NUMERICAL ANALYSIS PROJECT
MANUSCRIPT NA-92-18

NOVEMBER 1992

An Analysis of Local Error Control for Dissipative, Contractive and Gradient Dynamical Systems

by

A.M. Stuart
A.R. Humphries

NUMERICAL ANALYSIS PROJECT
COMPUTER SCIENCE DEPARTMENT
STANFORD UNIVERSITY
STANFORD, CALIFORNIA 94305

AN ANALYSIS OF LOCAL ERROR CONTROL FOR DISSIPATIVE, CONTRACTIVE AND GRADIENT DYNAMICAL SYSTEMS [1]

A.M. Stuart [2] and A.R Humphries[3] [4]

Abstract.

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 B which all trajectories eventually enter and remain inside. The exponentially contractive problems studied have a unique, globally attracting equilibrium point and thus they are also dissipative since the absorbing set 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. The aim is to find conditions under which numerical methods with local error control replicate these large-time dynamical features. The results are proved without recourse to asymptotic expansions for the truncation error.

Standard embedded Runge-Kutta pairs are analysed together with several non-standard error control strategies. These non-standard strategies are easy to implement and have desirable properties within certain of the classes of problems studied. 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 algebraically stable and explicit algebraically 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 algebraically 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 a positive power of 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 $\to 0$ as $\tau \to 0$. 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 whilst 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 Flows.

AMS SUBJECT CLASSIFICATIONS: 34C35, 34D05, 65L07, 65L20, 65L50.

January 27, 1993

[1] This work was funded by the Office of Naval Research under grant number N00014-92-J-1876 and by the National Science Foundation under grant number DMS-9201727

[2] Division of Applied Mechanics, Durand 252, Stanford University, Stanford CA 94305-4040, USA

[3] School of Mathematical Sciences, University of Bath, Bath BA2 7AY, UK.

[4] Department of Computer Science, Stanford University, Stanford CA 94305-2140, USA

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

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

where $u(t) \in \mathbb{R}^p$ for each $t \geq 0$, and $f\colon \mathbb{R}^p \to \mathbb{R}^p$ is assumed to be locally Lipschitz. We study variable time stepping strategies designed to control the local error incurred 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 explicit Runge-Kutta schemes 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:

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

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

and

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

Note that the sequence $\{V_n\}_{n=1}^{\infty}$ 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

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

or

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

where $\tau \ll 1$ is the error tolerance and e_0 is a scale factor to be specified later. The strategy (1.5) is known *as error per unit step* whilst the strategy (1.6) is known as *error per step.*

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

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

These matrices and vectors are assumed to be 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, $\bar{b}$ 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 extrapolation. The framework also includes methods where the error is estimated by step-halving: in this case the method for advancing $U_n \rightarrow U_{n+1}$ is simply to take two half steps of the method $U_n \rightarrow V_{n+1}$ and then form from this a method of order of accuracy one greater than that which takes U_n to V_{n+1}; hence $p = q + 1$.

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 $\mathbb{R}^p$ induced by the appropriate inner product – i.e. one inherited from one of the assumptions (D), (C), (E) or (G1-G5) whichwe now introduce. Here, and throughout the remainder of the paper,

$$B(v, \mathbf{6}) = \{\mathbf{u} \in \mathbb{R}^p : \|v - u\| < \delta\}: \tag{1.8}$$

The four conditions on the vector field $f(\bullet)$ whichwe consider are (D), (C), (E) and (G1–G5):

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

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

(E) $\exists r > 0 : \langle f(u), u\rangle < 0, \forall u \in \mathbb{R}^p$ satisfying $\|u\|^2 \geq r;$

(G1) $f(u) = -\nabla F(u)$, where $F \in C^2(\mathbb{R}^p, \mathbb{R});$
(G2) $F(u) \geq 0, \forall u \in \mathbb{R}^p$ and $|F(u)| \rightarrow \infty$ as $\|u\| \rightarrow \infty$.
(G3) $F(u) - F(v) \leq \langle f(u), v - u\rangle + c\|u - v\|^2, \forall\ u, v \in \mathbb{R}^p$.
(G4) Let $E = \{v \in \mathbb{R}^p : f(v) = 0\}$. Then E consists of isolated points.
(G5) There exists $D > 0$ such that $\|u - v\| \geq D$ for all $u, v \in E$ with $u \neq v$. Let

$$B(\delta) = \bigcup_{v \in E} B(v, \delta);$$

then for any $\delta > 0$ there exists $\varepsilon > 0$ such that

$$\inf_{x \notin B(\delta)} \|f(x)\| \geq \varepsilon.$$

See [19] for a review of the relevance of these classes of problems in numerical analysis and in applications. We now charaterise 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 3 a bounded absorbing set $\mathcal{B} \subset \mathbb{R}^p$ and, for each $U \in \mathbb{R}^p$, a time $t^* = t^*(U)$ such that $u(t) \in \mathcal{B}\ \forall t \geq t^*$.*

The following properties hold for (1.1):

Theorem ODE

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

(ii) under (C) every solution of (1.1) *satsifies u(t)* $\rightarrow 0$ *as* $t \rightarrow \infty$*. Thus* (1.1) *is dissipative with* $B = \bar{B}(0, \rho)$ *any* $\rho > 0$*;*

(iii) under (E), (1.1) *is dissipative with* $B = \bar{B}(0, r^{\frac{1}{2}})$;

(iv) under (G1-G5), for every $U \in \mathbb{R}^p$ $\exists v \in E$ *such that the solution of* (1.1) *satisfies* $u(t) \rightarrow v$ *as* $t \rightarrow \infty$*. If, in addition, E is bounded then* (1.1) *is dissipative with* $B = \{u \in \mathbb{R}^p : F(u) \leq \max_{v \in E} F(v) + p\}$*, any* $\rho > 0$*.*

Proof. The proof of (i) may be found in [19] and underlies much of the work in [20]. The proof of (ii) is straightforward. The proof of (iii) is similar to the proof of (i). The proof of (iv) may be found in [8]. □

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 [19]. 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 *algebraically stable* and in section 2 we construct *explicit* embedded pairs which are algebraically stable in this sense. Note that fixed step algebraically stable schemes are necessarily implicit. In addition, we prove an order barrier $\min\{p, q\} \leq 4$ for explicit algebraically stable embedded pairs of order (p, q), with non-negative b_i.

In section 3 we conisder 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 $\inf_{n \geq 0} \Delta t_n > 0$ since, without this, the time integration may terminate at a finite time.

It is known that fixed time-stepping algebraically stable Runge-Kutta methods define dissipative numerical methods for (D) and (C) respectively - see [1], [12], [19]. In section 4 we show that 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 a positive power of 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 sections 5 and 6 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 section 4.

We next consider condition (E); we state and prove a discrete analogue of Theorem

ODE(iii) in section 5 for the simplified error per unit step strategy (2.30), (2.31), (2.32) together with (1.5) – see Theorem E.

In section 6 we consider gradient systems under (G1)–(G5). (G1) is the standard gradient assumption, (G2) ensures global existence, uniqueness and boundedness of solutions to (1.1) whilst (G3) is equivalent to a one-sided Lipschitz condition [13]. (G4) and (G5) are structural stability conditions on the gradient system.

For τ sufficiently small, but independently of initial data, we prove in Theorem G1 that the simplified order *(p,* 1) error control scheme (2.24), (2.27) together with (1.5) forces the numerical solution to enter and remain in a ball centred on one of the equilibria in *E;* the radius of the ball $\rightarrow 0$ as $\tau \rightarrow 0$. If E is bounded then dissipativity follows. In addition, a modification of this error control is proposed which actually ensures that the solution is driven to an equilibrium point as $n \rightarrow \infty$. This is based on error per unit step control relative to a discrete time derivative – see Theorem G2.

Finally, in section 7, we present some numerical results to illustrate the theory.

The work contained here is inspired by the papers of [9] and [15] 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 an extension of the work of Stetter [18] (see also [11], [17]) where it is shown that, *over fixed time intervals* $0 \leq t \leq T$, the error is proportional to some positive power of the tolerance – essentially a convergence result for error controlled schemes as $\tau \rightarrow 0$; here we show that the "error" in the *asymptotic behaviour* $(t \rightarrow \infty)$ is proportional to a positive power of the tolerance, 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 [14] 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 [16], 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 towards substantiating the claim in [16].

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 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* algebraically stable embedded pairs exist but an order barrier of $\min\{p, q\} \leq 4$ is proved for such explicit methods with non-negative weights b_i. See Corollary 2.7.

- New simplified and computationally inexpensive embedded pairs are introduced with order *(p,* 1), p arbitrarily large, which are algebraically stable and computationally inexpensive. These embedded pairs may be explicit. See Example 2.9.

- New error control strategies are introduced for gradient systems for which the error control is relative to a discrete time-derivative. See section 6.

- For certain algebraically stable embedded pairs applied to dissipative, contractive and gradient systems we prove that the underlying long time behaviour of the differential equation is inherited by the error controlled scheme – see Theorems DC1, DC2 (section 4), Theorem E (section 5) and Theorems G1 and G2 (section 6).

- 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. 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 our knowledge this is the first rigorous treatment of error control strategies over long time intervals.

2. Algebraically Stable Embedded Pairs. 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

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

From (2.1), (1.4) it follows that

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

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

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

or

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

where

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

Using (2.5), equation (1.3) may be re-written as

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

Furthermore, (1.2), (2.2) gives

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

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

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

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, $\bar{b}$. In particular, recalling the definition of algebraic stability of a fixed-step method A, b from [1] and of DJ-reducibility from [10], we make the following definition for variable-step embedded pairs:

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

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

$$(2.8) \qquad \left\{ \begin{array}{c} \hat{B} = diag\{\hat{b}\}, \\ \hat{M} = \hat{B}\hat{A} + \hat{A}^T\hat{B} - \hat{b}\hat{b}^T, \\ \tilde{M} = \hat{B}A + A^T\hat{B} - \hat{b}\hat{b}^T. \end{array} \right\}$$

We shall 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

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

LEMMA 2.2. *Given $e_0 \in I \backslash (o)$ for which $\hat{B}$ is positive definite, the embedded pair A, b, $\bar{b}$ is algebraically stable if M is positive definite on V. Conversely if for each $e_0 \in I$ there exists $x \in V$ for which $x^T M x < 0$ then the embedded Runge-Kutta pair A, b, $\bar{b}$ is not 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

$$\begin{aligned} x^T\hat{M}x &= x^T\hat{B}\hat{A}x + x^T\hat{A}^T\hat{B}x - x^T\hat{b}\hat{b}^Tx \\ &= 2(\hat{B}x)^T\hat{A}x - (\hat{b}^Tx)^2 \\ &= 2(\hat{B}x)^T(Ax + e_0e(\bar{b} - b)^Tx) - (\hat{b}^Tx)^2 \\ &= 2(\hat{B}x)^TAx + 2e_0(\bar{b} - b)^Tx(x^T\hat{B}e) - (\hat{b}^Tx)^2 \\ &= x^T\tilde{M}x + 2e_0(\bar{b} - b)^Tx(x^T\hat{B}e). \end{aligned}$$

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

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

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

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

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

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

then

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

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

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

$$x^T\hat{M}x = x^T\tilde{M}x \ on \ V.$$

□

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 V. 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 V interleave the eigenvalues of $\tilde{M}$ [21] and so we trivially find from Lemma 2.2 that:

Corollary 2.3. *If $\tilde{M}$ has two negative eigenvalues for all $e_0 \in I$ then the embedded pair A, $b, \bar{b}$ is not algebraically stable.*

Lemma 2.2 and Corollary 2.3 are suggestive of an order barrier for explicit algebraically stable methods and we now prove that if A, b, $\bar{b}$ has order (p, q) with $\min\{p, q\} \geq 5$ and $b_i \geq 0$ then it is not algebraically stable. Preceding this theorem are two lemmas needed in the proof:

Lemma 2.4. *If A, $b, \bar{b}$ has order (p, q) with $\min\{p, q\} \geq 5$ then $\hat{A}$, $\hat{b}$ has order ≥ 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 tedious manipulations of the order conditions, using (2.2). □

The following definition will be needed:

DEFINITION 2.5. *The embedded pair A, b, $\bar{b}$ is 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 2.6. *Assume that the embedded pair A, b, $\bar{b}$ is explicit, irreducible, algebraically stable and has order (p, q) with* $\min\{p, q\} \geq 5$. *Let* $T = \{j \in 2 : \hat{b}_j = 0\}$. *Then:*

(i) $\exists J \geq 3 : T = \{j : 1 \leq j \leq J\}$.
(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) $b_j \neq 0$, $b_j \neq \bar{b}_j$ Vj $\in T$;
(iv) $a_{ij} = e_i b_j, \hat{a}_{ij} = 0, \forall i, j : i \notin T, j \in T$;
(v) $\exists j \in T : b_j <$ O.

Proof In the following we define

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

Since $\hat{A}$, $\hat{b}$, A, b and A, $\bar{b}$ have order at least five, it follows from the proof of Lemma IV.13.12 in [10] that

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

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

$$\hat{d}_i = 0 \;\; i \notin T. \tag{2.14}$$

Also

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

$$= \sum_{j=1}^{k} a_{ij} c_j + e_0 e_i (\sum_{j=1}^{k} c_j \bar{b}_j - c_j b_j) = \sum_{j=1}^{k} a_{ij} c_j$$

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

$$(2.15) \qquad d_i = 0 \; \forall i \notin T.$$

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

Because the method $\hat{A}, \hat{b}$ is algebraically stable it is DJ-reducible [10] to a method with $\hat{b}_i > 0$. Thus it follows that

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

and that

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

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

$$(2.18) \qquad b_j \neq 0 \; \textit{and} \; \bar{b}_j \neq b_j \; \forall \; j \in T.$$

Combining (2.16), (2.17) gives

$$(2.19) \qquad a_{ij} = e_i b_j, \forall i, j : i \notin T, j \in T.$$

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)

$$(2.20) \qquad d_2 = \sum_{j=1}^{k} a_{2j} c_j = -c_2^2/2 = 0$$

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 $1 \leq j \leq J$ 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),

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

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 $J \geq 3$. 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 decude that $J \geq 3$.

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

It follows automatically from Lemma 2.6(v) that

Corollary 2.7. *There are no explicit algebraically stable embedded pairs with non-negative weights b_i and order (p, q) satisfying* $\min\{p, q\} \geq 5$.

Proof. Assume to the contrary that $p, q \geq 5$ and the weights $b_i \geq 0$. By Lemma 2.5(v) we obtain a contradiction. □

Remark Most standard methods are constructed with the simplifying assumption that the $b_i \geq 0$ [3]; thus Corollary 2.7 maybe of interest.

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

Example 2.8.

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

$$A = (\quad ; \quad \left.\begin{matrix} 0 \\ 0 \end{matrix}\right), b = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \bar{b} = \begin{pmatrix} \frac{1}{2} \\ \frac{1}{2} \end{pmatrix}.$$

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 [10] to the backward Euler scheme and hence algebraically stable. □

Example 2.9.

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

$$\eta_1 = U_n,$$

$$\eta_2 = U_n + \Delta t_n f(\eta_1),$$

$$\eta_3 = U_n + \frac{\Delta t_n}{4}[f(\eta_1) + f(\eta_2)],$$

$$U_{n+1} = U_n + \frac{\Delta t_n}{2}[f(\eta_1) + f(\eta_2)],$$

$$V_{n+1} = U_n + \frac{\Delta t_n}{6}[f(\eta_1) + f(\eta_2)] + \frac{2\Delta t_n}{3} f(\eta_3).$$

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

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

The new method is DJ-reducible to the implicit mid-point rule and hence algebraically stable. □

Example 2.10.

Because of the order barrier established in Corollary 2.7, it is not possible to find explicit algebraically stable embedded pairs with order $(p, p+1)$, $p \geq 5$ and positive weights b_i. However it is possible to seek methods of order $(p, 1)$ for arbitrarily large p. Let

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

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

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

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

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

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

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

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

or

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

respectively so that the original scheme is close to the backward Euler scheme and hence the embedded pair is algebraically stable. Notice that whilst this error control is non-standard, 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. This does not greatly increase computational expense.

If (2.24) is the explicit Euler scheme and $\theta = 1/2$ then the method is simply the order $(1,2)$ pair of Example 2.8. 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],

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

The method (2.27) has

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

and

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

Clearly

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

however

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

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

EXAMPLE 2.11.

The methods of Example 2.10 can be generalised as follows. Let

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

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

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). This particular method will be considered in detail in section 5, with the time-step chosen according to the error per unit step criteria (1.5) with $e_0 = \theta^{-1}$. The assumption that $\phi \in [\frac{1}{2}, 1]$ is necessary and sufficient for the embedded pair to be 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 2.8. 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 2.9.

Notice that the methods of Example 2.9 correspond to choosing $\phi = 1$. Setting $\phi \neq 1$ allows higher order error control than is possible with the methods of Example 2.9, but at the cost of introducing an extra stage to the Runge-Kutta method.

The order barrier; $\min(p, q) \leq 2$ for (2.30)–(2.32) can be established by manipulating the order conditions. It is also easy to see that if $p \geq 2$ 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). In order 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 $n \geq 0$ (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 $n \geq 0$.

Furthermore, for the kind of problems in which are are interested here, the underlying differential equation has solutions defined for all $t \geq 0$. 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 – i.e. $\inf_{n \geq o} \Delta t_n > 0$.

In this section we describe a general framework in which we analyse these issues. Note that the error control criteria (1.5) or (1.6) determine Δt_n implicitly as a function of Un, once a method has been determined to ensure their satisfaction. Thus we may think of Δt_n in the following way:

$$\Delta t_n = \Gamma(U_n, \tau).$$

We shall not need to specify a particular Γ in this paper – we shall simply show that, under suitable conditions, (1.5) or (1.6) *can* be satisfied. However, it is worth noting that, in typical impmentations $\Gamma(\bullet, \tau)$ is a discontinuous function, since (1.5) or (1.6) is usually achieved through an iterative procedure in which the prospective time-step is determined by decreasing the candidate time-step by a constant factor, until (1.5) or (1.6) is satisfied.

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

NOTATION 3.1. *We denote the class of Lipschitz continuous functions mapping $I\!R^p$ into $I\!R^p$ and satisfying (D), (C), (E) or (G1–G5) by 3(D), 3(C), 3(E) and 3(G), respectively.*

DEFINITION 3.2. *Given an embedded pair* (1.2)–(1.4), (1.5) *or* (1.2)–(1.4), (1.6) *a sequence* $\{(U_n^T, V_n^T, \Delta t_n)\}_{n=0}^{\infty}$ *with* $(U_n^T, V_n^T, \Delta t_n) \in I\!R^{2p+1}$ *satisfying* (1.2)–(1.4), (1.5) *or* (1.2)–(1.4), (1.6) *is admissible if* $\inf_{n\geq 0} \Delta t_n > 0$. *An embedded pair is $\mathcal{F}(\bullet)-$ admissible if for every function $f \in \mathcal{F}(\bullet)$ and all $U \in I\!R^p$ there ezists $\tau^* = \tau^*(f, U)$ such that the embedded pair has an admissible sequence for each $\tau \in (0, \tau^*)$. The pair is $\mathcal{F}(\bullet)$ – globally admissible if a τ^* may be found which 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*.

We now address solvability of the Runge-Kutta equations. In the following it will be useful to define

$$\tilde{a} = \max_{1\leq i,j\leq k} |a_{ij}| \text{ and } \tilde{b} = \max_{1\leq i\leq k} |b_i|. \tag{3.1}$$

LEMMA 3.3. *For any $\gamma > 1$ let*

$$Q(X) = \{u \in I\!R^p : ||u - X|| \leq \Delta t\tilde{a}k\gamma||f(X)||\},$$

$L(X)$ be the Lipschitz constant for $f(\bullet)$ on Q(X) and let K(X) = $\sup_{u\in Q(X)} ||f(u)||$. Then, for all $\Delta t \in [0, \Delta t_c(X))$, where

$$\Delta t_c(X) = \min_{\Delta t\in I\!R^+} : \Delta t = \frac{1-\gamma^{-1}}{\tilde{a}kL(X)}, \tag{3.2}$$

there exists a unique solution $\{\eta_i\}_{i=1}^k, \eta_i \in I\!R^p$ of the equations

$$\eta_i = X + \Delta t\sum_{j=1}^{k} a_{ij} f(\eta_j) \tag{3.3}$$

satisfying $\eta_i \in Q(X)$. Furthermore if $\{\eta_i^l\}_{i=1}^k, l = 1, 2$ are solutions of (3.3) 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

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

and

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

Proof Note that the construction of Δt_c in (3.2) is slightly non-trivial since $L(X)$ depends upon Δt. Nonetheless it is clear that $\Delta t_c > 0$ and that, furthermore,

$$\Delta t < \frac{1-\gamma^{-1}}{\tilde{a}kL(X)} \tag{3.4}$$

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

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

If $\{\eta_i^l\}_{i=1}^k$, $l = 1, 2$ solve (3.3) then

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

Hence

$$\|\eta_i^1 - \eta_i^2\| = \|\sum_{j=1}^{k} a_{ij} (\Delta t^1 f\,(\eta_j^1) - \Delta t^2 f(\eta_j^2))\|$$

$$\leq \tilde{a} \sum_{j=1}^{k} [\Delta t^1 \|f(\eta_j^1) - f(\eta_j^2)\| + |\Delta t^1 - \Delta t^2| \|f(\eta_j^2)\|]$$

$$\leq \tilde{a} k L(X) \Delta t^1 \max_{1 \leq j \leq k} \|\eta_j^1 - \eta_j^2\| + \tilde{a} k K(X) |\Delta t^1 - \Delta t^2|.$$

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

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

□

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

$$\xi_i = X + \Delta t \sum_{j=1}^{k} a_{ij} f(\xi_j), \; i = \; 1, \dots, k \tag{3.5}$$

$$W = X + \Delta t \sum_{j=1}^{k} b_j f(\xi_j), \tag{3.6}$$

and

$$V = X + \Delta t \sum_{j=1}^{k} \bar{b}_j f(\xi_j). \tag{3.7}$$

Note that these functions are well-defined by Lemma 3.3 for any $X \in \mathbb{R}^p$ and any $\Delta t \in [0, \Delta t_c(X))$. Hence we may define $G : [0, \Delta t_c(X)) \times \mathbb{R}^p \to \mathbb{R}$ by

$$G(\Delta t, X) = \frac{\|W - V\|}{\Delta t}, \tag{3.8}$$

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

$$H(\Delta t, X) = \Delta t G(\Delta t, X). \tag{3.9}$$

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 3.4. *The functions $G(\Delta t, X)$ and $H(\Delta t, X)$ satisfy $G(0, X) = H(0, X) = 0$ $\forall X \in \mathbb{R}^p$ and are Lipschitz continuous 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 $\xi_j\,(0, X) = X$ $\forall X \in \mathbb{R}^p$ 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

$$|G(\Delta t^1, X) - G(\Delta t^2, X)|$$

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

$$\leq \|\sum_{j=1}^k (b_j - \bar{b}_j)[f(\xi_j\,(X, \Delta t^1)) - f(\xi_j\,(X, \Delta t^2))]\|$$

$$\leq kL(X) \max_{1 \leq j \leq k} |b_j - \bar{b}_j|\, \|\xi_j(X, \Delta t^1) - \xi_j(X, \Delta t^2)\|.$$

Thus, by Lemma 3.3,

$$|G(\Delta t^1, X) - G(\Delta t^2, X)| \leq C_3 L(X) K(X) |\Delta t^1 - \Delta t^2|, \tag{3.10}$$

with C_3 independent of X. Thus $G(\bullet, X)$ is Lipschitz.

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

We can use Lemma 3.4 to establish admissibility.

THEOREM 3.5. *Assume that $\exists \tau^* = \tau^*(U) > 0$ and a compact set $I = I(U) \subset \mathbb{R}^p$ 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 $\forall n \geq 0$. Then $\exists$ an admissible sequence satisfying* (1.2)–(1.4), (1.5).

Proof. Let

$$\Delta t_c = \inf_{X \in I(U)} \{\Delta t_c(X), \frac{\tau}{|e_0| C_3 L(X) K(X)}\} \tag{3.11}$$

where C_3 is defined in (3.10) and $\Delta t_c(X)$ in (3.4). Notice that $\Delta t_c > 0$ since both the Lipschitz constant for f and and $K(\bullet)$are bounded on any compact set. Now consider (1.2)–(1.5) with $\Delta t_n \equiv \Delta t_c$ $\forall n \geq 0$. Assume, for the purposes of induction, that $\exists$ 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 3.3 and

(3.11) $\exists$ a solution $\{\eta_i\}_{i=1}^{k}$ to (1.2) and thus a vector $U_{N+1} \in \mathbb{R}^p$ satisfying (1.3) with $n = N$ and $\Delta t_N = \Delta t_c$. By Lemma 3.4 and (3.11)

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

So, by construction, the error control criteria is satisfied. Thus $\exists$ 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 $\inf_{n \geq 0} \Delta t_n = \Delta t_c > 0$. □

The following Corollary is immediate. Furthermore, Theorem 3.7 and Corollary 3.8 follow similarly for the error control scheme (1.2)–(1.4), (1.6).

COROLLARY 3.6. *Assume that, for every function* $f \in \mathcal{F}(\bullet)$ *and all* $U \in \mathbb{R}^p$ $\exists \tau^* = \tau^*(f, U) > 0$ *and a compact set* $I = I(f, U) \subset \mathbb{R}^p$ *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 Qn* ≥ 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 3.7. *Assume that* $\exists \tau^* = r^*(U) > 0$ *and a compact set* $I = I(U) \subset \mathbb{R}^p$ *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* $\forall n \geq 0$*. Then 3 an admissible sequence satisfying* (1.2)–(1.4), (1.6).

COROLLARY 3.8. *Assume that, for every function* $f \in \mathcal{F}(\bullet)$ *and all* $U \in \mathbb{R}^p$ $\exists \tau^* = \tau^*(f,\ U) > 0$ *and a compact set* $I = I(f,\ U) \subset \mathbb{R}^p$ *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* $\forall n \geq 0$*. 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.*

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 Δt_{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 if often imposed by an actual implementation of an embedded pair in a software code to prevent enormous steps from begin taken – see [7] for a discussion of this point. Furthermore, we make the following assumption, noting that any algebraically stable method $\hat{A}$, $\hat{b}$ is DJ-reducible to one with positive weights [10]:

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

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

and

$$w^T d = 1$$

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

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 2.8 – 2.11 all satisfy this condition.

In order to clearly state the sense in which the numerical method inherits the properties of the differential equation for problems under (D), (C) and (E) we make the following definition.

Definition 4.1. *The embdedded pair* (1.2)–(1.4), (1.5) *or* (1.2)–(1.4), (1.6) *is said to be* $\mathcal{F}(\bullet)-$ *dissipative if it is 3(o)-admissible and if 3* $\tau_c = \tau_c(U)$ *and an absorbing set* $\mathcal{B}_\tau \subset \mathbb{R}^p$, *independent of U and uniformly bounded us* $\tau \longrightarrow 0$, *such that, for* $\tau \in (0, \tau_c)$ *and every admissible sequence with* inf $\Delta t_n \geq \bar{\Delta t} > 0$ $\exists n^* = n^*(U, \tau, \Delta\bar{t})$: $U_n \in \mathcal{B}_\tau$ $\forall n \geq n^*$. *The pair* is $\mathcal{F}(\bullet)$ – *globally dissipative if in addition,* τ_c *is indepdendent of U.*

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 algebraically stable and satisfies condition (K) and that* $\Delta t_n \leq \Delta t_{max}$ $\forall n \geq 0$. *Then the embedded pair is 3(D) – globally dissipative and 3(C) – globally dissipative. In the second case it follows that* $\|u\|^2 \leq c\tau$ *Vu* $\in \mathcal{B}_\tau$, *the absorbing set.*

Theorem DC2 *Consider* (1.2)–(1.4) *with error control* (1.6). *Assume that A, b,* $\bar{b}$ *is algebraically stable and satisfies condition (K) and that* $\Delta t_n \leq \Delta t_{max}$ $\forall n \geq 0$. *Furthermore, ussume that the unique solution of the Runge-Kutta equations* (1.2) *satisfying* $\eta_i \in \mathcal{Q}(U_n)$ *constructed in Lemma 3.3 is used for each n ≥ 0. Then the embedded pair is 3(D)- dissipative and 3(C)- dissipative. In the second tase it follows that* $\|u\|^2 \leq c\tau$ Vu $\in \mathcal{B}_\tau$, *the absorbing set.*

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

We now derive a preliminary lemma for the scheme (1.2)–(1.4), using the representation (2.6), (2.7). Our approach is motivated by the papers [1] and [12] 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 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], [10]. If such a non-trivial reduction is possible then we define a reduced Runge-Kutta method from $\hat{A}$, $\hat{b}$ by removing η_j, $j \in T$, (where T is defined in Lemma 2.6) 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 2.8 and 2.9) that the reducibility of the method $\hat{A}$, $\hat{b}$ does not imply the reducibility of the method A, b, $\bar{b}$.

Lemma 4.2. *Let the embedded pair A, b,* $\bar{b}$ *be algebraically stable and satisfy (K). Then, under the structural assumption (D) on f, solutions of the embedded pair* (1.2)–(1.4) *satisfy*

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

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

$$\|U_{n+1}\|^2 = \|U_n\|^2 + 2\Delta t_n \sum_{j=1}^{k} \hat{b}_j (U_n, fj) \ + \ \Delta t_n^2 \sum_{i,j=1}^{k} \hat{b}_i \hat{b}_j \langle f_i, f_j \rangle$$

$$+2\Delta t_n \langle E, U_n \rangle + 2\Delta t_n^2 \sum_{j=1}^{k} \hat{b}_j \ (E, \ f_j\rangle + \Delta t_n^2 \|E\|^2.$$

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

$$\|U_{n+1}\|^2 = \|U_n\|^2 + 2\Delta t_n \textstyle\sum_{i=1}^{k} \hat{b}_i \langle \eta_i, f_i \rangle$$

$$-2\Delta t_n^2 \textstyle\sum_{i,j=1}^{k} \hat{b}_i \hat{a}_{ij} \langle f_i, f_j \rangle - 2\Delta t_n^2 \sum_{i=1}^{k} \hat{b}_i e_i \langle E, f_i \rangle \tag{4.3}$$

$$+\Delta t_n^2 \textstyle\sum_{i,j=1}^{k} \hat{b}_i \hat{b}_j \langle f_i, f_j \rangle + 2\Delta t_n \langle E, U_n \rangle + 2\Delta t_n^2 \sum_{j=1}^{k} \hat{b}_j \ (E, \ f_j \rangle + \Delta t_n^2 \|E\|^2.$$

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

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

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

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

Noting that

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

and combining (4.3)–(4.5) gives

$$\|U_{n+1}\|^2 = \|U_n\|^2 + 2\Delta t_n \sum_{i=1}^{k} \hat{b}_i \langle \eta_i, f_i \rangle - \Delta t_n^2 \sum_{i,j=1}^{k} \hat{m}_{ij} \langle f_i - x_i E, \ f_j - x_j E \rangle - 2\Delta t_n^2 \sum_{i,j=1}^{k} \hat{m}_{ij} x_i \langle E, f_j \rangle$$

$$+\Delta t_n^2 \sum_{i,j=1}^{k} \hat{m}_{ij} x_i x_j \|E\|^2 + 2\Delta t_n^2 \sum_{j=1}^{k} \hat{b}_j (1-e_j)\langle E, f_j\rangle + \Delta t_n^2 \|E\|^2 + 2\Delta t_n \sum_{i=1}^{k} d_i \langle E, \eta_i\rangle$$

$$-2\Delta t_n^2 \sum_{i,j=1}^{\mathrm{k}} \hat{a}_{ij} d_i \langle E, f_j\rangle - 2\Delta t_n^2 \sum_{i=1}^{k} d_i e_i \|E\|^2.$$

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

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

we obtain from Lemma 4.2, in the error per unit step case (1.5) (which implies (2.3))

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

where

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

and we have assumed that At,, $\le$ At,,,. If we define

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

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

Lemma 4.3. *Assume that $\Delta t_n \le \Delta t_{max}$ Qn ≥ 0. Then, under the conditions of Lemma* 4.2, *the embedded Runge-Kuttu pair* (1.2)–(1.4), (1.5) *is 3(D) – globally admissible and 3(C) – globally admissible.*

Proof. Note that *3(D)* global admissibility implies $\mathcal{F}(C)$ 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

$$(4.9) \qquad R = \frac{\tilde{\alpha} + \rho}{\tilde{\beta}} + \Delta t_{max} C$$

where

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

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

and

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

Let

$$I(U) = \left\{ u \in I\!R^p : ||u||^2 \leq \max\{||U||^2, R\} \right\}.$$

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

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

then (4.6), which follows from Lemma 4.2, gives

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

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

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

then

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

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

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

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 3.6, since τ^* is independent of U. □

Proof of Theorem DC1 The *3(D)* and $\mathcal{F}(C)$ 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

$$\mathcal{B}_\tau = \{u \in I\!R^p : \|u\|^2 \leq R\}, \tag{4.11}$$

where R is defined by (4.9). Take any $p > 0$. Whilst

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

we have from (4.6)

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

Alternatively, if

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

it follows that

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

and then, as in the proof of Lemma 4.3, that

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

Now, if $\|U_n\|^2 \leq R$ and (4.12) holds then, from (4.13) we have that $\|U_{n+1}\|^2 \leq R$. On the other hand, if (4.14) holds then $\|U_{n+1}\|^2 \leq R$ 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 $\|U_{n+1}\|^2 > R\ \forall n \geq 0$, then (4.13) holds for all $n \geq 0$ and hence, since the sequence is admissible, $\exists \bar{\Delta t} > 0$:

$$\|U_N\|^2 \leq \|U\|^2 - 2\bar{\Delta t}\rho N,\ N \geq 0.$$

Letting $N \to \infty$ gives a contradiction. Thus we have *3(D)* and $\mathcal{F}(C)$ global dissipativity.

In the second case where (C) holds we have that $\alpha = 0$ so that $\tilde{\alpha}$, defined by (4.7), is $\mathcal{O}(\tau)$. The proof proceeds as for (D) except that now we take $\rho = \tau$ in the construction of R given by (4.9). Clearly $\gamma_i = \mathcal{O}(\tau^{\frac{1}{2}})$ and by the Lipschitz continuity of f it follows that

$$f^* = \max_i \max_{\|\eta_i\| \leq \gamma_i} \|f(\eta_i) - f(0)\| \leq c\tau^{\frac{1}{2}}$$

for some constant c independent of τ. Thus (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 subsection 4.1 to the error per unit step case. 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

$$I(U) = \{u \in \mathbb{R}^p : \|u\|^2 \leq \max\{\|U\|^2, R_1\}\}. \tag{4.17}$$

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 3.3,

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

and $L(X)$ is the Lipschitz constant for f described in Lemma 3.3.

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

LEMMA *4.4. Assume that $\Delta t_n \leq \Delta t_{max}$ $\forall n \geq 0$. Furthermore, assume that the unique solution of the Runge-Kutta equations* (1.2) *satisfying $\eta_i \in Q(U_n)$ constructed in Lemma 3.3 is used. Then, under the conditions of Lemma 4.2, the embedded Runge-Kuttu pair* (1.2)-(1.4), (1.6) *is 3(D)-admissible and 3(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 $\Delta t_N \geq \tau^2$ then $\|E\| \leq \tau$ and (4.6) follows just as in the error per unit step case. Thus, if $\Delta t_N \geq \tau^2$ we deduce that, as in the Proof of Theorem DC1, either

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

or

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

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

If $\Delta t_N \leq \tau^2$ 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 3.3,

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

Applying Lemma 3.3 we obtain

$$||\eta_j - U_{n+1}|| \leq ||\eta_j - U_n|| + ||U_{n+1} - U_n||$$

$$\leq \Delta t\tilde{a}k\gamma||f(U_n)|| + \Delta t\tilde{b}k\gamma||f(U_n)||.$$

Hence

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

Using $\tau \leq \tau_1^*$ we obtain

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

Thus, either (4.19) holds, or

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

which is equivalent to (4.20).

Hence we have shown that (4.19), (4.20) are true regardless of whether $\Delta t_n \leq \tau^2$ or $\Delta t_n \geq \tau^2$. From these it follows simply that $U_{N+1} \in I(U)$ and the induction is complete. The proof then follows from Corollary 3.8, 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

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

Because of the dependency of τ_1^* on U only $\mathcal{F}(D)-$ and $\mathcal{F}(C)-$ admissibility are obtained. □

5. Weaker Forms of Dissipativity. In this section we analyse error control schemes under the assumption (E). As in the previous section, we assume that there is an upper bound Δt_{max} on the time-step, which need *not* **be** small. We are unable to prove results for general algebraically stable Runge-Kutta pairs, and must work in a restricted class. Throughout this section we consider the explicit methods of Example 2.11 with the time-step chosen according to (1.5) with $e_0 = \theta^{-1}$.

We shall prove Theorem E for the error per unit step scheme (2.30), (2.31), (1.5), a discrete analogue of Theorem ODE(iii). However, the numerical analysis of this problem turns out to be more subtle than the analysis under the dissipativity assumption (D) because it is necessary to choose τ depending upon initial data. The discussion in Example 7.2 indicates why this is necessary.

Throughout the remainder of this section the definition of R_2 is given by (5.1) with r defined as in (E):

$$(5.1) \qquad R_2 = \left[r^{\frac{1}{2}} + \Delta t_{max}(1-\phi) \max_{\|\xi\| \leq r} \|f(\xi)\| + \Delta t_{max}(1-\phi)\tau\right]^2 .$$

THEOREM E *Consider the embedded pair* (2.30), (2.31) *with error control* (1.5). *Assume that* $\Delta t_n \leq \Delta t_{max}$ $\forall n \geq 0$. *Then the embedded pair is 3(E)-dissipative.*

It is possible to generalise Theorem E to error per step strategies of the form (1.6) as done in Theorem DC2 for the stronger form of dissipativity. However this is not particularly illuminating and we omit the details.

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

LEMMA 5.1. *Assume that* $\Delta t_n \leq \Delta t_{max}$ $\forall n \geq 0$. *Then the pair* (2.30), (2.31) *and* (1.5) *is F(E)-admissible.*

Proof Let

$$I(U) = \{u \in I\!R^p : \|u\|^2 \leq M\}$$

where

$$M = \max\{\|U\|^2, R_2\}\}$$

and R_2 is defined by (5.1). Since (2.30), (2.31) is explicit, f is Lipschitz and $\Delta t_n \leq \Delta t_{max}$ there exists $M' > 0$ such that if $\|U_n\|^2 \leq M$ then $\|U_{n+1}\|^2 \leq M'$. Let

$$X = \{x \in I\!R^p: r \leq \|x\|^2 \leq M'\}$$

and note that X is compact. Thus, since $\langle f(u), u\rangle < 0$ for all $u \in X$ there exists $\varepsilon > 0$ such that $\langle f(u), u\rangle \leq -\varepsilon$ for all $u \in X$. Now let

$$(5.2) \qquad \tau \leq \tau(M) = \frac{2\varepsilon}{1+M}.$$

For the purposes of induction, assume that $U_N \in I(U)$ noting that this is true for $n = 0$ by construction of $I(U)$. Note that, by (2.32),

$$(5.3) \quad \|U_{N+1}\| = \|\eta + (1-\phi)(U_{N+1} - U_N)\| \leq \|\eta\| + (1-\phi)\|U_{N+1} - U_N\|,$$

and that the defining equations (2.30), (2.31) and (1.5) imply that, since $e_0 = \theta^{-1}$,

$$(5.4) \qquad \|\tilde{f}(U_n; \Delta t_n) - f(\eta)\| \leq \tau.$$

Now using (2.30), (5.3), (5.4) and the fact that $\Delta t_n \leq \Delta t_{max}$ it follows that

$$(5.5) \qquad \|U_{N+1}\|^2 \leq R_2 \leq \max\{\|U\|^2, R_2\}, \quad if \ \|\eta\|^2 \leq r.$$

Now suppose that $||\eta||^2 \geq r$. Taking the inner product of (2.30) with η we obtain

$$(5.6) \qquad \langle U_{N+1} - U_N, \eta\rangle = \Delta t_N\langle f(\eta), \eta\rangle + \Delta t_N\langle \tilde{f}(U_N; \Delta t_N) - f(\eta), \eta\rangle.$$

Noting that

$$\begin{aligned} \langle U_{N+1} - U_N, \eta\rangle &= \langle U_{N+1} - U_N, (1-\phi)U_N + \phi U_{N+1}\rangle \\ &= \frac{1}{2}\left[||U_{N+1}||^2 - ||U_N||^2\right] + (\phi - \frac{1}{2})||U_{N+1} - U_N||^2 \\ &\geq \frac{1}{2}||U_{N+1}||^2 - \frac{1}{2}||U_N||^2 \end{aligned}$$

for $\phi \in [\frac{1}{2}, 1]$, together with (5.4) we see that (5.6) implies

$$(5.7) \qquad \frac{1}{2}||U_{N+1}||^2 \leq \frac{1}{2}||U_N||^2 - \Delta t_N\varepsilon + \Delta t_N\tau||\eta||$$

for $||\eta||^2 \geq r$. Since, by (2.32)

$$(5.8) \qquad \begin{aligned} &||\eta|| \leq (1-\phi)||U_N|| + \phi||U_{N+1}||, \\ &\Rightarrow ||\eta|| \leq \tfrac{1}{2} + \tfrac{1}{2}(1-\phi)||U_N||^2 + \tfrac{\phi}{2}||U_{N+1}||^2 \end{aligned}$$

we obtain,

$$(5.9) \qquad \begin{aligned} &(1 - \Delta t_N\tau\phi)||U_{N+1}||^2 \leq \\ &(1 + \Delta t_N\tau(1-\phi))||U_N||^2 - \Delta t_N(2\varepsilon - \tau), \quad ||\eta||^2 \geq r. \end{aligned}$$

Now since $\tau \leq \tau(M)$ and $||U_N||^2 \leq ||U||^2 \leq M$ by assumption, (5.2) implies that

$$\tau||U_N||^2 \leq 2\varepsilon - \tau,$$

and thus from (5.9),

$$(1 - \tau\Delta t_N\phi)||U_{N+1}||^2 \leq (1 - \tau\Delta t_N\phi)||U_N||^2,$$

which implies that

$$(5.10) \qquad ||U_{N+1}||^2 \leq ||U_N||^2 \leq \max\{||U||^2,\ R_2\},\ \ if\ \ ||\eta||^2 \geq r.$$

Equations (5.5), (5.10) complete the induction. Since $\tau(M)$ depends on U, Corollary 3.8 with $\tau^* = \tau(M)$ gives $\mathcal{F}(E)-$admissibility. □

Proof of Theorem E The $\mathcal{F}(E)-$admissibility follows from Lemma 5.1. Thus we need only exhibit an absorbing set for every admissible sequence. Let $\tau < \tau_c = \tau(M)$ defined by (5.2). Define

$$(5.11) \qquad \mathcal{B}_\tau = \{u \in I\!R^p : ||u||^2 \leq R_2\},$$

where R_2 is defined by (5.1). We show that $\mathcal{B}_\tau$ is positive invariant, by an argument identical to that in the Proof of Theorem DC1, using (5.5) and (5.10); thus if $U_n \in \mathcal{B}_\tau$ then $U_{n+1} \in \mathcal{B}_\tau$. Hence it remains to show that U_n enters $\mathcal{B}_\tau$ given $U \in I\!R^p\backslash\mathcal{B}_\tau$.

Assume to the contrary and note that this implies that $||U_n||^2 > R_2$ for all $n \geq 0$ and hence that $||\eta||^2 > r$ $\forall n \geq 0$ by (5.5). Then, by (5.7) we have

$$||U_{n+1}||^2 \leq ||U_n||^2 - 2\Delta t_n \varepsilon + 2\Delta t_n \tau ||\eta||, \quad for \quad ||\eta||^2 > r$$

Using (5.8) and noting that $||U_n||, ||U_{n+1}|| \leq ||U||$ by Lemma 5.1, it follows that

$$||U_{n+1}||^2 \leq ||U_n||^2 - 2\Delta t_n \varepsilon + 2\Delta t_n \tau ||U||. \tag{5.12}$$

Since $||U||^2 \leq M$, (5.2) implies that

$$\tau < \frac{2\varepsilon}{1 + ||U||^2} \leq \frac{\varepsilon}{||U||}.$$

Hence $\exists \varepsilon' > 0$ such that

$$\tau ||U|| - \varepsilon \leq -\varepsilon'.$$

Thus (5.12) implies that

$$||U_{n+1}||^2 \leq ||U_n||^2 - 2\Delta t_n \varepsilon'.$$

A contradiction argument as in Theorem DC1 shows that U_n enters $\mathcal{B}_r$ and the result follows. □

6. Gradient Systems. In this section we consider error control schemes for gradient systems satisfying (G1)–(G5). We are unable to prove results for arbitrary algebraically stable embedded pairs, but derive positive results for the order $(p, 1)$ embedded pair (2.24), (2.27) constructed in Example 2.10. We will impose an upper bound Δt_{max} on the time-step. Unlike previous sections, where for dissipative problems Δt_{max} could be taken to be arbirarily large, for gradient systems Δt_{max} will be bounded above in terms of the one-sided Lipschitz constant c appearing in (G3).

Recall that the equation (1.1) has the property that, under (G1-G5) all trajectories approach equilibria as $t \rightarrow \infty$. In subsection 6.1, 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* $\Delta t_n \leq \Delta t_{max} < 1/c$ $\forall n \geq 0$. *Then the embedded pair is 3(G) – globally admissible; furthermore, for any d > 0 there is r*(d) > 0 such that, for any* $\tau \in (0, \tau^*)$ *and any admissible sequence* $\exists N^* = N^*(U, \tau, d) > 0$, *and* $v \in E$: $||U_n - v|| \leq d$ $\forall n \geq N^*$. *If E is bounded then* (2.24)–(2.27) *is 3(G) – globally dissipative.*

This result is analogous to Theorem ODE(iv), noting that d can be made arbitrarily small by choice of τ. It is possible to generalise Theorem G1 to the error per step case as for the dissipative case in section 3 and also to implicit (2.24) but we do not give details here.

Note that (G5) is a natural condition to impose, and holds generically within the set of all gradient systems. For example, if

$$\liminf_{||x|| \rightarrow \infty} ||f(x)|| > 0$$

then, by continuity of f, (G5) is automatically satisfied for any $\delta > 0$.

From a dynamical systems viewpoint (G5) may be considered as a structural stability condition. If (G5) does not hold for any $\delta > 0$ then arbitrarily small perturbations of f can introduce new zeros to E, and hence alter the dynamics of the system. Regarding the numerical solution as a perturbation of the continuous system, we cannot expect to globally reproduce the dynamics of the underlying system if that system is not structurally stable.

Theorem G1 ensures that the solution approaches a small neighbourhood of an equilibrium which scales with the error tolerance. In accordance with the work of Hall [9] and Griffiths [15] **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 solutions is actually driven to equilibrium then we must improve (1.5). We consider a modification of the error control mechanism (1.5) specifically designed for gradient systems. We replace (1.5) by

$$||U_{n+1} - V_{n+1}|| \leq \theta\tau ||U_{n+1} - U_n||. \tag{6.1}$$

This is a form of error per unit step error control relative to an approximation of the time derivative – when the time derivative is small then the time-step is made small also. It is clear that this should drive the solution to equilibrium and we prove this in subsection 6.2. Specifically we shall show that:

THEOREM *G2 Consider* (2.24)- *(2.27) and* (6.1). *Assume that* $\Delta t_n \leq \Delta t_{max} < 1/c$ $\forall n \geq 0$*. Then the embedded pair is 3(G) − globally admissible; furthermore, for any d > 0 there is r*(d) > 0 such that, for any* $\tau \in (0, \tau^*)$ *and any admissible sequence* $\exists v \in E$:

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

If E is bounded, then (2.24)-(2.27) *is 3(G) − globally dissipative.*

Note that, with error control (6.1) 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.

6.1 Gradient Systems; Error Per Unit Step

Under (G4) the equilibria of (1.1) are isolated and hence countable. We label them

$$v_i, \quad i = 1, 2, \dots .$$

The following definitions and lemmas will be needed to prove Theorem G1. Choosing ε and δ given by (G5) we let

$$\Gamma \equiv \Gamma(\varepsilon) = \{x: ||f(x)|| < \varepsilon\}$$

and

$$B_i \equiv B_i(\varepsilon, 6) = \bar{B}(v_i, \delta) \bigcap \Gamma,$$

where $B(v, \delta)$ is defined by (1.8). Notice that $||f(x)|| < \varepsilon$ if $x \in B_i$ for some i, whilst, since ε is defined from δ by (G5), $|| f(x)|| \geq \varepsilon$ if $x \notin B_i$ for any i.

LEMMA 6.1. *Consider* (2.24)-(2.27) *and* (1.5) *under (G1)-(G5), and assume that* $\Delta t_n \leq \Delta t_{max} \leq 1/c$. *Let*

$$C_i \equiv C_i(\varepsilon, \delta, \tau) = \{u \in \mathbb{R}^p : ||u - v_i|| \leq \delta + \frac{1}{c}(\tau + \varepsilon)\}$$

If $U_{n+1} \in B_i$ *then* $U_n \in C_i$. *Furthermore, if* $u \in B_i$ *then*

$$|F(u) - F(v_i)| \leq c\delta^2$$

and if $u \in C_i$ *then*

$$|F(u) - F(v_i)| \leq c(\delta + \frac{1}{c}(\tau + \varepsilon))^2.$$

Proof. By (2.24)

$$U_n = U_{n+1} - \Delta t_n \tilde{f}(U_n; \Delta t_n).$$

Thus, if $U_{n+1} \in B_i$,

$$\begin{aligned} ||U_n - v_i|| &\leq ||U_{n+1} - v_i|| + \Delta t_n ||\tilde{f}(U_n; \Delta t_n)|| \\ &\leq \delta + \frac{1}{c}\Big[||f(U_{n+1}) + \tilde{f}(U_n; \Delta t_n) - f(U_{n+1})||\Big] \\ &\leq \delta + \frac{1}{c}\Big[||f(U_{n+1})|| + ||\tilde{f}(U_n; \Delta t_n) - f(U_{n+1})||\Big] \\ &\leq \delta + \frac{1}{c}[\varepsilon + \tau] \end{aligned}$$

by (2.28), which implies that $U_n \in C_i$ as required.

If $u \in B_i$ then by (G3)

$$\begin{aligned} |F(v_i) - F(u)| &\leq |\langle f(v_i), u - v_i\rangle + c||v_i - u||^2| \\ &\leq c\delta^2 \end{aligned}$$

since $f(v_i) = 0$. A similar proof holds for $u \in C_i$. □

For equation (1.1) under (G1) *F(u(t))* is a decreasing function of t for all $t \geq 0$. The following result shows that $F(U_n)$ is a decreasing function of n outside the union of all B_i.

LEMMA *6.2. Consider* (2.24)- *(2.27) and* (1.5) *under (G1)-(G5). Assume that* $\Delta t_n \leq \Delta t_{max} < 1/c$ *for all n > 0, and that*

$$\tau < \frac{\varepsilon(1 - c\Delta t_{max})}{2 - c\Delta t_{max}}. \tag{6.2}$$

Then

$$||f(U_{n+1})|| > \varepsilon \tag{6.3}$$

implies that

(6.4) $F(U_{n+1}) - F(U_n) \leq -\Delta t_n \Big[(1 - c\Delta t_{max})\varepsilon - (2 - c\Delta t_{max})\tau\Big] ||\tilde{f}(U_n; \Delta t_n)||.$

Furthermore suppose (6.3) holds for $M \leq n \leq N - 1$ *then*

$$(6.5) \qquad F(U_n) \leq F(U_M) - \Big[(1 - c\Delta t_{max})\varepsilon - (2 - c\Delta t_{max})\tau\Big] ||U_n - U_M||$$

for $M + 1 \leq n \leq N.$

Proof. Note that (2.24)- (2.27) imply (2.3). By (G3) and using (2.24), (2.28) we have

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

Now note that

$$\varepsilon < ||f(U_{n+1})|| \leq ||\tilde{f}(U_n; \Delta t_n)|| + ||f(U_{n+1}) - \tilde{f}(U_n; \Delta t_n)|| \leq ||\tilde{f}(U_n; \Delta t_n)|| + \tau,$$

by (1.5). Thus

$$(6.6) \qquad F(U_{n+1}) - F(U_n) \leq -\Delta t_n ||\tilde{f}(U_n; \Delta t_n)|| \Big[(1 - c\Delta t_n)(\varepsilon - \tau) - \tau\Big].$$

By (6.2) $\tau < \varepsilon$ and hence

$$\begin{aligned} (1 - c\Delta t_n)(\varepsilon - \tau) - \tau &\geq (1 - c\Delta t_{max})(\varepsilon - \tau) - \tau \\ &= (1 - c\Delta t_{max})\varepsilon - (2 - c\Delta t_{max})\tau \end{aligned}$$

and (6.4) follows from (6.6). Summing this for $n : M \leq n \leq N - 1$ gives

$$(6.7) F(U_n) \leq F(U_M) - \Big[(1 - c\Delta t_{max})\varepsilon - (2 - c\Delta t_{max})\tau\Big] \sum_{j=M}^{n-1} \Delta t_j ||\tilde{f}(U_j; \Delta t_j)||,$$

for $M + 1 \leq n \leq N$. *Now,* by (2.24),

$$U_n = U_M + \sum_{j=M}^{n-1} \Delta t_j \tilde{f}(U_j; \Delta t_j),$$

so that

$$(6.8) \qquad ||U_n - U_M|| \leq \sum_{j=M}^{n-1} \Delta t_j ||\tilde{f}(U_j; \Delta t_j)||.$$

Combining (6.7) and (6.8) gives (6.5) as required. □

We now use Lemma 6.2 to prove the following two lemmas which are fundamental in the proof of Theorem G1:

Lemma 6.3. *Consider* (2.24)-(2.27) *and* (1.5) *under (G1)-(G5). Assume that* $\Delta t_n \leq \Delta t_{max} < 1/c$ *for all* $n \geq 0$ *and that (6.2) holds. If* $U \notin B_j$ *for any* j *then either:*

(i) $\exists M, i$*:* $U_{M+1} \in B_i$, $U_M \in C_i$ *with* $F(v_i) \leq F(U) + c(\delta + \frac{1}{c}(\tau + \varepsilon))^2$*; or*

(ii) $U_n \notin B_i$*, for any* i*,* n *and then* $\Delta t_n \to 0$ *as* $n \to oo$ *and* $F(U_n) \leq F(U)\ \forall\ n \geq 0$*.*

Proof. Assume that, for $0 \leq n \leq M$ $U_n \notin B_j$ for any j and $\exists i : U_{M+1} \in B_i$. Then, by Lemma 6.1, $U_M \in C_i$. By Lemma 6.2 with $M = 0$ and $N \to M$ we have

$$F(U_M) \leq F(U) - \Big[(1 - c\Delta t_{max})\varepsilon - (2 - c\Delta t_{max})\tau\Big] \|U_M - U\| \leq F(U).$$

But, since $U_M \in C_i$, by Lemma 5.1

$$F(U_M) \geq F(v_i) - c(\delta + \frac{1}{C}(\tau + \varepsilon))^2$$

so that

$$F(v_i) \leq F(U) + c(\delta + \frac{1}{c}(\tau + \varepsilon))^2.$$

Finally, if $U_n \notin B_i$ for any i, n then $\|f(U_{n+1})\| > \varepsilon, \forall\ n \geq 0$. Hence, by (2.28) and (6.2),

$$\begin{aligned}\|\tilde{f}(U_n; \Delta t_n)\| &> \varepsilon - \tau \\ &> \frac{\varepsilon}{2 - c\Delta t_{max}}\end{aligned}$$

for all $n \geq 0$. Thus Lemma 6.2 gives

$$F(U_{n+1}) \leq F(U_n) - \Delta t_n \varepsilon \left[\frac{(1 - c\Delta t_{max})\varepsilon}{2 - c\Delta t_{max}} - \tau\right], \quad \forall n \geq 0$$

$$\Rightarrow F(U_n) \leq F(U) - \varepsilon \left[\frac{(1 - c\Delta t_{max})\varepsilon}{2 - c\Delta t_{mat}} - \tau\right] \sum_{j=0}^{n-1} \Delta t_n, \quad \forall n \geq 0.$$

Since F satisfies (G2) and (6.2) holds, we deduce that $\Delta t_n \to 0$ as $n \to \infty$ and that $F(U_n) \leq F(U)$, $\forall\ n \geq 0$. □

Given any $d > 0$, (G5) implies that there exists $\varepsilon^* > 0$ such that

$$\inf_{x \notin B(d/2)} \|f(x)\| \geq \varepsilon^*.$$

Now notice that if $\Delta t_{max} < 1/c$, it is possible (since inequalities (6.9) and (6.10) are strict if $\varepsilon = \delta = \tau^* = 0$) to satisfy

$$\delta + \frac{1}{c}(\tau^* + \varepsilon) \leq \frac{d}{2}, \tag{6.9}$$

and

$$(6.10) \qquad \delta + \frac{1}{c}(\tau^* + \varepsilon) \leq \sqrt{\frac{d}{4c}}\left[(1 - c\Delta t_{max})\varepsilon^* - (2 - c\Delta t_{max})\tau^*\right]$$

with $\min(\varepsilon, \delta, \tau^*) > 0$, and such that

$$(6.11) \qquad \tau^* < \frac{\varepsilon(1 - c\Delta t_{max})}{2 - c\Delta t_{max}}.$$

and

$$(6.12) \qquad \inf_{x \notin B(\delta)} \|f(x)\| \geq \varepsilon.$$

Moreover, note that given such ε, δ and τ^*, (6.9)–(6.12) also hold if τ^* is replaced by τ where $\tau \in (0, \tau^*)$. We now prove

LEMMA 6.4. *Consider* (2.24)– *(2.27) and* (1.5) *under (Gl)-(G5). For any $d > 0$ let $D_i = B(v_i, d)$ and assume that $\Delta t_n \leq \Delta t_{max} < 1/c$ for all $n \geq 0$. Then there exists $\tau^* > 0$ and $\delta > 0$ such that if $\tau < \tau^*$ and $U_M \in B_i$ then either*

(i) $U_n \in D_i$ for all $n \geq M$; or

(ii) $\exists N, j : U_{N+1} \in B_j$ with $F(v_j) < F(v_i)$; or

(iii) $\exists N : \forall n \geq N$, $U_N \notin B_j$, for any j and then $\Delta t_n \to 0$ as $n \to oo$ and

$$F(U_n) \leq F(v_i) + c\delta^2 \quad \forall\, n \geq N.$$

Proof. Without loss of generality we supppose $0 < d < D/2$. Choose ε, δ, $\tau^* > 0$ such that (6.9)–(6.12) hold. If $U_M \in B_i$ iterate until $U_n \notin B_i$. (If such an n does not exist then (i) holds since, by Lemma 6.1 and (6.9), $B_i \subset D_i$.) Re-label $(n - 1) \to M$; hence $U_M \in B_i$ and $U_{M+1} \notin B_i$.

If $U_n \notin B_j$, for any j, $\forall\, n \geq M + 1$ then, since (6.11) implies (6.2), Lemma 6.3(ii) implies $\Delta t_n \to 0$ as $n \to \infty$ and that

$$F(U_n) \leq F(U_M), \forall\, n \geq M.$$

Furthermore, by Lemma 6.1,

$$F(U_n) \leq F(U_M) \leq F(v_i) + c\delta^2$$

and thus (iii) holds.

Now suppose that $\exists N > M$ such that for $M < n < N$, $U_n \notin B_j$ for any j, and $U_N \in B_j$ for some j. Either $j = i$ or $j \neq i$.

Consider $j = i$ first; we show that in this case $U_n \in D_i$ for n such that $M \leq n \leq N$. For contradiction suppose that there exists n_2: $M < n_2 < N$ such that $U_{n_2} \notin D_i$. Let n_1 be the largest integer such that $M \leq n_1 < n_2$ and $U_{n_1} \in B(v_i, d/2)$. By Lemmas 6.1 and 6.2

$$(6.13) \qquad F(U_{n_1}) \leq F(U_M) \leq F(v_i) + c\delta^2.$$

By Lemma 6.1 $U_{N-1} \in C_i$ and hence

$$(6.14) \qquad F(v_i) - c(\delta + \frac{1}{c}(\tau + \varepsilon))^2 \leq F(U_{N-1}) < F(U_{n_2}).$$

Now note by construction of n_1 and ε^* that $\| f(U_n)\| \geq \varepsilon^*$ for $n_1 + 1 \leq n \leq n_2$ *and* hence by Lemma 6.2

$$(6.15) \quad F(U_{n_2}) \leq F(U_{n_1}) - \Big[(1 - c\Delta t_{max})\varepsilon^* - (2 - c\Delta t_{max})\tau\Big] \|U_{n_2} - U_{n_1}\|.$$

Combining (6.13)–(6.15) and noting that $\|U_{n_2} - U_{n_1}\| \geq d/2$ implies that

$$\begin{aligned} c\delta^2 + c(\delta + \frac{1}{c}(\tau + \varepsilon))^2 &> \frac{d}{2}\Big[(1 - c\Delta t_{max})\varepsilon^* - (2 - c\Delta t_{max})\tau\Big] \\ (\delta + \frac{1}{C}(\tau + \varepsilon))^2 &> \frac{d}{4c}\Big[(1 - c\Delta t_{max})\varepsilon^* - (2 - c\Delta t_{max})\tau\Big] \end{aligned}$$

which contradicts (6.10). Thus $U_n \in D_i$ for $M \leq n \leq N$. We can re-label $N \longrightarrow M$ and repeat the argument from the beginning of the proof. Either (i) or (iii) holds or we must consider $j \neq i$.

If $j \neq i$ then, by Lemma 6.1, $U_{N-1} \in C_j$. Let n_1 be the largest integer such that $M \leq n_1 < N$ and $U_{n_1} \in B(v_i, d/2)$. Letn_2 be the smallest integer such that $n_1 < n_2 \leq N - 1$ and $U_{n_2} \notin B(v_i, d)$. Note that by (6.9) and since $d < D/2$ that such n_1 and n_2 exist. Now note that by construction of n_1 and n_2 we have that (6.13) and (6.15) also hold in the case $i \neq j$. Also, since $U_{N-1} \in C_j$, by Lemma 6.1 we have

$$F(v_j) - c(\delta + \frac{1}{c}(\tau + \varepsilon))^2 \leq F(U_{N-1}) < F(U_{n_2}).$$

Combining (6.13)–(6.15) and noting that $\|U_{n_2} - U_{n_1}\| \geq d/2$ we obtain $F(v_j) < F(v_i)$ as required. □

We are now in a position to establish $\mathcal{F}(G)$ global admissibility. Let

$$(6.16) \qquad I^*(a) = \{u \in I\!R^p : F(u) \leq a\}.$$

By (G2) $I(\bullet)$ is bounded and it is clearly closed since F is continuous, and hence I is compact. Hence since the equilibria of f are isolated, by (G4), $I(a)$ contains only finitely many equilibria for any $a \in I\!R$. To show that all trajectories remain bounded define

$$(6.17) \qquad I(U) = I(F(U) + 2c(\delta + \frac{1}{c}(\tau^* + \varepsilon))^2)$$

where δ, ε, τ^* satisfy (6.9)–(6.12). With this definition we may now show:

Lemma 6.5. *Consider* (2.24)–(2.27) *and* (1.5) *under (Gl)-(G5) and assume that* $\Delta t_n \leq \Delta t_{max} < 1/c \ \forall \ n \geq 0$. *Then* $\exists \tau^*$, $\delta > 0$, *both independent of* U *such that if* $\tau < \tau^*$ *then* $U_n \in I(U) \ \forall \ n \geq 0$. *Hence the embedded pair* (2.24)– *(2.27) is 3(G) – globally admissible.*

Proof Let $M \geq 0$ be the least integer such that $U_M \in B_i$ for some i. (If there is no such integer then $F(U_n) \leq F(U) \ \forall n \geq 0$ and the result is trivial.) If $M > 0$ then by Lemma 6.1 $U_{M-1} \in C_i$ and $F(U_{M-1}) \leq F(U)$. Hence

$$F(v_i) \leq F(U) + c(\delta + \frac{1}{c}(\tau + \varepsilon))^2$$

(for $M \geq 0$), and if $U_n \in B_i$ then

$$(6.18) \qquad F(U_n) \leq F(U) + c\delta^2 \leq F(U) + 2c(\delta + \frac{1}{c}(\tau + \varepsilon))^2.$$

If $U_n \in B_i$ for all $n \geq M$ then the result follows. Otherwise, let N be the least integer such that $U_N \notin B_i$ and relabel $N \rightarrow M + 1$; hence $U_M \in B_i$ and $U_{M+1} \notin B_i$.

Now if $U_n \notin B_j$ for any j and for all $n \geq M$ then we are in case (iii) of Lemma 5.4 and $F(U_n) \leq F(U) + 2c(\delta + \frac{1}{c}(\tau + \varepsilon))^2$ for all $n \geq M$. Otherwise let N be the least integer such that $N > M$ and $U_N \in B_j$ for some j. Either $j = i$ or $j \neq i$.

Consider $j = i$ first. Then, by Lemma 6.1, $U_{N-1} \in C_i$ and by Lemma 6.2 and (6.18), since $U_M \in B_i$,

$$F(U_n) \leq F(U_M) \leq F(U) + 2c(\delta + \frac{1}{c}(\tau + \varepsilon))^2, \; n : M \leq n \leq N-1$$

whilst, by (6.18),

$$F(U_N) \leq F(U) + 2c(\delta + \frac{1}{c}(\tau + \varepsilon))^2$$

since $U_N \in B_i$.

Now consider $i \neq j$. Then, by Lemma 6.1, $U_{N-1} \in C_j$ and by Lemma 6.4 $F(v_j) < F(v_i)$. By Lemma 6.2 and (6.18), since $U_M \in B_i$

$$F(U_n) \leq F(U_M) \leq F(U) + 2c(\delta + \frac{1}{c}(\tau + \varepsilon))^2, M \leq n \leq N-1,$$

whilst, since $U_N \in B_j \subset C_j$ it follows that

$$\begin{aligned} F(U_n) &\leq F(v_j) + c(\delta + \frac{1}{c}(\tau + \varepsilon))^2 \\ &< F(v_i) + c(\delta + \frac{1}{c}(\tau + \varepsilon))^2 \\ &< F(U) + c\delta^2 + c(\delta + \frac{1}{c}(\tau + \varepsilon))^2 \\ &\leq F(U) + 2c(\delta + \frac{1}{c}(\tau + \varepsilon))^2. \end{aligned}$$

Repeating the above arguments inductively implies that

$$F(U_n) \leq F(U) + 2c(\delta + \frac{1}{c}(\tau + \varepsilon))^2$$

for all $n \geq 0$ as required. Since τ^* given by (6.11) is independent of U, $\mathcal{F}(G)$ global admissibility follows from Corollary 3.6. □

Proof of Theorem Gl The $\mathcal{F}(G)$ global admissibility follows from Lemma 6.5. Thus we examine the asymptotic behaviour of admissible sequences satisfying $\Delta t_n \geq \bar{\Delta t_n} > 0 \; \forall n \geq 0$. Since $\Delta t_n \geq \bar{\Delta t} > 0$, $\forall \; n \geq 0$, Lemma 6.3 shows that, without loss of generality, we may assume that $\exists i : U_0 \in B_i$ by re-labelling. By Lemma 6.4 we deduce that $\exists M, j : U_n \in D_j \; \forall \; n \geq M$ since, if not, by (ii) $\exists$ subsequences N_k, $j_k \rightarrow \infty$ with $U_{N_k} \in B_{j_k}$ and

$$F(v_{N_{k+1}}) < F(v_{N_k}).$$

But $U_n \in \bar{I}(U)$ for all n, and $\bar{I}(U)$ only contains finitely many equilibria so such sequences cannot exist.

Now, to establish dissipativity, let

$$\mathcal{B}_\tau = \{u \in I\!R^p : F(u) \leq \max_{v \in E} F(v) + 3c(\delta + \frac{1}{c}(\tau + \varepsilon))^2\}.$$

If

$$F(U) \leq \max_{v \in E} F(v) + c(\delta + \frac{1}{c}(\tau + \varepsilon))^2$$

then

$$U_n \in \mathcal{B}_\tau \ \forall n \geq 0$$

by Lemma 6.5. Alternatively, if

$$F(U) > \max_{v \in E} F(v) + c(\delta + \frac{1}{c}(\tau + \varepsilon))^2$$

then $U \notin C_i$, any i, by Lemma 6.1. Hence $U \notin B_i$, any i since $B_i \subset C_i$. By Lemma 6.3 $\exists M, i : U_M \in C_i$. Hence

$$F(U_M) \leq \max_{v \in E} F(v) + c(\delta + \frac{1}{c}(\tau + \varepsilon))^2$$

by Lemma 6.1. Letting $U_M \to U$ and applying Lemma 6.5 we have

$$U_n \in \mathcal{B}_\tau \ Vn \geq M.$$

The result follows. □

6.2 Gradient Systems; Relative Error Per Unit Step

We now consider the relative error strategy (2.24)–(2.27) and (6.1). Notice that (6.1) can be re-written as

$$(6.19) \qquad ||\tilde{f}(U_n; \Delta t_n) - f(U_{n+1})|| \leq \tau ||\tilde{f}(U_n; \Delta t_n)||.$$

The proof is similar to that for Theorem G1 but simplified because the function $F(U_n)$ will be shown to be non-increasing for all $n \geq 0$. We prove this result, together with boundedness of the solution sequence in the following Lemma; recall $I^*(\bullet)$ defined by (6.16).

Lemma 6.6. *Consider* (2.24)- *(2.27) under* (6.1). *Assume that (G1)-(G5) hold. If $\Delta t_n \leq \Delta t_{max} < 1/c \ \forall n \geq 0$ and $\tau < \tau^* \leq 1 - c\Delta t_{max}$ then there exists $\varepsilon^* > 0$ such that*

$$F(U_{n+1}) - F(U_n) \leq -\varepsilon^* \Delta t_n ||\tilde{f}(U_n; \Delta t_n)||^2, \ \forall n \geq 0.$$

Thus

$$U_n \in I^*(F(U)) \vee n \geq 0.$$

Proof As in the proof of Lemma 5.2 we have that

$$F(U_{n+1}) - F(U_n) \leq \langle \tilde{f}(U_n; \Delta t_n), U_n - U_{n+1}\rangle + \langle f\,(U_{n+1}) - \tilde{f}(U_n; At,), U_n - U_{n+1}\rangle$$
$$+ c||U_{n+1} - U_n||^2.$$

Applying (6.19) we obtain

$$F(U_{n+1}) - F(U_n) \leq -\frac{||U_{n+1} - U_n||^2}{\Delta t_n} + \frac{\tau}{\Delta t_n}||U_{n+1} - U_n||^2 + c||U_{n+1} - U_n||^2.$$

Thus

$$F(U_{n+1}) - F(U_n) \leq -\frac{(1 - \tau - c\Delta t_n)}{\Delta t_n}||U_{n+1} - U_n||^2.$$

If $\tau < \tau^*$ and $\Delta t_n < \Delta t_{max}$ then there exists $\varepsilon^* > 0$ such that $1 - \tau - c\Delta t_{max} \geq \varepsilon^*$ and hence it follows that

$$F(U_{n+1}) - F(U_n) \leq -\frac{\varepsilon^*}{\Delta t_n}||U_{n+1} - U_n||^2.$$

By (2.24) the first part of the result follows. Clearly

$$F(U_n) \leq \mathrm{F(U)},\ \forall n \geq 0$$

and hence the second part of the result follows automatically. □

LEMMA 6.7. *Consider* (2.24)- *(2.27) under* (6.1). *Assume that (G1)-(G5) hold. If $\Delta t \leq \Delta t_{max} < 1/c$ then the embedded pair is 3(G) – globally admissible.*

Proof Let $\tau^* = 1 - c\Delta t_{max}$ where τ^* is given in Lemma 6.6. Thus $U_n \in I^*(F(U))$ V $n \geq 0$. The error control (6.1) is to find $\Delta t = \Delta t_n$ such that, given $X = U_n$ *we* have

$$||W - V|| \leq \theta\tau||W - X||$$

where W and V are defined in (3.5), (3.6) and (3.7). Thus we define the function

$$R(X, \Delta t) := \frac{||W - V||}{||X - W||},$$

and, letting $d_i = b_i - \bar{b}_i$ we may write

$$R(X, \Delta t) = \frac{||\sum_{j=1}^{k} d_j f(\xi_j)||}{||\sum_{j=1}^{k} b_j f(\xi_j)||}.$$

Let

$$\tilde{d} = \max_{1 \leq i \leq k} |d_i|.$$

In order to show that the error control may be satisfied we show that $\exists C = C(X)$ independent of Δt such that $R(\Delta t; X) \leq C(X)\Delta t$ for Δt sufficiently small.

By Lemma 3.3,

$$||\xi_i - X|| \leq \Delta t \tilde{a} k \gamma ||f(X)||$$

for

$$\Delta t \leq (1-\gamma^{-1})/[\tilde{a} k L_I],$$

where

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

Noting that $\sum_{j=1}^k d_j = 0$, $\sum_{j=1}^k b_j = 1$ by consistency we deduce that

$$R(X, \Delta t) = \frac{||\sum_{j=1}^k d_j [f(\xi_j) - f(X)]||}{||f(X) + \sum_{j=1}^k b_j [f(\xi_j) - f(X)]||}$$

$$\leq \frac{\Delta t \tilde{a} \tilde{d} k^2 \gamma L_I ||f(X)||}{||f(X)|| - \Delta t \tilde{a} \tilde{b} k^2 \gamma L_I ||f(X)||}$$

$$= \frac{\Delta t \tilde{a} \tilde{d} k^2 \gamma L_I}{1 - \Delta t \tilde{a} \tilde{b} k^2 \gamma L_I} \leq C(X) \Delta t$$

for Δt sufficiently small. Hence, for any $\tau \in (0, \tau^*)$ and any $X \in I^*(F(U))$ $\exists \Delta t_c = \theta \tau / c$ such that the error control criteria is satisfied. Since the scheme is explicit, the existence of a solution sequence is guaranteed. Thus $\mathcal{F}(G)$ global admissibility follows. □

Lemma 6.8. *Consider* (2.24)–(2.27) *under* (6.1). *Assume that (Gl)-(G5) hold. If* $\Delta t_n \leq \Delta t_{max} < 1/c$, *and* $\tau < \tau^* \leq 1 - c\Delta t_{max}$ *then any admissible sequence satisfies*

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

and

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

Proof. The first result follows directly from Lemma 6.6 since Δt_n is bounded below by $\Delta \bar{t}_n > 0$ and (G2) holds. Now, from Lemma 6.6 and (6.19) if follows for some $\varepsilon^* > 0$ that

$$\begin{aligned} F(U_{n+1}) - F(U_n) &\leq -\varepsilon^* \Delta t_n ||f(U_{n+1}) + \tilde{f}(U_n; \Delta t_n) - f(U_{n+1})||^2 \\ &\leq -\varepsilon^* \Delta t_n \Big[||f(U_{n+1})|| - ||\tilde{f}(U_n; \Delta t_n) - f(U_{n+1})|| \Big]^2 \\ &\leq -\varepsilon^* \Delta t_n \Big[||f(U_{n+1})|| - \tau ||\tilde{f}(U_n; \Delta t_n)|| \Big]^2. \end{aligned}$$

To obtain a contradiction suppose that

$$\limsup_{n\to\infty} \|f(U_n)\| = l > 0,$$

which implies that $\|f(U_n)\| \geq \$1$ for infinitely many n. By the first part of the lemma we know that, for any $\varepsilon > 0 \; \exists \; N = N(\varepsilon) > 0$ such that

$$\|\tilde{f}(U_n; \Delta t_n)\| \leq \varepsilon, \; \forall n \geq N.$$

Taking $\varepsilon = \frac{l}{3\tau}$ we see that

$$F(U_{n+1}) - F(U_n) \leq -\varepsilon^* \, \Delta t_n \left[\frac{l}{3}\right]^2$$

for infinitely many $n \geq N$. But since $\Delta t \geq \bar{\Delta t} > 0$ and $F(U_{n+1}) \leq F(U_n) \; \forall n$, this implies that $F(U_n) \to -\infty$ as $n \to \infty$ which contradicts (G2). Thus

$$\limsup_{n\to\infty} \|f\,(U_n)\| = 0$$

as required. □

Proof of Theorem G2 From Lemma 6.8 and (2.24) we know that, for any admissible sequence,

$$\|f(U_{n+1})\|, \; \|U_{n+1} - U_n\| \to 0 \tag{6.20}$$

as $n \to \infty$, since $\Delta t_n \geq \bar{\Delta t} > 0$, $\forall n \geq 0$. Since the sequence $\{U_n\}_{n=0}^{\infty}$ is bounded, by Lemma 6.6 and (G2), we deduce that it has at least one convergent subsequence and, by (6.20), that every limit point is contained in E, defined in (G4). Let $v \in E$ be such a limit point. Note further that $v \in I(F(U))$ by Lemma 6.6. Thus all possible limit points lie in the intersection of E with the compact set $I(F(U))$. This implies that there are a finite number of possible limit points, say v_j, $j = 1, \ldots, J$. Let

$$\delta = \frac{1}{4} \min_{i \neq j} \|v_i - v_j\|,$$

$B_j = B(v_j\,, \delta)$, $B^+ = \bigcup_{j=1,\ldots,J} B_j$ and

$$B^- = Cl\{I(F(U)) \backslash B^+\}.$$

Notice that B^- is closed and bounded and hence compact. Now consider the given limit point v_i. Either the whole sequence converges to v_i and the proof is complete or $\exists \{m_j\} \to \infty$ such that

$$U_{m_j} \in B_i, \; U_{m_j+1} \notin B_i.$$

Since $\|U_{n+1} - U_n\| \to 0$ as $n \to \infty$ we can assume, without loss of generality, that $\|U_{n+1} - U_n\| \leq \delta$, $\forall n \geq m_0$. Hence $U_{m_j+1} \in B_i(v_i, 2\delta)$ and hence $U_{m_j+1} \in B^-$. But B^- is compact and hence the sequence U_{m_j+1} must contain a convergent subsequence with limit in B^-. This is a contradiction since B^- contains no points in E. The result follows.

To establish dissipativity, let

$$\mathcal{B}_\tau = \{u \in \mathbb{R}^p : F(u) \leq \max_{v \in E} F(v)\} \bigcup B(6)$$

where $B(S)$ is defined in (G5). Since $\mathcal{B}_r$ is formed as the union of two bounded sets it too is bounded. By (G5) it follows that

$$\inf_{x \notin \mathcal{B}_r} \|f(x)\| \geq \varepsilon.$$

Note that, for any admissible sequence $\exists \bar{\Delta t}_n > \mathbf{0} : \Delta t_n \geq \bar{\Delta t}_n \ Vn \geq 0$. Assume for the purposes of contradiction that $\mathcal{B}_r$ is not absorbing so that there is a sequence of integers $n_i \to \infty$ with $U_{n_i} \notin B$,. Then, by Lemma 6.6,

$$F(U_{n_i+1}) \leq F(U_{n_i}) \leq -\bar{\Delta t}_n \varepsilon^* \varepsilon^2.$$

Since (G2) holds this is a contradiction and dissipativity follows. **0**

REMARK Some of the arguments contained in this proof are similar to those used by Elliott [6] in the study of discretisations of the Cahn-Hilliard equation.

7. Numerical Results. In this section we describe numerical results which support our theoretical results. Throughout "energy" refers to the quantity $\|u\|$. In the Figures E_n denotes $\|U_n\|$.

EXAMPLE 7.1.

- Illustrating Theorem DCl We consider the Lorenz equations

$$\left.\begin{array}{rcl} x_t & = & 10(y - x), \\ y_t & = & 28x - \ y - xz, \\ z_t & = & -\frac{8}{3}z + xy, \end{array}\right\} \tag{7.1}$$

with initial data

$$x(0) = y(0) = z(0) = 100. \tag{7.2}$$

These equations satisfy (D) after translation of the origin; see [20]. For this example, $u = (x, y, z)^T$.

Figure 7.1 shows the result when applying the explicit Euler scheme to (7.1), (7.2) with fixed time-step $\Delta t = 1/200$. The energy is plotted and clearly starts to grow very rapidly; eventually the scheme breaks down after several more steps as the solution becomes unbounded for computational purposes. This contrasts with the behaviour of the underlying equation (1.1) for which, under (D), the solution should eventually lie in a bounded set.

We now apply the variable time-step scheme (2.22), (2.23) and (1.5) with $e_0 = 2$. The step selection mechanism is to choose

$$\Delta t_n = (0.8)^k \Delta t_{max} \tag{7.3}$$

where $\Delta t_{max} = 1/200$ and k is the minimum integer for which the error control criteria (1.5) is satisfied. The tolerance is set at $\tau = 0.002$. Figure 7.2 shows the behaviour of the energy and clearly the solution is forced into some bounded set with energy less than approximately 45, illustrating Theorem DCl. Figure 7.3 shows the time-steps selected and Figure 7.4 a plot of y against z from which the bounded set property of Theorem DCl is clear.

Example 7.2.

– Illustrating Theorem E We start with an example which shows that the dependence of τ on initial data is required in Theorem E and is not an artifact of the analysis. Consider equation (1.1) in dimension $p = 1$ where

$$f(u) = -\frac{1}{u}, |u| > 1$$

and with f constructed on $\{u : \|u\| \leq 1\}$ so that the function is Lipschitz on $\mathbb{R}$. We apply the scheme from Example 2.8 with error control (1.5) and $e_0 = \theta^{-1}$. Now let

$$U_n = (-1)^n q, \, \Delta t_n = 2q^2$$

where

$$q \geq 2/\tau. \tag{7.4}$$

With these choices, the explicit Euler scheme and the error control criteria are both satisfied. Clearly the solution does not enter a set bounded independently of initial data since it just oscillates with the amplitude of the initial data; hence no discrete analogue of the behaviour of (1.1) is possible unless τ depends on U. Notice that, to apply Theorem E with $U = q$ would require by (5.2) that $\tau < \frac{2}{1+q^2}$ contradicting (7.4) and thereby ruling out the undesirable periodic solution. □

Now consider the equations

$$\left.\begin{array}{rcl} x_t & = & y - \frac{x}{x^2+y^2}, \\ y_t & = & -x - \frac{y}{x^2+y^2}. \end{array}\right\} \tag{7.5}$$

The initial data is taken to be

$$x(0) = y(0) = 45. \tag{7.6}$$

These equations are not defined by a Lipschitz vector-field, but a simple modification close to the origin can be made to this end; we consider solutions bounded away from the origin so that explicit description of the modification need not be made. Notice that

$$x(y - \frac{x}{x^2+y^2}) + y(-x - \frac{y}{x^2+y^2}) = -1$$

so that $\langle f(u), u\rangle \leq -1$ outside a bounded set and (E) is satisfied – the precise value for r is determined by the modification to make the vector field Lipschitz.

Again (2.22), (2.23) and (1.5) are used to advance the solution. The time-step is chosen according to (7.3) with $\Delta t_{max} = 0.001$. Figure 7.5 shows the behaviour of the energy for $\tau = 0.1$ Notice that it is linearly *increasing* and this trend is continued as time evolves. The energy for the differential equation should be linearly *decreasing.* Figure 7.6 shows the behaviour of the energy in the case $\tau = 0.0004$ which is beneath the critical value for τ given by Theorem E which, for this initial data, is 0.00049. As predicted by the theorem, the energy is decreasing. Computational experiments reveal that the bound (5.2) for the critical value of τ is overly pessimistic and that the true critical value is actually close to 0.04. Nonetheless, the discussion at the start of this example shows that this value must be initial data dependent and our numerical experiments bear this out.

EXAMPLE 7.3.
– Illustrating Theorems G1 and G2 First consider the scalar equation

$$u_t = -u, \quad u(0) = U$$

This is a trivial example of a gradient system. If we apply the numerical scheme (2.22), (2.23) and (1.5) of Example 2.8 and take the maximum possble time-step (i.e equality in (1.5)) then

$$\mathrm{At}, = \frac{\tau}{|U_n|}, \quad U_{n+1} = (1 - \frac{\tau}{|U_n|})U_n.$$

Straightforward analysis shows that

$$|U_n| > \frac{\tau}{2} \Rightarrow |U_{n+1}| < |U_n|$$

whilst

$$|U_n| \leq \frac{\tau}{2} \Rightarrow |U_{n+1}| \leq \tau.$$

Thus it may be shown that an absorbing set for this problem is the interval $[-\tau, \tau]$. Figure 7.7 illustrates the behaviour of a solution sequence for initial data $U = 1$ and tolerance $\tau = 0.1$. This behaviour is essentially what is predicted by the analysis of Hall [9] and Griffiths [15] – notice that the solutions oscillate in a small neighbourhood of the (unique) equilibrium of the system. This agrees with Theorem G1.

If instead we apply the error control scheme (2.22), (2.23) and (6.1) with $\theta = \frac{1}{2}$, again with the maximum possible time-step (i.e. equality in (6.1)) then we obtain

$$\mathrm{At}_n = \tau, \quad U_n = (1 - \tau)^n U_0$$

and the solution converges to the origin as $n \longrightarrow \infty$ as predicted by Theorem G2.

Acknowledgements AMS is grateful to Kevin Burrage and John Butcher for discussions about order barriers and to David Griffiths and Arieh Iserles for discussions about the dynamics of error control. ARH is grateful to the SERC, UK and to Stanford University for financial support.

REFERENCES

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

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

[3] J. Butcher, *The numerical analysis of ordinary differential equations: Runge-Kuttu and general linear methods.* Wiley, Chichester, 1987.

[4] G. Dahlquist and R. Jeltsch, *Generalised disks of contructivity for explicit and implicit Runge-Kutta methods.* TRITA-NA Report 7906.

[5] K. Dekker and J.G Verwer, *Stability of Runge-Kuttu methods for stiff nonlinear equations.* North-Holland, Amsterdam, 1984.

[6] C.M. Elliott, *The Cahn-Hilliard model for the kinetics of phase separation.* Appears in "Mathematical models for phase change problems", ed J.F. Rodrigues, Birkhauser, 1989.

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

[8] J.K. Hale, *Asymptotic behuviour of dissipative systems.* AMS Mathematical Surveys and Monographs 25, Rhode Island, 1988.

[9] G. Hall, *Equilibrium states of Runge-Kutta schemes.* ACM Trans. on Math. Software **11**, 289–301.

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

[11] D.J. Higham, *The tolerance proportionality of adaptive ODE solvers.* University of Dundee Numerical Analysis Report NA/133, 1991.

[12] A.R. Humphries and A.M. Stuart, *Runge-Kutta methods for dissipative and gradient dynamical systems, 1992.* Submitted to SIAM J. Num. Anal.

[13] A.R. Humphries, *The dynamics of numerical methods for dynamical systems.* University of Bath, Phd thesis. In preparation.

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

[15] D.F. Griffiths, *The dynamics of some linear multistep methods with step-size control.* Appears in Numerical Analysis, eds: Griffiths, D.F. and Watson, G.A., Longman (1988).

[16] J.M. Sanz-Serna, *Numerical ordinary differential equations versus dynamical systems.* Appears in "The dynamics of numerics and the numerics of dynamics", Eds. D.S. Broomhead and A. Iserles, Clarendon Press, Oxford, 1992.

[17] L. Shampine, *Tolerance proportionality in ODE codes.* Appears in Numerical methods for ordinary differential equations, proceedings. Eds. A. Bellen, C. Gear and E. Russo. Springer-Verlag Lecture Notes 1386 (1987), 118–136.

[18] H.J. Stetter, *Tolerance proportionality in ODE-codes.* Appears in Proc. Second Conf. on Numerical Treatment of Ordinary Differential Equations, 109–123, Ed. R. Marz, Seminarberichte 32, Humboldt University, Berlin.

[19] A.M. Stuart and A.R. Humphries, *Model problems in numerical stability theory for initial value problems,* 1992. Submitted to SIAM Review.

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

[21] J. Wilkinson, *The algebraic eigenvalue problem. Clarendon Press, Oxford, 1965.*

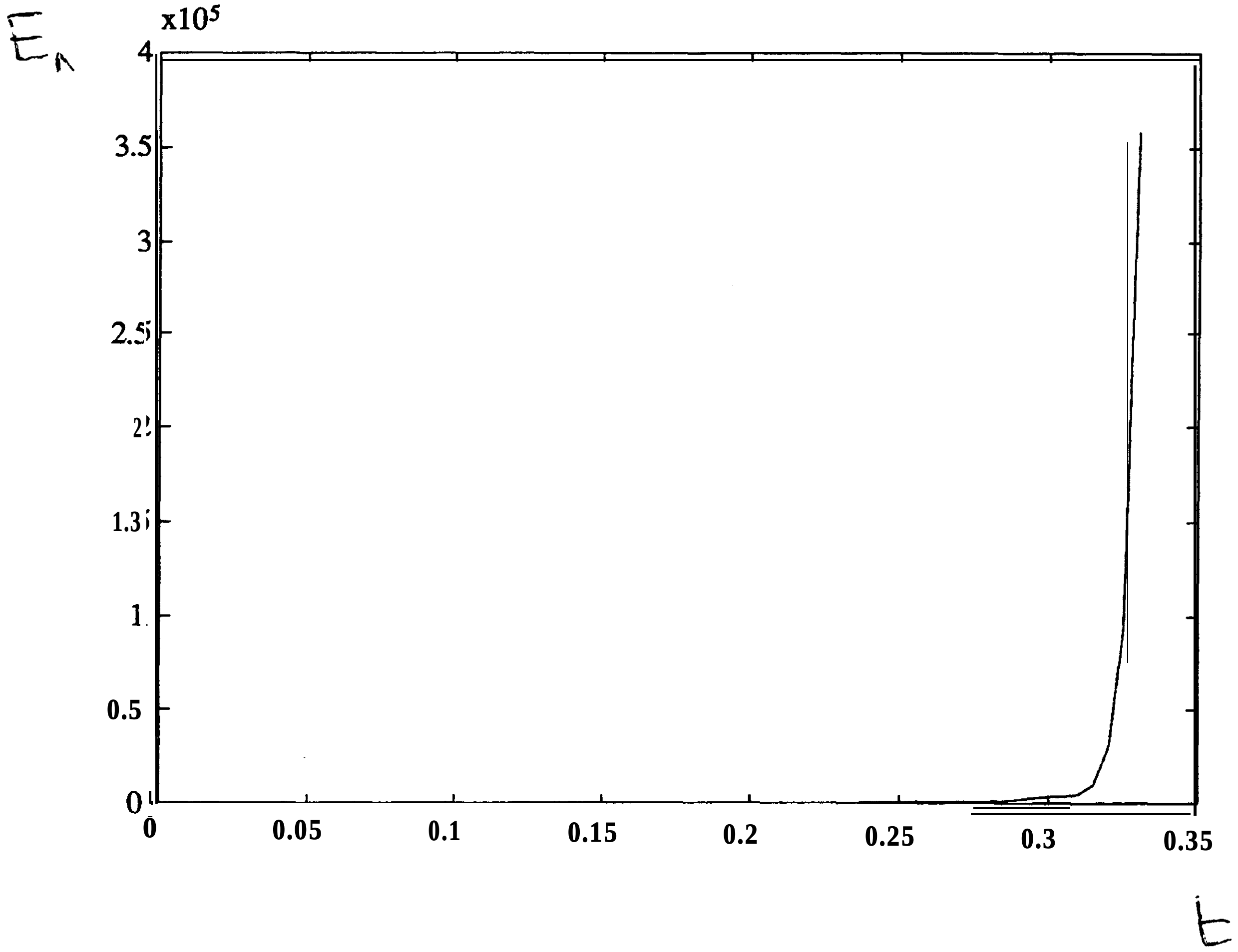

Figure 7-1

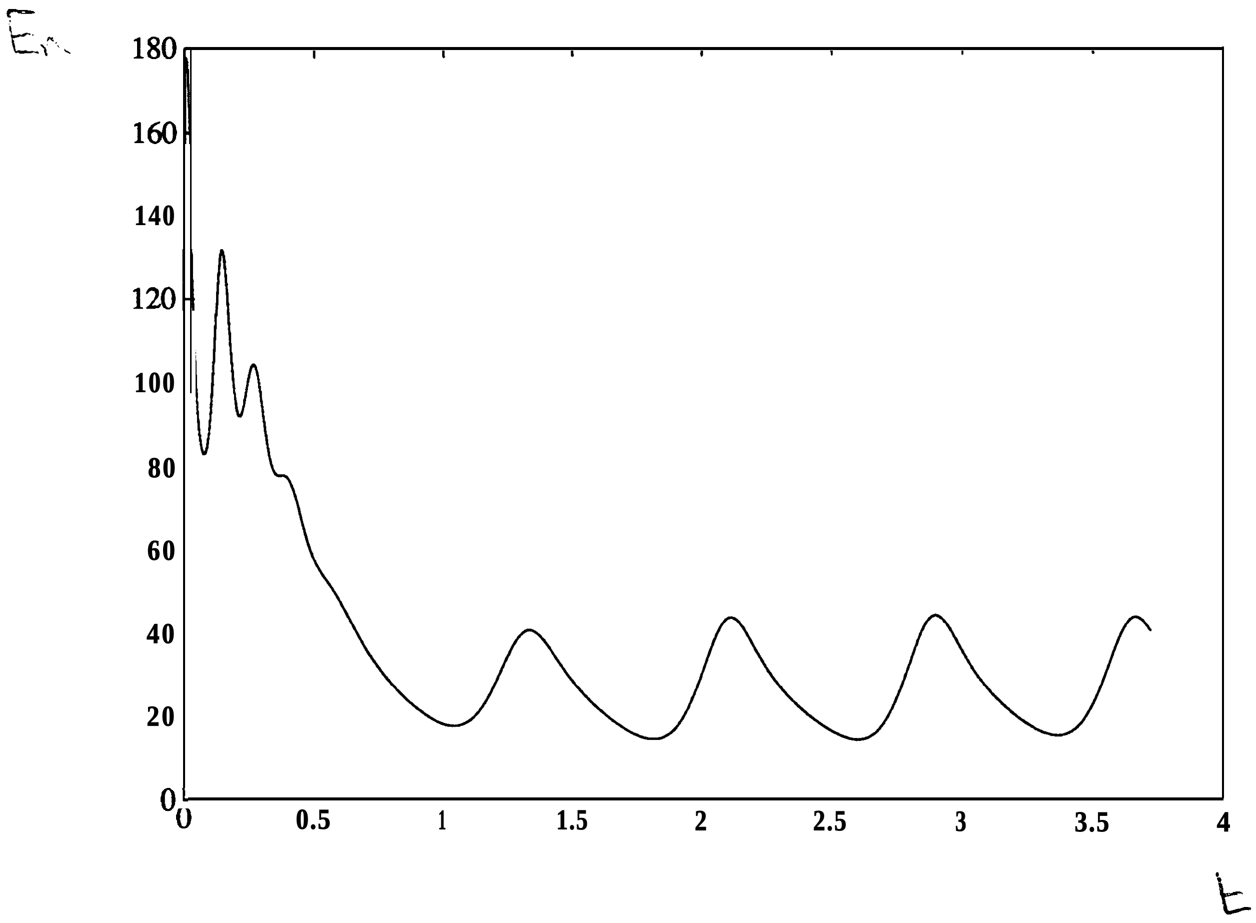

Figure 7.2

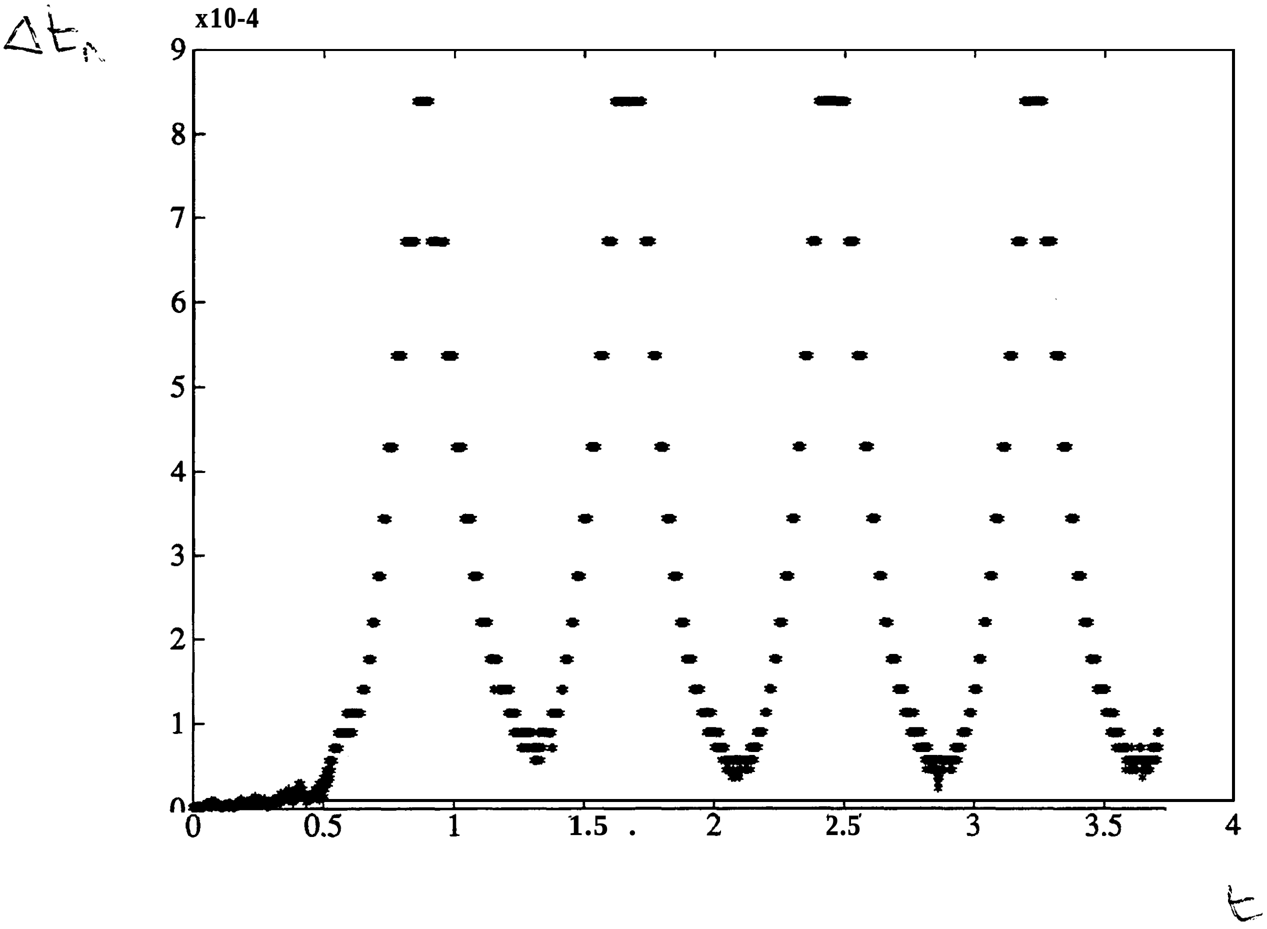

Figure 7.3

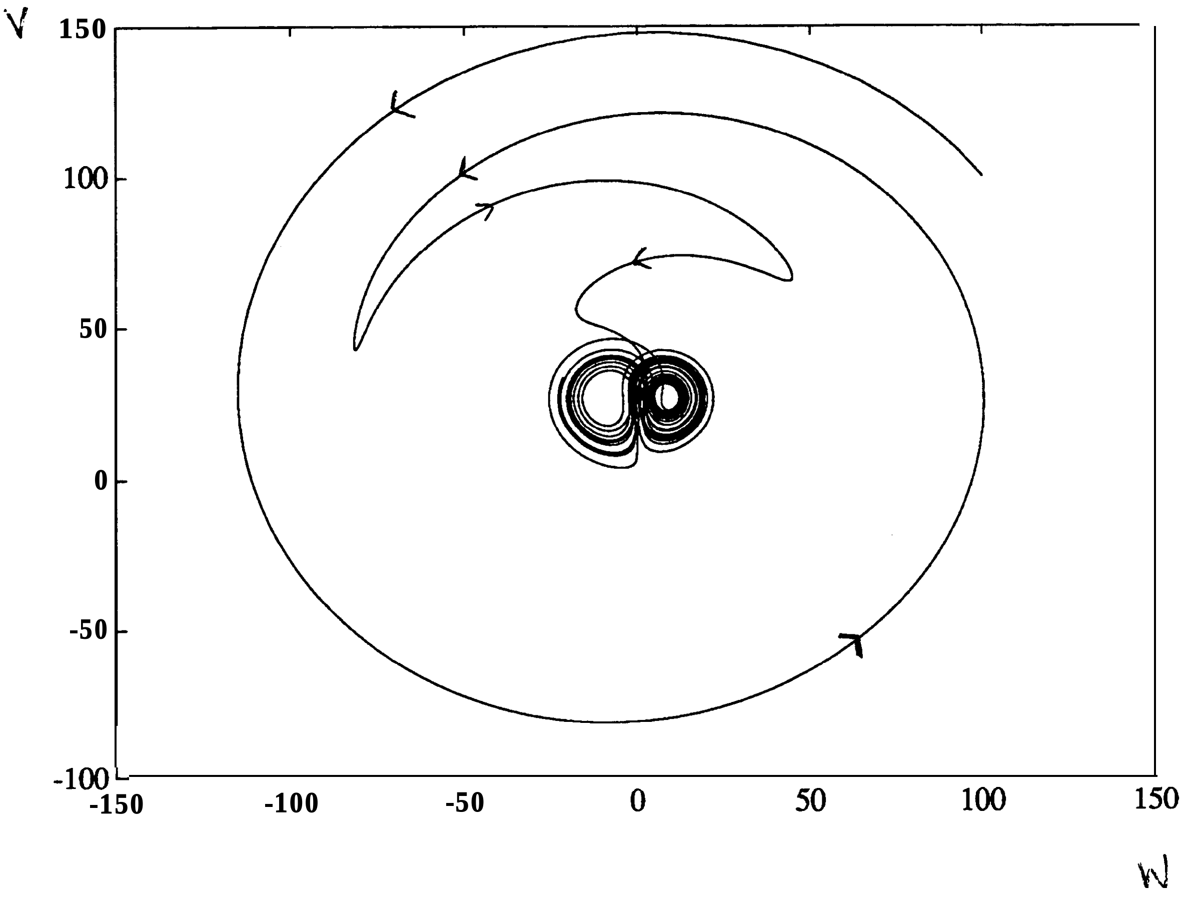

Figure 7.4

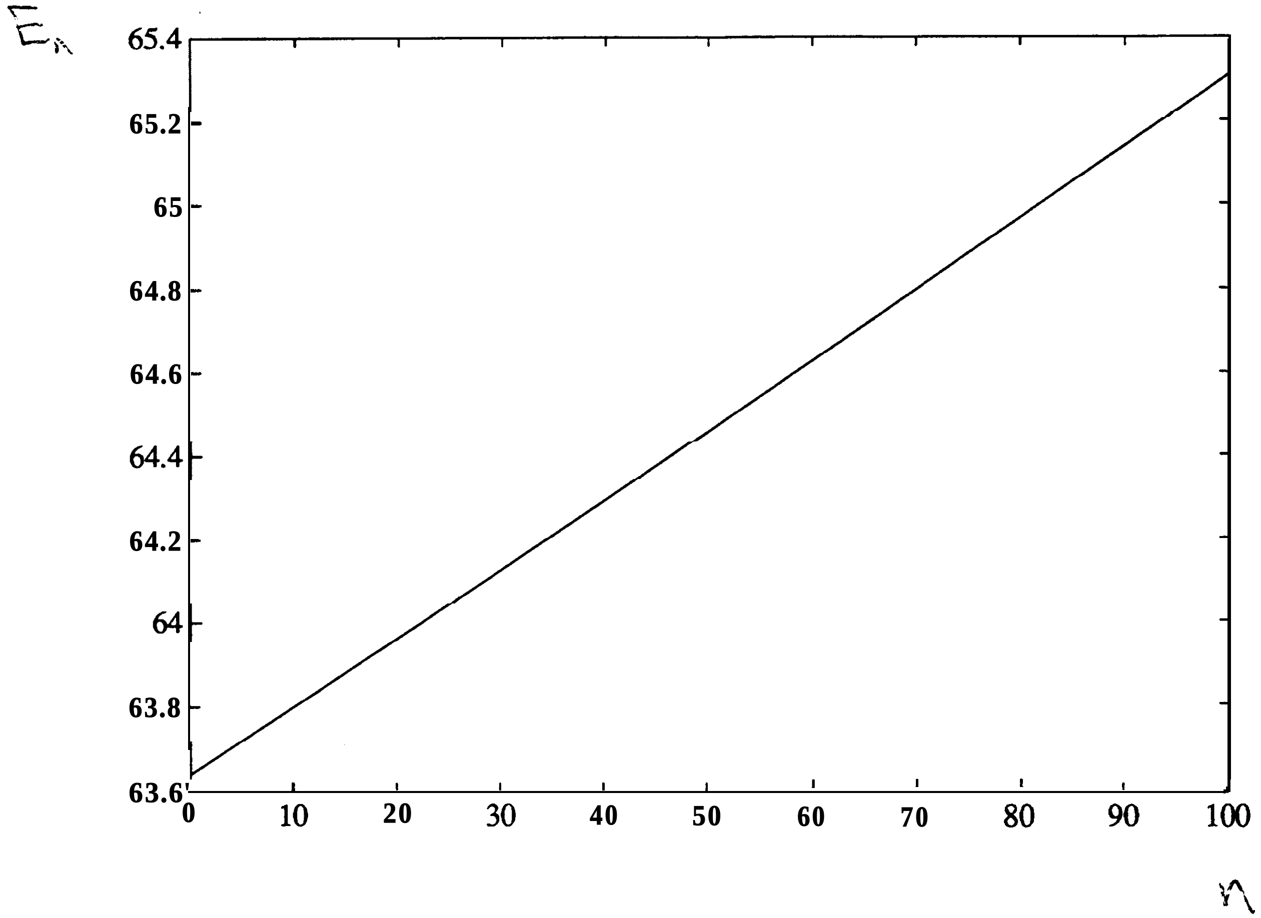
E_n
65.4
65.2
65
64.8
64.6
64.4
64.2
64
63.8
63.6
0
10
20
30
40
50
60
70
80
90
100
n

Figure 7-5

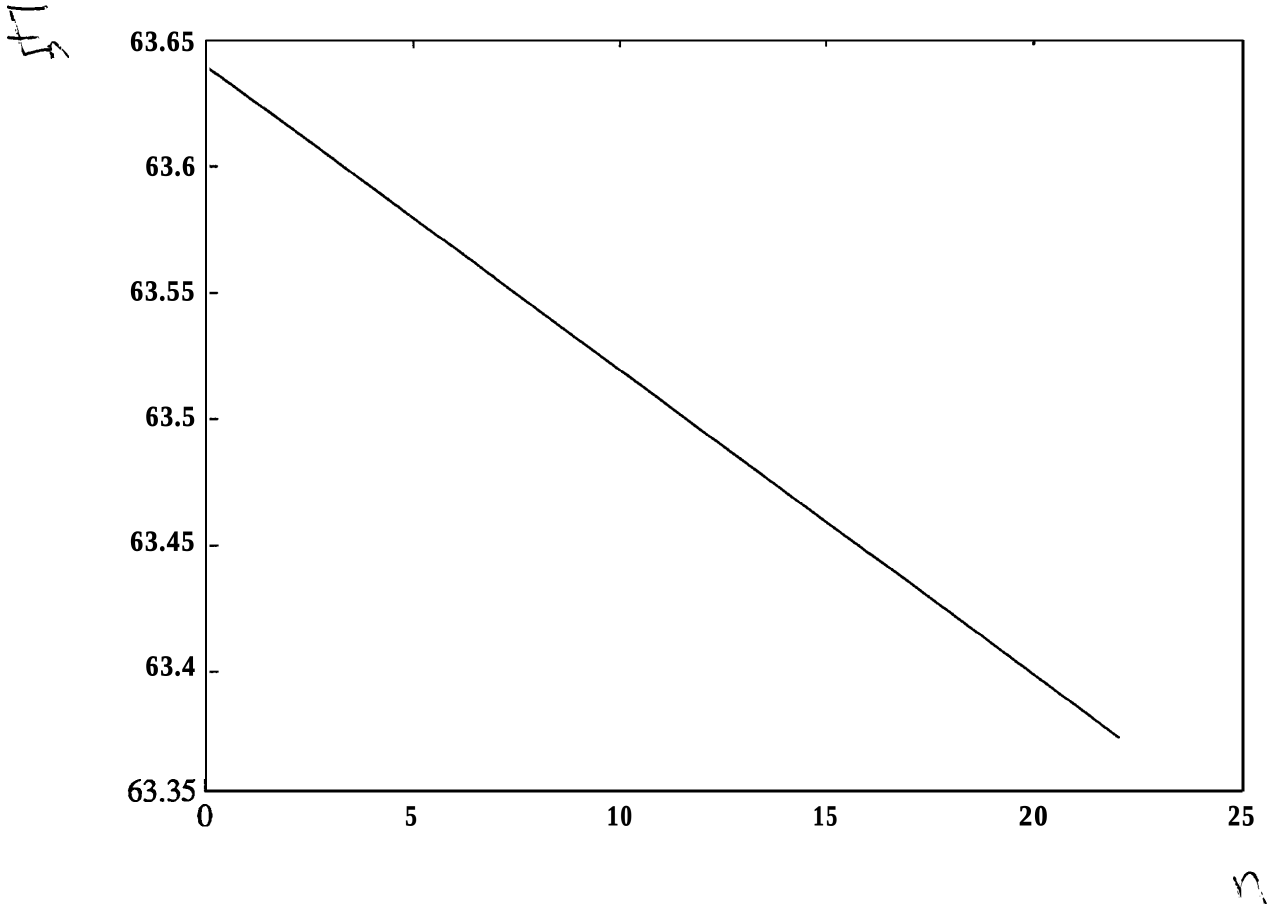

Figure 7.6.

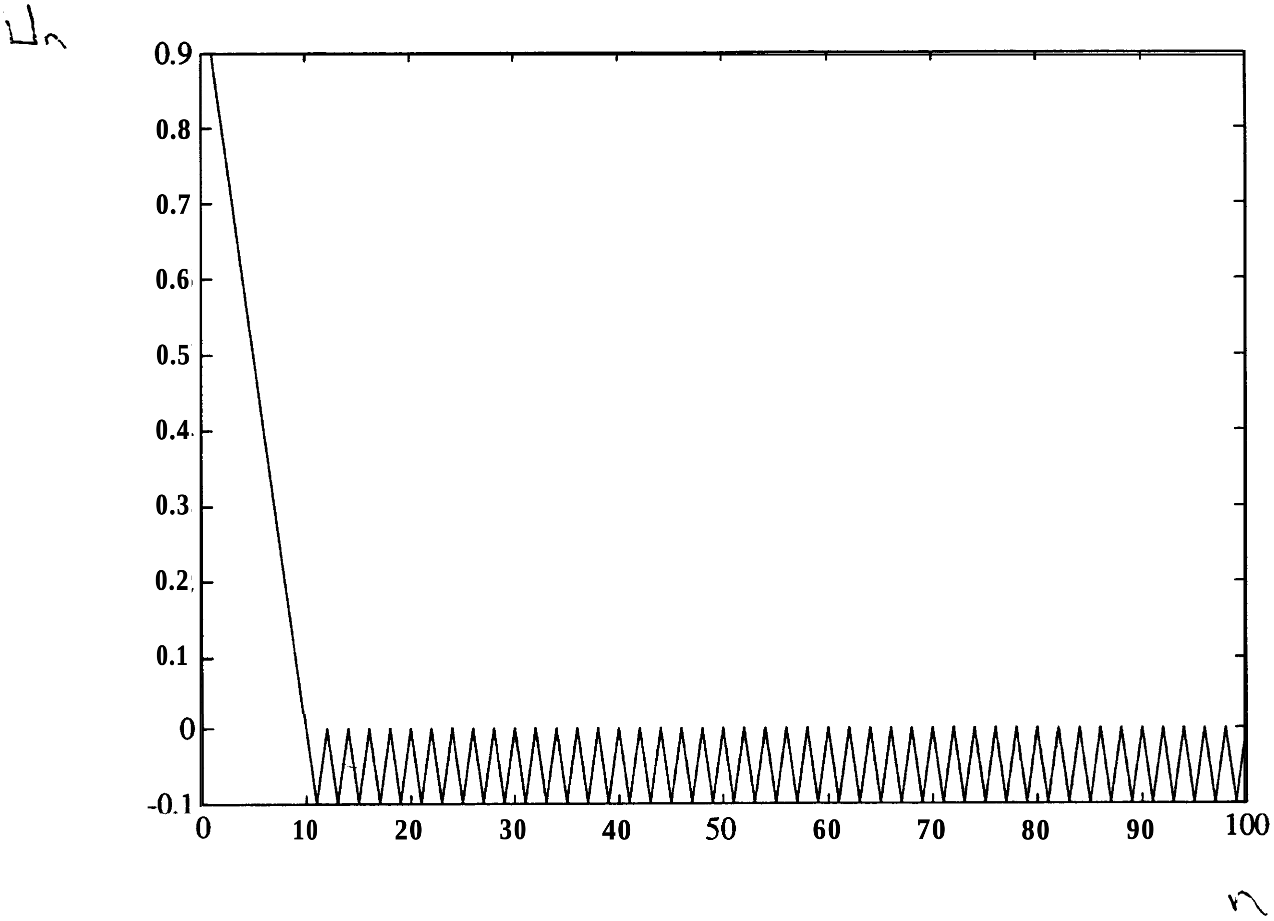

Figure 7-7.