SIAM J. NUMER. ANAL.
Vol. 14, No. 6, December 1977

DIAGONALLY IMPLICIT RUNGE-KUTTA METHODS FOR STIFF O.D.E.'S*

ROGER ALEXANDER†

Abstract. To be A-stable, and possibly useful for stiff systems, a Runge–Kutta formula must be implicit. There is a significant computational advantage in *diagonally implicit* formulae, whose coefficient matrix is lower triangular with all diagonal elements equal. We derive new, strongly S-stable diagonally implicit Runge–Kutta formulae of order 2 in 2 stages and of order 3 in 3 stages, and show that it is impossible for a strongly S-stable diagonally implicit method to attain order 4 in 4 stages. Merely A-stable diagonally implicit formulae, of order 3 in 2 stages and of order 4 in 3 stages, were previously known; we prove that no 4-stage method of this type has order 5. We describe a computer program for stiff differential equations which uses these methods, and compare them to each other and to the GEAR package.

"Dirk, a kind of Dagger used in the Highlands of Scotland."
. . . Samuel Johnson, Dictionary, 1755.

1. Introduction. Our task is to approximate the solution $y: [0, T] \to \mathbb{R}^m$ of an initial value problem

$$y' = f(t, y), \qquad y(0) = y_0 \tag{1.1}$$

in which $f: [0, T] \times \mathbb{R}^m \to \mathbb{R}^m$ is a sufficiently regular function. The idea of the Runge–Kutta methods is to get from t_n to $t_{n+1} = t_n + h$ (h is the current stepsize) by approximating the integral in

$$y(t_{n+1}) = y(t_n) + \int_{t_n}^{t_{n+1}} f(t, y(t))\, dt. \tag{1.2}$$

We choose quadrature points $\tau_1, \cdots, \tau_q$ and their weights $b_1, \cdots, b_q$; we intend to use the quadrature formula:

$$y(t_{n+1}) = y(t_n) + h \sum_{i=1}^{q} b_i f(t_n + \tau_i h, y(t_n + \tau_i h)) + \text{Error}. \tag{1.3}$$

Hereafter write $t_{n,i}$ for $t_n + \tau_i h$. Suppose we have an approximation y_n to $y(t_n)$; to use (1.3) we also need values $y_{n,i}$ to put in for $y(t_{n,i})$. Compute them also by numerical quadratures on the same nodes:

$$y_{n,i} = y_n + h \sum_{j=1}^{q} a_{ij} f(t_{n,j}, y_{n,j}) \qquad i = 1(1)q. \tag{1.4}$$

In general this is a set of implicit equations, which we solve and use in (1.3) for our next value of y:

$$y_{n+1} = y_n + h \sum_{i=1}^{q} b_i f(t_{n,i}, y_{n,i}). \tag{1.5}$$

* Received by the editors August 12, 1976, and in revised form January 17, 1977.

† Department of Mathematics, University of Colorado, Boulder, Colorado. Now at Department of Mathematical Sciences, Rensselaer Polytechnic Institute, Troy, New York 12181. This work was supported in part by the National Science Foundation under Grant MCS 76-06917. Computing time was provided in part by the Computing Center, University of Colorado.

Formulae (1.4) and (1.5) together define a Runge–Kutta (RK) method, which we designate by displaying its coefficients:

$$\begin{array}{ccc|c} a_{11} & \cdots & a_{1q} & \tau_1 \\ \vdots & \cdots & \vdots & \vdots \\ a_{q1} & \cdots & a_{qq} & \tau_q \\ \hline b_1 & \cdots & b_q & \end{array} \tag{1.6}$$

The traditional problem of choosing the q^2+2q coefficients in a q-stage method so as to obtain the highest possible order of accuracy, subject to stability or other constraints, leads to a nonlinear algebraic jungle, to which civilization and order were brought in the pioneering work of J. C. Butcher [2], [5], further refined in the thesis of M. Crouzeix [6]. Part of the purpose of this paper is to make their approach and their techniques better known.

The classical formulae of Runge and Kutta were *explicit*, that is, in (1.6) one had $a_{ij}=0$ for $i \leqq j$ so that in (1.5) $y_{n,i}$ is given explicitly in terms of the preceding $y_{n,j}$. The implicit RK methods introduced by Butcher [3], [4] became interesting for stiff problems when Ehle [7] showed the q-stage methods of order $2q$ to be A-stable. Unfortunately, to integrate a system of m differential equations, an implicit method with a full matrix (1.6) requires the solution of mq simultaneous implicit (in general nonlinear) equations at each time step. This is one disadvantage that contributed to the inferior performance of a program based on the 2-stage 4th-order method tested by Enright, Hull and Lindberg [8].

One way to circumvent this difficulty is to use a lower triangular matrix (a_{ij}) in (1.6): the equations (1.4) may then be solved in q successive stages, with only an m-dimensional system to be solved at each stage. Following Butcher, we call such a method *semi-implicit.*

There have been several investigations of semi-implicit RK methods [1], [12], [11], [13], [6]. The present work starts with the following idea: in solving (1.4) successively by Newton-type iterations one solves linear systems at each stage with a coefficient matrix of the form

$$I - ha_{ii}\, \partial f/\partial y.$$

If all a_{ii} are equal one may hope to use repeatedly the stored LU-factorization of a single such matrix. M. Crouzeix pointed out the usefulness of this idea for solving the linear differential equations arising from discretization of linear parabolic partial differential equations by the finite element method [6]. S. P. Nørsett made this idea the basis of his study of semi-implicit methods [13]. When all the a_{ii} are equal in a semi-implicit formula, we shall call it a diagonally implicit Runge–Kutta (DIRK) formula.

In the next section we survey the A-stable DIRK methods of maximum order in two and three stages, and prove that no four-stage DIRK formula has order five. In § 3 we derive new methods with stronger stability properties, and in § 4 we describe a comparison of several of these methods with each other and with the GEAR package.

2. A-stable methods. Referring to the general presentation of a Runge–Kutta formula (1.6), we make the following conventions:

q denotes the number of stages in a method,

p denotes the order of the method;

A is the $q \times q$ matrix (a_{ij});

T is the $q \times q$ matrix diag $(\tau_1, \cdots, \tau_q)$;

b is the q-dimensional vector (b_i);

e is the q-dimensional vector having all components equal to 1.

The symbol * denotes the transpose of a matrix.

It is easy to see that there is a unique $(q, p) = (1, 2)$ DIRK formula, the implicit midpoint rule, which is well known to be A-stable:

$$\begin{array}{c|c} \frac{1}{2} & \frac{1}{2} \\ \hline & 1 \end{array}. \tag{2.1}$$

Crouzeix [6] has determined all the 2-stage, third-order and 3-stage, fourth-order semi-implicit RK methods; from his work one can extract the following theorem.

THEOREM 1 (Crouzeix [6, Propositions 1.6, 1.8]). *For* $(q, p) = (2, 3)$ *and* $(q, p) = (3, 4)$ *there is exactly one A-stable DIRK formula. These are given by*

(2.2) $(q, p) = (2, 3)$,

$$\begin{array}{cc|c} \frac{1}{2}+\frac{1}{2\sqrt{3}} & 0 & \frac{1}{2}+\frac{1}{2\sqrt{3}} \\ -\frac{1}{\sqrt{3}} & \frac{1}{2}+\frac{1}{2\sqrt{3}} & \frac{1}{2}-\frac{1}{2\sqrt{3}} \\ \hline \frac{1}{2} & \frac{1}{2} & \end{array}$$

and

(2.3) $(q, p) = (3, 4)$, $\alpha = 2\cos(\pi/18)/\sqrt{3}$

$$\begin{array}{ccc|c} (1+\alpha)/2 & 0 & 0 & (1+\alpha)/2 \\ -\alpha/2 & (1+\alpha)/2 & 0 & 1/2 \\ 1+\alpha & -(1+2\alpha) & (1+\alpha)/2 & (1-\alpha)/2 \\ \hline 1/(6\alpha^2) & 1-1/(3\alpha^2) & 1/(6\alpha^2) & \end{array}$$

The formula (2.3) shows the strain of having A-stability plus fourth-order squeezed out of its three stages: the diagonal $(1+\alpha)/2 \sim 1.06$ so that at each stage an implicit equation must be solved over an interval longer than the stepsize, and the weights in the third row are roughly 2, -3 and 1, which can cause roundoff problems. Thus the following result should not be a surprise.

THEOREM 2. *There is no DIRK formula with* $(q, p) = (4, 5)$.

This result verifies part of a conjecture of [13]. Before giving the proof, we recall some of the facts about the order of RK methods and prove a lemma.

THEOREM 3 (Crouzeix [6, Thm. 1.6]). *Let* $p \leqq 5$. *In order that the RK method* (1.6) *be of order p for every sufficiently regular function* $f(t, y)$ *in* (1.1), *it is* necessary

that the relations (2.4.*i*), $i = 1(1)p$ *be satisfied.*

(2.4.1) $b^*e = 1,$

(2.4.2) $b^*Te = 1/2, \quad b^*Ae = 1/2,$

(2.4.3) $b^*T^2e = 1/3, \quad b^*TAe = 1/3, \quad b^*ATe = 1/6, \quad b^*A^2e = 1/6,$

(2.4.4)
$$b^*T^3e = 1/4, \quad b^*TATe = 1/8, \quad b^*AT^2e = 1/12, \quad b^*A^2Te = 1/24,$$
$$b^*T^2Ae = 1/4, \quad b^*TA^2e = 1/8, \quad b^*ATAe = 1/12, \quad b^*A^3e = 1/24,$$

(2.4.5)
$$b^*T^4e = 1/5, \quad b^*TAT^2e = 1/15, \quad b^*TA^2Te = 1/30,$$
$$b^*A^2T^2e = 1/60,$$
$$b^*T^3Ae = 1/5, \quad b^*TATAe = 1/15, \quad b^*TA^3e = 1/30,$$
$$b^*A^2TAe = 1/60,$$
$$b^*T^2ATe = 1/10, \quad b^*AT^3e = 1/20, \quad b^*ATATe = 1/40,$$
$$b^*A^3Te = 1/120,$$
$$b^*T^2A^2e = 1/10, \quad b^*AT^2Ae = 1/20, \quad b^*ATA^2e = 1/40,$$
$$b^*A^4e = 1/120.$$

LEMMA 1. *Let* (1.6) *be a semi-implicit RK formula with* $(q, p) = (4, 5)$. *Assume* $\delta = ATe - \frac{1}{2}T^2e \neq 0$. *Then*

$$A^*b = (I - T)b \tag{2.5}$$

and

$$A^*Tb = \tfrac{1}{2}(I - T^2)b \tag{2.6}$$

Remark. The "integration by parts" identities (2.5) and (2.6) hold in general if $q < p$ and all τ_i are distinct [6].

Proof. We first prove (2.5). Since $p = 5$, (2.4.1)–(2.4.5) show that b, Tb, T^2b and A^*b are orthogonal to $\delta \neq 0$, hence linearly dependent ($q = 4$!). So there are constants $\lambda_1, \lambda_2, \lambda_3, \lambda_4$ not all zero such that

$$\lambda_1 b + \lambda_2 Tb + \lambda_3 T^2 b + \lambda_4 A^* b = 0. \tag{2.7}$$

In this equation, take scalar products successively with e, Te, and T^2e, and use (2.4.1)–(2.4.5) to derive

$$\begin{aligned} \lambda_1 + \tfrac{1}{2}\lambda_2 + \tfrac{1}{3}\lambda_3 + \tfrac{1}{2}\lambda_4 &= 0, \\ \tfrac{1}{2}\lambda_1 + \tfrac{1}{3}\lambda_2 + \tfrac{1}{4}\lambda_3 + \tfrac{1}{6}\lambda_4 &= 0, \\ \tfrac{1}{3}\lambda_1 + \tfrac{1}{4}\lambda_2 + \tfrac{1}{5}\lambda_3 + \tfrac{1}{12}\lambda_4 &= 0. \end{aligned} \tag{2.8}$$

It is easy to show that $\lambda_1 = 0 \Rightarrow \lambda_2 = \lambda_3 = \lambda_4 = 0$. Since nonzero scalar factors do not matter in (2.7), set $\lambda_1 := 1$ and solve (2.8) for $\lambda_2, \lambda_3, \lambda_4$. This gives (2.5).

The proof of (2.6) is similar. By (2.4.1)–(2.4.5), b, TA^*b, A^*Tb and $A^{*2}b$ are orthogonal to the nonzero vector δ, so there are constants $\lambda_1, \lambda_2, \lambda_3, \lambda_4$ not all zero such that

$$\lambda_1 b + \lambda_2 TA^*b + \lambda_3 A^*Tb + \lambda_4 A^{*2}b = 0.$$

Take scalar products of this equation with e, Ae and A^2e and use (2.4.1)–(2.4.5) to derive

$$\lambda_1 + \tfrac{1}{6}\lambda_2 + \tfrac{1}{3}\lambda_3 + \tfrac{1}{6}\lambda_4 = 0,$$

$$\tfrac{1}{2}\lambda_1 + \tfrac{1}{12}\lambda_2 + \tfrac{1}{8}\lambda_3 + \tfrac{1}{24}\lambda_4 = 0,$$

$$\tfrac{1}{6}\lambda_1 + \tfrac{1}{40}\lambda_2 + \tfrac{1}{30}\lambda_3 + \tfrac{1}{120}\lambda_4 = 0.$$

This time we must have $\lambda_4 \neq 0$. Setting $\lambda_4 := 1$ and solving for $\lambda_1, \lambda_2, \lambda_3$ gives (2.6).

Proof of Theorem 2. We must have $\delta = ATe - \frac{1}{2}T^2e \neq 0$, else $a_{11} = 0$ so that the diagonal of A is zero and the method is explicit, which is known to be impossible. Thus if (1.6) be a DIRK formula having $(q, p) = (4, 5)$, the hypothesis of the lemma is satisfied. Look at the fourth components in (2.5). They are

$$a_{44}b_4 = (1 - \tau_4)b_4.$$

Now $b_4 \neq 0$ or our formula would actually have only 3 stages, and the reader may easily check that that is impossible.[1] Thus

$$a_{44} = 1 - \tau_4. \tag{2.9}$$

Now the equality of fourth components in (2.6) gives

$$a_{44}\tau_4 b_4 = \tfrac{1}{2}(1 - \tau_4^2)b_4.$$

Use $b_4 \neq 0$ and (2.9) to derive from this

$$(1 - \tau_4)\tau_4 = \tfrac{1}{2} - \tfrac{1}{2}\tau_4^2 \quad \text{or} \quad \tfrac{1}{2} - \tau_4 + \tfrac{1}{2}\tau_4^2 = 0$$

whence $\tau_4 = 1$. But then (2.9) shows $a_{44} = 0$, hence the diagonal of A is zero, and this is as impossible as before.

3. *S*-stability. A-stability is not the whole answer to the problem of stiff equations. In their work with large systems of stiff nonlinear equations Prothero and Robinson [14] found that A-stability of a method was no guarantee that it would give stable solutions, and that the accuracy of the solutions obtained often appeared to be unrelated to the order of the method used. Their analysis led to the introduction of a new stability concept.

DEFINITION. [14] A RK-method is *S-stable* if for any bounded function $g: [0, T] \to \mathbb{R}$ having a bounded derivative, and any positive constant λ_0, there is a positive constant h_0 such that the numerical solution (y_n) to the equation

$$y' = g'(t) + \lambda(y - g(t))$$

satisfies

$$\left| \frac{y_{n+1} - g(t_{n+1})}{y_n - g(t_n)} \right| < 1$$

[1] A 5th-order method is also 4th-order. But all DIRK formulae with $(q, p) = (3, 4)$ are known [6], and none is of order five.

provided $y_n \neq g(t_n)$, for all $0 < h < h_0$ and all complex λ with $\mathrm{Re}\,(-\lambda) \geqq \lambda_0$.

A RK method is *strongly S-stable* if

$$\frac{y_{n+1} - g(t_{n+1})}{y_n - g(t_n)} \to 0$$

as $\mathrm{Re}\,(-\lambda) \to \infty$ for all $h > 0$ such that $[t_n, t_{n+1}] \subset [0, T]$.

Notice that an S-stable method is A-stable (take $g \equiv 0$). The converse does not hold. Before we analyze the S-stability of our methods, we need some tools. For background see [14], [15].

To each RK method we associate the rational function which arises when the method is applied to the scalar test equation

$$y' = \lambda y, \qquad y(0) = y_0$$

with stepsize h. One computes $y_1 = R(h\lambda)y_0$, and in fact [15]

$$R(h\lambda) = 1 + h\lambda b^*(I - h\lambda A)^{-1}e.$$

Recall that a RK formula is A-stable if $|R(h\lambda)| < 1$ for $\mathrm{Re}\,(h\lambda) < 0$. We call a formula *stiffly A-stable* if it is A-stable and $\lim_{h\lambda \to \infty} R(h\lambda) = 0$. (Stiff A-stability has been called L-stability by B. L. Ehle, and strong A-stability by other writers.)

A semi-implicit formula with nonzero diagonal has an invertible matrix A; its $R(h\lambda)$ is holomorphic at infinity and we may put

$$\alpha_0 \equiv \lim_{h\lambda \to \infty} R(h\lambda) = 1 - b^*A^{-1}e.$$

We shall say that a q-stage semi-implicit RK method is *stiffly accurate* when $\tau_q = 1$ and $a_{qi} = b_i$, $i = 1(1)q$.

LEMMA 2. *A RK formula with invertible matrix A satisfying* $a_{qi} = b_i$, $i = 1(1)q$ *has* $\alpha_0 = 0$. (*In particular, if the formula is A-stable and stiffly accurate, it is stiffly A-stable*).

Proof. The result is immediate because $b^*A^{-1} = (0, \cdots, 0, 1)$, that is, the last row of A times A^{-1} gives the last row of the identity matrix.

We can now adapt the result of Prothero and Robinson to our situation.

THEOREM 4. [14, Thms. 2.1, 2.2]. *An A-stable semi-implicit RK formula with positive diagonal elements is S-stable if and only if* $|\alpha_0| < 1$. *An S-stable formula of this kind is strongly S-stable if and only if it is stiffly accurate.*

We apply these criteria to the methods derived by Crouzeix.

COROLLARY. *The A-stable methods with* $(q, p) = (2, 3)$ [6, *Prop.* 1.6] *are S-stable. The A-stable methods having* $(q, p) = (3, 4)$ [6, *Prop.* 1.8] *are S-stable except for the two methods for which* $\alpha_0 = 1$. *None of these methods is strongly S-stable.*

This shows, then, that to achieve stiff accuracy (required for strong S-stability) one must give up an order of ordinary accuracy. Strongly S-stable methods of orders three and four having $\tau_1 = 0$ were derived by R. Alt [1, summarized without proof in [6]]. K. Miller [12] and M. A. Kurdi [11] found strongly S-stable semi-implicit methods of order two, three and four (they called their methods "diagonally implicit," a term which is reserved here for the special case where all diagonal entries are equal in A). We now show how strong S-stability can be achieved with a DIRK formula.

THEOREM 5. *There are exactly two strongly S-stable DIRK formulae of order two in two stages, and exactly one strongly S-stable DIRK formula of order three in three stages. They are*

$$\begin{array}{cc|c} \alpha & 0 & \alpha \\ 1-\alpha & \alpha & 1 \\ \hline 1-\alpha & \alpha & \end{array} \qquad \text{with } \alpha = 1 \pm \tfrac{1}{2}\sqrt{2},$$

$$\begin{array}{ccc|c} \alpha & 0 & 0 & \alpha \\ \tau_2-\alpha & \alpha & 0 & \tau_2 \\ b_1 & b_2 & \alpha & 1 \\ \hline b_1 & b_2 & \alpha & \end{array} \qquad \alpha \text{ is the root of } x^3-3x^2+\tfrac{3}{2}x-\tfrac{1}{6}=0 \text{ lying in } (\tfrac{1}{6},\tfrac{1}{2}),$$

$$\tau_2 = (1+\alpha)/2,$$

$$b_1 = -(6\alpha^2-16\alpha+1)/4,$$

$$b_2 = (6\alpha^2-20\alpha+5)/4.$$

Proof. One verifies directly that these schemes are of order two and three respectively ([6] or [2]). Next, observe that in both cases $R(h\lambda)$ has all its poles at $\alpha > 0$; finally, simple algebra shows that $|R(iy)|^2 \leqq 1$ for real y so that A-stability follows from the maximum principle. We now show that these schemes are the only possible ones.

(a) A strongly S-stable DIRK formula with $(q, p) = (2, 2)$ necessarily has the form

$$\begin{array}{cc|c} \alpha & 0 & \tau \\ 1-\alpha & \alpha & 1 \\ \hline 1-\alpha & \alpha & \end{array}$$

Thus it is enough to show that $\tau = \alpha$ and that α must take one of the values specified.

Notice that α cannot be equal to 1. From the necessary conditions for $p = 2$ (2.4.2) we obtain

$$(1-\alpha)\cdot\alpha + \alpha\cdot 1 = \tfrac{1}{2} = (1-\alpha)\cdot\tau + \alpha\cdot 1$$

from which $\tau = \alpha = \frac{1}{2} - \alpha + \alpha^2$ and the result follows.

(b) A strongly S-stable DIRK formula with $(q, p) = (3, 3)$ necessarily has the form

$$\begin{array}{ccc|c} \alpha & 0 & 0 & \tau_1 \\ \beta & \alpha & 0 & \tau_2 \\ b_1 & b_2 & \alpha & 1 \\ \hline b_1 & b_2 & \alpha & \end{array} \tag{3.1}$$

We first show that α must be as specified. If the formula (3.1) is strongly S-stable and of order three, then

$$R(h\lambda)=(1+c_1h\lambda+c_2(h\lambda)^2)/(1-\alpha h\lambda)^3$$

which is $e^{h\lambda}+O(h\lambda)^4$ for small $h\lambda$ provided α is a root of

$$x^3-3x^2+\tfrac{3}{2}x-\tfrac{1}{6}=0, \tag{3.2}$$

and then

$$c_1=1-3\alpha,$$

$$c_2=\tfrac{1}{2}-3\alpha+3\alpha^2.$$

An algebraic computation shows that $|R(iy)|^2\leqq 1$ for all real y only when α is the root of (3.2) lying between 1/6 and 1/2.

Next we show that $Ae=Te$. If $Ae\neq Te$, then the necessary conditions (2.4.1)–(2.4.3) for $p=3$ show that b, Tb and A^*b lie in the two-dimensional subspace orthogonal to $Ae-Te$; hence there are numbers λ_1, λ_2, λ_3 not all zero such that

$$\lambda_1 b+\lambda_2 A^*b+\lambda_3 Tb=0. \tag{3.3}$$

Take scalar products with e and Ae in this equation and use (2.3.1)–(2.3.3) to derive

$$\lambda_1+\tfrac{1}{2}\lambda_2+\tfrac{1}{2}\lambda_3=0,$$

$$\tfrac{1}{2}\lambda_1+\tfrac{1}{6}\lambda_2+\tfrac{1}{3}\lambda_3=0.$$

Now $\lambda_1=0\Rightarrow\lambda_2=\lambda_3=0$. Put $\lambda_1:=1$ and solve; (3.3) becomes

$$A^*b=(I-T)b,$$

and the third component of this equation is

$$\alpha^2=0,$$

a contradiction.

This shows that the scheme (3.1) actually has the form given in the statement of the theorem; τ_2 is now determined by the requirement that the weight corresponding to the point 1 in the quadrature formula of the last line shall be α. The proof is now complete.

We conclude this section with an anti-result which shows that at least five stages are needed to achieve order four in a strongly S-stable DIRK formula.

THEOREM 6. *There is no strongly S-stable DIRK formula of order four in four stages.*

Proof. Such a formula would have to have the following form.

$$\begin{array}{cccc|c} \alpha & & & & \tau_1 \\ \beta & \alpha & & & \tau_2 \\ \gamma & \delta & \alpha & & \tau_3 \\ b_1 & b_2 & b_3 & \alpha & 1 \\ \hline b_1 & b_2 & b_3 & \alpha & \end{array} \tag{3.4}$$

As before, we first determine α. Since the formula is to be stiffly A-stable and of

order 4, we must have

$$R(h)=(1+c_1h+c_2h^2+c_3h^3)/(1-\alpha h)^4=e^h+O(h^5)$$

for small h. Hence α must be a root of

$$x^4-4x^3+3x^2-\tfrac{2}{3}x+\tfrac{1}{24}=0 \tag{3.5}$$

and we derive

$$c_1=1-4\alpha,$$
$$c_2=\tfrac{1}{2}-4\alpha+6\alpha^2,$$
$$c_3=\tfrac{1}{6}-2\alpha+6\alpha^2-4\alpha^3.$$

An algebraic computation shows that $|R(iy)|^2\leqq 1$ for all real y only if α is the root of (3.5) lying between 1/2 and 3/5.

Now $(A-\alpha I)^4=0$, and we use the requirement $p=4$ to obtain from (2.4.1)–(2.4.4),

$$\mu\equiv b^*A^4e=b^*(4\alpha A^3-6\alpha^2A^2+4\alpha^3A-\alpha^4)e=\tfrac{1}{24}-\tfrac{1}{2}\alpha+2\alpha^2-2\alpha^3,$$

where we have used the fact that α is a root of (3.5) to eliminate the fourth power term. Now $\mu\neq 0$. The best way to see this is to multiply (3.5) by 24 and check that the result does not factor over the integers; hence (3.5) is irreducible over the rationals and the only cubic polynomial in α with rational coefficients which vanishes is the zero polynomial.

We can now show that $Ae=Te$ is necessary, by the standard argument. If $\varepsilon=AE-Te\neq 0$, then (2.4.1)–(2.4.4) implies that the vectors b, A^*b, Tb, $A^{*2}b$ lie in the 3-dimensional subspace orthogonal to ε. Thus there are numbers $\lambda_1,\lambda_2,\lambda_3,\lambda_4$ not all zero such that

$$\lambda_1b+\lambda_2A^*b+\lambda_3Tb+\lambda_4A^{*2}b=0.$$

Take scalar products in this equation with e, Ae, A^2e to obtain

$$\begin{aligned}
\lambda_1+\tfrac{1}{2}\lambda_2+\tfrac{1}{2}\lambda_3+\tfrac{1}{6}\lambda_4&=0,\\
\tfrac{1}{2}\lambda_1+\tfrac{1}{6}\lambda_2+\tfrac{1}{3}\lambda_3+\tfrac{1}{24}\lambda_4&=0,\\
\tfrac{1}{6}\lambda_1+\tfrac{1}{24}\lambda_2+\tfrac{1}{8}\lambda_3+\mu\lambda_4&=0.
\end{aligned} \tag{3.6}$$

Now if $\lambda_1=0$, $\lambda_2,\lambda_3,\lambda_4$ must also be zero (this uses the fact that $288\mu-2\neq 0$). Set $\lambda_1:=1$ and solve (3.6) for $\lambda_2,\lambda_3,\lambda_4$ to obtain

$$A^*b=(I-T)b$$

the fourth component of which reads $\alpha^2=0$, a contradiction.

This establishes the necessity of $Ae=Te$. The putative fourth-order formula (3.4) must therefore look like

$$\begin{array}{cccc|c}
\alpha & & & & \alpha\\
\tau_2-\alpha & \alpha & & & \tau_2\\
\tau_3-\beta-\alpha & \beta & \alpha & & \tau_3\\
b_1 & b_2 & b_3 & \alpha & 1\\
\hline
b_1 & b_2 & b_3 & \alpha &
\end{array} \tag{3.7}$$

Here the nodes $(\alpha, \tau_2, \tau_3, 1)$ and their corresponding weights (b_1, b_2, b_3, α) define a quadrature formula which is correct for polynomials of degree less than or equal to three, by the first in each set of equations (2.4.1)–(2.4.4).

Let us first suppose all the nodes to be distinct. The weights are determined as soon as the nodes are chosen, so the requirement that α be the weight associated with the node 1 puts one constraint on the choice of nodes. Since nodes α and 1 are already fixed, we may take this constraint as determining τ_3 in terms of τ_2, or vice versa, from the requirement

$$\alpha(=b_4)=\frac{\frac{1}{4}-\frac{1}{3}(\alpha+\tau_2+\tau_3)+\frac{1}{2}(\alpha\tau_2+\alpha\tau_3+\tau_2\tau_3)-\alpha\tau_2\tau_3}{(1-\alpha)(1-\tau_2)(1-\tau_3)}. \tag{3.8}$$

Thus the scheme (3.7) is determined by the choice of β and either τ_2 or τ_3, keeping in mind that these last two must be distinct from each other and from α and 1.

In fact we can now show the impossibility of a fourth-order scheme (3.7) having distinct nodes, by showing that it is impossible to choose β and τ_2 or τ_3 so that (3.8) is satisfied together with

$$b^*ATe=\tfrac{1}{6}, \quad b^*TATe=\tfrac{1}{8}, \quad b^*AT^2e=\tfrac{1}{12}. \tag{3.9}$$

In fact, putting the scheme (3.7) into (3.9), using the requirements that the quadrature formula defined on α, τ_2, τ_3, and 1 should be of order three and that α is a root of (3.5) leads to

$$\begin{aligned} b_3(\tau_2-\alpha)\beta &= \tfrac{1}{6}-\tfrac{3}{2}\alpha+3\alpha^2-\alpha^3, \\ b_3\tau_3(\tau_2-\alpha)\beta &= \tfrac{1}{8}-\tfrac{7}{6}\alpha+\tfrac{5}{2}\alpha^2-\alpha^3, \\ b_3(\tau_2^2-\alpha^2)\beta &= \tfrac{1}{8}-\tfrac{4}{3}\alpha+\tfrac{7}{2}\alpha^2-2\alpha^3. \end{aligned} \tag{3.10}$$

Notice that b_3 cannot vanish, else α would satisfy a cubic equation with integral coefficients; thus the first two equations of (3.10) determine τ_3 in terms of α, but then (3.8) and the last of equations (3.10) lead to conflicting values for τ_2. Thus there is no fourth-order scheme (3.7) with all nodes distinct.

Now we come to the case where not all the nodes are distinct. Here it is easy to see that there must be three distinct nodes, so that there are three cases to consider: (i) τ_2 or $\tau_3=\alpha$, (ii) τ_2 or $\tau_3=1$, (iii) $\tau_2=\tau_3$, different from α and 1.

In all three cases, the three distinct nodes must be equal in some order to $\{(1+r_i)/2 | i=1, 2, 3\}$, where the r_i are roots of a cubic polynomial orthogonal on $[-1, 1]$ to the constant functions. This constraint, together with the fact that two of the nodes are α and 1, determines the third node to be

$$\tfrac{1}{2}(1-2\alpha)/(1-3\alpha)$$

In this case one computes the weight associated with the node 1 to be

$$(1-6\alpha+6\alpha^2)/(6(1-\alpha)(1-4\alpha)) \tag{3.11}$$

and if this were equal to α, then α would satisfy a cubic equation with integral coefficients. Thus cases (i) and (iii) are impossible. In case (ii), putting $\tau_3=1$ leads to equality of the right-hand sides of the first two equations in (3.10), hence to $\alpha=\frac{1}{6}$ or $\frac{1}{2}$, which is not true; finally, trying $\tau_2=1$ leads, by the first and last of equations (3.10), also to $\alpha=\frac{1}{6}$ or $\frac{1}{2}$. This exhausts the last possibility, and the proof

that there is no 4th-order strongly S-stable DIRK formula in 4 stages is complete.

It turns out that strongly S-stable DIRK formulae of order four with five stages come in several families, each depending on one or more parameters: perhaps unfortunately, having a choice means one must optimize. Our investigation of these methods is still in progress.

4. A DIRK program. This section describes the implementation of some DIRK formulae on the CDC 6400 at the Computing Center of the University of Colorado. We present the results of comparisons of different DIRK formulae with the GEAR package [10].

The DIRK program works with a fixed formula selected by the user at the start of the integration from among the following five formulae, designated by number of stages and order.

DIRK (1, 2). Implicit midpoint rule, (2.1).

DIRK (2, 3). Formula of Crouzeix, (2.2).

DIRK (3, 4). Formula of Crouzeix, (2.3).

DIRK (2, 2). Strongly S-stable formula of Theorem 5, $\alpha = 1 - \frac{1}{2}\sqrt{2}$.

DIRK (3, 3). Strongly S-stable formula of Theorem 5.

We note that DIRK (1, 2) is A-stable but not S-stable [14].

Integration is by the step-halving method to estimate error and adjust stepsize. First a step of size h is taken from y_n at time t_n to compute y_{n+1}. Next, this time step is repeated in two steps of size $h/2$, and $y_{n+2/2}$ is obtained. The estimate of the local truncation error in the more accurate value $y_{n+2/2}$ is taken to be

$$E_{n+1} := \|y_{n+1} - y_{n+2/2}\|/(2^p - 1)$$

where p is the order of the method, $\|\cdot\|$ is the weighted RMS norm

$$\|y\| = \left(\frac{1}{m}\sum_{i=1}^{m} (y^i/y^i_{\max})^2\right)^{1/2}$$

and $y^i_{\max}$ is the maximum modulus of the ith component so far in the integration. This choice is suitable for comparing the program to the GEAR package [10]. It is motivated by the idea that, if the user specifies a local error tolerance ε, then errors of order $\varepsilon \cdot y^i_{\max}$ were allowed when y^i was near $y^i_{\max}$, and it is not in general useful to require later errors to be smaller than that. If the problem is asymptotically stable, as many stiff systems are, it is easy to modify the program to give a true relative error test.

The asymptotic form of the local truncation error [15] is used to adjust the stepsize. The user specifies a tolerance for local error and then:

(i) if $E_{n+1} > \varepsilon$ the step is rejected and h is reduced to make the expected error $\sim\varepsilon/5$;

(ii) if $3\varepsilon/4 < E_{n+1} \leqq \varepsilon$, the step is accepted, but h is reduced to make the expected error on the next step $\sim\varepsilon/5$;

(iii) if $\varepsilon/10 < E_{n+1} \leqq 3\varepsilon/4$, the step is accepted and the same h is used for the next step;

(iv) if $E_{n+1} \leqq \varepsilon/10$, the step is accepted and h is increased to make an expected error on the next step $\sim\varepsilon/2$, provided

(a) at least $p+1$ successful steps have followed the last decrease in h,

(b) after a decrease, h is at most doubled on the next increase, while the largest increase allowed in any case is by a factor of 10, and

(c) the increase is by a factor of at least 1.3.

The implicit equations at each stage are solved by a Newton's method; the matrices

$$I - \alpha h\, \partial f/\partial y \quad \text{and} \quad I - \tfrac{1}{2}\alpha h\, \partial f/\partial y$$

are computed from the Jacobian supplied by the user, and their LU-factorizations are stored and used repeatedly, being updated every twenty steps or at stepsize changes. Starting values for the $y_{n,i}$ are obtained by a combination of linear interpolation and extrapolation using the stored derivatives. If after three Newton steps there is no convergence, we allow one update of the Jacobian.

The package has been tested on some problems in the battery devised by Enright, Hull and Lindberg [8] to illustrate different kinds of stiffness. The problems for which we present results here are listed in Table 1. Problems B1 and

TABLE 1

Problems from the battery of Enright, Hull, Lindberg [8].

B1:	$y_1' = -y_1 + y_2$	$y_1(0) = 1$	
	$y_2' = -100y_1 - y_2$	$y_2(0) = 0$	Integrate on [0, 20]
	$y_3' = -100y_3 + y_4$	$y_3(0) = 1$	Use $h_0 =$ 7.E−3
	$y_4' = -10000y_3 - 100y_4$	$y_4(0) = 0$	
B5:	$y_1' = -10y_1 + 100y_2$	$y_1(0) = 1$	Integrate on [0, 20]
	$y_2' = -100y_1 - 10y_2$	$y_2(0) = 1$	Use $h_0 =$ 1.E−2
	$y_3' = -4y_3$	$y_3(0) = 1$	
	$y_4' = -y_4$	$y_4(0) = 1$	
	$y_5' = -.5y_5$	$y_5(0) = 1$	
	$y_6' = -.1y_6$	$y_6(0) = 1$	
C1:	$y_1' = -y_1 + y_2^2 + y_3^2 + y_4^2$	$y_1(0) = 1$	Integrate on [0, 20]
	$y_2' = -10y_2 + 10(y_3^2 + y_4^2)$	$y_2(0) = 1$	Use $h_0 =$ 1.E−2
	$y_3' = -40y_3 + 40y_4^2$	$y_3(0) = 1$	
	$y_4' = -100y_4 + 2$	$y_4(0) = 1$	
C5:	$y_1' = -y_1 + 2$	$y_1(0) = 1$	Integrate on [0, 20]
	$y_2' = -10y_2 + 20y_1^2$	$y_2(0) = 1$	Use $h_0 =$ 1.E−2
	$y_3' = -40y_3 + 80(y_1^2 + y_2^2)$	$y_3(0) = 1$	
	$y_4' = -100y_4 + 200(y_1^2 + y_2^2 + y_3^2)$	$y_4(0) = 1$	

B5 are linear. Problem B1 has complex eigenvalues, and some of its components decay rapidly while others decay slowly. Problem B5 has complex eigenvalues of large modulus close to the imaginary axis. It is known that this is a difficult problem for the GEAR package, whose backward differentiation formulae are not

A-stable at orders higher than two. Problem C1 shows nonlinear coupling from transient components to smooth components. Problem C5 has nonlinear coupling from smooth components to transient components, and was especially difficult for the implicit Runge–Kutta method tested by Enright, Hull and Lindberg.

TABLE 2
Computer test results.

Description of Parameters:

ε is the tolerance for relative local truncation error, fixed by user

Time Central processor time, in seconds, required to complete the integration.

Max. Err Maximum absolute error during the integration, expressed in the RMS norm:

$$\sup_n \left(\frac{1}{m} \sum_{i=1}^{m} (y_n^i - y^i(t_n))^2 \right)^{1/2}.$$

No. Steps Number of steps required to complete the integration.

NFE Number of function evaluations = number of Newton iterations.

NJE Number of Jacobian evaluations
(for DIRK, each Jacobian evaluation means two LU-decompositions).

		Time	Max. Err	No. Steps	NFE	NJE
B1	$\varepsilon = 10^{-2}$					
	DIRK (1, 2)	2.322	1.100E−1	89	303	30
	DIRK (2, 2)	1.956	8.433E−2	67	435	28
	DIRK (2, 3)	2.100	8.419E−2	59	391	29
	DIRK (3, 3)	1.737	6.301E−2	47	454	24
	GEAR	6.013	4.525E−1	293	600	11
	$\varepsilon = 10^{-4}$					
	DIRK (2, 3)	6.628	2.390E−3	217	1,364	37
	DIRK (3, 3)	5.576	1.733E−3	163	1,521	33
	DIRK (3, 4)	5.986	1.740E−3	169	1,586	35
	GEAR	7.588	5.884E−3	397	726	19
	$\varepsilon = 10^{-6}$					
	DIRK (3, 3)	19.223	5.414E−5	542	4.956	41
	DIRK (3, 4)	18.230	4.252E−5	489	4,496	45
	GEAR	13.410	7.598E−5	710	1,222	24
B5	$\varepsilon = 10^{-2}$					
	DIRK (1, 2)	2.268	2.220E−2	76	256	22
	DIRK (2, 2)	1.912	2.174E−2	52	342	15
	DIRK (2, 3)	2.097	1.947E−2	47	313	15
	DIRK (3, 3)	1.956	8.173E−3	39	376	14
	GEAR	44.951	5.502E−2	2,387	4,753	6
	$\varepsilon = 10^{-4}$					
	DIRK (2, 3)	8.175	3.757E−4	191	1,211	28
	DIRK (3, 3)	7.307	2.327E−4	148	1,393	27
	DIRK (3, 4)	8.174	2.406E−4	151	1,429	28
	GEAR	48.029	4.191E−4	2,337	4,825	14

Table 2—*cont.*

		Time	Max. Err	No. Steps	NFE	NJE
	$\varepsilon = 10^{-6}$					
	DIRK (3, 3)	23.674	1.363E−5	479	4,408	31
	DIRK (3, 4)	23.911	5.779E−6	457	4,219	32
	GEAR	48.648	8.540E−6	2,577	4,198	16
C1	$\varepsilon = 10^{-2}$					
	DIRK (1, 2)	0.694	4.060E−3	22	86	12
	DIRK (2, 2)	0.656	2.394E−3	20	139	11
	DIRK (2, 3)	0.631	1.679E−3	20	143	10
	DIRK (3, 3)	0.751	3.143E−3	18	177	9
	GEAR	1.034	6.074E−3	57	101	13
	$\varepsilon = 10^{-4}$					
	DIRK (2, 3)	1.941	3.257E−5	53	390	27
	DIRK (3, 3)	1.653	8.344E−5	40	454	20
	DIRK (3, 4)	1.716	6.783E−5	40	457	20
	GEAR	2.131	1.166E−4	112	186	20
	$\varepsilon = 10^{-6}$					
	DIRK (3, 3)	4.802	4.073E−6	133	1,419	33
	DIRK (3, 4)	4.233	1.266E−6	109	1,259	37
	GEAR	4.113	1.763E−6	206	289	27
C5	$\varepsilon = 10^{-2}$					
	DIRK (1, 2)		8.290E−2	11	47	6
			2.914E+0	29	101	13
		1.208	4.446E+1	43	148	20
	DIRK (2, 2)		5.423E−2	6	55	3
			1.437E+0	14	110	7
		0.966	1.834E+1	27	188	13
	DIRK (2, 3)		5.206E−2	10	76	5
			3.943E+0	22	150	11
		1.374	1.036E+1	34	226	19
	DIRK (3, 3)		2.600E−2	9	93	5
			4.860E+0	14	149	8
		1.041	1.120E+1	24	240	14
	GEAR		1.258E−1	17	40	6
			2.084E+0	29	62	8
		1.143	3.798E+1	54	107	13
	$\varepsilon = 10^{-4}$					
	DIRK (2, 3)		2.566E−3	25	229	8
			9.024E−2	82	573	14
		4.595	5.493E−1	127	871	29
	DIRK (3, 3)		2.692E−3	13	194	4
			1.353E−1	48	513	11
		3.537	6.871E−1	82	822	19
	DIRK (3, 4)		2.249E−3	20	266	8
			1.109E−1	63	661	14
		4.474	5.061E−1	100	1,008	33
	GEAR		1.102E−3	37	80	8
			3.933E−2	67	134	12
		1.909	1.473E+0	109	217	18

Table 2—*cont.*

	Time	Max. Err	No. Steps	NFE	NJE
$\varepsilon = 10^{-6}$					
DIRK (3, 3)		2.209E−4	45	673	8
		4.254E−3	186	1,960	15
	12.397	3.819E−2	307	3,159	35
DIRK (3, 4)		3.280E−5	53	958	8
		7.666E−3	206	2,947	11
	17.328	2.316E−2	370	5,332	29
GEAR		1.362E−5	69	134	12
		1.451E−3	116	225	17
	3.738	2.487E−2	198	375	24

Since the solution of problem C5 grows very fast, our error test is a relative one. Absolute error is reported here, however, so we give results for each method at the first step to pass times .1, 1, and 20, where the magnitude of y_4 is about 100, 6,000, and 37,000 respectively.

Our results are presented in Table 2. The following is a summary of pertinent observations.

1. As could be expected, the DIRK methods, being A-stable, outdid the GEAR package on the highly oscillatory Problem B5.

2. At modest tolerances, DIRK performs comparably to GEAR, but as the accuracy requirement is tightened, GEAR, having fifth-order methods available, is superior.

3. DIRK requires more function and Jacobian calls but less internal overhead than GEAR. On large problems DIRK would be less efficient.

4. At low accuracy the strongly S-stable DIRK (2, 2) and DIRK (3, 3) work as well as DIRK (2, 3) and DIRK (3, 4), respectively. Only at fairly strict tolerances do the maximum-order methods do better, and on Problem C5 the strongly S-stable methods do better throughout (recall too the remarks about DIRK (3, 4) accompanying the statement of Theorem 1). This tends to confirm the observations of Prothero and Robinson on the importance of stiff accuracy: this might be an ingredient in the solution of the difficulties of implicit Runge–Kutta methods described in the Enright–Hull–Lindberg tests.

We conclude with some remarks about where further work is needed. First, there should be tests of strongly S-stable methods of order four, or higher, to see whether they can overcome the degradation in performance of low order methods at high accuracy. Second, other strategies for estimating error want study; S. P. Nørsett has implemented a scheme of Fehlberg type for a stiffly A-stable two-stage DIRK method of order two [13]. Finally, it seems desirable that a program based on RK methods choose its formula automatically, both for stability properties and order of accuracy. Literally the last published word on the subject known to me, however, appears in Gear's 1971 book [9, p. 81]: "No techniques are currently available for selecting the order in Runge–Kutta methods."

Acknowledgment. I would like to thank Keith Miller for many helpful discussions. I also thank the Editor and referee for their suggestions, which led to numerous improvements of the manuscript, and for bringing to my attention the reference [13].

REFERENCES

[1] R. ALT, *Méthodes A-stables pour l'intégration de systèmes différentielles mal conditionnés*, Thèse présentée a l'Université Paris VI, Paris, 1971.

[2] J. C. BUTCHER, *Coefficients for the study of Runge–Kutta integration processes*, J. Austral. Math. Soc., 3 (1963), pp. 185–201.

[3] ———, *Implicit Runge–Kutta processes*, Math. Comp., 18 (1964), pp. 50–64.

[4] ———, *Integration processes based on Radau quadrature formulas*, Ibid., 18 (1964), pp. 233–244.

[5] ———, *An algebraic theory of integration methods*, Ibid., 26 (1972), pp. 79–106.

[6] M. CROUZEIX, *Sur l'approximation des équations différentielles opérationnelles linéaires par des méthodes de Runge Kutta*, Thèse présentée a l'Université Paris VI, Paris, 1975.

[7] B. L. EHLE, *On Padé approximations to the exponential function and A-stable methods for the solution of initial value problems*, Thesis, Univ. of Waterloo, Waterloo, Ontario, Canada, 1969.

[8] W. ENRIGHT, T. HULL AND B. LINDBERG, *Comparing numerical methods for stiff systems of ordinary differential equations*, BIT, 15 (1975), pp. 10–48.

[9] C. W. GEAR, *Numerical Initial Value Problems in Ordinary Differential Equations*, Prentice-Hall, Englewood Cliffs, NJ, 1971.

[10] A. C. HINDMARSH, *GEAR: Ordinary differential equation system solver*, LRL Rep. UCID-30001 Revision 3, December, 1974.

[11] M. A. KURDI, *Stable high order methods for time discretization of stiff differential equations*, Thesis, Univ. of California, 1974.

[12] K. MILLER, *Diagonally-implicit Runge–Kutta methods for PDE's and stiff ODE's*, Unpublished lecture notes, Berkeley, CA, 1973.

[13] S. P. NORSETT, *Semi-explicit Runge–Kutta methods*, Mathematics and Computation No. 6/74, Univ. of Trondheim.

[14] A. PROTHERO AND A. ROBINSON, *On the stability and accuracy of one-step methods for solving stiff systems of ordinary differential equations*, Math. Comp., 28 (1974), pp. 145–162.

[15] H. J. STETTER, *Analysis of Discretization Methods for Ordinary Differential Equations*, Springer-Verlag, New York, 1973.