THE EFFECTIVE STABILITY OF ADAPTIVE TIMESTEPPING ODE SOLVERS

HARBIR LAMBA [*]

Abstract. We consider the behaviour of certain adaptive timestepping methods, based upon embedded explicit Runge-Kutta pairs, when applied to dissipative ODEs. It has been observed numerically that the standard local error controls can impart desirable stability properties, but this has only been rigorously verified for very special, low-order, Runge-Kutta pairs.

The rooted-tree expansion of a certain quadratic form, central to the stability theory of Runge-Kutta methods, is derived. This, together with key assumptions on the sequence of accepted timesteps and the local error estimate, provides a general explanation for the observed stability of such algorithms on dissipative problems. Under these assumptions, which are expected to hold for 'typical' numerical trajectories, two different results are proved. Firstly, for a large class of embedded Runge-Kutta pairs of order $(1,2)$, controlled on an error-per-unit-step basis, all such numerical trajectories will eventually enter a particular bounded set. This occurs for sufficiently small tolerances independent of the initial conditions. Secondly, for pairs of arbitrary orders $(p-1,p)$, operating under either error-per-step or error-per-unit-step control, similar results are obtained when an additional structural assumption (that should be valid for many cases of interest) is imposed on the dissipative vector field. Numerical results support both the analysis and the assumptions made.

Key words. Error control, stability, numerical integration, ordinary differential equations

AMS subject classifications. 65L06, 65L20

1. Introduction.

We consider adaptive timestepping ODE solvers applied to initial value problems for an autonomous system of ODEs

$$\frac{du}{dt} = f(u), \quad u(0) = U \tag{1.1}$$

where $u(t) \in \mathbb{R}^m$. Furthermore, the Lipschitz continuous vector field f satisfies the following structural assumption

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

where the norm $\|\cdot\|$ is that induced by the inner product $\langle\cdot,\cdot\rangle$.

A bounded closed set $\mathcal{B}$ is a *bounded absorbing set* for (1.1) if $\forall U \in \mathbb{R}^m$, $\exists t^* = t^*(U)$ such that $u(t) \in \mathcal{B}\ \forall t \geq t^*$. If a bounded absorbing set exists then (1.1) is termed *dissipative*. Under the structural assumption ($\mathbf{D}$), (1.1) is dissipative as stated in the following theorem [16].

THEOREM 1.1. *Let $\overline{B}(v,r)$ be the closed ball with centre v, radius r using the norm $\|\cdot\|$. Then assumption* ($\mathbf{D}$) *implies the existence of bounded absorbing sets $\mathcal{B} = \overline{B}(0, \sqrt{(\alpha+\epsilon)/\beta})\ \forall\epsilon > 0$.*

The structural assumption ($\mathbf{D}$) has played an important role in nonlinear stability theory, where the aim is to find conditions under which numerical schemes, when regarded as discrete dynamical systems, preserve various qualitative asymptotic features of the original ODE (such as the existence of bounded absorbing sets). However, the vast majority of this body of work only applies to methods employing a fixed timestep, whereas most algorithms used in practice allow the timesteps to change from one step to the next. In the algorithms considered here, the timesteps are chosen so as to control an estimate of the local (one-step) error and this adaptive timestepping approach can result in extremely impressive efficiency gains.

[*] Department of Mathematical Sciences, George Mason University, MS 3F2, 4400 University Drive, Fairfax, VA 22030, USA (hlamba@gmu.edu).

Even though the standard error controls were not designed with stability in mind, it has been observed that such adaptive timestepping algorithms often have much better stability properties than their fixed-timestepping counterparts. This paper addresses the questions of when, and how, the stepsizes induced by the local error control will confer desirable stability properties upon the adaptive numerical method.

As mentioned above, there have been many investigations into the stability properties of Runge-Kutta methods with a fixed timestep, under various structural assumptions (e.g. [2, 1, 5, 9, 18]). We now provide a brief outline of the relevant results for dissipativity. Consider a general (implicit or explicit) s-stage Runge-Kutta scheme for (1.1) with timestep h

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

$$U_{n+1} = U_n + h\sum_{i=1}^{s} b_i f(\eta_i) \tag{1.3}$$

and define the vector $b = (b_1, \ldots, b_s)^T$ and matrices A and B by $A(i,j) = a_{ij}$ and $B = \mathrm{diag}(b)$.

Equations (1.2) and (1.3), after standard manipulations (see, for example, [18]), imply that

$$\|U_{n+1}\|^2 = \|U_n\|^2 + 2h\sum_{i=1}^{s} b_i\langle \eta_i, f(\eta_i)\rangle - h^2\sum_{i,j=1}^{s} m_{ij}\langle f(\eta_i), f(\eta_j)\rangle$$

where $m_{ij} = M(i,j)$ with $M = BA + A^T B - b^T b$. Under the structural assumption (**D**), with B positive semi-definite, and using the same norm $\|\cdot\|$ and inner product $\langle\cdot,\cdot\rangle$, we obtain

$$\|U_{n+1}\|^2 \leq \|U_n\|^2 + 2h\sum_{i=1}^{s} b_i(\alpha - \beta\|\eta_i\|^2) - h^2\sum_{i,j=1}^{s} m_{ij}\langle f(\eta_i), f(\eta_j)\rangle. \tag{1.4}$$

The Runge-Kutta method is termed *algebraically stable* if the matrices M and B are both positive semi-definite. The condition on M ensures that the quadratic form

$$h^2\sum_{i,j=1}^{s} m_{ij}\langle f(\eta_i), f(\eta_j)\rangle \tag{1.5}$$

is non-negative while the condition on B is used to show that the quantities $b_i(\alpha - \beta\|\eta_i\|^2)$ are negative outside a ball of sufficiently large radius in the norm $\|\cdot\|$. Together these imply the existence of bounded absorbing sets for the discrete dynamical system defined by the numerical scheme *for all* $h > 0$. Thus the algebraic stability of the numerical scheme ensures that the property of dissipativity is transferred to the numerical approximation. However, M cannot be positive semi-definite for explicit Runge-Kutta methods and so all algebraically stable methods are necessarily implicit. Indeed, explicit Runge-Kutta methods using a fixed timestep often have very poor stability properties.

We now return to our discussion of adaptive schemes. While the quadratic form (1.5) cannot be forced to be non-negative for a non-algebraically-stable method, we shall show that, under certain conditions, the constraints imposed upon the timestep sizes by the local error control will also effectively bound the *magnitude* of (1.5) (as opposed to its sign). Then, for a Runge-Kutta method with B positive definite, outside a ball of sufficiently large radius the single-summation term in (1.4) will be shown to dominate and the norm of the numerical solution decrease. This idea underlies the approach introduced in this paper.

The class of adaptive schemes that will be analyzed is now defined. We set $U_0 = U$ and iteratively generate U_{n+1} from U_n using a timestep h_n. The equations defining a general embedded explicit Runge-Kutta pair with s stages are

$$\eta_i = U_n + h_n \sum_{j=1}^{s} a_{ij} f(\eta_j), \quad i = 1, \ldots, s, \tag{1.6}$$

$$V_{n+1} = U_n + h_n \sum_{i=1}^{s} b_i f(\eta_i), \tag{1.7}$$

$$W_{n+1} = U_n + h_n \sum_{i=1}^{s} \overline{b}_i f(\eta_i). \tag{1.8}$$

Such a Runge-Kutta pair, with orders $p-1$ and p, will be referred to as a $(p-1, p)$ pair. We shall assume that the higher-order method is represented by the weights $b_1, \ldots, b_s$ and the lower-order method by $\overline{b}_1, \ldots, \overline{b}_s$. Thus $U_{n+1} = V_{n+1}$ when the higher-order method is used to advance the solution (extrapolation mode) and $U_{n+1} = W_{n+1}$ otherwise (non-extrapolation mode). To complement the definitions of A, B and b, let $\overline{b} = (\overline{b}_1, \ldots, \overline{b}_s)^T$ and $\overline{B} = \mathrm{diag}(\overline{b})$.

The *local error estimate* $E(U_n, h_n)$ is defined as the difference between the two approximations,

$$E(U_n, h_n) := W_{n+1} - V_{n+1}.$$

The user defines a tolerance τ, and the timesteps must satisfy the following standard local error control

$$\|E(U_n, h_n)\| \leq \sigma(\tau, U_n) h_n^{\rho} \tag{1.9}$$

where $\rho = 0$ for error-per-step control and $\rho = 1$ for error-per-unit-step control. The quantity $\sigma(\tau, U_n)$ is a quantity closely related to the tolerance τ, and indeed may simply be equal to τ. However we wish to allow for the possibility of absolute, relative and mixed error controls. There are various ways in which this can be done but for simplicity we shall require only that there exists some constant $C_1 > 0$ such that

$$\sigma(\tau, u) \leq C_1 \tau \|u\| \quad \forall u \in \mathbb{R}^m. \tag{1.10}$$

It should be noted that absolute or mixed error controls will need to be modified on some neighbourhood of the origin in order to satisfy (1.10). However we will be concerned exclusively with trajectories that lie entirely outside of (large) balls centred upon the origin and the choice of (1.10) will help to streamline the analysis. For simplicity, the norms used in (1.9) and (1.10) and throughout the rest of the paper are the same as in (**D**).

We thus have four possible modes of operation depending upon the choice of solution-advancing method and type of error control. EPS and EPUS will denote error-per-step and error-per-unit step modes respectively in non-extrapolation mode while XEPS and XEPUS are their extrapolation counterparts. We also assume the existence of a maximum timestep $h_{\max}$, independent of τ, which is a very common feature of adaptive algorithms. Throughout, we assume that the vector field f is sufficiently smooth on $\mathbb{R}^m$. These smoothness requirements are determined only by the order of the Runge-Kutta methods used to form the local error estimate. Note that no further details of the algorithm need to be specified, in particular the way in which candidate timesteps are generated. All that is required is that the error control (1.9) is satisfied at every timestep.

While no explicit Runge-Kutta method can be algebraically stable, it has been observed [6, 18] that adaptive timestepping methods based upon explicit schemes do, for certain combinations of dissipative test problems and mode of operation, seem to have some desirable stability properties (see also [13] for a discussion of stability with regard to the existence of spurious fixed points). In particular, the numerical schemes appear to be dissipative. Of course, no amount of numerical testing can prove the existence of a bounded absorbing set for all initial data but the results do suggest that, with an extremely high degree of certainty, numerical trajectories enter and then remain within an 'absorbing set' close to $\overline{B}(0, \sqrt{\alpha/\beta})$.

There have been previous analyses of the behaviour of adaptive methods on dissipative ODEs that have attempted to explain this phenomenon. In [17], it was proved that very special, embedded explicit Runge-Kutta pairs generate a solution that, at each step, is a small perturbation of the solution generated by using a corresponding (implicit) algebraically stable method. In this way, the stability characteristics of this related scheme are transferred to the explicit pair. Such pairs were termed *essentially algebraically stable* and an order barrier for $(p-1, p)$ pairs, namely that $p \leq 5$, was proved. For these pairs, applied to ODEs satisfying $(\mathbf{D})$, under no additional assumptions and with an absolute error control, two different results were proved. The first, which is a discrete analogue of Theorem 1.1, stated that when such a pair is used in EPUS or XEPUS modes the numerical scheme has a bounded absorbing set for all sufficiently small tolerances τ, *independent of the initial data*. The second result, which is significantly weaker, states that for the same pairs operating in EPS or XEPS modes, each numerical trajectory will again eventually enter a particular bounded absorbing set but now the required tolerance does depend upon the initial data.

The independence of τ with respect to initial conditions is desirable, not just from a computational point of view, but also theoretically since it allows us to regard the numerical method, for a fixed sufficiently small tolerance, as a dynamical system with similar asymptotic behaviour to the underlying ODE for all initial conditions. However, the set of essentially algebraically stable pairs forms a very small subset of pairs currently employed and are necessarily of low-order.

A second analysis [8] took a different approach. There it was assumed, for a general adaptive method under EPUS control, that the *actual* one-step truncation errors $T(U_n, h_n)$ (rather than the one-step error estimates) were correctly controlled at every step, in particular that

$$T(U_n, h_n) \leq K(U)\tau h_n$$

occurred at every timestep, for some constant $K(U)$. Using this assumption that the error control works correctly, positive stability results were proved for general adaptive schemes but only in the much weaker sense that the required tolerance depended upon the initial data.

Any stability properties introduced to an explicit Runge-Kutta method via a local error control are due to the size of the accepted steps. However neither of the analyses described above explicitly consider the actual timestep sequences generated by the method (and they also only considered the case of absolute error control). In order to obtain tighter and/or more general results it is therefore natural to consider closely the timestep sequence itself, and this forms another motivation for our analysis.

The paper is organized as follows. In section 2, for a Runge-Kutta method of order r, an expansion of the quadratic form (1.5) is derived and the leading order term is proved to be at least $\mathcal{O}(h^{r+1})$. In Section 3 we then use this expansion, together with the corresponding expansion of the local error estimate $E(U_n, h_n)$ at each timestep, to state and justify our two key assumptions on the numerical trajectory. The first assumption takes the form of an upper bound on the timesteps used at each point in the phase space. The second assumption is that controlling the local error estimate also bounds the magnitude of the quadratic form (1.5) at each timestep. It must be emphasized that the justification for these assumptions is that they are expected to hold for every timestep along 'typical' numerical trajectories but it seems likely that for most vector fields satisfying (**D**) there will be 'atypical' numerical trajectories where, at one or more timesteps, they do not hold. Even when these extreme events occur, the fact that the assumptions hold for most of the timesteps should help to preserve the qualitative asymptotic features of the numerical trajectory.

We do not attempt to quantify the ways in which our assumptions can be violated and this is unsatisfactory from a rigorous mathematical viewpoint. However, using these assumptions, we shall gain valuable insights into how these algorithms behave on most simulations. The studies [15, 11, 12] have shown that even when considering the convergence to the exact solution, as $\tau \to 0$, of adaptive timestepping algorithms over finite time intervals and compact sets of initial data — arguably a more fundamental property — there are mechanisms that can give rise to the breakdown of convergence. These arise because of the possibility that the leading term of the error estimate may vanish at some point along the exact trajectory, resulting in a local increase in the size of the accepted timesteps and potential loss of convergence (or, more likely, a reduction in the rate of convergence). However, at least for generic vector fields, the probability of convergence failure is extremely small. These previous studies have therefore already demonstrated that a 'worst-case analysis' is not necessarily appropriate in the context of ODE solvers, since the very small probability of failure to converge is outweighed by the superior efficiency of adaptive algorithms. In fact the situation facing us here, where we are concerned with stability properties, is very much better than that for convergence properties. This is because convergence can be destroyed by a single 'bad' timestep whereas asymptotic qualitative properties are very likely to be robust in the presence of such extreme events. Nevertheless, it is hoped that the analysis presented here will stimulate further work into justifying or weakening the assumptions made.

In Section 4, we present the main results. Firstly, for embedded explicit Runge-Kutta pairs of order (1,2), operating in EPUS or XEPUS modes with B positive-definite, any numerical trajectory satisfying the assumptions of Section 3 will eventually enter a particular bounded set, for all sufficiently small τ independent of U. Secondly, motivated by the analysis, we introduce an additional structural assumption on the vector field f:

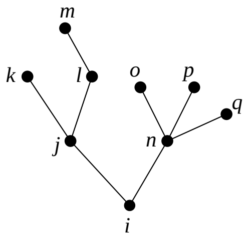
$(\mathbf{D'})\quad \exists\gamma > 0, R > 0 : \ \langle f(u), u\rangle \ \leq \ -\gamma\|f(u)\|\|u\| \quad \forall\|u\| \geq R.$

Intuitively this structural assumption implies that, for sufficiently large $\|u\|$, the vector field points inwards everywhere at some definite minimum non-zero angle and holds for many ODEs of interest. In particular, vector fields satisfying $(\mathbf{D})$ or $(\mathbf{D'})$ (or both) are not necessarily globally Lipschitz. Assuming that $(\mathbf{D})$, $(\mathbf{D'})$ and the assumptions on the numerical trajectory hold, then for sufficiently small τ independent of initial data, for arbitrary embedded $(p-1,p)$ pairs with B positive-definite in any mode of operation, a similar result is proved. Finally in Section 5, we present numerical results that support both the assumptions made in Section 3 and the results of Section 4.

2. Order Conditions and the matrix M.

The Taylor series expansions in powers of h of both the exact solution to (1.1) and the one-step Runge-Kutta approximation, over some time interval $[s, s+h]$, consist of multiples of expressions involving f and its higher derivatives which rapidly become very complicated. We therefore first recall some necessary definitions and terminology from the rooted tree description of Taylor series expansions. This theory was developed by Butcher and the reader is directed towards [3, 4] for full details of all the notation, definitions and results up to and including (2.3).

A rooted tree is an unlabeled connected graph containing no cycles and with one node identified as the 'root'. Each rooted tree with precisely n nodes corresponds uniquely to one term (of many) appearing at order h^n in the Taylor series. Each term is a multiple of an elementary differential of order n and this correspondence is achieved as follows. Let $f^i_{j_1,j_2,\ldots,j_r}$ denote the r^{th} partial derivative of the i^{th} component of f with respect to the components $j_1, j_2, \ldots, j_r$. Now attach the label i to the root of the tree and labels $j, k, l \ldots$ to the other nodes. Then for each node, write down f with a superscript equal to the label of that node and subscripts given by the other nodes that are directly connected to it on the side away from the root node. For example, the rooted tree



corresponds to the product $f^i_{jn} f^j_{kl} f^k f^l_m f^m f^n_{opq} f^o f^p f^q$ (using the summation convention over repeated indices) which is the i^{th} component of one particular elementary differential of order 9. Repeating the above process for each value of the index i provides each component of the elementary differential corresponding to the above (unlabeled) rooted tree. The elementary differential corresponding to a particular tree t will be denoted by the function $F(t) : \mathbb{R}^m \to \mathbb{R}^m$.

The set of all rooted trees, denoted by $\mathcal{T}$, is defined recursively as follows. The rooted tree consisting of a single node is defined as τ and any rooted tree t can be built up by joining trees $t_1, \ldots, t_k$ to a new root. The rooted tree t is then written as $t = [t_1, \ldots, t_k]$ (note that the order is unimportant) and m repetitions of a tree t_i are denoted by t_i^m.

We now recall some important functions that can be defined on the set $\mathcal{T}$. The function $\rho(t)$ is simply the number of nodes in t. The next three functions $\gamma(t), \sigma(t)$ and $\alpha(t)$ have important combinatorial interpretations ([3][Section 144]) and also allow for an elegant statement of Taylor series expansions. However the following recursive definitions, also due to Butcher, are the more relevant for our purposes:

$$\gamma(\tau) = 1, \quad = \gamma([t_1, \ldots, t_k]) = \rho([t_1, \ldots, t_k]) \prod_{j=1}^{k} \gamma(t_j) \tag{2.1}$$

and

$$\sigma(\tau) = 1, \quad \sigma([t_1^{n_1}, \ldots, t_k^{n_k}]) = n_1!n_2! \ldots n_k! \prod_{j=1}^{k} \sigma(t_j)^{n_j}$$

where the trees $t_1, \ldots, t_k$ are all distinct. Finally the function $\alpha(t)$ is defined by

$$\alpha(t) = \frac{\rho(t)!}{\gamma(t)\sigma(t)}.$$

In [3] it is then proved that the Taylor series for the exact solution of (1.1) at time $s + h$ is

$$u(s+h) = u(s) + \sum_{t \in \mathcal{T}} \frac{\alpha(t)}{\rho(t)!} h^{\rho(t)} F(t)(u(s)). \tag{2.2}$$

The one-step numerical approximation, $\tilde{u}(s + h)$ can also be expressed in terms of elementary differentials. For a given rooted tree t and Runge-Kutta method (determined by (1.2) and (1.3)) we define the elementary weight $\Phi(t)$ as follows. Label the root of the tree i and attach labels to the other vertices. For every edge connecting vertices u and v, write down a factor a_{uv} where u is the vertex closer to the root. Insert a final factor b_i, corresponding to the root, form the product of the above factors and then sum every index over all of the stages. Thus the elementary weight corresponding to the tree drawn above is

$$\Phi(t) = \sum_{i,j,k,l,m,n,o,p,q=1}^{s} b_i a_{ij} a_{jk} a_{jl} a_{lm} a_{in} a_{no} a_{np} a_{nq}.$$

Now the numerical approximation can be expanded as

$$\tilde{u}(s+h) = u(s) + \sum_{t \in \mathcal{T}} \frac{\gamma(t)\alpha(t)\Phi(t)}{\rho(t)!} h^{\rho(t)} F(t)(u(s)). \tag{2.3}$$

By comparing (2.2) and (2.3), Butcher proved that a necessary and sufficient condition for a Runge-Kutta method to be order precisely p is that $\Phi(t) = 1/\gamma(t)$ for all rooted trees t with $\rho(t) \leq p$, but not for at least one tree t with $\rho(t) = p + 1$.

The above definitions and results now enable us to prove a new expansion for the quadratic form (1.5).

Lemma 2.1. *Let the stages $\eta_1, \ldots, \eta_s$ be generated by a Runge-Kutta method of order r using a timestep h from a solution value u. Then there exists an integer $q \geq r + 1$ and scalar-valued functions $G_1(u)$ and $G_2(u, h)$ such that $G_2(u, 0) = 0$ and*

$$h^2 \sum_{i,j=1}^{s} m_{ij} \langle f(\eta_i), f(\eta_j) \rangle \;=\; h^q (G_1(u) + hG_2(u, h)). \tag{2.4}$$

Proof. Consider an arbitrary rooted tree t. Comparison of the Taylor series expansion of the numerical solution (2.3) with (1.3) shows that each of the terms $hf(\eta_i)$ can be expanded as

$$hf(\eta_i) = \sum_{t\in\mathcal{T}} \frac{\gamma(t)\alpha(t)\Phi_i(t)}{\rho(t)!} h^{\rho(t)} F(t)(u)$$

where each term $\Phi_i(t)$ is derived from $\Phi(t)$ by deleting both the factor b_i and the summation over the index i. Note that each $\Phi_i(t)$ has precisely $\rho(t)-1$ factors. Let us now fix trees T_1 and T_2 (not necessarily distinct) and consider the coefficient of $\langle F(T_1)(u), F(T_2)(u)\rangle$ in the expansion of (1.5). Using the definition of the matrix M, this is

$$\begin{aligned}(2-\mathcal{I}_{T_1=T_2})\, h^{\rho(T_1)+\rho(T_2)} \frac{\alpha(T_1)\alpha(T_2)\gamma(T_1)\gamma(T_2)}{\rho(T_1)!\rho(T_2)!} \sum_{i,j=1}^{s} [\Phi_i(T_1)\Phi_j(T_2)b_i a_{ij} + \Phi_i(T_1)\Phi_j(T_2)b_j a_{ji} \\ -\Phi_i(T_1)\Phi_j(T_2)b_i b_j] \qquad (2.5)\end{aligned}$$

where $\mathcal{I}_{T_1=T_2} = 1$ if $T_1 = T_2$ and 0 otherwise.

We now introduce some new notation. Given two trees $T_1 = [s_1, \dots s_m]$ and $T_2 = [t_1, \dots t_n]$ (where $m = 0$ or $n = 0$ correspond to $T_1 = \tau$ or $T_2 = \tau$ respectively) we define the tree $T_1 \nearrow T_2 := [s_1, \dots, s_m, T_2]$, which is the tree with $\rho(T_1) + \rho(T_2)$ nodes obtained by adding a single edge between the roots of T_1 and T_2 and keeping the root of T_1 as the root of the new tree. Similarly, $T_2 \nearrow T_1 := [t_1, \dots, t_n, T_1]$. Thus the first term in the summand of (2.5), after summation, corresponds to $\Phi(T_1 \nearrow T_2)$, the second term corresponds to $\Phi(T_2 \nearrow T_1)$ and the third term to $\Phi(T_1)\Phi(T_2)$.

Let us now assume that $\rho(T_1) + \rho(T_2) \le r$. Then this coefficient vanishes if

$$\Phi(T_1 \nearrow T_2) + \Phi(T_2 \nearrow T_1) = \Phi(T_1)\Phi(T_2). \qquad (2.6)$$

But since the Runge-Kutta method is of order r this is equivalent to the condition that

$$\frac{1}{\gamma(T_1 \nearrow T_2)} + \frac{1}{\gamma(T_2 \nearrow T_1)} = \frac{1}{\gamma(T_1)\gamma(T_2)}. \qquad (2.7)$$

This is easily proved via (2.1), the recursive definition of γ. For let us suppose first that $T_1 \ne \tau \ne T_2$. Then

$$\begin{aligned}\gamma(T_1) &= \rho(T_1)\gamma(s_1)\dots\gamma(s_m)\\ \gamma(T_2) &= \rho(T_2)\gamma(t_1)\dots\gamma(t_n)\\ \gamma(T_1 \nearrow T_2) &= [\rho(T_1)+\rho(T_2)]\rho(T_2)\gamma(s_1)\dots\gamma(s_m)\gamma(t_1)\dots\gamma(t_n)\\ \gamma(T_2 \nearrow T_1) &= [\rho(T_1)+\rho(T_2)]\rho(T_1)\gamma(t_1)\dots\gamma(t_n)\gamma(s_1)\dots\gamma(s_m)\end{aligned}$$

and (2.7) easily follows. The remaining cases when either $T_1 = \tau$ or $T_2 = \tau$ are also easily verified.

Thus

$$h^2 \sum_{i,j=1}^{s} m_{ij}\langle f(\eta_i), f(\eta_j)\rangle \;=\; h^q(G_1(u) + hG_2(u,h))$$

for some $q \geq r + 1$, with the function $G_1(u)$ being the sum of inner products of elementary differentials where the nodes in the corresponding rooted trees sum to precisely q, and the function $G_2(u, h)$ comprising the higher-order terms. $\square$

The possibility of $q > r + 1$ in the statement of Lemma 2.1 arises because, for a pair of trees T_1, T_2 with $\rho(T_1) + \rho(T_2) = n > r$, the Runge-Kutta method being of order n is a sufficient, but not a necessary, condition for (2.6) to be satisfied. This is in fact the case for the improved Euler (Heun) method where (2.6) is also satisfied for the unique pair of rooted trees whose nodes sum to 3. Thus $q = 4$ even though $r = 2$.

3. Assumptions. We now turn to the assumptions necessary for the analysis and results of Section 4. Once again, the purpose of these results is to provide an explanation of the observed behaviour of explicit Runge-Kutta pairs for 'typical' numerical trajectories of 'typical' vector fields. For the vast majority of adaptive schemes (i.e. apart from ones utilizing essentially algebraically stable pairs) it would appear that no results are possible without such assumptions. As mentioned in the introduction, similar problems arise when proving convergence results for adaptive algorithms, even for finite-time initial value problems on compact domains. This is because any method based upon a local error estimate can behave badly, even if only for a single timestep, by a sufficiently unfortunate (or devious) combination of vector field, solution value and candidate timestep. However, both of the assumptions stated and justified below are numerically verified for every single timestep used to advance the solutions in the numerical experiments of Section 5.

Assumption 1. *If the local error estimate is derived from a $(p-1, p)$ explicit Runge-Kutta pair then, for all sufficiently small $\tau > 0$, there exists a constant $K_1 > 0$, independent of U, such that for each accepted timestep h_n*

$$h_n^{p-\rho} \leq K_1 \frac{\sigma(\tau, U_n)}{\|f(U_n)\|}. \tag{3.1}$$

The intuitive reason for this assumption can be seen by following [15, 11] and expanding the local error estimate as

$$\begin{aligned} E(U_n, h_n) &= h_n^p \left(B_1(U_n) + h_n B_2(U_n, h_n)\right) & (3.2)\\ &= h_n^p \|f(U_n)\| \left(\tilde{B}_1(U_n) + h_n \tilde{B}_2(U_n, h_n)\right). & (3.3)\end{aligned}$$

In (3.3) the expansion has simply been rescaled by a factor of $\|f(U_n)\|$. Now let us suppose that the function $\|B_1(u)\|$ is bounded away from zero along the numerical trajectory. Then if the error control is working correctly (for sufficiently small τ), and the accepted timesteps are controlled by the (non-vanishing) leading-order term of the expansion (3.2), we see that (3.1) immediately follows.

In [15, 11, 10], rigorous proofs of the upper bound (3.1) on the sequence of accepted timesteps are obtained via induction arguments for sufficiently small τ, but only for numerical trajectories lying inside a predefined compact set on which $B_1(u)$ is bounded away from zero. By restricting ourselves to ODEs satisfying (**D**), we now argue that (3.1) will only fail to hold in exceptional cases, for any initial data.

Note first that under assumption (**D**), $f(u) \neq 0$ outside the ball $\overline{B}(0, \sqrt{\alpha/\beta})$. Thus, outside this ball, the leading order term of the error estimate can only vanish if $\tilde{B}_1$ does. But $\tilde{B}_1 : \mathbb{R}^m \to \mathbb{R}^m$ and so, for typical vector fields will only vanish at isolated points in the phase space. In order to obtain (3.1) from (3.3), we assume the existence of constants $K, K' > 0$ (independent of τ and U_n) for τ sufficiently small such that at each step of the numerical trajectory

$$\begin{aligned}\sigma(U_n,\tau) \geq \|E(U_n,h_n)\|/h_n^\rho &\geq K h_n^{p-\rho} \max\left(B_1(u), h_n B_2(U_n,h_n)\right)\\ &\geq K h_n^{p-\rho}\|B_1(u)\|\\ &\geq KK'\|f(U_n)\|h_n^{p-\rho}\end{aligned}$$

leading immediately to (3.1) with $K_1 = 1/KK'$. The existence of the constant $K > 0$ is equivalent to assuming that at each step no catastrophic cancellation occurs between B_1 and $h_n B_2$. In order to justify the existence of the constant $K' > 0$ we need to demonstrate that, under assumption (**D**), $\|\tilde{B}_1(u)\|$ does not tend to 0 as $\|u\| \to \infty$ in any direction. We achieve this by showing that at least one of the rescaled elementary differentials comprising $\tilde{B}_1(u)$ cannot vanish as $\|u\| \to \infty$.

Let us suppose that the lower-order method of the pair does not increase its order on linear constant-coefficient problems.[1] Then $B_1(u)$ must contain an elementary differential of the form $cf'(u)^{p-1}f(u)$ with coefficient $c \neq 0$ (the rooted trees corresponding to such elementary differentials are often referred to as 'tall trees' and contain no branches). Under the structural assumption (**D**), $\|f(u)\|$ must increase at least as fast as $\mathcal{O}(\|u\|)$ for sufficiently large $\|u\|$ in any given direction. Thus $\|f'(u)\|$ and $\|cf'(u)^{p-1}f(u)\|/\|f(u)\|$ cannot tend to 0 as $\|u\| \to \infty$ (although for pathological vector fields, $f'(u)$ may equal zero on arbitrarily large compact sets in the phase space). We now appeal once again to the principle that catastrophic cancellation (this time between the weighted and rescaled elementary differentials comprising $\tilde{B}_1(u)$) occurs negligibly often, giving $\tilde{B}_1(u) \not\to 0$ as $\|u\| \to \infty$ in any direction. This completes our justification of (3.1).

It should be noted that for a linear constant-coefficient ODE satisfying (**D**), $\|\tilde{B}_1(u)\|$ is a non-zero constant but for certain nonlinear problems we can expect $\|\tilde{B}_1(u)\|$ to grow as $\|u\|$ grows. Thus for particular classes of nonlinear problem it may be possible to strengthen the upper bound on the timestep sequence in Assumption 1 considerably (this is confirmed by numerical computations but we shall not explore this point further).

The second assumption states that the error control, which is of course designed to bound the local error estimate, also provides a bound on the magnitude of the quadratic form (1.5) for typical timesteps.

Assumption 2. *For all sufficiently small τ there exists a constant $K_2 > 0$, independent of U, such that at each timestep along the numerical trajectory*

$$\left| h_n^2 \sum_{i,j=1}^{s} m_{ij}\langle f(\eta_i), f(\eta_j)\rangle \right| \;\leq\; K_2\sigma(\tau,U_n)h_n^{1+\rho}\|f(U_n)\| \tag{3.4}$$

where the matrix M is the stability matrix for the higher-order method of the (p-1,p) Runge-Kutta pair.

[1] If this mild condition is violated then the method behaves substantially differently for such problems. Indeed if the $(p-1,p)$ pair has precisely p stages then the local error estimate $E(u,h) \equiv 0$ and the error control fails completely. The reader is directed to [11] for further discussion of this point. We simply note that most embedded pairs used in practice satisfy this criterion.

We start our justification of Assumption 2 by defining

$$\hat{E}(U_n, h_n) := \left| h_n \sum_{j=1}^{s} \langle E(U_n, h_n), f(\eta_j) \rangle \right| \tag{3.5}$$

$$= \left| h_n^2 \sum_{i,j=1}^{s} (b_i - \overline{b}_i) \langle f(\eta_i), f(\eta_j) \rangle \right|. \tag{3.6}$$

The enforcement of the local error control (1.9) now allows us to bound $\hat{E}(U_n, h_n)$ from above, since

$$\begin{aligned} \hat{E}(U_n, h_n) &\leq h_n \sum_{j=1}^{s} \|E\| \, \|f(\eta_j)\| \\ &\leq s\sigma(\tau, U_n) h_n^{1+\rho} \max_{j=1,\ldots,s} \|f(\eta_j)\|. \end{aligned} \tag{3.7}$$

We now compare the expansion (3.6) for $\hat{E}$ with that of the (absolute value of the) quadratic form for the higher-order method (1.5). From the proof of Lemma 2.1, (1.5) is a linear combination of inner products of elementary differentials. Reverting to the rooted tree description of elementary differentials, the only inner products $\langle F(T_1)(u), F(T_2)(u) \rangle$ appearing in the expansion are those for which $\rho(T_1) + \rho(T_2) \geq p + 1$, and their coefficients are of order $h_n^{\rho(T_1)+\rho(T_2)}$. The corresponding expansion for $\hat{E}$ contains those inner products $\langle F(T_1)(u), F(T_2)(u) \rangle$ for which $\max(\rho(T_1), \rho(T_2)) \geq p$, once again with coefficients of order $h_n^{\rho(T_1)+\rho(T_2)}$.

Note that the expansion of (1.5) therefore contains a (finite) number of additional inner products not appearing in that of $\hat{E}$. However these inner products are closely related to others that are common to both expansions and so our assumption reduces to the observation that control of the quantity $\hat{E}$ should effectively control (1.5) to within some constant. Assumption 2 now follows immediately from (3.7) by assuming that $\max_{j=1,\ldots,s} \|f(\eta_j)\|$ is always close to $\|f(U_n)\|$, which of course should be the case, barring any catastrophic cancellations in the formation of the error estimate.

In principle, Assumptions 1 and 2 could be weakened considerably by, for example, only requiring that (3.1) and (3.4) hold, for a given K_1 and K_2, on a sufficiently large proportion of the numerical timesteps. However an analysis resting on such assumptions would become far more difficult, without generating any new insights into the mechanisms leading to effective numerical stability.

4. Results.

Using Assumptions 1 and 2 we are now ready to prove the main results. We start by considering the case of embedded explicit Runge-Kutta pairs with order (1,2) in either EPUS or XEPUS modes.

THEOREM 4.1. *Consider an embedded explicit Runge-Kutta pair of order (1,2), under either EPUS or XEPUS control, where the higher-order method has positive weights. If Assumptions 1 and 2 are satisfied and the ODE (1.1) satisfies* (**D**) *then* $\exists \tau^* > 0$ *such that* $\forall \tau \leq \tau^*$ *the numerical trajectory eventually enters a compact set independent of* τ, U.

Proof. The tolerance τ is chosen sufficiently small such that Assumptions 1 and 2 are satisfied. We first consider advancing the numerical solution using the higher-order method (extrapolation mode). From (**D**) we have

$$\|V_{n+1}\|^2 = \|U_n\|^2 + 2h_n \sum_{i=1}^{s} b_i \langle \eta_i, f(\eta_i)\rangle - h_n^2 \sum_{i,j=1}^{s} m_{ij} \langle f(\eta_i), f(\eta_j)\rangle \tag{4.1}$$

$$\leq \|U_n\|^2 + 2h_n \sum_{i=1}^{s} b_i (\alpha - \beta\|\eta_i\|^2) - h_n^2 \sum_{i,j=1}^{s} m_{ij} \langle f(\eta_i), f(\eta_j)\rangle. \tag{4.2}$$

We now proceed by bounding the absolute value of the last term and, for sufficiently small τ, absorbing it into the previous one. From Assumptions 1 and 2 and (1.10),

$$\left| h_n^2 \sum_{i,j=1}^{s} m_{ij} \langle f(\eta_i), f(\eta_j)\rangle \right| \leq K_2 \sigma(\tau, U_n) h_n^2 \|f(U_n)\| \tag{4.3}$$

$$\leq K_1 K_2 h_n \sigma(\tau, U_n)^2 \tag{4.4}$$

$$\leq C_1^2 K_1 K_2 h_n \tau^2 \|U_n\|^2. \tag{4.5}$$

Now fix $0 < \tilde{\beta} < \beta$ and substitute (4.5) into the last term of (4.2) with

$$\tau < \sqrt{\frac{2b_1(\beta - \tilde{\beta})}{C_1^2 K_1 K_2}}.$$

Noting that $\eta_1 = U_n$ for explicit Runge-Kutta methods, we obtain

$$\|V_{n+1}\|^2 \leq \|U_n\|^2 + 2h_n \sum_{i=1}^{s} b_i (\alpha - \tilde{\beta}\|\eta_i\|^2) \tag{4.6}$$

The proof now proceeds exactly as in [17][Lemma 4.2 and Theorem DC1] by showing that the norm of the numerical solution strictly decreases until it enters, for any $\epsilon > 0$, the compact set $S = \overline{B}(0, \sqrt{\frac{\alpha+\epsilon}{\tilde{\beta}} + h_{\max} K})$ where

$$K = \max_{\|\eta_i\| \leq \gamma_i} \left(2 \sum_{i,j=1}^{s} b_i e_{ij} \langle \eta_i, f(\eta_i)\rangle + h_{\max} \sum_{i=1}^{s} b_i \left\| \sum_{j=1}^{s} e_{ij} f(\eta_j) \right\|^2 \right) \tag{4.7}$$

and

$$e_{ij} := b_j - a_{ij}, \qquad \gamma_i^2 := \frac{\alpha}{\tilde{\beta} b_i}.$$

We now consider the non-extrapolation case. From the local error control (1.9),

$$\begin{aligned} \|W_{n+1}\|^2 - \|V_{n+1}\|^2 &= \langle W_{n+1} + V_{n+1}, W_{n+1} - V_{n+1}\rangle \\ &\leq \|W_{n+1} + V_{n+1}\| \, \|W_{n+1} - V_{n+1}\| \\ &\leq \|W_{n+1} + V_{n+1}\| \sigma(\tau, U_n) h_n \\ &\leq 2\|V_{n+1}\| \sigma(\tau, U_n) h_n + \sigma^2(\tau, U_n) h_n^2. \end{aligned}$$

While the numerical trajectory is outside the compact set $\overline{B}(0, \sqrt{\frac{\alpha+\epsilon}{\tilde{\beta}} + h_{\max} K})$, we have already proved that, for sufficiently small τ, $\|V_{n+1}\| \leq \|U_n\|$ implying

$$\begin{aligned} \|W_{n+1}\|^2 - \|V_{n+1}\|^2 &\leq 2\|U_n\| \sigma(\tau, U_n) h_n + \sigma^2(\tau, U_n) h_n^2 \\ &\leq 2C_1 \tau \|U_n\|^2 h_n + C_1^2 \tau^2 \|U_n\|^2 h_n^2. \end{aligned} \tag{4.8}$$

Thus, the bound on $\|V_{n+1}\|^2$ from (4.6) may be invoked in (4.8) to give

$$\|W_{n+1}\|^2 \leq \|U_n\|^2 + 2h_n \sum_{i=1}^{s} b_i(\alpha - \tilde{\beta}\|\eta_i\|^2) + 2C_1\tau\|U_n\|^2 h_n + C_1^2\tau^2\|U_n\|^2 h_n h_{\max}. \quad (4.9)$$

The argument now concludes in a very similar fashion to the extrapolation case. After reducing the tolerance τ further if necessary, the last two terms of (4.9) can be absorbed into the preceding term by reducing $\tilde{\beta}$ once more, and then redefining (increasing) K to $\tilde{K}$ via (4.7). This then proves that the numerical solution enters a set $\overline{B}(0, \sqrt{\frac{\alpha+\epsilon}{\beta} + h_{\max}\tilde{K}})$ as required. □

Theorem 4.1 only states that numerical trajectories will enter a particular compact set, which is not necessarily close to the set $\overline{B}(0, \sqrt{\alpha/\beta})$. However, once a numerical trajectory has entered this set, finite-time convergence results, such as those in contained in [15, 12], can be applied to prove that typical numerical trajectories (possibly after a further reduction in τ) will enter and remain within $\mathcal{O}(\tau)$ of the absorbing set $B(0, \sqrt{\alpha/\beta})$ of the ODE (1.1). Furthermore in [10], and under additional assumptions, the existence of a (local) numerical attractor that is upper-semicontinuous to the global attractor of (1.1), can be proved.

We now prove a more general result, applicable to embedded Runge-Kutta pairs of any order and under any mode of operation. Note also that Assumption 1 is no longer required.

Theorem 4.2. *Consider an adaptive embedded Runge-Kutta pair of any order $(p-1, p)$, operating in EPS, XEPS, EPUS or XEPUS modes, where the higher-order method has positive weights. If Assumption 2 holds and the ODE (1.1) satisfies both* **(D)** *and* $(\mathbf{D}')$ *then $\exists\tau^* > 0$ such that $\forall\tau \leq \tau^*$ the numerical trajectory eventually enters a compact set independent of τ, U.*

Proof. Again we consider the extrapolation case first. From Assumption 2 and (1.10),

$$\left| h_n^2 \sum_{i,j=1}^{s} m_{ij}\langle f(\eta_i), f(\eta_j)\rangle \right| \leq K_2 C_1\|U_n\|\ \|f(U_n)\|\tau h_n^{1+\rho}$$

which upon substituting into (4.1) gives

$$\begin{aligned}\|V_{n+1}\|^2 &\leq \|U_n\|^2 + 2h_n \sum_{i=1}^{s} b_i\langle\eta_i, f(\eta_i)\rangle + K_2C_1\|U_n\|\ \|f(U_n)\|\tau h_n^{1+\rho} \\ &= \|U_n\|^2 + h_n b_1\langle\eta_1, f(\eta_1)\rangle + 2h_n \sum_{i=1}^{s} \hat{b}_i\langle\eta_i, f(\eta_i)\rangle \\ &\quad + K_2C_1\|U_n\|\ \|f(U_n)\|\tau h_n^{1+\rho}\end{aligned}$$

where $\hat{b}_1 = \frac{1}{2}b_1$ and $\hat{b}_i = b_i,\ i = 2, \ldots, s$. We now assume that $\|U_n\| > R$ and using $(\mathbf{D}')$ obtain

$$\begin{aligned}\|V_{n+1}\|^2 \leq \|U_n\|^2 &- h_n b_1\gamma\|U_n\|\ \|f(U_n)\| + 2h_n \sum_{i=1}^{s} \hat{b}_i\langle\eta_i, f(\eta_i)\rangle \\ &+ K_2C_1\|U_n\|\ \|f(U_n)\|\tau h_n^{1+\rho}\end{aligned}$$

Choosing

$$\tau \le \frac{b_1\gamma}{K_2 C_1 \max(1, h_{\max})^\rho}$$

and applying **(D)** we have

$$\|V_{n+1}\|^2 \le \|U_n\|^2 + 2h_n \sum_{i=1}^{s} \hat{b}_i(\alpha - \beta\|\eta_i\|^2).$$

Next we define $\tilde{\beta} = \beta/2$ to give

$$\begin{aligned} \|V_{n+1}\|^2 &\le \|U_n\|^2 + 2h_n \sum_{i=1}^{s} b_i(\alpha - \beta\|\eta_i\|^2) - h_n b_1(\alpha - \beta\|\eta_1\|^2) \\ &\le \|U_n\|^2 + 2h_n \sum_{i=1}^{s} b_i(\alpha - \tilde{\beta}\|\eta_i\|^2) - h_n b_1 \alpha \\ &\le \|U_n\|^2 + 2h_n \sum_{i=1}^{s} b_i(\alpha - \tilde{\beta}\|\eta_i\|^2). \end{aligned} \tag{4.10}$$

which is identical to (4.6). Thus the proof continues in a very similar manner to that of Theorem 4.1, via the construction of a compact set S outside of which

$$2h_n \sum_{i=1}^{s} b_i(\alpha - \tilde{\beta}\|\eta_i\|^2) \le 0.$$

While the numerical trajectory is outside the set $S \cup \overline{B}(0, R)$ its norm strictly decreases until the set is eventually entered.

For the non-extrapolation case, from (4.10) and (4.8) we once more obtain (4.9). Again, reducing $\tau, \tilde{\beta}$ and increasing K if necessary, the numerical trajectory eventually enters some compact set $S' \cup \overline{B}(0, R)$. □

5. Numerical Results. Some numerical examples are now presented to support Assumptions 1 and 2 and Theorems 4.1 and 4.2. We shall consider various embedded Runge-Kutta pairs in different operational modes. The algorithms used are all modifications of the ode23 routine supplied with MATLAB Version 4.2. This code was used (rather than, for example, the more sophisticated ODE routines in later MATLAB versions) because the timestep mechanism is particularly straightforward, containing only elements common to all such adaptive algorithms. Note that none of the previous analysis relies upon a detailed description of the timestep selection mechanism, merely that the local error control is satisfied.

Two examples of vector fields that satisfy both **(D)** and (**D′**), are the scalar ODE

$$u_t = -u|u| \tag{5.1}$$

and the linear constant-coefficient problem

$$\begin{aligned} x_t &= -y - \epsilon x \\ y_t &= x - \epsilon y \end{aligned} \tag{5.2}$$

for $\epsilon > 0$. Note that for scalar ODEs **(D)** implies (**D′**).

A vector field that satisfies **(D)** but not (**D**$'$) is, for $\epsilon > 0$,

$$\begin{aligned} x_t &= -y\sqrt{x^2+y^2} - \epsilon x \\ y_t &= x\sqrt{x^2+y^2} - \epsilon y \end{aligned} \tag{5.3}$$

while a vector field that satisfies neither, yet still has an absorbing set, is

$$\begin{aligned} x_t &= y - \epsilon\frac{x}{\sqrt{x^2+y^2}} \\ y_t &= -x - \epsilon\frac{y}{\sqrt{x^2+y^2}}. \end{aligned} \tag{5.4}$$

Up to this point, we have not considered how the numerical algorithm generates candidate timesteps since we only require that the error control is satisfied. However, for the sake of completeness, we shall explicitly describe the timestep selection mechanism used in the numerical simulations. This algorithm is based upon asymptotic considerations (see for example [14, 7, 12]) as the tolerance τ, and thus the timesteps, tend to zero. If h_{last} was the last attempted timestep (successful or otherwise), then the next attempted timestep is defined by

$$h_{\text{next}} = \min\left(h_{\max}, \theta\left(\frac{\sigma(\tau, U)}{E(U, h_{\text{last}})}\right)^{\frac{1}{p-\rho}} h_{\text{last}}\right)$$

where U is the most recent solution value. The constant $\theta < 1$ is a 'safety-factor' ensuring that, provided the exact solution lies in a compact set, the proportion of rejected timesteps along numerical approximations will tend to 0 as $\tau \to 0$.

We first consider the behaviour of order (1,2) pairs with error-per-unit-step control. Figure 5.1 plots the Euclidean norm of the numerical solution against integration time for the ODEs (5.1) through (5.4) using the embedded Runge-Kutta pair consisting of the Forward Euler and Heun methods, defined by

$$A = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}, \quad b = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad \overline{b} = \begin{pmatrix} \frac{1}{2} \\ \frac{1}{2} \end{pmatrix}, \tag{5.5}$$

in extrapolation mode with $\tau = 0.1$, $\theta = 0.9$ and the norm of the initial data set to $\|U\| = 10^5$. Here, as in all subsequent results, a relative error criterion defined by

$$\sigma(\tau, u) = \tau\|u\|_2$$

was used as this results in larger timesteps and thus provides a more severe (and, arguably, more relevant) test than a pure absolute error control. Even for this relatively large value of τ, the results are in agreement with Theorem 4.1. The reduction in norm of the numerical solution for sufficiently small τ is guaranteed for (5.1) — (5.3) since this pair is essentially algebraically stable. For (5.4) the norm of the solution increases with this value of τ. If τ is reduced sufficiently then stability of the numerical solution is recovered for this initial data but the instability reappears as $\|U\|$ is increased further i.e. τ depends upon the initial data. In Figure 5.2, we test Assumptions 1 and 2 by plotting the calculated values of

$$k_1(U_n, h_n) = \frac{h_n^{p-\rho}\|f(U_n)\|}{\sigma(\tau, U_n)}$$

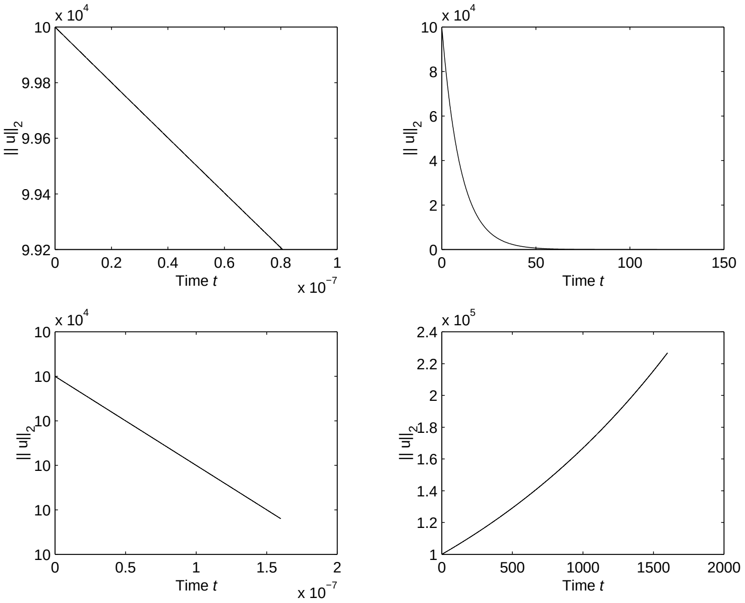


FIG. 5.1. *Figures a), b), c), d) plot the norms of the numerical solution using the embedded pair (5.5) in XEPUS mode for (5.1) — (5.4) respectively. The values of ϵ used in b), c) and d) are 0.1, 1, 1 and in each case* $\tau = 0.1$.

and

$$k_2(U_n, h_n) = \frac{\left| h_n^2 \sum_{i,j=1}^{s} m_{ij} \langle f(\eta_i), f(\eta_j) \rangle \right|}{\sigma(\tau, U_n) h_n^{1+\rho} \|f(U_n)\|}$$

The maxima of these quantities along the numerical trajectory are the effective values of K_1 and K_2 respectively and, if Assumptions 1 and 2 are justified, these quantities should remain bounded as $\|U_n\| \to \infty$. This is indeed the case for all four trajectories in Figure 5.1 and in Figure 5.2, k_1 and k_2 are plotted for just two of the test problems, namely (5.1) and (5.3) (for the linear ODE (5.2), these quantities are constant along the entire numerical trajectory).

Figure 5.3 is generated exactly as Figure 5.1 but using the non-essentially algebraically stable embedded pair

$$A = \begin{pmatrix} 0 & 0 \\ 2 & 0 \end{pmatrix}, \quad \overline{b} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad b = \begin{pmatrix} \frac{3}{4} \\ \frac{1}{4} \end{pmatrix}. \tag{5.6}$$

As can be seen, the results are very similar to those using the essentially algebraically stable (EAS) pair (5.5) and suggest that, although EAS pairs have guaranteed stability properties, there is little difference between EAS and non-EAS pairs in practice.

We now consider Theorem 4.2. Figure 5.4 is generated identically to Figure 5.1 except that now the method (5.5) is being used in XEPS mode rather than XEPUS. The interesting case is c), corresponding to the vector field (5.3) which satisfies **(D)** but not $(\mathbf{D'})$. Now the norm of the numerical solution increases rather than decreases and, for any given tolerance, this phenomenon appears to occur for sufficiently large initial data.

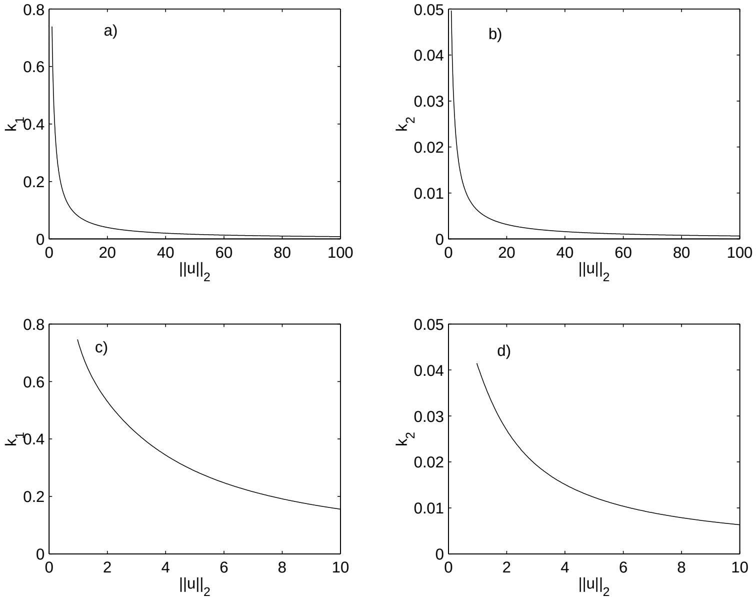


Fig. 5.2. *Figures a), b) show the computed values of k_1 and k_2 for a numerical trajectory of (5.1) while Figures c), d) are for (5.3). Apart from the initial data all the parameters are the same as used in Figure 1 a), c).*

Finally we present results for a higher-order pair. Figure 5.5 shows the numerical results obtained using the Fehlberg (4,5) pair in XEPS mode. Note that this embedded Runge-Kutta pair, whose coefficients are listed in [3][Page 306], does not satisfy the condition that the weights of the higher-order method are positive, but the results are similar to those obtained for other pairs that do satisfy this condition, suggesting that this condition could be weakened somewhat. Again, the importance of the additional structural assumption ($\mathbf{D'}$) is revealed in Figure c).

Acknowledgements

The author thanks one of the anonymous referees for their extremely careful reading of the original manuscript and many valuable suggestions.

REFERENCES

[1] K. Burrage and J.C. Butcher, *Stability criteria for implicit Runge-Kutta processes*, SIAM J. Num. Anal., 19 (1979), pp. 46–57.

[2] J.C. Butcher, *A stability property of implicit Runge-Kutta methods*, BIT, 15 (1975), pp. 358–361.

[3] ———, *The numerical analysis of ordinary differential equations*, Wiley, 1992.

[4] ———, *Numerical methods for ordinary differential equations*, Wiley, 2003.

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

[6] D. F. Griffiths, *The dynamics of some linear multistep methods with step-size control*, in Numerical analysis 1987 (Dundee, 1987), Longman Sci. Tech., 1988, pp. 115–134.

[7] E. Hairer, S.P. Nørsett, and G. Wanner, *Solving ordinary differential equations I. Nonstiff problems*, Springer-Verlag, Berlin, second ed., 1993.

[8] D.J. Higham and A.M. Stuart, *Analysis of the dynamics of local error control via a piecewise continuous residual*, BIT, 38 (1998), pp. 44–57.

[9] A.R. Humphries and A.M. Stuart, *Runge-Kutta methods for dissipative and gradient dy-*

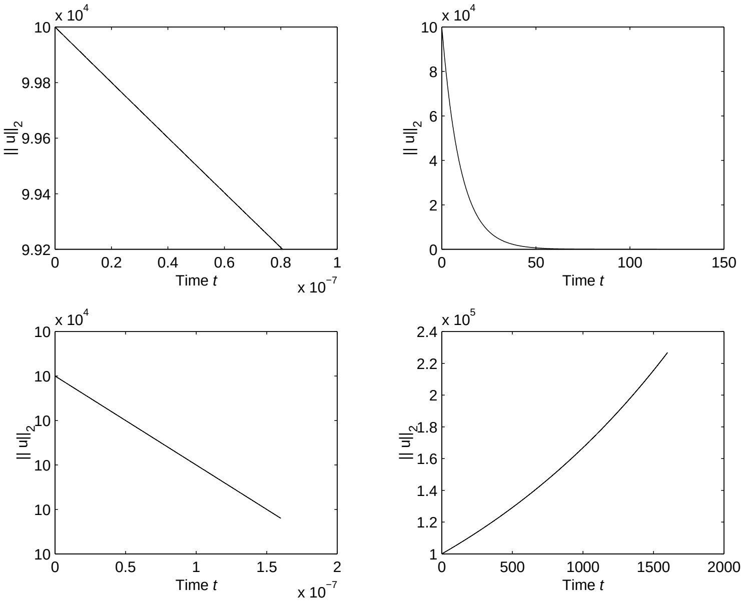


FIG. 5.3. *Figures a), b), c) d) are generated exactly as in Figure 1 but using the non-EAS pair (5.6).*

namical systems, SIAM J. Num. Anal., 31 (1994), pp. 1452–1485.

[10] H. LAMBA, *Dynamical systems and adaptive time-stepping in ODE solvers*, BIT, 40 (2000), pp. 314–335.

[11] H. LAMBA AND A.M. STUART, *Convergence results for the MATLAB ode23 routine*, BIT, 38 (1998), pp. 751–780.

[12] ———, *Convergence proofs for numerical IVP software*, in Dynamics of Algorithms, vol. 118 of IMA Volumes in Mathematics and its Applications, 1999, pp. 107–127.

[13] J. M. SANZ-SERNA, *Numerical ordinary differential equations vs. dynamical systems*, in The dynamics of numerics and the numerics of dynamics (Bristol, 1990), Oxford Univ. Press, New York, 1992.

[14] L.F. SHAMPINE, *Numerical Solution of Ordinary Differential Equations*, Chapman and Hall, New York, 1994.

[15] A.M. STUART, *Probabilistic and deterministic convergence proofs for software for initial value problems*, Numerical Algorithms, 14 (1997), pp. 227–260.

[16] A.M. STUART AND A.R. HUMPHRIES, *Model problems in numerical stability theory for initial value problems*, SIAM Review, 36 (1994), pp. 226–257.

[17] ———, *The essential stability of local error control for dynamical systems*, SIAM J. Num. Anal., 32 (1995), pp. 1940–1971.

[18] ———, *Dynamical systems and numerical analysis*, CUP, 1996.

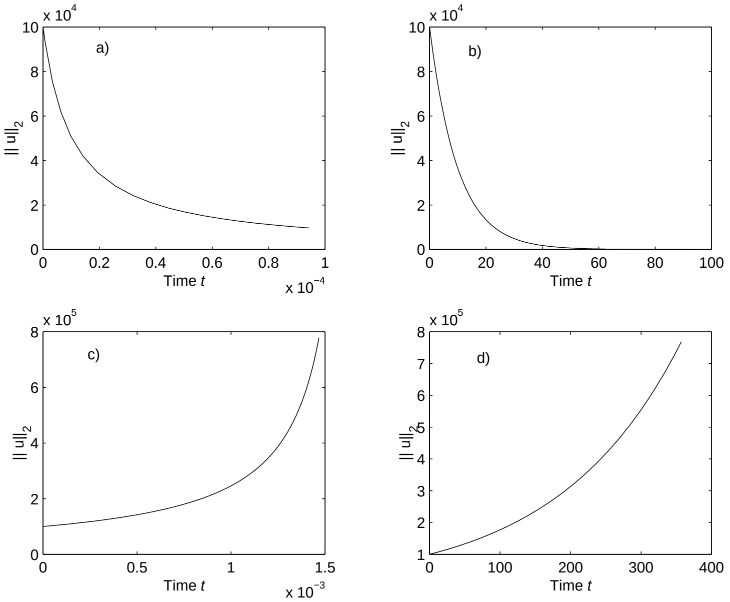


FIG. 5.4. *Figures a), b), c) d) are generated exactly as in Figure 1 but using the method (5.5) in XEPS rather than XEPUS mode.*

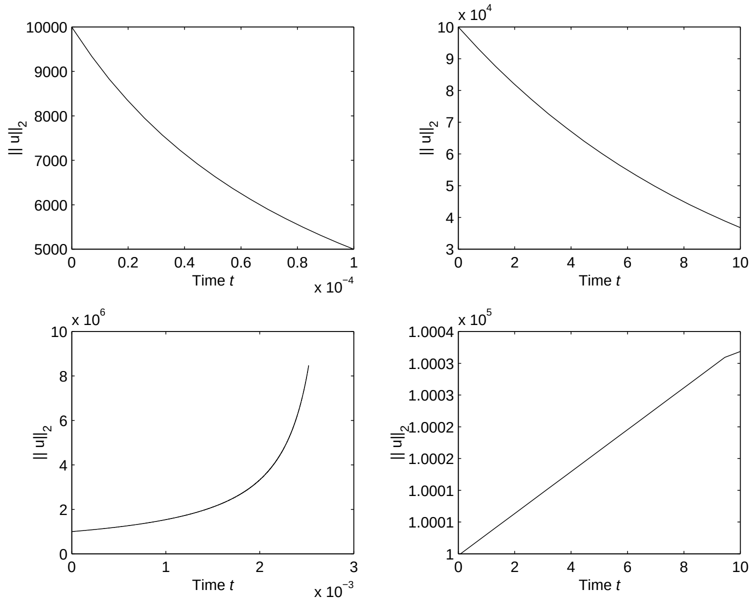


FIG. 5.5. *Figures a), b), c) d) are generated exactly as in Figure 5.1 but using the Fehlberg (4,5) pair in XEPS mode.*