SIAM J. NUMER. ANAL.
Vol. 32, No. 6, pp. 1940–1971, December 1995

011

THE ESSENTIAL STABILITY OF LOCAL ERROR CONTROL FOR DYNAMICAL SYSTEMS*

A. M. STUART† AND A. R. HUMPHRIES‡

Abstract. Although most adaptive software for initial value problems is designed with an accuracy requirement—control of the local error—it is frequently observed that stability is imparted by the adaptation. This relationship between local error control and numerical stability is given a firm theoretical underpinning.

The dynamics of numerical methods with local error control are studied for three classes of ordinary differential equations: *dissipative, contractive*, and *gradient* systems. Dissipative dynamical systems are characterised by having a bounded absorbing set $\mathcal{B}$ which all trajectories eventually enter and remain inside. The exponentially contractive problems studied have a unique, globally exponentially attracting equilibrium point and thus they are also dissipative since the absorbing set $\mathcal{B}$ may be chosen to be a ball of arbitrarily small radius around the equilibrium point. The gradient systems studied are those for which the set of equilibria comprises isolated points and all trajectories are bounded so that each trajectory converges to an equilibrium point as $t \to \infty$. If the set of equilibria is bounded then the gradient systems are also dissipative. Conditions under which numerical methods with local error control replicate these large-time dynamical features are described. The results are proved without recourse to asymptotic expansions for the truncation error.

Standard embedded Runge–Kutta pairs are analysed together with several nonstandard error control strategies. Both error per step and error per unit step strategies are considered. Certain embedded pairs are identified for which the sequence generated can be viewed as coming from a small perturbation of an algebraically stable scheme, with the size of the perturbation proportional to the tolerance τ. Such embedded pairs are *defined* to be *essentially algebraically stable* and explicit essentially stable pairs are identified. Conditions on the tolerance τ are identified under which appropriate discrete analogues of the properties of the underlying differential equation may be proved for certain essentially stable embedded pairs. In particular, it is shown that for dissipative problems the discrete dynamical system has an absorbing set $\mathcal{B}_\tau$ and is hence dissipative. For exponentially contractive problems the radius of $\mathcal{B}_\tau$ is proved to be proportional to τ. For gradient systems the numerical solution enters and remains in a small ball about one of the equilibria and the radius of the ball is proportional to τ. Thus the local error control mechanisms confer desirable global properties on the numerical solution. It is shown that for error per unit step strategies the conditions on the tolerance τ are independent of initial data while for error per step strategies the conditions are initial-data dependent. Thus error per unit step strategies are considerably more robust.

Key words. error control, algebraic stability, dissipativity, contractivity, gradient systems

AMS subject classifications. 34C35, 34D05, 65L07, 65L20, 65L50

1. Introduction.

In this paper we consider numerical approximation of the initial value problem

$$u_t = f(u), \quad u(0) = U, \tag{1.1}$$

where $u(t) \in \mathbb{R}^m$ for each $t \geqslant 0$, and $f: \mathbb{R}^m \to \mathbb{R}^m$ is assumed to be locally Lipschitz. We study variable time stepping strategies designed to control the local error incurred

* Received by the editors December 22, 1992; accepted for publication (in revised form) March 15, 1994.

† Scientific Computing and Computational Mathematics Program, Division of Applied Mechanics, Durand 252, Stanford University, Stanford, California 94305-4040 (stuart@sccm.stanford.edu). Supported by Office of Naval Research grant N00014-92-J-1876 and by the National Science Foundation under grant DMS-9201727.

‡ School of Mathematical Sciences, University of Bath, Bath, Avon BA2 7AY, United Kingdom and Scientific Computing and Computational Mathematics Program, Department of Computer Science, Stanford University, Stanford, California 94305-2140 (tony.humphries@bristol.ac.uk). Supported by the United Kingdom Science and Engineering Research Council, Stanford University, and the Office of Naval Research grant N00014-92-J-1876. Current address: School of Mathematics, University of Bristol, University Walk, Bristol BS8 1TW, United Kingdom.

at each step. In particular, our interest lies in the effect of the error control mechanism on the long-time dynamics of the problem (1.1) and in assessing whether, and in what sense, the dynamics are reproduced by the approximation scheme.

Embedded Runge–Kutta pairs are studied. Let t_n denote a sequence of (unequally spaced) grid points in time and let U_n denote an approximation to $u(t_n)$; then the embedded Runge–Kutta pair is defined as follows:

$$\eta_i = U_n + \Delta t_n \sum_{j=1}^{k} a_{ij} f(\eta_j), \quad i = 1, \ldots, k, \tag{1.2}$$

$$U_{n+1} = U_n + \Delta t_n \sum_{j=1}^{k} b_j f(\eta_j), \quad U_0 = U, \tag{1.3}$$

and

$$V_{n+1} = U_n + \Delta t_n \sum_{j=1}^{k} \bar{b}_j f(\eta_j). \tag{1.4}$$

For convenience we set $V_0 = U_0$. The sequence $\{V_n\}_{n=0}^{\infty}$ is introduced only to estimate the error so that the time step may be varied accordingly. The sequence $\{U_n\}_{n=0}^{\infty}$ is considered as the numerical approximation to $u(t)$ and it is the asymptotic features of this sequence that we shall study. The time step Δt_n is chosen so that either

$$\|U_{n+1} - V_{n+1}\| \leqslant \tau \Delta t_n / |e_0| \tag{1.5}$$

or

$$\|U_{n+1} - V_{n+1}\| \leqslant \tau^3 / |e_0|, \tag{1.6}$$

where $\tau \ll 1$ is an error tolerance and e_0 is a scale factor to be specified later. The strategy (1.5) is known as *error per unit step* while the strategy (1.6) is known as *error per step.* (The cubic dependence on τ in (1.6) streamlines the presentation of results and is, of course, simply a matter of definition.)

In the following it will be useful to define the matrix A and vectors $b, \bar{b}$ by

$$\{A\}_{ij} = a_{ij}, \quad b = (b_1, \ldots, b_k)^T, \quad \bar{b} = (\bar{b}_1, \ldots, \bar{b}_k)^T. \tag{1.7}$$

These matrices and vectors are usually chosen so that the difference of U_n and V_n provides an estimate of the error incurred over one step of the numerical method (1.2), (1.3) as is standard for embedded Runge–Kutta pairs [3]. We say that the scheme (1.2), (1.3), (1.4) has order (p, q) if A, b is an order p method and $A, \bar{b}$ is an order q method. In many software codes $q = p + 1$ and $\|U_{n+1} - V_{n+1}\|$ is an estimate of the local truncation error for U_{n+1}. However, $q = p - 1$ is sometimes used in codes so that the solution is advanced using the higher-order method, although the error estimate is only strictly valid for the lower-order scheme. This is known as *local extrapolation.*

In addition to studying these standard methods, we will also introduce some simple schemes with desirable properties where $q = 1$; for these methods the construction of V_{n+1} is computationally inexpensive. Furthermore, we analyse some simple modifications of standard error control strategies which are tailored to given structural assumptions about the differential equations.

To study the effect of local error control on large-time dynamics it is necessary to work with particular structural assumptions on the vector field $f(\bullet)$ which defines the differential equation (1.1). Throughout we assume that $\| \bullet \|$ denotes a norm in $\mathbb{R}^m$ induced by the appropriate inner product, i.e., one inherited from one of the assumptions (D), (C), or (G) which we now introduce. Here, and throughout the remainder of the paper,

$$B(v,\delta) = \{u \in \mathbb{R}^m : \|v-u\| < \delta\} \tag{1.8}$$

and

$$Q(\varepsilon) = \{v \in \mathbb{R}^m : \|f(v)\| \leqslant \varepsilon\}. \tag{1.9}$$

Note that $Q(0)$ comprises the set of equilibria of (1.1). The three conditions on the vector field $f(\bullet)$ which we consider are (D), (C), and (G):

(D) $\exists \alpha \geqslant 0, \beta > 0 : \langle f(u), u\rangle \leqslant \alpha - \beta\|u\|^2,\ \forall\, u \in \mathbb{R}^m$;

(C) $\exists \beta > 0 : \langle f(u) - f(v), u - v\rangle \leqslant -\beta\|u-v\|^2,\ \forall\, u, v \in \mathbb{R}^m$ and $f(0) = 0$;

(G) $\exists k > 0$ such that

- (G1) $f(u) = -\nabla F(u)$, where $F \in C^2(\mathbb{R}^m, \mathbb{R})$;
- (G2) $F(u) \geqslant 0, \forall u \in \mathbb{R}^m$ and $|F(u)| \to \infty$ as $\|u\| \to \infty$;
- (G3) $F(u) - F(v) \leqslant \langle f(u), v-u\rangle + k\|u-v\|^2, \forall\, u, v \in \mathbb{R}^m$;
- (G4) All members of $Q(0)$ are hyperbolic and $\|v\| \leqslant k\ \forall v \in Q(0)$;
- (G5) $\liminf_{\|v\|\to\infty} \|f(v)\| \geqslant k$.

See [24] for a review of the relevance of these classes of problems in numerical analysis and in applications. The report [25] contains similar analysis to that presented here for an additional structural assumption on $f(\bullet)$ weaker than, but strongly related to, (D); the report also contains a slight weakening of (G), which allows an unbounded set of equilibria. We now characterize the behaviour of (1.1) under these different structural assumptions on $f(\bullet)$; the following definition is fundamental.

DEFINITION 1.1. The equation (1.1) is said to be *dissipative* if there exists a bounded *absorbing set* $\mathcal{B} \subset \mathbb{R}^m$ and, for each $U \in \mathbb{R}^m$, a time $t^* = t^*(U)$ such that $u(t) \in \mathcal{B}\ \forall t \geqslant t^*$.

The following properties hold for (1.1).

THEOREM ODE (i) *under* (D), (1.1) *is dissipative with* $\mathcal{B} = \bar{B}(0, (\alpha+\rho)/\beta)$, *for any* $\rho > 0$;

(ii) *under* (C), *every solution of* (1.1) *satisfies* $u(t) \to 0$ *as* $t \to \infty$ *and thus* (1.1) *is dissipative with* $\mathcal{B} = \bar{B}(0,\rho)$, *for any* $\rho > 0$;

(iii) *under* (G), *for every* $U \in \mathbb{R}^m$ $\exists v \in Q(0)$ *such that the solution of* (1.1) *satisfies* $u(t) \to v$ *as* $t \to \infty$ *and thus* (1.1) *is dissipative with* $\mathcal{B} = \{u \in \mathbb{R}^m : F(u) \leqslant \max_{v\in Q(0)} F(v) + \rho\}$, *for any* $\rho > 0$.

Proof. The proof of (i) may be found in [24] and underlies much of the work in [26]. The proof of (ii) is straightforward. The proof of (iii) may be found in [10]. □

Throughout this paper our aim is to derive discrete analogues of Theorem ODE under the weakest possible assumptions on the tolerance τ. Note that, in fixed-step implementation, only implicit methods will replicate the behaviour of the ODE unless the time step is restricted in terms of initial data [24]. Thus it is of interest to derive explicit embedded pairs which yield discrete analogues of Theorem ODE without the tolerance τ being restricted in terms of initial data. The key to our analysis is the observation that, under certain conditions on the underlying Runge–Kutta method, the local error control ensures that the embedded pair is close to an algebraically stable Runge–Kutta scheme; the "closeness" is proportional to the error tolerance.

We call such embedded pairs *essentially algebraically stable* and in §2 we construct *explicit* embedded pairs which are essentially algebraically stable in this sense. Note that algebraically stable schemes are necessarily implicit. In addition, we prove an order barrier $\min\{p, q\} \leqslant 4$ for explicit essentially algebraically stable embedded pairs of order (p, q), with nonnegative b_i.

In §3 we consider the question of whether it is possible to find sequences $\{U_n\}_{n=0}^{\infty}$ and $\{\Delta t_n\}_{n=0}^{\infty}$ such that the error control schemes (1.2)–(1.4), (1.5) or (1.2)–(1.4), (1.6) are satisfied. In particular we determine conditions under which schemes admit sequences satisfying $\inf_{n\geqslant 0} \Delta t_n > 0$ since, without this, the time integration may terminate at a finite time. This simply boils down to proving boundedness of the numerical solution. However, the section is included since our aim is to describe a rigorous framework for the analysis of local error control. Since the details are somewhat technical, the basic idea (which is simple) is described in the text with details contained in the appendix.

It is known that algebraically stable Runge–Kutta methods implemented with a fixed time step define dissipative numerical methods for (D) and (C), respectively, for all $\Delta t > 0$ (see [1], [16], [24]). In §4 we show that essentially algebraically stable error per unit step and error per step embedded pairs also preserve the dissipativity of the underlying system. Under (D) there is an absorbing set $\mathcal{B}_\tau$ centred at the origin and under (C) this set has radius proportional to τ (see Theorems DC1 and DC2 which are discrete analogues of Theorem ODE (i) and (ii)).

For (D) and (C) we consider both error per step and error per unit step strategies. The error per unit step schemes have the advantage that the properties of the underlying differential equation are inherited for τ sufficiently small, but independent of initial data; this means that codes based on such a strategy are extremely robust since they operate effectively given *any* initial data. In contrast, the error per step strategies can only be guaranteed to mimic the differential equation if τ is bounded above in terms of the initial data U. In situations where a number of simulations of a system are made for a variety of initial conditions, it is extremely desirable to have codes which will operate robustly with respect to changes in the initial data. For this reason, codes which replicate the essential features of a problem for initial-data independent ranges of τ are useful.

In §5 we study gradient systems under (G). We consider only error per unit step strategies although it is straightforward to generalise the results to the error per step case as is done in §4. The assumptions (G1)–(G3) have the following interpretations: (G1) is the standard gradient assumption; (G2) ensures global existence, uniqueness, and boundedness of solutions to (1.1); and (G3) is equivalent to a one-sided Lipschitz condition [17]. Both (G4) and (G5) are structural stability conditions on the gradient system.

For gradient systems in a fixed-step implementation there are very few known schemes which preserve the gradient structure for Δt independent of initial data (see [6], [7], [16], and [17]). The simplest of these schemes is backward Euler. Thus in §5 we consider an error control method designed to keep the solution sequence close to that produced by the backward Euler scheme (see (2.24), (2.27), (1.5)). For τ sufficiently small, but independent of initial data, we prove in Theorem G1 that this simplified order $(p, 1)$ error control scheme forces the numerical solution to enter and remain in a ball centred on one of the equilibria in $Q(0)$; the radius of the ball is proportional to τ. Dissipativity follows from this. In addition, a modification of this error control is proposed which actually ensures that the solution is driven to an equilibrium point as $n \to \infty$. This is based on error per unit step control relative to a discrete time

derivative (see Theorem G2).

The work contained here is inspired by [11] and [9], where the dynamics of error controlled schemes are studied for linear decay problems; in particular they show that for such problems standard error control mechanisms drive the numerical solution to a neighbourhood of the origin which scales with the error tolerance. This motivates the results proved here for contractive and gradient systems.

The work is also related to the work of Stetter [22] (see also [13], [14], [21]) where it is shown that, *over fixed time intervals* $0 \leqslant t \leqslant T$, and assuming that the asymptotic expansion governing the local error is valid, the global error is proportional to some positive power of the tolerance; this is essentially a convergence result for error controlled schemes as $\tau \to 0$. Further analysis of time step control may be found in [19]. Here we show that the "error" in the *asymptotic behaviour* ($t \to \infty$) is proportional to a positive power of the tolerance; this is essentially a practical stability result for error controlled schemes. In our analysis we do not need asymptotic expansions for the truncation error to prove results; we simply use the closeness of the scheme to an algebraically stable one.

Recently there has been some interest in the subject of spurious solutions introduced by fixed time step discretisation (see [18] for a summary). One reason these spurious solutions are of interest is that they can exist for arbitrarily small Δt and thereby destroy the large-time properties of the underlying differential equation. However, in [20], a valid criticism of the body of literature on spurious solutions is voiced: in practice, error control mechanisms will prevent spurious solutions. Our work goes some way toward substantiating the claim in [20].

Numerical results illustrating the results contained here can be found in [25].

Summary. It is possible to make some progress in the rigorous analysis of error control strategies without the use of asymptotic error expansions. To this end

- We have introduced the notion of essentially algebraically stable embedded Runge–Kutta pairs. These are error control strategies which ensure that the solution is an $\mathcal{O}(\tau)$ perturbation of an algebraically stable scheme, where τ is the tolerance. It is shown that *explicit* essentially algebraically stable embedded pairs exist but an order barrier of $\min\{p, q\} \leqslant 4$ is proved for such explicit methods with nonnegative weights b_i. See Corollary 2.10.
- New simplified and computationally inexpensive embedded pairs are introduced with order $(p, 1)$, p arbitrarily large, which are essentially algebraically stable. These embedded pairs may be explicit. See Example 2.13.
- New error control strategies are introduced for gradient systems for which the error control is relative to a discrete time derivative. See §5.
- For certain essentially algebraically stable embedded pairs applied to dissipative, contractive, and gradient systems we prove that the underlying long-time behaviour of the differential equation (see Theorem ODE) is inherited by the error controlled scheme (see Theorems DC1, DC2 in §4 and Theorems G1 and G2 in §5).
- For error per unit step strategies we find that the underlying properties of classes (D), (C), and (G) are inherited for sufficiently small tolerance, but *independent of initial data.* This implies a strong degree of robustness for codes based on such strategies. The main technical difficulty in the analysis is to obtain results for τ, the tolerance, independent of initial data. The error per step strategies require initial-data dependent tolerance restrictions and are hence far less robust.
- Nowhere in the analysis do we actually describe how the time step is chosen to satisfy the error control criteria. Instead we prove that, at each step, the error control criteria

can be satisfied. (This amounts to showing boundedness of the solution sequence $\{U_n\}_{n=0}^{\infty}$; see §3.) Furthermore, under the appropriate structural assumptions on $f(\bullet)$ we also show that it is possible to find step-size sequences uniformly bounded from zero. This approach facilitates a straightforward approach to the analysis. To the best of our knowledge this is the first rigorous treatment of error control strategies over long time intervals.

• The analysis is preliminary in the sense that only a small number of standard embedded pairs have been shown to be essentially algebraically stable. However, no recourse is made to (possibly unwarranted) assumptions about the validity of the asymptotic expansions underlying the error control criteria. In a subsequent paper ([15]) the converse viewpoint is taken—a large class of error controlled schemes are analysed using similar mathematical ideas, under the assumption that the asymptotic expansion underlying the error control is valid. To close the gap between the analysis presented here and that in [15] would be extremely interesting.

2. Essentially algebraically stable embedded pairs. In this section we define and analyse stable embedded pairs. Roughly these are embedded pairs where the error control ensures that the scheme is close to a standard stable method. To this end recall the following matrices:

$$M = BA + A^T B - bb^T, \quad B = \operatorname{diag}\{b\}.$$

DEFINITION 2.1. The Runge–Kutta method A, b is *algebraically stable* if M and B are positive semi-definite. The Runge–Kutta method is *DJ-reducible* if for some nonempty index set $T \subset \{1, \ldots, s\}$,

$$b_j = 0,\ j \in T, \qquad a_{ij} = 0,\ i \notin T,\ j \in T,$$

and is said to be *DJ-irreducible* otherwise.

Recall that any algebraically stable method is DJ-reducible to a DJ-irreducible scheme with $b_i > 0, i = 1, \ldots, s$; see [5] or [12] for example.

We now find conditions under which the error control enables a given method to be considered as a small perturbation of an algebraically stable method. Given any scalar $e_0 \neq 0$ and any vector $e = (e_1, e_2, \ldots, e_k)^T$ we create a new Runge–Kutta method from the embedded pair (1.2)–(1.4) by defining

$$\hat{b} = (1 - e_0)b + e_0\bar{b} \tag{2.1}$$

and

$$\hat{A} = A + e_0 e(\bar{b} - b)^T. \tag{2.2}$$

We use the notation

$$\{\hat{A}\}_{ij} = \hat{a}_{ij}, \quad \hat{b} = (\hat{b}_1, \hat{b}_2, \ldots, \hat{b}_k)^T.$$

From (2.1), (1.4) it follows that

$$V_{n+1} = U_n + \Delta t_n \sum_{j=1}^{k} \left[\left(1 - \frac{1}{e_0} \right) b_j + \frac{1}{e_0} \hat{b}_j \right] f(\eta_j).$$

Thus the error controls (1.5) or (1.6) imply, respectively, that

$$\|E\| \leqslant \tau \tag{2.3}$$

or

$$\|E\| \leqslant \frac{\tau^3}{\Delta t_n}, \tag{2.4}$$

where

$$E = \sum_{j=1}^{k} [b_j - \hat{b}_j] f(\eta_j). \tag{2.5}$$

Using (2.5), equation (1.3) may be rewritten as

$$U_{n+1} = U_n + \Delta t_n \sum_{j=1}^{k} \hat{b}_j f(\eta_j) + \Delta t_n E. \tag{2.6}$$

Furthermore, (1.2), (2.2) give

$$\eta_i = U_n + \Delta t_n \sum_{j=1}^{k} \hat{a}_{ij} f(\eta_j) - e_0 e_i \Delta t_n \sum_{j=1}^{k} [\bar{b}_j - b_j] f(\eta_j),$$

and, by (2.1), we have $\hat{b} - b = e_0(\bar{b} - b)$ and hence

$$\eta_i = U_n + \Delta t_n \sum_{j=1}^{k} \hat{a}_{ij} f(\eta_j) + e_i \Delta t_n E. \tag{2.7}$$

Thus (2.6), (2.7) show that, under error control, the Runge–Kutta method (1.2), (1.3) is a perturbation of the new Runge–Kutta method defined by (2.1) and (2.2); the perturbation E is small and controlled by (2.3) or (2.4) depending upon the type of error control used. The basic idea behind this work is that, if the scalar e_0 and the vector e can be chosen to make the new Runge–Kutta method $\hat{A}, \hat{b}$ have desirable properties, then it may be possible to prove that those properties are also shared by the underlying embedded pair $A, b, \bar{b}$. Thus we give the following definition.

DEFINITION 2.2. The embedded Runge–Kutta pair (1.2)–(1.4) (briefly $A, b, \bar{b}$) is said to be *essentially algebraically stable* if there exists $e \in \mathbb{R}^k$ and $e_0 \in \mathbb{R}$ such that the Runge–Kutta method $\hat{A}, \hat{b}$ defined by (2.1), (2.2), is algebraically stable.

It is worth noting that if an embedded pair is essentially stable in non-extrapolation mode, then it is also essentially stable in extrapolation mode. The converse is also true. These facts are a simple consequence of Lemma 2.4 below. We require the following definition.

DEFINITION 2.3. Given any embedded pair (1.2)–(1.4) we define the *associated embedded pair*, found by interchanging the roles of b_j and $\bar{b}_j$, by considering (1.2) together with

$$U_{n+1} = U_n + \Delta t_n \sum_{j=1}^{k} \bar{b}_j f(\eta_j), \quad U_0 = U,$$

and

$$V_{n+1} = U_n + \Delta t_n \sum_{j=1}^{k} b_j f(\eta_j).$$

LEMMA 2.4. *If the embedded pair* (1.2)–(1.4) *is essentially algebraically stable, then the associated pair is essentially algebraically stable.*

Proof. Exchanging the roles of b and $\bar{b}$ in (2.1), (2.2) a short calculation shows that it is sufficient to replace e_0 and e by e_0^* and e^*, given by

$$e_0^* = 1 - e_0, \quad e^* = -\frac{e_0}{e_0^*} e$$

to establish the result. □

Note that it is possible for *explicit* embedded pairs to be essentially algebraically stable and we will give examples of such schemes. With this possibility in mind we now examine in detail the existence of essentially algebraically stable embedded Runge–Kutta pairs. In the following we need the matrices

$$\left\{ \begin{array}{c} \hat{B} = \operatorname{diag}\{\hat{b}\} \\ \hat{M} = \hat{B}\hat{A} + \hat{A}^T\hat{B} - \hat{b}\hat{b}^T \\ \tilde{M} = \hat{B}A + A^T\hat{B} - \hat{b}\bar{b}^T \end{array} \right\}. \tag{2.8}$$

We denote by $I \subset \mathbb{R}$ the closed interval for which $\hat{B}$ is positive semi-definite if $e_0 \in I$ and also define

$$\begin{aligned} \mathcal{S} &= \{x \in \mathbb{R}^k : x^T x = 1\}, \\ \mathcal{V} &= \{x \in \mathcal{S} : (\bar{b} - b)^T x = 0\}, \\ \mathcal{V}_\varepsilon &= \{x \in \mathcal{S} : |(\bar{b} - b)^T x| \leqslant \varepsilon\}. \end{aligned}$$

LEMMA 2.5. *Given $e_0 \in I\backslash\{0\}$ for which $\hat{B}$ is positive definite, the embedded pair $A, b, \bar{b}$ is essentially algebraically stable if $\tilde{M}$ is positive definite on $\mathcal{V}$. Conversely, if for each $e_0 \in I$ there exists $x \in \mathcal{V}$ for which $x^T\tilde{M}x < 0$, then the embedded Runge–Kutta pair $A, b, \bar{b}$ is not essentially algebraically stable.*

Proof. The Runge–Kutta method $\hat{A}, \hat{b}$ is algebraically stable if $\hat{M}, \hat{B}$ are positive semi-definite [1]. Now

$$\begin{aligned} x^T\hat{M}x &= x^T\hat{B}\hat{A}x + x^T\hat{A}^T\hat{B}x - x^T\hat{b}\hat{b}^T x \\ &= 2(\hat{B}x)^T\hat{A}x - (\hat{b}^T x)^2 \\ &= 2(\hat{B}x)^T(Ax + e_0 e(\bar{b} - b)^T x) - (\hat{b}^T x)^2 \\ &= 2(\hat{B}x)^T Ax + 2e_0(\bar{b} - b)^T x(x^T\hat{B}e) - (\hat{b}^T x)^2 \\ &= x^T\tilde{M}x + 2e_0(\bar{b} - b)^T x(x^T\hat{B}e). \end{aligned}$$

If $\tilde{M}$ is positive definite on $\mathcal{V}$ then, by continuity, for ε sufficiently small $\exists\delta > 0$ such that

$$x^T\hat{M}x \geqslant \delta \quad \text{on } \mathcal{V}_\varepsilon. \tag{2.9}$$

Furthermore, since $\mathcal{S}$ is a bounded set $\exists\gamma > 0$

$$x^T\tilde{M}x \geqslant -\gamma \quad \text{on } \{x \in \mathcal{S}\backslash\mathcal{V}_\varepsilon\}.$$

If we chose $\lambda > \gamma/(2e_0^2\varepsilon^2)$ and let e be the solution of

$$\hat{B}e = \lambda(\bar{b} - b)e_0,$$

then

$$x^T\hat{M}x = x^T\tilde{M}x + 2\lambda e_0^2[(\bar{b}-b)^T x]^2 > 0 \quad \text{on } \{x \in \mathcal{S}\backslash\mathcal{V}_\varepsilon\}. \tag{2.10}$$

The first part of the result follows since (2.9), (2.10) give lower positive bounds on $x^T M x$ on $\mathcal{S}$.

The second part of the result follows in a straightforward fashion since

$$x^T\hat{M}x = x^T\tilde{M}x \quad \text{on } \mathcal{V}. \qquad \square$$

This lemma shows that although there appear to be $k+1$ parameters to play with to ensure that $\hat{A}, \hat{b}$ is positive definite, in fact there is only one in almost all cases; this follows since e_0 is the only free parameter in $\tilde{M}$. Thus we now concentrate on studying $\tilde{M}$ on $\mathcal{V}$. (The question of the positivity of matrices on a linear subspace is considered in [23].) Notice that if A, b is explicit then, since $\tilde{M}$ is the algebraic stability matrix for the explicit Runge–Kutta method $A, \hat{b}$, it cannot be positive definite on $\mathbb{R}^k$. Furthermore, it is well known that the extreme values of the quadratic form $x^T\tilde{M}x$ on $\mathcal{V}$ interleave the eigenvalues of $\tilde{M}$ [27] and so from Lemma 2.5 we have the following result.

COROLLARY 2.6. *If $\tilde{M}$ has two negative eigenvalues for all $e_0 \in I$, then the embedded pair $A, b, \bar{b}$ is not essentially algebraically stable.*

Lemma 2.5 and Corollary 2.6 are suggestive of an order barrier for explicit essentially algebraically stable methods and we now prove that if $A, b, \bar{b}$ has order (p, q) with $\min\{p, q\} \geqslant 5$ and $b_i \geqslant 0$ then it is not essentially algebraically stable. Preceding this theorem are two lemmas needed in the proof.

LEMMA 2.7. *If $A, b, \bar{b}$ has order (p, q) with $\min\{p, q\} \geqslant 5$ then $\hat{A}, \hat{b}$ has order $\geqslant 5$.*

Proof. Note that if $\hat{A}, b$ and $\hat{A}, \bar{b}$ have order 5 then so does $\hat{A}, \hat{b}$ since $\hat{b}$ appears linearly in the order conditions [3]. Thus it is sufficient to show that $\hat{A}, b$ (and hence by an identical argument that $\hat{A}, \bar{b}$) has order 5. Noting that

$$\hat{c}_i = \sum_{j=1}^{k} \hat{a}_{ij} = \sum_{j=1}^{k} a_{ij} + e_0 e_i(\bar{b}_j - b_j) = \sum_{j=1}^{k} a_{ij} = c_i \tag{2.11}$$

the result follows from straightforward but lengthy manipulations of the order conditions, using (2.2). $\square$

The following definition will be needed.

DEFINITION 2.8. The embedded pair $A, b, \bar{b}$ is DJ-irreducible if no stages η_i can be simultaneously removed from both the methods A, b and $A, \bar{b}$ to yield an equivalent method with fewer stages.

LEMMA 2.9. *Assume that the embedded pair $A, b, \bar{b}$ is explicit, DJ-irreducible, essentially algebraically stable, and has order (p, q) with $\min\{p, q\} \geqslant 5$. Let $T = \{j \in \mathcal{Z} : \hat{b}_j = 0\}$. Then*

(i) $\exists J \geqslant 3 : T = \{j : 1 \leqslant j \leqslant J\}$;
(ii) $\sum_{j=1}^{k} \hat{a}_{ij} c_j = \sum_{j=1}^{k} a_{ij} c_j = c_i^2/2, i \notin T$;
(iii) $b_j \neq 0$, $b_j \neq \bar{b}_j \ \forall j \in T$;
(iv) $a_{ij} = e_i b_j, \hat{a}_{ij} = 0, \forall i, j : i \notin T, j \in T$;
(v) $\exists j \in T : b_j < 0$.

Proof. We define

$$d_i = \sum_{j=1}^{k} a_{ij} c_j - \frac{c_i^2}{2}, \quad \hat{d}_i = \sum_{j=1}^{k} \hat{a}_{ij} c_j - \frac{c_i^2}{2}. \tag{2.12}$$

Since $\hat{A}, \hat{b}$, A, b, and $A, \bar{b}$ have order at least 5, it follows from the proof of Lemma IV.13.12 in [12] that

$$\sum_{i=1}^{k} \hat{b}_i \hat{d}_i^2 = 0, \quad \sum_{i=1}^{k} b_i d_i^2 = 0, \quad \sum_{i=1}^{k} \bar{b}_i d_i^2 = 0. \tag{2.13}$$

Since $\hat{A}, \hat{b}$ is algebraically stable it follows that $\hat{b}_i \geqslant 0$ for all i. Let $T = \{j : \hat{b}_j = 0\}$. Thus

$$\hat{d}_i = 0 \quad \text{if } i \notin T. \tag{2.14}$$

Also

$$\begin{aligned} \sum_{j=1}^{k} \hat{a}_{ij} c_j &= \sum_{j=1}^{k} [a_{ij} c_j + e_0 e_i (\bar{b}_j - b_j) c_j] \\ &= \sum_{j=1}^{k} a_{ij} c_j + e_0 e_i \left(\sum_{j=1}^{k} c_j \bar{b}_j - c_j b_j \right) = \sum_{j=1}^{k} a_{ij} c_j \end{aligned}$$

since the methods A, b and $A, \bar{b}$ have order greater than 2. Thus, by (2.12) and (2.14), $d_i = \hat{d}_i \ \forall i$ and so

$$d_i = 0 \quad \forall i \notin T. \tag{2.15}$$

Equations (2.14), (2.15) establish (ii).

Because the method $\hat{A}, \hat{b}$ is algebraically stable it is DJ-reducible [12] to a method with $\hat{b}_i > 0$. Thus it follows that

$$\hat{a}_{ij} = a_{ij} + e_0 e_i (\bar{b}_j - b_j) = 0, \quad \forall i \notin T, j \in T \tag{2.16}$$

and that

$$\hat{b}_j = b_j + e_0 (\bar{b}_j - b_j) = 0, \quad \forall j \in T. \tag{2.17}$$

For the purposes of contradiction, let $j \in T$ and $b_j = 0$. Then $\bar{b}_j = 0$ since $e_0 \neq 0$ and it follows that $a_{ij} = 0 \ \forall i \notin T, j \in T$. But this contradicts the irreducibility of $A, b, \bar{b}$. Thus $b_j \neq 0$ and then $\bar{b}_j \neq b_j$ by (2.17). Hence

$$b_j \neq 0 \text{ and } \bar{b}_j \neq b_j \quad \forall \in T. \tag{2.18}$$

Combining (2.16), (2.17) gives

$$a_{ij} = e_i b_j, \quad \forall i, j : i \notin T, j \in T. \tag{2.19}$$

Equations (2.16)–(2.19) establish (iii) and (iv).

Now we characterise T. Clearly $2 \in T$ for if not we have by (2.15)

$$d_2 = \sum_{j=1}^{k} a_{2j} c_j = -c_2^2/2 = 0, \tag{2.20}$$

which is not possible for an irreducible explicit method. For the purposes of contradiction, let $1 \notin T$. Then, since the method is explicit $a_{12} = 0$ and (2.19) gives $e_1 = 0$

since $b_2 \neq 0$, by (2.18). Now $e_1 = 0$ implies $\hat{a}_{1j} = a_{1j} = 0 \; \forall j$, by (2.2). This gives a contradiction since $\hat{a}_{11} = 0$ implies

$$x^T \hat{M} x = -\hat{b}_1^2 < 0$$

if $x = (1, 0, \ldots, 0)^T$ and since $\hat{b}_1 > 0$ if $1 \notin T$; this violates the algebraic stability of $\hat{A}, \hat{b}$.

Thus $1, 2 \in T$. Assume that $j \in T$ for $1 \leqslant j \leqslant J$ and that $J + 1 \notin T$. For the purposes of contradiction, assume that $\exists j^* > J + 1 : j^* \in T$. Then, since the method is explicit, $a_{J+1,j^*} = 0$ and so, by (2.19), $e_{J+1} = 0$, since $b_{j^*} \neq 0$ by (2.18). Thus, by (2.2),

$$a_{J+1,j} = \hat{a}_{J+1,j} \quad \forall j.$$

If the vector x is defined by $\{x\}_i = \delta_{i,J+1}$ with the usual Kronecker-delta notation then it follows that $\hat{A}x$ is orthogonal to $\hat{B}x$

$$\{\hat{A}x\}_i = \hat{a}_{i,J+1}, \quad \{\hat{B}x\}_i = \hat{b}_i \delta_{i,J+1},$$

so that

$$(\hat{B}x)^T (\hat{A}x) = \hat{b}_{J+1} \hat{a}_{J+1,J+1} = \hat{b}_{J+1} a_{J+1,J+1} = 0,$$

since $a_{ii} = 0$ for explicit methods. Hence $x^T \hat{M} x = -\hat{b}_{J+1}^2 < 0$ and the contradiction follows since $\hat{b}_{J+1} > 0$ as $J + 1 \notin T$.

Finally, to complete (i), we need to show that $J \geqslant 3$. By (2.12), (2.13), and (2.15) we deduce that

$$\sum_{i=1}^{J} b_i d_i^2 = 0. \tag{2.21}$$

Note that $d_2 \neq 0$ for an irreducible explicit method by the argument following (2.20). Since $d_1 = 0$ for an explicit method, $d_2 \neq 0$ and $b_2 \neq 0$ by (iii), we deduce that $J \geqslant 3$.

To complete (v), note that since $b_2, d_2 \neq 0$ it follows that there exists $j \in T$ for which $b_j < 0$. □

Corollary 2.10 follows automatically from Lemma 2.9(v).

COROLLARY 2.10. *There are no explicit essentially algebraically stable embedded pairs with nonnegative weights b_i and order (p, q) satisfying* $\min\{p, q\} \geqslant 5$.

Proof. Assume to the contrary that $p, q \geqslant 5$ and the weights $b_i \geqslant 0$. By Lemma 2.9(v) we obtain a contradiction. □

Remark. Many standard methods are constructed with the simplifying assumption that the $b_i \geqslant 0$ [3]; thus Corollary 2.10 maybe of interest. However, some embedded pairs used in practice do not satisfy $b_i \geqslant 0$, leaving open the question of the existence of high-order essentially stable methods.

We now proceed to give some examples of essentially algebraically stable embedded pairs.

Example 2.11. One of the simplest error control strategies is to take the explicit Euler scheme

$$U_{n+1} = U_n + \Delta t_n f(U_n) \tag{2.22}$$

and then form the second-order accurate approximation

$$V_{n+1} = U_n + \frac{\Delta t_n}{2}[f(U_n) + f(U_{n+1})]. \tag{2.23}$$

This method has order $(1, 2)$. In the standard Butcher notation we have that

$$A = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}, \quad b = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad \bar{b} = \begin{pmatrix} \frac{1}{2} \\ \frac{1}{2} \end{pmatrix}.$$

If we take $e_0 = 2$ and $e = (0, 1)^T$ then

$$\hat{A} = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}, \quad \hat{b} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}.$$

The new method is DJ-reducible [12] to the backward Euler scheme. Thus (2.22), (2.23) are essentially algebraically stable.

We can also consider the scheme in extrapolation mode so that

$$V_{n+1} = U_n + \Delta t_n f(U_n)$$

and

$$U_{n+1} = U_n + \frac{\Delta t_n}{2}[f(U_n) + f(V_{n+1})].$$

By Lemma 2.4 this method is also essentially algebraically stable.

Example 2.12. As a second example we consider the Fehlberg order (2,3) method given by

$$\begin{aligned} \eta_1 &= U_n, \\ \eta_2 &= U_n + \Delta t_n f(\eta_1), \\ \eta_3 &= U_n + \frac{\Delta t_n}{4}[f(\eta_1) + f(\eta_2)], \\ U_{n+1} &= U_n + \frac{\Delta t_n}{2}[f(\eta_1) + f(\eta_2)], \\ V_{n+1} &= U_n + \frac{\Delta t_n}{6}[f(\eta_1) + f(\eta_2)] + \frac{2\Delta t_n}{3} f(\eta_3). \end{aligned}$$

In the standard Butcher notation we have that

$$A = \begin{pmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ \frac{1}{4} & \frac{1}{4} & 0 \end{pmatrix}, \quad b = \begin{pmatrix} \frac{1}{2} \\ \frac{1}{2} \\ 0 \end{pmatrix}, \quad \bar{b} = \begin{pmatrix} \frac{1}{6} \\ \frac{1}{6} \\ \frac{2}{3} \end{pmatrix}.$$

If we take $e_0 = \frac{3}{2}$ and $e = (0, 0, \frac{1}{2})^T$ then

$$\hat{A} = \begin{pmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & \frac{1}{2} \end{pmatrix}, \quad \hat{b} = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}.$$

The new method is DJ-reducible to the implicit midpoint rule and hence algebraically stable. As in the previous example, the scheme can be shown to be essentially algebraically stable in extrapolation mode by means of Lemma 2.4.

Example 2.13. Because of the order barrier established in Corollary 2.10, it is not possible to find explicit essentially algebraically stable embedded pairs with order $(p, p+1)$, $p \geqslant 5$ and positive weights b_i. However it is possible to seek methods of order $(p, 1)$ for arbitrarily large p. Let

$$U_{n+1} = U_n + \Delta t_n \tilde{f}(U_n; \Delta t_n), \quad U_0 = U, \tag{2.24}$$

denote any Runge–Kutta method where $\tilde{f}(U_n; \Delta t_n)$ is defined by the internal stages of the Runge–Kutta method by (1.2), (1.3). Thus in the case of explicit embedded pairs we have

$$\tilde{g}_i(u, \Delta t) := \sum_{j=1}^{i-1} a_{ij} f(u + \Delta t \tilde{g}_j(u, \Delta t)), \quad i = 1, \ldots, k, \tag{2.25}$$

$$\tilde{f}(u, \Delta t) := \sum_{j=1}^{k} b_j f(u + \Delta t \tilde{g}_j(u, \Delta t)). \tag{2.26}$$

Now define, for $\theta \in (0, 1]$,

$$V_{n+1} = U_n + \Delta t_n[(1-\theta)\tilde{f}(U_n; \Delta t_n) + \theta f(U_{n+1})]. \tag{2.27}$$

The error controls (1.5), (1.6) with $e_0 = \theta^{-1}$ then imply that

$$\|\tilde{f}(U_n; \Delta t_n) - f(U_{n+1})\| \leqslant \tau \tag{2.28}$$

or

$$\|\tilde{f}(U_n; \Delta t_n) - f(U_{n+1})\| \leqslant \tau^3 / \Delta t_n, \tag{2.29}$$

respectively, so that the original scheme is close to the backward Euler scheme and hence it may be shown that the embedded pair is essentially algebraically stable. (The details are omitted for brevity.)

Notice that while this error control is nonstandard, it is cheap to implement since $f(U_{n+1})$ must be calculated as the first function evaluation in the next step of any explicit method. Indeed the error controls (2.28) or (2.29) can be implemented directly without calculating V_{n+1} and could be used, for example, *in addition* to a standard error control mechanism based on an order $(p, p+1)$ pair. It is conceivable that, provided τ is chosen sensibly in relation to the tolerance arising from the standard error control code (that is, substantially larger), then the additional constraint on the choice of time step will not greatly increase computational expense. However this has not been verified.

If (2.24) is the explicit Euler scheme and $\theta = 1/2$, then the method is simply the order $(1, 2)$ pair of Example 2.11. However, if the method (2.24) has order $p > 1$ then (2.27) has order 1: Assume that (2.24) is defined by a $(k-1)$-stage Runge–Kutta method and let $\{b_i, c_i\}_{i=1}^{k-1}$ and $\{\bar{b}_i, \bar{c}_i\}_{i=1}^{k}$ denote the standard weights for the Runge–Kutta methods (2.24) and (2.27), respectively. If (2.24) has order $p > 1$ then, by [3],

$$\sum_{i=1}^{k-1} b_i = 1, \quad \sum_{i=1}^{k-1} b_i c_i = \frac{1}{2}.$$

The method (2.27) has

$$\bar{b}_i = (1-\theta)b_i,\ \bar{c}_i = c_i, \quad i = 1, \ldots, k-1$$

and

$$b_k = \theta, \quad c_k = 1.$$

Clearly

$$\sum_{i=1}^{k} \bar{b}_i = 1;$$

however

$$\sum_{i=1}^{k} \bar{b}_i c_i = \frac{1-\theta}{2} + \theta = \frac{1}{2} + \frac{\theta}{2} \neq \frac{1}{2}$$

since $\theta = 0$ is not admitted. Thus (2.27) has order 1.

Example 2.14. The methods of Example 2.13 can be generalised as follows. Let

$$U_{n+1} = U_n + \Delta t_n \tilde{f}(U_n; \Delta t_n), \quad U_0 = U, \tag{2.30}$$

$$V_{n+1} = U_n + \Delta t_n[(1-\theta)\tilde{f}(U_n; \Delta t_n) + \theta f(\eta)], \tag{2.31}$$

where

$$\eta = (1-\phi)U_n + \phi U_{n+1}, \tag{2.32}$$

and $\phi \in [\frac{1}{2}, 1]$. Equation (2.30) represents any explicit Runge–Kutta method defined by (1.2), (1.3) and $\tilde{f}$ is therefore defined by (2.25), (2.26). The assumption that $\phi \in [\frac{1}{2}, 1]$ is necessary and sufficient for the embedded pair to be essentially algebraically stable. If $\theta = \frac{1}{2}, \phi = 1$, and $\tilde{f}(u; \Delta t) \equiv f(u)$ then the method is the embedded (1,2) pair of Example 2.11. If $\phi = \frac{1}{2}, \theta = \frac{2}{3}$, and $\tilde{f}$ is appropriately chosen, then the method is the Fehlberg (2,3) pair described in Example 2.12.

Notice that the methods of Example 2.12 correspond to choosing $\phi = 1$. Setting $\phi \neq 1$ allows higher-order error control than is possible with the methods of Example 2.12, but at the cost of introducing an extra stage to the Runge–Kutta method.

The order barrier $\min(p, q) \leqslant 2$ for (2.30)–(2.32) can be established by manipulating the order conditions. It is also easy to see that if $p \geqslant 2$ and $\phi = \frac{1}{2}$ then $q = 2$ and hence that there exist schemes of order $(p, 2)$ for arbitrarily large p.

3. Satisfaction of error control criteria. The numerical approximation to (1.1) is given by a sequence $\{U_n\}_{n=0}^{\infty}$ generated by (1.2)–(1.3). To specify such a sequence, given initial data $U_0 = U$, it is necessary to show that there exists a sequence $\{\Delta t_n\}_{n=0}^{\infty}$ so that the Runge–Kutta equations (1.2) are solvable for every $n \geqslant 0$ (which is, of course, trivial, if the error control scheme is explicit) and so that the error control criteria (1.5) (or (1.6)) is satisfied for every $n \geqslant 0$.

Furthermore, for the kind of problems in which we are interested here, the underlying differential equation has solutions which are defined and remain bounded for all $t \geqslant 0$. For this reason it is important to show that the error control criteria may

be satisfied for a time step sequence $\{\Delta t_n\}_{n=0}^{\infty}$ uniformly bounded from zero, that is, $\inf_{n\geqslant 0}\Delta t_n > 0$.

In this section we describe a general framework in which we analyse these issues. The basic idea is simple: if $f(\bullet)$ is Lipschitz and the solution sequence remains bounded, then it is possible to satisfy the error control criteria with a step-size sequence uniformly bounded from zero. The technicalities are rather lengthy and relegated to an appendix. Statements of results and definitions are all that are given here.

We commence by defining appropriate classes of functions and making a definition.

NOTATION 3.1. We denote the class of Lipschitz continuous functions mapping $\mathbb{R}^m$ into $\mathbb{R}^m$ and satisfying (D), (C), or (G) by $\mathcal{F}(\mathbf{D}), \mathcal{F}(\mathbf{C})$, and $\mathcal{F}(\mathbf{G})$, respectively.

Since we do not specify how a solution sequence satisfying the error control criteria (1.5) or (1.6) is actually found (we simply prove that such sequences can be found) it is necessary to distinguish between sequences which have time steps uniformly bounded from zero and those which do not. Roughly an *admissible sequence* is one with time steps uniformly bounded from zero. A more precise definition follows.

DEFINITION 3.2. Given an embedded pair (1.2)–(1.4), (1.5) (respectively, (1.2)–(1.4), (1.6)) a sequence $\{(U_n^T, V_n^T, \Delta t_n)\}_{n=0}^{\infty}$ with $(U_n^T, V_n^T, \Delta t_n) \in \mathbb{R}^{2p+1}$ satisfying (1.2)–(1.4), (1.5) (respectively, (1.2)–(1.4), (1.6)) is *admissible* if $\inf_{n\geqslant 0}\Delta t_n > 0$. An embedded pair is $\mathcal{F}(\bullet)$-*admissible* if, for every function $f \in \mathcal{F}(\bullet)$ and all $U \in \mathbb{R}^m$, there exists $\tau^* = \tau^*(f, U)$ such that the embedded pair has an admissible sequence for each $\tau \in (0, \tau^*)$. The pair is $\mathcal{F}(\bullet)$-*globally admissible* if it is $\mathcal{F}(\bullet)$-admissible and $\tau^* = \tau^*(f)$ is independent of U.

Note that an $\mathcal{F}(\bullet)$-globally admissible embedded pair is considerably more robust than an $\mathcal{F}(\bullet)$-admissible embedded pair since a suitable τ can be found which is independent of initial data U. The following theorems and corollaries are proved in the appendix. Lemma A1 in the appendix, concerning solvability of the implicit Runge–Kutta equations, is frequently referred to in the text.

THEOREM 3.3. *Assume that* $\exists \tau^* = \tau^*(U) > 0$ *and a compact set* $I = I(U) \subset \mathbb{R}^m$ *such that for* $\tau \in (0, \tau^*)$ *any solution sequence* $\{U_n\}_{n=0}^{\infty}$ *satisfying* (1.2)–(1.4), (1.5) *remains in* I $\forall n \geqslant 0$. *Then, for any* $\tau \in (0, \tau^*)$, *there exists an admissible sequence satisfying* (1.2)–(1.4), (1.5).

COROLLARY 3.4. *Assume that, for every function* $f \in \mathcal{F}(\bullet)$ *and all* $U \in \mathbb{R}^m$, $\exists \tau^* = \tau^*(f, U) > 0$ *and a compact set* $I = I(f, U) \subset \mathbb{R}^m$ *such that for* $\tau \in (0, \tau^*)$ *any solution sequence* $\{U_n\}_{n=0}^{\infty}$ *satisfying* (1.2)–(1.4), (1.5) *remains in* I $\forall n \geqslant 0$. *Then* (1.2)–(1.4), (1.5) *is* $\mathcal{F}(\bullet)$-*admissible. If* τ^* *is independent of* U *then* (1.2)–(1.4), (1.5) *is* $\mathcal{F}(\bullet)$-*globally admissible.*

THEOREM 3.5. *Assume that* $\exists \tau^* = \tau^*(U) > 0$ *and a compact set* $I = I(U) \subset \mathbb{R}^m$ *such that for* $\tau \in (0, \tau^*)$ *any solution sequence* $\{U_n\}_{n=0}^{\infty}$ *satisfying* (1.2)–(1.4), (1.6) *remains in* I $\forall n \geqslant 0$. *Then, for any* $\tau \in (0, \tau^*)$, *there exists an admissible sequence satisfying* (1.2)–(1.4), (1.6).

COROLLARY 3.6. *Assume that, for every function* $f \in \mathcal{F}(\bullet)$ *and all* $U \in \mathbb{R}^m$ $\exists \tau^* = \tau^*(f, U) > 0$ *and a compact set* $I = I(f, U) \subset \mathbb{R}^m$ *such that for* $\tau \in (0, \tau^*)$ *any solution sequence* $\{U_n\}_{n=0}^{\infty}$ *satisfying* (1.2)–(1.4), (1.6) *remains in* I $\forall n \geqslant 0$. *Then* (1.2)–(1.4), (1.6) *is* $\mathcal{F}(\bullet)$-*admissible. If* τ^* *is independent of* U *then* (1.2)–(1.4), (1.6) *is* $\mathcal{F}(\bullet)$-*globally admissible.*

To clearly state the sense in which the numerical method inherits the properties of the differential equation for problems under (D), (C), and (G), we give the following definition, analogous to Definition 1.1 for (1.1).

DEFINITION 3.7. Let the embedded pair (1.2)–(1.4), (1.5) be $\mathcal{F}(\bullet)$-admissible

and denote $\inf_{n\geqslant 0}\Delta t_n = \overline{\Delta t}$ for a given admissible sequence. The embedded pair is said to be $\mathcal{F}(\bullet)$-*dissipative* if there exists a set $\mathcal{B}_\tau \subset \mathbb{R}^m$ and $\tau_c = \tau_c(f-U)$ such that, for every admissible sequence with $U_0 = U \in \mathbb{R}^m$ and $\tau \in (0,\tau_c)$, $\exists n^* = n^*(U,\tau,\overline{\Delta t})$ such that

$$U_n \in \mathcal{B}_\tau \quad \forall n \geqslant n^*.$$

The pair is $\mathcal{F}(\bullet)$-*globally dissipative* if it is $\mathcal{F}(\bullet)$-globally admissible and if, in addition, $\tau_c = \tau_c(f)$ is independent of U.

Note that, from the definition $\mathcal{B}_\tau$ is independent of U. It is also essential in applications that $\mathcal{B}_\tau$ be uniformly bounded as $\tau \to 0$; this has not been made part of the definition but is true in all cases in this paper. Note that absorbing sets are not unique so that it does not make sense to talk of convergence of absorbing sets as $\tau \to 0$. Nonetheless, it is possible to talk about convergence of the $\mathcal{B}_\tau$ with minimal possible radius. However we are unable to show convergence of the $\mathcal{B}_\tau$ of minimal possible radius to the $\mathcal{B}$ of minimal possible radius as $\tau \to 0$. An error proportional to the largest allowable time step remains as $\tau \to 0$.

4. Dissipative and contractive problems. In this section we analyse error control schemes under assumptions (D) and (C), respectively. We assume throughout that there is an upper bound $\Delta t_{\max}$ on the time step; this need not be small and can be thought of as an $\mathcal{O}(1)$ bound independent of τ. Such an upper bound is often imposed by an actual implementation of an embedded pair in a software code to prevent enormous steps from being taken (see [8] for a discussion of this point). Furthermore, we make the following assumption, recalling that any algebraically stable method $\hat{A},\hat{b}$ is DJ-reducible to one with positive weights [12]:

(K) *For the algebraically stable scheme $\hat{A},\hat{b}$ DJ-reduced so that $\hat{B}$ is positive definite, there exist vectors $x = (x_1,\dots,x_k)$ and $(d_1,\dots,d_k)$ such that*

$$\hat{A}^T d + \hat{M}x = \hat{b} - \operatorname{diag}\{e\}\hat{b}$$

and

$$w^T d = 1$$

where $w = (1,\dots,1)^T$.

Note that (K) requires that some linear combination of the columns of $\hat{M}$ and $\hat{A}$ is an invertible matrix. The schemes in Examples 2.11–2.14 all satisfy this condition.

We prove the following two results which show that the error control enforces discrete analogues of Theorem ODE(i), (ii). Notice that, for the error per unit step scheme, the upper bound on the tolerance is independent of initial data. This is not true for the error per step scheme.

THEOREM DC1. *Consider* (1.2)–(1.4) *with error control* (1.5). *Assume that $A,b,\bar{b}$ is essentially algebraically stable and satisfies condition* (K) *and that $\Delta t_n \leqslant \Delta t_{\max}$ $\forall n \geqslant 0$. Then the embedded pair is $\mathcal{F}(\mathbf{D})$-globally dissipative and $\mathcal{F}(\mathbf{C})$-globally dissipative. In the second case $\exists c > 0 : \|u\|^2 \leqslant c\tau^2$ $\forall u \in \mathcal{B}_\tau$, the absorbing set.*

THEOREM DC2. *Consider* (1.2)–(1.4) *with error control* (1.6). *Assume that $A,b,\bar{b}$ is essentially algebraically stable and satisfies condition* (K) *and that $\Delta t_n \leqslant \Delta t_{\max}$ $\forall n \geqslant 0$. Furthermore, assume that the unique solution of the Runge–Kutta equations* (1.2) *satisfying $\eta_i \in \mathcal{Q}(U_n)$ constructed in Lemma* A1 *is used for each $n \geqslant 0$. Then the embedded pair is $\mathcal{F}(\mathbf{D})$-dissipative and $\mathcal{F}(\mathbf{C})$-dissipative. In the second case $\exists c > 0 : \|u\|^2 \leqslant c\tau$ $\forall u \in \mathcal{B}_\tau$, the absorbing set.*

Note that Theorem DC1 is considerably stronger than DC2 since global dissipativity is achieved.

The property of the differential equation which follows under structural assumption (D) and which we wish to exploit in our numerical analysis is that

$$\frac{1}{2}\frac{d}{dt}\{\|u(t)\|^2\} \leqslant \alpha - \beta\|u(t)\|^2;$$

this implies that $\|u(t)\|^2$ is strictly decreasing outside a ball of sufficiently large radius, yielding dissipativity. We now derive a preliminary lemma for the scheme (1.2)–(1.4), using the representation (2.6), (2.7). This lemma is related to the property of the differential equation just described. Our approach is motivated by the papers [1] and [16] where similar manipulations are performed in the case $E \equiv 0$. Throughout we use the notation $f_j = f(\eta_j)$.

If $A, b, \bar{b}$ is essentially algebraically stable then the new Runge–Kutta method $\hat{A}, \hat{b}$ is algebraically stable by definition. Furthermore $\hat{A}, \hat{b}$ is DJ-reducible to a method with $\hat{B}$ positive definite [4], [12]. If such a nontrivial reduction is possible then we define a reduced Runge–Kutta method from $\hat{A}, \hat{b}$ by removing $\eta_j, j \in T$ (where T is defined in Lemma 2.9) from the definition. However we will use the same notation $\hat{A}, \hat{b}$ for the reduced method and the same index k for the number of stages. All subsequent manipulations of (2.6), (2.7) apply with $k, \hat{A}, \hat{b}$ given by reduced method. Notice (from Examples 2.11 and 2.12) that the reducibility of the method $\hat{A}, \hat{b}$ does not imply the reducibility of the method $A, b, \bar{b}$.

LEMMA 4.1. *Let the embedded pair $A, b, \bar{b}$ be essentially algebraically stable and satisfy* (K). *Then, under assumption* (D) *on f, solutions of the embedded pair* (1.2)–(1.4) *satisfy*

$$\begin{aligned}\|U_{n+1}\|^2 \leqslant \|U_n\|^2 &+ 2\Delta t_n \sum_{i=1}^{k} \hat{b}_i[\alpha - \beta\|\eta_i\|^2] \\ &+ 2\Delta t_n \sum_{i=1}^{k} d_i\langle E, \eta_i\rangle + C\Delta t_n^2\|E\|^2,\end{aligned} \tag{4.1}$$

where

$$C = \sum_{i,j=1}^{k} |\hat{m}_{ij}x_i x_j| + 2\sum_{j=1}^{k}|d_j e_j| + 1. \tag{4.2}$$

Proof. From (2.6) we obtain, with the notation $f_j = f(\eta_j)$,

$$\begin{aligned}\|U_{n+1}\|^2 = \|U_n\|^2 &+ 2\Delta t_n \sum_{j=1}^{k} \hat{b}_j\langle U_n, f_j\rangle + \Delta t_n^2 \sum_{i,j=1}^{k} \hat{b}_i\hat{b}_j\langle f_i, f_j\rangle \\ &+ 2\Delta t_n\langle E, U_n\rangle + 2\Delta t_n^2 \sum_{j=1}^{k} \hat{b}_j\langle E, f_j\rangle + \Delta t_n^2\|E\|^2.\end{aligned}$$

Now, from (2.7) we have that

$$\langle \eta_i, f_i\rangle = \langle U_n, f_i\rangle + \Delta t_n \sum_{j=1}^{k} \hat{a}_{ij}\langle f_i, f_j\rangle + e_i\Delta t_n\langle E, f_i\rangle.$$

Combining these expressions gives

$$
\begin{aligned}
(4.3)\qquad \|U_{n+1}\|^2 = \|U_n\|^2 &+ 2\Delta t_n \sum_{i=1}^{k} \hat{b}_i \langle \eta_i, f_i \rangle \\
&- 2\Delta t_n^2 \sum_{i,j=1}^{k} \hat{b}_i \hat{a}_{ij} \langle f_i, f_j \rangle - 2\Delta t_n^2 \sum_{i=1}^{k} \hat{b}_i e_i \langle E, f_i \rangle \\
&+ \Delta t_n^2 \sum_{i,j=1}^{k} \hat{b}_i \hat{b}_j \langle f_i, f_j \rangle \\
&+ 2\Delta t_n \langle E, U_n \rangle + 2\Delta t_n^2 \sum_{j=1}^{k} \hat{b}_j \langle E, f_j \rangle + \Delta t_n^2 \|E\|^2.
\end{aligned}
$$

Now note that, by assumption (K) on d and by (2.7),

$$
(4.4)\qquad U_n = \sum_{i=1}^{k} d_i U_n = \sum_{i=1}^{k} d_i \eta_i - \Delta t_n \sum_{i=1}^{k} d_i \sum_{j=1}^{k} \hat{a}_{ij} f_j - \Delta t_n \sum_{i=1}^{k} d_i e_i E.
$$

Recall $\hat{M}$ defined by (2.8) and let $\hat{m}_{ij} = \{\hat{M}\}_{ij}$. By the symmetry of $\hat{M}$ it follows that

$$
\begin{aligned}
(4.5)\qquad \sum_{i,j=1}^{k} \hat{m}_{ij} \langle f_i, f_j \rangle &= \sum_{i,j=1}^{k} \hat{m}_{ij} \langle f_i - x_i E, f_j - x_j E \rangle \\
&\quad + 2 \sum_{i,j=1}^{k} \hat{m}_{ij} x_i \langle E, f_j \rangle - \sum_{i,j=1}^{k} \hat{m}_{ij} x_i x_j \|E\|^2.
\end{aligned}
$$

Noting that

$$
2 \sum_{i,j=1}^{k} \hat{b}_i \hat{a}_{ij} \langle f_i, f_j \rangle = \sum_{i,j=1}^{k} [\hat{b}_i \hat{a}_{ij} + \hat{b}_j \hat{a}_{ji}] \langle f_i, f_j \rangle
$$

and combining (4.3)–(4.5) gives

$$
\begin{aligned}
\|U_{n+1}\|^2 = \|U_n\|^2 &+ 2\Delta t_n \sum_{i=1}^{k} \hat{b}_i \langle \eta_i, f_i \rangle - \Delta t_n^2 \sum_{i,j=1}^{k} \hat{m}_{ij} \langle f_i - x_i E, f_j - x_j E \rangle \\
&- 2\Delta t_n^2 \sum_{i,j=1}^{k} \hat{m}_{ij} x_i \langle E, f_j \rangle + \Delta t_n^2 \sum_{i,j=1}^{k} \hat{m}_{ij} x_i x_j \|E\|^2 \\
&+ 2\Delta t_n^2 \sum_{j=1}^{k} \hat{b}_j (1 - e_j) \langle E, f_j \rangle + \Delta t_n^2 \|E\|^2 + 2\Delta t_n \sum_{i=1}^{k} d_i \langle E, \eta_i \rangle \\
&- 2\Delta t_n^2 \sum_{i,j=1}^{k} \hat{a}_{ij} d_i \langle E, f_j \rangle - 2\Delta t_n^2 \sum_{i=1}^{k} d_i e_i \|E\|^2.
\end{aligned}
$$

Using the structural assumption (D) on f, the positivity of $\hat{M}$, and condition (K) on the method, we deduce that (4.1), (4.2) hold. This completes the proof. □

4.1. Error per unit step. We now prove Theorem DC1 through a basic lemma on admissibility. Recall that, for the DJ-reduced method which for simplicity we denote $\hat{A}, \hat{b}$, it is known that $\hat{b}_i > 0$ for all i. Now, using

$$|2d_i\langle E, \eta_i\rangle| \leqslant |d_i|\ \|E\|\ [1 + \|\eta_i\|^2]$$

we obtain from Lemma 4.1, in the error per unit step case (1.5) (which implies (2.3))

$$\|U_{n+1}\|^2 \leqslant \|U_n\|^2 + 2\Delta t_n \sum_{i=1}^{k} \hat{b}_i[\tilde{\alpha} - \tilde{\beta}\|\eta_i\|^2], \tag{4.6}$$

where

$$\tilde{\alpha} = \alpha + \frac{\tau^2 C \Delta t_{\max}}{2} + \frac{\tau}{2} \max_i \frac{|d_i|}{\hat{b}_i}, \quad \tilde{\beta} = \beta - \frac{\tau}{2} \max_i \frac{|d_i|}{\hat{b}_i}, \tag{4.7}$$

and we have assumed that $\Delta t_n \leqslant \Delta t_{\max}$. If we define

$$\tau^* = \min_i \frac{2\beta \hat{b}_i}{|d_i|} \tag{4.8}$$

then $\tilde{\beta} > 0$ provided that $\tau < \tau^*$.

LEMMA 4.2. *Assume that* $\Delta t_n \leqslant \Delta t_{\max}\ \forall n \geqslant 0$. *Then, under the conditions of Lemma* 4.1, *the embedded Runge–Kutta pair* (1.2)–(1.4), (1.5) *is* $\mathcal{F}(\mathbf{D})$*-globally admissible and* $\mathcal{F}(\mathbf{C})$*-globally admissible.*

Proof. Note that $\mathcal{F}(\mathbf{D})$ global admissibility implies $\mathcal{F}(\mathbf{C})$ global admissibility since assumption (C) implies (D) with $\alpha = 0$. Let τ^* be defined by (4.8), noting that it is independent of U. Given any $\rho > 0$, define

$$R = \frac{\tilde{\alpha} + \rho}{\tilde{\beta}} + \Delta t_{\max} K \tag{4.9}$$

where

$$K = \max_{\|\eta_i\| \leqslant \gamma_i} \left[2 \sum_{i,j=1}^{k} b_i e_{ij} \langle \eta_i, f(\eta_j)\rangle + \Delta t_{\max} \sum_{i=1}^{k} b_i \left\| \sum_{j=1}^{k} e_{ij} f(\eta_j) \right\|^2 \right],$$

$$e_{ij} = b_j - a_{ij},$$

and

$$\gamma_i^2 = \frac{\tilde{\alpha} + \rho}{\beta_i \hat{b}_i}. \tag{4.10}$$

Let

$$I(U) = \left\{ u \in \mathbb{R}^m : \|u\|^2 \leqslant \max\{\|U\|^2, R\} \right\}.$$

We show that any solution sequence must remain in $I(U)$. Noting that $U_0 \in I(U)$, we proceed by induction.

Assume that $U_N \in I(U)$. Now, if

$$\sum_{i=1}^{k} \hat{b}_i[\tilde{\alpha} - \tilde{\beta}\|\eta_i\|^2] \leqslant 0$$

then (4.6), which follows from Lemma 4.1, gives

$$\|U_{N+1}\|^2 \leqslant \|U_N\|^2 \leqslant \max\{\|U\|^2, R\}$$

and $U_{N+1} \in I(U)$ follows. Alternatively, if

$$\sum_{i=1}^{k} \hat{b}_i[\tilde{\alpha} - \tilde{\beta}\|\eta_i\|^2] \geqslant 0$$

then

$$\sum_{i=1}^{k} \hat{b}_i\|\eta_i\|^2 \leqslant \frac{\tilde{\alpha}}{\tilde{\beta}} \quad \Rightarrow \|\eta_i\|^2 \leqslant \frac{\tilde{\alpha}+\rho}{\hat{b}_i\tilde{\beta}}, \quad \rho > 0.$$

Now (1.2), (1.3) give

$$U_{n+1} = \eta_i + \Delta t_n \sum_{j=1}^{k} e_{ij} f(\eta_j)$$

and hence

$$\|U_{n+1}\|^2 = \|\eta_i\|^2 + 2\Delta t_n \sum_{j=1}^{k} e_{ij}\langle \eta_i, f(\eta_j)\rangle + \Delta t_n^2 \left\| \sum_{j=1}^{k} e_{ij} f(\eta_j) \right\|^2.$$

Noting that

$$\|U_{n+1}\|^2 = \sum_{i=1}^{k} \hat{b}_i \|U_{n+1}\|^2$$

we obtain $U_{N+1} \in I(U)$ and the inductive step follows. This completes the proof by Corollary 3.4, since τ^* is independent of U. □

Proof of Theorem DC1. The $\mathcal{F}(\mathbf{D})$- and $\mathcal{F}(\mathbf{C})$-global admissibility of the scheme are established in Lemma 4.2. Thus it remains to exhibit an absorbing set $\mathcal{B}_\tau$ for every admissible sequence.

Let $\tau_c = \tau^*$ defined by (4.8) and define

$$\mathcal{B}_\tau = \{u \in \mathbb{R}^m : \|u\|^2 \leqslant R\}, \tag{4.11}$$

where R is defined by (4.9). Take any $\rho > 0$. While

$$\sum_{i=1}^{k} \hat{b}_i[\tilde{\alpha} - \tilde{\beta}\|\eta_i\|^2] \leqslant -\rho \tag{4.12}$$

we have from (4.6)

$$\|U_{n+1}\|^2 \leqslant \|U_n\|^2 - 2\Delta t_n \rho. \tag{4.13}$$

Alternatively, if

$$\sum_{i=1}^{k} \hat{b}_i [\tilde{\alpha} - \tilde{\beta} \|\eta_i\|^2] \geqslant -\rho \tag{4.14}$$

it follows that

$$\sum_{i=1}^{k} \hat{b}_i \|\eta_i\|^2 \leqslant \frac{\tilde{\alpha} + \rho}{\tilde{\beta}}$$

and then, as in the proof of Lemma 4.2, that

$$\|U_{n+1}\|^2 \leqslant R. \tag{4.15}$$

Now, if $\|U_n\|^2 \leqslant R$ and (4.12) holds, then from (4.13), we have that $\|U_{n+1}\|^2 \leqslant R$. On the other hand, if (4.14) holds then $\|U_{n+1}\|^2 \leqslant R$ by (4.15). Hence the set $\mathcal{B}_\tau$ is positively invariant. It remains to show that iterates starting outside $\mathcal{B}_\tau$ enter $\mathcal{B}_\tau$ after a finite number of steps $n^*(U, \tau)$. A simple contradiction argument shows that this must occur, for if $\|U_{n+1}\|^2 > R \; \forall n \geqslant 0$, then (4.13) holds for all $n \geqslant 0$ and hence, since the sequence is admissible, $\exists \overline{\Delta t} > 0$

$$\|U_N\|^2 \leqslant \|U\|^2 - 2\overline{\Delta t} \rho N, \quad N \geqslant 0.$$

Letting $N \to \infty$ gives a contradiction. Thus we have $\mathcal{F}(\mathbf{D})$- and $\mathcal{F}(\mathbf{C})$-global dissipativity.

In the second case where (C) holds we have that $\alpha = 0$; using the fact that

$$\langle E, \eta_i \rangle \leqslant \frac{1}{2\delta^2} \|E\|^2 + \frac{\delta^2}{2} \|\eta_i\|^2$$

and making the choice $\delta : \delta^2 = \beta \hat{b}_i / d_i$ for each i we obtain (4.6) with

$$\tilde{\alpha} = \frac{\tau^2}{2} \left[C \Delta t_{\max} + \sum_{i=1}^{k} \frac{\beta d_i^2}{\hat{b}_i} \right], \quad \tilde{\beta} = \frac{\beta}{2}.$$

The proof proceeds as for (D) except that now we take $\rho = \tau^2$ in the construction of R given by (4.9). Clearly $\gamma_i = \mathcal{O}(\tau)$ and by the Lipschitz continuity of f it follows that

$$\max_i \max_{\|\eta_i\| \leqslant \gamma_i} \|f(\eta_i) - f(0)\| \leqslant \kappa \tau^{\frac{1}{2}}$$

for some constant κ independent of τ. Thus, since $f(0) = 0$, (4.9) shows that $R = c\tau$, c independent of τ, and this completes the proof. □

4.2. Error per step. We now extend the analysis of §4.1 to the error per step case. We require explicit bounds on the solutions of (1.2) in this subsection and hence frequently appeal to Lemma A1 of the appendix where a constructive existence theorem for η_i satisfying (1.2) is given.

We define R as in (4.9) and set

$$R_1 = \max \left\{ R, \frac{\alpha + \tau + \rho}{\beta - \tau} \right\}. \tag{4.16}$$

We may then define

$$I(U) = \left\{u \in \mathbb{R}^m : \|u\|^2 \leqslant \max\{\|U\|^2, R_1\}\right\}. \tag{4.17}$$

Now let

$$\tau_1^* = \min\left\{\tau^*, \beta, 2\left[(\tilde{a}+\tilde{b})\tilde{b}k^2\gamma L_I \sup_{u\in I(U)} \|f(u)\|\right]^{-1}\right\}, \tag{4.18}$$

where τ^* is defined by (4.8), the constants $\tilde{a}, \tilde{b}$, and γ are as in Lemma A1,

$$L_I = \sup_{X\in I(U)} L(X),$$

and $L(X)$ is the Lipschitz constant for f described in Lemma A1.

We now prove Theorem DC2 through a basic lemma on admissibility, paralleling the proof of Theorem DC1.

LEMMA 4.3. *Assume that* $\Delta t_n \leqslant \Delta t_{\max}\ \forall n \geqslant 0$. *Furthermore, assume that the unique solution of the Runge–Kutta equations* (1.2) *satisfying* $\eta_i \in \mathcal{Q}(U_n)$ *constructed in Lemma* A1 *is used. Then, under the conditions of Lemma* 4.2, *the embedded Runge–Kutta pair* (1.2)–(1.4), (1.6) *is* $\mathcal{F}(\mathbf{D})$*-admissible and* $\mathcal{F}(\mathbf{C})$*-admissible.*

Proof. Assume for the purposes of induction that

$$U_N \in I(U).$$

Clearly this is true for $N = 0$. Recall the bound (2.4) for $\|E\|$ under (1.6). Clearly, if $\Delta t_N \geqslant \tau^2$ then $\|E\| \leqslant \tau$ and (4.6) follows just as in the error per unit step case. Thus, if $\Delta t_N \geqslant \tau^2$ we deduce that, as in the proof of Theorem DC1, either

$$\|U_{N+1}\|^2 \leqslant \|U_N\|^2 - 2\Delta t_N \rho \tag{4.19}$$

or

$$\|U_{N+1}\|^2 \leqslant R_1, \tag{4.20}$$

since $R \leqslant R_1$ by (4.16).

If $\Delta t_N \leqslant \tau^2$ then we may exploit the size of Δt_N and work with the numerical method in the original form (1.2), (1.3). From (1.3) we obtain

$$U_{N+1} = U_N + \Delta t_N \sum_{j=1}^{k} b_j f(U_{N+1}) + \Delta t_N \sum_{j=1}^{k} b_j [f(\eta_j) - f(U_{N+1})].$$

Taking the inner product with U_{N+1} we obtain, from (D) and using $L(\bullet)$ as defined in Lemma A1,

$$\frac{1}{2}\|U_{N+1}\|^2 \leqslant \frac{1}{2}\|U_N\|^2 + \Delta t_N[\alpha - \beta\|U_{N+1}\|^2] + \Delta t_N \sum_{j=1}^{k} b_j L(U_N)\|\eta_j - U_{N+1}\|\ \|U_{N+1}\|.$$

Applying Lemma A1 we obtain

$$\begin{aligned}\|\eta_j - U_{n+1}\| &\leqslant \|\eta_j - U_n\| + \|U_{n+1} - U_n\| \\ &\leqslant \Delta t \tilde{a} k \gamma \|f(U_n)\| + \Delta t \tilde{b} k \gamma \|f(U_n)\|.\end{aligned}$$

Hence

$$\frac{1}{2}\|U_{N+1}\|^2 \leqslant \frac{1}{2}\|U_N\|^2+\Delta t_N[\alpha-\beta\|U_{N+1}\|^2]+\Delta t_N\tau^2(\tilde{a}+\tilde{b})\tilde{b}k^2L_I \sup_{u\in I(U)}\|f(u)\|\|U_{N+1}\|$$

and using $\tau \leqslant \tau_1^*$ we obtain

$$\frac{1}{2}\|U_{N+1}\|^2 \leqslant \frac{1}{2}\|U_N\|^2 + \Delta t_N[(\alpha+\tau)-(\beta-\tau)\|U_{N+1}\|^2].$$

Thus, either (4.19) holds or

$$\|U_{N+1}\|^2 \leqslant \frac{\alpha+\tau+\rho}{\beta-\tau} \leqslant R_1,$$

which is equivalent to (4.20).

Hence we have shown that (4.19), (4.20) are true regardless of whether $\Delta t_n \leqslant \tau^2$ or $\Delta t_n \geqslant \tau^2$. From these it follows simply that $U_{N+1} \in I(U)$ and the induction is complete. The proof then follows from Corollary 3.6, noting that τ_1^* depends on U. □

Proof of Theorem DC2. The proof is identical to that of Theorem DC1, noting that (4.19) and (4.20) form the basis for the induction; we take $\tau_c = \tau_1^*$ and

$$\mathcal{B}_\tau = \{u \in \mathbb{R}^m : \|u\|^2 \leqslant R_1\}.$$

Because of the dependency of τ_1^* on U only $\mathcal{F}(\mathbf{D})$- and $\mathcal{F}(\mathbf{C})$-admissibility are obtained. Note that, in the case of structural assumption (C) we take $\rho = \tau$ to obtain the result. □

5. Gradient systems. In this section we consider error control schemes for gradient systems satisfying (G). Recall that in a fixed time step implementation there are very few schemes that are known to be gradient stable (see [24]); the simplest gradient stable scheme is backward Euler. Hence we are unable to prove results for arbitrary essentially algebraically stable embedded pairs, but derive positive results for the order $(p, 1)$ embedded pair (2.24), (2.27) constructed in Example 2.13. This is possible since the error control forces the numerical method to behave as a small perturbation of the backward Euler scheme. We will impose an upper bound $\Delta t_{\max}$ on the time step. Unlike the previous section, where for dissipative problems $\Delta t_{\max}$ could be taken to be arbitrarily large, for gradient systems $\Delta t_{\max}$ will be bounded above in terms of the one-sided Lipschitz constant k appearing in (G3).

Recall that the equation (1.1) has the property that under (G) all trajectories approach equilibria as $t \to \infty$. In §5.1, we consider the error per unit step strategy (1.5) and prove the following result.

THEOREM G1. *Consider* (2.24)–(2.27) *and* (1.5). *Assume that* $\Delta t_n \leqslant 1/2k\ \forall n \geqslant 0$. *Then the embedded pair is* $\mathcal{F}(\mathbf{G})$*-globally admissible*; *furthermore, there exists* $\tau^* > 0$ *such that, for any* $\tau \in (0, \tau^*)$ *and any admissible sequence,* $\exists N^* = N^*(U,\tau) > 0$, $v = v(U,\tau) \in Q(0)$, *and* $K > 0$

$$\|U_n - v\| \leqslant K\tau \quad \forall n \geqslant N^*.$$

Thus (2.24)–(2.27) *under* (1.5) *is* $\mathcal{F}(\mathbf{G})$*-globally dissipative.*

This result is analogous to Theorem ODE(iii); however, rather than obtaining convergence to equilibrium we are guaranteed that solutions eventually enter and remain in a neighbourhood of an equilibrium point. It is possible to generalise Theorem

G1 to the error per step case as for the dissipative case in §3 and also to implicit (2.24) but we do not give details here. In accordance with the work of Hall [11] and Griffiths [9] on linear decay problems we know that the numerical solution may perform small oscillations about an equilibrium and hence that Theorem G1 is best possible for the error control (1.5). If we wish to ensure that the numerical solution is actually driven to equilibrium then we must modify (1.5). We consider a modification of the error control mechanism (1.5) specifically designed for gradient systems; we replace (1.5) by

$$\|U_{n+1} - V_{n+1}\| \leqslant \theta\tau\|U_{n+1} - U_n\|. \tag{5.1}$$

This is a form of error per unit step error control relative to an approximation of the time derivative, that is, when the time derivative is small then the time step is made small also. Clearly (5.1) may not be an appropriate error control in general but, once a trajectory of a gradient system has approached and remained inside a small neighbourhood of an equilibrium for a substantial time, it is natural to drive the solution to that equilibrium. It is clear that (5.1) should achieve this and in §5.2 we outline proof of the following theorem.

THEOREM G2. *Consider* (2.24)–(2.27) *and* (5.1). *Assume that* $\Delta t_n \leqslant 1/2k\ \forall n \geqslant 0$. *Then the embedded pair is* $\mathcal{F}(\mathbf{G})$*-globally admissible*; *furthermore, there exists* $\tau^* > 0$ *such that, for any* $\tau \in (0, \tau^*)$ *and any admissible sequence,* $\exists v = v(U, \tau) \in Q(0)$

$$\lim_{n\to\infty} \|U_n - v\| = 0.$$

Thus (2.24)–(2.27) *under* (5.1) *is* $\mathcal{F}(\mathbf{G})$*-globally dissipative.*

Note that, with error control (5.1) in Theorem G2, the structural stability assumption (G5) is not required in the proof and a modified statement could be made to reflect this fact.

5.1. Error per unit step. Since the set of equilibria is bounded and each member is hyperbolic, the set $Q(0)$ contains a finite number of points which we label $\{v_j\}_{j=1}^J$. Furthermore, there exists $\delta > 0$ such that

$$\min_{i\neq j} \|v_i - v_j\| \geqslant \delta. \tag{5.2}$$

We now prove that the set $Q(\varepsilon)$ is made up of a finite number of isolated neighbourhoods of equilibrium points, each of which may be inscribed in a ball with radius proportional to ε. For simplicity we drop the argument ε of Q, and the associated Q_j defined below, throughout the remainder of the section.

LEMMA 5.1. *There exist constants* $\varepsilon^*, C > 0$ *such that, for all* $\varepsilon \in [0, \varepsilon^*)$,

$$Q = \bigcup_{j=1}^{J} Q_j, \quad Q_i \bigcap Q_j = \emptyset,\ i \neq j, \quad Q_j \subseteq B(v_j, C\varepsilon),\ j = 1, \ldots, J.$$

Proof. By (G5) it follows that $\exists \varepsilon^*, R > 0$ such that, for all $\varepsilon \in [0, \varepsilon^*)$, $Q \in B(0, R)$. Then, by continuity of f it follows, possibly by further reduction of ε^*, that $\exists \eta > 0$ such that, for all $\varepsilon \in [0, \varepsilon^*)$,

$$Q = \bigcup_{j=1}^{J} Q_j, \quad Q_i \bigcap Q_j = \emptyset,\ i \neq j, \quad Q_j \subseteq B(v_j, \eta),\ j = 1, \ldots, J. \tag{5.3}$$

Now consider v satisfying

$$f(v) = \varepsilon w, \quad \|w\| \leqslant 1.$$

Note that v is of this form if and only if $v \in Q(\varepsilon)$. By (5.3) it follows that there exists $v_j \in Q(0) : \|v - v_j\| \leqslant \eta$. Now define $G : \mathbb{R}^m \times \mathbb{R} \mapsto \mathbb{R}^m$ by $G(v, \varepsilon) = f(v) - \varepsilon w$ and note that $G(v_j, 0) = 0$. Clearly $dG_v(v_j, 0)$ is invertible by (G4) and also $G(v, \varepsilon)$ is continuously differentiable with respect to v (by (G1)) and ε. Hence, by the implicit function theorem, we deduce that $\exists C_j > 0$ such that $\|v - v_j\| \leqslant C_j \varepsilon$, for all $v \in Q_j$. Taking the maximum over all $C_j, j = 1, \ldots, J$ we obtain the desired result. □

The proof of Theorem G1 now proceeds through a sequence of lemmas. Recall that differential equations in gradient form have the property that

$$\frac{d}{dt}\{F(u(t))\} = -\|u_t(t)\|^2,$$

which forms the basis of the proof that solutions approach equilibrium points as $t \to \infty$. The error controlled scheme has the property that $F(U_n)$ is nonincreasing except in small neighbourhoods of equilibria and this is the basis of our proof of Theorem G1.

LEMMA 5.2. *Let $v \in Q$. Then there exists an integer $j : 1 \leqslant j \leqslant J$ such that $v \in Q_j$ and*

$$|F(v) - F(v_j)| \leqslant C(1 + kC)\varepsilon^2 \quad \forall \varepsilon \in [0, \varepsilon^*).$$

Proof. The existence of a $j : v \in Q_j$ follows from Lemma 5.1. By (G3) and Lemma 5.1 we have

$$F(v_j) - F(v) \leqslant \langle f(v_j), v - v_j \rangle + k\|v_j - v\|^2 \leqslant kC^2\varepsilon^2.$$

Similarly

$$\begin{aligned} F(v) - F(v_j) &\leqslant \langle f(v), v_j - v \rangle + k\|v_j - v\|^2 \\ &\leqslant \|f(v)\|\|v_j - v\| + k\|v_j - v\|^2 \leqslant C\varepsilon^2 + kC^2\varepsilon^2. \end{aligned}$$

This completes the proof. □

In the remainder we set $\varepsilon = 4\tau$ and assume that

$$\tau \in (0, \varepsilon^*/4) \quad \text{and} \quad 2\Delta t_n k \leqslant 1. \tag{5.4}$$

LEMMA 5.3. *Let (5.4) hold. Assume that there exist positive integers N, M with $N \leqslant M$ such that $U_n \notin Q(\varepsilon)$ for $n = N + 1, \ldots, M$ and $U_N, U_{M+1} \in Q(\varepsilon)$. Then*

$$F(U_M) - F(U_N) \leqslant -\frac{\tau}{2} \sum_{n=N}^{M-1} \|U_{n+1} - U_n\|.$$

Proof. By (G3) we have, from (5.4),

$$\begin{aligned} &F(U_{n+1}) - F(U_n) \leqslant \langle f(U_{n+1}), U_n - U_{n+1} \rangle + k\|U_{n+1} - U_n\|^2 \\ &= \langle \tilde{f}(U_n; \Delta t_n), U_n - U_{n+1} \rangle + k\|U_{n+1} - U_n\|^2 + \langle f(U_{n+1}) - \tilde{f}(U_n; \Delta t_n), U_n - U_{n+1} \rangle \\ &\leqslant -\frac{\|U_{n+1} - U_n\|^2}{2\Delta t_n} + \tau\|U_n - U_{n+1}\| \\ &= -\frac{\|U_{n+1} - U_n\|}{2}[\|\tilde{f}(U_n; \Delta t_n)\| - 2\tau]. \end{aligned}$$

By assumption we have that

$$\|f(U_{n+1})\| \geqslant 4\tau, \quad n = N, \ldots, M-1.$$

Hence, by (1.5) (which implies (2.28)), we have

$$\|\tilde{f}(U_n; \Delta t_n)\| \geqslant 3\tau, \quad n = N, \ldots, M-1, \tag{5.5}$$

so that

$$\|\tilde{f}(U_n; \Delta t_n)\| - 2\tau \geqslant \tau, \quad n = N, \ldots, M-1.$$

Thus

$$F(U_{n+1}) - F(U_n) \leqslant -\frac{\tau}{2}\|U_{n+1} - U_n\|, \quad n = N, \ldots, M-1. \tag{5.6}$$

The result follows. □

LEMMA 5.4. *Let* (5.4) *hold. Assume that there exists an integer* M *such that* $U_M \notin Q(\varepsilon)$ *and* $U_{M+1} \in Q(\varepsilon)$. *Then*

$$F(U_{M+1}) - F(U_M) \leqslant \frac{65\tau^2}{4k}.$$

Proof. By (2.24), (1.5) we have

$$\|U_{M+1} - U_M\| = \Delta t_n \|\tilde{f}(U_M; \Delta t_M)\| \leqslant \Delta t_n [\|f(U_{M+1})\| + \tau] \leqslant \frac{5\tau}{2k}. \tag{5.7}$$

Also, by (G3),

$$F(U_{M+1}) - F(U_M) \leqslant \|f(U_{M+1})\| \|U_{M+1} - U_M\| + k\|U_{M+1} - U_M\|^2.$$

Combining these two completes the proof since $U_{M+1} \in Q(\varepsilon) = Q(4\tau)$. □

LEMMA 5.5. *Under the same conditions as Lemma* 5.3, *with the assumption that* $U_N \in Q_j, U_{M+1} \in Q_j$, *it follows that there exists* $K_1 > 0$ *such that*

$$\|U_n - U_N\| \leqslant K_1, \quad n = N+1, \ldots, M.$$

Proof. Note that Lemmas 5.3 and 5.4 give

$$F(U_{M+1}) - F(U_N) \leqslant -\frac{\tau}{2} \sum_{n=N}^{M-1} \|U_{k+1} - U_k\| + \frac{65\tau^2}{4k}. \tag{5.8}$$

Hence, for $N+1 \leqslant n \leqslant M$ we have that

$$\begin{aligned} \|U_n - U_N\| &\leqslant \sum_{k=N}^{n-1} \|U_{k+1} - U_k\| \leqslant \sum_{k=N}^{M-1} \|U_{k+1} - U_k\| \\ &\leqslant \frac{2}{\tau}\left[\frac{65\tau^2}{4k} + F(U_N) - F(U_{M+1})\right]. \end{aligned}$$

Note that U_N and $U_{M+1} \in Q_j$; hence, applying Lemma 5.2 and noting that $\varepsilon = 4\tau$ gives the required result. □

LEMMA 5.6. *Under the same conditions as Lemma* 5.3, *with the assumption that* $U_N \in Q_j, U_{M+1} \in Q_i$, *and* $i \neq j$, *it follows that there exists* $\kappa > 0$ *such that*

$$F(v_i) - F(v_j) \leqslant -\frac{\delta\tau}{2} + \kappa\tau^2.$$

Hence, possibly by further reduction of ε^*, *we have*

$$F(v_i) - F(v_j) \leqslant -\frac{\delta\tau}{4}.$$

Proof. By Lemma 5.2 we have that

$$F(U_N) - F(v_j) \leqslant 16C(1 + kC)\tau^2$$

and

$$F(v_i) - F(U_{M+1}) \leqslant 16C(1 + kC)\tau^2.$$

Thus by (5.8) it follows that

$$F(v_i) - F(v_j) \leqslant -\frac{\tau}{2}\sum_{k=N}^{M-1} \|U_{k+1} - U_k\| + \frac{65\tau^2}{4k} + 32C(1 + kC)\tau^2.$$

Also, using (5.2), (5.7), and Lemma 5.1 we obtain

$$\begin{aligned}\delta &\leqslant \|v_i - v_j\| \leqslant \|U_M - U_N\| + \|U_{M+1} - U_M\| + \|U_{M+1} - v_i\| + \|U_N - v_j\| \\ &\leqslant \sum_{k=N}^{M-1} \|U_{k+1} - U_k\| + \frac{5\tau}{2k} + 8C\tau.\end{aligned}$$

Putting these estimates together gives the desired result. □

Proof of Theorem G1. Let

$$\bar{F} = \max_{v \in Q(0)} F(v)$$

and assume throughout this proof that $\tau \in (0, \varepsilon^*/4)$. Note that, from (5.5), (5.6), it follows that if $U_{n+1} \notin Q(4\tau)$ then

$$F(U_{n+1}) - F(U_n) \leqslant -3\tau^2\Delta t_n/2. \tag{5.9}$$

Note also that if $U_{n+1} \in Q(4\tau)$, then by Lemma 5.2 we have

$$F(U_{n+1}) - \bar{F} \leqslant 16C(1 + kC)\tau^2. \tag{5.10}$$

Assume for the purposes of induction that

$$F(U_n) \leqslant I(U) := \max\{F(U_0), \bar{F} + 16C(1 + kC)\tau^2\}. \tag{5.11}$$

Clearly this holds for $n = 0$. Assume it is true for $n = N$. If $U_{N+1} \notin Q(4\tau)$, then (5.9) gives

$$F(U_{N+1}) \leqslant F(U_N) \leqslant I(U).$$

On the other hand, if $U_{N+1} \in Q(4\tau)$, then (5.10) gives

$$F(U_{N+1}) \leqslant \bar{F} + 16C(1+kC)\tau^2 \leqslant I(U).$$

Hence (5.11) holds by induction for all $n \geqslant 0$. By (G2) this implies a global bound on the solution sequences provided $\tau \in (0, \varepsilon^*/4)$ and Corollary 3.4 yields global $\mathcal{F}(\mathbf{G})$-admissibility.

Now consider an admissible sequence with $\Delta t_n \geqslant \overline{\Delta t}\ \forall n \geqslant 0$. By (5.9) and (G2) we deduce that, if $U_m \notin Q(4\tau)$ then there exists $M \leqslant 2I(U)/(3\tau^2\overline{\Delta t})$ such that $U_M \in Q(4\tau)$. Let m_j denote the (possibly infinite) sequence of integers such that $U_{m_j} \notin Q(4\tau)$ and $U_{m_j+1} \in Q(4\tau)$. If the m_j is a finite set of integers then the proof is complete by Lemma 5.1. If the m_j comprise an infinite set then $m_j \to \infty$ as $j \to \infty$. By Lemma 5.6 we deduce that there exists an integer $k \geqslant 0$ and integer $i \in [0, J]$ such that

$$U_{m_j} \notin Q_i(4\tau), \quad U_{m_j+1} \in Q_i(4\tau) \quad \forall j \geqslant k,$$

since otherwise we obtain a contradiction to the fact that $F(\bullet) \geqslant 0$ by (G2). By Lemmas 5.1 and 5.5 we deduce that

$$\|U_n - v_i\| \leqslant (K_1 + C)\tau \quad \forall n \geqslant m_{k+1}.$$

This completes the proof. □

5.2. Relative error per unit step. We now consider the relative error strategy (2.24)–(2.27) and (5.1). Notice that (5.1) can be rewritten as

$$\|\tilde{f}(U_n; \Delta t_n) - f(U_{n+1})\| \leqslant \tau \|\tilde{f}(U_n; \Delta t_n)\|. \tag{5.12}$$

The proof is similar to that for Theorem G1 but simplified because the function $F(U_n)$ is nonincreasing for all $n \geqslant 0$. We omit the proof of the theorem but sketch the essential details. The details of the proof can be found in [25] where slightly more general structural conditions (G), including the case considered here, are studied.

First note that under the error control (5.1) analogous manipulations to those used in the proof of Lemma 5.3 show that, for τ sufficiently small, there exists $\xi > 0$ such that

$$F(U_{n+1}) - F(U_n) \leqslant -\xi \Delta t_n \|\tilde{f}(U_n; \Delta t_n)\|^2 \quad \forall n \geqslant 0. \tag{5.13}$$

This automatically gives boundedness of the solution sequence from (G2).

Second, note that the concept of admissibility, and related theorems showing that boundedness implies admissibility, can be extended to the error control (5.12) using techniques analogous to those in the appendix.

Finally, using (5.13) it is possible to show that

$$\lim_{n\to\infty} \tilde{f}(U_n; \Delta t_n) = 0$$

and that

$$\lim_{n\to\infty} f(U_{n+1}) = 0.$$

This allows us to show that all solution sequences are bounded and that any accumulation point of the sequence $\{U_n\}_{n=0}^{\infty}$ is contained in $Q(0)$. Arguing similarly to the proof of Theorem 4.3 in [16] we deduce that, in fact, the whole sequence converges to a point in $Q(0)$.

6. Appendix. In this appendix we provide the mathematical detail leading to the proofs of Theorems 3.3 and 3.5 and Corollaries 3.4 and 3.6. We start by addressing the solvability of the Runge–Kutta equations. In the following it will be useful to define

$$\tilde{a} = \max_{1\leqslant i,j\leqslant k} |a_{ij}| \quad \text{and} \quad \tilde{b} = \max_{1\leqslant i\leqslant k} |b_i|. \tag{6.1}$$

Note in the following that $L(X)$ depends upon Δt.

LEMMA A1. *For any $\gamma > 1$ let*

$$\mathcal{Q}(X) = \{u \in \mathbb{R}^m : \|u - X\| \leqslant \Delta t \tilde{a} k \gamma \|f(X)\|\},$$

$L(X)$ be the Lipschitz constant for $f(\bullet)$ on $\mathcal{Q}(X)$, and let $K(X) = \sup_{u\in\mathcal{Q}(X)} \|f(u)\|$. Let $\Delta t_c(X)$ be the minimum over all Δt satisfying

$$k\tilde{a}\Delta t L(X) = 1 - \gamma^{-1}, \tag{6.2}$$

or be ∞ if no such Δt exists. Then, for all $\Delta t \in [0, \Delta t_c(X))$, there exists a unique solution $\{\eta_i\}_{i=1}^k, \eta_i \in \mathbb{R}^m$ of the equations

$$\eta_i = X + \Delta t \sum_{j=1}^{k} a_{ij} f(\eta_j) \tag{6.3}$$

satisfying $\eta_i \in \mathcal{Q}(X)$. Furthermore if $\{\eta_i^l\}_{i=1}^k, l = 1, 2$ are solutions of (6.3) corresponding to distinct values $\Delta t = \Delta t^1$ and $\Delta t = \Delta t^2$, respectively, $\Delta t^l \in [0, \Delta t_c(X)), l = 1, 2$ then

$$\|\eta_i^1 - \eta_i^2\| \leqslant \tilde{a} k \gamma K(X) |\Delta t^1 - \Delta t^2|.$$

Finally, if $U_n = X$, then

$$\|U_{n+1} - X\| \leqslant \Delta t \tilde{b} k \gamma \|f(X)\|.$$

Proof. Note that the construction of Δt_c in (6.2) is slightly nontrivial since $L(X)$ depends upon Δt. Nonetheless it is clear that $\Delta t_c > 0$ and that, furthermore,

$$\Delta t < \frac{1 - \gamma^{-1}}{\tilde{a} k L(X)} \tag{6.4}$$

for all $\Delta t \in (0, \Delta t_c)$.

The existence of a solution satisfying the appropriate bound on the η_i follows from a contraction mapping argument, similar to that in [2] and here based on the iteration scheme

$$\xi_i^{k+1} = X + \Delta t \sum_{j=1}^{k} a_{ij} f(\xi_j^k), \quad i = 1, \ldots, K.$$

If $\{\eta_i^l\}_{i=1}^k$, $l = 1, 2$ solve (6.3) then

$$\eta_i^l = X + \Delta t^l \sum_{j=1}^{k} a_{ij} f(\eta_j^l), \quad i = 1, \ldots, k, \ l = 1, 2.$$

Hence

$$\begin{aligned}\|\eta_i^1 - \eta_i^2\| &= \|\sum_{j=1}^{k} a_{ij}(\Delta t^1 f(\eta_j^1) - \Delta t^2 f(\eta_j^2))\| \\ &\leqslant \tilde{a}\sum_{j=1}^{k}[\Delta t^1 \|f(\eta_j^1) - f(\eta_j^2)\| + |\Delta t^1 - \Delta t^2|\|f(\eta_j^2)\|] \\ &\leqslant \tilde{a}kL(X)\Delta t^1 \max_{1\leqslant j\leqslant k}\|\eta_j^1 - \eta_j^2\| + \tilde{a}kK(X)|\Delta t^1 - \Delta t^2|.\end{aligned}$$

Since this is true for any i and since $\tilde{a}kL(X)\Delta t^1 \leqslant (1-\gamma^{-1})$ it follows that

$$\max_{1\leqslant j\leqslant k}\|\eta_j^1 - \eta_j^2\| \leqslant \gamma\tilde{a}kK(X)|\Delta t^1 - \Delta t^2|.$$

The final bound follows from (1.3) since each $\eta_i \in Q(X)$. □

Next we discuss whether it is possible to satisfy (1.5) or (1.6). To this end, define $\xi_i, V, W : \mathbb{R}\times\mathbb{R}^m \to \mathbb{R}^m$ which are functions of Δt and X satisfying

$$\xi_i = X + \Delta t\sum_{j=1}^{k} a_{ij} f(\xi_j), \quad i = 1,\dots,k, \tag{6.5}$$

$$W = X + \Delta t\sum_{j=1}^{k} b_j f(\xi_j), \tag{6.6}$$

and

$$V = X + \Delta t\sum_{j=1}^{k} \bar{b}_j f(\xi_j). \tag{6.7}$$

Note that these functions are well defined by Lemma A1 for any $X \in \mathbb{R}^m$ and any $\Delta t \in [0, \Delta t_c(X))$. Hence we may define $G : [0,\Delta t_c(X))\times\mathbb{R}^m \to \mathbb{R}$ by

$$G(\Delta t, X) = \frac{\|W - V\|}{\Delta t} \tag{6.8}$$

and $H : [0,\Delta t_c(X))\times\mathbb{R}^m \to \mathbb{R}$ by

$$H(\Delta t, X) = \Delta t G(\Delta t, X). \tag{6.9}$$

The functions $G(\bullet, U_n)$ and $H(\bullet, U_n)$ must be made sufficiently small in order to satisfy the error controls (1.5) or (1.6), respectively. Thus their properties are important.

LEMMA A2. *The functions $G(\Delta t, X)$ and $H(\Delta t, X)$ satisfy $G(0,X) = H(0,X) = 0\ \forall X \in \mathbb{R}^m$ and are Lipschitz in $\Delta t \in [0,\Delta t_c(X))$.*

Proof. Since $\sum_{j=1}^{k} b_j = \sum_{j=1}^{k}\bar{b}_j = 1$ and $\xi_j(0,X) = X\ \forall X \in \mathbb{R}^m$ it follows that $G(0,X) = 0$. We now show that $G(\bullet, X)$ is Lipschitz continuous in $\Delta t \in [0,\Delta t_c(X))$. Note that

$$\begin{aligned}&|G(\Delta t^1, X) - G(\Delta t^2, X)| \\ &\quad = \left|\left\|\sum_{j=1}^{k}(b_j - \bar{b}_j) f(\xi_j(X,\Delta t^1))\right\| - \left\|\sum_{j=1}^{k}(b_j - \bar{b}_j) f(\xi_j(X,\Delta t^2))\right\|\right|\end{aligned}$$

$$\leqslant \left\| \sum_{j=1}^{k} (b_j - \bar{b}_j)[f(\xi_j(X, \Delta t^1)) - f(\xi_j(X, \Delta t^2))] \right\|$$
$$\leqslant kL(X) \max_{1 \leqslant j \leqslant k} |b_j - \bar{b}_j| \, \|\xi_j(X, \Delta t^1) - \xi_j(X, \Delta t^2)\|.$$

Thus, by Lemma A1,

$$|G(\Delta t^1, X) - G(\Delta t^2, X)| \leqslant CL(X)K(X)|\Delta t^1 - \Delta t^2|, \tag{6.10}$$

with C independent of X. Thus $G(\bullet, X)$ is Lipschitz.

The properties of $H(\bullet, X)$ follow immediately from those of $G(\bullet, X)$ since $H(\Delta t, X) = \Delta t G(\Delta t, X)$. □

Finally we prove Theorem 3.3.

Proof of Theorem 3.3. Let

$$\Delta t_c = \inf_{X \in I(U)} \left\{ \Delta t_c(X), \frac{\tau}{|e_0| CL(X)K(X)} \right\}, \tag{6.11}$$

where C is defined in (6.10) and $\Delta t_c(X)$ is defined in (6.2). Notice that $\Delta t_c > 0$ since both the Lipschitz constant for f and $K(\bullet)$ are bounded on any compact set. Now consider (1.2)–(1.5) with $\Delta t_n \equiv \Delta t_c \; \forall n \geqslant 0$. Assume, for the purposes of induction, that there exist solution sequences $\{U_n\}_{n=0}^{N}$ and $\{\Delta t_n\}_{n=0}^{N-1}$ satisfying (1.2)–(1.5) for $n = 0, \ldots, N-1$ with $\Delta t_n \equiv \Delta t_c$. Then, by assumption $U_N \in I(U)$ and hence, by Lemma A1 and (6.11), there exists a solution $\{\eta_i\}_{i=1}^{k}$ to (1.2) and thus a vector $U_{N+1} \in \mathbb{R}^m$ satisfying (1.3) with $n = N$ and $\Delta t_N = \Delta t_c$. By Lemma A2 and (6.11)

$$|G(\Delta t_c, U_N)| = |G(0, U_N) - G(\Delta t_c, U_N)| \leqslant \tau / |e_0|.$$

So, by construction, the error control criteria is satisfied. Thus there exist solution sequences $\{U_n\}_{n=0}^{N+1}$ and $\{\Delta t_n\}_{n=0}^{N}$ satisfying (1.2)–(1.5) for $n = 0, \ldots, N$ with $\Delta t_n \equiv \Delta t_c$. The inductive hypothesis is true for $N = 1$ by an identical argument since $U \in I(U)$ and hence an admissible sequence has been constructed satisfying $\inf_{n \geqslant 0} \Delta t_n = \Delta t_c > 0$. □

Corollary 3.4 is immediate. Furthermore the proofs of Theorem 3.5 and Corollary 3.6 follow similarly for the error control scheme (1.2)–(1.4), (1.6).

Acknowledgements. We are grateful to Kevin Burrage, John Butcher, Rob Corless, David Griffiths, Des Higham, and Arieh Iserles for a number of helpful suggestions.

REFERENCES

[1] K. Burrage and J. Butcher, *Stability criteria for implicit Runge–Kutta processes.* SIAM J. Numer. Anal., 16 (1979), pp. 46–57.

[2] J. Butcher, *Implicit Runge–Kutta processes*, Math. Comp., 18 (1964), pp. 50–64.

[3] ———, *The Numerical Analysis of Ordinary Differential Equations: Runge–Kutta and General Linear Methods*, Wiley, Chichester, 1987.

[4] G. Dahlquist and R. Jeltsch, *Generalised Disks of Contractivity for Explicit and Implicit Runge–Kutta Methods*, TRITA-NA report 7906, 1979.

[5] K. Dekker and J.G. Verwer, *Stability of Runge–Kutta Methods for Stiff Nonlinear Equations*, North-Holland, Amsterdam, 1984.

[6] C.M. Elliott, *The Cahn-Hilliard model for the kinetics of phase separation*, in Mathematical Models for Phase Change Problems, J.F. Rodrigues, ed., Birkhäuser, Boston, 1989.

[7] C.M. Elliott and A.M. Stuart, *The global dynamics of discrete semilinear parabolic equations*, SIAM J. Numer. Anal., 30 (1993), pp. 1622–1663.

[8] W.H. Enright, T.E. Hull, and B. Lindberg, *Comparing numerical methods for stiff systems of O.D.E.s*, BIT, 15 (1975), pp. 10–48.

[9] D.F. Griffiths, *The dynamics of some linear multistep methods with step-size control*, in Numerical Analysis, D. F. Griffiths and G.A. Watson, eds., Longman Scientific and Technical, 1988.

[10] J.K. Hale, *Asymptotic behaviour of dissipative systems*, Mathematical Surveys and Monographs vol. 25, American Mathematical Society, Providence, RI, 1988.

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

[12] E. Hairer and G. Wanner, *Solving Ordinary Differential Equations II. Stiff and Differential-Algebraic Problems*, Springer, New York, 1991.

[13] D. J. Higham, *Global error versus tolerance for explicit Runge–Kutta methods,* IMA J. Numer. Anal., 11 (1991), pp. 457–480

[14] ———, *Error control for initial value problems with discontinuities and delays*, Appl. Numer. Math., 12 (1993), pp. 315–330.

[15] D. J. Higham and A. M. Stuart, *Analysis of the dynamics of local error control via a piecewise continuous residual*, Numer. Math., submitted, 1995.

[16] A.R. Humphries and A.M. Stuart, *Runge–Kutta methods for dissipative and gradient dynamical systems*, SIAM J. Numer. Anal., 31 (1994), pp. 1452–1485.

[17] A.R. Humphries, *Numerical Analysis of Dynamical Systems*, Ph.D. thesis, University of Bath, Bath, United Kingdom, 1993.

[18] A. Iserles, A.T. Peplow, and A.M. Stuart, *A unified approach to spurious solutions introduced by time discretisation*, SIAM J. Numer. Anal., 28 (1991), pp. 1723–1751.

[19] C. Johnson, *Error estimates and adaptive time step control for a class of one-step methods for stiff ordinary differential equations*, SIAM J. Numer. Anal., 25 (1988), pp. 908–926.

[20] J.M. Sanz-Serna, *Numerical ordinary differential equations versus dynamical systems*, in The Dynamics of Numerics and the Numerics of Dynamics, D.S. Broomhead and A. Iserles, eds., Clarendon Press, Oxford, 1992.

[21] L. Shampine, *Tolerance proportionality in ODE codes*, in Numerical Methods for Ordinary Differential Equations (Proceedings), Lecture Notes in Mathematics, Springer-Verlag, Berlin, New York, 1987, pp. 118–136.

[22] H.J. Stetter, *Tolerance proportionality in ODE-codes*, in Proc. Second Conf. on Numerical Treatment of Ordinary Differential Equations, R. März, ed., Humboldt University, Berlin, 1980.

[23] J. Stewart, *Positive definite functions and generalizations: a historical survey*, Rocky Mountain J. Math., 6 (1976), pp. 409–434.

[24] A.M. Stuart and A.R. Humphries, *Model problems in numerical stability theory for initial value problems*, SIAM Review, 36 (1994), pp. 226–251.

[25] ———, *An analysis of local error control for dissipative, contractive and gradient dynamical systems*, Numerical Analysis Report NA-92-18, Stanford University, Stanford, CA, 1992.

[26] R. Temam, *Infinite Dimensional Dynamical Systems in Mechanics and Physics*, Springer, New York, 1989,

[27] J. Wilkinson, *The Algebraic Eigenvalue Problem*, Clarendon Press, Oxford, 1965,