IMA Journal of Numerical Analysis (1988) **8**, 305–310

Analysis of Stepsize Selection Schemes for Runge–Kutta Codes

G. HALL AND D. J. HIGHAM

Department of Mathematics, University of Manchester, Manchester M13 9PL, England

[Received 1 September 1987 and in revised form 3 February 1988]

Conditions on Runge–Kutta algorithms can be obtained which ensure smooth stepsize selection when stability of the algorithm is restricting the stepsize. Some recently derived results are shown to hold for a more general test problem.

1. Introduction

IN THE papers [1, 2], a theory was developed for analysis of the step-selection mechanisms of commonly used Runge–Kutta codes. Numerical experiments show that, for certain choices of problems and methods, when the step-size is restricted by stability, such mechanisms produce smooth solutions. In other cases, an erratic sequence of steps is observed, some inside and some outside the stability region of the method, including frequent step rejections, with consequent loss of smoothness in solution components. Conditions on the method were obtained, using simple test problems, which enabled this behaviour to be better understood and predicted. The purpose of this paper is to show that these conditions are still valid for a more general test problem.

We consider the problem

$$y' = Ay, \qquad A \in \mathbb{R}^{s \times s}, \tag{1.1}$$

where A is assumed to be nondefective. The real Jordan form of A ([3: p. 242]) is then

$$PAP^{-1} = B = \operatorname{diag}(B_1, \ldots, B_t) \tag{1.2}$$

where $P \in \mathbb{R}^{s \times s}$ and each B_k is either a 1×1 block containing a real eigenvalue of A, or a 2×2 block of the form

$$B_k = \begin{bmatrix} \beta & -\gamma \\ \gamma & \beta \end{bmatrix},$$

corresponding to a complex conjugate pair of eigenvalues $\beta \pm i\gamma$, with $\gamma \neq 0$.

An explicit Runge–Kutta algorithm applied to (1.1) takes the form (see [1])

$$y_{n+1} = s(h_n A)y_n, \qquad e_{n+1} = e(h_n A)y_n, \qquad h_{n+1} = (\theta\delta_{\text{tol}}/\|e_{n+1}\|)^{1/q}h_n, \tag{1.3}$$

where the polynomials s and e are characteristic of the method. The above form assumes that control is by the criterion of absolute error per step, aiming at a fraction θ of the tolerance δ_{tol}. If the criterion of error per unit step is used, then

$$e_{n+1} = A\bar{e}(h_n A)y_n$$

for a certain polynomial $\hat{e}$. Our conclusions then remain valid with e replaced by $\hat{e}$.

In Section 2, we consider the case where the dominant eigenvalues, relative to the stability region of the method, are a complex conjugate pair. This means that this pair of eigenvalues would be the first to give a violation of the condition of absolute stability as the stepsize is increased. The corresponding analysis for the case of a dominant real eigenvalue was already given in [1]. Section 3 briefly reviews this result and includes comments on the interpretation of the two cases.

2. Dominant complex eigenvalues

Suppose that the dominant eigenvalues are λ and $\bar{\lambda}$, where $\lambda = \beta + i\gamma$, and let these be the eigenvalues of the 2×2 block B_k in (1.2). Denote the eigensystem of B by

$$((\lambda, \boldsymbol{u}); (\bar{\lambda}, \bar{\boldsymbol{u}}); (\lambda_i, \boldsymbol{u}_i) : i = 3, \ldots, s),$$

where

$$\boldsymbol{u} = [0, \ldots, 0, \tfrac{1}{2}\sqrt{2}, -\tfrac{1}{2}i\sqrt{2}, 0, \ldots, 0]^{\mathrm{T}}$$

so that

$$\boldsymbol{u}^{\mathrm{T}}\boldsymbol{u} = 0, \qquad \bar{\boldsymbol{u}}^{\mathrm{T}}\boldsymbol{u} = 1,$$

and, from the form of B,

$$\boldsymbol{u}^{\mathrm{T}}\boldsymbol{u}_i = \bar{\boldsymbol{u}}^{\mathrm{T}}\boldsymbol{u}_i = 0 \quad (i = 3, \ldots, s).$$

Further, if the corresponding eigenvectors of A are $\boldsymbol{v}$, $\bar{\boldsymbol{v}}$ and $\boldsymbol{v}_i$ $(i = 3, \ldots, s)$, where $\boldsymbol{v} = P^{-1}\boldsymbol{u}$ etc., then these relations can be expressed as

$$\boldsymbol{v}^{\mathrm{T}}P^{\mathrm{T}}P\boldsymbol{v} = 0, \qquad \bar{\boldsymbol{v}}^{\mathrm{T}}P^{\mathrm{T}}P\boldsymbol{v} = 1, \qquad \boldsymbol{v}^{\mathrm{T}}P^{\mathrm{T}}P\boldsymbol{v}_i = \bar{\boldsymbol{v}}^{\mathrm{T}}P^{\mathrm{T}}P\boldsymbol{v}_i = 0 \quad (i = 3, \ldots, s). \tag{2.1}$$

For the analysis of this section, the norm in (1.3) is the P norm, defined by

$$\| \bullet \|_P = \| P \bullet \|_2.$$

The practical effects of this choice are discussed in Section 3.

Lemma 1. *Let $t(\bullet)$ be any polynomial with real coefficients, and consider $\boldsymbol{w} = \alpha\boldsymbol{v} + \bar{\alpha}\bar{\boldsymbol{v}}$. Then*

$$\| t(hA)\boldsymbol{w} \|_P = |t(h\lambda)| \, \|\boldsymbol{w}\|_P.$$

Proof

$$\begin{aligned} \| t(hA)\boldsymbol{w} \|_P^2 &= [\alpha t(h\lambda)\boldsymbol{v}^{\mathrm{T}} + \bar{\alpha} t(h\bar{\lambda})\bar{\boldsymbol{v}}^{\mathrm{T}}]P^{\mathrm{T}}P[\alpha t(h\lambda)\boldsymbol{v} + \bar{\alpha} t(h\bar{\lambda})\bar{\boldsymbol{v}}] \\ &= 2\alpha\bar{\alpha} t(h\lambda) t(h\bar{\lambda}) \quad \text{(from (2.1))}. \end{aligned}$$

The special case $t \equiv 1$ gives $\|\boldsymbol{w}\|_P^2 = 2\alpha\bar{\alpha}$, and the result follows. □

Using this lemma, we can write down a particular solution of the recurrence (1.3), referred to as the steady-state solution. This is given by

$$|s(h_{\mathrm{L}}\lambda)| = 1, \qquad |e(h_{\mathrm{L}}\lambda)| \, \|\boldsymbol{w}_n\|_P = \theta\delta_{\mathrm{tol}}. \tag{2.2}$$

Here the stepsize is constant at h_L, and the solution $w_n = \alpha_n v + \bar{\alpha}_n \bar{v} \in \mathbb{R}^s$ is of constant norm: $\|w_n\|_P = (2\alpha_n \bar{\alpha}_n)^{\frac{1}{2}} =: d_L$, since it follows from Lemma 1 and (2.2) that $\|w_{n+1}\|_P = \|w_n\|_P = d_L$. The points $h_L\lambda$ and $h_L\bar{\lambda}$ are on the boundary of the stability region of the method, and the dominance assumption implies that

$$|s(h_L\lambda_i)| < 1 \quad (i = 3, \dots, s).$$

Remark. In interpreting results obtained for the test problem (1.1) on general problems, the solution y of (1.1) represents the global error (see [1]). Hence the global error in the steady-state solution is in the span of the dominant eigenvectors v and $\bar{v}$.

Whether the steady-state solution defined by (2.2), with its associated smoothness, will be realized in a practical computation depends on the stability of this solution of (1.3) with respect to small perturbations in the stepsize and solution components. In the remainder of this section, we carry out this analysis.

LEMMA 2. *Let $y = z + \varepsilon$, where ε is a vector of small perturbations. Then, to first order of small quantities,*

$$\|y\|_P = \|z\|_P(1 + z^T P^T P\varepsilon / z^T P^T P z).$$

Proof

$$\begin{aligned} \|y\|_P^2 = \|Py\|_2^2 &= (z^T + \varepsilon^T)P^T P(z + \varepsilon) \\ &= \|z\|_P^2 + 2z^T P^T P\varepsilon + O(\varepsilon^T\varepsilon) \\ &= \|z\|_P^2(1 + 2z^T P^T P\varepsilon / z^T P^T P z) + O(\varepsilon^T \varepsilon), \end{aligned}$$

and the result follows. □

In the following, p' will denote the derivative of the polynomial p.

THEOREM 1. *Consider a general real perturbation of the solution* (2.2), *which may be written in the form*

$$h_n = h_L(1 + \varepsilon_n), \qquad y_n = \alpha_n(1 + \delta_n)v + \bar{\alpha}_n(1 + \bar{\delta}_n)\bar{v} + \sum_{i=3}^{s} \delta_n^{(i)} v_i. \tag{2.3}$$

Then, to first order of small quantities, the perturbations are propagated according to

$$\varepsilon_{n+1} = \left(1 - \frac{1}{q}\,\mathrm{Re}\,\frac{h_L\lambda e'(h_L\lambda)}{e(h_L\lambda)}\right)\varepsilon_n - \frac{1}{q}\,\mathrm{Re}\,\delta_n, \qquad \delta_{n+1} = \delta_n + \varepsilon_n \frac{h_L\lambda s'(h_L\lambda)}{s(h_L\lambda)},$$

$$\delta_{n+1}^{(i)} = s(h_L\lambda_i)\delta_n^{(i)} \quad (i = 3, \dots, s).$$

(*Note that, if some v_i $(i = 3, \dots, s)$ are complex, then the summation $\sum_{i=3}^{s} \delta_n^{(i)} v_i$ will include pairs of complex conjugate terms.*)

Proof. (In the following we neglect all but first-order terms in small quantities without further comment.)

$$\begin{aligned}\mathbf{y}_{n+1} &= s\big(h_{\mathrm{L}}(1+\varepsilon_n)A\big)\mathbf{y}_n\\ &= [s(h_{\mathrm{L}}A)+\varepsilon_n h_{\mathrm{L}} A s'(h_{\mathrm{L}}A)]\mathbf{y}_n\\ &= \alpha_n(1+\delta_n)s(h_{\mathrm{L}}\lambda)\mathbf{v}+\bar{\alpha}_n(1+\bar{\delta}_n)s(h_{\mathrm{L}}\bar{\lambda})\bar{\mathbf{v}}+\sum_{i=3}^{s}\delta_n^{(i)}s(h_{\mathrm{L}}\lambda_i)\mathbf{v}_i\\ &\quad+\varepsilon_n[\alpha_n h_{\mathrm{L}}\lambda s'(h_{\mathrm{L}}\lambda)\mathbf{v}+\bar{\alpha}_n h_{\mathrm{L}}\bar{\lambda}s'(h_{\mathrm{L}}\bar{\lambda})\bar{\mathbf{v}}].\end{aligned}$$

Comparing with

$$\mathbf{y}_{n+1}=\alpha_{n+1}(1+\delta_{n+1})\mathbf{v}+\bar{\alpha}_{n+1}(1+\bar{\delta}_{n+1})\bar{\mathbf{v}}+\sum_{i=3}^{s}\delta_{n+1}^{(i)}\mathbf{v}_i,$$

where $\mathbf{w}_{n+1}=\alpha_{n+1}\mathbf{v}+\bar{\alpha}_{n+1}\bar{\mathbf{v}}$ and $\alpha_{n+1}=s(h_{\mathrm{L}}\lambda)\alpha_n$, we obtain

$$\delta_{n+1}=\delta_n+\varepsilon_n h_{\mathrm{L}}\lambda s'(h_{\mathrm{L}}\lambda)/s(h_{\mathrm{L}}\lambda),\qquad \delta_{n+1}^{(i)}=s(h_{\mathrm{L}}\lambda_i)\delta_n^{(i)},$$

as required.

Similarly we obtain

$$\mathbf{e}_{n+1}=e(h_{\mathrm{L}}\lambda)\alpha_n(1+\varphi_n)\mathbf{v}+e(h_{\mathrm{L}}\bar{\lambda})\bar{\alpha}_n(1+\bar{\varphi}_n)\bar{\mathbf{v}}+\sum_{i=3}^{s}e(h_{\mathrm{L}}\lambda_i)\delta_n^{(i)}\mathbf{v}_i,$$

where $\varphi_n=\delta_n+\varepsilon_n h_{\mathrm{L}}\lambda e'(h_{\mathrm{L}}\lambda)/e(h_{\mathrm{L}}\lambda)$, which we write as

$$\mathbf{e}_{n+1}=e(h_{\mathrm{L}}A)\mathbf{w}_n+\alpha_n\varphi_n e(h_{\mathrm{L}}\lambda)\mathbf{v}+\bar{\alpha}_n\bar{\varphi}_n e(h_{\mathrm{L}}\bar{\lambda})\bar{\mathbf{v}}+\sum_{i=3}^{s}e(h_{\mathrm{L}}\lambda_i)\delta_n^{(i)}\mathbf{v}_i.$$

Using Lemma 2 and (2.2), we find

$$\|\mathbf{e}_{n+1}\|_P=\theta\delta_{\mathrm{tol}}(1+\gamma_n/e(h_{\mathrm{L}}\lambda)e(h_{\mathrm{L}}\bar{\lambda})d_{\mathrm{L}}^2),$$

where

$$\begin{aligned}\gamma_n &= [\alpha_n e(h_{\mathrm{L}}\lambda)\mathbf{v}^{\mathrm{T}}+\bar{\alpha}_n e(h_{\mathrm{L}}\bar{\lambda})\bar{\mathbf{v}}^{\mathrm{T}}]P^{\mathrm{T}}P\Big(\alpha_n\varphi_n e(h_{\mathrm{L}}\lambda)\mathbf{v}+\bar{\alpha}_n\bar{\varphi}_n e(h_{\mathrm{L}}\bar{\lambda})\bar{\mathbf{v}}\\ &\quad+\sum_{i=3}^{s}e(h_{\mathrm{L}}\lambda_i)\delta_n^{(i)}\mathbf{v}_i\Big)\\ &= \alpha_n\bar{\alpha}_n e(h_{\mathrm{L}}\lambda)e(h_{\mathrm{L}}\bar{\lambda})(\varphi_n+\bar{\varphi}_n)\end{aligned}$$

(from (2.1)), and hence

$$\|\mathbf{e}_{n+1}\|_P=\theta\delta_{\mathrm{tol}}(1+\mathrm{Re}\,\varphi_n),$$

since $2\alpha_n\bar{\alpha}_n=d_{\mathrm{L}}^2$. Note that the perturbations $\delta_n^{(i)}$ have no effect, to first order, on $\|\mathbf{e}_{n+1}\|_P$. Finally, from (1.3),

$$h_{\mathrm{L}}(1+\varepsilon_{n+1})=(1+\mathrm{Re}\,\varphi_n)^{-1/q}h_{\mathrm{L}}(1+\varepsilon_n),$$

so that $\varepsilon_{n+1}=\varepsilon_n-(1/q)\,\mathrm{Re}\,\varphi_n$, which completes the proof. □

Using Theorem 1 we may generalize a result which was proved in [2] for a 2 × 2 test problem. In the following, ρ(•) denotes the spectral radius.

THEOREM 2. *The steady-state solution of* (1.3) *defined by conditions* (2.2) *is stable* *with respect to small perturbations if* $\rho(C) < 1$ *where*

$$C = \begin{bmatrix} 1 - \frac{1}{q} \operatorname{Re} \frac{h_{\mathrm{L}}\lambda e'(h_{\mathrm{L}}\lambda)}{e(h_{\mathrm{L}}\lambda)} & -\frac{1}{q} \\ \operatorname{Re} \frac{h_{\mathrm{L}}\lambda s'(h_{\mathrm{L}}\lambda)}{s(h_{\mathrm{L}}\lambda)} & 1 \end{bmatrix}.$$

Proof. Let $\delta_n = \rho_n + \mathrm{i}\sigma_n$. Applying Lemma 2 to (2.3), we obtain

$$\begin{aligned} \|y_n\|_P &= d_{\mathrm{L}}\left[1 + (\alpha_n v^{\mathrm{T}} + \bar{\alpha}_n \bar{v}^{\mathrm{T}}) P^{\mathrm{T}} P\left(\alpha_n \delta_n v + \bar{\alpha}_n \bar{\delta}_n \bar{v} + \sum_{i=3}^{s} \delta_n^{(i)} v_i\right) \Big/ 2\alpha_n \bar{\alpha}_n\right] \\ &= d_{\mathrm{L}}(1 + \tfrac{1}{2}\delta_n + \tfrac{1}{2}\bar{\delta}_n) \quad \text{(using (2.1))} \\ &= d_{\mathrm{L}}(1 + \rho_n). \end{aligned}$$

Therefore, from Theorem 1, perturbations to the constant stepsize and constant norm of the steady-state solution are propagated according to

$$\begin{pmatrix} \varepsilon_{n+1} \\ \rho_{n+1} \end{pmatrix} = C \begin{pmatrix} \varepsilon_n \\ \rho_n \end{pmatrix},$$

giving the desired result. □

3. Conclusions

In the case where the dominant eigenvalue is real, the test problem (1.1) was analyzed and Theorem 2 proved in [1]. The important distinction to make in this case is that the result was shown to hold using any vector norm in (1.3). There is therefore a clear advantage in using algorithms with $\rho(C) < 1$ at $h_{\mathrm{L}}\lambda$ where $[h_{\mathrm{L}}\lambda, 0]$ is the interval·of absolute stability along the negative real axis.

The analysis of Section 2 shows that, for the complex case, it is necessary in general to use the P norm to enable the smoothing effects of the condition $\rho(C) < 1$ to be strictly realized. This is, of course, not a practical possibility. However, for a number of widely used test problems, we have $P = I$, and therefore the beneficial effects of $\rho(C) < 1$ will be realized in the Euclidean norm, thus affecting results on comparisons of methods. More generally, we have $\|\bullet\|_P = \|\bullet\|_2$ whenever P is orthogonal, and hence whenever A is normal. (In this case, (1.2) is the real Schur decomposition.) There are therefore practical problems on which the condition $\rho(C) < 1$ and the use of the Euclidean norm produce smooth solutions. The behaviour when the Euclidean norm is used and P is not orthogonal, or when the max norm is used, was briefly illustrated in [2]; there can be beneficial effects although a true steady-state solution does not exist.

The authors are currently involved in the derivation of practical Runge–Kutta formulae taking this theory into account.

Acknowledgements

We thank Nick Higham for commenting on the manuscript and supplying reference [3]. The work of the second author was supported by a SERC Research Studentship.

References

1. Hall, G. 1985 Equilibrium states of Runge–Kutta Schemes. *ACM Trans on Math. Software* **11,** 289–301.
2. Hall, G. 1986 Equilibrium states of Runge–Kutta Schemes—II. *ACM Trans on Math. Software* **12,** 183–192.
3. Lancaster, P., & Tismenetsky, 1985 *The Theory of Matrices* (2nd). Academic Press, New York.