Riv. Mat. Univ. Parma (2) 9 (1968), 109-127

DIRAN SARAFYAN (*)

Estimation of Errors for the Approximate Solution of Differential Equations and their Systems. (**)

1. - Introduction.

In the numerical solution of initial value problems

$$(1)\qquad \begin{cases} \dfrac{\mathrm{d}y}{\mathrm{d}x} = f(x,\ y) \\ \\ x = x_0\,, \quad y = y_0\,, \end{cases}$$

we are still far from possessing rigorous error estimation methods for the general case [3]. For higher order differential equations and systems of differential equations the situation is worse. It is our purpose to bring some substantial improvements to these important problems.

We shall reach our goal through the use of pseudo-iterative formulas, which are known to constitute a special class of RUNGE-KUTTA formulas endowed with an error estimating internal property ([5], [6], [7]). Subsequently, we shall generalize our error estimation method by using RUNGE-KUTTA formulas which are no longer of pseudo-iterative type. However it is not necessary to adopt this approach; once the main idea is grasped one can just as well start directly with the use of RUNGE-KUTTA formulas proper. We are merely following the order with which various ideas occured to us.

Furthermore, we shall define the near-optimal stepsize and give a practical rule for its determination. This will enable one to check or curb the truncation as well as the round-off errors over an extended interval.

(*) Indirizzo: Department of Mathematics, Louisiana State University, New Orleans, U.S.A..

(**) Ricevuto: 5-V-1968.

Although we are not concerned directly with numerical quadratures, a by-product of this work will be a new error estimation method for the former. This method is as effective as the known error estimation methods for quadratures, but as it will be seen it is far more practical.

Finally, we felt useful to start this investigation with a short critical survey of a few of the most well-known methods or formulas dealing with errors in the approximate solution of (1).

2. – A formula which is frequently mentioned in the literature is that of BIEBERBACH's [2]. This formula provides rigorous upper-bounds for the use of the classical fourth order RUNGE-KUTTA formula

$$\tilde{y}_4(x_0 + h) = y_0 + \frac{1}{6}[k_0 + 2k_1 + 2k_2 + k_3]\,, \tag{2}$$

where

$$\begin{cases} k_0 = h\,f(x_0\,,\; y_0) \\ k_1 = h\,f\left(x_0 + \frac{1}{2}h\,,\;\; y_0 + \frac{1}{2}k_0\right) \\ k_2 = h\,f\left(x_0 + \frac{1}{2}h\,,\;\; y_0 + \frac{1}{2}k_1\right) \\ k_3 = h\,f(x_0 + h\,,\;\; y_0 + k_2)\,. \end{cases}$$

Let $y(x)$, $y_0 = y(x_0)$, represent the exact solution of (1) and let

$$f \in C^4, \qquad |f| \leqq N\,, \qquad \left|\frac{\partial^r f}{\partial x^i\,\partial y^s}\right| \leqq MN^{1-s}\,,$$

with $0 < r \leqq 4$ and $0 \leqq s \leqq 4$.

Then BIEBERBACH's formula is written

$$|\,y(x_0 + h) - \tilde{y}_4(x_0 + h)\,| \leqq MN\,(3.7 + 5.4\,M + 1.3\,M^2 + 0.017\,M^3)\,h^5. \tag{3}$$

It is seen that however rigorous this method may be, it is of little practical value.

Assume now the function f in (1) exempt of y. In this case (2) reduces to SIMPSON's rule.

For the latter we have a well known error formula which is:

$$y(x_0 + h) - \tilde{y}_4(x_0 + h) = -\frac{1}{90}\left(\frac{h}{2}\right)^5 f^{(4)}(\bar{x}) , \tag{4}$$

or

$$| y(x_0 + h) - \tilde{y}_4(x_0 + h) | = \frac{h^5}{2880} f^{(4)}(\bar{x}) .$$

In general $f^{(4)}(x)$ is not a constant function and $\bar{x}$ is unknown. Thus in order to make the formula practical we write in the form:

$$| y(x_0 + h) - \tilde{y}_4(x_0 + h) | \leqq \frac{h^5}{2880} S = T_1 , \tag{5}$$

where $S = \max | f^{(4)}(x) |$.

On the other hand (3) can be written now

$$| y(x_0 + h) - \tilde{y}_4(x_0 + h) | \leqq 3.7 \; H h^5 \; (1 + \text{positive terms}) = T_2 , \tag{6}$$

where $H = \max | f^{(r)}(x) |$, $1 \leqq r \leqq 4$.

Definitely $H \geqq S$ and $3.7 = 10656/2880$. It follows then from a comparison between (5) and (6) that $T_2 > 10656 \; T_1$, that is BIEBERBACH's bound for the truncation error is at least 10656 times larger than that provided by SIMPSON's error formula (5).

As far as LOTKIN's method is concerned [4], it provides a bound which is only about 300 times larger than SIMPSON's. However, LOTKIN's bound is valid only when h is sufficiently small.

All these are well known facts and they can be found, for instance, in [**1**] (particularly p. 326).

The assumption that the function f in (1) is independent of y led us to some kind of grading or scaling of SIMPSON's, BIEBERBACH's and LOTKIN's formulas according to the accuracy of the approximations which they provide. We shall apply this process to pseudo-iterative formulas one of which is:

$$\tilde{y}_5(x_0 + h) = y_0 + \frac{1}{336} [14 \; k_0 + 35 \; k_3 + 162 \; k_4 + 125 \; k_5] , \tag{7 a}$$

where

$$\text{(7 b)}\quad \left\{ \begin{array}{ll} k_0 = h\, f(x_0\,,\; y_0)\,, & \tilde{y}_1(x_0 + h) = y_0 + k_0 \\ k_1 = h\, f\left(x_0 + \frac{1}{2}\, h\,,\; y_0 + \frac{1}{2}\, k_0\right)\,, & \tilde{y}_2(x_0 + h) = y_0 + k_1 \\ k_2 = h\, f\left(x_0 + \frac{1}{2}\, h\,,\; y_0 + \frac{1}{4}\,(k_0 + k_1)\right) & \\ k_3 = h\, f(x_0 + h\,,\; y_0 - k_1 + 2\, k_2)\,, & \tilde{y}_4(x_0 + h) = y_0 + \frac{1}{6}\,(k_0 + 4 k_2 + k_3) \\ k_4 = h\, f\left(x_0 + \frac{2}{3}\, h\,,\; y_0 + \frac{1}{27}\,(7\, k_0 \;+ 10\, k_1 + k_3)\right) & \\ k_5 = h\, f\left(x_0 + \frac{2}{10}\, h,\; y_0 + \frac{16}{10000}\,(28\, k_0 - 125\, k_1 + 546\, k_2 + 54\, k_3 - 378\, k_4\right)\,. & \end{array} \right.$$

The above formula appears to be one of the best from the families of fifth order pseudo-iterative formulas as established in [7]. And since we are in search of everincreasing accuracies, we shall center our attention on this formula.

In the considered case where f is exempt of y, the fifth order formula (7 a) and the imbedded fourth order in it, can be written:

$$\text{(8)}\qquad \tilde{y}_5(x_0 + h) = y_0 + \frac{h}{336}\left[14\, f(x_0) + 35\, f(x_0 + h) + \right.$$
$$\left. 162\, f\left(x_0 + \frac{2}{3}\, h\right) + 125\, f\left(x_0 + \frac{2}{10}\, h\right)\right],$$

$$\text{(9)}\qquad \tilde{y}_4(x_0 + h) = y_0 + \frac{h}{6}\left[f(x_0) + 4\, f\left(x_0 + \frac{1}{2}\, h\right) + f(x_0 + h)\right]\,.$$

The functions $\tilde{y}_4\,(x_0 + h)$, $\tilde{y}_5\,(x_0 + h)$ and $y(x_0 + h)$ have the following TAYLOR series expansions:

$$\tilde{y}_4(x_0 + h) = y_0 + h\, f_0 + \frac{h^2}{2\,!}\, f_0' + \frac{h^3}{3\,!}\, f_0'' + \frac{h^4}{4\,!}\, f_0''' + \frac{5}{576}\, h^5\, f_0^{(4)} + \ldots,$$

$$\tilde{y}_5(x_0 + h) = y_0 + h\, f_0 + \frac{h^2}{2\,!}\, f_0' + \frac{h^3}{3\,!}\, f_0'' + \frac{h^4}{4\,!}\, f_0''' + \frac{h^5}{5\,!}\, f_0^{(4)} + \frac{151}{108000}\, h^6\, f_0^{(5)} + \ldots,$$

$$y\;(x_0 + h) = y_0 + h\, f_0 + \frac{h^2}{2\,!}\, f' + \frac{h^3}{3\,!}\, f_0'' + \frac{h^4}{4\,!}\, f_0''' + \frac{h^5}{5\,!}\, f_0^{(4)} + \frac{h^6}{6\,!}\, f_0^{(5)} + \ldots\;.$$

Then we find

$$(10) \qquad y(x_0 + h) - \tilde{y}_4(x_0 + h) = -\frac{h^5}{2880} f^{(4)}(\bar{x}),$$

$$(11) \qquad \tilde{y}_5(x_0 + h) - \tilde{y}_4(x_0 + h) = -\frac{h^5}{2880} f^{(4)}(\bar{x}_1),$$

$$(12) \qquad y(x_0 + h) - \tilde{y}_5(x_0 + h) = -\frac{h^6}{108000} f^{(5)}(\bar{x}_2).$$

It is seen that (10) is no other than (4). This is obvious, because in the case under consideration the imbedded fourth order formula is reduced to SIMPSON's rule now represented by (9). Besides this, the right hand members of (10) and (11) are alike except for the fact that the points $\bar{x}$ and $\bar{x}_1$ may be different. But these points belong to the same open interval, more precisely to the intervals $(x_0 - h, x_0)$ or $(x_0, x_0 + h)$, $h > 0$, according to whether we wish to progress to the left or right of x_0.

In general, h is small and $f^{(4)}(x)$ is continuous. Under these conditions it is reasonable to assume $f^{(4)}(\bar{x}) = f^{(4)}(\bar{x}_1)$. Then the combination of (10) and (11) yields:

$$\tilde{y}_5(x_0 + h) - \tilde{y}_4(x_0 + h) \approx y(x_0 + h) - \tilde{y}_4(x_0 + h).$$

In other words, the approximate value $\tilde{y}_4(x_0 + h)$, given by the imbedded fourth order formula, agrees with the exact value $y(x_0 + h)$ to the same accuracy as to which $\tilde{y}_5(x_0 + h)$ and $\tilde{y}_4(x_0 + h)$ agree with each other. This is the rule which we have stated earlier for the general case, that is relative to equation (1), with an entirely different but less rigorous approach ([7], pp. 2-7). We shall rediscover again this rule in the general case following a rigorous treatment.

Thus the pseudo-iterative formula (7 a) offers for $\tilde{y}_4$ the error estimate $\tilde{y}_5 - \tilde{y}_4$ which in the case of $y' = f(x)$ may almost coincide with the absolute exact error. This error estimation process does not involve any derivatives but it requires two additional evaluations of $f(x)$, namely, $f\left(x_0 + \frac{2}{3} h\right)$ and $f\left(x_0 + \frac{2}{10} h\right)$.

On the other hand SIMPSON's error formula (5) requires the analytical derivation of $f^{(4)}(x)$ which usually is not convenient especially in digital computer operations. It requires also the determination of $S = \max | f^{(4)}(x) |$ which may well be quite involved. Furthermore, since $S \geqslant |f^{(4)}(\bar{x})|$, the substitution of the latter absolute value by S, although a practical necessity, may substantially reduce the effectiveness of the resulting formula (5).

As far as the accuracy of $\widetilde{y}_5$ is concerned, combining (10) and (12) we find

$$\left| \frac{y(x_0 + h) - \widetilde{y}_5(x_0 + h)}{y(x_0 + h) - \widetilde{y}_4(x_0 + h)} \right| = \frac{h}{37.5} \left| \frac{f^{(5)}(\bar{x}_2)}{f^{(4)}(\bar{x})} \right| .$$

In ordinary cases $f^{(4)}(x)/4!$ and $f^{(5)}(x)/5!$ do not differ too much from each other numerically. Thus, the preceding relation can be written

$$| y(x_0 + h) - \widetilde{y}_5(x_0 + h) | \approx \frac{h}{7.5} | y(x_0 + h) - \widetilde{y}_4(x_0 + h) | .$$

This relation shows that under the precited suitable conditions the absolute error in $\widetilde{y}_4$ is by a factor of $h/(7.5)$ larger than that in $\widetilde{y}_5$.

3. – Assume $f(x, y) \in C^5$. Let as before $\widetilde{y}_5(x_0 + h)$ be a 5-th order approximation for $y(x_0 + h)$ obtained through the use of the pseudo-iterative formula (7 a).

The application of TAYLOR's theorem gives

$$y(x_0 + h) - \widetilde{y}_5(x_0 + h) = h^6 M_5(h) , \tag{13}$$

where $M_5(h)$, called a constant of proportionality, is actually a function of h.

The step-size $h \neq 0$. From (13) we have

$$\frac{y(x_0 + h) - \widetilde{y}_5(x_0 + h)}{h^6} = M_5(h), \tag{14}$$

Since the left side of (14) is a continuous function of h, so must be $M_5(h)$. Thus a small change in the step-size h will produce a small change in the value of $M_5(h)$.

Likewise relative to the imbedded 4-th order formula we can write

$$y(x_0 + h) - \widetilde{y}_4(x_0 + h) = h^5 M_4(h) , \tag{15}$$

where $M_4(h)$ is a continuous function of h.

Let c, $0 < c \neq 1$, be an arbitrarily selected constant near unity.

Replacing h by ch in (13) and (15) we obtain:

$$y(x_0 + ch) - \widetilde{y}_5(x_0 + ch) = c^6 h^6 M_5(ch) , \tag{16}$$

$$y(x_0 + ch) - \widetilde{y}_4(x_0 + ch) = c^5 h^5 M_4(ch) . \tag{17}$$

The combination of (13) and (15) yields:

$$d(h) = \tilde{y}_5(x_0 + h) - \tilde{y}_4(x_0 + h) = h^5 M_4(h) - h^6 M_5(h)\,. \tag{18}$$

Likewise the combination of (16) and (17) yields:

$$d(ch) = \tilde{y}_5(x_0 + ch) - \tilde{y}_4(x_0 + ch) = c^5 h^5 M_4(ch) - c^6 h^6 M_5(ch)\,. \tag{19}$$

Let

$$X = h^5 M_4(h)\,, \qquad Y = h^6 M_5(h)\,.$$

Since $M_4(h)$ is continuous, we know, for a given $\varepsilon' > 0$ there corresponds a $\delta > 0$ such that

$$| M_4(h) - M_4(ch) | < \varepsilon' \qquad \text{whenever} \quad | h - ch | < \delta$$

or

$$| X - h^5 M_4(ch) | < h^5 \varepsilon' = \varepsilon \qquad \text{whenever} \quad h | c - 1 | < \delta\,.$$

Thus if ε is of the desired or accepted accuracy in our computations and approximations, then for a chosen c by taking h sufficiently small we can set: $X = h^5 M_4(ch)$. Likewise we can set $Y = h^6 M_5(ch)$.

The substitution of X's and Y's in (18) and (19) gives:

$$\left\{\begin{array}{l} X - Y = d(h) \\ \\ c^5 X - c^6 Y = d(ch)\,. \end{array}\right. \tag{20}$$

This system of linear equations has as solution:

$$X = \frac{d(ch) - c^6\, d(h)}{c^5\,(1 - c)}\,, \qquad Y = \frac{d(ch) - c^5\, d(h)}{c^5\,(1 - c)}\,;$$

that is

$$h^5 M_4 = \frac{d(ch) - c^6\, d(h)}{c^5\,(1 - c)}\,, \qquad h^6 M_5 = \frac{d(ch) - c^5\, d(h)}{c^5\,(1 - c)}\,. \tag{21}$$

The substitution from (21) into the equations (13), (15), (16) and (17) yields:

$$y(x_0 + h) - \tilde{y}_5(x_0 + h) = \frac{d(ch) - c^5\, d(h)}{c^5\,(1-c)}\,, \tag{22 a}$$

$$y(x_0 + h) - \tilde{y}_4(x_0 + h) = \frac{d(ch) - c^6\, d(h)}{c^5\,(1-c)}\,, \tag{22 b}$$

$$y(x_0 + ch) - \tilde{y}_5(x_0 + ch) = \frac{c[d(ch) - c^5\, d(h)]}{1-c}\,, \tag{23 a}$$

$$y(x_0 + ch) - \tilde{y}_4(x_0 + ch) = \frac{d(ch) - c^6\, d(h)}{1-c}\,. \tag{23 b}$$

It must be pointed out that the second equation in (20) is only approximately true. It follows then that the formulas (22 a, b) and (23 a, b) must also be only approximately true.

Taking this fact into consideration we can state the following:

Theorem 1. *Let $\tilde{y}_4(x_0 + h)$ and $\tilde{y}_5(x_0 + h)$ be the fourth and fifth order approximations provided by a fifth order pseudo-iterative formula. Furthermore, let:*

$$d(h) = \tilde{y}_5(x_0 + h) - \tilde{y}_4(x_0 + h)\,,$$
$$d(ch) = \tilde{y}_5(x_0 + ch) - \tilde{y}_4(x_0 + ch)\,,$$

where c, $0 < c \neq 1$, is a constant near unity. Then, designating by $\tilde{e}(x_0\,, y_0;\; i;\; h)$ an approximation for the true error

$$e(x_0\,, y_0;\; i;\; h) = y(x_0 + h) - \tilde{y}_i(x_0 + h) \qquad (i = 4, 5),$$

we have

$$\tilde{e}(x_0\,, y_0;\; 5;\; h) = \frac{1}{1-c}\left[\frac{d(ch)}{c^5} - d(h)\right], \tag{24 a}$$

$$\tilde{e}(x_0\,, y_0;\; 4;\; h) = \frac{1}{1-c}\left[\frac{d(ch)}{c^5} - c\, d(h)\right], \tag{24 b}$$

$$\tilde{e}(x_0\,, y_0;\; 5;\; ch) = \frac{c}{1-c}\,[d(ch) - c^5\, d(h)]\,, \tag{25 a}$$

$$\tilde{e}(x_0\,, y_0;\; 4;\; ch) = \frac{1}{1-c}\,[d(h) - c^6\, d(h)]\,. \tag{25 b}$$

Henceforth, whenever the context is clear enough to avoid confusion, for the sake of simplicity we shall designate the true and the estimated errors, that is $e(x_0, y_0; i; h)$ and $\tilde{e}(x_0, y_0; i; h)$, $i = 4, 5$, either by $e_i(h)$ and $\tilde{e}_i(h)$ or by $e_i(x_0)$ and $\tilde{e}_i(x_0)$ or merely by e_i and $\tilde{e}_i$, respectively.

4. – For the sake of convenience let us agree to progress to the right of the initial point (x_0, y_0).

Let P designate the point $(x_0, 0)$ and let $Q(x_q, 0)$ and $R(x_r, 0)$ be two distinct points on the x-axis to the right of P.

Assume $x_q = x_1 = x_0 + h$ and $x_r = x_0 + ch$. Then: $h = x_q - x_0$, $ch = x_r - x_0$. Consequently,

$$c = \frac{x_r - x_0}{x_q - x_0} \qquad \text{and} \qquad 1 - c = \frac{x_q - x_r}{x_q - x_0}.$$

With this h and this c the application of (24 a) gives:

$$\tilde{e}_5(x_q) = \frac{d(x_r) - \left(\dfrac{x_r - x_0}{x_q - x_0}\right)^5 d(x_q)}{\left(\dfrac{x_r - x_0}{x_q - x_0}\right)^5 \dfrac{x_q - x_r}{x_q - x_0}} = \frac{x_q - x_0}{x_q - x_r}\left[\left(\frac{x_q - x_0}{x_r - x_0}\right)^5 d(x_r) - d(x_q)\right]. \tag{26}$$

Assume now that we choose $x_r = x_1 = x_0 + h$ and $x_q = x_0 + ch$. Then: $h = x_r - x_0$ and $ch = x_q - x_0$. Consequently,

$$c = \frac{x_q - x_0}{x_r - x_0} \qquad \text{and} \qquad 1 - c = \frac{x_r - x_q}{x_r - x_0}.$$

With this new h and new c the application of (25 a) gives:

$$\left\{\begin{aligned} \tilde{e}_5(x_q) &= \frac{\dfrac{x_q - x_0}{x_r - x_0}\, d(x_q) - \left(\dfrac{x_q - x_0}{x_r - x_0}\right)^6 d(x_r)}{\dfrac{x_r - x_q}{x_r - x_0}} \\ &= \frac{x_q - x_0}{x_q - x_r}\left[\left(\frac{x_q - x_0}{x_r - x_0}\right)^5 d(x_r) - d(x_q)\right]. \end{aligned}\right. \tag{27}$$

It is seen that (26) and (27) are identical. In other words the formulas (24 a) and (25 a) provide the same error estimate to e_5 at x_q. We shall say that the formulas (24 a) and (25 a) are equivalent.

In like manner we can show that (24 b) and (25 b) are equivalent.

Calling $x_0 + h$ and $x_0 + ch$ the point of approximation and the auxiliary point of approximation, respectively, these results can be stated as follows:

Theorem II. *Let x_q and x_r be two points in a right (or left) neighborhood of x_0. Take one of these two points as the point of approximation and the other as the auxiliary point of approximation. Determine with the appropriate formulas of Theorem* I, $\tilde{e}_i(h)$ *and* $\tilde{e}_i(ch)$, $i = 4, 5$. *Interchange the role of these points, that is, consider now as the point of approximation the former auxiliary point of approximation. Find* $\tilde{e}_i(h')$ *and* $\tilde{e}_i(c'h')$, $i = 4, 5$, *where c' and h' are the new c and new h. Then we have:*

$$\tilde{e}_i(h) = \tilde{e}_i(c'h') , \qquad \tilde{e}_i(ch) = \tilde{e}_i(h') , \qquad i = 4, 5.$$

Let us illustrate this error estimation method with the initial value problem ([7], pp. 37-48):

$$\frac{dy}{dx} = \frac{2y}{1+x}; \qquad x_0 = 0, \ \ y_0 = 1 .$$

For $h = 2^{-n}$, $n = 5, 4, 3, 2, 1, 0$, the results are summarized in the following Tables 1-6.

TABLE 1

$$\begin{aligned} y &= 1.063\,476\,562\,500\,000 \\ \tilde{y}_5 &= 1.063\,476\,562\,400\,801 \\ \tilde{y}_4 &= 1.063\,476\,555\,495\,786 \\ d(h) = \tilde{y}_5 - \tilde{y}_4 &= 0.000\,000\,006\,905\,014 \\ e_5 = y - \tilde{y}_5 &= 0.000\,000\,000\,099\,199 \end{aligned}$$

c	$\tilde{e}_5(h)$, $h = 0.031\ 250$
0.5	0.000 000 000 548 966
0.6	0.000 000 000 540 766
0.7	0.000 000 000 535 111
0.8	0.000 000 000 530 687
0.9	0.000 000 000 526 928
1.1	0.000 000 000 520 449
1.2	0.000 000 000 517 507
1.3	0.000 000 000 514 685
1.4	0.000 000 000 511 955
1.5	0.000 000 000 509 297
1.6	0.000 000 000 506 695
1.7	0.000 000 000 504 142
1.8	0.000 000 000 501 631
1.9	0.000 000 000 499 156
2.0	0.000 000 000 496 714

TABLE 2

$$\begin{aligned} y &= 1.128\,906\,250\,000 \\ \tilde{y}_5 &= 1.128\,906\,244\,065 \\ \tilde{y}_4 &= 1.128\,906\,038\,999 \\ d(h) = \tilde{y}_5 - \tilde{y}_4 &= 0.000\,000\,205\,066 \\ e_5 = y - \tilde{y}_5 &= 0.000\,000\,005\,934 \end{aligned}$$

c	$\tilde{e}_5(h)$, $h = 0.062\ 500$
0.5	0.000 000 031 789
0.6	0.000 000 031 457
0.7	0.000 000 031 139
0.8	0.000 000 030 831
0.9	0.000 000 030 530
1.1	0.000 000 029 947
1.2	0.000 000 029 663
1.3	0.000 000 029 384
1.4	0.000 000 029 110
1.5	0.000 000 028 839
1.6	0.000 000 028 574
1.7	0.000 000 028 312
1.8	0.000 000 028 054
1.9	0.000 000 027 800
2.0	0.000 000 027 550

TABLE 3

$$y = 1.265\ 625\ 000\ 000$$
$$\tilde{y}_5 = 1.265\ 624\ 673\ 166$$
$$\tilde{y}_4 = 1.265\ 618\ 992\ 695$$
$$d(h) = \tilde{y}_5 - \tilde{y}_4 = 0.000\ 005\ 680\ 471$$
$$e_5 = y - \tilde{y}_5 = 0.000\ 000\ 326\ 833$$

c	$\tilde{e}_5(h)$, $h = 0.125\ 000$
0.5	0.000 001 763 256
0.6	0.000 001 729 452
0.7	0.000 001 696 733
0.8	0.000 001 665 036
0.9	0.000 001 634 310
1.1	0.000 001 575 602
1.2	0.000 001 547 545
1.3	0.000 001 520 306
1.4	0.000 001 493 853
1.5	0.000 001 468 156
1.6	0.000 001 443 185
1.7	0.000 001 418 913
1.8	0.000 001 395 313
1.9	0.000 001 372 361
2.0	0.000 001 350 033

TABLE 4

$$y = 1.562\ 500\ 000$$
$$\tilde{y}_5 = 1.562\ 484\ 253$$
$$\tilde{y}_4 = 1.562\ 345\ 679$$
$$d(h) = \tilde{y}_5 - \tilde{y}_4 = 0.000\ 138\ 574$$
$$e_5 = y - \tilde{y}_5 = 0.000\ 015\ 747$$

c	$\tilde{e}_5(h)$, $h = 0.250\ 000$
0.5	0.000 086 402
0.6	0.000 083 241
0.7	0.000 080 265
0.8	0.000 077 459
0.9	0.000 074 810
1.1	0.000 069 937
1.2	0.000 067 694
1.3	0.000 065 567
1.4	0.000 063 549
1.5	0.000 061 632
1.6	0.000 059 809
1.7	0.000 058 074
1.8	0.000 056 423
1.9	0.000 054 848
2.0	0.000 053 346

TABLE 5

$$y = 2.250\ 000\ 000$$
$$\tilde{y}_5 = 2.249\ 393\ 939$$
$$\tilde{y}_4 = 2.246\ 666\ 666$$
$$d(h) = \tilde{y}_5 - \tilde{y}_4 = 0.002\ 727\ 273$$
$$e_5 = y - \tilde{y}_5 = 0.000\ 606\ 060$$

c	$\tilde{e}_5(h)$, $h = 0.500\ 000$
0.5	0.003 414 191
0.6	0.003 184 628
0.7	0.002 978 887
0.8	0.002 793 782
0.9	0.002 626 641
1.1	0.002 337 564
1.2	0.002 212 080
1.3	0.002 097 354
1.4	0.001 992 182
1.5	0.001 895 522
1.6	0.001 806 470
1.7	0.001 724 239
1.8	0.001 648 139
1.9	0.001 577 569
2.0	0.001 511 995

TABLE 6

$$y = 3.999\ 999\ 999$$
$$\tilde{y}_5 = 3.983\ 333\ 334$$
$$\tilde{y}_4 = 3.944\ 444\ 444$$
$$d(h) = \tilde{y}_5 - \tilde{y}_4 = 0.038\ 888\ 890$$
$$e_5 = y - \tilde{y}_5 = 0.016\ 666\ 665$$

c	$\tilde{e}_5(h)$, $h = 1.000\ 000$
0.5	0.096 767
0.6	0.085 566
0.7	0.076 279
0.8	0.068 498
0.9	0.061 914
1.1	0.051 463
1.2	0.047 275
1.3	0.043 623
1.4	0.040 418
1.5	0.037 589
1.6	0.035 079
1.7	0.032 842
1.8	0.030 838
1.9	0.029 036
2.0	0.027 409

In these investigated examples with $c = 0.5 - 2.0$, $c \neq 1$, as shown in the Tables 1-6, the best results are always obtained with $c = 2.0$. In this case the formula (24 a) reduces to

$$\tilde{e}_5(h) = d(h) - \frac{d(2h)}{32}. \tag{28}$$

The introduction of additional decimal figures to c does not in general appear to affect noticeably the value of $\tilde{e}_5$. This can be attributed mainly to round-off errors. These errors are inherited from $\tilde{y}_5(h)$, $\tilde{y}_4(h)$, $\tilde{y}_5(ch)$ and $\tilde{y}_4(ch)$, and are also originated by the error estimation formulas proper.

Indeed, the closer c is to unity the closer $M_5(ch)$ is to $M_5(h)$, but also the larger the numerical value of the factor $1/(1-c)$ in (24 a, b) and (25 a, b) becomes.

Thus if $c = 2$, $1/(1-c) = -1$, but if $c = 0.999$ then $1/(1-c) = 1000$.

It follows that with the latter value of c, the round-off errors present in the bracket in formulas (24 a, b) and (25 a, b) will automatically be magnified a thousandfold. Thus by taking c closer to unity the gain resulting from the increased tendency of $M_5(h)$ and $M_4(h)$ to behave like constant functions may well be offset by this magnification of round-off errors. And this can be verified experimentally. Some of our experimental results are condensed in the Table 7, which can be considered as a complement of Table 1, the step-size being the same in both of them.

TABLE 7

$y = 1.063\,476\,562\,500\,000$
$\tilde{y}_5 = 1.063\,476\,562\,400\,801$
$\tilde{y}_4 = 1.063\,476\,555\,495\,786$
$d(h) = \tilde{y}_5 - \tilde{y}_4 = 0.000\,000\,006\,905\,014$
$e_5 = y - \tilde{y}_5 = 0.000\,000\,000\,099\,199$

c	$\tilde{e}_5(h)$, $h = 0.031\,250$
0.999 999 1	0.000 000 000 601 532
0.999 999 2	0.000 000 000 663 218
0.999 999 3	0.000 000 000 742 527
0.999 999 4	0.000 000 000 663 231
0.999 999 5	0.000 000 000 552 215
0.999 999 6	0.000 000 000 663 244
0.999 999 7	0.000 000 000 848 288
0.999 999 8	0.000 000 001 218 370
0.999 999 9	0.000 000 001 218 376
1.000 000 1	0.000 000 000 108 166
1.000 000 2	0.000 000 000 108 173
1.000 000 3	0.000 000 000 108 180
1.000 000 4	0.000 000 000 385 742
1.000 000 5	0.000 000 000 330 237
1.000 000 6	0.000 000 000 293 237
1.000 000 7	0.000 000 000 425 412
1.000 000 8	0.000 000 000 385 768
1.000 000 9	0.000 000 000 354 936
1.000 001 0	0.000 000 000 441 293
2.000 000 0	0.000 000 000 496 714

It can be seen from this table that, contrary to one's expectations, the two values $c = 1 \pm 10^{-7}$, do not generate values for $\tilde{e}_5$ which are close to each other.

Indeed, with $c = 1 + 10^{-7} = 1.000\,000\,1$ we obtain a better than 10-decimal figure approximation for e_5. However, with the other value, i. e., $c = 0.999\,999\,9$, the obtained approximation is correct to eight decimal figures only. As c increases with equal increments of 10^{-7}, the corresponding values of $\tilde{e}_5$ show fluctuations which can be attributed only to round-off errors. Taking c very close to unity is a hazardous undertaking since, as pointed out before and now experimentally verified, we do not necessarily do any better than with $c = 2$, unless the round-off errors in $\tilde{y}_5(h)$, $\tilde{y}_4(h)$, $\tilde{y}_5(ch)$ and $\tilde{y}_4(ch)$ are made negligible. This can be achieved either by adding more decimals to our calculations or taking smaller stepsize. The first solution does not present any particular difficulty but is subject to the limitations of the available computer. The second approach will be treated in the next section.

5. – Suppose that for some step-size h', $d(h') \neq 0$ but $d(h'/2) = 0$. We then can write: $d(h'/2^i) = 0 \quad (i = 1, 2, \ldots)$.

Consider on the x-axis the four consecutive points

$$Q(x_0 + h',\ 0), \quad R\left(x_0 + \frac{h'}{2}, 0\right), \quad S\left(x_0 + \frac{h'}{4}, 0\right) \quad \text{and} \quad T\left(x_0 + \frac{h'}{8}, 0\right).$$

With the pair of points Q and R and $c = 2$, that is considering R as the point of approximation, the formula (24 a) gives at $x = x_0 + \dfrac{h'}{2}$, $\tilde{e}_5(h'/2) = -\dfrac{d(h')}{32} \neq 0$. On the other hand, with the pair of points R and S and $c = 1/2$, at the same point $x = x_0 + \dfrac{h'}{2}$ the formula (24 a) yields $\tilde{e}_5(h'/2) = 0$.

Had the round-off errors been negligible and had (24 a, b), (25 a, b) been exact formulas, we would not get two distinct values for $\tilde{e}_5$ at $x = x_0 + \dfrac{h'}{2}$.

However, with either pair of points R and S or S and T the formula (24 a) gives at $x = x_0 + \dfrac{h'}{4}$ the same value for $\tilde{e}_5$; more precisely, $\tilde{e}_5(h'/4) = 0$.

This ideal outcome can not be attributed to coincidence. Indeed, starting from this point $x_0 + 2^{-2}h'$ at all consecutive points $x_0 + 2^{-n}h'$ $(n = 3, 4, \ldots)$ the same ideal situation will prevail, that is, we shall have $\tilde{e}_5(2^{-n}h') = 0$ $(n = 3, 4, \ldots)$.

We may thus consider the step-size $h = h'/4$ as being sufficiently small to assume the variations of $M(h)$'s and the round-off errors arising from the use of pseudo-iterative formula (7 a) and the trunction errors as negligible. And thus truly now $\tilde{e}_5(h'/4) = 0$. It follows that with the number of decimal digits retained in our numerals (and calculations) we finally have the ideal result of $\tilde{y}_5(x_0 + h) = y(x_0 + h)$.

The round-off errors and the truncation errors, as well as the errors originating from the formulas (24 a, b), start to become negligible with the use of this ideal step-size, $h = h'/4$. This will be refered to as the *near optimal step-size* and will be designated by h^0. Thus we can announce the following:

Definition. *If for some* h, $d(h) \neq 0$ *but* $d\left(\frac{h}{2}\right) = 0$, *then we shall refer to* $h^0 = \frac{h}{4}$ *as the near-optimal step-size.*

The new point thus obtained, namely $P_1(x_0 + h^0, \tilde{y}_5(x_0 + h^0))$ is an exact point just like $P_0(x_0, y_0)$. Therefore, departing from P_1 as a new but exact initial point and repeating the process of finding a near-optimal step-size we determine another exact point P_2 and so on. However, one will find out that in general the near-optimal step-sizes behave like a constant over some small interval; and that after some j applications of (7 a), at the end of this interval we still have $\tilde{y}_5(x_0 + j\, h^0) = y(x_0 + j\, h^0)$.

In particular, h^0 the near-optimal step-size found at the start, can be used for the determination of the p pivotal exact points necessary for the starting of any predictor-corrector process.

As an illustrative example, let us consider again the initial value problem given earlier on page 118. When $h = 0.00781250$ we find $d(h) = 10^{-11}$ and $d\,(h/2) = 0$. Thus $h^0 = h/4 = 0.001953125$.

Indeed, with this near-optimal step-size, after 512 consecutive applications of (7 a) we find $\tilde{y}_5(1) = y(1) = 4.000\,000\,000\,00$; that is, the exact value to 11 decimal figures.

The computed results are listed in the Table 8 which is self-explanatory.

In any box of the second and third column of this table are listed consecutively $y(h)$, $\tilde{y}_5(h)$, $\tilde{y}_4(h)$ and $y(1)$, $\tilde{y}_5(1)$ and $\tilde{y}_4(1)$, respectively.

TABLE 8

h	$y(h)$, $\tilde{y}_5(h)$, $\tilde{y}_4(h)$	$y(1)$, $\tilde{y}_5(1)$, $\tilde{y}_4(1)$

1.000 000 000 00	4.000 000 000 00
	3.983 333 334 55
	3.944 444 444 44
		4.000 000 000 00
		3.983 333 334 55
		3.944 444 444 44

0.500 000 000 00	2.250 000 000 00
	2.249 393 939 69
	2.246 666 666 67
		4.000 000 000 00
		3.998 755 918 63
		3.997 647 392 81

0.250 000 000 00	1.562 500 000 00
	1.562 484 253 03
	1.562 345 679 01
		4.000 000 000 00
		3.999 939 840 97
		3.999 907 257 84

0,125 000 000 00	1.265 625 000 00
	1.265 624 673 17
	1.265 618 992 70
		4.000 000 000 00
		3.999 997 697 98
		3.999 996 712 21

0.062 500 000 00	1.128 906 250 00
	1.128 906 244 07
	1.128 906 039 00
		4.000 000 000 00
		3.999 999 921 12
		3.999 999 890 81

0.031 250 000 00	1.063 476 562 50
	1.063 476 562 40
	1.063 476 555 50
		4.000 000 000 00
		3.999 999 997 49
		3.999 999 996 55

0.015 625 000 00	1.031 494 140 63
	1.031 494 140 62
	1.031 494 140 40
		4.000 000 000 00
		3.999 999 999 95
		3.999 999 999 92

0.007 812 500 00	1.015 686 035 16
	1.015 686 035 16
	1.015 686 035 15
		4.000 000 000 00
		4.000 000 000 02
		4.000 000 000 01

0.003 906 250 00	1.007 827 758 79
	1.007 827 758 79
	1.007 827 758 79
		4.000 000 000 00
		4.000 000 000 01
		4.000 000 000 01

$\underline{h_0 = 0.001\ 953\ 125\ 00}$	$\underline{1.003\ 910\ 064\ 70}$ 1.003 910 064 70 1.003 910 064 70	$\underline{4.000\ 000\ 000\ 00}$ 4.000 000 000 00 4.000 000 000 00

0.000 976 562 50	1.001 954 078 67
	1.001 954 078 67
	1.001 954 078 67
		4.000 000 000 00
		4.000 000 000 00
		4.000 000 000 00

0.000 488 281 25	1.000 976 800 92
	1.000 976 800 92
	1.000 976 800 92
		4.000 000 000 00
		4.000 000 000 00
		4.000 000 000 00

6 – The application of TAYLOR's theorem to $d(h)$ gives

$$d(h) = \tilde{y}_5(x_0 + h) - \tilde{y}_4(x_0 + h) = M\,h^5 .$$

It is seen that $d(h)$ decreases with decreasing h and that $d(h) \to 0$ as $h \to 0$.

Assume now that for some step-size h, the approximate values $\tilde{y}_5(x_0 + h)$ and $\tilde{y}_4(x_0 + h)$ have been computed, the calculations being carried out to p decimal figures.

Let E designate the tolerated maximum absolute error.

Set

$$E = | \tilde{y}_5(x_0 + h) - \tilde{y}_4(x_0 + h) | .$$

Then $| d(ch) | \leqslant E$, $c < 1$. Since all errors smaller than E are negligible, it follows from (24 a, b) that $\tilde{e}_5(h) = \tilde{e}_4(h) = 0$, that is, $y(x_0 + h) = \tilde{y}_5(x_0 + h) = \tilde{y}_4(x_0 + h)$, approximately, and with a tolerance of E. We have thus rediscovered again the familiar *rule*, namely: the approximation $\tilde{y}_5(x_0 + h)$ agrees with $y(x_0 + h)$ to the same accuracy as to which $\tilde{y}_5(x_0 + h)$ and $\tilde{y}_4(x_0 + h)$ agree with each other.

A modified but more practical version of this rule is the following:

Rule of thumb. *Assume that for some step-size h, the approximations $\tilde{y}_5(x_0 + h)$ and $\tilde{y}_4(x_0 + h)$ have their integral parts and their leading n decimal digits coinciding, the calculations being carried out to p, $p > n$, decimal figures. Then $\tilde{y}_5(x_0 + h)$ and $\tilde{y}_4(x_0 + h)$ also agree up to their n-th decimal digit with $y(x_0 + h)$.*

It must be pointed out that since $d(h) = O(h^5)$ and $d(ch) = O(h^5)$, the formula (24 a) indicates $\tilde{e}_5(h) = O(h^5)$. On the other hand, since $e_5(h) = O(h^6)$, it follows then that $| \tilde{e}_5(h) | > | e_5(h) |$.

This rapid and simple error estimation method based on this rule does not require any additional substitution except that necessary for the evaluation of $\tilde{y}_5(x_0 + h)$; the other more accurate estimation method based upon formula (24 a) necessitates five additional substitutions.

7. – The formulas (24 a, b) and the related error estimation method can be extended easily to systems of differential equations and differential equations of order $n \geqslant 2$.

For instance, consider the system ([7], p. 44-50)

$$\begin{cases} \dfrac{\mathrm{d}y}{\mathrm{d}x} = f^1(x,\ y,\ z) \\[2ex] \dfrac{\mathrm{d}z}{\mathrm{d}x} = f^2(x,\ y,\ z) \end{cases}$$

subject to the condition: $y(x_0) = y_0$, $z(x_0) = z_0$.

The corresponding pseudo-iterative formulas are:

$$(29)\quad \begin{cases} \tilde{y}_5(x_0 + h) = y_0 + \dfrac{1}{336}\,(14\,k_0^1 + 35\,k_3^1 + 162\,k_4^1 + 125\,k_5^1) \\[2ex] \tilde{y}_4(x_0 + h) = y_0 + \dfrac{1}{6}\,(k_0^1 + 4\,k_2^1 + k_3^1)\,, \end{cases}$$

$$\begin{cases} \tilde{z}_5(x_0 + h) = y_0 + \dfrac{1}{336}\,(14\,k_0^2 + 35\,k_3^2 + 162\,k_4^2 + 125\,k_5^2) \\[2ex] \tilde{z}_4(x_0 + h) = y_0 + \dfrac{1}{6}\,(k_0^2 + 4\,k_2^2 + k_3^2)\,, \end{cases}$$

where

$$\begin{cases} k_0^1 = h\,f^1(x_0\,,\ y_0\,,\ z_0) \\[1ex] k_0^2 = h\,f^2(x_0\,,\ y_0\,,\ z_0)\,, \end{cases}$$

$$\begin{cases} k_1^1 = h\,f^1\left(x_0 + \dfrac{1}{2}\,h\,,\ y_0 + \dfrac{1}{2}\,k_0^1\,,\ z_0 + \dfrac{1}{2}\,k_0^2\right) \\[2ex] k_1^2 = h\,f^2\left(x_0 + \dfrac{1}{2}\,h\,,\ y_0 + \dfrac{1}{2}\,k_0^1\,,\ z_0 + \dfrac{1}{2}\,k_0^2\right), \end{cases}$$

$$\begin{cases} k_2^1 = h\,f^1\left(x_0 + \dfrac{1}{2}\,h\,,\ y_0 + \dfrac{1}{4}\,(k_0^1 + k_1^1)\,,\ z_0 + \dfrac{1}{4}\,(k_0^2 + k_1^2)\right) \\[2ex] k_2^2 = h\,f^2\left(x_0 + \dfrac{1}{2}\,h\,,\ y_0 + \dfrac{1}{4}\,(k_0^1 + k_1^1)\,,\ z_0 + \dfrac{1}{4}\,(k_0^2 + k_1^2)\right), \end{cases}$$

$$\begin{cases} k_3^1 = h\,f^1(x_0 + h\,,\ y_0 - k_1^1 + 2\,k_2^1\,,\ z_0 - k_1^2 + 2\,k_2^2) \\[1ex] k_3^2 = h\,f^2(x_0 + h\,,\ y_0 - k_1^1 + 2\,k_2^1\,,\ z_0 - k_1^2 + 2\,k_2^2)\,, \end{cases}$$

$$\begin{cases} k_4^1 = h\,f^1\left(x_0 + \dfrac{2}{3}\,h,\ y_0 + \dfrac{1}{27}\,(7\,k_0^1 + 10\,k_1^1 + k_3^1),\ z_0 + \dfrac{1}{27}\,(7\,k_0^2 + 10 k_1^2 + k_3^2)\right) \\[2ex] k_4^2 = h\,f^2\left(x_0 + \dfrac{2}{3}\,h,\ y_0 + \dfrac{1}{27}\,(7\,k_0^1 + 10\,k_1^1 + k_3^1),\ z_0 + \dfrac{1}{27}\,(7\,k_0^2 + 10 k_1 + k_3)\right), \end{cases}$$

$$\left\{\begin{aligned}
k_5^1 &= h\, f^1\Big(x_0 + \frac{2}{10}\,h\,,\; y_0 + \frac{16}{10000}\,(28\,k_0^1 - 125\,k_1^1 + 546\,k_2^1 + 54\,k_3^1 - 378\,k_4^1)\,, \\
&\qquad z_0 + \frac{16}{10000}\,(28\,k_0^2 - 125\,k_1^2 + 546\,k_2^2 + 54\,k_3^2 - 378\,k_4^2)\Big) \\
k_5^2 &= h\, f^2\Big(x_0 + \frac{2}{10}\,h\,,\; y_0 + \frac{16}{10000}\,(28\,k_0^1 - 125\,k_1^1 + 546\,k_2^1 + 54\,k_3^1 - 378\,k_4^1), \\
&\qquad z_0 + \frac{16}{10000}\,(28\,k_0^2 - 125\,k_1^2 + 546\,k_2^2 + 54\,k_3^2 - 378\,k_4^2)\Big).
\end{aligned}\right.$$

The error estimation formulas (24 a, b) become now:

$$\tilde{e}_5(y,\; h) = \frac{1}{1-c}\left[\frac{d(y,\; ch)}{c^5} - \; d(y,\; h)\right],$$

$$e_4(y,\; h) = \frac{1}{1-c}\left[\frac{d(y,\; ch)}{c^5} - c\, d(y,\; h)\right],$$

$$\tilde{e}_5(z,\; h) = \frac{1}{1-c}\left[\frac{d(z,\; ch)}{c^5} - \; d(z,\; h)\right],$$

$$\tilde{e}_4(z,\; h) = \frac{1}{1-c}\left[\frac{d(z,\; ch)}{c^5} - c\, d(z,\; h)\right],$$

where now

$$\left\{\begin{aligned}
&\tilde{e}_i(y,\; h) \approx e_i(y,\; h) = y(x_0 + h) - \tilde{y}_i(x_0 + h) &&\qquad (i = 4,\; 5) \\
&\tilde{e}_i(z,\; h) \approx e_i(z,\; h) = z(x_0 + h) - \tilde{z}_i(x_0 + h) &&\qquad (i = 4,\; 5),
\end{aligned}\right.$$

$$\begin{aligned}
d(y,\; h) &= y_5(h) - \tilde{y}_4(h)\,, \\
d(z,\; h) &= \tilde{z}_5(h) - \tilde{z}_4(h)\,, \\
d(y,\; ch) &= \tilde{y}_5(ch) - \tilde{y}_4(ch)\,, \\
d(z,\; ch) &= \tilde{z}_5(ch) - y_4(ch)\,.
\end{aligned}$$

As far as differential equations of order $n \geqslant 2$ or systems of higher order differential equations are concerned, they can be written as a system of first order differential equations to which pseudo-iterative formulas apply. The reader in each case can readily write the associated error formulas as above.

It must be mentioned that the formulas (24 a, b), (28) and (29) and the related error estimation method are not restricted to pseudo-iterative formulas. Any two formulas of fourth and fifth order can be used instead. But this will require at least seven substitutions instead of six.

References.

[1] H. A. Antosiewicz and W. Gautschi, *Numerical methods in ordinary differential equations*, Survey of Numerical Analysis, McGraw-Hill, New York 1962 (cf. pp. 323-326).

[2] L. Bieberbach, *On the remainder of the Runge-Kutta formula of the theory of ordinary differential equations*, Z. Angew. Math. Physik **2** (1951), 233-248.

[3] L. Collatz, *The Numerical Treatment of Differential Equations*, Third Edition, Springer, Berlin 1960 (cf. p. 51).

[4] M. Lotkin, *On the Accuracy of Runge-Kutta's Method*, Math. Tables and Other Aids to Computation **5** (1961), 128-133.

[5] D. Sarafyan, *Error estimation for Runge-Kutta methods*, Notices Amer. Math. Soc. **12** (1965), p. 572.

[6] D. Sarafyan, *Runge-Kutta formulas in pseudo-iterative form*, Notices Amer. Math. Soc. **13** (1966), p. 224.

[7] D. Sarafyan, *Error estimation for Runge-Kutta methods through pseudo-iterative formulas*, Riv. Mat. Univ. Parma (2) **9** (1968), 1-42.

Résumé.

Dans un article antérieur l'auteur a établi les formules pseudo-itératives de Runge-Kutta. Ces formules donnent non seulement des approximations au cinquième ordre, mais encore des estimations des erreurs sans nécessiter de nouvelles évaluations de la fonction.

Ici on développe une méthode qui améliore les résultats précédents. Cette nouvelle méthode pour estimer les erreurs est géneralisée à touts les types de formules de Runge-Kutta et étendue aussi aux systèmes d'équations différentielles ordinaires.

* * *