MATHEMATICS OF COMPUTATION, VOLUME 27, NUMBER 122, APRIL, 1973

A Modified Butcher Formula for Integration of Stiff Systems of Ordinary Differential Equations

By H. Nosrati

Abstract. An A-stable one-step integration formula, called the modified Butcher (MB), is presented and is shown to have an order of accuracy $p = 3$, when the differential system is linear, and $p = 2$ otherwise. A method for evaluating the local truncation error of the formula is also suggested. Finally, the main features of this formula, vis-a-vis the trapezoidal, are compared.

I. Introduction. The characterization of the dynamic behavior of many processes (e.g. chemical and nuclear reactor kinetics, circuit analysis and control systems) often gives rise to a set of N first-order differential equations of the form

$$(1) \qquad \dot{X}(t) = f[X(t)], \qquad X(t_0) = X_0, \qquad t \in [t_0, t_f].$$

These equations usually exhibit the stiffness property, i.e., locally, their Jacobian matrix $J(t) = \partial f/\partial X$ contains widely separated eigenvalues. Their efficient solution requires the use of special integration formulas which allow relatively large step-sizes without becoming unstable. In the literature, integration formulas enjoying such a property in the entire left-half of the $h\lambda$-plane, are called A-stable integration formulas (e.g. see [1]). It is well known that such formulas are necessarily implicit [1] and their proper implementation requires the solution, in general, of a set of nonlinear algebraic equations at each integration step.

Dahlquist [1] has shown that the order of accuracy p of an A-stable linear multistep integration formula cannot exceed 2. For example, with the exception of the backward Euler formula with $p = 1$, the trapezoidal, Gear's [2] and those recently introduced by Genin [3] are of second order. Outside the linear multistep class, A-stable integration formulas do exist with order of accuracy $p > 2$. These include the set of one-step methods of Liniger and Willoughby [4] and the implicit Runge-Kutta two-stage process of Butcher [5]. For another example, see [1]. It must be noted that linear multistep integration formulas with $p > 2$ exist which are stable except for a specified region of the left-half of the $h\lambda$-plane [2], [3].

The purpose of this paper is to report the results of an investigation on the existence of other A-stable integration methods. The research begins with an examination of the implicit Runge-Kutta two-stage process of Butcher and the subsequent need for its modification. The result is an A-stable one-step integration formula with the order of accuracy $p = 3$ if the system of equations in (1) is linear and $p = 2$

Received May 30, 1972.

AMS (MOS) subject classifications (1970). Primary 65L05.

Key words and phrases. Stiff system of ordinary differential equations, A-stability, Newton-Raphson iterations, Butcher's implicit Runge-Kutta process.

otherwise. But, as will be shown, in many cases, even for a nonlinear system of differential equations, the error of the formula is smaller than those of trapezoidal or Gear's second-order formula.

II. Derivation of Main Results. Consider the following implicit Runge-Kutta two-stage integration formula of Butcher [5]:

$$g_1 = f[y_n + h(b_{11}g_1 + b_{12}g_2)], \tag{2-1}$$

$$g_2 = f[y_n + h(b_{21}g_1 + b_{22}g_2)], \tag{2-2}$$

$$y_{n+1} = y_n + h(b_1 g_1 + b_2 g_2). \tag{2-3}$$

This formula has an order of accuracy $p = 4$ provided that $b_{11} = b_{22} = \frac{1}{4}$, $b_{12} = \frac{1}{4} - \sqrt{3}/6$, $b_{21} = \frac{1}{4} + \sqrt{3}/6$, $b_1 = b_2 = \frac{1}{2}$. In (2), $f[\cdot]$ represents the same function as the one on the right-hand side of (1) and the symbol y_n is used to denote the computed value of the solution at a given time $t_n = t_0 + nh$. Furthermore, X_n $[=X(t_n)]$ is used to denote the true solution of (1) evaluated at time $t = t_n$, h denotes the stepsize and $\dot{y}_n = f[y_n]$.

Although (2) is attractive as an A-stable integration formula of high accuracy (e.g. see Ehle [6]), its implementation requires solution of two sets of implicit systems of equations. This is a definite disadvantage. Moreover, suppose that it is desired to use (2) to integrate the following single first-order differential equation

$$\dot{x}(t) = \lambda x(t), \qquad x(t_0) = x_0 \tag{3}$$

in which λ is a complex number such that Re $\lambda < 0$. Substitution of (3) into (2) yields

$$y_{n+1} = \frac{12 + 6h\lambda + (h\lambda)^2}{12 - 6h\lambda + (h\lambda)^2} y_n, \qquad y_0 = x_0. \tag{4}$$

Now, for values of Re $h\lambda$ in the far left-half of the $h\lambda$-plane, the true solution $x(t) = e^{\lambda(t-t_0)}x_0$ is nearly zero whereas the computed solution obtained from (4) may not be unless y_n is already sufficiently close to zero. This situation, although not critical, can become cause for concern, for, in a typical stiff system of equations, the parasitic effects of the transient can still affect the behavior of the dc solution.

The preceding arguments tend to support the need for modification of formula (2), though such a modification will affect the order of accuracy p of the formula. Nevertheless, the modification will be affected to achieve the following objectives:

(i) Preserving A-stability and self-starting nature of the formula.

(ii) The need to solve at most one set of implicit equations.

(iii) When the formula is used to compute a solution to (3), then the computed solution should tend to zero as Re $h\lambda \rightarrow \infty$.

(iv) The formula should have the highest order of accuracy consistent with the above objectives.

Pursuant to objective (ii), formula (2) is changed to the following form:

$$g_1 = f[y_n + h(a_{11}g_1 + a_{12}g_2)], \tag{5-1}$$

$$g_2 = f[y_n + ha_{21}g_1], \tag{5-2}$$

$$y_{n+1} = y_n + ha_1 g_1. \tag{5-3}$$

The four parameters a_{11}, a_{12}, a_{21}, and a_1 are chosen to realize the remaining objectives. This choice will make Eqs. (5) third order if $f[\cdot]$ is linear and second order otherwise. A routine calculation yields the following values for the parameters:

$$a_{11} = \tfrac{2}{3}, \qquad a_{12} = -\tfrac{1}{6}, \qquad a_{21} = 1, \qquad a_1 = 1.$$

Now, since $a_{21} = a_1$, it follows that $g_2 = \dot{y}_{n+1}$, and hence a final expression for y_{n+1} is obtained as

$$y_{n+1} = y_n + hf[\tfrac{1}{3}y_n + \tfrac{2}{3}y_{n+1} - \tfrac{1}{6}h\dot{y}_{n+1}]. \tag{6}$$

When $f[\cdot]$ is linear and of the form given in (3), Eq. (6) yields

$$y_{n+1} = \frac{1 + \frac{1}{3}\lambda h}{1 - \frac{2}{3}\lambda h + \frac{1}{6}(\lambda h)^2}\, y_n \tag{7}$$

which is of the form required by (iii). Moreover, in this case, (6) may be written as

$$\begin{aligned} y_{n+1} &= y_n + \tfrac{1}{3}h\{f[y_n] + 2f[y_{n+1}]\} - \tfrac{1}{6}h^2 f[\dot{y}_{n+1}], \\ y_{n+1} &= y_n + \tfrac{1}{3}h[\dot{y}_n + 2\dot{y}_{n+1}] - \tfrac{1}{6}h^2\ddot{y}_{n+1}. \end{aligned} \tag{8}$$

It is noted that formulation (8) is a special case of the class of A-stable one-step integration formulas of Liniger and Willoughby [4]. It has an order of accuracy $p = 3$, but, because of the presence of the second derivative term, its use is limited to cases where the Jacobian of the system in (1) is readily available.

III. The Error Term. The local truncation error of (6) is, in the general case,

$$T_{n+1} = (h^3/24)\phi_n + O(h^4) \tag{9}$$

where ϕ_n denotes the principal error function of (6) and is given by

$$\phi_n = [((\partial^2 f/\partial X^2)f)f](X_n). \tag{10}$$

When $f[\cdot]$ is linear, the local truncation error of (6) becomes

$$T_{n+1} = (h^4/72)X_{n+1}^{(4)} + O(h^5) \tag{11}$$

where $X^{(4)}(t) = (d^4/dt^4)X(t)$. Equations (9)–(11) can be readily verified by the substitution of the true solution $X(t)$ in place of y in (6), expansion of $f[\cdot]$ in a Taylor series about X_n and collection of all the terms corresponding to like powers of h. Also in (10), the term $\partial^2 f/\partial X^2$ is, in general, a third rank tensor with components $\partial^2 f_i/\partial X_j \partial x_k$, $i, j, k = 1, 2, \cdots, n$. It vanishes, however, when $f[\cdot]$ is linear.

In controlling the local truncation error, it is necessary to compute the principal error function (9) or the fourth derivative of $X(t)$ in (11) as may be appropriate. Neither task is easy or desirable. As an alternative, an appropriate predictor formula may be used. For example, the following formula

$$y_{n+1}^{P} = y_{n-1} + 2hf[\tfrac{1}{3}y_{n-1} + \tfrac{2}{3}y_n + \tfrac{1}{3}h\dot{y}_n] \tag{12}$$

may be employed to predict a value to the solution at each step. For the general case, (12) has a local truncation error

$$T_{n+1}^{P} = \tfrac{1}{3}h^3\phi_n + O(h^4). \tag{13}$$

When $f[\cdot]$ is linear, the local truncation error of (12) is

$$T^P_{n+1} = \tfrac{1}{9}h^4 X^{(4)}_{n+1} + O(h^5). \tag{14}$$

If the symbols P_{n+1} and C_{n+1} are used to denote the predicted and corrected values as obtained from (12) and (6), respectively, then the local truncation error of (6), for $f[\cdot]$ linear or otherwise, is approximately given by

$$T_{n+1} = \tfrac{1}{7}(P_{n+1} - C_{n+1}). \tag{15}$$

A more flexible version of (12) is

$$y^P_{n+1} = y_{n-1} + Rf[RS(y_n - y_{n-1} - h_2\dot{y}_n) + y_{n-1} + \tfrac{1}{2}R\dot{y}_n] \tag{16}$$

where

$$h_1 = t_{n+1} - t_n, \qquad h_2 = t_n - t_{n-1},$$
$$R = h_1 + h_2, \qquad S = (2h_2 - h_1)/3h_2^2.$$

In this case, any double computation will be eliminated when a change in the step size h is contemplated. However, Eq. (15) for the local truncation error must be changed to

$$T_{n+1} = \frac{K^3(P_{n+1} - C_{n+1})}{3K^3 + 3(K+1)(K-1)^2 + 4} \quad \text{if } f[\cdot] \text{ is nonlinear,} \tag{17}$$

$$T_{n+1} = \frac{K^4(P_{n+1} - C_{n+1})}{2K^4 + 4K^3 + 1} \quad \text{if } f[\cdot] \text{ is linear,} \tag{18}$$

where K is defined by $h_1 = Kh_2$.

IV. Corrector Solution by Recursive Iteration. When $f[\cdot]$ is linear, the solution of (6) for y_{n+1} presents no serious problem. However, for the general case, an alternative to repeated back substitution into (6) itself is use of the Newton-Raphson iteration method. When (6) is expressed in the form

$$F[y_{n+1}] = y_{n+1} - y_n - hf[\tfrac{1}{3}y_n + \tfrac{2}{3}y_{n+1} - \tfrac{1}{6}h\dot{y}_{n+1}] = 0, \tag{19}$$

then a sequence $\{y^i_{n+1}\}^r_{i=1}$ may be obtained from

$$y^{i+1}_{n+1} = y^i_{n+1} - B_i^{-1}F_i \tag{20}$$

where y^i_{n+1} denotes the value of y_{n+1} at the ith iteration, B_i is the Jacobian of $F[y^i_{n+1}]$ and $F_i = F[y^i_{n+1}]$. The explicit expression for B_i is given by*

$$B_i = I - \tfrac{2}{3}h\hat{J}_i(I - \tfrac{1}{4}hJ_i) \tag{21}$$

where $\hat{J}_i$ and J_i are the Jacobian of $f[V]$ evaluated at $V = \frac{1}{3}y_n + \frac{2}{3}y^i_{n+1} - \frac{1}{6}h\dot{y}^i_{n+1}$ and $V = y^i_{n+1}$, respectively. Normally, the sequence of iterations produced by (20) is terminated when the condition

$$||y^{i+1}_{n+1} - y^i_{n+1}|| \leqq C\,||y^i_{n+1}|| \tag{22}$$

for some constant C and an appropriate norm is satisfied. It can be shown [7] that

* See Appendix I for derivation.

the sequence $\{y^1_{n+1}, y^2_{n+1}, \cdots, y^r_{n+1}, \cdots\}$ will converge to a solution y_{n+1}, such that $F[y_{n+1}] = 0$ in the sense of (22), when y^0_{n+1} satisfies the following condition for a reasonably small number d:

$$||y_{n+1} - y^0_{n+1}|| \leqq d. \tag{23}$$

V. Concluding Remarks. The Modified Butcher (MB) formula (6), like the trapezoidal, is a one-step formula requiring two function evaluations. Whereas the trapezoidal formula, for sufficiently large values of [Re $h\lambda$] oscillates, formula (6) does not. For the trapezoidal, the order of accuracy is $p = 2$ regardless of the nature of $f[\cdot]$. Formula (6) has $p = 2$ when $f[\cdot]$ is nonlinear, and $p = 3$ otherwise. On the other hand, when using the Newton-Raphson iteration scheme (20), for each iteration, formula (6) requires one more function and Jacobian evaluations than the trapezoidal, in addition to one matrix multiplication. However, with an appropriate choice of step size, it may be possible, in many problems, to calculate only J_i or $\hat{J}_i$. Moreover, in many sparse systems, two evaluations of the Jacobian and one matrix multiplication will not generally be very serious.

It is conjectured that even for nonlinear systems, the local truncation error of (6) will be generally smaller than that of the trapezoidal. However, accurate and practical estimation of the error, and hence step size control, of formula (6) is a problem as it generally is with other formulas. A method has been suggested in Section III (Eqs. (12)-(18)), but it has the drawback of requiring two function evaluations. Clearly, a less costly predictor would greatly enhance the appeal of formula (6). As was noted earlier, for the special case of linear $f[\cdot]$, formula (8) of Liniger and Willoughby and the present formula are identical. However, for a general $f[\cdot]$, although (8) has an order of accuracy $p = 3$, it requires the knowledge of the system Jacobian at each step of integration. Consequently, its use is limited to the cases where the Jacobian can be easily computed. On the other hand, formula (6) makes no such requirement, although its use in conjunction with the Newton-Raphson iteration scheme depends on the knowledge of the Jacobian. Here, however, an exact knowledge of the Jacobian is not as critical as in formula (8), so that a scheme such as Broyden's method [8], [9] may suffice.

Preliminary tests, using formulas (6) and (12) along with sequence (20), have proved successful. Further tests comparing performance of (6) and other integration formulas of about the same order of accuracy are under way and the results will be the subject of a future report.

Appendix I

Determination of the coefficient matrix B_i in (20).

Write formula (6) in the form

$$F(y_{n+1}) = y_{n+1} - y_n - hf[\tfrac{1}{3}y_n + \tfrac{2}{3}y_{n+1} - \tfrac{1}{6}h\dot{y}_{n+1}] = 0. \tag{I-1}$$

Let

$$Z_{n+1} = \tfrac{1}{3}y_n + \tfrac{2}{3}y_{n+1} - \tfrac{1}{6}hf[y_{n+1}]. \tag{I-2}$$

Then (I-1) becomes

$$F(y_{n+1}) = y_{n+1} - y_n - hf[Z_{n+1}] = 0. \tag{I-3}$$

By use of the chain rule differentiation,

$$\frac{\partial F}{\partial y_{n+1}} = I - h\frac{\partial f}{\partial Z_{n+1}}\cdot\frac{\partial Z_{n+1}}{\partial y_{n+1}}, \tag{I-4}$$

$$\frac{\partial Z_{n+1}}{\partial y_{n+1}} = \tfrac{2}{3}I - \frac{h}{6}\frac{\partial f}{\partial y_{n+1}}, \tag{I-5}$$

where I denotes the $n \times n$ identity matrix. Let $w = f[V]$ and denote the Jacobian of $f[V]$, evaluated at $V = V^i$, by $J(V^i)$, i.e.

$$J(V^i) = \left.\frac{\partial f}{\partial V}\right|_{V=V^i}.$$

Then write

$$\hat{J}_i = J(Z^i_{n+1}) = J(\tfrac{1}{3}y_n + \tfrac{2}{3}y^i_{n+1} - \tfrac{1}{6}h\dot{y}^i_{n+1}) = \left.\frac{\partial f}{\partial Z_{n+1}}\right|_{Z_{n+1}=Z_{n+1}{}^i},$$

$$J_i = J(y^i_{n+1}) = \left.\frac{\partial f}{\partial y_{n+1}}\right|_{y_{n+1}=y_{n+1}{}^i}.$$

From (I-4), (I-5) and above relations, the expression for B_i becomes

$$B_i = \left.\frac{\partial F}{\partial y_{n+1}}\right|_{y_{n+1}{}^i} = I - h\hat{J}_i[\tfrac{2}{3}I - \tfrac{1}{6}hJ_i] = I - \tfrac{2}{3}h\hat{J}_i[I - \tfrac{1}{4}hJ_i].$$

Acknowledgement. The author would like to thank the referee for his helpful comments and suggestions for improving this paper.

MBLE Research Laboratory
av. Van Becelaere 2
1170 Brussels, Belgium

1. G. G. Dahlquist, "A special stability problem for linear multi-step methods," *Nordisk Tidskr. Informationsbehandling,* v. 3, 1963, pp. 27–43. MR **30** #715.

2. C. W. Gear, *The Numerical Integration of Stiff Ordinary Differential Equations,* Report #221, Dept. Comp. Sci., University of Illinois, Urbana, Ill., January 1967.

3. Y. Genin, *A New Approach to the Synthesis of Stiffly Stable Linear Multistep Formulas,* Report #R188, MBLE Research Laboratory, Brussels, March 1972.

4. W. Liniger & R. A. Willoughby, "Efficient integration methods for stiff systems of ordinary differential equations," *SIAM J. Numer. Anal.*, v. 7, 1970, pp. 47–66. MR **41** #4809.

5. J. C. Butcher, "Implicit Runge-Kutta processes," *Math. Comp.*, v. 18, 1964, pp. 50–64. MR **28** #2641.

6. B. L. Ehle, "High order *A*-stable methods for the numerical solution of systems of D. E.'s," *Nordisk Tidskr. Informationsbehandling,* v. 8, 1968, pp. 276–278. MR **39** #1119.

7. E. Isaacson & H. B. Keller, *Analysis of Numerical Methods,* Wiley, New York, 1966. MR **34** #924.

8. C. G. Broyden, "A class of methods for solving nonlinear simultaneous equations," *Math. Comp.*, v. 19, 1965, pp. 557–593. MR **33** #6825.

9. C. G. Broyden, "A new method of solving nonlinear simultaneous equations," *Comput. J.*, v. 12, 1969, pp. 94–99. MR **39** #6509.