Journal of Computational and Applied Mathematics 29 (1990) 25–33
North-Holland

Embedded Runge–Kutta formulae with stable equilibrium states

D.J. HIGHAM * and G. HALL
Department of Mathematics, University of Manchester, Manchester, United Kingdom M13 9PL

Received 12 May 1988
Revised 15 August 1988

Abstract: The equilibrium theory of Hall and Higham (1988) can be used to determine whether a Runge–Kutta algorithm will perform smoothly when stability restricts the stepsize. In this paper we show that current high quality order 4, 5 pairs do not behave well in this respect, and we determine the extent to which the overall quality must be compromised in order for the equilibrium conditions to be satisfied. Three new formulae are presented and their properties are compared with those of existing formulae.

Keywords: Runge–Kutta, embedded formulae, stability, stepsize selection.

1. Introduction

Explicit Runge–Kutta pairs can be effective tools for solving nonstiff and mildly stiff initial-value problems. In the latter case, when stability restricts the stepsize, the analysis given in [4–6] for a simple test problem leads to conditions which ensure efficient behaviour of the stepsize control mechanism. In this report we present three Runge–Kutta pairs which are well-behaved in the above sense. These are embedded 4,5 pairs from the 4-parameter family of Dormand and Prince [1].

We introduce the general formula pair below and outline properties which a good quality pair must possess. (For more details, see [1,9].) In Section 2 we show how the results of [4–6] can be used to influence the choice of parameters. We derive three new formula pairs and compare their properties with those of existing formulae. Numerical results which support the theory are given in Section 3.

We are concerned with the numerical solution of the initial-value problem

$$y' = f(x,\ y), \quad y(x_0) \text{ given}, \quad f : \mathbb{R} \times \mathbb{R}^m \to \mathbb{R}^m,$$

using an s-stage embedded Runge–Kutta 4,5 pair. The approximate solution, $\hat{y}_n \approx y(x_n)$, is advanced to $x_n + h_n$ by forming

$$\hat{y}_{n+1} = \hat{y}_n + \sum_{i=1}^{s} \hat{b}_i k_i,$$

* Supported by a SERC Research Studentship.

along with

$$\delta_{n+1} = y_{n+1} - \hat{y}_{n+1},$$

where

$$k_1 = h_n f(x_n,\ \hat{y}_n), \qquad k_i = h_n f\left(x_n + c_i h_n,\ \hat{y}_n + \sum_{j=1}^{i-1} a_{ij} k_j\right), \quad i = 2, \ldots, s,$$

and

$$y_{n+1} = \hat{y}_n + \sum_{i=1}^{s} b_i k_i.$$

We assume here that the pair is being used in local extrapolation mode: this means that $\hat{y}_{n+1}$ and y_{n+1} come from the 5th and 4th-order formulae respectively. This is by far the most popular choice in modern codes. The quantity δ_{n+1} gives an estimate for the local error in y_{n+1} which can be compared with the user-supplied tolerance TOL. After a successful step the next stepsize may be chosen according to an absolute error-per-step criterion

$$h_{n+1} = h_n \left(\frac{\gamma\ \mathrm{TOL}}{\| \delta_{n+1} \|} \right)^{1/5}, \tag{1}$$

with $0 < \gamma < 1$ a safety factor.

For a sufficiently smooth function f the local error in $\hat{y}_{n+1}$ may be written

$$le_{n+1} = h_n^6 \sum_{j=1}^{r_6} \hat{\tau}_j^{(6)} F_j^{(6)} + h_n^7 \sum_{j=1}^{r_7} \hat{\tau}_j^{(7)} F_j^{(7)} + \mathrm{O}\left(h_n^8\right), \tag{2}$$

where the elementary differentials $F_j^{(k)}$ depend only on f, and the truncation error coefficients $\hat{\tau}_j^{(k)}$ depend only on the 5th-order formula. Hence, in order for $\hat{y}_{n+1}$ and δ_{n+1} to be as accurate as possible it is desirable to have a small leading term in (2). Although this term is problem-dependent, good results were obtained in [1] by minimising the truncation error norm,

$$\| \hat{\tau}^{(6)} \|_2 =: A^{(6)}.$$

Further, to help justify the asymptotics used in the stepsize selection mechanism (1), Prince and Dormand [9] require that

$$\frac{\| \tau^{(6)} \|_2}{\| \tau^{(5)} \|_2} =: B^{(6)} \quad \text{and} \quad \frac{\| \tau^{(6)} - \hat{\tau}^{(6)} \|_2}{\| \tau^{(5)} \|_2} =: C^{(6)}$$

be fairly small ($\approx < 1$) and also $\tau_j^{(5)} \neq 0$, $j = 1, \ldots, r_5$, where $\tau_j^{(k)}$ are the kth-order truncation error coefficients of the 4th-order formula. Finally, to reduce the chance of excessive roundoff errors we should avoid having large coefficients $\{a_{ij}\}$, $\{\hat{b}_i\}$ and $\{b_i - \hat{b}_i\}$.

2. The new formulae

The behaviour of a Runge–Kutta pair when stability is restricting the stepsize has received much attention. Considering a linear test problem, Shampine [10] shows that the method will take an average stepsize h, where $h\lambda$ lies on the absolute stability boundary and λ is the

dominant eigenvalue of the Jacobian. Further analysis is given in [4] and extended in [5,6]. It is shown that an equilibrium state exists in which the method continues with constant stepsize h. This equilibrium state is realised in practice provided that it is stable with respect to small perturbations. If the equilibrium state is unstable then $h_n\lambda$ oscillates about the stability boundary and frequent step rejections occur, resulting in wasted function evaluations. Also, the global error fluctuates wildly. Hence the cost of the integration and the quality of the solution depend not only on the size of the absolute stability region but also on the stability of the equilibrium state. A sufficient condition for a stable equilibrium at a point $h\lambda$ on the stability boundary is [6]

$$\rho(C) < 1, \quad \text{where } C = \begin{pmatrix} 1 - \frac{1}{5}\mathscr{R}\left\{\dfrac{h\lambda E'(h\lambda)}{E(h\lambda)}\right\} & -\frac{1}{5} \\ \mathscr{R}\left\{\dfrac{h\lambda S'(h\lambda)}{S(h\lambda)}\right\} & 1 \end{pmatrix}, \tag{3}$$

and the polynomials S and E are characteristic of the method. We denote $\rho(C)$ at the point where $\arg(h\lambda) = \theta$ by μ_θ, for $\pi \geqslant \theta \geqslant \frac{1}{2}\pi$. When λ is nonreal the result is only strictly valid when a certain Euclidean-type norm is used in (1). However, in practice condition (3) seems to be beneficial in more general circumstances (see [5]). In the important special case where λ is real, the result holds for any choice of norm and hence $\mu_\pi < 1$ is an extremely desirable property, guaranteeing smooth, efficient solutions on many practical problems.

Therefore, in the search described below we sought methods which have $\mu_\theta < 1$ either at $\theta = \pi$ or over a wider range of θ. In particular we wished to determine the extent to which these extra constraints affect the truncation and absolute stability properties.

As a basis for our search we made use of the Dormand and Prince model [1]. Here the coefficients $\{a_{ij}\}$, $\{\hat{b}_i\}$, $\{b_i\}$ and $\{c_i\}$ are generated from the free parameters c_3, c_4, c_5 and b_7. The formulae produced have 7 stages with the "first-same-as-last" property—after a successful step the final function evaluation is the one required at the start of the next step. With this model the polynomial S in (3), which also determines the absolute stability region, has the form

$$S(z) = \sum_{i=0}^{5} \frac{z^i}{i!} + \frac{c_4(2 - 5c_3)}{240} z^6, \tag{4}$$

and it can be shown [8] that

$$E(z) = z^5 - \frac{c_3 - 2c_4(1 - 5c_3^2) + 2c_4^2(2 - 5c_3)(15c_3^2 - 14c_3 + 4)}{3c_3 - 2c_4(3 + 2c_3 - 20c_3^2) + 20c_4^2(2 - 5c_3)(1 - 3c_3 + 3c_3^2)} z^6$$
$$- \frac{c_4(2 - 5c_3)(c_3 - c_4(10c_3^2 - 8c_3 + 2))}{2(3c_3 - 2c_4(3 + 2c_3 - 20c_3^2) + 20c_4^2(2 - 5c_3)(1 - 3c_3 + 3c_3^2))} z^7. \tag{5}$$

(Since $E(z)$ appears in (3) in the form $E'(z)/E(z)$ we have rescaled to give a leading coefficient of unity.) From (4) and (5) we see that the absolute stability and equilibrium properties depend only on c_3 and c_4. Given these two values the term $A^{(6)}$ is a quadratic in c_5. Hence in performing an extensive computer search our approach was to vary c_3 and c_4 over a range where the absolute stability region has a reasonable size. For each pair $\{c_3, c_4\}$ the equilibrium properties

Table 1
Coefficients of RK5(4)7FEq1

c_i	a_{ij}						$\hat{b}_i$	b_i
0							$\frac{1}{12}$	$\frac{2}{15}$
$\frac{2}{9}$	$\frac{2}{9}$						0	0
$\frac{1}{3}$	$\frac{1}{12}$	$\frac{1}{4}$					$\frac{27}{32}$	$\frac{27}{80}$
$\frac{1}{2}$	$\frac{1}{8}$	0	$\frac{3}{8}$				$-\frac{4}{3}$	$-\frac{2}{15}$
$\frac{3}{5}$	$\frac{91}{500}$	$-\frac{27}{100}$	$\frac{78}{125}$	$\frac{8}{125}$			$\frac{125}{96}$	$\frac{25}{48}$
1	$-\frac{11}{20}$	$\frac{27}{20}$	$\frac{12}{5}$	$-\frac{36}{5}$	5		$\frac{5}{48}$	$\frac{1}{24}$
1	$\frac{1}{12}$	0	$\frac{27}{32}$	$-\frac{4}{3}$	$\frac{125}{96}$	$\frac{5}{48}$	0	$\frac{1}{10}$

were examined and $c_5 \in (0, 1)$ was chosen to minimise $A^{(6)}$. Once a promising triple $\{c_3, c_4, c_5\}$ had been located, b_7 was chosen to give acceptable values of $B^{(6)}$, $C^{(6)}$ and $\tau^{(5)}$.

Using this technique we obtained the three Runge–Kutta pairs presented in Tables 1–3. Our nomenclature comes from [2]. The first two formulae, RK5(4)7FEq1 and RK5(4)7FEq2, have stable equilibrium states when the dominant eigenvalue is real, the latter being only just stable with $\mu_\pi = 0.998$. The third formula, RK5(4)7FEq3, has $\mu_\theta < 1$ for $\pi \geqslant \theta \geqslant 1.005(\frac{1}{2}\pi)$. (Although formulae were found to exist with $\mu_\theta < 1$ for $\pi \geqslant \theta \geqslant \frac{1}{2}\pi$, we chose to sacrifice the $\theta \approx \frac{1}{2}\pi$ case for

Table 2
Coefficients of RK5(4)7FEq2

c_i	a_{ij}						$\hat{b}_i$	b_i
0							$\frac{181}{2700}$	$\frac{11377}{154575}$
$\frac{2}{13}$	$\frac{2}{13}$						0	0
$\frac{3}{13}$	$\frac{3}{52}$	$\frac{9}{52}$					$\frac{656903}{1846800}$	$\frac{35378291}{105729300}$
$\frac{5}{9}$	$\frac{12955}{26244}$	$-\frac{15925}{8748}$	$\frac{12350}{6561}$				$\frac{19683}{106400}$	$\frac{343359}{1522850}$
$\frac{3}{4}$	$-\frac{10383}{52480}$	$\frac{13923}{10496}$	$-\frac{176553}{199424}$	$\frac{505197}{997120}$			$\frac{34112}{110565}$	$\frac{535952}{1947645}$
1	$\frac{1403}{7236}$	$-\frac{429}{268}$	$\frac{733330}{309339}$	$-\frac{7884}{8911}$	$\frac{104960}{113967}$		$\frac{67}{800}$	$\frac{134}{17175}$
1	$\frac{181}{2700}$	0	$\frac{656903}{1846800}$	$\frac{19683}{106400}$	$\frac{34112}{110565}$	$\frac{67}{800}$	0	$\frac{1}{12}$

Table 3
Coefficients of RK5(4)7FEq3

c_i	a_{ij}						$\hat{b}_i$	b_i
0							$\frac{1247}{10890}$	$\frac{21487}{185130}$
$\frac{11}{45}$	$\frac{11}{45}$						0	0
$\frac{11}{30}$	$\frac{11}{120}$	$\frac{11}{40}$					$\frac{57375}{108053}$	$\frac{963225}{1836901}$
$\frac{55}{56}$	$\frac{106865}{87808}$	$-\frac{408375}{87808}$	$\frac{193875}{43904}$				$-\frac{1229312}{1962015}$	$-\frac{39864832}{33354255}$
$\frac{9}{10}$	$\frac{79503}{121000}$	$-\frac{1053}{440}$	$\frac{147753}{56870}$	$\frac{27048}{710875}$			$\frac{125}{207}$	$\frac{2575}{3519}$
1	$\frac{89303}{78045}$	$-\frac{2025}{473}$	$\frac{994650}{244541}$	$-\frac{2547216}{28122215}$	$\frac{475}{2967}$		$\frac{43}{114}$	$\frac{4472}{4845}$
1	$\frac{1247}{10890}$	0	$\frac{57375}{108053}$	$-\frac{1229312}{1962015}$	$\frac{125}{207}$	$\frac{43}{114}$	0	$-\frac{1}{10}$

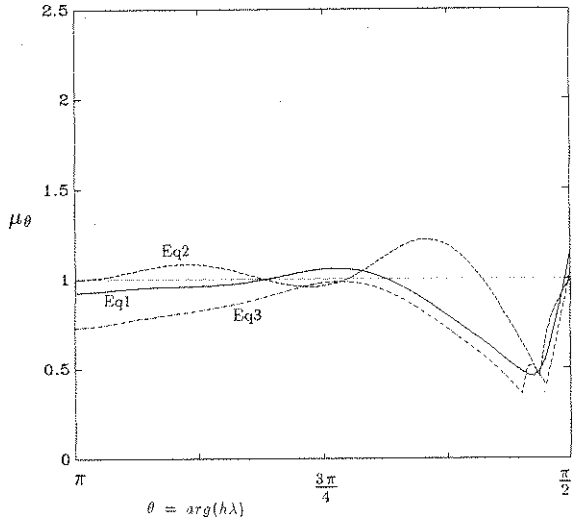

Fig. 1. μ_θ values for RK5(4)7FEq1, RK5(4)7FEq2 and RK5(4)7FEq3.

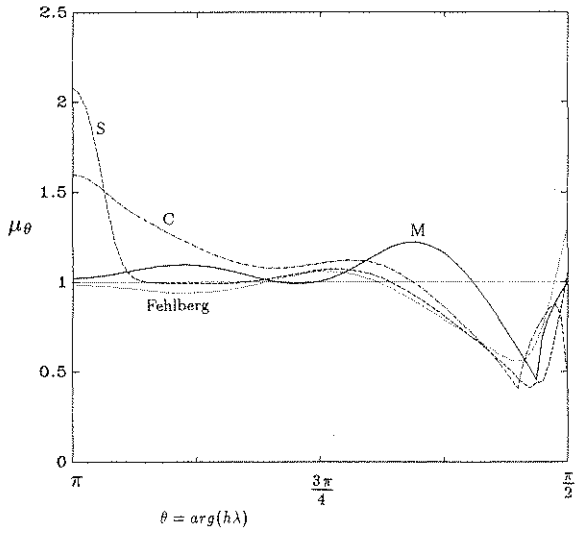

Fig. 2. μ_θ values for RK5(4)7FS, RK5(4)7FC, RK5(4)7FM and the Fehlberg pair.

the sake of overall quality.) Figures 1–4 and Table 4 show the main characteristics of the new formulae along with those of the following Dormand–Prince pairs [1,2]:

RK5(4)7FM which has a near-optimal error term, $A^{(6)}$;

RK5(4)7FS which has a large absolute stability region;

RK5(4)7FC which offers a compromise between error and stability properties.

For simplicity we will refer to the above formulae as Eq1, Eq2, Eq3, M, S and C. We also include the commonly used Fehlberg formula [3] which is a true 6-stage 4,5 pair. (Note that a failed step with this pair costs 5 function evaluations whereas a failed step with a member of the

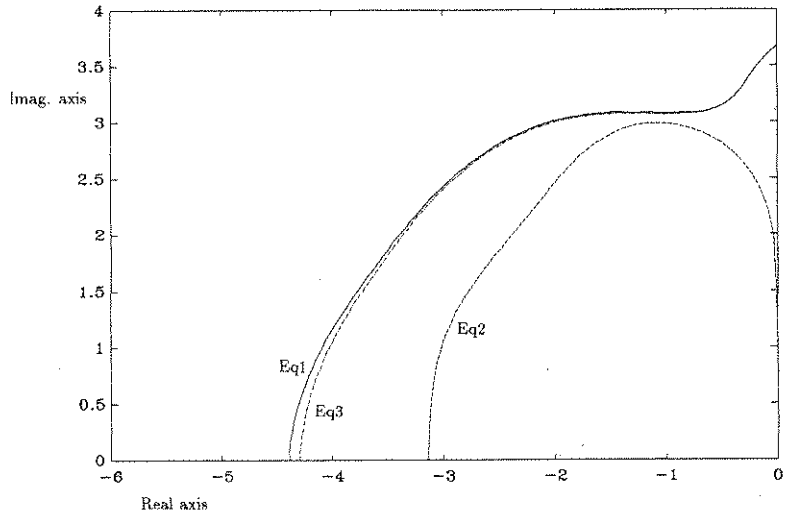

Fig. 3. Absolute stability boundaries of RK5(4)7FEq1, RK5(4)7FEq2 and RK5(4)7FEq3.

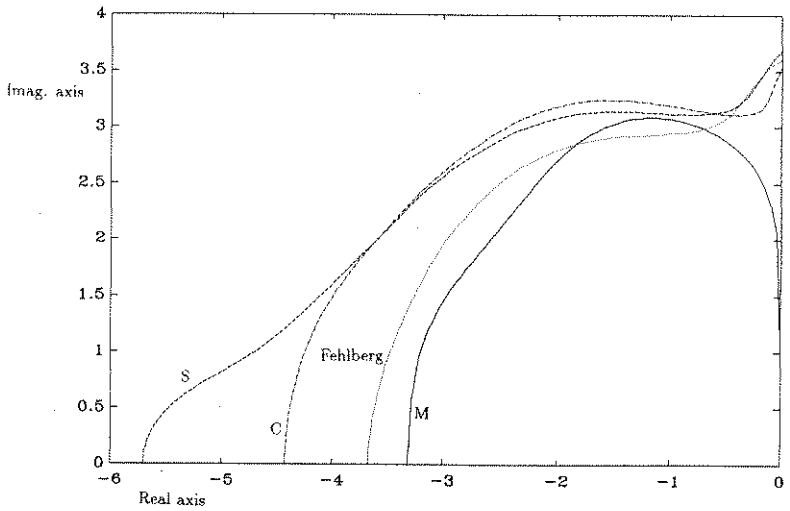

Fig. 4. Absolute stability boundaries of RK5(4)7FS, RK5(4)7FC, RK5(4)7FM and the Fehlberg pair.

Dormand–Prince family costs 6 function evaluations.) The M pair is widely acknowledged as the most efficient 4,5 pair for general use. We give results for the S and C pairs since they were specially designed to offer advantages on mildly stiff problems.

From Fig. 2 we see that the M, S and C pairs have $\mu_\pi > 1$, and that their μ_θ curves are generally quite poor. In particular, the benefits offered by the extended absolute stability boundaries near the real axis for the S and C pairs are outweighed by the large μ_θ values (typically a step rejection will occur after every two or three steps—see Section 3). For the Fehlberg pair, $\mu_\theta < 1$ at $\theta = \pi$ and over most of $[\frac{1}{2}\pi, \pi]$ and the stability region is larger than those of M and Eq2, but smaller than those of S, C, Eq1 and Eq3.

Although the M pair has a μ_π value which is close to 1, we were unable to satisfy the condition $\mu_\pi < 1$ without a significant loss in overall efficiency, as measured by $A^{(6)}$. The Eq2 pairs comes within a factor 2.5 of the "optimum" $A^{(6)}$ value, and Eq1, with its larger stability region, comes within a factor 5. To produce the excellent equilibrium properties of Eq3, a further increase in $A^{(6)}$ proved necessary. Note, however, that each of the new formulae has a smaller $A^{(6)}$ value than the Fehlberg pair.

Table 4
Truncation and real equilibrium values

Formula	$A^{(6)}$	$B^{(6)}$	$C^{(6)}$	μ_π
RK5(4)7FEq1	$1.80 \cdot 10^{-3}$	1.7	1.1	0.925
RK5(4)7FEq2	$9.38 \cdot 10^{-4}$	1.0	0.14	0.998
RK5(4)7FEq3	$2.49 \cdot 10^{-3}$	1.0	0.38	0.731
RK5(4)7FM	$3.99 \cdot 10^{-4}$	1.5	0.19	1.02
RK5(4)7FS	$1.81 \cdot 10^{-3}$	5.0	1.3	2.08
RK5(4)7FC	$1.49 \cdot 10^{-3}$	2.8	1.2	1.60
Fehlberg	$3.36 \cdot 10^{-3}$	3.2	0.16	0.985

We conclude that the Eq3 pair is superior to the Fehlberg pair since it has smaller leading truncation coefficients, a larger stability region and better equilibrium properties, We also consider the Eq2 pair, based on our search procedures, as the optimal 4,5 formula subject to the condition $\mu_\pi < 1$. Similar remarks apply to the Eq3 pair with the stronger restriction $\mu_\theta < 1$, $\frac{1}{2}\pi < \theta \leqslant \pi$. In general though, comparison of Runge–Kutta formulae is a subjective process which depends strongly upon how much emphasis one places on the different criteria. Clearly, for problems where stability restrictions are known not be become active the M pair must remain the method-of-choice among 4,5 formulae. It is worth stressing, however, that the Eq3 pair is significantly more efficient and reliable on mildly stiff problems than any other known pair.

3. Numerical results

To illustrate the equilibrium theory, we solved the following system:

$$y' = Ay, \quad \text{where } A = \begin{pmatrix} R\cos\theta & -R\sin\theta & 1 \\ R\sin\theta & R\cos\theta & 2 \\ 0 & 0 & -1 \end{pmatrix},$$

in which the Jacobian A has eigenvalues $\{Re^{\pm i\theta}, -1\}$. We chose $R = 10^4$ and used θ as a parameter in order to vary the argument of the dominant eigenvalue(s) in the complex plane. The seven Runge–Kutta pairs mentioned in Section 2 were implemented in locally extrapolated error-per-step mode using the stepsize mechanism (1) with Euclidean vector norm and a local error tolerance of TOL $= 10^{-3}$. An initial condition of $y(0) = (-10^{-4}, 10^{-4}, 2)^T$ ensured that the fast transients affected the numerical stability rather than the accuracy of the methods. The initial stepsize was chosen to put $h\lambda$ near the absolute stability boundary and 500 steps were taken. Allowing 20 steps for the stepsize selection mechanism to settle down, we recorded N_{FAIL},

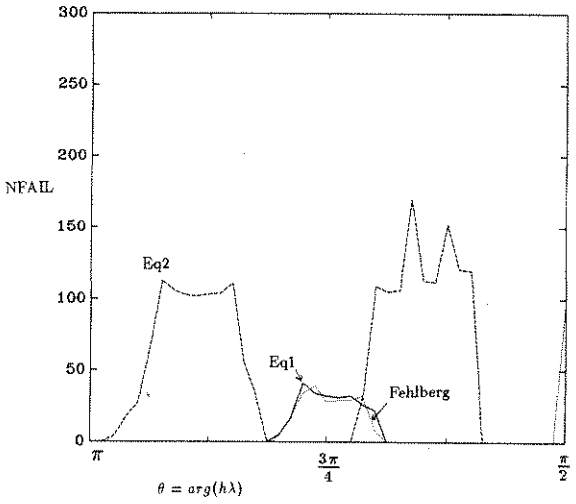

Fig. 5. Step failures with RK5(4)7FEq1, RK5(4)7FEq2 and the Fehlberg pair on a linear problem. (For RK5(4)7FEq3, $N_{\text{FAIL}} \equiv 0$ except at $\theta = \frac{1}{2}\pi$.)

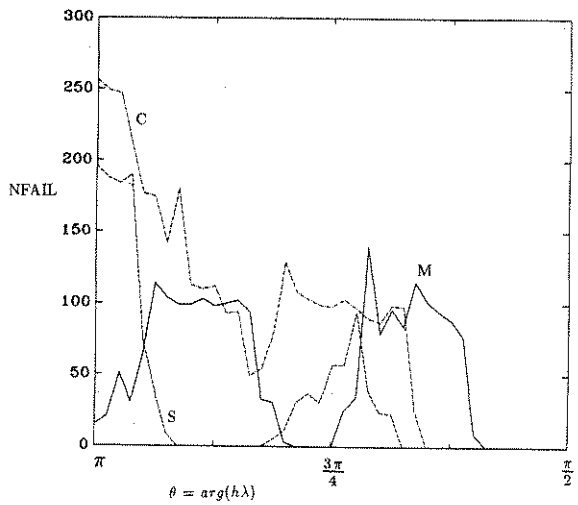

Fig. 6. Step failures with RK5(4)7FS, RK5(4)7FC and RK5(4)7FM on a linear problem.

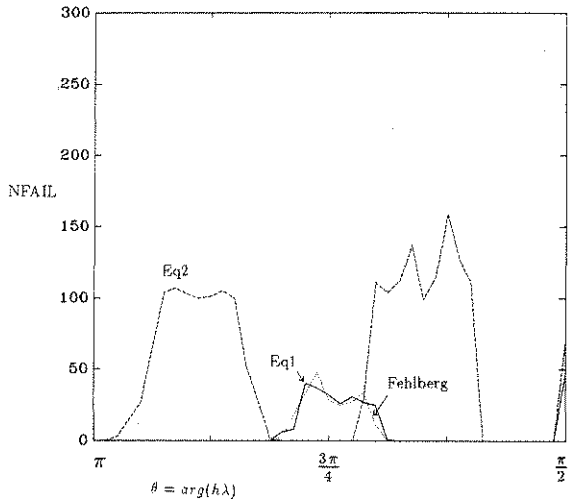

Fig. 7. Step failures with RK5(4)7FEq1, RK5(4)7FEq2 and the Fehlberg pair on the Krogh problem. (For RK5(4)7FEq3, $N_{\mathrm{FAIL}} \equiv 0$.)

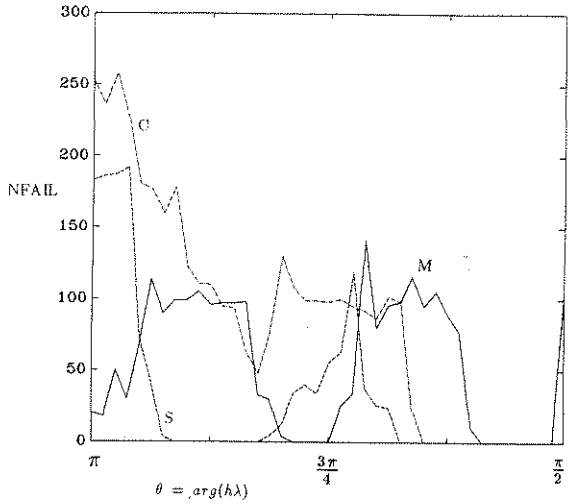

Fig. 8. Step failures with RK5(4)7FS, RK5(4)7FC and RK5(4)7FM on the Krogh problem.

the number of failed steps between x_{20} and x_{500}. Results for 41 equally spaced values of θ in $[\pi, \frac{1}{2}\pi]$ are given in Figs. 5 and 6—piecewise linear interpolation has been used to emphasise the pattern of failures. We see that the behaviour is in almost exact agreement with that predicted by the equilibrium plots in Figs. 1 and 2. Exceptions occur at $\theta = \frac{1}{2}\pi$; for example RK5(4)7FEq1 records no failures here although $\mu_{\pi/2} > 1$ from Fig. 1. The explanation is that the absolute stability boundary crosses the imaginary axis at more than one point and the method has found an alternative equilibrium state.

Similar results were obtained using a nonlinear problem of Krogh [7]:

$$y' = -By + U^{\mathrm{T}}\left(\tfrac{1}{2}z_1^2 - \tfrac{1}{2}z_2^2,\ z_1 z_2,\ z_3^2,\ z_4^2\right)^{\mathrm{T}},$$

where

$$z = Uy, \quad B = U^{\mathrm{T}} \begin{pmatrix} -10\cos\theta & -10\sin\theta & 0 & 0 \\ 10\sin\theta & -10\cos\theta & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0.5 \end{pmatrix} U,$$

and U is the orthogonal matrix with diagonal elements of $-\frac{1}{2}$ and all other elements equal to $\frac{1}{2}$. Here the Jacobian has eigenvalues which approach $\{-|10\cos\theta| \pm \mathrm{i}\, 10\sin\theta, -1, -0.5\}$ as $x \to \infty$. Using the solution given by Krogh, which corresponds to the initial condition $y(0) = (0, -2, -1, -1)^{\mathrm{T}}$, we started the integration at a point where stability restrictions occurred. As above, the number of step failures N_{FAIL} between x_{20} and x_{500} were recorded for 41 equally spaced θ values in $[\pi, \frac{1}{2}\pi]$. These are plotted in Figs. 7 and 8. Again, for each formula pair, the pattern of step failures closely matches the equilibrium plot.

If the infinity norm is used instead of the Euclidean norm, or the system is altered so that the dominant subsystem has a nonnormal Jacobian, then, strictly, the equilibrium theory is no longer applicable. In practice we have observed that the former change makes little difference to the overall pattern of rejected steps, although the latter does have a marked effect.

We have also carried out precision-work tests (global error versus number of function evaluations) on some nonstiff test problems for the seven Runge–Kutta pairs. Our results are in broad agreement with those of [1,2]—the size of the truncation coefficients, in particular $A^{(6)}$, strongly influences the efficiency.

In summary, we have developed three new order 4,5 pairs with special properties which allow them to perform extremely efficiently on a restricted problem class. We have also quantified the compromises in overall quality which these extra features necessitate. A similar investigation of higher order pairs is currently being undertaken.

Acknowledgements

We thank Dr. P.J. Prince for supplying the coefficients in equation (5) and Dr. N.J. Higham for commenting on the manuscript.

References

[1] J.R. Dormand and P.J. Prince, A family of embedded Runge–Kutta formulae, *J. Comput. Appl. Math.* **6** (1980) 19–26.

[2] J.R. Dormand and P.J. Prince, A reconsideration of some embedded Runge–Kutta formulae, *J. Comput. Appl. Math.* **15** (1986) 203–211.

[3] E. Fehlberg, Klassische Runge–Kutta-Formeln vierter und niedriger Ordnung mit Schrittweiten-Kontrolle und ihre Anwendung auf Wärmeleitungsprobleme, *Computing* **6** (1970) 61–71.

[4] G. Hall, Equilibrium states of Runge–Kutta schemes, *ACM Trans. Math. Software* **11** (1985) 289–301.

[5] G. Hall, Equilibrium states of Runge–Kutta schemes: part II, *ACM Trans. Math. Software* **12** (1986) 183–192.

[6] G. Hall and D.J. Higham, Analysis of stepsize selection schemes for Runge–Kutta codes, *IMA J. Numer. Anal.* **8** (1988) 305–310.

[7] F.T. Krogh, On testing a subroutine for the numerical integration of ordinary differential equations, *J. Assoc. Comput. Mach.* **4** (1973) 545–562.

[8] P.J. Prince, Private communication, 1986.

[9] P.J. Prince and J.R. Dormand, High order embedded Runge–Kutta formulae, *J. Comput. Appl. Math.* **7** (1981) 67–75.

[10] L.F. Shampine, Stiffness and non-stiff differential equation solvers, in: L. Collatz, Ed., *Numerische Behandlung von Differential Gleichungen*, Internat. Ser. Numer. Math. **27** (Birkhäuser, Basel, 1975) 287–301.