}{1-|h|\varepsilon_1'}|h|\right\}E_n'+\frac{|h|}{1-|h|\varepsilon_1'}\omega'^{1)}.$$

This has the form analogous to (2.40). Consequently, according to (2.42), we have:

$$E_n'\leqq\Gamma_n'\overset{\text{def}}{=}E_0'\cdot\exp\left[\frac{M'+\varepsilon_1'+\varepsilon_2'}{1-|h|\varepsilon_1'}|t_n-t_o|\right]+\frac{1}{M'+\varepsilon_1'+\varepsilon_2'}\left\{\left(\exp\frac{M'+\varepsilon_1'+\varepsilon_2'}{1-|h|\varepsilon_1'}|t_n-t_0|\right)-1\right\}\omega' \quad (n=0,\ 1,\ 2,\cdots). \tag{4.48}$$

The Γ_n' are the desired bounds of E_n'.

Now, let us transform the difference equation (4.41) to an equation of the differential form. To do this, we construct a continuous function $E'(t)$ corresponding to the solution E_n' so that

$$\begin{cases} E'(t_n)=E_n', \\ E'(t)=E_n'+\dfrac{t-t_n}{h}(E_{n+1}'-E_n') & \text{for } t\in[t_n,\ t_{n+1})^{2)}. \end{cases} \tag{4.49}$$

Also, corresponding to ω_n', we construct a step function $\omega'(t)$ so that

$$\omega'(t)=\omega_n' \qquad \text{for } t\in[t_n,\ t_{n+1}). \tag{4.50}$$

Then, from (4.41), we have:

$$\frac{dE'(t)}{dt}=\frac{1}{h}\left(\frac{|A+hG_n'|\pm h\varepsilon_2'}{1-|h|\varepsilon_1'}-1\right)E'(t_n)\pm\frac{1}{1-|h|\varepsilon_1'}\omega'(t) \qquad \text{for } t\in[t_n,\ t_{n+1}), \tag{4.51}$$

where the upper signs are taken for $h>0$ and the lower signs are taken for $h<0$[3].

Now, the first term in the right-hand side of (4.51) is equal to

$$\frac{1}{1-|h|\varepsilon_1'}\left\{\frac{|A+hG_n'|-1}{h}\pm(\varepsilon_1'+\varepsilon_2')\right\}E'(t_n),$$

which, for small $|h|$, due to (2.29), is nearly equal to

$$\frac{1}{1-|h|\varepsilon_1'}\left\{\mu_A^{\pm}[G'(t)]\pm(\varepsilon_1'+\varepsilon_2')\right\}E'(t). \tag{4.52}$$

Moreover their difference is estimated by using Lemmas 3, 5 and 6 of 4.1 as follows:

---

1) Here $E_n'\geqq 0$ $(n=0,\ 1,\ 2,\cdots)$ as is seen from (4.41) and (4.42).

2) This means that $t_n\leqq t<t_{n+1}$ or $t_n\geqq t>t_{n+1}$ according as $h>0$ or $h<0$.

3) This convention is kept in the sequel whenever the double signs are used.

$$(4.53)\qquad \left|\frac{1}{h}\left(\frac{|\Delta+hG'_n|\pm h\varepsilon'_2}{1-|h|\varepsilon'_1}-1\right)E'(t_n)-\frac{\mu_\Delta^{\pm}[G'(t)]\pm(\varepsilon'_1+\varepsilon'_2)}{1-|h|\varepsilon'_1}E'(t)\right|\leqq\Delta(t),$$

where

$$(4.54)\qquad \begin{cases} \varepsilon_3=\max\left|\dfrac{|\Delta+hG'_n|-1}{h}-\mu_\Delta^{\pm}[G'_n]\right|=o(1) \quad \text{as } |h|\to 0. \\ \varepsilon_4=\begin{cases} |h|\,|T|\,|T^{-1}|L_2\sum\limits_{l=1}^{k}(|\beta_0|\,|\alpha_l|+\beta_l|) \\ \qquad\qquad \text{for a simple multi-step formula,} \\ |h|\,|T|\,|T^{-1}|L_2\sum\limits_{l=1}^{k-1}\{(|\beta_{-1}|\,|\hat{\alpha}_0|+|\beta_0|)|\alpha_l|+|\beta_{-1}|\,|\hat{\alpha}_l|+|\beta_l|\} \\ \qquad\qquad \text{for a compound multi-step formula,} \\ |h|L_2 \qquad \text{for the general Runge-Kutta formula,} \end{cases} \end{cases}$$

and $\Delta(t)$ is a step function such that

$$(4.55)\qquad \Delta(t)=\left\{\frac{\varepsilon_3+\varepsilon_4}{1-|h|\varepsilon'_1}+|h|\left(\frac{M'+\varepsilon'_1+\varepsilon'_2}{1-|h|\varepsilon'_1}\right)^2\right\}\Gamma'_n + |h|\frac{M'+\varepsilon'_1+\varepsilon'_2}{(1-|h|\varepsilon'_1)^2}\omega'(t) \qquad \text{for } t\in[t_n, t_{n+1}).$$

Hence, from (4.51) and (4.53), in each interval $[t_n, t_{n+1})$, we obtain differential inequalities as follows:

$$(4.56)\qquad \begin{cases} \dfrac{dE'(t)}{dt}\leqq\dfrac{\mu_\Delta^{+}[G'(t)]+(\varepsilon'_1+\varepsilon'_2)}{1-h\varepsilon'_1}E'(t)+\dfrac{1}{1-h\varepsilon'_1}\omega'(t)+\Delta(t) & \text{for } h>0, \\ \dfrac{dE'(t)}{dt}\geqq\dfrac{\mu_\Delta^{-}[G'(t)]-(\varepsilon'_1+\varepsilon'_2)}{1+h\varepsilon'_1}E'(t)-\dfrac{1}{1+h\varepsilon'_1}\omega'(t)-\Delta(t) & \text{for } h<0. \end{cases}$$

Then, since $E'(t)$ and $\{\mu_\Delta^{\pm}[G'(t)]\pm(\varepsilon'_1+\varepsilon'_2)\}/(1-|h|\varepsilon'_1)$ are all continuous in the interval $|t-t_0|\leqq L$, the above differential inequalities are solved as follows:

$$(4.57)\qquad E'(t)\leqq E'_0\,e^{H^{\pm}(t)}\pm e^{H^{\pm}(t)}\int_{t_0}^{t}e^{-H^{\pm}(\tau)}\left\{\frac{1}{1-|h|\varepsilon'_1}\omega'(\tau)+\Delta(\tau)\right\}d\tau,$$

where

$$(4.58)\qquad H^{\pm}(t)=\int_{t_0}^{t}\frac{\mu_\Delta^{\pm}[G'(\tau)]\pm(\varepsilon'_1+\varepsilon'_2)}{1-|h|\varepsilon'_1}d\tau.$$

Thus, by (4.42) and (4.43), we have:

$$(4.59)\qquad |e_n|\leqq E_n\overset{\text{def}}{=\!=}|T|\,|T^{-1}|\,|\boldsymbol{e}_0|e^{H^{\pm}(t_n)} \pm|T|e^{H^{\pm}(t_n)}\int_{t_0}^{t_n}e^{-H^{\pm}(\tau)}\left\{\frac{1}{1-|h|\varepsilon'_1}\omega'(\tau)+\Delta(\tau)\right\}d\tau \qquad (n=0, 1, 2,\cdots).$$

This is the desired estimate of errors.

4.4 Comparison of various estimates of errors

As is seen from (4.22),

$$\gamma_n=|T|\,|T^{-1}|\,|e_0|\exp\left[\frac{K_1'+K_2'}{1-|h|K_1'}|t_n-t_0|\right] + \frac{|T|}{K_1'+K_2'}\left\{\left(\exp\frac{K_1'+K_2'}{1-|h|K_1'}|t_n-t_0|\right)-1\right\}\varepsilon' \qquad (n=0, 1, 2,\cdots) \tag{4.60}$$

is an estimate of errors. Also, as is seen from (4.48),

$$\Gamma_n=|T|\,|T^{-1}|\,|e_0|\exp\left[\frac{M'+\varepsilon_1'+\varepsilon_2'}{1-|h|\varepsilon_1'}|t_n-t_0|\right] + \frac{|T|}{M'+\varepsilon_1'+\varepsilon_2'}\left\{\left(\exp\frac{M'+\varepsilon_1'+\varepsilon_2'}{1-|h|\varepsilon_1'}|t_n-t_0|\right)-1\right\}\omega' \qquad (n=0, 1, 2,\cdots) \tag{4.61}$$

is also an estimate of errors.

But, by ((4.16)−(4.21)) and ((4.45), (4.47)), the quantities $K_1'+K_2'$, M', ε' and ω' are as follows:

for a simple multi-step formula,

$$\begin{cases} K_1'+K_2'=|T|\,|T^{-1}|M(|\beta_0|+\sum_{l=1}^{k}|\beta_l|), \\ M'=|T|\,|T^{-1}|M(|\beta_0|\sum_{l=1}^{k}|\alpha_l|+\sum_{l=1}^{k}|\beta_l|), \end{cases}$$

$$\begin{cases} \varepsilon'=|T^{-1}|(\xi+\zeta), \\ \omega'=|T^{-1}|\,|(\xi+\zeta)/(1-|h|M|\beta_0|); \end{cases}$$

for a compound multi-step formula,

$$\begin{cases} K_1'+K_2'=|T|\,|T^{-1}|M\{(|\beta_{-1}|\,|\hat{\alpha}_0|+|\beta_0|)+|\beta_{-1}|\sum_{l=1}^{k-1}|\hat{\alpha}_l|+\sum_{l=1}^{k-1}|\beta_l| \\ \qquad\qquad +|h|M|\beta_{-1}|\sum_{l=0}^{k}|\hat{\beta}_l|\}, \\ M'=|T|\,|T^{-1}|M\{(|\beta_{-1}|\,|\hat{\alpha}_0|+|\beta_0|)\sum_{l=1}^{k-1}|\alpha_l|+|\beta_{-1}|\sum_{l=1}^{k-1}|\hat{\alpha}_l|+\sum_{l=1}^{k-1}|\beta_l|\}, \end{cases}$$

$$\begin{cases} \varepsilon' = |T^{-1}| \{\xi+\zeta+(\hat{\xi}+\hat{\zeta})M|\beta_{-1}|\}, \\ \omega' = |T^{-1}| \{\xi+\zeta+(\hat{\xi}+\hat{\zeta})L_4|\beta_{-1}|\}/\{1-|h|M(|\beta_{-1}|\,|\hat{\alpha}_0|+|\beta_0|)\} \\ \qquad (L_4 = M+\gamma L_1 \sum_{l=0}^{k-1} |\hat{\alpha}_l| + |h|L_3+(\hat{\xi}+\hat{\zeta})L_1); \end{cases}$$

for the general Runge-Kutta formula,

$$\begin{cases} K_1'+K_2' = MW, \\ M' = M \end{cases}$$

$$\begin{cases} \varepsilon' = (\xi+\zeta+\eta W), \\ \omega' = (\xi+\zeta+\eta W). \end{cases}$$

Hence, we see that, for small $|h|$ and $|\boldsymbol{e}_0|$, γ_n *and* Γ_n *have no essential difference between them*, since, for such h and $\boldsymbol{e}_0$, ε_1' and ε_2' are small compared with M'.

Next, let us compare Γ_n with the estimate E_n given by (4.59).
To do this, we first prove a lemma.

Lemma. *If* $g(t)$ *is a continuously differentiable function such that*

$$\text{(4.62)} \qquad \begin{cases} \dfrac{dg(t)}{dt} \leqq g_1 & \text{for } t \geqq t_0, \text{ or} \\ \dfrac{dg(t)}{dt} \geqq -g_2 & \text{for } t \leqq t_0, \end{cases}$$

then

$$\text{(4.63)} \qquad \left| e^{g(t)} \int_{t_0}^{t} e^{-g(\tau)} d\tau \right| \leqq \begin{cases} \dfrac{1}{g_1}(e^{g_1|t-t_0|}-1) & \text{for } t \geqq t_0, \text{ or} \\ \dfrac{1}{g_2}(e^{g_2|t-t_0|}-1) & \text{for } t \leqq t_0 \end{cases}$$

respectively, where, for $g_1=0$ *or* $g_2=0$, *the right-hand sides of* (4.63) *mean*

$$\lim_{g_i \to 0} \frac{1}{g_i}(e^{g_i|t-t_0|}-1) = |t-t_0| \qquad (i=1, 2).$$

Proof. Put

$$\text{(4.64)} \qquad y(t) = \left| e^{g(t)} \int_{t_0}^{t} e^{-g(\tau)} d\tau \right|.$$

Then this satisfies the differential equation

$$(4.65)\qquad \begin{cases} \dfrac{dy}{dt}=g'y+1 & \text{for } t\geqq t_0\,, \\ \dfrac{dy}{dt}=g'y-1 & \text{for } t\leqq t_0 \end{cases} \qquad \left(g'=\frac{dg(t)}{dt}\right)$$

and also the initial condition

$$(4.66)\qquad y(t_0)=0.$$

Of course, by (4.65), we mean that, at $t=t_0$, the right derivative of $y(t)$ is equal to $+1$ and its left derivative is equal to -1. Corresponding to the equation (4.65), let us consider the equation

$$(4.67)\qquad \begin{cases} \dfrac{dY}{dt}=g_1Y+(1+\varepsilon) & \text{for } t\geqq t_0\,, \\ \dfrac{dY}{dt}=-g_2Y-(1+\varepsilon) & \text{for } t\leqq t_0\,, \end{cases}$$

where ε is an arbitrary positive number. And let $Y_\varepsilon(t)$ be a solution of (4.67) (in the same meaning as $y(t)$ is a solution of (4.65)) satisfying the initial condition

$$(4.68)\qquad Y_\varepsilon(t_0)=0.$$

Then, comparing ((4.67), (4.68)) with ((4.65), (4.66)), from (4.62), we have

$$(4.69)\qquad y(t)\leqq Y_\varepsilon(t),$$

because

$$(4.70)\qquad Y_\varepsilon(t)=\begin{cases} \dfrac{1+\varepsilon}{g_1}\{e^{g_1|t-t_0|}-1\}\geqq 0 & \text{for } t\geqq t_0\,, \\ \dfrac{1+\varepsilon}{g_2}\{e^{g_2|t-t_0|}-1\}\geqq 0 & \text{for } t\leqq t_0\,. \end{cases}$$

Here, of course, for $g_1=0$ or $g_2=0$, the right-hand sides mean

$$\lim_{g_i\to 0}\frac{1+\varepsilon}{g_1}\{e^{g_i|t-t_0|}-1\}=(1+\varepsilon)|t-t_0|\geqq 0.$$

Now ε is an arbitrary positive number, consequently, letting $\varepsilon\to 0$, from (4.69) and (4.70), we obtain (4.63). This proves the lemma.

Let us return to the estimate E_n given by (4.59).

By Lemma 3 of 4.1 and (4.44), let us assume that

$$(4.71)\qquad \mu_\Delta^+[G'(t)]\leqq C^+\leqq M',\quad \mu_\Delta^-[G'(t)]\geqq -C^-\geqq -M'.$$

Then, from (4.58), follows

$$H^{\pm}(t) \leqq \frac{C^{\pm}+\varepsilon_1'+\varepsilon_2'}{1-|h|\varepsilon_1'}|t-t_0| \leqq \frac{M'+\varepsilon_1'+\varepsilon_2'}{1-|h|\varepsilon_1'}|t-t_0|, \tag{4.72}$$

and

$$\begin{cases} \dfrac{dH^{+}(t)}{dt} \leqq \dfrac{C^{+}+\varepsilon_1'+\varepsilon_2'}{1-|h|\varepsilon_1'} \leqq \dfrac{M'+\varepsilon_1'+\varepsilon_2'}{1-|h|\varepsilon_1'}, \\ \dfrac{dH^{-}(t)}{dt} \geqq -\dfrac{C^{-}+\varepsilon_1'+\varepsilon_2'}{1-|h|\varepsilon_1'} \geqq -\dfrac{M'+\varepsilon_1'+\varepsilon_2'}{1-|h|\varepsilon_1'}. \end{cases} \tag{4.73}$$

Consequently, by the above lemma, from the expression of E_n given by (4.59), we have:

$$E_n \leqq V_n^{\pm}+\hat{V}_n^{\pm} \leqq \Gamma_n+\hat{\Gamma}_n, \tag{4.74}$$

where

$$\begin{cases} V_n^{\pm} = |T|\,|T^{-1}|\,|\boldsymbol{e}_0| \cdot \exp\left[\dfrac{C^{\pm}+\varepsilon_1'+\varepsilon_2'}{1-|h|\varepsilon_1'}|t_n-t_0|\right] \\ \qquad + \dfrac{|T|}{C^{\pm}+\varepsilon_1'+\varepsilon_2'}\left\{\left(\exp\dfrac{C^{\pm}+\varepsilon_1'+\varepsilon_2'}{1-|h|\varepsilon_1'}|t_n-t_0|\right)-1\right\}\omega', \\ \hat{V}_n^{\pm} = \dfrac{|T|(1-|h|\varepsilon_1')}{C^{\pm}+\varepsilon_1'+\varepsilon_2'}\left\{\left(\exp\dfrac{C^{\pm}+\varepsilon_1'+\varepsilon_2'}{1-|h|\varepsilon_1'}|t_n-t_0|\right)-1\right\} \times \max_{[t_0,t_n]} \Delta(t), \\ \hat{\Gamma}_n = \dfrac{|T|(1-|h|\varepsilon_1')}{M'+\varepsilon_1'+\varepsilon_2'}\left\{\left(\exp\dfrac{M'+\varepsilon_1'+\varepsilon_2'}{1-|h|\varepsilon_1'}|t_n-t_0|\right)-1\right\} \times \max_{[t_0,t_n]} \Delta(t). \end{cases} \tag{4.75}$$

Now, by (4.55), for small $|h|$, $\max \Delta(t)$ is small compared with Γ_n. Therefore, from the rightest side of (4.74), we have

$$E_n \leqq \Gamma_n \qquad \text{approximately.} \tag{4.76}$$

This says E_n always gives a better estimate of errors than Γ_n.

Then, has E_n any serious difference from Γ_n ?

To answer this question, let us conider $V_n^{\pm}$ and $\hat{V}_n^{\pm}$. Evidently the functions of the forms

$$e^{g|t_n-t_0|} \quad \text{and} \quad \frac{1}{g}(e^{g|t_n-t_0|}-1)$$

are both monotonically increasing with respect to g in $(-\infty, \infty)$. Consequently, from (4.71) and (4.75),

$$V_n^{\pm} \leqq \Gamma_n \quad \text{and} \quad \hat{V}_n^{\pm} \leqq \hat{\Gamma}_n,$$

and moreover the differences $\Gamma_n - V_n^{\pm}$ and $\hat{\Gamma}_n - \hat{V}_n^{\pm}$ become larger and larger as $M'-C^{\pm}$ increase. In particular, these differences become quite marked when the $C^{\pm}$ become negative. This fact says that E_n can be quite different

from Γ_n, consequently from γ_n. This says that E_n is a more precise estimate of errors than Γ_n and γ_n.

4.5 Remarks

1° Once E_n has been found in the above way, we can improve this E_n further in the following way:

i) replacing γ by $\max E_n$, we calculate ε_1 and ε_2 again by (4.24) or (4.28) or (4.32),

ii) replacing Γ'_n by $E_n/|T|$, we calculate $\Delta(t)$ again by (4.55),

iii) for ε_1, ε_2 and $\Delta(t)$ obtained newly, we calculate E_n again by (4.59).

This process can be continued indefinitely so long as the new E_n is smaller than the old E_n.

2° When actual computation of $H^{\pm}(t)$ is difficult, as is readily seen from (4.74), we can take

$$V_n^{\pm}+\hat{V}_n^{\pm}$$

as an estimate of errors choosing $C^{\pm}$ as small as possible. As is mentioned in the preceding paragraph, even this gives a considerably better estimate than Γ_n and γ_n if $C^{\pm}$ are chosen sufficiently small.

3° When the different multi-step formulas are applied to each component of the given differential system, we can get also the similar estimates of errors if we replace the scalar coefficients $\alpha_1, \alpha_2, \cdots, \alpha_k$; $\beta_{-1}, \beta_0, \beta_1, \cdots, \beta_k$; $\hat{\alpha}_0, \hat{\alpha}_1, \cdots, \hat{\alpha}_{k-1}$; $\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_k$ by the diagonal matrices whose diagonal elements are respectively the different scalar coefficients $\alpha_1, \alpha_2, \cdots, \alpha_k$; $\beta_{-1}, \beta_0, \beta_1, \cdots, \beta_k$; $\hat{\alpha}_0, \hat{\alpha}_1, \cdots, \hat{\alpha}_{k-1}$; $\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_k$.

4° As is readily seen, the results of the present paper are valid also for the complex differential system derived from the real differential system by the complex linear transformation.

5° In this section, for generality, we are concerned with the complex differential system mentioned in the above section.

For multi-step formulas for which (3.42) is valid, from (2.5),

$$A_0=\begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 & 1 \\ 0 & 0 & \cdots & 0 & 1 \end{pmatrix},$$

consequently, as T_0 in (2.9), we may take

$$T_0=\begin{pmatrix} 1 & 0 & \cdots\cdots & 0 & 1 \\ 1 & \vdots & & \delta & 0 \\ \vdots & \vdots & \delta^2 & & \vdots \\ \vdots & 0 & \iddots & & \vdots \\ 1 & \delta^{k-2} & \cdots\cdots & & 0 \\ 1 & 0 & \cdots\cdots\cdots & & 0 \end{pmatrix}.$$

For such T_0, from (4.35), we readily see that

$$G'(t)=T^{-1}G(t)\,T=(G'^{\nu}_{\mu}(t)),$$

where

$$G^{\nu}_{\mu l}=\frac{1}{h}\,(A^{\nu}_{\mu l}-\delta^{\nu}_{\mu}\,\alpha_l)$$

and

$$G'^{\nu}_{\mu}=\begin{pmatrix} \sum_{l=1}^{k} G^{\nu}_{\mu l} & \delta^{k-2}G^{\nu}_{\mu 2} & \cdots & \delta G^{\nu}_{\mu k-1} & G^{\nu}_{\mu k} \\ * & * & \cdots\cdots & * & * \\ \vdots & \vdots & & \vdots & \vdots \\ * & * & \cdots\cdots & * & * \end{pmatrix}.$$

Then, by Lemma 2 of **4.1**, we see that,

for Adams' formulas,

$$\mu^{\pm}_{A}[G'(t)]=\binom{\max}{\min}_{\nu}\Big[\Re F^{\nu}_{\nu}(t)\pm\sum_{\mu\neq\nu}|F^{\nu}_{\mu}(t)| \pm(\delta^{k-2}|\beta_2|+\cdots+\delta|\beta_{k-1}|+|\beta_k|)\sum_{\mu=1}^{N}|F^{\nu}_{\mu}(t)|\Big]$$

and, for the compound multi-step formulas satisfying (3.42),

$$\mu^{\pm}_{A}[G'(t)]=\binom{\max}{\min}_{\nu}\Big[\Re F^{\nu}_{\nu}(t)\pm\sum_{\mu\neq\nu}|F^{\nu}_{\mu}(t)| \pm(\delta^{k-1}|\beta_{-1}\,\hat{\alpha}_2+\beta_2|+\cdots+\delta|\beta_{-1}\,\hat{\alpha}_{k-1}+\beta_{k-1}|)\sum_{\mu=1}^{N}|F^{\nu}_{\mu}(t)|\Big].$$

For the general Runge-Kutta formulas, from (1.73), $A_0=I$ and $G(t)=F(t)$, consequently, by Corollary of Lemma 2 of **4.1**, it readily follows that

$$\mu^{\pm}_{A}[G(t)]=\binom{\max}{\min}_{\nu}\Big[\Re F^{\nu}_{\nu}(t)\pm\sum_{\mu\neq\nu}|F^{\nu}_{\mu}(t)|\Big].$$

Examples.

For Adams' formula

$$x_{n+4}=x_{n+3}+\frac{h}{720}(251\,\dot{x}_{n+4}+646\,\dot{x}_{n+3}-264\,\dot{x}_{n+2}+106\,\dot{x}_{n+1}-19\,\dot{x}_n),$$

$$\mu_{\Delta}^{\pm}[G'(t)]=\begin{pmatrix}\max\\ \min\\ \nu\end{pmatrix}\Big[\Re F_{\nu}^{\nu}(t)\pm\sum_{\mu\neq\nu}|F_{\mu}^{\nu}(t)| \pm\frac{1}{720}(264\,\delta^2+106\,\delta+19)\sum_{\mu=1}^{N}|F_{\mu}^{\nu}(t)|\Big].$$

For the compound multi-step formula

$$\begin{cases} x_{n+4}=x_{n+3}+\dfrac{h}{720}(-19\,\hat{\dot{x}}_{n+5}+346\,\dot{x}_{n+4}+456\,\dot{x}_{n+3}-74\,\dot{x}_{n+2}+11\,\dot{x}_{n+1}),\\ \hat{x}_{n+5}=x_{n+4}+\dfrac{h}{720}(1901\,\dot{x}_{n+4}-2774\,\dot{x}_{n+3}+2616\,\dot{x}_{n+2}-1274\,\dot{x}_{n+1}+251\,\dot{x}_n), \end{cases}$$

$$\mu_{\Delta}^{\pm}[G'(t)]=\begin{pmatrix}\max\\ \min\\ \nu\end{pmatrix}\Big[\Re F_{\nu}^{\nu}(t)\pm\sum_{\mu\neq\nu}|F_{\mu}^{\nu}(t)| \pm\frac{1}{720}(74\,\delta^2+11\,\delta)\sum_{\mu=1}^{N}|F_{\mu}^{\nu}(t)|\Big].$$

4.6 Numerical examples

By way of example, let us compute the various estimates of errors of some integration formulas applied to the Cauchy problem such that the given equations are

$$\frac{dx}{dt}=\varepsilon x \qquad (x\text{: scalar},\ \varepsilon=\pm 1) \tag{4.77}$$

and the initial condition is

$$x(0)=1. \tag{4.78}$$

Evidently the true solutions of the above problem are

$$x(t)=\exp(\varepsilon t). \tag{4.79}$$

Let us consider the solutions in the domain

$$D:\ 0\leqq x\leqq 1.7,\quad 0\leqq t\leqq 0.5.$$

Then, since

$$F(x,\,t)=\varepsilon, \tag{4.80}$$

it readily follows from (4.10) and (4.11) that

(4.81) $$L_1=L_2=0,\quad M_1=1.7,\quad M=1.$$

The integration formulas taken into consideration in this paragraph are

I: $$x_{n+4}=x_{n+3}+\frac{h}{720}(251\,\dot{x}_{n+4}+646\,\dot{x}_{n+3}-264\,\dot{x}_{n+2}+106\,\dot{x}_{n+1}-19\,\dot{x}_n)$$

(Adams' formula);

II: $$x_{n+4}=x_{n+2}+\frac{h}{90}(29\,\dot{x}_{n+4}+124\,\dot{x}_{n+3}+24\,\dot{x}_{n+2}+4\,\dot{x}_{n+1}-\dot{x}_n);$$

III: $$\begin{cases} x_{n+4}=x_{n+3}+\dfrac{h}{720}(-19\,\hat{\dot{x}}_{n+5}+346\,\dot{x}_{n+4}+456\,\dot{x}_{n+3}-74\,\dot{x}_{n+2}+11\,\dot{x}_{n+1}), \\ \hat{x}_{n+5}=x_{n+4}+\dfrac{h}{720}(1901\,\dot{x}_{n+4}-2774\,\dot{x}_{n+3}+2616\,\dot{x}_{n+2}-1274\,\dot{x}_{n+1}+251\,\dot{x}_n) \end{cases}$$

(compound multi-step formula);

IV: $$x_{n+1}=x_n+\frac{h}{6}(k_{n1}+2k_{n2}+2k_{n3}+k_{n4}),\quad \text{where}$$

$$\begin{cases} k_{n1}=f(x_n,\ t_n), \\ k_{n2}=f(x_n+\dfrac{h}{2}k_{n1},\ t_n+\dfrac{1}{2}h), \\ k_{n3}=f(x_n+\dfrac{h}{2}k_{n2},\ t_n+\dfrac{1}{2}h), \\ k_{n4}=f(x_n+h\,k_{n3},\ t_n+h), \end{cases}$$

(Runge-Kutta formula).

As is readily seen, the truncation errors of these formulas are estimated as follows:

$$|T_n|\leqq\frac{3}{160}h^6\max|x^{\mathrm{VI}}|\qquad\text{for I;}$$

$$|T_n|\leqq\frac{7}{648}h^6\max|x^{\mathrm{VI}}|\qquad\text{for II;}$$

$$|T_n|\leqq\frac{11}{1440}h^6\max|x^{\mathrm{VI}}|,\quad|\hat{T}_n|\leqq\frac{95}{288}h^6\max|x^{\mathrm{VI}}|\qquad\text{for III;}$$

$$|T_n|\leqq Ch^5\qquad\text{for IV.}$$

Here C is a constant which can be computed for any given differential equation.

Now, for the solutions of the present problem lying in the domain D,

(4.82) $$\begin{cases} \max|x^{\mathrm{VI}}|=\begin{cases} 1.7 & \text{for } \varepsilon=1, \\ 1 & \text{for } \varepsilon=-1; \end{cases} \\ C=\dfrac{1}{120}\max|x|=\begin{cases} 1.7/120 & \text{for } \varepsilon=1, \\ 1/120 & \text{for } \varepsilon=-1. \end{cases} \end{cases}$$

Consequently, if we take

$$h=0.01,$$

then, by (4.13), we have:

$$\text{for I,}\quad \zeta=\begin{cases}0.031875\times10^{-10} & (\varepsilon=1),\\ 0.018750\times10^{-10} & (\varepsilon=-1);\end{cases}$$

$$\text{for II,}\quad \zeta=\begin{cases}0.018364\times10^{-10} & (\varepsilon=1),\\ 0.010802\times10^{-10} & (\varepsilon=-1);\end{cases}$$

$$\text{for III,}\quad \zeta=\begin{cases}0.012986\times10^{-10} & (\varepsilon=1),\\ 0.0076389\times10^{-10} & (\varepsilon=-1),\end{cases}$$

$$\hat{\zeta}=\begin{cases}0.0056076\times10^{-10} & (\varepsilon=1),\\ 0.0032986\times10^{-10} & (\varepsilon=-1);\end{cases}$$

$$\text{for IV,}\quad \zeta=\begin{cases}1.41667\times10^{-10} & (\varepsilon=1),\\ 0.83333\times10^{-10} & (\varepsilon=-1).\end{cases}$$

To get the estimates of round-off errors, we shall prove a

Lemma. *When the vector equation*

$$x=X(x)$$

is solved numerically by the method of iteration, it holds that

$$|x-X(x)|\leqq\frac{1+S}{1-S}\,\varepsilon$$

in the state of numerical convergence[1], *where* ε *is a bound of the round-off errors arising in the computation of* $X(x)$ *and* S *is a positive constant less than* 1 *such that*

$$|X(x')-X(x'')|\leqq S|x'-x''|$$

for any two values x' and x'' *of* x.

Proof. Let $\hat{x}$ be a true solution of the given equation, and let x_M, $x_{M+1},\cdots$, x_{M+m} be the computed values of x in the state of numerical convergence. Then it readily follows that

1) When the given equation is solved numerically by the method of iteration, in the process of computation, there always appears the state in which a certain number of computed values of x are repeated. Such a state is called *the state of numerical convergence* [17].

$$
\begin{aligned}
&|x_{M+1}-\hat{x}| \leqq S|x_M-\hat{x}| + \varepsilon,\\
&|x_{M+2}-\hat{x}| \leqq S|x_{M+1}-\hat{x}| + \varepsilon,\\
&\qquad\vdots\\
&|x_{M+m}-\hat{x}| \leqq S|x_{M+m-1}-\hat{x}| + \varepsilon,\\
&|x_M-\hat{x}| \leqq S|x_{M+m}-\hat{x}| + \varepsilon.
\end{aligned}
$$

Consequently it follows that

$$|x_M-\hat{x}| \leqq S^{m+1}|x_M-\hat{x}| + \varepsilon(1+S+S^2+\cdots+S^m),$$

namely that

$$|x_M-\hat{x}| \leqq \frac{\varepsilon}{1-S}.$$

Then

$$|x_M-X(x_M)| \leqq |x_M-\hat{x}| + |X(\hat{x})-X(x_M)| \leqq \frac{1+S}{1-S}\varepsilon.$$

This proves the lemma.

Then, if the computation is rounded off correctly to ten decimal places, for the round-off errors, we have:

for I, $S=\frac{251}{720}h$, $\xi_n=\xi=50.34983\times10^{-10}$;

for II, $S=\frac{29}{90}h$, $\xi_n=\xi=50.32326\times10^{-10}$;

for III, $S=\frac{1901}{720}h$, $\xi_n=\xi=52.71188\times10^{-10}$,

$\hat{\xi}_n=\hat{\xi}=0.52712\times10^{-10}$;

for IV, $S=0$, $\xi_n=\xi=50.00000\times10^{-10}$.

By the definition of the matrix T, we can take T so that,

for I and III, $T=\begin{pmatrix} 1 & 0 & 0 & 1\\ 1 & 0 & \delta & 0\\ 1 & \delta^2 & 0 & 0\\ 1 & 0 & 0 & 0 \end{pmatrix}$, $T^{-1}=\begin{pmatrix} 0 & 0 & 0 & 1\\ 0 & 0 & \delta^{-2} & -\delta^{-2}\\ 0 & \delta^{-1} & 0 & -\delta^{-1}\\ 1 & 0 & 0 & -1 \end{pmatrix}$;

for II, $T=\begin{pmatrix} 1 & 1 & 0 & 1\\ 1 & -1 & \delta & 0\\ 1 & 1 & 0 & 0\\ 1 & -1 & 0 & 0 \end{pmatrix}$, $T^{-1}=\begin{pmatrix} 0 & 0 & \frac{1}{2} & \frac{1}{2}\\ 0 & 0 & \frac{1}{2} & -\frac{1}{2}\\ 0 & \delta^{-1} & 0 & -\delta^{-1}\\ 1 & 0 & -1 & 0 \end{pmatrix}$,

for IV, $T=T^{-1}=I$.

Then, if we take

$$\delta=0.9,$$

we see from (4.54) that

for I, $|T|=2$, $|T^{-1}|=2.46914$,

$$\mu_A^+[G_n']=\begin{cases} 1.45589 & (\varepsilon=1), \\ -0.54411 & (\varepsilon=-1), \end{cases}$$

$\varepsilon_3=0$;

for II, $|T|=3$, $|T^{-1}|=2.22222$,

$\mu_A^+[G_n']=1.44778$,

$\varepsilon_3=0$;

for III, $|T|=2$, $|T^{-1}|=2.46914$,

$$\mu_A^+[G_n']=\begin{cases} 1.09700 & (\varepsilon=1), \\ -0.90300 & (\varepsilon=-1), \end{cases}$$

$\varepsilon_3=0$;

for IV, $|T|=|T^{-1}|=1$,

$\mu_A^+[G_n']=\varepsilon$,

$\varepsilon_3=0$.

Using the above values, we compute the error estimates γ_n, Γ_n and E_n by means of (4.60), (4.61) and (4.59). The results are shown in the table of the next page. In this table, $\tilde{E}_n$ are the values of E_n improved by the process mentioned in **1°** of **4.5**.

The reason why $E_n>\Gamma_n$ for $n=40, 50$ in the case of II is due to the fact that $\max \Delta(t)$ is not small compared with Γ_n. Indeed their values are

$$\max_{0\leqq t\leqq 0.4}\Delta(t)=\begin{cases} 3849.32891\times 10^{-10} & (\varepsilon=1), \\ 3849.32663\times 10^{-10} & (\varepsilon=-1), \end{cases}$$

$$\max_{0\leqq t\leqq 0.5}\Delta(t)=\begin{cases} 14846.89895\times 10^{-10} & (\varepsilon=1), \\ 14845.06739\times 10^{-10} & (\varepsilon=-1). \end{cases}$$

Anyhow the table of the next page shows the superiority of the estimates E_n compared with γ_n and Γ_n.

Formulas	ε	n	$10^{10}\gamma_n$	$10^{10}\Gamma_n$	$10^{10}E_n$	$10^{10}\tilde{E}_n$	$10^{10}e_n$
I	+1	10	47	46	35	34	0
		20	156	152	92	79	−1
		30	425	408	222	146	0
		40	1083	1030	578	256	+1
		50	2698	2536	1632	462	+1
	−1	10	47	46	34	34	+1
		20	156	152	81	69	−1
		30	425	408	176	109	−2
		40	1083	1028	412	154	0
		50	2697	2536	1050	206	0
II	+1	10	87	84	62	56	0
		20	419	396	262	163	−1
		30	1736	1608	1263	488	−2
		40	6952	6288	6507	—	−1
		50	27706	24324	33346	—	−1
	−1	10	87	84	62	56	0
		20	419	396	262	163	0
		30	1735	1608	1263	488	0
		40	6951	6288	6507	—	0
		50	27702	24321	33342	—	0
III	+1	10	42	42	34	34	0
		20	117	114	76	72	−1
		30	259	248	138	120	0
		40	529	500	244	181	+1
		50	1039	972	449	261	+1
	−1	10	42	42	30	30	+1
		20	117	114	61	57	−1
		30	259	248	101	84	−2
		40	529	500	162	111	0
		50	1039	972	270	137	0
IV	+1	10	6	6	6	—	−1
		20	12	12	12	—	−2
		30	19	19	19	—	−2
		40	26	26	26	—	−1
		50	34	34	35	—	−1
	−1	10	6	6	5	—	+1
		20	12	12	10	—	−1
		30	18	18	14	—	−2
		40	26	26	17	—	0
		50	34	34	21	—	0

References

[1] Collatz, L.: *Numerische Behandlung von Differentialgleichungen*, 2 Auflage, Berlin (1955).

[2] Carr, John W., III: *Error bounds for the Runge-Kutta single-step integration process*, J. Assoc. Comput. Machinery, **5** (1958), 39-44.

[3] Dahlquist, Germund: *Convergence and stability in the numerical integration of ordinary differential equations*, Math. Scand., **4** (1956), 33-53.

[4] Dahlquist, Germund: *Stability and error bounds in the numerical integration of ordinary differential equations*, Kungl. Tekniska Högskolans Handlingar, Stockholm, Nr. 130 (1959).

[5] Galler, B. A., and Rozenberg, D. P.: *A generalization of a theorem of Carr on error bounds for Runge-Kutta procedures*, J. Assoc. Comput. Machinery, **7** (1960), 57-60.

[6] Gill, S.: *A process for the step-by-step integration of differential equations in an automatic digital computing machine*, Proc. Cambridge Phil. Soc., **47** (1951), 96-108.

[7] Hamming, R. W.: *Stable predictor-corrector methods for ordinary differential equations*, J. Assoc. Comput. Machinery, **6** (1959), 37-47.

[8] Hildebrand, F. B.: *Introduction to numerical analysis*, New York (1956).

[9] Hull, T. E., and Luxemburg, W. A. J.: *Numerical methods and existence theorems for ordinary differential equations*, Numerische Math., **2** (1960), 30-41.

[10] Lotkin, Mark: *The propagation of error in numerical integrations*, Proc. Amer. Math. Soc., **5** (1954), 869-887.

[11] Milne, W. E., and Reynolds, R. R.: *Stability of a numerical solution of differential equations*, J. Assoc. Comput. Machinery, **6** (1959), 196-203.

[12] Milne, W. E., and Reynolds, R. R.: *Stability of a numerical solution of differential equations*—Part II, J. Assoc. Comput. Machinery, **7** (1960), 46-56.

[13] Rutishauser, H.: *Über die Instabilität von Methoden zur Integration gewöhnlicher Differentialgleichungen*, Z. Angew. Math. Physik, **3** (1952), 65-74.

[14] Todd, John: *Notes on modern numerical analysis. I. Solution of differential equations by recurrence relations*, Math. Tables and Other Aids to Computation, **4** (1950), 39-44.

[15] Uhlmann, Werner: *Fehlerabschätzungen bei Anfangswertaufgaben gewöhnlicher Differentialgleichungssysteme 1. Ordnung*, Z. angew. Math. Mech., **37** (1957), 88-99.

[16] Uhlmann, Werner: *Fehlerabschätzungen bei Anfangswertaufgaben einer gewöhnlichen Differentialgleichung höherer Ordnung*, Z. angew. Math. Mech., **37** (1957), 99-111.

[17] Urabe, Minoru: *Convergence of numerical iteration in solution of equations*, J. Sci. Hiroshima Univ., Ser. A, **19** (1956), 479-489.

[18] Urabe, M., Yanagiwara, H., and Shinohara, Y.: *Periodic solutions of van der Pol's equation with damping coefficient* $\lambda=2\sim10$, J. Sci. Hiroshima Univ., Ser. A, **23** (1960), 325-366.

[19] Wilf, Herbert S.: *A stability criterion for numerical integration*, J. Assoc. Comput. Machinery, **6** (1959), 363-365.

Department of Mathematics, Faculty of Science,
Hiroshima University
and MRC, United States Army,
University of Wisconsin.