Numerical Algorithms 14 (1997) 227–260

Probabilistic and deterministic convergence proofs for software for initial value problems

A. M. Stuart*

Scientific Computing and Computational Mathematics Program, Division of Mechanics and Computation, Durand 257, Stanford University, Stanford, CA 94305-4040, USA

The numerical solution of initial value problems for ordinary differential equations is frequently performed by means of adaptive algorithms with user-input tolerance τ. The time-step is then chosen according to an estimate, based on small time-step heuristics, designed to try and ensure that an approximation to the local error commited is bounded by τ. A question of natural interest is to determine how the global error behaves with respect to the tolerance τ. This has obvious practical interest and also leads to an interesting problem in mathematical analysis. The primary difficulties arising in the analysis are that: (i) the time-step selection mechanisms used in practice are discontinuous as functions of the specified data; (ii) the small time-step heuristics underlying the control of the local error can break down in some cases. In this paper an analysis is presented which incorporates these two difficulties.

For a mathematical model of an error per unit step or error per step adaptive Runge–Kutta algorithm, it may be shown that in a certain probabilistic sense, with respect to a measure on the space of initial data, the small time-step heuristics are valid with probability one, leading to a probabilistic convergence result for the global error as $\tau \to 0$. The probabilistic approach is only valid in dimension $m > 1$; this observation is consistent with recent analysis concerning the existence of spurious steady solutions of software codes which highlights the difference between the cases $m = 1$ and $m > 1$. The breakdown of the small time-step heuristics can be circumvented by making minor modifications to the algorithm, leading to a deterministic convergence proof for the global error of such algorithms as $\tau \to 0$. An underlying theory is developed and the deterministic and probabilistic convergence results proved as particular applications of this theory.

Keywords: error control, convergence.

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

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

1. Introduction

In this paper we consider the approximation of initial-value problems for ordinary differential equations by means of adaptive time-step software. We prove convergence results framed in terms of the tolerance τ. A fairly complete mathematical model of the software code is studied, incorporating step-rejections, due to violation of an estimated error bound, together with upper bounds on the maximum step-size and maximum step-size ratio. Thus we obtain a discontinuous dynamical system governing the evolution of the approximation of the solution itself, together with the time increments.

The heuristics underlying the algorithm are designed to ensure that the global error is bounded by a power of the tolerance τ. However, the heuristics break down for certain trajectories. The purpose of this paper is to include the possibility of this breakdown in a rigorous analysis of the discontinuous dynamical system governing the adaptive algorithm. Two types of results are proven. In the first type we prove that the probability of the heuristics breaking down is small with respect to a measure on the space of initial data; this leads to a proof of convergence with probability one. In the second type of result we make minor modifications of the standard algorithm; these modifications are designed to ensure that the algorithm actually behaves correctly even when the heuristics underlying it break down.

The use of probabilistic reasoning in numerical analysis has been fairly widespread in the context of linear algebra, starting with the work of Demmel [4], leading to more recent work in, for example, [5, 8] and [15]. Demmel's work itself was partially motivated by Smale's pioneering analysis of Newton's method; see [10]. However, to the best of our knowledge a probabilistic approach to the analysis of numerical methods has not been undertaken in the context of differential equations. The first result in the literature concerning the behaviour of the global error with respect to tolerance τ appears to be [12]; this work has been developed further in, for example [7, 9]. The work of [12] forms the basis for the work presented here but we extend in three important ways. Firstly, in [12] the *assumption* is made that the leading term in the expansion for the local error estimate does not vanish along the trajectory being approximated; this assumption does not hold for all trajectories and, furthermore, for trajectories close to those which violate the assumption, certain constants appearing in [12] will be large; this issue is not addressed in [12]. Secondly, the paper [12] relies on an asymptotic expansion for the error in powers of the time-step Δt and only the analysis of the leading order term is given in detail; here we control the complete error, leading to more precise estimates on how small the tolerance and initial time-step need to be for the analysis to hold. This issue is related to the first point since these upper bounds on the tolerance and initial time-step may be particularly severe for solutions close to trajectories along which the leading term in the error estimate disappears. Thirdly, [12] employs a simplified model of the step-size selection mechanism that does not incorporate step-size rejection and maximum step-size and step-size ratio bounds explicitly in the analysis.

The first of these points is addressed in [3], although the second and third points are not addressed there. In [3] the basic step-size selection procedure is appended

with a computational estimate of the leading term in the error estimate and, whenever this is small, the step-size selection procedure is modified. Using this modification of the algorithm the authors of [3] improve upon the results of [12].

The paper [13] is also an important contribution to the rigorous analysis of automatic step-size control. However, in the context that interests us here that work is slightly lacking in two main respects: firstly, the maximum step-size is assumed *a priori* to be bounded by a positive power of the tolerance τ, something which is not true for most software codes used in practice – see [6]; secondly, as in [3, 7, 9] and [12], step-size rejection is not included in the analysis.

We now introduce the mathematical background in which our results are framed, starting with the initial-value problem which we wish to approximate. Consider the equation

$$\frac{\mathrm{d}u}{\mathrm{d}t} = f(u), \quad u(0) = U, \tag{1.1}$$

where $f \in C^\infty(\mathbb{R}^m, \mathbb{R}^m)$. Thus we have a local solution $u(t) = S(U,t)$ defined, for every $U \in \mathbb{R}^m$, and t in an interval $I_1 = I_1(U) \subseteq \mathbb{R}$. Furthermore, on its interval of existence we have $S(U,\cdot) \in C^\infty(I_1, \mathbb{R}^m)$. Thus for $(U,t) \in \mathbb{R}^m \times I_1(U)$, we may form the Taylor series expansion

$$\begin{aligned} S(U,t) = &\sum_{j=0}^{r+1} \frac{1}{j!} \beta_j^{(0)}(U,0) t^j \\ &+ \frac{t^{r+2}}{(r+1)!} \int_0^1 (1-s)^{r+1} \beta_{r+2}^{(0)}(U; st)\, \mathrm{d}s, \quad \forall r \in \mathbb{Z}^+, \end{aligned} \tag{1.2}$$

where

$$\beta_j^{(0)}(u,t) = \frac{\partial^j}{\partial t^j}\big\{S(u,t)\big\}.$$

For simplicity we assume in the remainder of the paper that f and all its derivatives are uniformly bounded on $\mathbb{R}^m$. This simplifies the analysis and statement of results but is not actually necessary for the results to hold.

We also consider two explicit Runge–Kutta methods approximating the flow generated by (1.1). Consider the equations

$$\eta_i = U + t \sum_{j=1}^{l} a_{ij} f(\eta_j), \quad i = 1, \ldots, l. \tag{1.3}$$

Since the η_i represent the internal stages of an explicit Runge–Kutta method, it follows that $a_{ij} = 0,\ i \leqslant j$. Hence $\eta_i(U,t) \in C^\infty(\mathbb{R}^m \times \mathbb{R}, \mathbb{R}^m)$. Given $b_i^{(k)}$ for $i = 1, \ldots, l$ and $k = 1, 2$ we define, for $k = 1, 2$,

$$S^{(k)}(U,t) = U + t \sum_{i=1}^{l} b_i^{(k)} f\big(\eta_i(U,t)\big) \tag{1.4}$$

noting that $S^{(k)}(U,t) \in C^\infty(\mathbb{R}^m \times \mathbb{R}, \mathbb{R}^m)$. The mapping $S^{(1)}$ will be used to advance the solution, whilst $S^{(2)}$ will be used to control the error. For all $(U,t) \in \mathbb{R}^m \times \mathbb{R}$ we may form the Taylor series expansions

$$S^{(k)}(U,t) = \sum_{j=0}^{r+1} \frac{1}{j!} \beta_j^{(k)}(U,0) t^j + \frac{t^{r+2}}{(r+1)!} \int_0^1 (1-s)^{r+1} \beta_{r+2}^{(k)}(U; st)\, \mathrm{d}s, \quad \forall r \in \mathbb{Z}^+, \tag{1.5}$$

where

$$\beta_j^{(k)}(u,t) = \frac{\partial^j}{\partial t^j} \{ S^{(k)}(u,t) \}.$$

Equations (1.3) and (1.4) with $k = 1, 2$ define two-distinct Runge–Kutta methods. We assume that the two Runge–Kutta methods have order s and q_1. To be more precise we shall assume that

$$\begin{aligned} \beta_j^{(1)}(u,0) &\equiv \beta_j^{(0)}(u,0), \quad j = 0, \ldots, s, \\ \beta_j^{(2)}(u,0) &\equiv \beta_j^{(0)}(u,0), \quad j = 0, \ldots, q_1, \\ \beta_j^{(1)}(u,0) &\equiv \beta_j^{(2)}(u,0), \quad j = 0, \ldots, q_2, \end{aligned} \tag{1.6}$$

and that agreement between the β's does not occur at the next order in each of the three cases. Note that $q_2 \geqslant \min\{s, q_1\}$ and that, if $s \neq q_1$, then $q_2 = \min\{s, q_1\}$. We do not specify which of the two methods has higher order; this allows us to consider methods operating both in extrapolation ($s > q_1$) and non-extrapolation ($s \leqslant q_1$) modes.

The numerical method for the approximation of (1.1) is now described. Let U_n denote our approximation to $u(t_n)$ where

$$t_n = \sum_{j=0}^{n-1} \Delta t_j.$$

The sequences $\{U_n\}$ and $\{\Delta t_n\}$ are generated as follows. Define

$$E(u,t) = \left\| S^{(1)}(u,t) - S^{(2)}(u,t) \right\| / t^\rho. \tag{1.7}$$

(Throughout the paper $\|\cdot\|$ will denote the Euclidean norm on $\mathbb{R}^m$.) Here $\rho = 0$ if we consider error per-step (EPS) and $\rho = 1$ if we consider error per unit step (EPUS). Thus

$$E(u,t) = \mathcal{O}(t^q),$$

where $q = q_2 + 1 - \rho$. Then U_n and Δt_n are generated so that

$$\begin{aligned} U_{n+1} &= S^{(1)}(U_n; \Delta t_n), \quad U_0 = U, \\ \Delta t_{n+1} &= \beta^k \Delta t_{n+1}^{(0)}, \quad \Delta t_0^{(0)} = \Delta t_{\text{init}}, \end{aligned} \tag{1.8}$$

where $k = k(U_n, \Delta t_n)$ is the minimal non-negative integer such that

$$E(U_{n+1}, \Delta t_{n+1}) \leqslant \tau \tag{1.9}$$

and

$$\Delta t_{n+1}^{(0)} = \min\left\{\theta\left(\frac{\tau}{E(U_n, \Delta t_n)}\right)^{1/q}\Delta t_n, \alpha\Delta t_n, D\right\}. \tag{1.10}$$

Here $0 < \beta < 1$, $0 < \theta < 1 < \alpha$ and $D \in \mathbb{R}^+$. Throughout, θ and D will be fixed independently of τ. For all results in sections 4 and 5 we will assume that α is fixed independently of τ whilst for the results in section 6 only we will assume that $\alpha \to 1_+$ as $\tau \to 0_+$. Thus, unless explicitly stated, α will be assumed independent of τ; this is the case for most software used in practice.

Note that, by (1.8) and (1.10), it follows that

$$\Delta t_{n+1} \leqslant \Delta t_{n+1}^{(0)} \leqslant \min\{\alpha\Delta t_n, D\},$$

so that a maximum step-size ratio of α and a maximum step-size of D are imposed by the algorithm. Note also that Δt_{n+1} is discontinuous as a function of Δt_n and U_n, the potential discontinuities being introduced through the selection of the integer k. In summary, we have a dynamical system of the form

$$\begin{aligned} U_{n+1} &= \Phi(U_n, \Delta t_n), \\ \Delta t_{n+1} &= \Gamma(U_n, \Delta t_n), \end{aligned} \tag{1.11}$$

where Φ is smooth and Γ discontinuous as functions of their arguments. In fact, we will prove implicitly that such a function Γ is well-defined in the course of the paper. We will need the following definition of *truncation error* in the remainder of the paper:

$$T(u,t) = S^{(1)}(u,t) - S(u,t). \tag{1.12}$$

In section 2 we introduce some background properties and notation for the underlying Runge–Kutta methods (1.3), (1.4). In section 3 we prove a number of basic results concerning the adaptive algorithm (1.7)–(1.10) which are used in subsequent sections to prove our main theorems. In section 4 we state a basic convergence result of $\mathcal{O}(\tau^{s/q})$ for the adaptive algorithm (see theorem 4.1) and use it to study linear constant coefficient differential equations. The case $s < q_1$ and $\rho = 1$ (so that $q = q_2 = s$) leads to an error per unit step code with global error $\mathcal{O}(\tau)$. Similarly, the case $s < q_1$ and $\rho = 0$ so that $q_2 = s$ and $q = s + 1$ leads to an error per step code with global error $\mathcal{O}(\tau^{s/(s+1)})$. In section 5 we use theorem 4.1 to prove that, with probability one, the adaptive algorithm converges on a general class of nonlinear problems and that the global error is $\mathcal{O}(\tau^{s/q})$ as $\tau \to 0$. Actually we prove more, estimating the probability that the error constant in the global error falls below a given specified number – see theorem 5.1. The probabilistic approach is only valid in dimension $m > 1$; this observation is consistent with recent very interesting constructions of spurious steady

solutions of software codes which highlight the difference between the cases $m = 1$ and $m > 1$ – see [1].

In section 6 we present certain modifications to the algorithm (1.7)–(1.10) which allow the global error to be controlled in the exceptional cases shown to have small probability in section 5. This leads to several deterministic convergence results, namely theorems 6.2 and 6.4. In practice we believe that the probabilistic convergence results of section 5 are of more value than the deterministic results of section 6. This is because the modifications to the basic algorithm which we propose to obtain deterministic results are not currently used in practice; furthermore, they are in any case only of use for certain exceptional cases of small probability. It would be of interest to study the effect on these cases of small probability of the modifications to the basic algorithm (1.7)–(1.10) proposed in [3]; currently [3] does not include the effect of step-size rejections. The reader interested only in probabilistic convergence results need only study section 2, section 3, up to and including assumption 3.5, and sections 4 and 5.

We have chosen to analyse an algorithm which faithfully represents most of the important features of real adaptive Runge–Kutta based algorithms for the solution of initial-value problems. Nonetheless, any writer of software code will be able to find features not addressed in this analysis. We mention three such features and indicate how the analysis given here could be extended to include them.

The first is the fact that the error estimate in real codes typically has an absolute and a relative component. Specifically, the constraint (1.9) is replaced by

$$E(U_{n+1}, \Delta t_{n+1}) \leqslant \tau \max\left\{a, b\|U_{n+1}\|\right\} \tag{1.13}$$

for some fixed $a, b > 0$. This change can be studied by similar techniques to those used here since, provided bounded solutions are studied,

$$\sup_{t\in[0,T]} \left\|u(t)\right\| \leqslant c$$

and, for U_n which are close to $u(t_n)$ for τ small, (1.13) implies that

$$E(U_{n+1}, \Delta t_{n+1}) \leqslant \tau \max\{a, bc\} + \mathrm{o}(\tau).$$

The similarity of this inequality to (1.9) enables adaptation of the analysis given in this paper to the case where (1.13) is used rather than (1.9).

The second is the fact that many Runge–Kutta codes use Richardson extrapolation to estimate the error. Specifically we have

$$S^{(1)}(u,t) = S(u,t) + T(u,t)$$

and $T(u,t) = \mathcal{O}(t^{s+1})$. Thus the error estimate is

$$E(u,t) = \left\|S^{(1)}(u,t) - S^{(1)}\big(S^{(1)}(u,t/2),t/2\big)\right\|/t^{\rho}.$$

A little manipulation shows that

$$E(u,t) = \mathcal{O}(t^{s+1-\rho})$$

and that $E(u,t)$ is proportional to $f(u)$. Hence the methodology presented in this paper may be applied to this situation.

The third is the fact that only explicit Runge–Kutta methods are considered. To study implicit methods, an additional criterion would have to be added to the adaptation of the time-step, namely to choose it sufficiently small that the implicit equations have a unique solution in a small neighbourhood of the solution at the previous time-step. If this is done then similar techniques to those used here could be applied. However, implicit methods are typically used for stiff problems and the question of deriving stiffness independent error estimates would require special attention.

2. Properties of the underlying methods

Here we describe the basic properties of $S(\cdot,\cdot)$ and the $S^{(k)}(\cdot,\cdot)$ which we require in the remaining sections. We start by defining *elementary differentials*. Let

$$\begin{aligned} e^{(i)} &= (0,0,\ldots,1,0,\ldots,0)^T \in \mathbb{R}^m, \\ K^{(j)} &= \big(K_1^{(j)},\ldots,K_m^{(j)}\big)^T \in \mathbb{R}^m, \\ f(u) &= \big(f_1(u),\ldots,f_m(u)\big)^T \in \mathbb{R}^m, \end{aligned}$$

where $f_j(u):\mathbb{R}^m \mapsto \mathbb{R}$ and $f(u):\mathbb{R}^m \mapsto \mathbb{R}^m$. In the above, only the ith entry of the vector $e^{(i)}$ is non-zero. We define the Mth Fréchet derivative of f, namely $f^{(M)}:\mathbb{R}^{Mm} \mapsto \mathbb{R}^m$, by

$$f^{(M)}(z)\big(K^{(1)},\ldots,K^{(M)}\big) = \sum_{i=1}^{m}\sum_{j_1=1}^{m}\cdots\sum_{j_M=1}^{m} \frac{\partial^M f_i(z)}{\partial z_{j_1}\cdots\partial z_{j_M}} K_{j_1}^{(1)}\cdots K_{j_M}^{(M)} e^{(i)}.$$

The elementary differentials of f are denoted $F_s:\mathbb{R}^m \mapsto \mathbb{R}^m$ and their order by r_s; these are defined recursively by:

(i) $f(u)$ is the only elementary differential of order 1;

(ii) if $F_s(u)$, $s = 1,\ldots,M$, are elementary differentials of order r_s, $s = 1,\ldots,M$, respectively then $f^{(M)}(u)(F_1(u),\ldots,F_M(u))$ is an elementary differential of order $1 + \sum_{s=1}^{M} r_s$.

Lemma 2.1. For $k = 0,1,2$ and each j for which they are defined, the $\beta_j^{(k)}(U,0)$ are linear combinations of elementary differentials of order j.

Proof. See Butcher [2]. □

Lemma 2.2. For each $j \geqslant 1$ for which $\beta_j^{(k)}(u,t)$ are defined and for $k = 1, 2$, there exist l $p \times p$ matrices $d_{i,j}^{(k)}(u,t)$, and a $p \times p$ matrix $d_j^{(0)}(u,t)$, $i = 1, \ldots, l$, such that

$$\beta_j^{(k)}(u,t) = \sum_{i=1}^{l} d_{i,j}^{(k)}(u,t) f\big(\eta_i(u,t)\big), \quad k = 1, 2,$$
$$\beta_j^{(0)}(u,t) = d_j^{(0)}(u,t) f(u).$$

Proof. See lemma 4.6.4 of [14] for $k = 1, 2$. For $k = 0$ the result follows by the chain rule. □

Lemma 2.3. Let $f(u) = Au$ for some $m \times m$ matrix A. Then for $k = 1, 2$ and $j = 1, 2, \ldots$ we have real numbers $c_j^{(k)}$ such that

$$\beta_j^{(k)}(u,0) = c_j^{(k)} A^j u.$$

Proof. If $f(u) = Au$ then $f^{(M)}(z)$ is identically zero for $M > 1$. It follows that the only elementary differential of order j is $A^j u$ and lemma 2.1 gives the desired result. □

In the case where linear problems are considered (section 4) s (see the discussion after equation (1.6)) will be taken to be the effective order of the method when applied to the class of linear autonomous problems. This may be higher than the order on the general class of problems (1.1) because of the special structure of the truncation error predicted by lemma 2.3. Similar considerations apply to q_1, q_2 and q.

Now define

$$B_1(u) := \big[\beta_{q_2+1}^{(1)}(u,0) - \beta_{q_2+1}^{(2)}(u,0)\big]/(q_2+1)!, \tag{2.1}$$
$$B_2(u,t) := \int_0^1 \big[\beta_{q_2+2}^{(1)}(u,ts) - \beta_{q_2+2}^{(2)}(u;ts)\big] \frac{(1-s)^{q_2+1}}{(q_2+1)!}\, \mathrm{d}s,$$

$$b_1(u) = \begin{cases} B_1(u)/\|f(u)\|, & \|f(u)\| \neq 0, \\ 0, & \|f(u)\| = 0; \end{cases} \tag{2.2}$$
$$b_2(u,t) = \begin{cases} B_2(u,t)/\|f(u)\|, & \|f(u)\| \neq 0, \\ 0, & \|f(u)\| = 0. \end{cases}$$

Note also that, by (1.5) with $r = q_2$ and (1.6)–(1.7),

$$E(u,t) = t^q \big\|B_1(u) + tB_2(u,t)\big\| = t^q \big\|f(u)\big\| \big\|b_1(u) + tb_2(u,t)\big\|. \tag{2.3}$$

Thus the leading order term in the error estimate $E(u,t)$ is zero when $f(u)=0$ or when $b_1(u)=0$. Such points, and their neighbourhoods, will be of crucial importance in the remainder of the paper. We briefly outline why this is so. Assume that $f(U_n)$ and $b_1(U_n)$ are non-zero for all n. It then follows from (1.7) and (1.9) that

$$\Delta t_n^q \big\|f(U_n)\big\|\big\|b_1(U_n)\big\| \leqslant \tau + \mathcal{O}\big(\Delta t_n^{q+1}\big)$$

so that, provided Δt_n is small, $\Delta t_n \leqslant C\tau^{1/q}$. It then follows from (1.6), (1.12) that the local truncation error is of order $\Delta t_n \tau^{s/q}$. A straightforward Gronwall argument proves that the global error behaves like $\tau^{s/q}$. This is the essence of the result in [12] where $f(u)$ and $b_1(u)$ are assumed to be bounded away from zero on any trajectory under consideration. Since we do not wish to make this assumption, the behaviour of the algorithm near to points where $f(u)$ and $b_1(u)$ disappear will be of crucial importance to us.

In the remainder of the section we prove some results concerning the functions now defined.

Lemma 2.4. Let $J \subset \mathbb{R}^m$ be bounded. Then, if $dB_1(\cdot)$ denotes the Jacobian of $B_1(\cdot)$,

$$\sup_{u\in J}\big\|b_1(u)\big\| < \infty, \qquad \sup_{(u,t)\in J\times[0,D]}\big\|b_2(u,t)\big\| < \infty,$$

$$\sup_{(u,v,s)\in J\times J\times[0,1]}\big\|dB_1\big(su+(1-s)v\big)\big\| < \infty.$$

Proof. We consider the EPUS case $\rho = 1$ and $q_2 = q$; the EPS case $q_2 = q-1$ is similar. Since $S^{(k)}(U,t) \in C^\infty(\mathbb{R}^m \times \mathbb{R}, \mathbb{R}^m)$ it follows that

$$\beta_j^{(k)}(U,t) \in C^\infty\big(\mathbb{R}^m \times \mathbb{R}, \mathbb{R}^m\big)$$

for $j = 1, \ldots, q+2$; thus lemma 2.2 gives an expression for $\beta_j^{(k)}(u,t)$ for $j = 1, \ldots, q+2$. By lemma 4.2.6 in [14] we have that there are constants $c_i = c_i(u)$ such that

$$\big\|f\big(\eta_i(u,t)\big)\big\| \leqslant (1+c_i t)\big\|f(u)\big\|. \tag{2.4}$$

Thus lemma 2.2 gives

$$\big\|\beta_j^{(1)}(u,t) - \beta_j^{(2)}(u,t)\big\| \leqslant k_j(u,t)\big\|f(u)\big\| \tag{2.5}$$

for some functions $k_j(u,t)$ which are bounded for $(u,t) \in J \times [0,D]$ and $j = 1, \ldots, q+2$. Hence, since by (2.3),

$$\big\|b_1(u)\big\| = \big\|f(u)\big\|^{-1}\big\|B_1(u)\big\|, \qquad \big\|b_2(u,t)\big\| = \big\|f(u)\big\|^{-1}\big\|B_2(u,t)\big\|,$$

the first two results follow from (2.2). Also $\beta_{q+1}^{(k)}(U,t) \in C^1(\mathbb{R}^m \times \mathbb{R}, \mathbb{R}^m)$ and hence

$$dB_1(w) := \frac{1}{(q+1)!}\, d\beta_{q+1}^{(1)}(w,0) - d\beta_{q+1}^{(2)}(w,0) \in C(\mathbb{R}^m \times \mathbb{R}, \mathbb{R}^m).$$

The third result follows. □

The final result of this section concerns the truncation error given by (1.12).

Lemma 2.5. Let $J \subset \mathbb{R}^m$ be bounded. There is a constant $K = K(J, D)$ such that, for all $u \in J$, $t \in [0, D]$,

$$T(u,t) \leqslant K\big\|f(u)\big\|t^{s+1}.$$

Proof. By (1.2), (1.5) and (1.6) we have

$$\big\|S^{(1)}(u,t) - S(u,t)\big\| \leqslant C\big\|\beta_{s+1}^{(1)}(u,t) - \beta_{s+1}^{(0)}(u,t)\big\|t^{s+1}.$$

Hence (2.5) gives the desired result. □

3. Properties of the error control

In this section we prove three basic lemmas concerning the behaviour and error control properties of the algorithm (1.7)–(1.10) in the following three situations: (a) $\|B_1(S(U,t))\|$ is bounded away from zero; (b) $\|B_1(S(U,t))\|$ is small for some t because $\|b_1(S(U,t))\|$ is small for some t; (c) $\|B_1(u)\|$ is small for some t because $\|f(S(U,t))\|$ is small for some t. These are lemmas 3.4, 3.7 and 3.8 respectively. In subsequent sections we use these lemmas as building blocks for our theorems. Sections 4 and 5 require only material from this section up to and including assumption 3.5, whilst section 6 requires all the material from this section.

In lemma 3.3, we prove that the time-step Δt_n is bounded away from zero uniformly in n and hence that any finite time T can be reached by the algorithm.

We start by introducing some notation. As mentioned in the previous section the neighbourhoods of points where $f(u)$ and $b_1(u)$ disappear will be crucial in our analysis and this motivates the following. Let

$$\begin{aligned} \Psi(\varepsilon) &:= \big\{u \in \mathbb{R}^m\colon\ \big\|b_1(u)\big\| < \varepsilon\big\}, \\ \Psi(0) &:= \big\{u \in \mathbb{R}^m\colon\ \big\|b_1(u)\big\| = 0\big\} = \bigcap_{\varepsilon>0} \Psi(\varepsilon) \end{aligned} \tag{3.1}$$

and

$$\begin{aligned} \Gamma(\delta) &:= \big\{u \in \mathbb{R}^m\colon\ \big\|f(u)\big\| < \delta\big\}, \\ \Gamma(0) &:= \big\{u \in \mathbb{R}^m\colon\ \big\|f(u)\big\| = 0\big\} = \bigcap_{\delta>0} \Gamma(\delta). \end{aligned} \tag{3.2}$$

Define, given sets $I, J \subset \mathbb{R}^m$,

$$\chi = \Gamma(\delta) \cup \Psi(\varepsilon),$$

$$J_{\varepsilon,\delta} = J\backslash\chi, \qquad I_{\varepsilon,\delta} = I\backslash\chi, \qquad J_\varepsilon = J\backslash\Psi(\varepsilon), \qquad I_\varepsilon = I\backslash\Psi(\varepsilon).$$

We also define the constants $C_3, \ldots, C_9$ (depending on bounded $J \subset \mathbb{R}^m$ and D) by

$$C_3 = \sup_{u\in J} \|b_1(u)\|, \qquad C_4 = \sup_{(u,t)\in J\times[0,D]} \|b_2(u,t)\|,$$

$$C_5 = \sup_{(u,v,s)\in J\times J\times[0,1]} \|dB_1\big(su+(1-s)v\big)\|, \qquad C_6 = \sup_{u\in J} \|f(u)\|,$$

$$C_7 = \big\{C_6[C_3 + DC_4]\big\}^{-1/q}, \qquad C_8 = Lip\big\{f(u),\ u\in J\big\},$$

$$C_9 = \sup_{(u,t)\in J\times[0,D]} E(u,t)/t^q, \qquad C_{10} = \min\left\{\frac{\theta^{2q}}{\alpha^q}, \beta^q\right\}\frac{1}{C_9}.$$

These constants are finite by lemma 2.4, the smoothness of $f(u)$ and (2.3).

In the following either or both of δ and ε may be fixed independently of τ or may be chosen proportional to a positive power of τ. Thus we set

$$\delta = \delta_0\tau^a, \qquad \varepsilon = \varepsilon_0\tau^b, \tag{3.3}$$

for some $\delta_0, \varepsilon_0 > 0$ and $a, b \geqslant 0$. We *always* assume that a and b in (3.3) are chosen so that

$$\tau/(\delta\varepsilon^{1+q}) \to 0 \quad \text{as } \tau \to 0. \tag{3.4}$$

This automatically holds if $a = b = 0$. Only the case $a = b = 0$ will be used in sections 4 and 5. In the case where $a = 0$ in (3.3), then (3.4) implies that

$$\tau/(\delta^2\varepsilon^2) \to 0 \quad \text{as } \tau \to 0, \tag{3.5}$$

since $q \geqslant 1$.

We now consider a solution, or families of solutions, of (1.1) defined for $t \in [0, T]$ and satisfying $u(t) \in I$, where $I \subset \mathbb{R}^m$ is bounded. Let $J = \overline{\mathcal{N}}(I, d)$ for some fixed $d > 0$, independent of τ. We will show that our numerical approximation lies in J in lemmas 3.4, 3.7 and 3.8; the theorems proved using those lemmas will hence involve constants depending upon J.

The first lemma of this section bounds the time-step selected by the algorithm at a point where $\|B_1(u)\|$ is bounded away from zero.

Lemma 3.1. Let $u \in J_{\varepsilon,\delta}$. Then all $t \in \mathbb{R}^+$ satisfying $E(u,t) \leqslant \tau$ and $t \in [0, \varepsilon q/C_4(q+1)]$ also satisfy $t^q \leqslant (q+1)\tau/\delta\varepsilon$.

Proof. If $E(u,t) \leqslant \tau$ and $u \in J_{\varepsilon,\delta}$ then (2.3) gives

$$\|b_1(u) + tb_2(u,t)\|t^q \leqslant \tau/\delta.$$

But

$$\|b_1(u) + tb_2(u,t)\| \geqslant \|b_1(u)\| - t\|b_2(u,t)\| \geqslant \varepsilon - C_4 t \geqslant \varepsilon/(q+1).$$

Thus $\varepsilon t^q \leqslant (q+1)\tau/\delta$ and the result follows. □

The second lemma simply uses induction to extend the previous result over several steps.

Lemma 3.2. Assume that there exist integers $M = M(\tau)$ and $N = N(\tau)$ such that $U_n \in J_{\varepsilon,\delta}$ for $n = M, \ldots, N$. Then if $\Delta t_M^{(0)} \leqslant \varepsilon q / C_4(q+1)$ it follows that, for τ sufficiently small,

$$\Delta t_n^q \leqslant (q+1)\tau/\delta\varepsilon, \quad \forall n = M, \ldots, N.$$

Proof. Note that if

$$\Delta t_n^{(0)} \leqslant \frac{\varepsilon q}{C_4(q+1)}$$

then

$$\Delta t_n \leqslant \frac{\varepsilon q}{C_4(q+1)}$$

by (1.8), since $\beta < 1$.

We use induction. Clearly the result holds for $n = M$ by lemma 3.1. Assume that it holds for $n = p \leqslant N-1$. Then, by (1.10) and (3.4),

$$\Delta t_{p+1}^{(0)} \leqslant \alpha \left(\frac{(q+1)\tau}{\delta\varepsilon} \right)^{1/q} \leqslant \frac{\varepsilon q}{C_4(q+1)}$$

for τ sufficiently small. Lemma 3.1 completes the induction. □

Define $\phi = \tau^{1/q}/C_9^{1/q}$ and note that

$$E(v,t) \leqslant \tau, \quad \forall v \in J,\ t \in (0,\phi), \tag{3.6}$$

provided that τ is sufficiently small to ensure that $\phi \leqslant D$. The next lemma gives a bound on the time-step Δt_n from below; the bound is expressed in terms of ϕ.

Lemma 3.3. Assume that there exists an integer $N = N(\tau)$ such that $U_n \in J$ for $n = 0, \ldots, N$. Then, for all τ sufficiently small,

$$\Delta t_n \geqslant \min\left\{ \frac{\theta^2\phi}{\alpha}, \beta\phi, \Delta t_0 \right\}, \quad n = 0, \ldots, N. \tag{3.7}$$

Proof. Clearly (3.7) holds for $n = 0$. Assume that it holds for $n = m \leqslant N-1$ for induction. Let $\widetilde{E}_m := E(U_m, \Delta t_m)$. If

$$\widetilde{E}_m \geqslant \tau(\theta/\alpha)^q \tag{3.8}$$

then, by definition of C_9,

$$\tau(\theta/\alpha)^q \leqslant C_9 \Delta t_m^q, \tag{3.9}$$

so that $\Delta t_m \geqslant \theta\phi/\alpha$. By (3.8), (3.9) and (1.10)

$$\Delta t_{m+1}^{(0)} = \min\left\{ \theta\big(\tau/\widetilde{E}_m\big)^{1/q} \Delta t_m, D \right\} \geqslant \min\left\{ \theta \Delta t_m, D \right\} \geqslant \min\left\{ \theta^2\phi/\alpha, D \right\}.$$

For τ sufficiently small we obtain

$$\Delta t_{m+1}^{(0)} \geqslant \theta^2\phi/\alpha.$$

If $\Delta t_{m+1}^{(0)} < \phi$ then by (3.6) $\Delta t_{m+1} = \Delta t_{m+1}^{(0)} \geqslant \theta^2\phi/\alpha$ whilst if $\Delta t_{m+1}^{(0)} \geqslant \phi$ then $\Delta t_{m+1} \geqslant \beta\phi$ by (3.6). Hence, if (3.8) holds then we have

$$\Delta t_{m+1} \geqslant \min\{\theta^2\phi/\alpha, \beta\phi\} \tag{3.10}$$

and (3.7) holds with $n = m+1$. If (3.8) does not hold then

$$\widetilde{E}_m \leqslant \tau(\theta/\alpha)^q \tag{3.11}$$

and, by (1.10),

$$\Delta t_{m+1}^{(0)} = \min\{\alpha\Delta t_m, D\}. \tag{3.12}$$

If $\Delta t_{m+1} < \beta\phi$ then $\Delta t_{m+1}^{(0)} = \Delta t_{m+1}$; this follows since, by (1.8),

$$\Delta t_{m+1} = \beta^k \Delta t_{m+1}^{(0)},$$

where $k = 0$ or where there is an integer $k > 0$ such that

$$\begin{aligned} E\big(U_{m+1}, \beta^j \Delta t_{m+1}^{(0)}\big) &> \tau, \quad 0 \leqslant j < k, \\ &\leqslant \tau, \quad j = k. \end{aligned} \tag{3.13}$$

But $\beta^{k-1}\Delta t_{m+1}^{(0)} = (1/\beta)\Delta t_{m+1} < \phi$ so that (3.6) shows (3.13) is impossible so that $k = 0$. Thus, if (3.11) holds we have from (3.12), for τ sufficiently small,

$$\Delta t_{m+1} \geqslant \min\{\beta\phi, \alpha\Delta t_m, D\} = \min\{\beta\phi, \Delta t_m\}. \tag{3.14}$$

This completes the proof. □

We now consider the numerical error. Let $u_n = u(t_n)$ and define $E_n = \|u_n - U_n\|$ (not to be confused with the functions $E(\cdot,\cdot)$ and $\widetilde{E}_m$ defined from it). Recall that s is the order of the method $S^{(1)}(\cdot,\cdot)$ used to advance the solution.

Lemma 3.4. Let $a = 0$ in (3.3) and assume that $u(t) \in I_{2\varepsilon,2\delta}$ for $T_L \leqslant t \leqslant T_0$. Also let $M = M(\tau),\ N = N(\tau)$ be such that $T_L \leqslant t_M \leqslant t_{N-1} \leqslant T_0$. Assume in addition that there exists a constant K such that, for all τ sufficiently small,

$$\Delta t_M^{(0)} \leqslant \frac{\varepsilon q}{C_4(q+1)} \quad \text{and} \quad E_M \leqslant K\left(\frac{\tau}{\delta\varepsilon}\right)^{s/q}.$$

Then there exist $K_1, K_2 \geqslant 0$ such that, for $n = M, \ldots, N$,

$$E_n \leqslant \left[E_M + K_1\left(\frac{\tau}{\delta\varepsilon}\right)^{s/q}(t_n - t_M)\right]\exp\big[K_2(t_n - t_M)\big] \tag{3.15}$$

and

$$U_n \in J_{\varepsilon,\delta} \quad \text{and} \quad \Delta t_n^q \leqslant (q+1)\tau/\delta\varepsilon, \tag{3.16}$$

for all τ sufficiently small.

Proof. We prove first that if (3.15) holds for $n = M, \dots, m$ then (3.16) holds for $n = M, \dots, m$. We have $E_n \leqslant K_3(\tau/\delta\varepsilon)^{s/q}$ so that by (3.4) we have, for τ sufficiently small, $E_n \leqslant d$. Since $u_m \in I$ it follows that $U_m \in J$. Also

$$\begin{aligned}\big\|f(U_n)\big\| &\geqslant \big\|f(u_n)\big\| - \big\|f(U_n) - f(u_n)\big\| \geqslant 2\delta - C_8\|U_n - u_n\| \\ &\geqslant 2\delta - C_8K_3\tau/\delta\varepsilon \geqslant \delta\end{aligned}$$

by (3.5), for τ sufficiently small. Also

$$\begin{aligned}\big\|B_1(U_n)\big\| &\geqslant \big\|B_1(u_n)\big\| - C_5\|U_n - u_n\| = \big\|b_1(u_n)\big\|\big\|f(u_n)\big\| - C_5\|U_n - u_n\| \\ &\geqslant \big\|b_1(u_n)\big\|\big\|f(U_n)\big\| - \big\|b_1(u_n)\big\|\big\|f(u_n) - f(U_n)\big\| - C_5\|U_n - u_n\| \\ &\geqslant 2\varepsilon\big\|f(U_n)\big\| - (C_3C_8 + C_5)\|U_n - u_n\|. \qquad (3.17)\end{aligned}$$

By (3.5) we may choose τ sufficiently small that

$$(C_3C_8 + C_5)\frac{K_3\tau}{\delta\varepsilon} \leqslant \delta\varepsilon.$$

Then

$$(C_3C_8 + C_5)\|U_n - u_n\| \leqslant \delta\varepsilon \leqslant \big\|f(U_n)\big\|\varepsilon$$

so that $\|B_1(U_n)\| \geqslant \varepsilon\|f(U_n)\|$. Thus $\|b_1(U_n)\| \geqslant \varepsilon$ as required. Hence $U_n \in J_{\varepsilon,\delta}$ for $n = M, \dots, m \geqslant M$ if (3.15) holds for $n = M, \dots, m$. It follows that, if (3.15) holds for $n = M, \dots, m$ then

$$\Delta t_n^q \leqslant (q+1)\tau/\delta\varepsilon, \quad n = M, \dots, m, \qquad (3.18)$$

by lemma 3.2.

Thus, to prove the lemma, it is sufficient to prove that if (3.15) holds for $n = M, \dots, m$ then it holds for $n = m + 1$, using (3.18). To this end note that

$$\begin{aligned}E_{m+1} &= \big\|S(u_m;\Delta t_m) - S^{(1)}(U_m;\Delta t_m)\big\| \\ &\leqslant \big\|S^{(1)}(u_m;\Delta t_m) - S^{(1)}(U_m;\Delta t_m)\big\| + \big\|S^{(1)}(u_m;\Delta t_m) - S(u_m;\Delta t_m)\big\| \\ &\leqslant (1 + K_2\Delta t_m)E_m + K_3\Delta t_m^{s+1}.\end{aligned}$$

This last line of the calculation follows from the Lipschitz constant and truncation error bounds for the Runge–Kutta method – see [14], theorem 4.6.7, for example; thus $K_2,\ K_3$ depend upon J. By (3.18) we have

$$\Delta t_m^s \leqslant \big[(q+1)\tau/\delta\varepsilon\big]^{s/q}.$$

Thus there are constants $K_1,\ K_2$ depending upon J such that

$$E_{m+1} \leqslant (1 + K_2\Delta t_m)E_m + K_1\Delta t_m(\tau/\delta\varepsilon)^{s/q}.$$

Since $1 + x \leqslant \mathrm{e}^x$ and $1 \leqslant \mathrm{e}^x$ for all $x \geqslant 0$ the inductive hypothesis gives

$$\begin{aligned}E_{m+1} &\leqslant \big[E_M + K_1(\tau/\delta\varepsilon)^{s/q}(t_m - t_M)\big]\exp\big[K_2(t_{m+1} - t_M)\big] \\ &\quad + K_1\Delta t_m(\tau/\delta\varepsilon)^{s/q}\exp\big[K_2(t_{m+1} - t_M)\big] \\ &= \big[E_M + K_1(\tau/\delta\varepsilon)^{s/q}(t_{m+1} - t_M)\big]\exp\big[K_2(t_{m+1} - t_M)\big].\end{aligned}$$

This completes the induction and (3.15) is established. □

Remarks. (i) If $\varepsilon > 0$ and $\delta > 0$ are fixed then, for τ sufficiently small and $u(t) \in I_{\varepsilon,\delta}$, it may be shown that no step-rejections occur so that $k = 0$ in (1.8). Lemma 3.4 is thus simply a refinement of Stetter's theory from [12] (where step-size rejection is ignored and the leading term in the local error is assumed to be bounded from zero), with the addition of maximum step-size and step-size ratios to the step-size selection mechanism. Estimating the probabilities that ε and δ take specified positive values leads to the probabilistic convergence results of section 5.

(ii) If the numerical solution enters $\Gamma(\delta)$ or $\Psi(\varepsilon)$ then the step-size selection mechanism may allow the time-step to increase to a value where the asymptotic considerations underlying the theory valid for $u(t) \in I_{\varepsilon,\delta}$ fail. It has proved impossible in this situation to obtain a convergence theory unless modifications (not used in practice) are made to the step-size selection mechanism. These modifications are given in assumptions 6.1 and 6.3 and lead to the deterministic convergence results of section 6.

(iii) Lemma 3.3 addresses a question raised in [11] which is this: do practical step-size selection mechanisms ensure that, for any given tolerance τ, any finite time T may be reached by the code in a finite number of steps? The answer is in the affirmative – under natural smoothness assumptions, sequences where the time-step decreases in a geometric fashion are not produced.

In the next two lemmas, and in sections 5 and 6, we will make the following assumption which implies that $\Psi(\varepsilon)$ and $\Gamma(\delta)$ are disjoint for sufficiently small ε and δ.

Assumption 3.5. There is a constant $\varepsilon_c > 0$ such that, for each $\varepsilon \in [0, \varepsilon_c)$, the set $\Psi(\varepsilon)$ is the disjoint union of a countable set of neighbourhoods $\{\Psi_i\}_{i=1}^{M}$ with $M \leqslant \infty$, each containing a point $z_i \in \mathbb{R}^m$ at which $b_1(z_i) = 0$. Thus, for each $\varepsilon \in [0, \varepsilon_c)$,

$$\Psi(\varepsilon) = \bigcup_{i=1}^{M} \Psi_i, \quad \Psi_i \cap \Psi_j = \emptyset, \ \forall i \neq j.$$

Furthermore, for any finite integer M_0 there are constants C_1 and C_2 such that, for all $\varepsilon \in [0, \varepsilon_c)$,

$$\Psi_i \subseteq B(z_i, C_1\varepsilon), \quad i = 1, \ldots, M_0, \qquad \min_{1 \leqslant i \leqslant M_0} \left\| f(z_i) \right\| \geqslant C_2.$$

Important remark. This assumption will hold for generic $f(u)$ within the class of sufficiently smooth functions. To understand that this is so in the linear case, for example, see section 5.1. To make such a genericity statement precise in the general case would require placing a probability measure on an appropriate space of functions in which f lies and showing that assumption 3.5 holds with probability one; to do this would complicate the paper unnecessarily and would not yield deeper insight. Thus we simply assume that assumption 3.5 holds. The simple reason why it does hold generically is that points where $B_1(u) = 0$ will typically be isolated, since $B_1 : \mathbb{R}^m \mapsto \mathbb{R}^m$ and such functions picked at random will have isolated zeros. (Note that $B_1(u)$ depends upon f and its derivatives so that picking f at random induces a

random choice of B_1.) Since the assumption 3.5 holds whenever the zeros of $B_1(u)$ are isolated, the intuition that the assumption is generic follows. □

In the following we take $M_0 < \infty$ to be sufficiently large that $\{z_i\}_{i=1}^{M_0}$ are the only equilibria contained in J.

Lemma 3.6. Let assumption 3.5 hold and assume that $u(t) \in \Psi(2\varepsilon) \cap I$ for $T_0 < t < T_R$. Then, for ε sufficiently small,

$$T_R - T_0 \leqslant 8C_1\varepsilon/C_2.$$

Proof. By assumption 3.5, $u(t) \in \Psi_i$, $u(t) \in B(z_i, 2C_1\varepsilon)$ for $T_0 < t < T_R$ so that

$$\|u(T_R) - u(T_0)\| \leqslant 4C_1\varepsilon.$$

Now

$$u(T_R) = u(T_0) + \int_{T_0}^{T_R} f\big(u(s)\big)\,\mathrm{d}s$$

which implies that

$$u(T_R) - u(T_0) = \int_{T_0}^{T_R} f(z_i)\,\mathrm{d}s + \int_{T_0}^{T_R} \big[f\big(u(s)\big) - f(z_i)\big]\,\mathrm{d}s.$$

But $\|f(z_i)\| \geqslant C_2$ and so

$$C_2|T_R - T_0| \leqslant \left\| \int_{T_0}^{T_R} f(z_i)\,\mathrm{d}s \right\| \leqslant 4C_1\varepsilon + 2|T_R - T_0|C_8C_1\varepsilon. \tag{3.19}$$

For ε sufficiently small the result follows. □

Lemma 3.7. Let assumption 3.5 hold, let $b > 0$ in (3.3), assume that $u(t) \in \Psi(2\varepsilon) \cap I$ for $T_0 < t < T_R$ and let $N = N(\tau)$, $Q = Q(\tau)$ be such that $T_0 < t_N < t_{Q-1} < T_R \leqslant t_Q$. If there is a constant K such that, for all τ sufficiently small,

$$E_N \leqslant K(\tau/\delta\varepsilon)^{s/q}, \qquad \Delta t_N^q \leqslant K\tau/\delta\varepsilon,$$

then there exists a constant $K_3 > 0$ such that, for all τ sufficiently small,

$$\Delta t_p \leqslant \Delta t_{\max,p} := \Delta t_N + (\alpha - 1)(t_p - t_N), \tag{3.20}$$

$$E_p \leqslant \big\{E_N + K_3\Delta t_{\max,p-1}^s(t_p - t_N)\big\} \exp\big[K_2(t_p - t_N)\big] \tag{3.21}$$

for $p = N, \ldots, Q$.

Proof. We have $\Delta t_i \geqslant 0,\ \alpha\Delta t_i \geqslant \Delta t_{i+1}$ for all $i \geqslant 0$, by (1.8), (1.10). Thus

$$\sum_{i=N}^{p-1} \Delta t_{i+1} \leqslant \alpha \sum_{i=N}^{p-1} \Delta t_i,$$

which implies that

$$\sum_{i=N}^{p-1} \Delta t_i + \Delta t_p - \Delta t_N \leqslant \alpha \sum_{i=N}^{p-1} \Delta t_i.$$

Hence (3.20) follows.

Now

$$\begin{aligned}(t_Q - t_N) &\leqslant (T_R - T_0) + (T_0 - t_N) + (t_Q - T_R) \leqslant (T_R - T_0) + \Delta t_{Q-1} \\ &\leqslant (T_R - T_0) + \Delta t_N + (\alpha - 1)(t_{Q-1} - t_N) \\ &\leqslant (T_R - T_0) + \Delta t_N + (\alpha - 1)(T_R - T_0) \leqslant \Delta t_N + \alpha(T_R - T_0).\end{aligned}$$

By choosing $K\tau \leqslant \delta\varepsilon^{q+1}$ (which is possible by (3.4)) we have $\Delta t_N^q \leqslant \varepsilon^q$ and so, also using lemma 3.6, it follows that there is a constant $K_4 > 0$:

$$(t_Q - t_N) \leqslant K_4\varepsilon. \tag{3.22}$$

Note that (3.21) holds for $p = N$. For induction assume that it holds for $p = N, \ldots, P < Q$. Then, by (3.4), (1.8), (1.10), (3.22) we have

$$E_p \leqslant [\varepsilon^s + K_3 D^s K_4 \varepsilon] \exp[K_2 K_4 \varepsilon], \quad p = N, \ldots, P < Q.$$

Hence, for τ (and hence ε as $b > 0$) sufficiently small,

$$\|U_p\| \leqslant \big\|u(t_p)\big\| + \|E_p\| \leqslant \big\|u(t_p)\big\| + d$$

so that $U_p \in J,\ p = N, \ldots, P$. Similarly to the proof of lemma 3.4, we thus have

$$E_{m+1} \leqslant (1 + K_2\Delta t_m)E_m + K_3\Delta t_m^{s+1}, \quad m = N, \ldots, P. \tag{3.23}$$

Applying theorem A in the appendix gives

$$E_{P+1} \leqslant \left[E_N + K_3 \sum_{j=N}^{P} \Delta t_j^{s+1}\right] \exp\left\{K_2 \sum_{j=N}^{P} \Delta t_j\right\}.$$

But, for $N \leqslant j \leqslant P$, equation (3.20) gives

$$\Delta t_j \leqslant \Delta t_N + (\alpha - 1)(t_j - t_N) \leqslant \Delta t_N + (\alpha - 1)(t_P - t_N) = \Delta t_{\max,P}.$$

Hence (3.21) holds for $p = P + 1$ and the induction is complete. □

Lemma 3.8. Let assumption 3.5 hold and let $a, b > 0$ be chosen so that $\delta = (\tau/\delta\varepsilon)^{s/q}$. Assume that $u(t) \in I_{2\varepsilon}$ for $T_L \leqslant t \leqslant T_0$ and let $M = M(\tau),\ N = N(\tau)$ be such that $T_L \leqslant t_M \leqslant t_{N-1} \leqslant T_0$. Assume that

$$\Delta t_M^{(0)} \leqslant \frac{\varepsilon q}{C_4(q+1)} \tag{3.24}$$

and also that

$$\Delta t_p \leqslant \frac{\varepsilon q}{\alpha C_4(q+1)} \tag{3.25}$$

if $\|f(U_p)\| \leqslant \delta,\ \|f(U_{p+1})\| > \delta$. Finally, assume there is a constant K such that, for all τ sufficiently small, $E_M \leqslant K\delta$. Then there exist $K_1, K_2 \geqslant 0$ such that, for $n = M, \ldots, N$,

$$E_n \leqslant \big[E_M + K_1\delta(t_n - t_M)\big] \exp\big[K_2(t_n - t_M)\big], \tag{3.26}$$

$$U_n \in J_\varepsilon, \tag{3.27}$$

$$\Delta t_n^q \leqslant (q+1)\tau/\delta\varepsilon \quad \text{if } \big\|f(U_n)\big\| > \delta, \tag{3.28}$$

for all τ sufficiently small.

Proof. We prove first that if (3.26) holds for $n = M, \ldots, m \leqslant N$ then (3.27), (3.28) hold for $n = M, \ldots, m$. We have $E_n \leqslant K_5\delta$, so that, by reducing τ sufficiently,

$$\|U_n\| \leqslant \big\|u(t_n)\big\| + K_5\delta \leqslant \big\|u(t_n)\big\| + d$$

and so $U_n \in J$. First assume that

$$\big\|f(U_n)\big\| \geqslant C_2/2.$$

Then, by the same argument that yields (3.17), we have

$$\big\|B_1(U_n)\big\| \geqslant 2\varepsilon\big\|f(U_n)\big\| - (C_3C_8 + C_5)\|U_n - u_n\|.$$

But

$$\|U_n - u_n\| \leqslant K_5\delta = K_5(\tau/\delta\varepsilon)^{s/q}.$$

By (3.4) it follows that, since $s \geqslant 1$,

$$\frac{\|U_n - u_n\|}{\varepsilon} \to 0 \quad \text{as } \tau \to 0$$

and hence, for τ sufficiently small,

$$\frac{(C_3C_8 + C_5)\|U_n - u_n\|}{\varepsilon} \leqslant \frac{C_2}{2} \leqslant \big\|f(U_n)\big\|$$

so that

$$\|b_1(U_n)\| = \frac{\|B_1(U_n)\|}{\|f(U_n)\|} \geqslant \varepsilon$$

and $U_n \in J_\varepsilon$. On the other hand, if

$$\|f(U_n)\| < C_2/2$$

it is not possible that $U_n \in \Psi(\varepsilon)$, for if it were then $U_n \in \Psi_i$ for some i and, by assumption 3.5,

$$\frac{C_2}{2} \geqslant \|f(U_n)\| \geqslant \|f(z_i)\| - \|f(U_n) - f(z_i)\| \geqslant \|f(z_i)\| - C_8 C_1 \varepsilon$$

so that $\|f(z_i)\| < C_2$ for ε sufficiently small. This contradicts assumption 3.5. Hence (3.27) follows from (3.26). That (3.28) follows from (3.26) may be seen from the assumption on Δt_p if $U_{p+1} \in J \backslash \Psi(\delta)$ and $U_p \in \Psi(\delta)$, followed by application of lemma 3.2, since (3.25) implies $\Delta t_{p+1}^{(0)} \leqslant \varepsilon q / C_4(q+1)$.

Since (3.27), (3.28) hold we have that

$$\begin{aligned} E_{m+1} &= \|S(u_m; \Delta t_m) - S^{(1)}(U_m; \Delta t_m)\| \\ &\leqslant \|S(u_m; \Delta t_m) - S(U_m; \Delta t_m)\| + \|S(U_m; \Delta t_m) - S^{(1)}(U_m; \Delta t_m)\|. \end{aligned}$$

By continuity of the semigroup $S(\cdot, t)$ and by lemma 2.5 we obtain

$$E_{m+1} \leqslant (1 + K_2 \Delta t_m) E_m + K_3 \|f(U_m)\| \Delta t_m^{s+1},$$

where, without loss of generality, we have chosen K_2, K_3 as given in the proof of lemma 3.4. Now, if $U_m \in \Gamma(\delta)$, then there is $K_6 > 0$ such that

$$K_3 \|f(U_m)\| \Delta t_m^{s+1} \leqslant K_6 \delta \Delta t_m,$$

whilst if $U_m \in J_\varepsilon \backslash \Gamma(\delta) = J_{\varepsilon,\delta}$ then (3.28) gives

$$K_3 \|f(U_m)\| \Delta t_m^{s+1} \leqslant K_6 (\tau / \delta \varepsilon)^{s/q} \Delta t_m = K_6 \delta \Delta t_m.$$

Hence

$$E_{m+1} \leqslant (1 + K_2 \Delta t_m) E_m + K_6 \delta \Delta t_m$$

and an inductive step as in the proof of lemma 3.4 gives (3.26) for $n = M, \ldots, m+1$. Since (3.27), (3.28) follow from (3.26) this completes the inductive proof. □

4. Basic convergence theorem and applications

We employ the results of section 3 to obtain a basic convergence theorem concerning the approximation of (1.1) by (1.7)–(1.10). In the following we set $T < \infty$ and

$$\eta = \eta(U,T) := \inf_{0 \leqslant t \leqslant T} \big\| f\big(S(U,t)\big) \big\|,$$

$$\zeta = \zeta(U,T) := \inf_{0 \leqslant t \leqslant T} \big\| b_1\big(S(U,t)\big) \big\|.$$

Let B denote a set with the property that there is a bounded set $I = I(B,T)$ such that

$$\bigcup_{U \in B} S(U,t) \subseteq I, \quad \forall t \in [0,T].$$

The basic theorem assumes that $\eta, \zeta > 0$. Note that, since $T < \infty$, $\eta > 0$ occurs automatically provided that U is not an equilibrium point; on the other hand, determining whether $\zeta > 0$ would appear to be very hard in general. For linear problems defined through an invertible matrix it will be shown to hold. Furthermore, we shall show in section 5 that, in a certain probabilistic sense, $\eta, \zeta > 0$ is very likely to occur. Thus the following theorem will eventually be of more use than might appear at first reading.

Theorem 4.1. Assume that $\|f(U)\| \neq 0$ and that $\zeta(U,T) > 0$. Then there are constants $K = K(B,T)$, $\tau_c = \tau_c(U,B,T)$ and $\gamma = \gamma(U,B,T)$ such that, for all such $U \in B$, the sequences $\{U_n\}$ and $\{\Delta t_n\}$ generated by the algorithm (1.7)–(1.10), together with the truncation error (1.12), satisfy

$$\big\| u(t_n) - U_n \big\| \leqslant K \left(\frac{\tau}{\eta \zeta} \right)^{s/q},$$

$$\min \big\{ C_{10} \tau, \Delta t_0^q \big\} \leqslant \Delta t_n^q \leqslant \frac{4(q+1)\tau}{\eta \zeta}$$

and

$$\big\| T(U_n; \Delta t_n) \big\| \leqslant K \left(\frac{\tau}{\eta \zeta} \right)^{s/q} \Delta t_n$$

for all $0 \leqslant t_n \leqslant T$, $\tau \in (0, \tau_c)$ provided that $\Delta t_{\text{init}} \leqslant \gamma$.

Proof. We apply lemma 3.4 with $a = b = 0$, $T_L = 0$, $T_0 = T$, $\delta = \eta/2$ and $\varepsilon = \zeta/2$. Note that $\zeta > 0$ by assumption and that $\eta > 0$ since $\|f(U)\| \neq 0$. The result on U_n and the upper bound on Δt_n follow, provided

$$\Delta t_0^{(0)} < \gamma = \frac{\zeta q}{2 C_4 (q+1)},$$

where C_4 depends on B through J and ζ depends on U and T. The bound on the truncation error follows from the estimate on the time-steps. The lower bound on Δt_n follows from lemma 3.3. □

We now consider various cases where theorem 4.1 applies directly. The first of these is the class of constant coefficient linear systems

$$u_t = Au, \quad u(0) = U, \tag{4.1}$$

where the $m \times m$ matrix A has entries $a_{ij} = \{A\}_{ij}$ satisfying

$$\|A\|_F^2 := \sum_{i,j=1}^{n} a_{ij}^2 = 1. \tag{4.2}$$

Such a normalization can always be achieved by scaling time. We use the theory of section 4 to prove results about this problem. The following lemma, concerning solutions of (4.1), will be useful in this regard:

Lemma 4.2. Let A be an invertible $m \times m$ matrix. All solutions of (4.1) subject to (4.2) satisfy

$$\eta(U,T) := \inf_{0 \leqslant t \leqslant T} \|Au(t)\| \geqslant \frac{\mathrm{e}^{-T}\|U\|}{\|A^{-1}\|}.$$

Proof. By (4.1)

$$Au_t = A^2 u.$$

Hence, using $\langle \cdot, \cdot \rangle$ to denote the inner product inducing the Euclidean norm $\| \cdot \|$, we have

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|Au\|^2 = \langle Au, Au_t \rangle = \langle Au, A^2 u \rangle \geqslant -\|Au\|.\|A^2 u\|.$$

But, since $\| \cdot \| \leqslant \| \cdot \|_F$, we have

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|Au\|^2 \geqslant -\|Au\|^2\|A\| \geqslant -\|A\|_F\|Au\|^2 = -\|Au\|^2.$$

Integrating gives

$$\|Au\| \geqslant \mathrm{e}^{-t}\|AU\|.$$

Also, $\|v\| \leqslant \|A^{-1}\| \cdot \|Av\|$ so that

$$\|Au\| \geqslant \mathrm{e}^{-t}\|U\|/\|A^{-1}\|.$$

The result follows. □

By lemma 2.3 we know that the truncation error on linear problems may be of higher order than on nonlinear problems. Henceforth in this section we assume that s and q_1 have been chosen to be the *effective* orders of the methods in question when applied to the class of linear problems of the form (4.1). Note that these orders will be greater than or equal to the actual orders when applied to the fully nonlinear problem (1.1). Similar considerations apply to q_2 and q.

Lemma 4.3. The functions $B_1(u)$ and $b_1(u)$ satisfy

$$B_1(u) = aA^{q_2+1}u, \qquad b_1(u) = aA^{q_2+1}u/\|Au\| \tag{4.3}$$

for some $a \neq 0$, when (1.7)–(1.10) is applied to (4.1). Hence, if A is non-singular,

$$\zeta := \inf_{0\leqslant t\leqslant T} \big\|b_1\big(u(t)\big)\big\| > 0$$

for all solutions of (4.1). Specifically

$$\zeta \geqslant \frac{|a|}{\|A^{-1}\|^{q_2}}.$$

Proof. From lemma 2.3, the definitions of $B_1(u)$ and $b_1(u)$ and from (1.6), the result (4.3) follows. Now

$$\|Au\| = \|A^{-q_2}A^{q_2+1}u\| \leqslant \|A^{-1}\|^{q_2}\|A^{q_2+1}u\|.$$

Hence

$$\big\|b_1(u)\big\| = |a|\frac{\|A^{q_2+1}u\|}{\|Au\|} \geqslant \frac{|a|}{\|A^{-1}\|^{q_2}}. \qquad \square$$

The following results are a consequence of theorem 4.1 and lemmas 4.2, 4.3.

Corollary 4.4. For method (1.7)–(1.10) applied to (4.1) with A invertible, the set $\Psi(\varepsilon)$ is empty for all ε sufficiently small. Hence assumption 3.5 is automatically satisfied.

Theorem 4.5. Assume that the set $B \subset \mathbb{R}^m$ is bounded and that A is invertible. There are constants $C = C(B,T),\ \tau_c = \tau_c(U,B,T)$ and $\gamma = \gamma(U,B,T)$ such that, for all $U \in B\backslash 0$ the algorithm (1.7)–(1.10) applied to (4.1) satisfies

$$\big\|u(t_n) - U_n\big\| \leqslant C\left(\frac{\|A^{-1}\|^{q_2+1}\tau}{\|U\|}\right)^{s/q},$$

for all $0 \leqslant t_n \leqslant T,\ \tau \in (0,\tau_c)$, provided $\Delta t_0^{(0)} \leqslant \gamma$.

A direct analysis of the linear problem would be interesting and might yield more information than our analysis here which is a corollary of a general nonlinear analysis.

There are some nonlinear problems where the specific choice of f and of the numerical method imply that $\zeta > 0$ for all initial data because $b_1(u)$ has no zeros. In such situations theorem 4.1 may be applied directly. As an illustration consider the following embedded Runge–Kutta pair:

$$\eta_1 = U, \qquad \eta_2 = U + tf(\eta_1),$$

$$S^{(1)}(U,t) = U + tf(\eta_1), \qquad S^{(2)}(U,t) = U + \frac{t}{2}\big[f(\eta_1) + f(\eta_2)\big].$$

From expansions in powers of t we see that

$$\begin{aligned} S^{(1)}(U,t) &= U + tf(U), \\ S^{(2)}(U,t) &= U + tf(U) + \frac{t^2}{2} df(U)f(U) + \mathcal{O}(t^3). \end{aligned} \tag{4.4}$$

It follows from (1.7) that

$$E(u,t) = \frac{t}{2} df(u)f(u) + \mathcal{O}(t^2).$$

Thus, by definition (2.2), (2.3) we see that

$$B_1(u) = df(u)f(u)$$

and that

$$b_1(u) = \frac{df(u)f(u)}{\|f(u)\|}.$$

With this expression we show that, for this particular error control method and some specific choices of vector fields, the set $\Psi(0)$ is empty. Consider the equations

$$\begin{aligned} x_t &= x + x(x^2 + y^2), \quad x(0) = X, \\ y_t &= y + y(x^2 + y^2), \quad y(0) = Y. \end{aligned} \tag{4.5}$$

Thus the Jacobian of the vector field $f(\cdot)$ governing the flow is

$$\begin{pmatrix} 1 + 3x^2 + y^2 & 2xy \\ 2xy & 1 + x^2 + 3y^2 \end{pmatrix}. \tag{4.6}$$

The determinant of this matrix is

$$(1 + 3x^2 + y^2)(1 + x^2 + 3y^2) - 4x^2y^2 = 1 + 4(x^2 + y^2) + 3(x^2 + y^2)^2,$$

which is clearly non-zero for all real x and y. Hence $b_1(u)$ is bounded away from zero for all $u \in \mathbb{R}^2$. Similar examples may be found in [3].

The preceding discussion shows that for certain adaptive algorithms of the form (1.7)–(1.10) applied to certain specific vector fields there are no points where $b_1(u) = 0$ so that theorem 4.1 may be applied directly. However, it is difficult in general to determine whether this situation holds for a given method and a given vector field. This is the motivation for the probabilistic considerations of the next section.

5. Probabilistic convergence results

Theorem 4.1 applies whenever η and ζ are bounded away from zero and gives convergence of the method (1.7)–(1.10). The values of η and ζ depend in practice upon the choice of initial data. In this section we take the viewpoint that it is therefore worthwhile to calculate the probability that η and ζ are bounded below by a certain amount, with respect to random choice of the initial data.

Let $B(0;R)$ be an open ball in $\mathbb{R}^m$ centred at 0 with radius R. We consider the problem (1.1), and its numerical approximation (1.7)–(1.10), with initial data U chosen at random uniformly on $B(0;R)$. Thus we have a probability triplet $(B(0;R), \mathcal{B}^m, \mathcal{L}^m)$ where $\mathcal{B}^m$ are the Borel sets of $B(0;R)$ and $\mathcal{L}^m$ is m-dimensional Lebesgue measure restricted to $B(0,R)$ and normalized to have total mass 1 on $B(0,R)$. We will use $P(\cdot)$ to denote the probability of an event. For simplicity of notation we will also let *Vol*$(\cdot)$ denote the m-dimensional Lebesgue measure of a set in $\mathbb{R}^m$ since the measure will essentially be volume whenever we use it.

Fix $T > 0$. For each numerical approximation of (1.1) let N be any integer such that $t_N \leqslant T$. Given $\omega \in B(0;R)$ we may define the family of random variables $X(\cdot, \Delta t_{\text{init}})$ by

$$X(\omega; \Delta t_{\text{init}}) := \limsup_{\tau \to 0} \sup_{N:\ 0 \leqslant t_N \leqslant T} \frac{\|u(t_N) - U_N\|}{\tau^{s/q}};$$

thus the initial guess for the time-step, Δt_{init}, yields the family of random variables whose union form a stochastic process. We now consider the following family of events, parameterized by ε:

$$Y_\varepsilon = \left\{\omega \in B(0;R) \mid \exists \Delta t_c > 0:\ X(\omega; \Delta t_{\text{init}}) \leqslant \varepsilon^{-1},\ \forall \Delta t_{\text{init}} \in (0, \Delta t_c]\right\}.$$

If the event Y_ε occurs then the numerical error behaves like

$$\left\|u(t_N) - U_N\right\| \leqslant \frac{2\tau^{s/q}}{\varepsilon}, \quad \forall N:\ 0 \leqslant t_N \leqslant T, \tag{5.1}$$

for Δt_{init} and τ sufficiently small, and the method converges at the rate to be expected from the small time-step heuristics underlying it. It is hence of interest to calculate the probability of Y_ε given the random choice of initial data $\omega \in B(0;R)$. In this section we prove the following theorem. The reader is encouraged to study the *Important remark* following assumption 3.5 concerning its genericity. Note also that the zeros of f are hyperbolic for generic $f(u)$; in the proof, however, hyperbolicity can be replaced by the weaker assumption that the set $\Gamma(\delta)$ is contained in the disjoint union of non-intersecting balls of radius $\mathcal{O}(\delta)$, for all δ sufficiently small.

Theorem 5.1. Let assumption 3.5 hold, assume that all zeros of f are hyperbolic, and let $m > 1$. Consider the approximation of (1.1) by the numerical method (1.7)–(1.10)

with U chosen at random uniformly (with respect to Lebesgue measure) in $B(0;R)$. Then there is $\varepsilon_c > 0$ and $C = C(T) > 0$ such that, for any $\varepsilon \in (0, \varepsilon_c)$,

$$P\{Y_\varepsilon\} \geqslant 1 - C\varepsilon^l,$$

where

$$l = \frac{(m-1)mq}{(2m-1)s}.$$

Hence, with probability 1, there is $\varepsilon > 0$ such that the numerical method converges as predicted by (5.1) as $\tau \to 0$, for all Δt_{init} sufficiently small.

Actually the proof shows that the set of initial conditions of small measure for which a global error bound of the form $\tau^{s/q}/\varepsilon$ cannot be obtained is contained in a finite number of sets with volumes of size $\mathcal{O}(\varepsilon^l)$.

The idea behind the proof of this theorem is simple: from theorem 4.1 we need to estimate the probability that $(\eta\zeta)^{s/q}$ is greater than $\mathcal{O}(\varepsilon)$. Now $\|f(u)\|$ and $\|b_1(u)\|$ are only small in small neighbourhoods of the points where $f(u)$ and $b_1(u)$ are zero and by hypothesis these points are isolated. By the definition of η and ζ we see that these quantities are only small when the solution enters the small neighbourhoods of the points where f and b_1 are zero. Thus proving the theorem boils down to estimating the probabilities of entering these small neighbourhoods and then putting the information together to obtain the required probability. Estimating the probability that a randomly chosen set of initial data points enters a small neighbourhood during the time T forward evolution of a semigroup is the same as estimating the probability that the backward time T evolution of the small neighbourhood contains the set of initial data points. Lemma 5.2 is fundamental in this regard; we prove this lemma and then use it to obtain the proof of theorem 5.1.

For the purposes of this section we define the action of the group $S(\cdot, t)$ on sets by defining, for a set $B \subseteq \mathbb{R}^m$,

$$S(B,t) = \bigcup_{U \in B} S(U,t).$$

We also define the diameter of a set B by

$$\mathit{diam}(B) = \sup_{x,y \in B} \|x - y\|.$$

Lemma 5.2. Given $x \in \mathbb{R}^m$ and $\varepsilon > 0$, let $B = B(x, \varepsilon/2)$ and define

$$E(t) = \bigcup_{0 \leqslant \tau \leqslant t} S(B,\tau).$$

Then there are constants $\varepsilon_c > 0$ and $C = C(x,T) > 0$ such that, for all $\varepsilon \in (0, \varepsilon_c)$,

$$\mathit{Vol}\big\{E(t)\big\} \leqslant C\Big[\sup_{0 \leqslant t \leqslant T} \big\|f\big(S(x,t)\big)\big\| + \varepsilon\Big]\varepsilon^{m-1}, \quad \forall t \in [0,T].$$

Proof. We proceed by approximation. Let $\Delta t \ll 1$ and define

$$B_n = S(B, n\Delta t), \qquad E_n = \bigcup_{0 \leqslant m \leqslant n} B_m.$$

By continuity of $S(U, \cdot)$ it follows that, for any fixed $t \in \mathbb{R}^+$,

$$\lim_{N \to \infty,\ N\Delta t = t} \big| \mathit{Vol}\{E(t)\} - \mathit{Vol}\{E_N\} \big| = 0. \tag{5.2}$$

Now $E_n = E_{n-1} \cup B_n$ and $B_{n-1} \subseteq E_{n-1}$ so that

$$\begin{aligned} \mathit{Vol}\{E_n\} &= \mathit{Vol}\{E_{n-1}\} + \mathit{Vol}\{B_n \backslash [B_n \cap E_{n-1}]\} \\ &\leqslant \mathit{Vol}\{E_{n-1}\} + \mathit{Vol}\{B_n \backslash [B_n \cap B_{n-1}]\}. \end{aligned} \tag{5.3}$$

Our objective is to estimate the last term in this inequality and iterate to prove the desired result.

We first show that there exists $C_1 = C_1(T)$ such that

$$\mathit{diam}(B_j) \leqslant C_1 \varepsilon, \quad \forall j\colon\ 0 \leqslant j\Delta t \leqslant T. \tag{5.4}$$

To see that this is true, let $u,\ v$ be two arbitrary points in the set $S(B, t)$ for some $t \in [0, T]$. Thus there are points $\tilde{u}, \tilde{v} \in B$ such that $u = S(\tilde{u}, t),\ v = S(\tilde{v}, t)$ and, by continuity of $S(\cdot, t)$, it follows that there exists $C_1 = C_1(T)$ such that

$$\|u - v\| \leqslant C_1(T) \|\tilde{u} - \tilde{v}\|.$$

But $\|\tilde{u} - \tilde{v}\| \leqslant \mathit{diam}(B) = \varepsilon$ and, since $u,\ v$ are arbitrary in $S(B, t)$, it follows that

$$\mathit{diam}\big(S(B, t)\big) \leqslant C_1(T)\varepsilon.$$

The result (5.4) follows.

Now note that $B_n = S(B_{n-1}, \Delta t)$ so that, if $y \in B_n$ then there exists $z \in B_{n-1}$ such that

$$y = z + \int_0^{\Delta t} f\big(S(z, \tau)\big)\, \mathrm{d}\tau.$$

Since $S(z, \tau) = z + \mathcal{O}(\tau)$ it follows that

$$y = z + \int_0^{\Delta t} f(z)\, \mathrm{d}\tau + \mathcal{O}(\Delta t^2).$$

By (5.4) we know that

$$\big\| z - S\big(x, (n-1)\Delta t\big) \big\| \leqslant C_1 \varepsilon$$

and hence that

$$\begin{aligned}\|y-z\| &\leqslant \Delta t\big\|f(z)\big\| + \mathcal{O}(\Delta t^2)\\ &\leqslant \big[\Delta t\big\|f\big(S\big(x,(n-1)\Delta t\big)\big)\big\| + LC_1\Delta t\varepsilon\big] + \mathcal{O}(\Delta t^2),\end{aligned}$$

where L is the Lipschitz constant for f. Since y is an arbitrary point in B_n and since $z \in B_{n-1}$ we have

$$B_n \subseteq \overline{\mathcal{N}}(B_{n-1}, \delta\Delta t), \tag{5.5}$$

where

$$\delta = \sup_{0\leqslant t\leqslant T} \big\|f(S(x,t)\big\| + LC_1\varepsilon + \mathcal{O}(\Delta t). \tag{5.6}$$

Now, by the group property of $S(\cdot,t)$ we deduce that

$$\overline{\mathcal{N}}(B_{n-1},\delta\Delta t)\backslash B_{n-1} = S\big(B^*,(n-1)\Delta t\big)$$

for some set $B^* \subset \mathbb{R}^m$.

We now prove that there exists $C_2 = C_2(T)$ such that

$$B^* \subseteq \overline{\mathcal{N}}(B, C_2\delta\Delta t)\backslash B. \tag{5.7}$$

Let $a \in B^*$. Then let $b = S(a,(n-1)\Delta t) \in \overline{\mathcal{N}}(B_{n-1},\delta\Delta t)\backslash B_{n-1}$. Hence there exists $c \in B$ with $d = S(c,(n-1)\Delta t) \in B_{n-1}$ and $\|b-d\| \leqslant \delta\Delta t$. By continuity of $S(\cdot,t)$ in negative t it follows that there exists $C_2 = C_2(T)$ such that $\|a-c\| \leqslant C_2\delta\Delta t$. Thus, since a is arbitrary, (5.7) holds. Now let

$$\mathcal{V}(t) = \mathit{Vol}\big\{S(B^*,t)\big\}.$$

Since there exists $\kappa > 0$ such that

$$\big|\nabla.f(u)\big| \leqslant \kappa, \quad \forall u \in \mathbb{R}^m,$$

we have

$$\mathcal{V}(t) \leqslant \mathrm{e}^{\kappa t}\mathcal{V}(0).$$

But, since B is a ball in $\mathbb{R}^m$, the volume of a neighbourhood can be calculated as the product of the surface area and the width of the neighbourhood, and (5.7) gives

$$\mathcal{V}(0) \leqslant \mathit{Vol}\big\{\mathcal{N}(B, C_2\delta\Delta t)\backslash B\big\} = C_4\varepsilon^{m-1}\delta\Delta t,$$

where $C_4 = C_4(T)$. Thus there exists $C_5 = C_5(T)$ such that

$$\mathit{Vol}\big\{\mathcal{N}(B_{n-1},\delta\Delta t)\backslash B_{n-1}\big\} = \mathcal{V}\big((n-1)\Delta t\big) \leqslant C_5\varepsilon^{m-1}\delta\Delta t.$$

By (5.3) and (5.5) we have

$$\begin{aligned}\mathit{Vol}\{E_n\} &\leqslant \mathit{Vol}\{E_{n-1}\} + \mathit{Vol}\big\{\mathcal{N}(B_{n-1},\delta\Delta t)\backslash B_{n-1}\big\}\\ &\leqslant \mathit{Vol}\{E_{n-1}\} + C_5\varepsilon^{m-1}\delta\Delta t.\end{aligned}$$

Thus, for $N\Delta t \leqslant T$, we have

$$Vol\{E_N\} \leqslant Vol\{E_0\} + TC_5(T)\varepsilon^{m-1}\delta.$$

Using the expression (5.6) for δ and taking the limit $\Delta t \to 0,\ N \to \infty$ with $N\Delta t \leqslant T$ gives the required result, after noting that $E_0 = B$ and hence has volume of $\mathcal{O}(\varepsilon^m)$. □

We now prove the probabilistic convergence theorem:

Proof of theorem 5.1. In the following we let η denote the random variable $\eta(\cdot, T)$ and ζ denote the random variable $\zeta(\cdot, T)$. Note that $\eta,\ \zeta$ are dependent positive-valued scalar random variables. We define the set

$$A_\varepsilon := \left\{ \omega \in B(0;R) \;\middle|\; \frac{K}{(\eta\zeta)^{s/q}} \leqslant \varepsilon^{-1} \right\},$$

where $K = K(B(0;R), T)$ is given by theorem 4.1. Thus

$$P\{Y_\varepsilon\} \geqslant P\{A_\varepsilon\}. \tag{5.8}$$

We see that

$$P\{A_\varepsilon\} = 1 - P\big\{\eta\zeta < (K\varepsilon)^{q/s}\big\}. \tag{5.9}$$

By the law of total probability, for any $p \in (0,1)$, we have

$$\begin{aligned} P\big\{\eta\zeta < (K\varepsilon)^{q/s}\big\} = {} & P\big\{\eta\zeta < (K\varepsilon)^{q/s} \mid \eta \geqslant (K\varepsilon)^{pq/s}\big\} P\big\{\eta \geqslant (K\varepsilon)^{pq/s}\big\} \\ & + P\big\{\eta\zeta < (K\varepsilon)^{q/s} \mid \eta < (K\varepsilon)^{pq/s}\big\} P\big\{\eta < (K\varepsilon)^{pq/s}\big\}. \end{aligned}$$

Let

$$\delta_1 = (K\varepsilon)^{(1-p)q/s}, \qquad \delta_2 = (K\varepsilon)^{pq/s}.$$

Then

$$P\big\{\eta\zeta < (K\varepsilon)^{q/s}\big\} \leqslant P\big\{\zeta < \delta_1 \mid \eta \geqslant \delta_2\big\} + P\big\{\eta < \delta_2\big\}.$$

Thus, since

$$P(A \mid B) = P(A \cap B)/P(B) \leqslant P(A)/P(B)$$

for any events A and B,

$$P\big\{\eta\zeta < (K\varepsilon)^{q/s}\big\} \leqslant \frac{P\{\zeta < \delta_1\}}{P\{\eta \geqslant \delta_2\}} + P\{\eta < \delta_2\},$$

so that

$$P\big\{\eta\zeta < (K\varepsilon)^{q/s}\big\} \leqslant \frac{P\{\zeta < \delta_1\}}{1 - P\{\eta < \delta_2\}} + P\{\eta < \delta_2\}. \tag{5.10}$$

Given a time T and randomly chosen initial condition U we have that

$$P\{\zeta < \delta_1\} = \frac{Vol\{(\bigcup_{-T \leqslant t \leqslant 0} S(\Psi(\delta_1), t)) \cap B(0;R)\}}{Vol\{B(0;R)\}} \tag{5.11}$$

and

$$P\{\eta < \delta_2\} = \frac{Vol\{(\bigcup_{-T\leqslant t\leqslant 0} S(\Gamma(\delta_2), t)) \cap B(0; R)\}}{Vol\{B(0; R)\}}. \tag{5.12}$$

(This follows from the argument that the probability that a given set enters a particular small neighbourhood under forward evolution is equal to the probability that the small neighbourhood contains that set under backward evolution.) Note that, by assumption 3.5 and the fact that all zeros of f are hyperbolic, $\Psi(\delta_1)$ and $\Gamma(\delta_2)$ comprise disjoint components contained in balls of radius δ_1 and δ_2 respectively. Note also that each disjoint component of $\Gamma(\delta_2)$ contains a point x: $f(S(x,t)) = 0$ for all t. By lemma 5.2, and equations (5.11), (5.12), we deduce that there are constants C_i, $i = 1, \ldots, 4$, such that

$$P\{\zeta < (K\varepsilon)^{(1-p)q/s}\} \leqslant C_1\delta_1^{m-1} = C_2\varepsilon^{(1-p)(m-1)q/s}$$

and

$$P\{\eta < (K\varepsilon)^{pq/s}\} \leqslant C_3\delta_2^m = C_4\varepsilon^{pmq/s}.$$

We choose p to balance these two terms. Thus with $p = (m-1)/(2m-1)$ we see from (5.9) and (5.10) that, for all ε sufficiently small,

$$P\{Y_\varepsilon\} \geqslant 1 - C\varepsilon^l,$$

where l is given in the statement of the theorem. This completes the proof. □

The fact that the error estimate depends crucially upon the dimension being greater than one is not simply a product of the analysis. The numerical experiments of [3] indicate that in dimension one the error can behave badly for a set of initial data of positive measure. Furthermore, the work of [1] shows that in dimension one spurious steady solutions can be produced by adaptive algorithms for time integration whilst in dimension greater than one this is extremely unlikely.

6. Modifications of the basic algorithm and deterministic convergence results

Recall that theorem 4.1 is useful in all but a small set of exceptional cases as made precise by theorem 5.1. In this section we try to modify the algorithm (1.7)–(1.10) so that these exceptional cases can be incorporated into the analysis to obtain convergence results.

We start by trying to eliminate the assumption on ζ made in theorem 4.1. This can be done at the expense of making assumption 3.5 about the set $\Psi(\varepsilon)$, making the assumption that $\alpha \to 1$ as $\tau \to 0$ and accepting a reduced rate of convergence in τ. Read the *Important remark* following assumption 3.5 on the genericity of the assumptions.

In the following, define

$$\beta = (1 + q + q/s)^{-1}(s+1)$$

and note that $\beta \leqslant s/q$ as $s \geqslant 1$.

Assumption 6.1. The constant $\alpha > 1$ appearing in (1.10) satisfies $\alpha \to 1$ as $\tau \to 0$.

This assumption is perhaps a reasonable one to make whilst the code is in its asymptotic regime $\tau \to 0$ since the time-steps should not be allowed to change much from step to step once τ is sufficiently small. However, codes used in practice do not typically satisfy this constraint.

Theorem 6.2. Let assumptions 3.5 and 6.1 hold and assume that $\|f(U)\| \neq 0$. Then there are constants $K = K(B,T)$, $\tau_c = \tau_c(U,B,T)$ and $\gamma = \gamma(B,T)$ such that, for all such $U \in B$, the algorithm (1.7)–(1.10) and truncation error (1.12) satisfy

$$\big\|u(t_n) - U_n\big\| \leqslant K\tau^{\beta}/\eta^{s/q}$$

and

$$\big\|T(U_n, \Delta t_n)\big\| \leqslant K\left(\frac{\tau^{\beta s/(s+1)}}{\eta^{s/q}}\right)\Delta t_n$$

for all $0 \leqslant t_n \leqslant T$, $\tau \in (0, \tau_c)$ provided that $\Delta t_0^{(0)} \leqslant \gamma\tau^{\beta/s}$.

Proof. Here $a = 0$, $b > 0$ in (3.3) and we let $\delta = \eta$. We consider the two cases $u(t) \in I_{2\varepsilon,2\delta}$ and $u(t) \in \Psi(2\varepsilon) \cap I$. We then apply lemmas 3.4 and 3.7 respectively in these two regimes. In order to balance the error from each lemma we set

$$\tau = \varepsilon^{1+(s+1)q/s} = \varepsilon^{q+1+q/s}.$$

Then

$$(\tau/\varepsilon)^{s/q} = \varepsilon^{s+1} \quad \text{and} \quad \tau^{\beta q/s} = \frac{\tau}{\varepsilon}. \tag{6.1}$$

Also

$$\tau/\varepsilon^{q+1} = \varepsilon^{q/s} \to 0 \quad \text{as } \tau \to 0 \tag{6.2}$$

so that, since δ is fixed, (3.4) holds and the theory of section 3 applies. The time-step predicted by (3.20) of lemma 3.7 then satisfies, for some $\theta = \theta(\tau) > 0$ satisfying $\theta \to 0$ as $\tau \to 0$,

$$\Delta t_p \leqslant \theta\varepsilon, \quad p = N, \ldots, Q, \tag{6.3}$$

by virtue of (3.22), (6.2) and assumption 6.2. Thus the error predicted by lemma 3.7 is $\mathcal{O}(\varepsilon^{s+1})$ which, by (6.1), balances the error predicted by lemma 3.4, since δ is fixed independently of τ. The error is thus of $\mathcal{O}(\tau^{\beta}/\delta^{s/q})$ and $\mathcal{O}(\tau^{\beta})$ in lemmas 3.4 and 3.7, respectively.

The solution may alternately pass through $I_{2\varepsilon,2\delta}$ and $\Psi(2\varepsilon)$ any number of times on the interval $[0,T]$. In order to apply the two lemmas in this way we need to show that the requisite conditions on the time-step are satisfied. Under the assumption on $\Delta t_0^{(0)}$ we deduce that

$$\Delta t_0^q \leqslant \gamma^q \frac{\tau}{\varepsilon}$$

by (6.1) so that the initial condition required on Δt_0 for lemma 3.7 is satisfied if $U \in \Psi(2\varepsilon)$, by appropriate choice of γ, depending upon B and T through J. The assumption on $\Delta t_0^{(0)}$ also shows that

$$\Delta t_0^{(0)} \leqslant \varepsilon q C_4 (q+1)$$

by virtue of (6.2) so that lemma 3.4 may be applied if $U \in I_{2\varepsilon,2\delta}$.

It remains to show that we can pass from $I_{2\varepsilon,2\delta}$ to $\Psi(2\varepsilon)$ applying lemmas 3.4 and 3.7 respectively. After exiting $I_{2\varepsilon,2\delta}$, lemma 3.4 gives $\Delta t_N^q \leqslant (q+1)\tau/\delta\varepsilon$ so that lemma 3.7 applies. Aftex exiting $\Psi(2\varepsilon)$, lemma 3.7 shows that, by (6.3), τ sufficiently small ensures that

$$\Delta t_Q^{(0)} \leqslant \alpha \Delta t_{Q-1} \leqslant \varepsilon q / C_4 (q+1)$$

so that lemma 3.4 may be applied.

This completes the proof except for the bound on the truncation error. This last step follows from the bound on the time-steps proved during the course of the lemmas. The slight loss of accuracy as compared with the global error occurs because, when estimating the global error, (3.22) is used to further reduce the error. □

Finally, we try to eliminate the dependence of theorem 6.2 on η and the assumption that $\|f(U)\| \neq 0$. This can be achieved at the expense of an additional assumption on the method and still further reduction in the rate of convergence. Let

$$\sigma = (s+q+2+q/s)^{-1}(s+1). \tag{6.4}$$

The extra assumption is now given. The basic point here is that, in the neighbourhood of an equilibrium point, the time-step can be very large and errors still small. Assumption 6.3 ensures that large choices of the time-step do not persist after the neighbourhood of the equilibrium point is left behind.

Assumption 6.3. The minimal non-negative integer k in (1.8) is chosen so that if $\|f(U_n)\| \leqslant \tau^\sigma$, $\|f(U_{n+1})\| > \tau^\sigma$ then, in addition to (1.9) holding,

$$\Delta t_n \leqslant h(\tau) \tau^{\sigma/(s+1)},$$

where $h(\tau) \to 0$ as $\tau \to 0$.

We may now prove:

Theorem 6.4. Let assumptions 3.5, 6.1 and 6.3 hold. Then there are constants $K = K(B,T)$, $\tau_c = \tau_c(B,T)$ and $\gamma = \gamma(B)$ such that, for all $U \in B$, the algorithm (1.7)–(1.10) and the truncation error (1.12) satisfy

$$\big\|u(t_n) - U_n\big\| \leqslant K\tau^\sigma$$

and

$$\big\|T(U_n, \Delta t_n)\big\| \leqslant K\tau^{\sigma s/(s+1)} \Delta t_n$$

for all $0 \leqslant t_n \leqslant T$, $\tau \in (0, \tau_c)$, provided that $\Delta t_0^{(0)} \leqslant \gamma \tau^{\sigma/s}$.

Proof. Here $a, b > 0$ in (3.3) and we consider $u(t) \in I_{2\varepsilon}$ and $u(t) \in \Psi(2\varepsilon) \cap I$. In order to balance the errors from lemmas 3.7 and 3.8 we set

$$\left(\frac{\tau}{\delta\varepsilon}\right)^{s/q} = \delta = \varepsilon^{s+1},$$

giving $\tau = \varepsilon^{s+q+2+q/s},\ \delta = \varepsilon^{s+1}$ so that

$$\left(\frac{\tau}{\delta\varepsilon}\right)^{s/q} = \tau^{\sigma} \tag{6.5}$$

and

$$\frac{\tau}{\delta\varepsilon^{q+1}} = \varepsilon^{q/s} \to 0 \quad \text{as } \tau \to 0. \tag{6.6}$$

Thus (3.4) holds and the theory of section 3 may be applied. As in the proof of theorem 6.2, the time-step predicted by lemma 3.7 thus satisfies (6.3) for some $\theta > 0$ satisfying $\theta \to 0$ as $\tau \to 0$. Thus the error predicted by lemma 3.7 is $\mathcal{O}(\varepsilon^{s+1}) = \mathcal{O}(\tau^{\sigma})$, balancing the error predicted by lemma 3.8.

By assumption we have that

$$\Delta t_0 \leqslant \Delta t_0^{(0)} \leqslant \gamma \left(\frac{\tau}{\delta\varepsilon}\right)^{1/q}$$

so that lemma 3.7 applies if $U \in \Psi(2\varepsilon)$. Also, by (6.6),

$$\Delta t_0^{(0)} \leqslant \frac{\varepsilon q}{C_4(q+1)}$$

for τ sufficiently small so that lemma 3.8 applies if $U \in I_{2\varepsilon}$. After exiting $I_{2\varepsilon}$ lemma 3.8 gives $\Delta t_N^q \leqslant (q+1)\tau/\delta\varepsilon$ so that lemma 3.7 applies. After exiting $\Psi(2\varepsilon)$, lemma 3.7 shows that, by (6.3), for all τ sufficiently small,

$$\Delta t_Q^{(0)} \leqslant \alpha \Delta t_{Q-1} \leqslant \varepsilon q / C_4(q+1)$$

so that lemma 3.8 may be applied. This completes the proof. □

As an illustration we apply theorem 6.4 to the study of linear problems with A invertible so that assumption 3.5 holds automatically by corollary 4.4. We obtain the following:

Theorem 6.5. Assume that B is bounded and that A is invertible and that, in addition, the algorithm satisfies assumptions 6.1 and 6.3. Then there are constants $C = C(B,T)$, $\tau^* = \tau^*(B,T)$ and $\gamma^* = \gamma^*(B)$ such that, for all $U \in B$, the solution $u(t)$ of the linear equation (4.1), and its numerical approximation by the algorithm (1.7)–(1.10), satisfy

$$\left\|u(t_n) - U_n\right\| \leqslant C\tau^{\sigma}$$

for all $0 \leqslant t_n \leqslant T,\ \tau \in (0, \tau_c)$, provided $\Delta t_0^{(0)} \leqslant \gamma^* \tau^{\sigma/s}$.

Appendix

Theorem A. Consider sequences $\{\Delta t_n\}_{n=M}^{N-1}$, $\{P_n\}_{n=M}^{N-1}$ and $\{G_n\}_{n=M}^{N-1}$ satisfying

$$\sum_{n=M}^{N-1} \Delta t_n = r, \qquad \sum_{n=M}^{N-1} \Delta t_n G_n = a_1(r), \qquad \sum_{n=M}^{N-1} \Delta t_n P_n = a_2(r)$$

and

$$\Delta t_n, P_n, G_n \geqslant 0, \quad n = M, \ldots, N-1.$$

If the sequence $\{E_n\}_{n=M}^{N}$ satisfies

$$E_{n+1} \leqslant (1 + \Delta t_n G_n) E_n + \Delta t_n P_n, \quad n = M, \ldots, N-1,$$

then

$$E_N \leqslant \big[E_M + a_2(r)\big] \exp\big\{a_1(r)\big\}, \quad n = M, \ldots, N.$$

Proof. Define $\{Q_n\}_{n=0}^{N}$ by

$$Q_{n+1} = (1 + \Delta t_n G_n)^{-1} Q_n, \quad Q_M = 1.$$

Then

$$E_{n+1} \leqslant \frac{Q_n E_n}{Q_{n+1}} + \Delta t_n P_n$$

so that

$$Q_{n+1} E_{n+1} \leqslant Q_n E_n + \Delta t_n P_n Q_{n+1}.$$

Hence

$$Q_N E_N \leqslant Q_M E_M + \sum_{n=M}^{N-1} \Delta t_n P_n Q_{n+1},$$

which implies that

$$Q_N E_N \leqslant E_M + \sum_{n=M}^{N-1} \Delta t_n P_n = E_M + a_2(r). \tag{A1}$$

But

$$Q_N^{-1} = \prod_{n=M}^{N-1} (1 + \Delta t_n G_n)$$

so that, noting that $1 + x \leqslant \mathrm{e}^x$ for all positive x, we have

$$Q_N^{-1} \leqslant \prod_{n=M}^{N-1} \exp(\Delta t_n G_n) = \exp\left(\sum_{n=M}^{N-1} \Delta t_n G_n \right) = \exp\big(a_1(r)\big). \tag{A2}$$

Combining (A1) and (A2) gives the desired result. □

Acknowledgements

The author is grateful to Des Higham, Harbir Lamba and to several participants in the Georgia Tech Conference on Dynamical Numerical Analysis, for a number of useful suggestions.

References

[1] M. A. Aves, D. F. Griffiths and D. J. Higham, Does error control suppress spuriosity?, to appear in SIAM J. Num. Anal.
[2] J. C. Butcher, *The Numerical Analysis of Ordinary Differential Equations* (Wiley, New York, 1992).
[3] M. Calvo, D. J. Higham, J. I. Montijano and L. Randez, Stepsize selection for tolerance proportionality in explicit Runge–Kutta codes, Advances in Computational Mathematics, to appear.
[4] J. W. Demmel, On condition numbers and the distance to the nearest ill-posed problem, Num. Math. 51 (1987) 251–289.
[5] A. Edelman, On the Distribution of a scaled condition number, Math. Comp. 58 (1992) 185–190.
[6] G. Hall, Equilibrium states of Runge–Kutta schemes, ACM Trans. on Math. Software 11 (1985) 289–301.
[7] D. J. Higham, Global error versus tolerance for explicit Runge–Kutta methods, IMA J. Numer. Anal. 11 (1991) 457–480
[8] R. Schrieber and L. N. Trefethen, Average-case stability of gaussian elimination, SIAM J. Matrix Anal. 11 (1990) 335–360.
[9] L. F. Shampine, Tolerance proportionality in ODE codes, in: *Numerical Methods for Ordinary Differential Equations* (*Proceedings*), eds. A. Bellen, C. Gear and E. Russo, Lecture Notes in Mathematics 1386 (Springer, Berlin, 1987) pp. 118–136.
[10] S. Smale, The fundamental theorem of algebra and complexity theory, Bull. Amer. Math. Soc. 4 (1981) 1–35.
[11] H. J. Stetter, Considerations concerning s theory for ODE-solvers, in: *Numerical Treatment of Differential Equations*, eds. R. Burlisch, R. Grigorieff and J. Schröder, Lecture Notes in Mathematics 631 (Springer, Berlin, 1976).
[12] H. J. Stetter, Tolerance proportionality in ODE-codes, in: *Proc. Second Conf. on Numerical Treatment of Ordinary Differential Equations*, ed. R. März, Seminarberichte 32, Humboldt University, Berlin (1980).
[13] D. Stoffer and K. Nipp, Invariant curves for variable step-size integrators, BIT 31 (1991) 169–180 and BIT 32 (1992) 367–368.
[14] A. M. Stuart and A. R. Humphries, *Dynamical Systems and Numerical Analysis* (Cambridge Univ. Press, 1996).
[15] M.-C. Yeung and T. F. Chan, Probabilistic analysis of Gaussian elimination without pivoting, Preprint, UCLA (1995).