BIT 38:1 (1998), 44–57

# ANALYSIS OF THE DYNAMICS OF LOCAL ERROR CONTROL VIA A PIECEWISE CONTINUOUS RESIDUAL *

D. J. HIGHAM[1]† and A. M. STUART[2]‡

[1] *Department of Mathematics, University of Strathclyde*
*Glasgow, G1 1XH, Scotland. email: d.j.higham@strath.ac.uk*

[2] *Scientific Computing and Computational Mathematics Program*
*Division of Mechanics and Computation, Stanford University*
*Stanford, CA 94305-4040, USA. email: stuart@sccm.stanford.edu*

## Abstract.

Positive results are obtained about the effect of local error control in numerical simulations of ordinary differential equations. The results are cast in terms of the local error tolerance. Under the *assumption* that a local error control strategy is successful, it is shown that a continuous interpolant through the numerical solution exists that satisfies the differential equation to within a small, piecewise continuous, residual. The assumption is known to hold for the MATLAB ode23 algorithm [10] when applied to a variety of problems.

Using the smallness of the residual, it follows that at any finite time the continuous interpolant converges to the true solution as the error tolerance tends to zero. By studying the perturbed differential equation it is also possible to prove discrete analogs of the long-time dynamical properties of the equation—dissipative, contractive and gradient systems are analysed in this way.

*AMS subject classification:* 34C35, 34D05, 65L07, 65L20, 65L50.

*Key words:* Error control, continuous interpolants, dissipativity, contractivity, gradient systems.

## 1 Introduction.

Here we study adaptive numerical algorithms for the initial value problem

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

where $u(t) \in \mathbb{R}^m$ for each $t \geq 0$; throughout $f: \mathbb{R}^m \mapsto \mathbb{R}^m$ is assumed to be locally Lipschitz. We study adaptive algorithms and *assume* that they control the true local error at each step. Under this assumption we prove that the

*Received July 1996. Revised July 1997. Communicated by Gustaf Söderlind.

†Supported by the Engineering and Physical Sciences Research Council under grants GR/H94634 and GR/K80228.

‡Supported by the Office of Naval Research under grant N00014-92-J-1876 and by the National Science Foundation under grant DMS-9201727.

long-time behaviour of certain classes of differential equation is replicated by the adaptive algorithm. The assumption we make is a significant one but it has recently been proved to hold for the MATLAB ode23 algorithm [10] when applied to (1.1) with $f(u) = -Au$ and $A$ invertible and also for certain nonlinear problems approximated by the same code [9].

To examine the effect of local error control on large time dynamics it is necessary to work with particular structural assumptions on the vector field $f(\bullet)$ that defines the differential equation (1.1). We work with three structural assumptions on $f(\bullet)$ all of which imply that equation (1.1) is dissipative in the general sense of the following definition:

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

This definition is taken from [6] and [17] and is widely employed in the modern literature of nonlinear semigroups.[1]

The three structural assumptions that we make are now detailed. Throughout, we assume that $\|\bullet\|$ denotes a norm in $\mathbb{R}^m$ induced by the appropriate inner product—that is, the one inherited from assumptions **(D)**, **(C)** or **(G)** below. We let

$$B(v,\delta) := \{u \in \mathbb{R}^m : \ \|v-u\| < \delta\}, \tag{1.2}$$

$$\mathcal{N}(X,\delta) := \{u \in \mathbb{R}^m : \ \exists x \in \bar{X} : \|x-u\| < \delta\}, \tag{1.3}$$

and

$$Q(\varepsilon) := \{v \in \mathbb{R}^m : \ \|f(v)\| \leq \varepsilon\}. \tag{1.4}$$

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

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

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

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

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

[1] The class of contractive problems (1.1), **(C)** defined herein are also sometimes termed dissipative in the numerical analysis literature, a nomenclature consistent with that in certain branches of semigroup theory [11]. However, in order to distinguish between the assumptions **(D)**, **(C)** we have chosen to follow the terminology of [6, 17] used in Definition 1.1.

The class of problems (1.1), **(D)** denotes dissipative problems with dissipativity induced via an inner-product and for which the rate of absorption into the absorbing set is at least exponentially fast; the Lorenz equations provide a natural example. The class (1.1), **(C)** denotes contractive problems, in the norm induced by the inner-product, with rate of contraction which is at least exponentially fast. In the terminology of [3], the condition **(C)** implies the existence of a negative one-sided Lipschitz constant. The class **(G)** denotes a natural class of gradient systems; the condition **(G3)** is equivalent to the existence of a positive one-sided Lipschitz condition—see [15], Lemma 2.8.19. The problem with $f(u) = -Au$ and $A$ positive-definite symmetric is a member of each of the three problem classes. Furthermore, all three problem classes are dissipative in the general sense of Definition 1.1:

THEOREM **ODE** [6, 15, 17]

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

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

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

Our aim is to derive discrete analogues of Theorem ODE for error-controlled numerical schemes. We make the ***assumption*** (discussed below) that the scheme successfully controls the ***local*** error-per-unit-step, and we show that this confers good ***global*** properties on the numerical solution. Our results are expressed in terms of the tolerance parameter; our assumptions do not require the time steps to be uniformly small as the error tolerance approaches zero. Indeed, typically they will not be small in neighbourhoods of equilibria.

This work is motivated by the analysis in [16], which concerns a class of adaptive Runge–Kutta schemes termed ***essentially algebraically stable.*** For these schemes, positive results about long-time behaviour are derived for sufficiently small tolerances but ***independently of initial data***; furthermore no assumptions are made about the error control other than that the sequence of time steps is uniformly bounded from above. However, the class of methods to which the analysis in [16] applies is very restrictive. The results that we derive in this paper apply to ***all*** time-stepping methods under the assumption that the local-error-per-unit-step is controlled by the particular adaptation strategy used—see Assumption 2.1. Although the results proved here apply to a wider class of methods than those used in [16], they are weaker than those in [16] in the sense that the tolerance must be chosen sufficiently small, ***depending upon initial data.*** The method of analysis differs from that in [16]: here we construct a continuous interpolant, $\eta(t)$, through the numerical solution, rather than dealing directly with the discrete approximation. We use the fact that the error control forces $\eta(t)$ to satisfy the system (1.1) to within a small residual. This facilitates a straightforward analysis, mimicking, to a large extent, the analysis for (1.1).

We mention that in [1] Aves et al. adopt a different approach to the study of long-term behaviour of error-controlled methods. That work is concerned solely with spurious fixed points, and it does not require any structural assumptions about $f(\bullet)$. The main conclusions in [1] are similar in spirit to those presented here: positive results can be derived about the effect of error control, for small tolerances. Specifically, whilst locally adaptive schemes may possess spurious fixed points in one dimension, these are shown to be typically unstable. In higher dimensions such spurious fixed points will typically not exist.

In section 2 we introduce some notation and definitions used throughout and state (and discuss) the basic assumption concerning the error control. In section 3 the numerical solution at any finite time is shown to converge to the true solution as the tolerance tends to zero. The numerical approximation under (**D**) and (**C**) is studied in section 4, and discrete counterparts of the properties of the true solution (see Theorem ODE (i),(ii)) are established for the error-controlled approximation. Section 5 contains similar discrete counterparts of the results for gradient systems satisfying (**G**)—see Theorem ODE (iii). The main contribution of the paper is contained in section 5 where we prove a strong result concerning the dynamics of the discretization (mimicking continuous time properties studied in [2]) from which a discrete analog of Theorem ODE (iii) is a simple corollary. The results in sections 3 and 4 are stated for completeness; their proofs are straightforward and may be found in [8].

## 2 Background.

We start with some notation and terminology used throughout.

DEFINITION 2.1. *We denote by* $S(U,t) : \mathbb{R}^m \times \mathbb{R}^+ \mapsto \mathbb{R}^m$ *the solution operator for* (1.1), *so that* $u(t) = S(U,t)$.

We suppose that an adaptive time-stepping method is used to compute approximations $U_n \approx u(t_n)$ with $U_0 = U$. The time step $\Delta t_n = t_{n+1} - t_n$, and hence $t_n = \sum_{j=0}^{n-1} \Delta t_j$. A user-supplied *tolerance parameter* $\tau > 0$ determines the time steps and so, indirectly, controls the accuracy of the process.

DEFINITION 2.2. *The local error over a step from* $t_n$ *to* $t_{n+1}$ *is defined to be* $\mathrm{LE}_{n+1} := U_{n+1} - S(U_n, \Delta t_n)$.

Our assumptions about the overall numerical method are as follows.

ASSUMPTION 2.1. *There is a constant* $D > 0$ *and, for any* $U \in \mathbb{R}^m$ *constants* $K = K(U) > 0$ *and* $\tau_c = \tau_c(U) > 0$ *such that, for all* $\tau \in (0, \tau_c)$, *the method produces a solution sequence* $\{\Delta t_n, U_n\}_{n=0}^{\infty}$ *with*

$$\sum_{n=0}^{\infty} \Delta t_n = \infty, \tag{2.1}$$

$$\sup_{n\geq 0}\{\Delta t_n\} \leq D, \tag{2.2}$$

$$\|\mathrm{LE}_{n+1}\| \leq K\tau\Delta t_n, \quad \forall n \geq 0. \tag{2.3}$$

Condition (2.1) is needed to ensure that the process does not break down at a finite time. (If, for example, $\Delta t_n = 1/n^2$ then the limit $n \to \infty$ does not correspond to the limit $t \to \infty$.) Condition (2.1) will be satisfied by the MATLAB ode23 code [10], for example, provided the computed sequence $\{U_n\}_{n=0}^{\infty}$ remains in a bounded set [9]. Condition (2.2) is reasonable since most codes have a maximum allowable time step constraint and $D$ will then be independent of $U$—MATLAB ode23, for example, has a maximum allowable time-step proportional to the length of the time integration; for the unbounded case $t \in [0, \infty)$, $D$ can be obtained by repeatedly using the code on finite time intervals.

The third condition, (2.3), is our key assumption—the adaptive method must successfully control the local-error-per-unit-step in terms of the tolerance. Most adaptive algorithms attempt, either directly or indirectly, to control some measure of the local error over each step. This is usually done by applying two formulas and taking the difference to obtain an approximation of the true local error, the ***local error estimate***. Unfortunately the local error estimate can vanish at places where the true local error is non-zero and it is then possible for (2.3) to fail. However, for linear problems $f(u) = -Au$ with $A$ invertible it is proved in [9], Corollary 3.4 that (2.3) holds for the MATLAB ode23 code provided that the computed sequence $\{U_n\}_{n=0}^{\infty}$ remains in a compact set. For typical nonlinear problems approximated by MATLAB ode23, (2.3) will hold for all but a set of initial data of small Lebesgue measure—see Theorem 3.12 and section 4.2 of [9]. Thus we believe that (2.3) is a good assumption to make since it is intuitively reasonable and it is provably true in certain circumstances. Note that by working with Assumption 2.1 we are able to derive results about adaptive time-step software without studying the details of the time-step selection mechanism.

The condition (2.3) was used in the early work on tolerance proportionality by Shampine [12]. Stetter investigated conditions under which local error estimates could break down, meaning that (2.3) might not hold; he then proceeded to study tolerance proportionality on the assumption that this break-down did not occur [13]. Some recent work of Stoffer and Nipp [14] proves interesting, positive results about the behaviour of variable time step integrators in the neighbourhood of a periodic solution; specifically, convergence of an approximate invariant curve to the true periodic solution is established as $\tau \to 0$. The work in [14] makes the ***assumption*** that $\Delta t_n \leq K\tau^{\nu}$ for some $\nu > 0$, which implies (2.3); however, unlike (2.3), this assumption requires the time step to be uniformly small for all time, something which is not true in general, in particular for solutions which pass close to equilibria [7]. Note that for the MATLAB ode23 code, (2.3) can hold for solutions passing arbitrarily near equilibria without $\Delta t_n$ necessarily being small. This is because local errors for Runge-Kutta methods are proportional to $f(U_n)$, which is small near equilibria.

## 3 Basic error estimate.

We now construct a piecewise continuous interpolant to the numerical solution sequence. The same construction was used by Stetter [13] to study global error behaviour in a different context. The construction is a purely theoretical device since it employs the local error (which is not known) to form the interpolant. Our main use of this theoretical tool will be to facilitate analysis of the numerical algorithm by use of *continuous time* analysis; this enables us to build directly on the methods of analysis used for the underlying equation (1.1). Since the interpolant passes through the numerical data this analysis yields direct information about the computed solution sequence.

In this section we will show that, under Assumption 2.1, the interpolant satisfies a perturbed version of the original system (1.1). Bounding the residual in terms of the error tolerance allows the global error to be assessed. Similar results may already be found in the literature (see [12, 13]) and we state them here as a basic introduction to our methodology.

DEFINITION 3.1. *Given $\{\Delta t_n, U_n\}_{n=0}^{\infty}$ we define the numerical interpolant $\eta(t)$ by $\eta(0) = U$ and*

$$\eta(t) := S(U_n, t - t_n) + \frac{(t - t_n)}{\Delta t_n}\mathrm{LE}_{n+1}, \quad \forall t \in (t_n, t_{n+1}]. \tag{3.1}$$

By construction $\eta(t_{n+1}) = U_{n+1}$ and as $t$ tends to $t_n$ from above $\eta(t)$ tends to $U_n$. Hence $\eta(t)$ is continuous for $t > 0$. The first derivative $\eta_t(t)$, however, is not continuous in general. For definiteness, we will define $\eta_t(t_n)$ by taking the limit from the left; that is,

$$\eta_t(t_n) := \lim_{t \to t_n{}^-} \eta_t(t).$$

The following result shows that the numerical interpolant satisfies an equation which is a small perturbation of (1.1).

THEOREM 3.1. *Suppose that Assumption 2.1 holds, and that there is a bounded set $X$ and time $T > 0$ such that $\eta(t) \in X \subset I\!R^m$ for all $t \in [0, T]$. If $L$ is the Lipschitz constant for $f$ on the set $Y := \mathcal{N}(X, \delta)$ for some $\delta > 0$, $K := \sup_{U \in Y} K(U) < \infty$ and $\tau \in (0, \delta/DK)$, then*

$$\|\eta_t - f(\eta)\| \leq (1 + LD)K\tau, \quad \forall t \in [0, T]. \tag{3.2}$$

PROOF. See [8]. □

The next result shows that, at any finite time $T$, the numerical interpolant converges to the true solution of (1.1) as the tolerance is decreased. The result is a minor extension of the "fundamental lemma" (Theorem 10.2 of [5]) to allow for the fact that we do not have a global Lipschitz constant for $f$.

THEOREM 3.2. *Suppose that Assumption 2.1 holds, and that there is a bounded set $X$ and time $T > 0$ such that $u(t) \in X \subset I\!R^m$ for all $t \in [0, T]$. If $L$ is the Lipschitz constant for $f$ on the set $Y := \mathcal{N}(X, \delta)$ for some $\delta > 0$ and $K := \sup_{U \in Y} K(U) < \infty$, then there exists $\tau^* = \tau^*(\delta, L, T)$ and $C = C(L, t)$ such that, if $\tau \in (0, \tau^*)$, we have*

$$\|\eta(t) - u(t)\| \leq C(L, t)\tau \quad \forall t \in [0, T].$$

PROOF. See [8]. □

## 4 Dissipative and contractive problems.

Here we prove that the numerical interpolant mimics the asymptotic behaviour of solutions to (1.1) when applied to dissipative and contractive systems. We start with the dissipative system and then give a result concerning contractive problems as a corollary.

The following result shows that the numerical method is dissipative with an absorbing set which may be chosen to be the same as the absorbing set for the underlying differential equation. Note however that $\tau$ must be chosen sufficiently small, dependent upon initial data, to obtain this result. Note also that, in the proof, $\tau$ must be chosen proportional to $\delta$.

THEOREM 4.1. *Suppose that* (1.1) *satisfies Assumption* (**D**) *and is approximated by a numerical method satisfying Assumption* 2.1. *For any* $R, \delta > 0$, *there exists* $\tau_c = \tau_c(R, \delta) > 0$ *such that, if* $\tau \in (0, \tau_c)$, *there is a* $T = T(R, \delta) > 0$ *with the property that, for any* $U \in B(0, R)$,

$$\eta(t) \in \bar{B}(0, \sqrt{\alpha/\beta + \delta}) \quad \forall t \geq T.$$

PROOF. See [8]. □

The following is a counterpart of Theorem ODE (ii) for the continuous interpolant. However, rather than establishing convergence of the numerical interpolant to zero, we establish convergence into a ball of radius proportional to $\tau$. Such a result is "best possible" in the sense that the work of Griffiths [4] and Hall [7] indicates that exact solutions of error control schemes exist which oscillate in neighbourhoods of equilibria whose diameters tend to zero with the tolerance.

THEOREM 4.2. *Suppose that* (1.1) *satisfies Assumption* (**C**) *and is approximated by a numerical method satisfying Assumption* 2.1. *There is a universal constant* $\kappa > 0$ *so that for any* $U \in B(0, R)$, *there exists* $\tau_c = \tau_c(R) > 0$ *and* $T = T(R)$ *such that, if* $\tau \in (0, \tau_c)$,

$$\eta(t) \in B(0, \kappa\tau) \quad \forall t \geq T.$$

PROOF. See [8]. □

For contractive problems satisfying (**C**), it is also possible to show the much stronger result that $\eta(t)$ remains uniformly close to $S(t, U)$ for all $t > 0$.

THEOREM 4.3. *Suppose that Assumption* 2.1 *and* (**C**) *hold. Then there is a bounded set* $X$ *such that* $u(t) \in X \subset \mathbb{R}^m$ *for all* $t \geq 0$. *If* $L$ *is the Lipschitz constant for* $f$ *on the set* $Y := \mathcal{N}(X, \delta)$ *for some* $\delta > 0$ *and* $K := \sup_{U \in Y} K(U) < \infty$, *then there exists* $\tau^* = \tau^*(\delta, L)$ *and* $C = C(L)$ *such that, if* $\tau \in (0, \tau^*)$, *we have*

$$\|\eta(t) - u(t)\| \leq C(L)\tau \quad \forall t \geq 0.$$

PROOF. The proof of Theorem 3.4 in [8] can be adapted by exploiting one-sided Lipschitz constants. □

We finish this section with a few remarks concerning the dynamics of (1.1) under Assumption (**D**), and its numerical approximation, inside the set

$$\mathcal{B} := \bar{B}(0, \sqrt{\alpha/\beta + \delta}). \tag{4.1}$$

Recall that a compact invariant set $\mathcal{A}$ is a *global attractor* for the solution operator $S(\bullet, t)$ if it attracts every bounded open neighbourhood of itself. See [15] for precise definitions of these dynamical systems concepts in the context of numerical analysis. By applying, for example, Theorems 2.8.8 and 2.8.13 in [15] it is straightforward to show that (1.1) under (**D**) has a global attractor $\mathcal{A} \subseteq \mathcal{B}$. The set of points comprising the attractor provides a description of where the complicated dynamics lie within the set $\mathcal{B}$. We now show that under our error control assumptions the numerical approximations eventually lie close to the true attractor $\mathcal{A}$.

THEOREM 4.4. *Suppose that* (1.1) *satisfies Assumption* (**D**) *and is approximated by a numerical method satisfying Assumption* 2.1. *Then equation* (1.1) *has a global attractor* $\mathcal{A}$ *contained in* $\mathcal{B}$ *given by* (4.1). *Furthermore, for any* $\varepsilon > 0$ *there exist* $T, \tau_c > 0$ *such that for* $U_0 \in \mathcal{B}$ *and any* $\tau \in (0, \tau_c)$,

$$\eta(t) \in \mathcal{N}(\mathcal{A}, \varepsilon) \quad \forall t \geq T.$$

PROOF. See [8]. □

Note that Theorem 4.1 shows that all solutions starting in a bounded set $E$ eventually enter $\mathcal{B}$ for $\tau$ sufficiently small, depending upon $E$. Thus Theorem 4.4 is a statement about the long time dynamics of *all* numerical solutions, provided $\tau$ is sufficiently small.

We have *not* proved that the numerical approximation has an attractor. There are two reasons for this. First, to do so we would need to define a dynamical system on $\mathbb{R}^m \times \mathbb{R}^+$ generating sequences $(U_n^T, \Delta t_n)^T \in \mathbb{R}^{m+1}$ and this requires specification of a time step mechanism; in contrast the viewpoint taken in this paper is to say as much as possible about sequences satisfying Assumption 2.1 without specifying the time step selection mechanism. Second, even if a time step selection mechanism is chosen, it will typically be discontinuous. Hence the standard theory of dynamical systems does not apply and, for example, important properties such as invariance of limit sets no longer hold.

## 5 Gradient systems.

Recall Theorem ODE(iii). Here we prove that for gradient systems, the numerical interpolant mimics this key property of the true solution. Specifically we show that the numerical solution eventually enters and remains in a small ball about an equilibrium point; the ball has radius proportional to $\tau$. As in the case of Theorem 4.2, this is consistent with the work of Griffiths [4] and Hall [7].

We will assume without explicit statement that (**G**) holds throughout Lemmas 5.1–5.5. The equilibrium points are isolated and finite in number under (**G**) and are labelled $\{v_i\}_{i=1}^J$. Because they are isolated we may define

$$0 < \gamma := \min_{k \neq l} \|v_k - v_l\|. \tag{5.1}$$

In the following we will use $Q$ to denote $Q(\varepsilon)$.

The first two lemmas are Lemmas 5.1 and 5.2 of [16] respectively.

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

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

LEMMA 5.2. *There is a constant $\varepsilon^* > 0$ such that, if $\varepsilon \in [0, \varepsilon^*)$ and $v \in Q$, then there exists $v_j \in Q(0)$ such that $v \in Q_j$ and*

$$|F(v) - F(v_j)| \leq C(1 + kC)\varepsilon^2.$$

Note that, without loss of generality, we have taken the same $\varepsilon^*$ in both the previous lemmas. Next we prove three lemmas that allow us to reach the final theorem. To help orient the reader, we first summarize these lemmas loosely in words, and explain how they combine to prove the result.

Our main aim is to show that $\eta(t)$ is eventually contained in a ball of radius proportional to $\tau$ around some equilibrium point. We know from Lemma 5.1 that, for small $\varepsilon$, $Q \equiv Q(\varepsilon)$ is made up of distinct neighbourhoods $\{Q_j\}_{j=1}^J$ that are contained in balls of radius proportional to $\varepsilon$ around the equilibrium points. Taking $\varepsilon$ as some multiple of $\tau$, it would therefore be sufficient to show that $\eta(t)$ is eventually contained in $Q$ for all $t$ sufficiently large. Lemma 5.3 below shows that if $\eta(t)$ is outside $Q$ then $\| \eta_t(t) \|$ can be bounded away from zero, and $F(\eta(t))$ is strictly monotonic decreasing. But by (**G2**), $F$ is bounded below, and so it follows that $\eta(t)$ cannot remain outside $Q$ indefinitely. If $\eta(t)$ eventually remains inside one neighbourhood, $Q_j$, then we are done. Otherwise, $\eta(t)$ must leave and re-enter $\bigcup_{j=1}^J Q_j$ indefinitely. Lemma 5.4 shows that if $\eta(t)$ leaves and re-enters the *same* neighbourhood $Q_i$, without visiting any other $Q_j$, $j \neq i$, in the meantime, then $\eta(t)$ cannot stray more than an $\mathcal{O}(\tau)$ distance from the equilibrium point $v_i$, as required. The remaining possibility is that $\eta(t)$ continues to visit *different* members of $\{Q_j\}_{j=1}^J$. However, Lemma 5.5 shows that if $Q_j$ and $Q_i$, with $j \neq i$, are visited in succession, then $F(v_i) \leq F(v_j) - C_2\tau$, for some constant $C_2$. Since $F$ is bounded below, this cannot continue indefinitely—there must be a *final* neighbourhood to be visited, and Lemma 5.4 then becomes applicable.

LEMMA 5.3. *Suppose that Assumption* 2.1 *holds, and that there is a bounded set $X$ and time $T > 0$ such that $\eta(t) \in X \subseteq \mathbb{R}^m$ for all $t \in [0, T]$. Let $L$ denote* *the Lipschitz constant for $f$ on the set $Y := \mathcal{N}(X, \delta)$, $E = (1 + LD)K$ and $\varepsilon = 3E\tau$. Then, if*

$$\tau < \tau^* := \min\left\{\frac{\delta}{DK}, \frac{\varepsilon^*}{3E}\right\},$$

*$t \in [0, T]$ and $\eta(t) \notin Q$ it follows that*

$$\|\eta_t(t)\| \geq 2E\tau$$

*and*

$$\frac{d}{dt}\{F(\eta(t))\} \leq -\|\eta_t(t)\|E\tau.$$

PROOF. Theorem 3.1 holds. Using the property (**G1**),

$$\langle\eta_t(t), \eta_t(t)\rangle = -\langle\nabla F(\eta(t)), \eta_t(t)\rangle + \langle r(t), \eta_t(t)\rangle.$$

So,

$$\frac{d}{dt}\{F(\eta(t))\} = \langle\nabla F(\eta(t)), \eta_t(t)\rangle = -\langle\eta_t(t), \eta_t(t)\rangle + \langle r(t), \eta_t(t)\rangle.$$

From the Cauchy-Schwarz inequality, it follows that

$$(5.2)\quad \frac{d}{dt}\{F(\eta(t))\} \leq -\|\eta_t(t)\|^2 + \|r(t)\|\|\eta_t(t)\| \leq -\|\eta_t(t)\|(\|\eta_t(t)\| - E\tau),$$

using the bound from Theorem 3.1.

Finally, with $\varepsilon = 3E\tau$ and $\eta(t) \notin Q$, we have $\|f(\eta(t))\| > 3E\tau$, so that, from (3.2),

$$\|\eta_t(t)\| \geq \|f(\eta(t))\| - \|r(t)\| > 2E\tau.$$

Using this in (5.2) completes the proof. □

In the remainder of the analysis we take $\varepsilon = 3E\tau$.

LEMMA 5.4. *Suppose that Assumption* 2.1 *holds, and that there is a bounded set $X$ and time $T > 0$ such that $\eta(t) \in X \subseteq \mathbb{R}^m$ for all $t \in [0, T]$. Let $L$ denote the Lipschitz constant for $f$ on the set $Y := \mathcal{N}(X, \delta)$, $E = (1 + LD)K$ and $\varepsilon = 3E\tau$. Then, if*

$$\tau < \tau^* := \min\left\{\frac{\delta}{DK}, \frac{\varepsilon^*}{3E}\right\},$$

*$t \in [0, T]$, $\eta(t_\pm) \in Q_i$, with $t_- < t_+ \leq T$ and $\eta(t) \notin Q$ for $t \in (t_-, t_+)$, there exists $C_1 > 0$ such that*

$$\|\eta(t) - v_i\| \leq C_1\tau, \quad \text{for all } t \in [t_-, t_+].$$

PROOF. From Lemma 5.2, setting $\bar{c} := C(1 + kC)9E^2$, we have

$$\begin{aligned} |F(\eta(t_-)) - F(v_i)| &\leq \bar{c}\tau^2, \\ |F(\eta(t_+)) - F(v_i)| &\leq \bar{c}\tau^2. \end{aligned}$$

So, from the triangle inequality,

$$|F(\eta(t_+)) - F(\eta(t_-))| \leq 2\bar{c}\tau^2.$$

But

$$|F(\eta(t_+)) - F(\eta(t_-))| = \left| \int_{t_-}^{t_+} \frac{d}{dt}\{F(\eta(t))\}\, dt \right| \geq E\tau \int_{t_-}^{t_+} \|\, \eta_t(t) \,\|\, dt,$$

using Lemma 5.3. Hence,

$$2\bar{c}\tau^2 \geq E\tau \int_{t_-}^{t_+} \|\, \eta_t(t) \,\|\, dt,$$

giving

$$\frac{2\bar{c}\tau}{E} \geq \int_{t_-}^{t_+} \|\, \eta_t(t) \,\|\, dt. \tag{5.3}$$

Now, using (5.3), for any $t \in [t_-, t_+]$,

$$\|\, \eta(t) - \eta(t_-) \,\| \leq \int_{t_-}^{t_+} \|\, \eta_t(t) \,\|\, dt \leq \frac{2\bar{c}\tau}{E}.$$

Lemma 5.1 gives $\|\, \eta(t_-) - v_i \,\| \leq C\varepsilon$, and hence

$$\|\, \eta(t) - v_i \,\| \leq \|\, \eta(t) - \eta(t_-) \,\| + \|\, \eta(t_-) - v_i \,\| \leq \frac{2\bar{c}\tau}{E} + C\varepsilon = (\frac{2\bar{c}}{E} + 3CE)\tau.$$

So, the result holds with $C_1 = 2\bar{c}/E + 3CE$. □

LEMMA 5.5. *Suppose that Assumption* 2.1 *holds, and that there is a bounded set $X$ and time $T > 0$ such that $\eta(t) \in X \subseteq \mathbb{R}^m$ for all $t \in [0, T]$. Let $L$ denote the Lipschitz constant for $f$ on the set $Y := \mathcal{N}(X, \delta)$, $E = (1 + LD)K$ and $\varepsilon = 3E\tau$. Then, if*

$$\tau < \bar{\tau} := \min\left\{\tau^*, \frac{E\gamma}{4[3C^2E^2 + \bar{c}]}\right\},$$

*$\eta(t_+) \in Q_i$, $\eta(t_-) \in Q_j$ with $t_- < t_+ < T$, $i \neq j$, and $\eta(t) \notin Q$ for $t \in (t_-, t_+)$, then there exists $C_2 > 0$ such that*

$$F(v_i) \leq F(v_j) - C_2\tau.$$

PROOF. Let $\varepsilon \in (0, \varepsilon^*)$ for $\varepsilon^*$ given by Lemmas 5.1 and 5.2. We have

$$F(\eta(t_+)) - F(\eta(t_-)) = \int_{t_-}^{t_+} \frac{d}{dt}\{F(\eta(t))\}\, dt \leq -E\tau \int_{t_-}^{t_+} \|\, \eta_t(t) \,\|\, dt,$$

from Lemma 5.3. Hence,

$$F(\eta(t_+)) - F(\eta(t_-)) \leq -E\tau \|\, \eta(t_+) - \eta(t_-) \,\|.$$

So, using Lemma 5.1, the triangle inequality and (5.1),

$$F(\eta(t_+)) - F(\eta(t_-)) \le -E\tau\left[\|\, v_j - v_i \,\| - 2C\varepsilon\right] \le -E\tau\gamma + 6CE^2\tau^2. \tag{5.4}$$

Also, using Lemma 5.2 with $\bar{c}$ given in the proof of Lemma 5.4,

$$\begin{aligned} F(\eta(t_+)) &\ge F(v_i) - \bar{c}\tau^2, \\ -F(\eta(t_-)) &\ge -F(v_j) - \bar{c}\tau^2, \end{aligned}$$

which, together, imply that

$$F(\eta(t_+)) - F(\eta(t_-)) \ge F(v_i) - F(v_j) - 2\bar{c}\tau^2. \tag{5.5}$$

Now, combining (5.4) and (5.5) gives

$$F(v_i) - F(v_j) \le -E\tau\gamma + 6CE^2\tau^2 + 2\bar{c}\tau^2 = -\tau\big[E\gamma - 6C^2E^2\tau - 2\bar{c}\tau\big].$$

By choosing $\tau$ sufficiently small, we can ensure that $E\gamma - 6C^2E^2\tau - 2\bar{c}\tau \ge E\gamma/2$, and hence the required result holds with $C_2 = E\gamma/2$. □

The desired analog of Theorem ODE(iii) follows from (ii) of the next theorem after noting that $t_N^+ = \infty$. Actually we prove more than a discrete analog of Theorem ODE (iii). The work of Babin and Vishik [2] shows that, for (1.1) under **(G)**, the solution passes through a finite number of small neighbourhoods of equilibria, ordered by decreasing Lyapunov function $F(\bullet)$. Such a result is true for the numerical approximation, with neighbourhoods of $\mathcal{O}(\tau)$, as the following theorem shows.

THEOREM 5.6. *Suppose that* (1.1) *satisfies Assumptions (G) and is approximated by a numerical method satisfying Assumption 2.1. Then there are constants* $\kappa_1, \kappa_2$ *and* $\tau_c = \tau_c(U) > 0$ *such that, if* $\tau \in (0, \tau_c)$*, there is an integer* $N$ *and times* $\{t_i^\pm\}_{i=1}^N$ *satisfying:*

*(i)* $t_i^- < t_i^+ < t_{i+1}^-, i = 1, \dots, N-1, 0 \le t_1^-, t_N^- < t_N^+ = \infty;$

*(ii)* $\eta(t) \in B(v_i, \kappa_1\tau), \quad \forall t \in (t_i^-, t_i^+);$

*(iii)* $F(v_i) \le F(v_{i-1}) - \kappa_2\tau, \quad i = 2, \dots, N.$

PROOF. Let

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

and define

$$Z := \{u \in \mathbb{R}^m : F(u) \le \{F(U), \bar{F} + \delta\}\}$$

noting that this is a compact set by **(G2)**. Let $X = \mathcal{N}(Z, \delta)$ and let $L$ denote the Lipschitz constant for $f$ on $\mathcal{N}(X, \delta)$. We define $E, \varepsilon, \bar{\tau}$ as in the previous lemmas of this section. Let

$$\tau_c := \min\left\{\bar{\tau}, \sqrt{\frac{\delta}{\bar{c}}}\right\}.$$

Note that $\eta(0) = U \in Z$ so that, by continuity there exists $t^* > 0$ such that $\eta(t) \in X$ for all $t \in [0, t^*)$. On this interval Lemmas 5.3, 5.4 and 5.5 hold with $X = \mathcal{N}(Z, \delta)$. Now consider $t \in [0, t^*)$. If $\eta(t) \in Q(\tau)$ then Lemma 5.2 gives, for some $v_j \in Q(0)$,

$$F(\eta(t)) \leq F(v_j) + \bar{c}\tau^2$$

so that

$$F(\eta(t)) \leq \bar{F} + \delta. \tag{5.6}$$

If $\eta(t) \notin Q(\tau)$ then either there exists $t_c < t$ such that $\eta(t_c) \in Q(\tau)$ and $\eta(s) \notin Q(\tau)$ for all $s \in (t_c, t]$ or $\eta(s) \notin Q(\tau)$ for all $s \in [0, t]$. In the first case

$$F(\eta(t)) \leq F(\eta(t_c)) \leq \bar{F} + \delta \tag{5.7}$$

by Lemmas 5.2 and 5.3 and the same argument leading to (5.6); in the second case

$$F(\eta(t)) \leq F(\eta(0)) = F(U) \tag{5.8}$$

by Lemma 5.3. Points (5.6), (5.7) and (5.8) show that $\eta(t) \in Z$ for all $t \in [0, t^*)$. Hence we may take $t^*$ arbitrarily large for otherwise we have a contradiction to the continuity of $\eta(t)$. It follows that $\eta(t) \in Z$ for all $t \in (0, \infty)$.

We now have that Lemmas 5.3, 5.4 and 5.5 hold for any $T \in [0, \infty)$. Let $i = 1, \ldots, N$ be an index set labelling those $v_i$ such that $\eta(t) \in Q_i$ for some $t \geq 0$. Note that $N \leq J$ and is hence finite. Define

$$\begin{aligned} t_i^- &= \inf\{t \in \mathbb{R}^+ : \eta(t) \in Q_i\}, \\ t_i^+ &= \sup\{t \in \mathbb{R}^+ : \eta(t) \in Q_i\}, \end{aligned}$$

for each $i = 1, \ldots, N$. By Lemma 5.5 it follows that there is no integer $j \neq i$ such that $\eta(t) \in Q_j$ for $t \in (t_i^-, t_i^+)$. Furthermore, by Lemmas 5.1 and 5.4 it follows that there exists $\kappa_1 = \max\{C_1, C\} > 0$ such that

$$\eta(t) \in B(v_i, \kappa_1\tau) \quad \forall t \in (t_i^-, t_i^+).$$

Possibly by further reduction of $\tau_c$ we can ensure that

$$B(v_i, \kappa_1\tau) \bigcap B_j(v_j, \kappa_1\tau) = \emptyset, \quad \forall i \neq j, \tag{5.9}$$

since (5.1) holds. Now note that there exists at least one integer $i$ such that (ii) holds for $t \in (t_i^-, t_i^+)$ since, if not, then $\eta(t) \notin Q$ for all $t \in \mathbb{R}^+$ and then, by Lemma 5.3,

$$\frac{d}{dt}\{F(\eta(t))\} \leq -2E^2\tau^2, \quad \forall t \in \mathbb{R}^+;$$

this contradicts (**G2**). Now, by (5.9), we may re-order the intervals so that $t_i^+ < t_{i+1}^-$ as required and point (iii) follows by Lemma 5.5. Note that $t_N^+ = \infty$, since otherwise, by Lemma 5.3,

$$\frac{d}{dt}\{F(\eta(t))\} \leq -2E^2\tau^2, \quad \forall t \geq t_N^+,$$

again contradicting (**G2**). □

## REFERENCES

1. M. A. Aves, D. F. Griffiths and D. J. Higham, *Does error control suppress spuriosity?* SIAM J. Num. Anal., 34 (1997), pp. 756–778.
2. A. V. Babin and M. I. Vishik, *Attractors of Evolution Equations*, Studies in Applied Mathematics 25, North-Holland, New York, 1992.
3. K. Dekker and J. G Verwer, *Stability of Runge-Kutta Methods for Stiff Nonlinear Differential Equations*, North-Holland, Amsterdam, 1984.
4. D. F. Griffiths, *The dynamics of some linear multistep methods with step-size control*, in Numerical Analysis 1987, D. F. Griffiths and G. A. Watson eds., Longman, 1988, pp. 115–134.
5. E. Hairer, S. P. Nørsett and G. Wanner, *Solving Ordinary Differential Equations I*, Springer-Verlag, New York, 1987.
6. J. K. Hale, *Asymptotic Behaviour of Dissipative Systems*, AMS Mathematical Surveys and Monographs 25, Rhode Island, 1988.
7. G. Hall, *Equilibrium states of Runge-Kutta schemes*, ACM Trans. Math. Software, 11 (1985), pp. 289–301.
8. D. J. Higham and A. M. Stuart, *Analysis of the dynamics of local error control via a piecewise continuous residual*, Stanford University Technical Report SCCM-95-03, 1995.
9. H. Lamba and A. M. Stuart, *Convergence results for the* MATLAB *ode23 routine*, Preprint, 1997.
10. The Math Works, Inc., *MATLAB User's Guide*, Natick, Massachusetts, 1992.
11. A. Pazy, *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Springer-Verlag, New York, 1983.
12. L. F. Shampine, *Tolerance proportionality in ODE codes*, in Numerical Methods for Ordinary Differential Equations (Proceedings), A. Bellen, C. W. Gear and E. Russo eds., Springer-Verlag, Lecture Notes 1386, 1987, pp. 118–136.
13. H. J. Stetter, *Tolerance proportionality in ODE-codes*, in Proc. Second Conf. on Numerical Treatment of Ordinary Differential Equations, R. März ed., Seminarberichte 32, Humboldt University, Berlin, 1980.
14. D. Stoffer and K. Nipp, *Invariant curves for variable step-size integrators*, BIT, 31 (1991), pp. 169–180 and (*Erratum*) BIT, 32 (1992), pp. 367–368.
15. A. M. Stuart and A. R. Humphries, *Dynamical Systems and Numerical Analysis*, Cambridge University Press, 1996.
16. A. M. Stuart and A. R. Humphries, *The essential stability of local error control for dynamical systems*, SIAM J. Num. Anal., 32 (1995), pp. 1940–1971.
17. R. Temam, *Infinite Dimensional Dynamical Systems in Mechanics and Physics*, Springer, New York, 1989.