# Semi-Implicit Runge-Kutta Procedures with Error Estimates for the Numerical Integration of Stiff Systems of Ordinary Differential Equations

### J. R. CASH

Imperial College, London, England

ABSTRACT A-stable, semi-implicit Runge-Kutta procedures requiring at most one Jacobian evaluation per time step are developed for the approximate numerical integration of stiff systems of ordinary differential equations. A simple procedure for estimating the local truncation error is described and, with the help of this estimate, efficient integration procedures are derived. The algorithms are illustrated by direct application to a particular example

KEY WORDS AND PHRASES semi-implicit Runge-Kutta procedure, stiff differential equations, A-stability, Richardson extrapolation

CR CATEGORIES: 5 17

#### 1. Introduction

In this paper we are concerned with the approximate numerical integration of stiff systems of first-order ordinary differential equations of the form

$$\dot{x} = f(x), \quad x(t_0) = x_0$$
 (1.1)

by means of an R-stage, semi-implicit Runge-Kutta procedure of the form

$$x_{n+1} - x_n = h \sum_{r=1}^{R} w_r K_r,$$

$$K_1 = f(x_n) + \alpha_1 h A(x_n) K_1,$$

$$K_r = f(x_n + h \sum_{s=1}^{r-1} b_{rs} K_s) + \alpha_r h A(x_n + h \sum_{s=1}^{r-1} \beta_{rs} K_s) K_r, \quad r = 2, 3, \dots, R,$$

$$(1.2)$$

where  $A(x_n) = (\partial/\partial x) f(x_n)$ .

These procedures, which were originally suggested by Rosenbrock [1], allow A-stable algorithms suitable for the numerical integration of stiff systems of equations to be developed while still maintaining computational efficiency, i.e. avoiding the necessity to iterate It can be seen from the formulation of scheme (12) that each K, is given as the solution of a linear system of algebraic equations and so may be determined using any standard direct method, such as LU decomposition, for example. Several algorithms of the general form (1.2) have been proposed for the numerical integration of stiff systems of equations [2–4] but, apart from the algorithm due to Haines [4], attempts to obtain an error estimate applicable to systems of equations which may be used in a step control procedure are noticeable by their absence. In this paper we develop L-stable [5, p. 236] and A-stable semi-implicit Runge-Kutta formulas of orders 2 and 3, respectively, which

Copyright © 1976, Association for Computing Machinery, Inc General permission to republish, but not for profit, all or part of this material is granted provided that ACM's copyright notice is given and that reference is made to the publication, to its date of issue, and to the fact that reprinting privileges were granted by permission of the Association for Computing Machinery.

Author's address: Department of Mathematics, Imperial College of Science and Technology, Exhibition Road, London SW7 2RH, England

456 J. R. CASH

are formulated in such a way that an estimate of the local truncation error may be obtained at each step point with what is in general a negligible additional amount of computation. This estimate may be used to develop a simple step control procedure which allows the estimated local truncation error to be kept below a predetermined upper bound at each step point in the range of integration.

The algorithms developed in Sections 2 and 3 are demonstrated in Section 4 by application to a particular example

## 2. A Second-Order Scheme

In this section we consider the derivation of a second order, L-stable scheme of the form

$$x_{n+1} - x_n = hw_1K_1 + hw_2K_2,$$

$$K_1 = [I - ahA(x_n)]^{-1}f(x_n),$$

$$K_2 = [I - ahA(x_n)]^{-1}f(x_n + hb_1K_1),$$
(2.1)

where  $A(x_n) = (\partial/\partial x) f(x_n)$ , a,  $w_1$ ,  $w_2$ , and  $b_1$  are constants to be determined and h is the step length of integration with  $t_{n+1} = t_n + h$ . This scheme is of the general form (1.2) with  $\alpha_1 = \alpha_2 = a$  and  $\beta_{21} = 0$ ,  $b_{11} = b$ . It turns out in practice that when semi-implicit Runge-Kutta procedures are used to solve stiff systems of ordinary differential equations, the solution of the linear systems of algebraic equations, which implicitly define the  $K_1$ , forms a large proportion of the total computation cost. We note, however, that the quantities  $K_1$  and  $K_2$  appearing in scheme (2.1) have the same coefficient matrix, so only one matrix factorization is required at each time step. In order to simplify the notation, we shall derive a scheme for the case where the quantity f appearing in (1.1) is a scalar, but it may be verified that the scheme derived is immediately applicable to systems of equations. Any scheme of the form (2.1) is of second order if its coefficients satisfy the relations

$$w_1 + w_2 = 1, \quad \frac{1}{2} - a = w_2 b_1, \tag{2.2}$$

with the principal term in the local truncation error being

$$(a^{2} - a + \frac{1}{6}) \cdot (1/3!)h^{3}f(x_{n})A^{2}(x_{n}) + \frac{1}{2}h^{3}(\frac{1}{3} - b_{1}^{2}w_{2})f^{2}(x_{n})f_{xx}(x_{n}).$$
 (2.3)

Since relation (2.2) defines two equations in four unknowns, there will clearly be two "free" parameters in any solution of (2.2). For convenience we will denote the second-order formula of the form (2.1) with particular coefficients  $w_1$ ,  $w_2$ , a, and  $b_1$  used at the step point  $t_n$  to obtain a solution at  $t_{n+1}$  by  $R_2(t_n, w_1, w_2, a, b_1, h)$ . We shall regard the two parameters a and  $b_1$  as being free with the other two parameters being constrained to satisfy (2.2). We shall choose one of the free parameters so that the scheme  $R_2(t_n, w_1, w_2, a, b_1, h)$  is L-stable, and we shall choose the other free parameter at a later stage so that the resulting integration schemes are such that the procedure for estimating the local truncation error described later in this section is immediately applicable.

We first of all choose the coefficient a so that  $R_2(t_n, w_1, w_2, a, b_1, h)$  is L-stable. Applying (2.1) to the scalar test equation  $x = \lambda x$ , where  $\lambda$  is a complex constant with negative real part, we obtain the expression

$$x_{n+1} = \{[1 + (1-2a)q + (a^2 - 2a + \frac{1}{2})q^2]/(1-aq)^2\}x_n, q = \lambda h.$$
 (2.4)

Clearly  $R_2(t_n, w_1, w_2, a, b_1, h)$  is L-stable if the coefficient a satisfies the relation

$$a^2 - 2a + \frac{1}{2} = 0 (2.5)$$

and the rational approximation

$$R = [1 + (1 - 2a)q]/(1 - aq)^{2}$$

is A-acceptable [5, p. 237]. It is easy to verify that if  $a = 1 + 1/\sqrt{2}$  then the rational

approximation

$$R = \left[1 - (1 + \sqrt{2})q\right] / \left[1 - 2\left(1 + \frac{1}{\sqrt{2}}\right)q + \left(1 + \frac{1}{\sqrt{2}}\right)^2 q^2\right]$$

is indeed A-acceptable, showing the scheme  $R_2(t_n, w_1, w_2, 1 + 1/\sqrt{2}, b_1, h)$  to be L-stable.

We now describe the procedure which we shall use to obtain an estimate of the local truncation error as the integration proceeds. The procedure which we propose is a modified form of Richardson extrapolation whereby the solution at a particular point is found using two step lengths of size h and one step length of size 2h, and from the two solutions obtained at this point an estimate of the local truncation error may be determined. This procedure, although normally fairly accurate, usually requires about 50 percent extra computation in order to obtain the solution with step 2h. The crucial point which we wish to make in this paper is, however, that if we use the two integration procedures  $R_2(t_n, w_1, w_2, 1 + 1/\sqrt{2}, b_1, h)$  and  $R_2(t_{n-1}, \bar{w}_1, \bar{w}_2, \frac{1}{2} + 1/(2\sqrt{2}), b_1/2, 2h)$  then the quantities  $K_1$  and  $K_2$  are the same for both integration procedures, and so very little extra computation is required to calculate the solution with step length 2h. For the formula  $R_2(t_{n-1}, \bar{w}_1, \bar{w}_2, \frac{1}{2} + 1/(2\sqrt{2}), b_1/2, 2h)$  to be of the second order the coefficients must satisfy the two relations

$$\bar{w}_1 + \bar{w}_2 = 1, \quad 1 - (1 + 1/\sqrt{2}) = \bar{w}_2 b_1, \quad (2.6)$$

and it may easily be verified that the resulting scheme is A-stable and has a local truncation error with principal part given by

$$(2h)^{3} \{ (\frac{1}{4}a^{2} - \frac{1}{2}a + \frac{1}{6})[A^{2}(x_{n})f(x_{n})/3!] + (\frac{1}{3} - \frac{1}{4}b_{1}^{2}\bar{w}_{2}) \cdot \frac{1}{2}f^{2}(x_{n})(\partial^{2}/\partial x^{2})f(x_{n}) \}.$$
 (2.7)

Equations (2.6) and (2.2) now define four equations in five unknowns, so we still have one free parameter at our disposal. We shall choose this parameter so that expressions (2.3) and (2.7) may be used to provide an estimate of the local truncation error.

Suppose now that the free parameter is chosen so that

$$(\frac{1}{4}a^2 - \frac{1}{2}a + \frac{1}{6})/(\frac{1}{3} - \frac{1}{4}b_1^2\bar{w}_2) = (a^2 - a + \frac{1}{6})/(\frac{1}{3} - b_1^2w_2) \equiv \tau, \text{ say.} \quad (2.8)$$

It follows from relation (2.3) that

$$x(t_n) = R_2(t_n, w_1, w_2, a, b_1, h) + h^3(a^2 - a + \frac{1}{6})(a_{n-1} + a_{n-2}) + O(h^4)$$

and from relation (2.7) that

$$x(t_n) = R_2(t_{n-1}, \bar{w}_1, \bar{w}_2, \frac{1}{2}a, \frac{1}{2}b_1, 2h) + 8h^3(\frac{1}{4}a^2 - \frac{1}{2}a + \frac{1}{6})a_{n-2} + O(h^4),$$

where  $a_n = (1/3!)A^2(x_n)f(x_n) + (1/2\tau)f^2(x_n) (\partial^2/\partial x^2)f(x_n)$ . Assuming that  $a_n$  is effectively a constant, c, over the two steps  $[t_{n-1}, t_n]$  and  $[t_n, t_{n+1}]$ , simple manipulation shows that

$$R_2(t_n, w_1, w_2, a, b_1, h) - R_2(t_{n-1}, \bar{w}_1, \bar{w}_2, \frac{1}{2}a, \frac{1}{2}b_1, 2h) = 2h^3(-a + \frac{1}{2})c + \cdots$$

$$\epsilon_{n} \equiv x(t_{n}) - R_{2}(t_{n}, w_{1}, w_{2}, a, b_{1}, h) 
= 2h^{3}(a^{2} - a + \frac{1}{6})c + \cdots 
= \frac{a^{2} - a + \frac{1}{6}}{-a + \frac{1}{2}} \left\{ R_{2}(t_{n}, w_{1}, w_{2}, a, b_{1}, h) - R_{2}(t_{n-1}, \overline{w}_{1}, \overline{w}_{2}, a/2, b_{1}/2, 2h) \right\} + \cdots$$

Neglecting the terms of order  $O(h^4)$ , the quantity

$$\{(a^{2}-a+\frac{1}{6})/(-a+\frac{1}{2})\}\$$

$$\cdot\{R_{2}(t_{n}, w_{1}, w_{2}, a, b_{1}, h) - R_{2}(t_{n-1}, \bar{w}_{1}, \bar{w}_{2}, \frac{1}{2}a, \frac{1}{2}b_{1}, 2h)\}$$
 (2.9)

458 j. r. cash

now serves as a computable estimate for the local truncation error committed in stepping from  $t_n$  to  $t_{n+1}$ .

Equations (2.2), (2.6), and (2.8) now give five equations in five unknowns; these equations may be solved to give

$$b_1 = -2.306019375$$
,  $w_1 = 0.4765409197$ ,  $w_2 = 0.5234590803$ .  
 $\bar{w}_1 = 0.6933647701$ ,  $\bar{w}_2 = 0.3066352299$ ,

Our assumption that the quantity  $a_n$  is effectively a constant over the two steps  $[t_{n-1}, t_n]$  and  $[t_n, t_{n+1}]$  requires further comment at this stage. This assumption is clearly valid if h is sufficiently small compared with A(x) as is often the case in the initial (transient) phase. Since the problem is stiff, however, the local errors are damped, often severely so, and as a consequence an often more realistic assumption in the asymptotic (nontransient) phase is that only the local truncation error occurring at the second step  $[t_n, t_{n+1}]$  need be considered, and our estimate (2.9) for the local truncation error can be modified accordingly. It is worth noting, however, that the procedure developed earlier in this section, based on the assumption that  $a_n$  is a constant over two successive steps, has the virtue of being conservative in that the local truncation error is usually overestimated rather than underestimated.

### 3. Third-Order Equations

In this section we extend the procedures developed in Section 2 to third-order, three-stage semi-implicit Runge-Kutta procedures of the form

$$\begin{array}{ll} x_{n+1} - x_n &= h(w_1K_1 + w_2K_2 + w_3K_3), & \text{where} \\ K_1 &= f(x_n) + ahA(x_n)K_1, \\ K_2 &= f(x_n + hb_1K_1) + ahA(x_n)K_2, \\ K_3 &= f(x_n + hb_2K_1 + hb_3K_2) + ahA(x_n)K_3. \end{array} \tag{3.1}$$

$$w_{1} + w_{2} + w_{3} = 1,$$

$$\frac{1}{2} - a = w_{2}b_{1} + w_{3}(b_{2} + b_{3}),$$

$$\frac{1}{6} = \frac{1}{2}b_{1}^{2}w_{2} + \frac{1}{2}w_{3}(b_{2} + b_{3})^{2},$$

$$a^{2} - a + \frac{1}{6} = w_{3}b_{3}b_{1}$$
(3.2)

and

$$\bar{w}_{1} + \bar{w}_{2} + \bar{w}_{3} = 1, 
1 - a = \bar{w}_{2}b_{1} + \bar{w}_{3}(b_{2} + b_{3}), 
\frac{1}{6} = \frac{1}{8}b_{1}^{2}\bar{w}_{2} + \frac{1}{8}(b_{2} + b_{3})^{2}\bar{w}_{3}, 
\frac{1}{4}a^{2} - \frac{1}{2}a + \frac{1}{6} = \frac{1}{4}\bar{w}_{3}b_{3}b_{1}.$$
(3.3)

Equations (3.2) and (3.3) define eight equations in ten unknowns so, in general, we have two "free" parameters at our disposal. In order to use the step control procedure developed in Section 2, on the assumption that the local truncation error over two consecu-

tive steps is effectively a constant, we choose these two free parameters so that

$$\begin{array}{l} (a^3 - \frac{3}{2}a^2 + \frac{1}{2}a - \frac{1}{2}\frac{1}{2})/(\frac{1}{2}4 - \frac{1}{6}w_2b_1^3 - \frac{1}{6}w_3(b_2 + b_3)^3) \\ = (\frac{1}{8}a^3 - \frac{3}{8}a^2 + \frac{1}{4}a - \frac{1}{2}\frac{1}{2})/(\frac{1}{2}4 - \frac{1}{4}8\bar{w}_2b_1^3 - \frac{1}{4}8\bar{w}_8(b_2 + b_3)^3) \equiv \tau, \quad \text{say,} \\ \text{and } (-\frac{1}{2}a + \frac{1}{6} - w_3b_3b_1(b_1 + b_2 + b_3))/(a^3 - \frac{3}{2}a^2 + \frac{1}{2}a - \frac{1}{2}4) \\ = (-\frac{1}{4}a + \frac{1}{6} - \frac{1}{8}\bar{w}_3b_3b_1(b_1 + b_2 + b_3))/(\frac{1}{8}a^3 - \frac{3}{8}a^2 + \frac{1}{4}a - \frac{1}{2}4) \equiv \gamma, \quad \text{say.} \end{array}$$

Equations (3.2), (3.3), and (3.4) now give ten equations in ten unknowns; these may be solved to yield

$$a=0.8670738051,$$
  $w_1=0.9215174816,$   $w_2=0.1703752788,$   $w_3=-0.09189276043,$   $\bar{w}_1=0.1510038779,$   $\bar{w}_2=0.2847611470,$   $\bar{w}_3=0.5642349751,$   $b_1=-1.593640495,$   $b_2=0.6888190852,$   $b_3=0.3510545776.$  (3.5)

It may easily be verified that both of the integration procedures  $R_3(t_n, w_1, w_2, w_3, b_1, b_2, b_3, a, h)$  and  $R_3(t_{n-1}, \bar{w}_1, \bar{w}_2, \bar{w}_3, \frac{1}{2}b_1, \frac{1}{2}b_2, \frac{1}{2}b_3, \frac{1}{2}a, 2h)$  are A-stable, and a computable estimate,  $\epsilon_n$ , of the local truncation error when stepping from  $t_n$  to  $t_{n+1}$ , on the assumption that the local truncation error committed when stepping from  $t_{n-1}$  to  $t_n$  is effectively the same as that committed when stepping from  $t_n$  to  $t_{n+1}$ , is

$$\epsilon_n = \mu \{ R_3(t_{n-1}, \bar{w}_1, \bar{w}_2, \bar{w}_3, \frac{1}{2}b_1, \frac{1}{2}b_2, \frac{1}{2}b_3, \frac{1}{2}a, 2h) \\
- R_3(t_n, w_1, w_2, w_3, b_1, b_2, b_3, a, h) \} / (1 - \mu),$$

where

$$\mu = (-\frac{1}{2}a + \frac{1}{6} - w_3b_3b_1(b_1 + b_2 + b_3))/(8[-\frac{1}{4}a + \frac{1}{6} - \frac{1}{8}\bar{w}_3b_3b_1(b_1 + b_2 + b_3)]).$$

#### 4. Numerical Results

In this section we consider the numerical integration of the system

$$x_1 = 0.01 - (x_1^2 + 1001x_1 + 1001)(0.01 + x_1 + x_2), \quad x_1(0) = 0,$$
  
 $\dot{x}_2 = 0.01 - (1 + x_2^2)(0.01 + x_1 + x_2), \quad x_2(0) = 0,$  (4.1)

in the range  $0 \le t \le 100$ . This system, originally considered by Liniger and Willoughby [6], has been suggested as a test problem by Bjurel et al. [7]. In Table I we list some of the results obtained for the integration of (4.1) using the integration procedure  $R_2(t_n, w_1, w_2, a, b, h)$  with the procedure  $R_2(t_{n-1}, \bar{w}_1, \bar{w}_2, \frac{1}{2}a, \frac{1}{2}b, 2h)$  also being used to provide an estimate of the local truncation error. Initially the step length of integration chosen was .1\*10<sup>-5</sup> and the strategy adopted was to double the step length if  $\max_i |\epsilon_{i,n}| < 10^{-10}$ , where  $\epsilon_n = \{^{\epsilon_{1,n}}_{i,2,n}\}$ , and to halve it if  $\max_i |\epsilon_{i,n}| > 10^{-9}$ . As can be seen from Table I the procedure developed in Section 2 provides a conservative estimate of the local truncation error and allows us to develop a simple step control procedure based

Estimated ŧ Solution obtained True local error Step size local error 10-6 2\*10-6  $.2748*10^{-10}$  $.2749*10^{-10}$  $x_1 = -0.1997976622*10^{-4}$ 0 2001417704\*10-10 .2766\*10-13 .2768\*10-18 .2185\*10-9 6\*10-6 .2183\*10-9  $x_1 = -0.5981814751*10^{-4}$ 2\*10-6  $.2200*10^{-12}$  $0.1798835197*10^{-9}$  $2197*10^{-12}$ 10-6  $x_1 = -0.9949576697*10^{-4}$ .2174\*10-9 .2176\*10-9 2\*10-6  $.2188*10^{-12}$  $0.4987827785*10^{-9}$ .2191\*10-13 0183358  $x_1 = -0.1006189858*10^{-1}$ .2962\*10-9 .3052\*10~9 29\*10-6  $0.8301827931*10^{-4}$ 3036\*10-12 .3082\*10-12 14.293203  $x_1 = -0.8756200958*10^{-1}$ .3001\*10-9 .6350\*10~9 221\*10-6  $0.7758388251*10^{-1}$ .2246\*10-9 .2380\*10-9

TABLE I

TABLE II

t	Solution obtained	True local error	Estimated local error	Step size
2*10-5	$x_1 = -0 \ 1979918305*10^{-3}$	.153*10-10	.167*10-10	10-5
	$x_2 = 0.1986559395*10^{-8}$	$.164*10^{-13}$	$154*10^{-13}$	
6*10-5	$x_i = -0.5821716667*10^{-3}$	$.240*10^{-9}$	.253*10-9	2*10-5
	$x_2 = 0.1764097724*10^{-7}$	$.250*10^{-12}$	$234*10^{-12}$	
10*10-5	$x_1 = -0.9511431031*10^{-3}$	231*10-9	.242*10-9	$2*10^{-5}$
	$x_2 = 0.4835541392*10^{-7}$	$240*10^{-12}$	$225*10^{-12}$	
14*10-5	$x_1 = -0.1305519277*10^{-2}$	222*10-9	232*10-9	$2*10^{-5}$
	$x_2 = 0.9353329237*10^{-7}$	$.230*10^{-12}$	$.216*10^{-12}$	
39 332	$x_1 = -0.4348338434$	$.111*10^{-7}$	$.391*10^{-7}$	1.6384
	$x_2 = 0.4248691006$	$.241*10^{-8}$	$.263*10^{-8}$	

on this estimate. The results given in Table I are for the first, second, third, 1300th, and 1400th steps, and as can be seen the step size is increased as the steady state solution is reached. In Table II we list the results obtained for the integration of (4.1) using the integration procedure  $R_3(t_n, w_1, w_2, w_3, b_1, b_2, b_3, a, h)$  with the procedure  $R_3(t_{n-1}, \bar{w}_1, \bar{w}_2, \bar{w}_3, \frac{1}{2}b_1, \frac{1}{2}b_2, \frac{1}{2}b_3, \frac{1}{2}a, 2h)$  also being used to provide an estimate for the local truncation error. The initial step length chosen was  $.1*10^{-4}$ , and initially the step length was doubled if  $\max_{\cdot} |\epsilon_{1,n}| < \frac{1}{2}10^{-10}$  and halved if  $\max_{\cdot} |\epsilon_{1,n}| > 10^{-9}$ . After fifty steps, when a relatively steady state had been reached, the error tolerance was raised to  $10^{-8}$  and the step length was doubled if  $\max_{\cdot} |\epsilon_{1,n}| < 10^{-8}$  and halved if  $\max_{\cdot} |\epsilon_{1,n}| > 10^{-7}$ . The results given in Table II are for the first four and the fiftieth steps; again it can be seen that the step size is increased as the integration proceeds. We note with this example that the error in  $x_2$  is underestimated in the transient phase and overestimated in the steady state phase whereas the error in  $x_1$ , the estimate of which determines the step size initially, is overestimated throughout.

ACKNOWLEDGMENT. The author is very grateful to the referees for many helpful suggestions and comments.

## REFERENCES

- 1 ROSENBROCK, H H. Some general implicit processes for the numerical solution of differential equations. Comput J 5 (1963), 329-330
- 2 CALAHAN, D. Numerical solution of linear systems with widely separated time constants Proc. IEEE 55 (Nov 1967), 2016-2017.
- 3. ALLEN, R H, AND POTTLE, C Stable integration methods for electronic circuit analysis with widely separated time constants Proc. Sixth Annual Allerton Conf. on Circuit and System Theory, T. Trick and R T. Chien, Eds., pp. 311-320
- HAINES, C Implicit integration processes with error estimates for the numerical solution of differential equations. Comput. J. 12 (1969), 183-187.
- LAMBERT, J.D. Computational Methods in Ordinary Differential Equations Wiley, New York, 1973.
- LINIGER, W, AND WILLOUGHBY, R. Efficient numerical integration of stiff systems of ordinary differential equations. Res. Rep. RC 1970, IBM Thomas J. Watson Research Center, Yorktown Heights, NY, 1967.
- 7 BJUREL, G, DAHLQUIST, G, LINDBERG, B, LINDE, S, AND ODEN, L. Survey of stiff ordinary differential equations Comput. Sci. Rep. NA 7011, Dep. of Information Processing, Royal Inst. of Technology, Stockholm, Sweden, 1970.

RECEIVED SEPTEMBER 1974; REVISED SEPTEMBER 1975