

STABILITY OF RUNGE-KUTTA METHODS USED IN MODULAR INTEGRATION

B. OWREN

ABSTRACT. A pair of Runge-Kutta methods is applied to a system of ordinary differential equations in a modular fashion known as time point relaxation. For a class of two by two linear systems with constant coefficients, the concept of coupling stability is introduced. This is way of measuring the loss of stability due to the decoupling of the system into two scalar subsystems. The strategy for handling the interactions between the two modules is controlled by a parameter, where certain choices of the parameter corresponds to the Gauss-Jacobi and Gauss-Seidel method. Results are obtained for the case when Runge-Kutta methods in general are applied with only one iteration per time step. The case with several iteration is investigated for the well-known θ -methods.

1. INTRODUCTION

“*Modular integration* is a technique connected to dynamic simulation. Modules representing different parts of a process integrate their variables over given time periods, after each there is an exchange of variables between the modules. The integration between each exchange takes place in separate local integrators”

This definition of modular integration is given by Iversen [3]. In numerical analysis, the idea of splitting systems of ordinary differential equations into subsystems which are integrated independently over a time window, is more commonly known as waveform relaxation or time point relaxation. The ideas and principles behind these techniques are to a large extent based on the same as those of modular integration. The differences in these paradigms are mainly to be found in the problems which they aim to resolve. Modular integration has been extensively used in simulation of large and complex dynamical systems where subsystems or modules arise in a natural way, for instance from a physical perspective. It might then be desirable to simulate each module separately and to handle the interactions between modules as forcing terms or input/output flow. The specification of the interface of a module needs only to include information at the end of each time window, and only those variables which affect the other modules. The local integrator for a module can be chosen according to the properties of that particular subsystem. Many of the recent papers in waveform and time point relaxation are concerned with the potential of parallelism when such methods are applied to large systems of equations. It has also been prevalent in these papers to assume that the same method is used for all subsystems. Some authors have considered partitioned integration methods where the system is split into a stiff and a non-stiff part to which different methods can be applied. For an account of these strategies, see Enright and Kamel [2], Weiner et al. [5] and the references therein.

From the numerical analysts point of view, it is important to consider how the choice of methods and their implementation can be optimized with respect to accuracy and stability. The approach of this paper, is based on the assumption that the local integrators are given, and it is our aim to investigate how various strategies for handling the interaction between the modules affect the overall stability of the resultant dynamic simulation. We shall confine the scope of the discussion to the strategy of time point relaxation with Runge-Kutta methods as introduced by Lie and Skålin [4]. An investigation of stability of time point relaxation with identical Runge-Kutta methods applied to the test equation $u' = \lambda u - \mu v$, $v' = \mu y + \lambda v$, $\lambda, \mu \in \mathbb{R}$ can be found in Bellen et al. [1]. Consider instead the general system

$$(1) \quad \begin{bmatrix} u' \\ v' \end{bmatrix} = \begin{bmatrix} m_{11} & m_{12} \\ m_{21} & m_{22} \end{bmatrix} \cdot \begin{bmatrix} u \\ v \end{bmatrix}$$

One idea is to break the couplings by integrating two subsystems $u' = m_{11}u + g_1$ and $v' = m_{22}v + g_2$ where g_1 and g_2 are obtained from a previous iteration or a previous step. We will always assume that $m_{11} < 0$ and $m_{22} < 0$ such that with constant forcing terms g_1 and g_2 the solution of each subsystem will be stable, or more precisely, tend to the constants $-g_i/m_{ii}$, $i = 1, 2$ for u_n, v_n respectively. The solution of (1) tends to zero if and only if the eigenvalues of the coefficient matrix have negative real parts. Thus, any similarity transform of this matrix will result in a system with the same asymptotic stability properties. For many modular methods, their stability properties only depend on m_{11} , m_{22} , and the quantity $\frac{m_{12}m_{21}}{m_{11}m_{22}}$. In this paper we shall put $\xi = m_{11}$, $\eta = m_{22}$ and $\gamma = \frac{m_{12}m_{21}}{m_{11}m_{22}}$

and then apply a similarity transform to (1) given by the matrix $\text{diag}(\frac{m_{12}}{m_{11}}, 1)$ to obtain the system

$$(2) \quad \begin{bmatrix} u' \\ v' \end{bmatrix} = \begin{bmatrix} \xi & \xi \\ \gamma\eta & \eta \end{bmatrix} \cdot \begin{bmatrix} u \\ v \end{bmatrix}$$

To avoid unnecessary use of notation, the symbols u and v are reinstated in (2). We need to impose the condition $\gamma < 1$ to ensure that the eigenvalues of this new system have negative real parts. Notice that (2) allows for subsystems with different time constants as opposed to the model equations used in [1]. This is important in investigating the use of modular integration on problems where the subsystems have a varying degree of stiffness.

In [1] a more advanced strategy is used for exchanging information between subsystems. Instead of passing only one number per subsystem in each exchange, their approach is based on interpolating stage values from a previous iteration. In doing this, they obtain an order of accuracy which is consistent with the order of each method, provided a sufficient number of iterations are performed in each step. They analyze both the Gauss-Jacobi and Gauss-Seidel algorithm for exchange of information.

Although we will mainly use the first strategy, in Section 2, we shall introduce the concept *coupling stability* or C_α -stability in a way so as to be applicable also when the methods of Bellen et al.[1] are adapted to (2). This new stability definition is useful for comparing the stability properties of the modular method with the methods for each subsystem, i.e. to what extent is the coupling strategy affecting the overall stability of the modular method. We also consider in particular the case where we use only one iteration (one exchange

of information between subsystems) per step when two, possibly different Runge-Kutta methods are applied to each subsystem. In Section 3, we consider k iterations with the so called θ -methods which can be seen as a subclass of the Runge-Kutta methods.

2. ONE ITERATION WITH GENERAL RK METHODS

A Runge-Kutta method is generally defined for initial value problems which, in their autonomous form can be written as

$$(3) \quad y' = f(y), \quad t \geq t_0, \quad y(t_0) = y_0$$

where $y_0, y(t) \in \mathbb{R}^N$ and $f : \mathbb{R}^N \rightarrow \mathbb{R}^N$ is a nonlinear vector function. The Runge-Kutta method can then be defined by the formulas

$$\begin{aligned} Y_r &= y_n + h \sum_{j=1}^s a_{rj} f(Y_j), \quad r = 1, \dots, s, \\ y_{n+1} &= y_n + h \sum_{r=1}^s b_r f(Y_r). \end{aligned}$$

Here h is the stepsize used in the integration. We denote by A the $s \times s$ matrix whose rj -element is a_{rj} and we let $b = (b_1, \dots, b_s)^T$. When a Runge-Kutta method is applied to the system $y' = My$, $M \in \mathbb{R}^{N \times N}$ one obtains the formula

$$y_{n+1} = R(hM)y_n$$

where the rational function $R(z) = 1 + zb^T(I - zA)^{-1}e$ and $e = (1, \dots, 1)^T$. Clearly, as n increases, the sequence y_n tend to zero if and only if the eigenvalues of $R(hM)$ are in the unit disk. When two Runge-Kutta methods are applied to (2) in a modular way, it is desirable to measure which values of each stability function R_1 and R_2 that cause the approximations u_n and v_n to tend to zero as n increases. If there had been no coupling between the two equations in (2), say $u' = \xi u + c_1$ and $v' = \eta v + c_2$ where c_1, c_2 are real constants, the numerical approximations u_n and v_n would agree asymptotically with the exact solution, satisfying $\lim_{n \rightarrow \infty} u_n = -c_1/\xi$ and $\lim_{n \rightarrow \infty} v_n = -c_2/\eta$ if and only if $|R_1(h\xi)| < 1$ and $|R_2(h\eta)| < 1$. These uncoupled equations are actually special cases of the test equation $y'(t) = \lambda y(t) + g(t)$ as discussed by Zennaro [6]. Letting $\lambda < 0$ and $g(t)$ be a real continuous function, Zennaro calls the Runge-Kutta method $A_f(0)$ -stable if

$$(4) \quad |y_{n+1}| \leq \max \left\{ |y_n|; \max_{1 \leq i \leq s} \frac{|g(t_n + c_i h)|}{-\operatorname{Re}(\lambda)} \right\} \quad \text{for all } \lambda < 0.$$

Similarly, he defines the region of $A_f(0)$ -stability to be the maximum segment $[-r, 0]$ such that (4) holds whenever $-r < \lambda h < 0$. The definition is later generalized to include continuous Runge-Kutta (CRK) methods, and results are given in the case that the CRK method is a linear interpolant.

To proceed, we shall think of R_1 and R_2 as coordinate axes in \mathbb{R}^2 . One could say that the above uncoupled system is stable for all values of $h\xi$ and $h\eta$ such that with $r_1 = R_1(h\xi)$ and $r_2 = R_2(h\eta)$, the point $(r_1, r_2) \in (-1, 1) \times (-1, 1)$. Since we shall require both ξ and η to be negative, it will suffice to consider only the rectangular subset of this square consisting of pairs $(r_1, r_2) \in R_1(\mathbb{R}^-) \times R_2(\mathbb{R}^-)$. Thus, to quantify the loss of stability due to the decoupling of (2) by means of a modular integration technique, we need to characterize the subset of the above rectangle for which the resultant numerical approximations u_n and v_n tend to zero. Naturally, the asymptotic behaviour of u_n and v_n also depends on the value of the coupling parameter γ in (2). Since modular integration is likely to work

best when the size of the coupling is small, we shall also allow for a restriction on the range of γ . Thus, we introduce another parameter $\alpha \geq 0$ and for a fixed value of α , we consider the stability properties of the modular method applied to (2) when $\gamma \in (-\alpha, 1)$ as we recall that $\gamma < 1$ is necessary for the exact solution of the overall system (2) to be asymptotically stable. For any subset $M \subseteq \mathbb{R}$ and rational function $R(z)$, we let $R^{-1}(M) = \{z : R(z) = r \text{ for some } r \in M\}$. We are now ready to give the following definition of coupling stability.

Definition 1. Let $r_1 \in R_1(\mathbb{R}^-) \cap (-1, 1)$, $r_2 \in R_2(\mathbb{R}^-) \cap (-1, 1)$ and $\alpha \geq 0$ be given. We shall say that the pair of methods (M_1, M_2) with stability functions R_1 and R_2 is C_α -stable at (r_1, r_2) if when applied to (2)

$$\lim_{n \rightarrow \infty} u_n = \lim_{n \rightarrow \infty} v_n = 0 \quad \forall (h\xi, h\eta) \in R_1^{-1}(r_1) \times R_2^{-1}(r_2) \cap \mathbb{R}^- \times \mathbb{R}^- \text{ and } -\alpha < \gamma < 1.$$

Similarly, we define the region of C_α stability to be the set

$$S_\alpha = \{(r_1, r_2) : (M_1, M_2) \text{ is } C_\alpha\text{-stable at } (r_1, r_2)\}$$

Finally, we shall say that the pair of methods is AC_α -stable if

$$S_\alpha \supseteq (R_1(\mathbb{R}^-) \times R_2(\mathbb{R}^-)) \cap ((-1, 1) \times (-1, 1)).$$

These stability definitions should not be confused with the conventional ones associated with the scalar test equation $y' = \lambda y$. It might at first seem unnatural that the stability regions are expressed in terms of (r_1, r_2) instead of $(h\xi, h\eta)$. However, in studying modular integration methods we want to measure their merits *relative* to each of the methods involved. We are interested in the extent to which stability is lost due to the strategy used for handling input/output between the modules. The stability region as suggested here can be transformed to the $(h\xi, h\eta)$ plane by applying the inverse of the stability functions R_1 and R_2 to each point (r_1, r_2) in the C_α region of stability.

Notice that, for instance, it is possible for a modular method consisting of two explicit Runge-Kutta methods to be AC_α -stable.

From the definition we see that $\alpha_1 > \alpha_2$ implies $S_{\alpha_1} \subseteq S_{\alpha_2}$.

For what remains of this section, we consider the modular method obtained by applying two Runge-Kutta methods separately to the equations

$$(5) \quad \begin{aligned} u' &= \xi u + \xi v_n \\ v' &= \eta v + \eta\gamma((1-q)u_n + qu_{n+1}) \end{aligned}$$

in each step. Thus, if $q = 0$ the two methods can be applied independently (Gauss-Jacobi) while if $q \neq 0$ the methods must be applied one after the other (Gauss-Seidel). We list some stability properties for these methods.

Theorem 2. Consider the modular method obtained by applying two, possibly different Runge-Kutta methods to (5). Let $S_m \subset [-1, 1] \times [-1, 1]$ be the closed region confined by the lines $r_1 = 1$, $r_2 = 1$, $r_2 = -r_1$. Then we have

(a): If $q = 0$ (Gauss-Jacobi), then $S_\alpha \subseteq S_m$ for all $\alpha \geq 0$.

(b): If $0 \leq q_1 \leq q_2 \leq \frac{1}{2}$ then $S_{\alpha, q_1} \subseteq S_{\alpha, q_2}$

(c): If

$$\frac{\alpha + 1}{\alpha} - \frac{\sqrt{\alpha + 1}}{\alpha} \leq q_1 \leq q_2$$

then $S_{\alpha, q_2} \subseteq S_{\alpha, q_1}$ for all $\alpha > 0$.

(d): For any q we have

$$\bigcap_{\alpha > 0} S_\alpha = \emptyset$$

Proof. The resulting discrete formulas are

$$(6) \quad \begin{bmatrix} u_{n+1} \\ v_{n+1} \end{bmatrix} = \begin{bmatrix} r_1 & -(1-r_1) \\ -\gamma(1-r_2)(1-q(1-r_1)) & r_2 + \gamma q(1-r_1)(1-r_2) \end{bmatrix} \cdot \begin{bmatrix} u_n \\ v_n \end{bmatrix}$$

where $r_1 = R_1(h\xi)$ and $r_2 = R_2(h\eta)$. We need to consider the conditions for which the spectral radius of the matrix in (6) is less than one. We compute the characteristic polynomial

$$(7) \quad p(\lambda) = \lambda^2 - (r_1 + r_2 + \gamma q(1-r_1)(1-r_2))\lambda + r_1 r_2 - \gamma(1-r_1)(1-r_2)(1-q)$$

The Routh-Hurwitz criterion yields the conditions

$$(g_1) \quad g_1(r_1, r_2, \gamma) := (1-r_1)(1-r_2)(1-\gamma) > 0$$

$$(8) \quad (g_2) \quad g_2(r_1, r_2, q, \gamma) := (1+r_1)(1+r_2) + \gamma(1-r_1)(1-r_2)(2q-1) > 0$$

$$(g_3) \quad g_3(r_1, r_2, q, \gamma) := 1 - r_1 r_2 + \gamma(1-r_1)(1-r_2)(1-q) > 0$$

Since $\gamma < 1$ by assumption, clearly (g_1) is satisfied for any h . We prove (a) by showing that if $(r_1, r_2) \notin S_m$, then (g_2) fails to hold. To this end, we rewrite (g_2) with $q = 0$ into the form

$$\gamma < 1 + 2 \frac{r_1 + r_2}{(1-r_1)(1-r_2)}$$

If this is to hold for γ near 1, it is necessary that $r_1 + r_2 \geq 0$ so for any $\alpha > 0$, $S_\alpha \subseteq S_m$.

To prove (b), let $0 \leq q_1 \leq q_2 \leq 1/2$ and $\alpha \geq 0$ be given and assume that (8) holds for $r_1, r_2 \in (-1, 1)$, q_1 and $\gamma \in (-\alpha, 1)$. We prove that it holds for r_1, r_2, q_2, α . Because (g_2) and (g_3) both are linear in γ , it suffices to check that they hold for $\gamma = 1$ in (g_2) and $\gamma = -\alpha$ in (g_3) . We have

$$g_2(r_1, r_2, q_2, 1) - g_2(r_1, r_2, q_1, 1) = 2(1-r_1)(1-r_2)(q_2 - q_1) \geq 0$$

and

$$g_3(r_1, r_2, q_2, -\alpha) - g_3(r_1, r_2, q_1, -\alpha) = \alpha(1-r_1)(1-r_2)(q_2 - q_1)$$

so these arbitrarily chosen $r_1, r_2 \in S_{\alpha, q_1}$ also belong to S_{α, q_2} .

(c) is proved in a similar way. Observe first that the lower bound on $l(\alpha) := (\alpha + 1 - \sqrt{\alpha + 1})/\alpha$ increases monotonically from $1/2$ to 1 as α varies from 0 to ∞ . Assume for a moment that f_1 and f_2 are positive numbers, and that $f_1 f_2 > (2q - 1)\alpha$ with $q \geq l(\alpha)$. Then $(f_1 - f_2)^2 = (f_1 + f_2)^2 - 4f_1 f_2 \geq 0$. Thus

$$(f_1 + f_2)^2 \geq 4f_1 f_2 > 4(2q - 1)\alpha \geq 4(\alpha + 2 - 2\sqrt{\alpha + 1}) = (2(\sqrt{\alpha + 1} - 1))^2 \geq (2(1 - q)\alpha)^2$$

so that $f_1 + f_2 > 2(1 - q)\alpha$. Considering (g_2) with $q \geq 1/2$ and (g_3) , we get the inequalities

$$f_1 f_2 > (2q - 1)\alpha, \quad f_1 + f_2 > 2(1 - q)\alpha \quad \text{where} \quad f_i = \frac{1 + r_i}{1 - r_i}, \quad i = 1, 2$$

From the above it follows that the second inequality dominates when $q \geq l(\alpha)$. As in the proof of (b), we obtain $(r_1, r_2) \in S_{\alpha, q_2}$ implies $(r_1, r_2) \in S_{\alpha, q_1}$ if $q_2 \geq q_1$.

To prove part (d) pick any $(r_1, r_2) \in (-1, 1) \times (-1, 1)$ and observe that if $q < 1$, (g_3) is violated for γ near $-\alpha$ if $\alpha > (1 - r_1 r_2) / ((1 - r_1)(1 - r_2)(1 - q))$. Similarly, if $q > 1/2$, (g_2) fails to hold for values of γ near $-\alpha$ if $\alpha > (1 + r_1)(1 + r_2) / ((1 - r_1)(1 - r_2)(2q - 1))$. \square

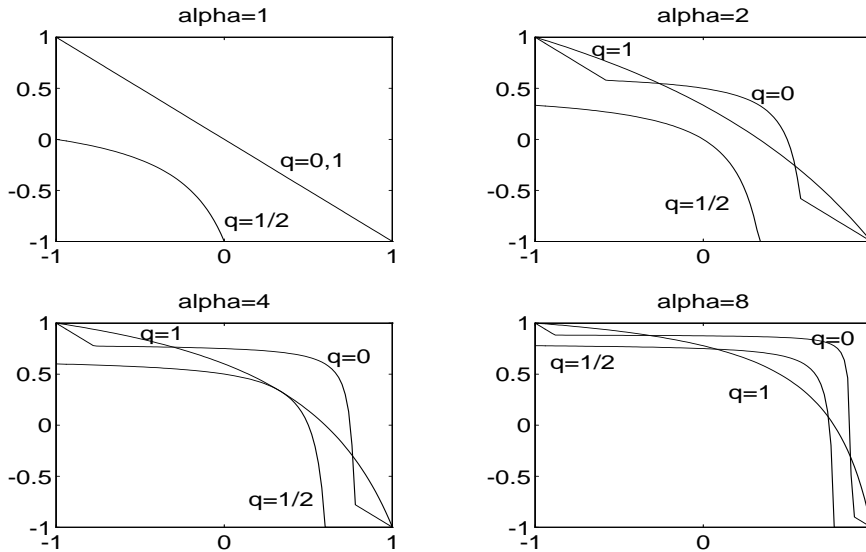


FIGURE 1. C_α regions of stability

Figure 1 shows the C_α stability region for four different values of α . In each subplot the regions corresponding to the $q = 0, 1/2, 1$ are the areas above the legended curves.

3. ITERATIONS WITH θ -METHODS

In the previous section, we applied only one iteration with two general Runge-Kutta methods. In this section, we shall consider the θ -methods which applied to the initial value problem (3) can be defined by the formula

$$(9) \quad y_{n+1} = y_n + h(1 - \theta)f(y_n) + h\theta f(y_{n+1})$$

where θ is typically chosen in the interval $[0, 1]$. Notice that the choices $\theta = 0, 1, 1/2$ results in the well-known methods Euler, Backward Euler and Trapezoidal Rule respectively. We shall investigate the case when two θ -methods (i.e. two possibly different choices of θ in (9)) are combined to constitute a modular method which is applied to (2). We allow several iterations, i.e. we consider an iteration scheme of the form

$$(10) \quad \begin{aligned} u_{n+1}^{(k)} &= u_n + h\xi \left((1 - \theta_1)u_n + \theta_1 u_{n+1}^{(k)} \right) + h\xi \left((1 - \theta_1)v_n + \theta_1 v_{n+1}^{(k-1)} \right) \\ v_{n+1}^{(k)} &= v_n + h\eta\gamma \left((1 - \theta_2)u_n + \theta_2 \left((1 - q)u_{n+1}^{(k-1)} + qu_{n+1}^{(k)} \right) \right) + h\eta \left((1 - \theta_2)v_n + \theta_2 v_{n+1}^{(k)} \right) \end{aligned}$$

for $k \geq 1$ where $u_{n+1}^{(0)} = u_n$ and $v_{n+1}^{(0)} = v_n$. Like in Section 2 we see that $q = 0$ corresponds to a Gauss-Jacobi type of iteration, while when $q \neq 0$ it is intended that the formulas are applied in sequence.

Let us first recall that the stability function of a θ -method is given as

$$(11) \quad R(z) = \frac{1 + (1 - \theta)z}{1 - \theta z}$$

And observe that for $\theta > 0$ it is bounded and monotone on $(-\infty, 0)$, $1 - 1/\theta < R(z) < 1$, thus the inverse function exists. Let us rewrite (10) in matrix form, where we use $r_1 = R_1(h\xi)$ and $r_2 = R_2(h\eta)$. We get a system of the form

$$(12) \quad V \cdot \begin{bmatrix} u_{n+1}^{(k)} \\ v_{n+1}^{(k)} \end{bmatrix} = U \cdot \begin{bmatrix} u_n \\ v_n \end{bmatrix} + W \cdot \begin{bmatrix} u_{n+1}^{(k-1)} \\ v_{n+1}^{(k-1)} \end{bmatrix}$$

where the 2×2 matrices V, U and W are given as

$$V = \begin{bmatrix} 1 & 0 \\ (1 - r_2)\gamma\theta_2q & 1 \end{bmatrix}, \quad U = \begin{bmatrix} r_1 & -(1 - \theta_1)(1 - r_1) \\ -\gamma(1 - r_2)(1 - \theta_2) & r_2 \end{bmatrix}$$

and

$$W = \begin{bmatrix} 0 & -\theta_1(1 - r_1) \\ -(1 - r_2)(1 - q)\theta_2\gamma & 0 \end{bmatrix}$$

Thus, we have

$$\begin{bmatrix} u_{n+1}^{(k)} \\ v_{n+1}^{(k)} \end{bmatrix} = B_k \cdot \begin{bmatrix} u_n \\ v_n \end{bmatrix} \quad \text{where} \quad B_k = V^{-1}U + V^{-1}WB_{k-1}.$$

If $B_\infty := \lim_{k \rightarrow \infty} B_k$ exists, we have $B_k - B_\infty = V^{-1}W(B_{k-1} - B_\infty)$, thus convergence is equivalent to the spectral radius of $V^{-1}W$ being less than one. By computing the characteristic polynomial of this matrix and by using the Routh-Hurwitz criterion we obtain

Proposition 3. *Let $\beta := \gamma\theta_1\theta_2(1 - r_1)(1 - r_2)$. The iteration (10) converges to*

$$(13) \quad \begin{bmatrix} u_{n+1}^{(\infty)} \\ v_{n+1}^{(\infty)} \end{bmatrix} = (I - V^{-1}W)^{-1}V^{-1}U \cdot \begin{bmatrix} u_n \\ v_n \end{bmatrix}$$

if and only if

$$\beta \in \left(-\frac{1}{1 - q}, 1 \right) \quad \text{for } 0 \leq q \leq \frac{2}{3},$$

$$\beta \in \left(-\frac{1}{2q - 1}, 1 \right) \quad \text{for } \frac{2}{3} < q \leq 1.$$

Hence, the largest interval of convergence for β is achieved with $q = 2/3$ for which the scheme converges if and only if $\beta \in (-3, 1)$.

Observe that for the θ -methods, we have

$$(1 - R(z))\theta = -\frac{\theta z}{1 - \theta z}$$

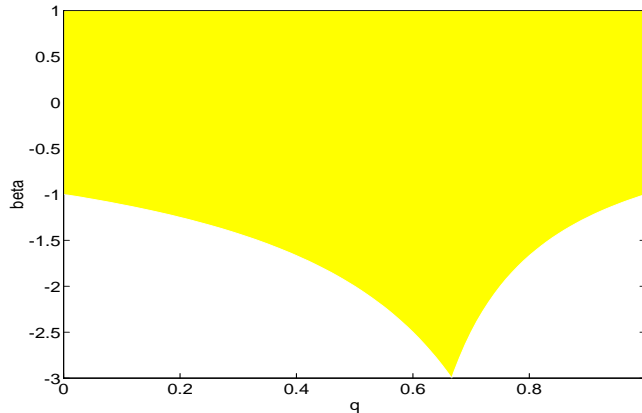


FIGURE 2. Interval of convergence in terms of q

which decreases monotonically from 1 to 0 for $z \in (-\infty, 0)$. Thus β takes values between 0 and γ for all pairs $(h\xi, h\eta) \in \mathbb{R}^- \times \mathbb{R}^-$.

To analyze the stability of the limit method (13), one can use a similar approach to arrive at

Proposition 4. *The limit method (13) is AC_α stable for all $\alpha \geq 0$ if and only if $\theta_1 \geq 1/2$ and $\theta_2 \geq 1/2$. Moreover, there is no $\alpha \geq 0$ such that with either $\theta_1 < 1/2$ or $\theta_2 < 1/2$ the resultant limit method is AC_α -stable.*

It should be noted that this strong stability result for $\theta_1, \theta_2 \geq 1/2$ cannot be exploited in full by using the iteration (10) to arrive at the limit method, since the condition for convergence is by far the more restrictive. Observe also that the above result with $\theta = \theta_1 = \theta_2$ agrees with the well-known result that θ -methods are A -stable if and only if $\theta \geq 1/2$.

We shall consider stability of (10) for $q = 0$ and $q = 1$ using a fixed number of iterations in each step. We have the following result

Theorem 5. *Assume that $\theta_1 \geq 1/2$, $\theta_2 \geq 1/2$, $q = 0$ and $0 \leq \alpha \leq 1$ are given.*

- (a): *The method (10) with k iterations is AC_1 -stable if $k = 4m$ where m is any positive integer.*
- (b): *Let $\theta = \min\{\theta_1, \theta_2\}$. The method with $k = 4m - 2$ iterations is AC_α -stable for $0 < \alpha < 1$ if $\theta > 1/2$ and*

$$2m - 1 > \frac{\ln(2\theta - 1)}{\ln \alpha}$$

- (c): *The method (10) with k an odd number of iterations is not AC_α stable for any $\alpha \geq 0$ unless $\theta_1 = \theta_2 = 1$.*

Theorem 6. *Assume that $\theta_1 \geq 1/2$, $\theta_2 \geq 1/2$, $q = 1$ and $0 \leq \alpha \leq 1$ are given. Then the method (10) with k iterations is AC_α -stable for all even k . If $\theta_1 > 1/2$, $\theta_2 \geq 1/2$, $0 \leq \alpha < 1$ and $q = 1$ then the method with k iterations is AC_α -stable for all*

$$k > \frac{\ln(2\theta_1 - 1)}{\ln \alpha}$$

The condition that $\alpha \leq 1$ in the above theorems is imposed to ensure that the scheme converges for any $(h\xi, h\eta) \in \mathbb{R}^- \times \mathbb{R}^-$.

Notice the resemblance between the case $k = 4m - 2$ for $q = 0$ and k odd for $q = 1$. The formula indicates that in these cases one needs twice as many iterations in the $q = 0$ case to obtain AC_α stability as in the $q = 1$ case. However, in the latter case, the condition is imposed on θ_1 while the $q = 0$ case is still more restrictive since the condition involves $\min\{\theta_1, \theta_2\}$.

To ease the proofs of the above theorems, we shall first prove the following lemma

Lemma 7. *Let $\theta_1 \geq 1/2$, $\theta_2 \geq 1/2$ and $\gamma < 1$ be given and put $\theta = \min\{\theta_1, \theta_2\}$. Then the eigenvalues of B_∞ are contained in the disk $\{z : |z - (1 - \frac{1}{2\theta})| < \frac{1}{2\theta}\}$.*

Proof. Compute the characteristic polynomial $p(\lambda)$ of B_∞ . Assume for instance that $\theta = \theta_1 \leq \theta_2$, and set $\hat{p}(\lambda) = p(2\theta\lambda - (2\theta - 1))$. Then use the Routh-Hurwitz conditions to assert that the roots of $\hat{p}(\lambda)$ lie within the unit circle. \square

Proof of Theorem 5. With $q = 0$, $V = I$ so $V^{-1}W = W$. Thus, $W^2 = \beta I$ where $\beta = (1 - r_1)(1 - r_2)\theta_1\theta_2\gamma$. Recall that $-1 \leq -\alpha < \gamma < 1$, so $\beta \in (-\alpha, 1) \subseteq (-1, 1)$. Hence $B_\infty = (I - W)^{-1}U = (1 - \beta)^{-1}(I + W)U$, and

$$B_k = \begin{cases} \beta^{k/2}I + (1 - \beta^{k/2})B_\infty, & \text{for } k \text{ even} \\ \beta^{(k-1)/2}(U + W) + (1 - \beta^{(k-1)/2})B_\infty, & \text{for } k \text{ odd} \end{cases}$$

Letting λ_k and λ_∞ be eigenvalues of B_k and B_∞ , we get for even k

$$(14) \quad \lambda_k = \beta^{k/2} + (1 - \beta^{k/2})\lambda_\infty$$

In the case that $k = 4m$, we have $0 \leq \beta^{2m} < 1$ and

$$|\lambda_{4m}| = |\beta^{2m} + (1 - \beta^{2m})\lambda_\infty| \leq \beta^{2m} + (1 - \beta^{2m})|\lambda_\infty| < 1$$

since $|\lambda_\infty| < 1$ by Proposition 4. This proves the part **(a)** of Theorem 5.

Now, assume that $k = 4m - 2$ and set $\kappa = \beta^{2m-1}$. If $\gamma \geq 0$ then $\kappa \geq 0$ and **(b)** follows as in the previous case. Assume instead that $\gamma < 0$ i.e. $\kappa < 0$. From (14) we get

$$\begin{aligned} |\lambda_k| &= |\kappa + (1 - \kappa)\lambda_\infty| = |1 - \frac{1}{2\theta} + \frac{\kappa}{2\theta} + (1 - \kappa)(\lambda_\infty - (1 - \frac{1}{2\theta}))| \\ &< \frac{1}{2\theta}(1 - \kappa) + |1 - \frac{1}{2\theta}(1 - \kappa)| \end{aligned}$$

where we have used Lemma 7 and the triangle inequality. Thus, if $-\kappa \leq 2\theta - 1$ then $|\lambda_k| < 1$. So **(b)** follows by observing that $-\kappa < (-\gamma)^{2m-1} < \alpha^{2m-1}$.

To prove **(c)**, observe first from (11) that $\theta(1 - R(z)) = \frac{\theta z}{1 - \theta z}$. Since $\lim_{z \rightarrow -\infty} \frac{\theta z}{1 - \theta z} = 1$, r_1 and r_2 can be chosen such that $d := \theta_1(1 - r_1) = \theta_2(1 - r_2)$ is arbitrarily close to 1. It is then possible, by choosing γ sufficiently close to 1, to make $\kappa := \beta^{m-1} = (\alpha d^2)^{m-1}$ arbitrarily close to 1. Recall that

$$B_{2m-1} = \kappa(U + W) + (1 - \kappa)B_\infty$$

By continuity, one can for any $\epsilon > 0$ find $\delta_1 > 0$ such that $|\lambda(B_{2m-1}) - \lambda(U + W)| < \epsilon$ for each d and γ such that $\max\{1 - d, 1 - \gamma\} < \delta_1$. Let M be the matrix obtained by setting $\gamma = 1$, $d = 1$ in $U + W$. M has eigenvalues $\lambda_1 = 1$ and $\lambda_2 = 1 - 1/\theta_1 - 1/\theta_2$. Again by continuity, it follows that δ_2 can be found such that $|\lambda(U + W) - \lambda_2| < \epsilon$ if $\max\{1 - d, 1 - \gamma\} < \delta_2$. Hence

$$|\lambda(B_{2m-1}) - \lambda_2| \leq |\lambda(B_{2m-1} - \lambda(U + W))| + |\lambda(U + W) - \lambda_2| < 2\epsilon$$

for $\max\{1 - d, 1 - \gamma\} < \max\{\delta_1, \delta_2\}$. So the result follows by choosing $0 < \epsilon < (-1 - \lambda_2)/2 = (1/\theta_1 + 1/\theta_2 - 2)/2$. \square

Proof of Thm 6. In the case that $q = 1$, we obtain another simple version of the matrix $V^{-1}W$.

$$\tilde{W} := V^{-1}W = \begin{bmatrix} 0 & -\theta_1(1 - r_1) \\ 0 & \beta \end{bmatrix}$$

with $\beta = (1 - r_1)(1 - r_2)\theta_1\theta_2\gamma$. Hence, now $\tilde{W}^k = \beta^{k-1}\tilde{W}$ and we obtain

$$B_k = \beta^{k-1}(\tilde{U} + \tilde{W}) + (1 - \beta^{k-1})B_\infty$$

where $\tilde{U} = V^{-1}U$. Letting $T := \text{trace}(B_k)$ and $D := \det(B_k)$ we must ensure the positivity of the quantities $1 - T + D$, $1 + T + D$ and $1 - D$. Some calculations show that

$$1 - T + D = \frac{1 - \beta^k}{1 - \beta}(1 - r_1)(1 - r_2)(1 - \gamma)$$

which is positive for all $r_1, r_2, \gamma, |\beta|$ which are less than unity. For the second condition we compute

$$1 + T + D = \frac{1 - \beta^k}{1 - \beta} \left((1 + r_1)(1 + r_2) - \frac{(2\theta_1 - 1)(2\theta_2 - 1)}{\theta_1\theta_2} \right) + \frac{(2\theta_2 - 1)(2\theta_1 - 1 + \beta^k)}{\theta_1\theta_2}.$$

Since $r_i \in (1 - 1/\theta_i, 1)$, $i = 1, 2$ it is sufficient to impose

$$(2\theta_2 - 1)(2\theta_1 - 1 + \beta^k) \geq 0$$

This condition is satisfied for each $\beta \in (-1, 1)$ if k is even and for all $\beta \in [0, 1)$ if k is odd. Notice that it also holds unconditionally if $\theta_2 = 1/2$. If $\beta < 0$ (corresponding to $\gamma < 0$), k is odd and $\theta_2 > 1/2$ we obtain the condition

$$(15) \quad \alpha^k \leq 2\theta_1 - 1$$

again since $\beta \in (\gamma, 0)$ and $\gamma \in (-\alpha, 1)$.

We proceed by considering $1 - D$

$$1 - D = \frac{1 - \beta^k}{1 - \beta} \left(1 - r_1r_2 - \frac{\theta_1 + \theta_2 - 1}{\theta_1\theta_2} \right) + \frac{\theta_1 + \theta_2 - 1}{\theta_1\theta_2} + \frac{\beta^k}{\theta_1\theta_2}(1 - \theta_2)$$

Again by considering the range of r_1 and r_2 we conclude that the critical case is when r_1 and r_2 are both negative in which case the first part of the above expression is positive and we are left with the condition

$$\theta_1 + \theta_2 - 1 + \beta^k(1 - \theta_2) \geq 0$$

As in the previous condition, this is satisfied for any k if $\beta \in [0, 1)$ and for even k if $\beta \in (-1, 1)$. If $\beta < 0$ and k is odd, then we must have

$$(16) \quad \alpha^k \leq \frac{\theta_1 + \theta_2 - 1}{1 - \theta_2}$$

By comparing (15) with (16) we find that (15) is the critical inequality whenever $\theta_1(1 - 2\theta_2) > 0$. Thus, it is only necessary to consider (16) when $\theta_2 = 1/2$, in which case it takes the form $\alpha^k \leq 2\theta_1 - 1$. In conclusion, the method with k iterations and $q = 1$ is AC_1 -stable for all even k , and for odd k it is AC_α stable whenever $\alpha^k \leq 2\theta_1 - 1$. \square

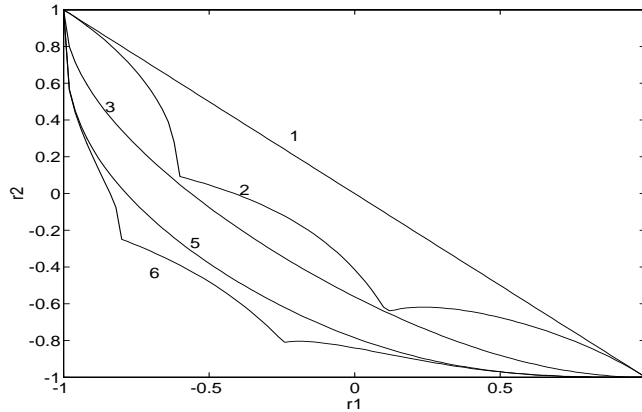


FIGURE 3. C_1 -stability regions for $\theta_1 = \theta_2 = 1/2$, $q = 0$, $k = 1, 2, 3, 5, 6$.

Figure 3 shows the C_1 regions of stability when the trapezoidal rule is used for both methods ($\theta_1 = \theta_2 = 1/2$) and the iteration is done in a Gauss-Jacobi fashion. The legended curves show the stability regions for $k = 1, 2, 3, 5, 6$ where k is the number of iterations. The region is the area in $(-1, 1) \times (-1, 1)$ above the curve.

4. CONCLUDING REMARKS

By introducing a *relative* stability concept, we have tried to quantify the loss of stability due to the breaking of the couplings by modular integration in a two by two linear system of ODEs. Our setting is more general than [1] as our test system include their system as a special case. This generalization allows for subsystems with varying degree of stiffness. However, in this paper we have dealt with modular methods which are less advanced than those in [1]. However, the basic definition, that of coupling stability can be generalized to account for the methods based on interpolating stage values as in [1]. We believe that there is a potential for studying the methods of Bellen et al. in this new framework, and that will be the subject of forthcoming papers.

Acknowledgement. I would like to thank Dr Harald H. Simonsen, Dr Anne Kværnø and Prof. Syvert P. Nørsett for several fruitful discussions and for commenting on an early version of this manuscript.

REFERENCES

1. A. Bellen, Z. Jackiewicz, and M. Zennaro. Time-point relaxation Runge-Kutta methods for ordinary differential equations. *J. of Comp. Appl Mathematics*, 45:121–137, 1993.
2. W.H. Enright and M. Kamel. Automatic partitioning of stiff systems and exploiting the resulting structure. *ACM-TOMS*, 5:374–385, 1979.
3. Torleif Iversen. Modular Integration – An Overview. Technical Report STF48 A93022, SINTEF Automatic Control, May 1993.
4. I. Lie and R. Skålin. Relaxation-based Integration by Runge-Kutta Methods and its Applications to the Moving Finite Element Method. In J.R. Cash and I. Gladwell, editors, *Computational ordinary differential equations*, pages 357–368, 1992.
5. R. Weiner, M. Arnold, P. Rentrop, and K. Strehmel. Partitioning strategies in Runge-Kutta type methods. *IMA J. of Num. An.*, 13:303–319, 1993.
6. M. Zennaro. Contractivity of Runge-Kutta methods with respect to forcing terms. *Appl. Num. Math.*, 10:321–345, 1993.

SINTEF, INDUSTRIAL MATHEMATICS, N-7034 TRONDHEIM, NORWAY

E-mail address: bryn@simasintef.no