SIAM J. NUMER. ANAL.
Vol. 25, No. 4, August 1988





# ERROR ESTIMATES AND ADAPTIVE TIME-STEP CONTROL FOR A CLASS OF ONE-STEP METHODS FOR STIFF ORDINARY DIFFERENTIAL EQUATIONS*

CLAES JOHNSON†

**Abstract.** We prove new optimal a priori error estimates for a class of implicit one-step methods for stiff ordinary differential equations obtained by using the discontinuous Galerkin method with piecewise polynomials of degree zero and one. Starting from these estimates we propose a new algorithm for automatic time-step control and we discuss the relation between this algorithm and earlier algorithms implemented in packages for the numerical solution of stiff ordinary differential equations.

**Key words.** automatic time-step control, error estimates, stiff ordinary differential equations (ODEs), discontinuous Galerkin, one-step method

**AMS(MOS) subject classifications.** 65L05, 65L50, 65L60

**Introduction.** In this note we consider the problem of constructing algorithms for automatic time-step control for numerical methods for initial value problems for a class of stiff ordinary differential equations. Typically, exact solutions of stiff initial value problems are nonsmooth in (initial) transients but become smoother with increasing time. Efficient time stepping methods for computing approximate solutions of such problems require the time steps to adaptively be chosen small in transients and increasingly large as the exact solution becomes smoother.

Our objective is to construct adaptive time-stepping algorithms for stiff problems which in particular satisfy the following criteria. Here $\delta > 0$ is a given tolerance, $e(t)$ is the error in the approximate solution at time $t$, $|\cdot|$ is the Euclidean norm, and $I = (0, T)$ is a given time interval.

(a) For $t \in I$ we have $|e(t)| \sim \delta$.

(b) The cost of the time-step control is comparatively small.

(c) The algorithm can be theoretically justified.

(d) No (or only very rough) a priori information of the exact solution is required.

To be able to satisfy (d), the necessary information concerning the smoothness of the exact solution must be obtained from the computed approximate solution as the computation proceeds.

A program for constructing adaptive methods satisfying (a)-(d) was initiated in Johnson [10] and was continued in the context of parabolic problems in Eriksson and Johnson [7], Johnson, Nie, and Thomée [11] and Eriksson, Johnson, and Lennblad [8]. The present paper is a revised version of [10].

Let us now, for the moment in a nonprecise way, describe the basic ideas in our approach to adaptivity for stiff problems. Comparisons with earlier adaptive methods implemented in packages for stiff ordinary differential equations will be made in § 3 below. Consider an initial value problem of the following form: Find $y = (y_1, \cdots, y_M) : (0, T) \to \mathbb{R}^M$, $M \geqq 1$, such that

$$\dot{y}(t) + f(y(t)) = g(t), \qquad t \in (0, T), \tag{0.1a}$$

$$y(0) = y_0, \tag{0.1b}$$

* Received by the editors November 26, 1984; accepted for publication (in revised form) July 28, 1987. This research was partly supported by the Swedish Board for Technical Development (STU).

† Department of Mathematics, Chalmers University of Technology, 412 96 Göteborg, Sweden.

where $f=(f_1,\cdots,f_M):\mathbb{R}^M\to\mathbb{R}^M$, $g:\mathbb{R}\to\mathbb{R}^M$, $y_0\in\mathbb{R}^M$, and $T\in(0,\infty]$ are given. We assume that (0.1) is stiff which, roughly speaking, corresponds to the fact that the eigenvalues of the Jacobian matrix $f'(y(t))$, $f'=(\partial f_i/\partial y_k)$, are distributed in a sector $\{z\in\mathbb{C}\colon \operatorname{Re} z\in[0,\Lambda],\ \arg z\in(-\alpha,\alpha)\}$ in the complex plane $\mathbb{C}$, where $0\leqq\alpha<\pi/2$ and $\Lambda\gg 1$. The simplest example is given by $f(x)=Ax$, where $A$ is a symmetric positive definite $M\times M$-matrix with eigenvalues distributed in a large interval $[0,\Lambda]$. Below we will consider a certain class of stiff problems of the form (0.1) with precise assumptions on $f$, $g$, and $y$.

To discretize (0.1) in time let $0=t_0<t_1<\cdots t_N=T$ be a subdivision of $I=(0,T)$ into time intervals $I_n=(t_{n-1},t_n]$ of length $k_n=t_n-t_{n-1}$. We will consider a special class of implicit one-step methods for (0.1) obtained by discretizing in time using the discontinuous Galerkin method with piecewise polynomials of degree $q$ (for short, the DG $(q)$-method below). In the case $f(x)=Ax$ and $g=0$, where $A$ is a constant $M\times M$-matrix, these methods reduce to the following well-known class of methods:

$$Y_n=r(k_nA)\,Y_{n-1},\qquad n=1,2,\cdots,N, \tag{0.2a}$$

$$Y_0=y_0, \tag{0.2b}$$

where $r(\lambda)$ is the subdiagonal Padé approximation of $e^{-\lambda}$ of order $p=2q+1$ and $Y_n$ is an approximation of $y(t_n)$. In the general case the DG $(q)$-method for (0.1) can be formulated as follows: For $n=1,2,\cdots,N$, given $Y_{n-1}^-$ find $Y\equiv Y|_{I_n}\in[P_q(I_n)]^M$ such that

$$\int_{I_n}(\dot Y+f(Y),v)\,dt+(Y_{n-1}^+,v_{n-1}^+)=(Y_{n-1}^-,v_{n-1}^+)+\int_{I_n}(g,v)\,dt \tag{0.3}$$

$$\forall v\in[P_q(I_n)]^M,$$

where $(\cdot,\cdot)$ denotes the usual scalar product in $\mathbb{R}^M$,

$$Y_0^-=y_0,\qquad v_n^\pm=v(t_n)^\pm,\qquad v(t)^\pm=\lim_{s\to 0^\pm}v(t+s),$$

and $P_q(I_n)$ denotes the set of polynomials of degree at most $q$ on $I_n$. For example, in the case $q=0$ the above method can be formulated as follows with $Y_n\equiv Y_n^-$: For $n=1,2,\cdots$, given $Y_{n-1}\in\mathbb{R}^M$ find $Y_n\in\mathbb{R}^M$ such that

$$Y_n+k_nf(Y_n)=Y_{n-1}+\int_{I_n}g\,dt, \tag{0.4a}$$

$$Y_0=y_0, \tag{0.4b}$$

which is a variant of the classical backward Euler method with an average of the right-hand side replacing the usual value $g(t_n)$.

In the case $q=1$ we have the following characterization of $Y$ on each interval $I_n$: $Y(t)=(1-r)U_0+rU_1$, with $r=(t-t_{n-1})/k_n$, $t\in I$, where $(U_0,U_1)\in\mathbb{R}^M\times\mathbb{R}^M$ satisfies

$$U_1+k_n\int_0^1 f((1-r)U_0+rU_1)\,dr=Y_{n-1}^-+k_n\int_0^1 g(t_{n-1}+k_nr)\,dr, \tag{0.5a}$$

$$U_1-U_0+2k_n\int_0^1 f((1-r)U_0+rU_1)r\,dr=2k_n\int_0^1 g(t_{n-1}+k_nr)r\,dr. \tag{0.5b}$$

Error estimates for (0.3) in the case of nonstiff ordinary differential equations were first given in Delfour, Hager, and Trochu [5].

*Remark* 1. To evaluate the integrals in (0.3)-(0.5) we would in practice use numerical quadrature which would introduce additional quadrature errors. For simplicity we do not in this note consider the effect on the global error of these quadrature errors and we leave this problem to subsequent work. Let us just note that we have a situation parallel to that in finite element analysis of, e.g., elliptic problems where one first derives error estimates assuming that all integrals are evaluated exactly and then considers the effect of numerical quadrature on the discrete scheme. This has proven to be a convenient way of simplifying the analysis by separating the full discretization into two steps, namely the discretization of the differential equation using piecewise polynomials and the evaluation of the resulting integrals in the discrete problem. If this separation is not used, then important structural aspects of the analysis may get lost. Note that in (0.4) the analysis of the effect of quadrature is very simple since no integrals involving the unknown $Y$ occur. Using the monotonicity of $f$ (see below) we easily find in this case that the part of the global error due to quadrature is bounded by the sum of the local quadrature errors resulting from the integrals on the right-hand side of (0.4a) involving the given function $g$. □

Our adaptive algorithms are based on almost optimal a priori error estimates for (0.3) of basically the following form

$$|e|_n \leqq C_n \max_{m \leqq n} k_m^p |y^{(p)}|_m, \qquad n = 1, 2, \cdots, \tag{0.6}$$

where $e = y - Y$, $y^{(p)} = d^p y / dt^p$, $p = q+1$ or $p = 2q+1$, $|v|_m = \max_{t \in I_m} |v(t)|$, $I_m = (t_{m-1}, t_m)$, $|\cdot|$ is the Euclidean norm in $\mathbb{R}^M$ and $C_n$ is a constant including a logarithmic dependence on $t_n/k_n$ and also a "mild" dependence on $y$ (in case (0.1) is nonlinear) to be made precise below. Note that (0.6) in fact represents a new improved type of error estimate for stiff problems. Earlier estimates typically are of the form (see, e.g., Frank, Schneid, and Ueberhuber [9] and Dahlquist [2], [3])

$$|e|_n \leqq C \sum_{m=1}^{n} k_m^p \int_{I_m} |y^{(p+1)}| \, dt, \tag{0.7}$$

where the global error is bounded by a sum of so-called local truncation errors (see § 3 below) involving derivatives of order $p+1$ (note that only derivatives of order $p$ occur in (0.6)). An attempt to improve earlier error estimates for linear multistep methods of the form (0.7) in the direction of (0.6) is made in the recent paper by Nevanlinna and Jeltsch [14].

Suppose now $\delta > 0$ is a given *tolerance* and suppose we want the error $e = y - Y$ to satisfy

$$|e|_n \leqq \delta \quad \text{for } n = 1, 2, \cdots. \tag{0.8}$$

Relying on (0.6) we are then led to seek to choose the time steps $k_m$ so that, replacing for simplicity the constants $C_n$ by a common constant $C$,

$$C k_m^p |y^{(p)}|_m \cong \delta$$

or

$$k_m \cong \left( \frac{\delta}{C |y^{(p)}|_m} \right)^{1/p}. \tag{0.9}$$

Our adaptive algorithm is based on a computational form of (0.9) where the unknown quantities $C$ and $|y^{(p)}|_m$ are successively estimated through the computed solution $Y_n$. In § 3 below we compare our adaptive algorithms based on (0.6) with earlier adaptive

algorithms based on (0.7). Stated briefly, the advantages of (0.6) as compared with (0.7) are two-fold. First, the form of (0.6) with a maximization on the right-hand side is more suitable for error control than (0.7) with instead a sum to control. Second, (0.6) is easier to implement and to justify theoretically since derivatives of order $p$ and not $p+1$ are involved.

Let us now consider in more detail the simple model case when $f(x) = Ax$ with $A$ symmetric, positive definite, $g = 0$, and $r(\lambda) = (1+\lambda)^{-1}$ corresponding to the DG (0)-method, or in classical terms the backward Euler method:

$$Y_n = (I + k_n A)^{-1} Y_{n-1}.$$

In this case (0.6) holds with $p = 1$ and $C_n = C(1+\log(t_n/k_n))$ where $C$ is a positive constant depending only on $\max_{n\geqq 2}(k_{n-1}/k_n)$ (in particular $C$ is independent of $M$ and the distribution of eigenvalues of $A$ on the positive real axis). A short proof of this estimate is given in § 1 below. We now replace in (0.9) with $p = 1$ the unknown quantity $|\dot{y}|_m$ by the computable quantity $k_m^{-1}|Y_m - Y_{m-1}|$, and we are led to the following remarkably simple (and apparently new) algorithm for adaptive choice of time steps for the backward Euler method: For $m = 1, 2, \cdots$, given $Y_{m-1}$ choose $k_m$ so that

$$C|Y_m - Y_{m-1}| = \delta. \tag{0.10}$$

To determine $k_m$ from (0.10) we have in principle to solve a (nonlinear) equation since $Y^m$ depends on $k_m$. In practice one may expect to be able to predict $k_m$ accurately by (cf. [8], [11])

$$k_m = \frac{\delta k_{m-1}}{C|Y_{m-1} - Y_{m-2}|}, \tag{0.11}$$

which corresponds to estimating $|\dot{y}|_m$ by $k_{m-1}^{-1}|Y_{m-1} - Y_{m-2}|$.

To be able to verify that (0.10) will imply the global error control (0.8), we would need an a posteriori error estimate of the form

$$|e_n| \leqq C_n \max_{m\leqq n} |Y_m - Y_{m-1}|. \tag{0.12}$$

Now, it is in fact possible to prove under certain natural assumptions that an estimate of basically the form (0.12) is valid in the case $f(x) = Ax$ (see [11]). Thus, for the backward Euler method with step control according to (0.10) or (0.11), we can prove in the model case $f(x) = Ax$ that (a) in our above list is satisfied. Clearly (b) and (d) hold in this case and as indicated also (c). As a result we have that our program (a)-(d) is fully realized in the case of linear constant coefficient problems and the implicit Euler method or DG (0)-method (see [11], where numerical results supporting the theory also are presented).

Our aim is now to obtain corresponding results for classes of nonlinear problems and/or higher order time discretization methods. In this direction so far we have obtained the following results. In [7] an a priori estimate of the form (0.6) with $p = 1$ for DG (0) for a nonlinear parabolic problem was proved. A priori estimates for DG (1) for linear parabolic problems and associated time- (and space-) step control are given in [8] together with results of numerical experiments. In this paper we prove a priori estimates of basically the form (0.6) for DG (0) and DG (1) for a class of nonlinear stiff ordinary differential equations. A posteriori analogues of these a priori estimates will appear in subsequent work together with more extensive numerical tests.

An outline of the content of this note is as follows: In § 1 we give, using a spectral method, a short proof of an a priori error estimate of the form (0.6) for stiffly stable

methods of the form (0.2) in the model case $f(x) = Ax$ with $A$ symmetric, positive definite and independent of time and $g = 0$. The purpose here is mainly didactic since spectral methods are difficult to employ in nonlinear or variable coefficient problems. In § 2 we prove using variational techniques the a priori estimates for DG (0) and DG (1) for a class of nonlinear stiff ordinary differential equations. In § 3 we compare our adaptive technique with earlier techniques implemented in packages for stiff ordinary differential equations. For numerical results we refer to [8] and [11].

**1. A priori estimates for linear constant coefficient problems using spectral methods.** We consider the following problem:

$$\dot{y} + Ay = 0, \qquad t \in (0, T), \tag{1.1a}$$

$$y(0) = y_0, \tag{1.1b}$$

where $A$ is a positive definite $M \times M$-matrix and $y_0 \in \mathbb{R}^M$. Let $0 < \lambda_1 \leqq \cdots \leqq \lambda_M$ be the eigenvalues of $A$ with corresponding normalized eigenvectors $\varphi^j \in \mathbb{R}^M$, $j = 1, \cdots, M$. The solution of (1.1) is given by

$$y(t) = \sum_{j=1}^{M} (y_0, \varphi^j)\, e^{-\lambda_j t} \varphi^j, \tag{1.2}$$

where $(\cdot, \cdot)$ is the usual scalar product in $\mathbb{R}^M$.

Let now $r(\lambda)$ be a rational approximation of $e^{-\lambda}$ such that

$$e^{-\lambda} - r(\lambda) = \mathcal{O}(\lambda^{p+1}) \quad \text{as } \lambda \to 0, \tag{1.3a}$$

$$|r(\lambda)| < 1 \quad \text{for } \lambda > 0, \tag{1.3b}$$

$$r(\lambda) \to 0 \quad \text{as } \lambda \to \infty. \tag{1.3c}$$

For use below we note that (1.3b) and (1.3c) together imply that for some $c > 0$ (see Thomée [16])

$$|r(\lambda)| \leqq \frac{1}{1 + c\lambda}, \qquad \lambda > 0. \tag{1.4}$$

Let the sequence $Y_n \in \mathbb{R}^M$, $n = 0, 1, 2, \cdots$, be defined by (0.2), that is,

$$Y_n = r(k_n A)\, Y_{n-1}, \qquad n = 1, 2, \cdots, \tag{1.5a}$$

$$Y_0 = y_0. \tag{1.5b}$$

We shall prove the following result.

THEOREM 1. *Let $\bar{C} > 0$ be given and suppose that the time-step sequence $\{t_n\}$ satisfies $k_{n-1} \leqq \bar{C} k_n$, $n = 2, 3, \cdots$. Let $\{Y_n\}$ be given by (1.5) where $r(\lambda)$ satisfies (1.3). Then there exists a constant $C$ depending only on $\bar{C}$ and $r$ (thus $C$ is in particular independent of $M$, $A$, $y_0$, and $T$) such that for $n = 1, 2, \cdots$*

$$|Y_n - y_n| \leqq C\left(1 + \log \frac{t_n}{k_n}\right) \max_{m \leqq n} k_m^p |A^p y_{m-1}|, \tag{1.6}$$

*where $y_n = y(t_n)$.*

*Proof.* Given $n$ define for $m = 1, \cdots, n-1$,

$$R_n(t_m, A) = \prod_{j=m+1}^{n} r(k_j A),$$

and set $R_n(t_n, A) = I =$ Identity. We have the following representation of the error $Y - y$:

$$Y_n - y_n = \sum_{m=1}^{n} R_n(t_m, A)(r(k_m A) - e^{-k_m A})\, e^{-t_{m-1}A} y_0.$$

Thus, since $y^{m-1} = e^{-t_{m-1}A} y_0$,

$$\begin{aligned} |Y_n - y_n| \leqq & \sum_{m=1}^{n-1} |AR_n(t_m, A)| k_m \max_{m \leqq n-1} |(r(k_m A) - e^{-k_m A}) A^{-1} y_{m-1} k_m^{-1}| \\ & + |(r(k_n A) - e^{-k_n A})\, e^{-t_{n-1}A} y_0|. \end{aligned} \tag{1.7}$$

For the first factor in the sum we have

$$\begin{aligned} \sum_{m=1}^{n-1} |AR_n(t_m, A)| k_m &= \sum_{m=1}^{n-1} |AR_n(t_m, A)(t_n - t_{m-1})| \frac{k_m}{t_n - t_{m-1}} \\ &\leqq \max_{m \leqq n-1} |AR_n(t_m, A)(t_n - t_{m-1})| \sum_{m=1}^{n-1} \frac{k_m}{t_n - t_{m-1}} \\ &\leqq \max_{m \leqq n-1} \max_{\lambda \geqq 0} |\lambda R_n(t_m, \lambda)(t_n - t_{m-1})| \log \frac{t_n}{k_n}. \end{aligned} \tag{1.8}$$

Recalling (1.4) we easily see that for $\lambda \geqq 0$ and $m = 1, \cdots, n-1$,

$$|R_n(t_m, \lambda)| \leqq (1 + c(t_n - t_m)\lambda)^{-1}.$$

Together with the fact that, by the assumption $k_{n-1} \leqq \bar{C} k_n$,

$$t_n - t_{m-1} \leqq C(t_n - t_m),$$

this proves that

$$\max_{m \leqq n-1} \max_{\lambda \geqq 0} |R_n(t_m, \lambda)(t_n - t_{m-1})\lambda| \leqq C. \tag{1.9}$$

For the second factor in the sum in (1.7) we have by (1.3a), (1.3b) for $m = 1, \cdots, n-1$

$$\begin{aligned} |(r(k_m A) - e^{-k_m A}) A^{-1} y_{m-1} k_m^{-1}| &\leqq |(r(k_m A) - e^{-k_m A})(k_m A)^{-(p+1)}| k_m^p |A^p y_{m-1}| \\ &\leqq \max_{\lambda \geqq 0} |(r(k_m \lambda) - e^{-k_m \lambda})(k_m \lambda)^{-(p+1)}| k_m^p |A^p y_{m-1}| \\ &\leqq C k_m^p |A^p y_{m-1}|, \end{aligned}$$

and similarly

$$|(r(k_n A) - e^{-k_n A}) y_n| \leqq C k_n^p |A^p y_{n-1}|.$$

Combining these estimates with (1.7)-(1.9) we obtain the statement of the theorem. □

*Remark.* The standard error estimate for (1.5), which corresponds to (0.7), reads

$$|Y_n - y_n| \leqq C \sum_{m=1}^{n} k_m^{p+1} |A^{p+1} y_{m-1}|.$$

In the proof of this estimate we use the following weaker stability condition instead of (1.9):

$$\max_{m \leqq n-1} \max_{\lambda \geqq 0} |R_n(t_m, \lambda)| \leqq 1. \qquad \square$$

**2. A priori error estimates for DG (0) and DG (1).**

**2.1. Assumptions and statement of main result.** We shall now derive a priori error estimates of the form (0.6) with $p = q + 1$ for DG $(q)$, $q = 0, 1$, applied to the nonlinear

problem (0.1) under precise assumptions on $f$ and the given exact solution $y(t)$. We also prove for DG (1) an estimate of basically the form (0.6) with $p=3=2q+1$ and with $|e|_n$ replaced by $|e_n^-|$ showing that DG (1) is third-order accurate at the discrete time levels $t_n$. As discussed above these estimates may be used as the theoretical basis for adaptive time-step control.

We use the following notation:

$$A(t)=f'(y(t)), \qquad \dot{A}(t)=\frac{d}{dt}A(t),$$

$$\hat{A}=\frac{1}{2}(A+A^*),$$

$$B_n=\frac{1}{k_n}\int_{I_n} B(t)\,dt \quad \text{for } B=\hat{A}, A^*,$$

where $f'$ denotes the Jacobian of $f$ and $A^*$ is the transpose of $A$. Further, $(\cdot,\cdot)$ denotes the usual scalar product in $\mathbb{R}^M$ with corresponding norm $|\cdot|$ and $\|\cdot\|$ denotes a significantly stronger norm (the "energy norm") to be commented on below.

We shall use below the following interpolant $\tilde{Y}\in V_h^{(q)}$ of $y$, defined for $q=0$ by

$$\tilde{Y}_n^- = y(t_n), \qquad n=1,\cdots,N, \tag{2.1a}$$

and for $q=1$ by (2.1a) together with

$$\int_{I_n} (\tilde{Y}-y)\,dt=0, \qquad n=1,\cdots,N, \tag{2.1b}$$

that is,

$$\tilde{Y}|_{I_n} = y(t_n)+(t-t_n)\frac{2}{k_n^2}\int_{I_n}(y(t)-y(t_n))\,dt.$$

We assume that there are positive constants $c$, $C$, and $\gamma$ with $\gamma$ sufficiently small, such that for all $v, w, u\in\mathbb{R}^M$, $n=1,\cdots,N$,

$$c\|v-w\|^2 \leqq (f(v)-f(w), v-w) \leqq C\|v-w\|^2, \tag{2.2a}$$

$$c\|v\|^2 \leqq (A(t)v, v) \leqq C\|v\|^2, \quad t\in I, \tag{2.2b}$$

$$|(\dot{A}(t)v, v)| \leqq C|\dot{y}(t)|\,\|v\|^2, \quad t\in I, \tag{2.2c}$$

$$|((f'(y(t)+u)-f'(y(t))w, v)| \leqq C\min(|u|\,\|w\|\,\|v\|, |u||w||A^*(t)v|), \quad t\in I, \tag{2.2d}$$

$$|(\hat{B}(t)-\hat{B}_n)v| \leqq C|\dot{y}|_{L_1(I_n)}|A^*(t)v|, \quad t\in I_n, \quad B=\hat{A}, A^*, \tag{2.2e}$$

$$(A^*(t)v, \hat{A}(t)v) \geqq c|A^*(t)v|^2, \quad t\in I, \tag{2.2f}$$

$$\int_I |\dot{y}(t)|\,dt \leqq C, \tag{2.3a}$$

$$|\dot{y}|_{L_1(I_n)} \equiv \int_{I_n}|\dot{y}|\,dt \leqq \gamma, \tag{2.3b}$$

$$\left(\int_I \|y-\tilde{Y}\|^2\,dt\right)^{1/2} \leqq \gamma, \tag{2.3c}$$

$$k_{n-1} \leqq Ck_n, \qquad n=1,\cdots,N. \tag{2.4}$$

Finally, we assume that $|g(t)|$ is integrable on $I$.

We now briefly comment on the above assumptions which, on the one hand, model basic features of stiff nonlinear initial value problems and, on the other hand,

are chosen so as to allow a fairly nontechnical analysis leading to optimal error estimates. The assumptions (2.2a), (2.2b) reflect the monotonicity of $f$ and may be viewed as giving an intrinsic definition of the energy norm $\|\cdot\|$. The assumption that $\|\cdot\|$ is a much stronger norm than $|\cdot|$ reflects the stiffness of the problem. Assumptions (2.2c)-(2.2e) state natural boundedness and continuity properties of $f'$. Further (2.2f) states that the nonsymmetric part $\frac{1}{2}(A-A^*)$ of $A$ in some sense is dominated by the symmetric part $\hat{A}$. If $A=\lambda\in\mathbb{C}$ then (2.2f) corresponds to $\lambda\in\{z\in\mathbb{C}\colon \arg z\in(-\alpha,\alpha)\}$ where $0\leqq\alpha<\pi/2$. The assumption (2.2f) may be replaced by the assumption $(Av,w)^2\leqq C(A(t)v,v)(A(t)w,w)$ corresponding to angleboundedness (cf. Remark 3 below). This possibility was suggested to us by O. Nevanlinna (see also Nevanlinna and Odeh [13]). With only trivial changes in the proofs to follow we may generalize the lower bound in (2.2a) to

$$(f(v)-f(w),v-w)\geqq c\|v-w\|^2-C|v-w|^2, \tag{2.5}$$

and similarly we may add lower order terms to the right-hand sides in (2.2d)-(2.2f).

The assumptions (2.3a) and (2.3c) put a certain (mild) restriction on the regularity of the exact solution $y(t)$, but still allow $y(t)$ to have, e.g., an initial transient. Further, (2.3b) puts a restriction on the time steps $k_n$ which guarantees that $y(t)$ does not change significantly on a single time interval.

In the literature on numerical methods for stiff ordinary differential equations usually more general assumptions than those stated above are used (see, e.g., Burrage and Butcher [1] and Dahlquist [4]). In fact, in most cases the basic assumptions are that $f$ is monotone in the sense that for $v,w\in\mathbb{R}^M$

$$(f(v)-f(w),v-w)\geqq 0, \tag{2.6a}$$

and that $f'$ is "unbounded" in the sense that

$$\Lambda(t)=\max_{\substack{v\in\mathbb{R}^M\\ v\neq 0}}\frac{|f'(y(t))v|}{|v|}\gg 1. \tag{2.6b}$$

However, if the monotonicity assumption is reduced to (2.6a) (as compared to (2.2a) or (2.5)), then in fact valuable available information may be discarded. This can be seen already in the case $f(y)=Ay$ with $A$ a diagonal matrix with diagonal elements $\lambda_i\geqq 0$, $i=1,\cdots,M$, where $\Lambda\equiv\max_i\lambda_i\gg 1$. If we in this case only use the fact that

$$(Av,v)\geqq 0 \tag{2.7}$$

corresponding to (2.6a), instead of

$$(Av,v)\geqq\|v\|^2\equiv\sum_i\lambda_i v_i^2, \tag{2.8}$$

corresponding to our assumption (2.2a), (2.2b), then we would lose strong information since some $\lambda_i$ are large. In fact, only under the assumption (2.6a) is it probably impossible to perform a precise error analysis and to derive sharp error estimates of the form (0.6).

To sum up, it seems that in order to be able to obtain sharp error estimates which can be used as a basis for rational methods for automatic time-step control, it is necessary to go beyond (2.6) and use more of the structural properties of stiff problems.

We are now ready to state the main a priori error estimate of this paper. Here and below

$$L_n=\left(1+\log\frac{t_n}{k_n}\right)^{1/2},\qquad |v|_I=\max_{t\in I}|v(t)|.$$

The existence of a unique solution of (0.3) follows, e.g., from Brouwer's fixed point theorem and the monotonicity of $f$.

THEOREM 2. *Suppose* (2.2)-(2.4) *are satisfied. For* $q=0$, 1, *let* $Y$ *be the solution given by the* DG $(q)$*-method* (0.3) *applied to* (0.1). *Then there is a constant* $C$ *such that for* $n=1, \cdots, N$,

$$|e|_n \leqq CL_n \max_{m \leqq n} k_m^{q+1} |y^{(q+1)}|_m. \tag{2.9}$$

*Further for* $q=1$

$$|e_n^-| \leqq CL_n \left( \max_{m \leqq n} (k_m^3 |A_m y^{(2)}|_m + k_m^3 |y^{(2)}|_m |\dot{y}|_m) + |e|_I^2 \right). \tag{2.10}$$

The proof of this result will be divided into the following steps showing the overall logic of the argument:

(a) a basic error estimate;
(b) an error representation formula;
(c) a strong stability estimate for a linearized backward problem;
(d) end of proof in the case $q=0$;
(e) end of proof in the case $q=1$.

In the analysis it is convenient to write the DG $(q)$-method (0.3) on compact form as follows: Find $Y \in V_h^{(q)}$ such that

$$B(Y, v) = L(v) \quad \forall v \in V_h^{(q)}, \tag{2.11}$$

where

$$\begin{aligned} &V_h^{(q)} = \{v : I \to \mathbb{R}^M : v|_{I_n} \in [P_q(I_n)]^M,\ n=1, \cdots, N\}, \\ &B(w, v) = \sum_{n=1}^{N} \int_{I_n} (\dot{w} + f(w), v)\, ds + \sum_{n=2}^{N} ([w_{n-1}], v_{n-1}^+) + (w_0^+, v_0^+), \\ &L(v) = \int_I (g, v)\, dt + (y_0, v_0^+), \\ &[w_n] = w_n^+ - w_n^-. \end{aligned} \tag{2.12}$$

We note that

$$B(y, v) - B(Y, v) = L(v) - L(v) = 0 \quad \forall v \in V_h^{(q)}. \tag{2.13}$$

We also note that by integrating by parts over each subinterval $I_n$, we obtain the following alternate expression for $B(w, v)$:

$$B(w, v) = \sum_{n=1}^{N} \int_{I_n} (-(w, \dot{v}) + (f(w), v))\, dt - \sum_{n=1}^{N-1} (w_n^-, [v_n]) + (w_N^-, v_N^-). \tag{2.14}$$

In particular, adding (2.12) and (2.14) we see that

$$\begin{aligned} B(v, v-w) - B(w, v-w) &= \int_I (f(v) - f(w), v-w)\, dt \\ &\quad + \frac{1}{2} \left( |(v-w)_0^+|^2 + \sum_{n=1}^{N-1} |[(v-w)_n]|^2 + |(v-w)_N^-|^2 \right) \\ &\geqq c \int_I \|v-w\|^2\, dt \\ &\quad + \frac{1}{2} \left( |(v-w)_0^+|^2 + \sum_{n=1}^{N-1} |[(v-w)_n]|^2 + |(v-w)_N^-|^2 \right), \end{aligned} \tag{2.15}$$

where the last inequality follows from (2.1a).

**2.2. A basic error estimate.** Recall that $\tilde{Y} \in V_h^{(q)}$ is the interpolant of $y$ defined by (2.1) and set $\theta = Y - \tilde{Y} \in V_h^{(q)}$. Using (2.15), (2.13), (2.2a), and the definition of $\tilde{Y}$, we have with $\eta = y - \tilde{Y}$

$$
\begin{aligned}
c \int_I \|\theta\|^2 \, dt &\leqq B(Y, \theta) - B(\tilde{Y}, \theta) = B(y, \theta) - B(\tilde{Y}, \theta) \\
&= \int_I (f(y) - f(\tilde{Y}), \theta) \, dt \leqq C \int_I \|\eta\| \, \|\theta\| \, dt.
\end{aligned} \tag{2.16}
$$

Thus

$$
\int_I \|\theta\|^2 \, dt \leqq C \int_I \|\eta\|^2 \, dt,
$$

so that since $e = \eta - \theta$ and recalling (2.3c), we have

$$
\int_I \|e\|^2 \, dt \leqq C \int_I \|\eta\|^2 \, dt \leqq C\gamma^2. \tag{2.17}
$$

**2.3. An error representation formula.** We now derive an error representation formula based on a discrete solution of a linear "backward" problem obtained by linearization along the exact solution $y(t)$. We first note that (2.13) can be written in the form

$$
\tilde{D}(e, v) = 0 \quad \forall v \in V_h^{(q)}, \tag{2.18}
$$

where

$$
\begin{aligned}
\tilde{D}(w, v) = \sum_{n=1}^{N} \Bigg[ \int_{I_n} (\dot{w}, v) \, dt + \int_{I_n} \left( \int_0^1 f'(ry + (1-r)Y) \, dr \, w, v \right) dr \Bigg] \\
+ \sum_{n=2}^{N} ([w_{n-1}], v_{n-1}^+) + (w_0^+, v_0^+).
\end{aligned}
$$

Next, we introduce the related bilinear form

$$
\begin{aligned}
D(w, v) &\equiv \sum_{n=1}^{N} \int_{I_n} (\dot{w} + A(t)w, v) \, dt + \sum_{n=2}^{N} ([w_{n-1}], v_{n-1}^+) + (w_0^+, v_0^+) \\
&= \sum_{n=1}^{N} - \int_{I_n} (w, \dot{v} + A^*(t)v) \, dt - \sum_{n=2}^{N} (w_{n-1}^-, [v_{n-1}]) + (w_N^-, v_N^-).
\end{aligned} \tag{2.19}
$$

Let now $s \in (t_{N-1}, t_N]$ and let $Z \in V_h^{(q)}$ be defined by

$$
D(w, Z) = (w(s)^-, \theta(s)^-) \quad \forall w \in V_h^{(q)}, \tag{2.20}
$$

where $\theta = Y - \tilde{Y}$ and $\tilde{Y} \in V_h^{(q)}$ is defined by (2.1), that is, $Z$ is a discontinuous Galerkin solution of the linear "backward" problem

$$
\begin{aligned}
&-\dot{z} + A^*(t)z = 0, \qquad 0 < t < s, \\
&z(s) = \theta(s)^- = (Y - \tilde{Y})(s)^-.
\end{aligned} \tag{2.21}
$$

Now taking $w = \theta = Y - \tilde{Y}$ in (2.20), we get using (2.18)

$$\begin{aligned}|\theta(s)^-|^2 = D(Y-\tilde{Y},Z) &= \tilde{D}(Y-\tilde{Y},Z)+(D-\tilde{D})(Y-\tilde{Y},Z)\\ &= \tilde{D}(y-\tilde{Y},Z)+(D-\tilde{D})(Y-\tilde{Y},Z)\\ &= D(\eta,Z)+(\tilde{D}-D)(\eta,Z)+(D-\tilde{D})(Y-\tilde{Y},Z)\\ &= D(\eta,Z)+(\tilde{D}-D)(e,Z).\end{aligned}$$

We thus arrive at the following representation formula for the quantity $\theta = Y - \tilde{Y}$:

$$|\theta(s)^-|^2 = D(\eta,Z)+(\tilde{D}-D)(e,Z), \tag{2.22}$$

where $Z$ satisfies (2.20) and $s \in (t_{N-1}, t_N]$.

**2.4. A strong stability estimate for a linearized backward problem.** We now prove a strong stability estimate for the solution $Z$ of (2.20) which is the key technical part of the proof of Theorem 2.

LEMMA 1. *There is a constant $C$ such that if $Z \in V_h^{(q)}$, $q = 0, 1$, satisfies* (2.20), *then*

$$\left(\int_I \|Z\|^2\,dt + \sum_{n=1}^N (t_N - t_{n-1})\int_{I_n} |A^*(t)Z|^2\,dt\right)^{1/2} \leqq C|\theta(s)^-|, \tag{2.23}$$

$$\int_I |A^*(t)Z|\,dt \leqq CL_N|\theta(s)^-|, \tag{2.24a}$$

$$\sum_{n=1}^N \int_{I_n} |\dot{Z}|\,dt \leqq CL_N|\theta(s)^-|. \tag{2.24b}$$

To prove this result we first note that (2.24a) follows from (2.23) via Cauchy's inequality:

$$\sum_{n=1}^N \int_I |A^*Z|\,dt \leqq \left(\sum_{n=1}^N (t_N - t_{n-1})^{-1}k_n\right)^{1/2}\left(\sum_{n=1}^N (t_N - t_{n-1})\int_{I_n}|A^*Z|^2\,dt\right)^{1/2}.$$

Further (2.24b) follows from (2.24a) since taking in (2.20), $v(t) = (t - t_n)\dot{Z}$ for $t \in I_n$ shows that

$$\sum_{n=1}^N \int_{I_n} |\dot{Z}|\,dt \leqq C\left(\int_I |A^*Z|\,dt + |\theta(s)^-|\right).$$

It thus remains to prove (2.23); to facilitate the reading of this proof, we shall first prove an analogous result for the following forward variant of the continuous problem (2.21): Find $x\colon (0, T) \to \mathbb{R}^M$ such that

$$\dot{x} + A(t)x = 0, \qquad 0 < t < T, \tag{2.25a}$$

$$x(0) = g. \tag{2.25b}$$

The assumptions for (2.25) correspond to those for (2.21), i.e., $A^*$ is replaced by $A$ in (2.2).

LEMMA 2. *Suppose* (2.2b), (2.2c), (2.2f), *and* (2.3a) *are satisfied. Then there is a constant $C$ such that if $x$ satisfies* (2.25), *then*

$$\left(\int_0^T \|x\|^2\,dt + \int_0^T t|A(t)x|^2\,dt\right)^{1/2} \leqq C|g|.$$

*Proof.* Multiplying (2.25a) by $x(t)$ and integrating we get using (2.2b)

$$0=\int_0^T \frac{1}{2}\frac{d}{dt}|x|^2\,dt+\int_0^T (Ax,x)\,dt \geqq \frac{1}{2}|x(T)|^2-\frac{1}{2}|g|^2+c\int_0^T \|x\|^2\,dt$$

so that

$$\int_0^T \|x\|^2\,dt \leqq C|g|^2. \tag{2.26}$$

Next, multiplying (2.25a) by $t\hat{A}(t)x$ we find that

$$\frac{1}{2}\frac{d}{dt}(t(x,A(t)x))+t(A(t)x,\hat{A}(t)x)=\frac{1}{2}(x,\hat{A}(t)x)+\frac{1}{2}t(x,\dot{A}(t)x).$$

Integrating we thus have for $0<s<T$ using (2.2b), (2.2c), (2.2f)

$$s\|x(s)\|^2+\int_0^s t|A(t)x|^2\,dt \leqq C\int_0^s \|x\|^2\,dt+C\int_0^s |\dot{y}(t)|t\|x(t)\|^2\,dt.$$

Thus by Gronwall's inequality we get recalling (2.26) and (2.3a)

$$T\|x(t)\|^2+\int_0^T t|A(t)x|^2\,dt \leqq C\int_0^T \|x\|^2\,dt \leqq C|g|^2,$$

which proves the lemma. □

We can now give the proof of Lemma 1 which is a discrete version of the proof of Lemma 2.

*Proof of Lemma* 1. For convenience we prove a result analogous to (2.23) for the following discrete version of the forward problem (2.25): Given $g\in\mathbb{R}^M$ and $s\in[0,t_1)$ find $X\in V_h^{(q)}$ such that

$$D(X,v)=(g,v(s)^+)\quad \forall v\in V_h^{(q)}. \tag{2.27}$$

We thus want to prove that if $X$ satisfies (2.27), then

$$\left(\int_I \|X\|^2\,dt+\sum_{n=1}^N t_n\int_{I_n}|A(t)X|^2\,dt\right)^{1/2}\leqq C|g|, \tag{2.28}$$

under the assumptions (2.2) and (2.3) with again $A^*$ replaced by $A$, together with the following analogue of (2.4) for the forward problem (2.27):

$$k_n \leqq Ck_{n-1},\qquad n=2,\cdots,N. \tag{2.29}$$

We now first take $v=X$ in (2.27) to get, evaluating the time derivative term,

$$\begin{aligned}(g,X(s)^+)&=\sum_{n=1}^N\int_{I_n}((\dot{X},X)+(AX,X))\,dt+\sum_{n=2}^N([X_{n-1}],X_{n-1}^+)+(X_0^+,X_0^+)\\&=\frac{1}{2}\left(|X_N^-|^2+\sum_{n=1}^{N-1}(|X_n^+|^2-2(X_n^+,X_n^-)+|X_n^-|^2)+|X_0^+|^2\right)+\int_I (AX,X)\,dt,\end{aligned}$$

so that by (2.2b)

$$|X_0^+|^2+\int_I \|X\|^2\,dt \leqq C(g,X(s)^+). \tag{2.30}$$

Next we choose $v = t_n \hat{A}_n X(t)$ in (2.27) to get

$$\sum_{n=1}^{N} \left( \int_{I_n} (\dot{X}, t_n \hat{A}_n X)\, dt + t_n \int_{I_n} (A(t)X, \hat{A}(t)X)\, dt \right)$$
$$+ \sum_{n=2}^{N} t_n([X_{n-1}], \hat{A}_n X_{n-1}^+) + t_1(X_0^+, \hat{A}_1 X_0^+)$$
$$= \sum_{n=1}^{N} t_n \int_{I_n} (A(t)X, (\hat{A}(t) - \hat{A}_n)X)\, dt + t_1(g, \hat{A}_1 X(s)^+)$$
$$\equiv R_1 + R_2,$$

with obvious notation. Evaluating the time derivative term, we thus have

$$R_1 + R_2 = \sum_{n=1}^{N} t_n \int_{I_n} (A(t)X, \hat{A}(t)X)\, dt + \sum_{n=1}^{N} \frac{1}{2} t_n((X_n^-, \hat{A}_n X_n^-) - (X_{n-1}^+, \hat{A}_n X_{n-1}^+))$$
$$+ \sum_{n=2}^{N} t_n(X_{n-1}^+, \hat{A}_n X_{n-1}^+) - t_n(X_{n-1}^-, \hat{A}_n X_{n-1}^+) + t_1(X_0^+, \hat{A}_1 X_0^+)$$
$$= \sum_{n=1}^{N} t_n \int_{I_n} (A(t)X, \hat{A}(t)X)\, dt + \frac{1}{2} t_N(X_N^-, \hat{A}_N X_N^-)$$
$$+ \frac{1}{2} \sum_{n=2}^{N} t_n((X_{n-1}^+, \hat{A}_n X_{n-1}^+) - 2(X_{n-1}^-, \hat{A}_n X_{n-1}^+) + (X_{n-1}^-, \hat{A}_n X_{n-1}^-))$$
$$+ \frac{1}{2} \sum_{n=2}^{N} (t_{n-1}(X_{n-1}^-, \hat{A}_{n-1} X_{n-1}^-) - t_n(X_{n-1}^-, \hat{A}_n X_{n-1}^-)) + \frac{1}{2} t_1(X_0^+, \hat{A}_1 X_0^+)$$
$$= \sum_{n=1}^{N} t_n \int_{I_n} (A(t)X, \hat{A}(t)X)\, dt + \frac{1}{2} t_N(X_N^-, \hat{A}_N X_N^-)$$
$$+ \frac{1}{2} \sum_{n=2}^{N} t_n([X_{n-1}], \hat{A}_n[X_{n-1}]) - R_3 + \frac{1}{2} t_1(X_0^+, \hat{A}_1 X_0^+),$$

where

$$R_3 = \frac{1}{2} \sum_{n=2}^{N} (t_{n-1}(X_{n-1}^-, (\hat{A}_n - \hat{A}_{n-1})X_{n-1}^-) + (t_n - t_{n-1})(X_{n-1}^-, \hat{A}_n X_{n-1}^-)).$$

Thus, by (2.2b), (2.2d)-(2.2f), we have

$$t_N \|X_N^-\|^2 + \sum_{n=1}^{N} t_n \int_{I_n} |A(t)X|^2\, dt \leqq C(R_1 + R_2 + R_3)$$
$$\leqq C\left( \sum_{n=1}^{N} |\dot{y}|_{L_1(I_n)} t_n \int_{I_n} |A(t)X|^2\, dt + \frac{1}{2\varepsilon} |g|^2 + \frac{\varepsilon}{2} t_1^2 |\hat{A}_1 X(s)^+|^2 \right.$$
$$\left. + \sum_{n=2}^{N} t_{n-1} \|X_{n-1}^-\|^2 \gamma_n + \sum_{n=2}^{N} \|X_{n-1}^-\|^2 k_n \right),$$

where $\varepsilon > 0$ and

$$\gamma_n = \int_{t_{n-2}}^{t_n} |\dot{y}(t)|\, dt.$$

Using now the easy-to-prove inverse estimates

$$|\hat{A}_1 X(s)^+|^2 \leqq \frac{C}{t_1} \int_{I_1} |\hat{A}_1 X(t)|^2\, dt, \qquad \|X_{n-1}^-\|^2 \leqq \frac{C}{k_{n-1}} \int_{I_{n-1}} \|X\|^2\, dt,$$

together with (2.29), and assuming that the constant $\gamma$ in (2.3b) and $\varepsilon$ are sufficiently small, we find that

$$t_N\|X_N^-\|^2+\sum_{n=1}^{N} t_n\int_{I_n}|A(t)X|^2\,dt\leqq C\left(|g|^2+\int_I\|X\|^2\,dt\right)+C\sum_{n=2}^{N}t_{n-1}\|X_{n-1}^-\|^2\gamma_n.$$

This inequality clearly holds with $N$ replaced by $n=1,\cdots,N$ and is of the form

$$a_n+b_n\leqq C\sum_{m=2}^{n}b_{m-1}\gamma_m+\bar{C}\equiv\psi_n,$$

where

$$a_n=t_n\|X_n^-\|^2,\qquad b_n=\sum_{m=1}^{n}t_m\int_{I_m}|A(t)X|^2\,dt,\qquad \bar{C}=C\left(|g|^2+\int_I\|X\|^2\,dt\right).$$

We have $\psi_1=\bar{C}$ and for $n=2,\cdots,N$,

$$\psi_n-\psi_{n-1}=Cb_{n-1}\gamma_n\leqq C\psi_{n-1}\gamma_n,$$

that is,

$$\psi_n\leqq(1+C\gamma_n)\psi_{n-1}$$

so that by (2.3a)

$$a_n+b_n\leqq\bar{C}\exp\left(C\sum_{m=2}^{n}\gamma_m\right)\leqq C\bar{C},\qquad n=1,\cdots,N,$$

and hence in particular

$$\sum_{n=1}^{N}t_n\int_{I_n}|A(t)X|^2\,dt\leqq C\left(|g|^2+\int_I\|X\|^2\,dt\right).\tag{2.31}$$

To be able to complete the proof using (2.30) we also need (in the case $q=1$) some control of $\dot{X}$ on $I_1$ to bound $|X(s)^+|$. Taking $v=t_1\dot{X}$ on $I_1$ and $v(t)=0$ for $t_1<t<t_N$ in (2.27), we easily get

$$t_1^2|\dot{X}|_{I_1}^2\leqq C\left(t_1\int_{I_1}|AX|^2\,dt+|g|^2\right),$$

so that

$$\begin{aligned}|X(s)^+|^2&\leqq(|X_0^+|^2+t_1^2|\dot{X}|_{I_1}^2)\\&\leqq C\left(|X_0^+|^2+t_1\int_{I_1}|AX|^2\,dt+|g|^2\right).\end{aligned}$$

Combining now (2.31) and (2.30), we find that

$$|X_0^+|^2+\int_I\|X\|^2\,dt+\sum_{n=1}^{N}t_n\int_{I_n}|\hat{A}X|^2\,dt\leqq C|g|^2,$$

and (2.28) follows also in the case $q=1$. This completes the proof of Lemma 1. □

**2.5. End of proof in the case $q=0$.** Recalling the error representation (2.22) with $s=\bar{t}_N$ and using the property (2.1a) of the interpolant $\tilde{Y}$, we have

$$\begin{aligned}|e_N^-|^2&=\int_I(\eta,A^*Z)\,dt+(\tilde{D}-D)(e,Z)\\&\leqq|\eta|_I\int_I|A^*Z|dt+\int_I|e|\,\|e\|\,\|Z\|\,dt,\end{aligned}$$

where we also used (2.2d). Thus using (2.23) and the fact that $e(t) = y(t) - Y_N^- = y(t) - y(t_N) + y(t_N) - Y_N^- = \eta(t) - e_N^-$ for $t \in I_N$, we have

$$c|e|_N^2 \leqq |\eta|_I \int_I |A^*Z|\, dt + |\eta|_N^2 + \sum_{n \leqq N} |e|_n^2 \int_{I_n} \|e\|^2\, dt.$$

Hence by a discrete Gronwall inequality and (2.17) with $\gamma$ sufficiently small,

$$c|e|_N^2 \leqq |\eta|_I \int_I |A^*Z|\, dt + |\eta|_I^2,$$

so that finally by (2.24)

$$\begin{aligned} |e|_N &\leqq CL_N |\eta|_I \leqq CL_N \max_{n \leqq N} \int_{I_n} |\dot{y}|\, dt \\ &\leqq CL_N \max_{n \leqq N} k_n |\dot{y}|_n, \end{aligned}$$

which proves (2.9) in the case $q = 0$.

**2.6. End of proof in the case $q = 1$.** By the error representation (2.22) with $s$ ranging over $(t_{N-1}, t_N]$ we find as in the case $q = 0$

$$|e|_N \leqq CL_N |\eta|_I \leqq CL_N \max_{n \leqq N} k_n^2 |y^{(2)}|_n. \tag{2.32}$$

To prove the "third-order" estimate (2.10) we choose $s = t_N$ in (2.22) to get using (2.2d), (2.2e), (2.24), and (2.1),

$$\begin{aligned} |e_N^-|^2 &= \int_I (\eta, A^*Z)\, dt + (\tilde{D} - D)(e, Z) \\ &\leqq \int_I (A\eta, Z)\, dt + \int_I |e|^2 |A^*Z|\, dt \\ &= \sum_{n=1}^N \left( \int_{I_n} (A_n\eta, Z_n^- + (t - t_n)\dot{Z})\, dt + \int_{I_n} (\eta, (A - A_n)^* Z(t))\, dt \right) + CL_N |e|_I^2 |e_N^-| \\ &\leqq CL_N \max_{n \leqq N} |A_n\eta|_n k_n |e_N^-| + CL_N \max_{n \leqq N} |\eta|_n |\dot{y}|_{L_1(I_n)} |e_N^-| + CL_N |e|_I^2 |e_N^-|, \end{aligned}$$

so that using (2.32)

$$|e_N^-| \leqq CL_N \max_{n \leqq N} k_n^3 (|A_n y^{(2)}|_n + |\dot{y}|_n |y^{(2)}|_n) + CL_N |e|_I^2,$$

which proves (2.10). This completes the proof of Theorem 2.

**3. Comparison with earlier methods for time-step control.** In this section we briefly review some earlier techniques for time-step control for stiff ordinary differential equations and compare with the new approach in this note.

For simplicity of discussion let us then first consider the backward Euler method (0.4) for (0.1). We recall that our time-step control in this case reads (cf. (0.10)):

$$|Y_n - Y_{n-1}| \sim \frac{\delta}{C}, \qquad n = 1, \cdots, N, \tag{3.1}$$

and is based on the optimal a priori error estimate

$$|e|_n \leqq CL_n \max_{m \leqq n} k_m |\dot{y}|_m, \qquad n = 1, \cdots, N. \tag{3.2}$$

The usual error estimate for the backward Euler method is of the form (cf. (0.7)),

$$|e|_n \leqq C \sum_{m=1}^{n} k_m \int_{I_m} |y^{(2)}(t)|\, dt \leqq C \sum_{m=1}^{n} k_m^2 |y^{(2)}|_m. \tag{3.3}$$

The right-hand side in (3.3) is basically a sum of so-called *local truncation errors* $\varepsilon_m$ defined by

$$\varepsilon_m \equiv k_m \left| \dot{y}(t_m) - \frac{y(t_m) - y(t_{m-1})}{k_m} \right| \leqq \frac{1}{2} k_m^2 |y^{(2)}|_m. \tag{3.4}$$

Thus, the usual error estimate (3.3) states that the global error $|e|_n$ can be estimated by the sum of the local truncation errors $\varepsilon_m$ for $m \leqq n$, that is assuming $C = 1$ for simplicity,

$$|e|_n \leqq \sum_{m \leqq n} \varepsilon_m. \tag{3.5}$$

Note that in the derivation of (3.2) given above we do *not* use the concept of local truncation error as defined in (3.4), which makes it possible to avoid the second derivative $y^{(2)}$ occurring in (3.3) via (3.4). Thus our error analysis is different from the usual analysis where the local truncation error is a fundamental concept. Note that if $\dot{y}$ is rapidly oscillating then the right-hand side of (3.3) will be much larger than the right-hand side of (3.2), and thus (3.2) in certain cases is sharper than (3.3). On the other hand, if $\dot{y}$ is not oscillating then the two right-hand sides may be of the same size.

Earlier methods for time-step control for the backward Euler method seek their theoretical basis in (3.5) (cf., e.g., Shampine and Gear [15]). Two different methods motivated by (3.5) have been proposed namely the *error per unit step* and the *error per step* method. In the error per unit step method we seek to choose the time steps $k_m$ so that the local truncation error per unit step is of order $\delta$, that is,

$$\frac{\varepsilon_m}{k_m} \sim \delta, \tag{3.6}$$

in which case by (3.5),

$$|e|_n \leqq C \sum_{m \leqq n} \varepsilon_m = \sum_{m \leqq n} \frac{\varepsilon_m}{k_m} k_m \sim \delta, \tag{3.7}$$

resulting in global error control as desired. However, error per unit step control may lead to inefficient time-step sequences (cf., e.g., Lindberg [12]) with unnecessarily small time steps in, e.g., an initial transient and therefore this method is usually not recommended.

In the error per step method we introduce a new parameter $\varepsilon$ the so-called *local error tolerance*, and we seek to choose the time steps $k_m$ so that

$$\tfrac{1}{2} k_m^2 |y^{(2)}|_m \sim \varepsilon, \qquad m = 1, \cdots, N. \tag{3.8}$$

In this case we have by (3.5) for the global error

$$|e|_N \leqq N\varepsilon. \tag{3.9}$$

The error per step method, where we thus seek to make all the local truncation errors roughly equal to the local error tolerance $\varepsilon$, seems to produce reasonable time-step

sequences in many cases. To implement the error per step method we face the following two problems:

(3.10) How do we choose $\varepsilon$ for a given $\delta$?

(3.11) How do we estimate $|y^{(2)}|_m$ to determine $k_m$ via (3.8)?

Let us first consider (3.10). With (3.9) as a basis for the error control we want to choose $\varepsilon$ so that $N\varepsilon \sim \delta$. However the total number of steps $N$ is not known in advance and thus it is not obvious how to choose $\varepsilon$ for a given $\delta$. One way to solve this problem would seem to be to first guess $\varepsilon$, perform a first computation choosing $k_m$ to satisfy (3.8), check if $N\varepsilon \leqq \delta$ and if not repeat the computation with a (suitably chosen) smaller $\varepsilon$ until $N\varepsilon \leqq \delta$.

We now turn to problem (3.11). To estimate $|y^{(2)}|_m$ various approaches have been used. A usual idea in the case of Runge–Kutta-type methods is to compare the computed $Y_m$ with the result $Y_m^*$ of a *higher order method* with the hope that

$$\varepsilon_m \sim |Y_m^* - Y_m|. \tag{3.12}$$

This would follow if we assume that $Y_{m-1} = y(t_{m-1})$ from the fact that if $y$ is sufficiently smooth then $|y(t_m) - Y_m^*| = \mathcal{O}(k_m^3)$ while $|y(t_m) - Y_m| = \varepsilon_m = \mathcal{O}(k_m^2)$. Alternatively, $Y_m^*$ could be the result of taking two time steps of size $k_m/2$ with the original method starting from $Y_{m-1}$. Another possiblity would be to estimate $|y^{(2)}|_m$ by a difference quotient using, e.g., the values $y_m$, $y_{m-1}$ and $y_{m-2}$.

To sum up we may say that it appears in principle to be possible to answer (3.10)–(3.11) and thus use the error per step method (3.8) with $\varepsilon$ determined through repeated calculations and $|y^{(2)}|_m$ estimated, e.g., by comparison with a higher order or more accurate method. Such an algorithm could be expected to pick a correct time-step sequence in a case when (3.3) is a tight estimate. However, the indicated method is not entirely satisfactory since repeated solution would be required to determined $\varepsilon$ *and* a higher order computationally more expensive method would be used *only* for the step control. These drawbacks seem to have so far prevented use of the error per step method in the indicated form in usual adaptive ordinary differential equation codes.

We note that in the new method (3.1) the problems related to the error per step method just stated do not occur: In (3.1) there is no local tolerance $\varepsilon$ to determine and instead of $|y^{(2)}|_m$ it is sufficient to estimate $|\dot{y}|_m$ which is much easier computationally.

We now turn to an adaptive method, local extrapolation [17], that seems to be frequently used in existing codes and appears to perform in a satisfactory way in many cases. This method may be described starting from the basic error per step method discussed above where a higher order method is used to estimate the local truncation error as indicated. The new method is simply obtained from the basic error per step method using the following modifications:

(3.11a) Use the values of the higher order method;

(3.11b) Choose $\varepsilon = \delta$.

These modifications seem to eliminate the drawbacks of the basic error per step method discussed above; now the results of the higher order method seem to be properly used and we have a rule for choosing $\varepsilon$! Note in particular that if (3.11a) is used, then the time steps will in fact be determined by comparison with a *lower order* method. The heuristic motivation for (3.11b) appears to be that the use of the results of the higher

order method allows us to "forget" the factor $N$ in (3.9) and thus choose the local tolerance equal to the global tolerance. However, no theoretical justification of the method in the case of stiff problems seems to be available.

We note that (3.1) may be viewed, if we want, as based on comparing the result $Y_n$ of the original first-order method with the result $Y_n^*$ of again a lower order method, namely the trivial zero order method $Y_n^* = Y_{n-1}$. Thus, also the time-step control in (3.1) may be viewed as obtained through comparison with a lower order method.

We now conclude with an argument that may help to give a theoretical justification for the stiff problems considered here of the modified error per step method (3.11) by viewing this method from the perspective of our approach to time-step control. Suppose then that we again start out with a first-order method like the backward Euler method with the conventional error estimate

$$|e|_I \leqq C \sum_{n \leqq N} k_n^2 |y^{(2)}|_n.$$

The basic conventional error per step strategy for this method would read

$$\varepsilon_n = k_n^2 |y^{(2)}|_n \sim \varepsilon \tag{3.13}$$

with $\varepsilon$ a local error tolerance to choose. Suppose next that to estimate $\varepsilon_n$ we introduce a higher order method and suppose we then actually use the results of this method with still the step control monitored by (3.13), and suppose we also choose $\varepsilon \sim \delta$. This method then would correspond to the modified error per step method (3.12) for which theoretical support seems to be lacking. Suppose now that the higher order method in fact admits an optimal error estimate of the form

$$|e|_I \leqq C \max_{n \leqq N} k_n^2 |y^{(2)}|_n, \tag{3.14}$$

which we have seen holds for example for DG (1) (see Theorem 2). The time-step control according to our new strategy for this method would then be

$$k_n^2 |y^{(2)}|_n \sim \frac{\delta}{C}. \tag{3.15}$$

If we now compare (3.13) and (3.15), we see that the conventional strategy (3.13) with $\varepsilon \sim \delta$ in fact would coincide with our new strategy (3.15). We thus arrive at the conclusion that, in fact, it may be possible to justify the modified error per step method (3.12) by using the theory proposed in this note.

*Remark* 2. The quantity $|y^{(p)}|_m$ in the time-step control (0.9) based on (0.6) may be estimated using a suitable difference quotient involving, e.g., the computed values $Y_{m-1}^-, Y_{m-2}^-, \cdots, Y_{m-1-p}^-$ (cf. (0.11) and [11] for the case $p = 1$, and [8] for $p = 2, 3$). The constant $C_m$ in (0.9) corresponding to the constant $CL_n$ in (2.9), which is related to (2.24a) through the representation formula (2.22), can probably be estimated by solving the discrete dual problem (2.20) with $A(t) = f(y(t))$ replaced by $f(Y(t))$ and $\theta(s)^-$ replaced by a general $v \in \mathbb{R}^M$ with $|v| = 1$. Thus, it seems to be possible to obtain a full quantitative control of the global error on a given tolerance level by repeated calculations, where $f'(y(t))$ is replaced by $f'(Y(t))$ with $Y(t)$ a previous approximation of $y(t)$. We plan to investigate this possiblity in detail in subsequent work. Note that appropriate constants $C$ in (2.9), which apply to the case $f(x) = Ax$ with $A$ positive definite, are given in [8].

*Remark* 3. Consider the problem (2.25) under the assumptions (2.2a) and (2.2b). Suppose there is a constant $C$ such that

$$(A(t)v, w)^2 \leqq C(A(t)v, v)(A(t)w, w) \quad \forall v, w \in \mathbb{R}^M, \quad t \in I, \tag{3.16a}$$

$$|\dot{A}(t)v| \leqq C|\dot{y}(t)||A(t)v| \quad \forall v \in \mathbb{R}^M, \quad t \in I. \tag{3.16b}$$

Then there is a constant $C$ such that

$$|A(t)x(t)| \leqq \frac{C}{t}|g|, \qquad t \in I. \tag{3.17}$$

To prove (3.17) we multiply (2.25a) by $t^2A^*(t)A(t)x(t)$ to get using (3.16a) with $v = x$ and $w = Ax$ and using also (3.16b)

$$\frac{1}{2}\frac{d}{dt}(t^2|Ax|^2) + t^2(A^2x, Ax) = t(Ax, Ax) + t^2(\dot{A}x, Ax)$$

$$\leqq \frac{1}{2}t^2(A^2x, Ax) + C(Ax, x) + C|\dot{y}(t)|t^2|Ax|^2.$$

Using Gronvall's inequality, (2.2a) and recalling (2.26), we obtain (3.17).

In a similar way one can prove an analogue of (3.17) for the discrete problem (2.20) from which (2.24a), (2.24b) follow with now $L_N = 1 + \log(t_n/k_n)$. This leads to a proof of Theorem 2 with (2.2f) replaced by (3.15a). □

## REFERENCES


[1] K. BURRAGE AND J. G. BUTCHER, *Stability criteria for implicit Runge-Kutta methods*, SIAM J. Numer. Anal., 16 (1979), pp. 46-57.

[2] G. DAHLQUIST, *Error analysis for a class of methods for stiff nonlinear initial value problems*, in Numerical Analysis, G. A. Watson, ed., Dundee, 1975, Lecture Notes in Mathematics 506, Springer-Verlag, Berlin, New York, Heidelberg, 1976.

[3] ———, *On the control of the global error in stiff initial value problems*, in Numerical Analysis, Dundee 1981, Lecture Notes in Mathematics 912, Springer-Verlag, Berlin, pp. 38-49.

[4] ———, *G-stability is equivalent to A-stability*, BIT, 18 (1978), pp. 384-401.

[5] M. DELFOUR, W. HAGER, AND F. TROCHU, *Discontinuous Galerkin methods for ordinary differential equations*, Math. Comp., 36 (1981), pp. 455-473.

[6] K. ERIKSSON, C. JOHNSON, AND V. THOMÉE, *Time discretization of parabolic problems by the discontinuous Galerkin method*, RAIRO MAN, 19 (1985), pp. 611-643.

[7] K. ERIKSSON AND C. JOHNSON, *Error estimates and automatic time step control for nonlinear parabolic problems*, I, SIAM J. Numer. Anal., to appear.

[8] K. ERIKSSON, C. JOHNSON, AND J. LENNBLAD, *Optimal error estimates and adaptive time and space step control for linear parabolic problems*, Technical Report, Chalmers University of Technology, Göteborg, Sweden, 1986; SIAM J. Numer. Anal., submitted.

[9] R. FRANK, J. SCHNEID, AND C. UEBERHUBER *The concept of B-convergence*, SIAM J. Numer. Anal., 18 (1981), pp. 573-780.

[10] C. JOHNSON, *Error estimates and automatic time step control for numerical methods for stiff ordinary differential equations*, Technical Report 1984-27, Chalmers University of Technology, Göteborg, Sweden, 1984.

[11] C. JOHNSON, Y. Y. NIE, AND V. THOMÉE, *An a posteriori error estimate and automatic time step control for a backward Euler discretization of a parabolic problem*, Technical Report, Chalmers University of Technology, Göteborg, Sweden, 1985; SIAM J. Numer. Anal., submitted.

[12] B. LINDBERG, *Characterization of optimal stepsize sequences for methods for stiff differential equations*, SIAM J. Numer. Anal., 14 (1977), pp. 859-889.

[13] O. NEVANLINNA AND F. ODEH, *Multiplier techniques for linear multistep methods*, Numer. Funct. Anal. Optim., 3 (1981), pp. 377-423.

[14] O. NEVANLINNA AND R. JELTSCH, *Error bounds for multistep methods revisited*, preprint.

[15] L. F. SHAMPINE AND C. W. GEAR, *A user's view of solving stiff ordinary differential equations*, SIAM Rev., 21 (1979), pp. 1-17.

[16] V. THOMÉE, *Galerkin finite element methods for parabolic problems*, Lecture Notes in Mathematics 1054, Springer-Verlag, Berlin, New York, Heidelberg, 1984.

[17] L. F. SHAMPINE AND M. K. GORDON, *Computer Solution of Ordinary Differential Equations*, W. H. Freeman, San Francisco, 1975.