Global Error Control for the Continuous Galerkin Finite Element Method for Ordinary Differential Equations

Donald Estep
Donald French

CRPC-TR93383
November 1993

Center for Research on Parallel Computation
Rice University
P.O. Box 1892
Houston, TX 77251-1892

GLOBAL ERROR CONTROL FOR THE CONTINUOUS GALERKIN FINITE ELEMENT METHOD FOR ORDINARY DIFFERENTIAL EQUATIONS

DONALD ESTEP AND DONALD FRENCH

ABSTRACT. We analyse a continuous Galerkin finite element method for the integration of initial value problems in ordinary differential equations. We derive quasi-optimal a priori and a posteriori error bounds. We use these results to construct a rigorous and robust theory of global error control. We conclude by exhibiting the properties of the error control in a series of numerical experiments.

§1. INTRODUCTION

Our main purpose is to outline a rigorous theory of global error control for the continuous Galerkin finite element method for

$$\begin{cases} \dot{y} + f(y,t) = 0, & 0 < t \leq T, \\ y(0) = y_0 \in \mathbf{R}^d, & d \geq 1. \end{cases} \tag{1.1}$$

The continuous Galerkin (cG) method produces a continuous piecewise polynomial approximation Y. It has been used previously for certain evolution problems (see [10],[11], [12]) because it often has the property of preserving an "energy" naturally associated to the differential equation. We are interested in adaptive error control for the cG method in order to achieve *accuracy* and *efficiency* in computations. On one hand, it is computationally impractical and even impossible to use a uniform (small) step-size on many problems. Examples are systems obtained from a method of lines discretization of a partial differential equation and problems which require computations over long time intervals. On the other hand, it is generally impossible to a priori choose step-sizes that guarantee accuracy. However, we show that information obtained from the approximation can be used to make computations of a specified accuracy.

The theory of adaptive error control we describe is based on a combination of rigorous a priori and a posteriori error analyses. This is the same approach that

1991 *Mathematics Subject Classification.* 65L05, 65L50, 65L60.

Key words and phrases. a posteriori error bounds, a priori error bounds, adaptive error control, continuous Galerkin finite element method, global error control, one step method, ordinary differential equation, stiff problems.

The work of the first author was supported in part by the National Science Foundation Cooperative Agreement No. CCR-8809615 and Grant No. DMS-9208684

The work of the second author was supported in part by the Army Research Office through Grant No. 28535-MA

Typeset by AMS-TEX

has been used in an ongoing project to develop a theory of adaptive error control for approximations of partial differential equations (see [2]–[8], [14] and references therein).

A priori error bounds measure the error by quantities that reflect the regularity of the solution and the stability properties of the numerical scheme. The usual derivation for a difference scheme is based on estimation of the truncation error by means of Taylor's theorem. In contrast, we use Galerkin orthogonality to compare the cG approximation to other approximations in the finite element space. Hence, we obtain optimal order results rather than the usual sub-optimal bounds derived for difference schemes. In addition, we prove that the second order cG approximation is superconvergent at time nodes, i.e. has an extra order of accuracy at those points.

While a priori error bounds describe the convergence properties of an approximation, they are not directly useful for error control because they involve unknown information about the solution. Instead, we use a posteriori error bounds as adaptive criteria for choosing step-sizes. An a posteriori bound measures the error by *computable* quantities that depend on the regularity of the approximation and the stability properties of the solution. Suppose that the interval $[0,T]$ is partitioned into N subintervals I_m of length k_m, and that q denotes the order of the cG approximation. Our a posteriori bounds have the form

$$|Y(t_n) - y(t_n)| \leq S(t_n) \max_{m \leq n} k_m^{q+1} \max_{I_m} |D_t^q f(Y(t),t)|, \tag{1.2}$$

for $1 \leq n \leq N$, where $|\cdot|$ denotes the Euclidean norm on $\mathbf{R}^d$, D_t^q denotes the q^{th} order time derivative and $S(t_n)$ depends on t_n but not on any k_m. Note that $k_m^{q+1} \max_{I_m} |D_t^q f(Y(t),t)|$ is computed on each interval and measures the local regularity of the approximation. We call $S(t)$ the stability factor and it is a measure of the accumulation of error. It is given by a semi-norm on the solution of the linear dual problem to (1.1) obtained by linearizing (1.1) around the solution. We show that $S(t_n)$ can be approximated using the linear problem obtained by linearizing around the approximation. Hence, the bound involves only information that can be obtained from the approximation. If Y is computed so as to keep the quantities on the right-hand side of (1.2) below a given tolerance, then the error is also kept below the tolerance.

This theory of adaptive error control is completely different from the standard theories for difference schemes which are based on error estimates that are asymptotic in the limit of small step-size. Hence, we avoid some difficulties associated to this approach. While the asymptotic estimates are valid only when the error is small, there is no computational criteria for determining if the asymptotic regime has been reached with the chosen step-sizes. Thus, a small asymptotic estimate does not imply that the error is small. In fact, the criteria of choosing steps so as to keep the asymptotic estimates valid is generally harsher than computing approximations of a given accuracy. For example, this is essentially the root of the issue of choosing the error-per-step or the error-per-unit-step criteria for the widely-used strategy called local error control. Note that in this context, the goal of adaptive error control is to use as large as steps as possible while producing an approximation of the desired accuracy. Finally, asymptotic estimates require extra regularity of the

solution, which is of particular concern in applications to nonlinear initial-boundary value problems in partial differential equations.

This approach to error analysis and adaptive error control was initiated by Johnson in [14], which contains an a priori analysis of the discontinuous Galerkin (dG) method for autonomous ordinary differential equations. The dG method produces a discontinuous piecewise polynomial approximation that is well suited for stiff, dissipative problems. Eriksson and Johnson made complete a priori and a posteriori analyses of the dG method for linear parabolic problems in [3] as well as outlined a theory of error control. Estep did the same for the dG method for non-autonomous ordinary differential equations in [8]. This analysis has been extended in several directions in recent years, see [2]–[7].

We would like to extend the theory to cover general numerical methods for a variety of equations and this paper is a step towards this goal. It is natural to consider the cG method as an alternative to the dG method because its stability properties make it more suitable for equations with oscillatory and periodic solutions than the dG method (see §2 and §4). The analysis we present here follows the same lines as the analysis in [8], however the technical details are altered to account for the differences between the cG and dG methods. In particular, we deal with difficulties associated to the fact that in the cG method, the approximation space and the test space are different.

The paper is arranged as follows. In §2, we introduce notation and describe the cG method. In §3, we present the a priori and a posteriori results. In §4, we outline the strategy for adaptive error control based on the a posteriori result as well as discuss some technical points concerning implementation. We demonstrate the adaptive method on four test problems, including the difficult two-body problem. In particular, we present plots of the error-to-bound ratio as a measure of reliability and efficiency. We also make a comparison with the dG method on these test cases. We present the proofs of the a priori results in §5 and of the a posteriori results in §6.

§2. The scheme and notation

The finite element method is based on a variational formulation of (1.1) that reads: find $y \in C_1((0,T))$ such that

$$\begin{cases} \displaystyle\int_0^T (\dot{y}, v)dt + \int_0^T (f(y(t),t), v(t))dt = 0, \\ y(0) = y_0, \end{cases} \tag{2.1}$$

for all $v \in C_1((0,T))$, where $C_p([0,T])$ denotes the set of functions with continuous derivatives of order p and less on $[0,T]$.

We construct a piecewise polynomial approximation Y to y. We partition $[0,T]$ into

$$0 =: t_0 < t_{1/2} < t_1 < t_{3/2} < t_2 < \cdots < t_N =: T,$$

setting $k_m := t_m - t_{m-1}$, $t_{m-1/2} := t_m - k_m/2$, $I_m := [t_{m-1}, t_m]$, and $k := \max_m k_m$. We choose the finite element space $C^{(q)} = C^{(q)}([0,T])$ of continuous

functions that are polynomials of degree q on each interval I_m, i.e.

$$C^{(q)} := \{U \in C_0([0,T]) : U|_{I_m} \in \mathcal{P}^{(q)}(I_m), 1 \le m \le N\},$$

where $\mathcal{P}^{(q)}(I_m)$ denotes the set of polynomials in $\mathbf{R}^d$ of degree q on I_m. Because of the continuity, a function in $C^{(q)}$ has only q degrees of freedom on each interval. Accordingly, we define the test space

$$\mathcal{D}^{(q-1)} = \mathcal{D}^{(q-1)}([0,T]) := \{U : U|_{I_m} \in \mathcal{P}^{(q-1)}(I_m), 1 \le m \le N\}.$$

Since these functions may be discontinuous, we let $U_m^{+,-}$ denote the left- and right-hand limits of $U \in \mathcal{D}^{(q-1)}$ at t_m and $[U]_m := U_m^+ - U_m^-$ the jump in value.

For $1 \le n \le N$, the cG approximation $Y \in C^{(q)}$ solves

$$\begin{cases} \sum_{m=1}^{n} \int_{I_m} (\dot{Y}(t) + f(Y(t),t), V(t))dt = 0, \\ Y(0) = y_0. \end{cases}$$

for all $V \in \mathcal{D}^{(q-1)}$. Y can be computed interval by interval as well; for $1 \le m \le n$, it solves

$$\begin{cases} \int_{I_m} (\dot{Y}, V)dt + \int_{I_m} (f(Y(t),t), V(t))dt = 0, \\ \lim_{t \to t_{m-1}^-} Y(t) = \lim_{t \to t_{m-1}^+} Y(t). \end{cases} \tag{2.2}$$

for all $V \in \mathcal{P}^{(q-1)}(I_m)$.

When $q = 1$, Y is the piecewise linear function

$$Y|_{I_m} = Y_{m-1}\frac{(t - t_m)}{-k_m} + Y_m\frac{(t - t_{m-1})}{k_m}$$

with coefficient Y_m determined by

$$Y_m + \int_{I_m} f(Y(t),t)dt = Y_{m-1}. \tag{2.3}$$

When $q = 2$, Y is the piecewise quadratic function on I_m

$$\begin{aligned} Y|_{I_m} = {} & Y_{m-1}\frac{2}{k_m^2}(t - t_{m-1/2})(t - t_m) \\ & - Y_{m-1/2}\frac{4}{k_m^2}(t - t_{m-1})(t - t_m) + Y_m\frac{2}{k_m^2}(t - t_{m-1})(t - t_{m-1/2}), \end{aligned}$$

with coefficients determined by

$$\begin{cases} Y_m + \int_{I_m} f(Y(t),t)dt = Y_{m-1}, \\ Y_{m-1/2} + \int_{I_m} f(Y(t),t)\frac{5t_m + t_{m-1} - 6t}{4k_m}dt = Y_{m-1}. \end{cases} \tag{2.4}$$

Existence and uniqueness can be shown for k sufficiently small.

Remark 2.1. Consider $f(y,t) = \lambda y$. The $q = 1$ scheme (with uniform step) is

$$Y_m = \frac{1 - \lambda k/2}{1 + \lambda k/2} Y_{m-1},$$

which agrees with the second order trapezoidal rule at nodes. When $q = 2$, the cG approximant agrees at nodes with the fourth order Runge-Kutta scheme

$$Y_m = \frac{1 - \lambda k/2 + \lambda^2 k^2/12}{1 + \lambda k/2 + \lambda^2 k^2/12} Y_{m-1}. \tag{2.5}$$

Both cG schemes are A-stable; stable for $\mathcal{R}e(\lambda) \geq 0$ and unstable for $\mathcal{R}e(\lambda) < 0$. In particular, when $\mathcal{R}e(\lambda) > 0$, y decreases in size as time increases and the cG approximations have the same property without restriction on the step-size, as would be needed for an explicit scheme for example. However, note that for fixed k, the factors above tend to one in magnitude as $|\lambda| \to \infty$. This contrasts both with the behavior of the solution and the discontinuous Galerkin method. On the other hand, when $\mathcal{R}e(\lambda) = 0$ and y is purely oscillatory, the cG approximants are neither increasing or decreasing in magnitude, indicating that the cG method might be a good choice for problems with periodic and oscillatory solutions. The discontinuous Galerkin method is not particularly suitable for such problems because the approximants exhibit numerical damping. See §4 for further discussion on this issue.

We also note that the cG schemes preserves discrete versions of the conservation properties or Lyapunov functionals which the original system might possess (see [12]). This property can be used to analyze the stability properties and the long time behavior of the numerical scheme (see [10] and [11]).

The cG schemes are not equivalent to any standard Runge-Kutta schemes when applied to truly nonlinear, non-autonomous problems.

We use the following notation. For an time interval I, we let

$$|g|_I := \sup_{t \in I} |g|.$$

In addition,

$$e := y - Y$$

$$f_y(y,t) := \frac{\partial f}{\partial y}, f_t(y,t) := \frac{\partial f}{\partial t}, f_{yt}(y,t) := \frac{\partial f_y}{\partial t}, \text{ etc.},$$

$$A(t) := f_y(y(t),t), \dot{A} := \frac{d}{dt}A(t), \hat{A} := \frac{1}{2}(A + A^*),$$

$$u^{(p)} := \frac{d^p u}{dt^p},$$

$$\mathcal{N}_\delta := \{(u,t) : u \in C_0([0,T]) \text{ and } |u(t) - y(t)| \leq \delta \text{ for } 0 \leq t \leq T\}.$$

We employ several interpolants and projections of u in $C^{(q)}$. $\tilde{U}$ satisfies

$$\tilde{U}(0) = u(0),$$

and on I_m, $1 \le m \le n$,

$$\begin{cases} \dot{\tilde{U}}(t_m) = \dot{u}(t_m), & q = 1, 2, \\ \dot{\tilde{U}}(t_{m-1}) = \dot{u}(t_{m-1}), & q = 2. \end{cases}$$

The following error bound is easy to show

$$|\dot{\tilde{U}} - \dot{u}|_{I_m} \le \begin{cases} k_m |\tilde{u}|_{I_m}, & q = 1, 2, \\ k_m^2 |u^{(3)}|_{I_m}, & q = 2. \end{cases} \tag{2.6}$$

$\mathcal{P}_{\mathcal{D}}$ denotes the L^2 projection of u into $\mathcal{D}^{(q-1)}$; in other words, $\mathcal{P}_{\mathcal{D}} u \in \mathcal{D}^{(q-1)}$ satisfies

$$\int_0^T (\mathcal{P}_{\mathcal{D}} u, V) dt = \int_0^T (u, V) dt,$$

for all $V \in \mathcal{D}^{(q-1)}$. In places, we abuse notation to let $\mathcal{P}_{\mathcal{D}}$ denote $\mathcal{P}_{\mathcal{D}}|_{I_m}$. By standard results,

$$|\mathcal{P}_{\mathcal{D}} u - u|_{I_m} \le \begin{cases} \int_{I_m} |\dot{u}| dt \le k_m |\dot{u}|_{I_m}, & q = 1, 2, \\ C k_m \int_{I_m} |\tilde{u}| dt \le C k_m^2 |\tilde{u}|_{I_m}, & q = 2. \end{cases} \tag{2.7}$$

For $u \in C_{p+1}(I)$, $0 \le p \le q$, we let $\mathcal{I}_q u \in C^{(q)}$ denote the interpolant such that on I_m, $1 \le m \le n$,

$$\begin{cases} \mathcal{I}_q u = u \text{ at } t_{m-1}, t_m, & q = 1, 2, \\ \mathcal{I}_q u = u \text{ at } t_{m-1/2}, & q = 2. \end{cases}$$

From standard theory (see [16]), there is a constant $C = C(p)$ such that

$$\left| \frac{d^r}{dt^r} (u - \mathcal{I}_q u) \right|_{I_m} \le C k_m^{p+1-r} |y^{(p+1)}|_{I_m}, \tag{2.8}$$

for $0 \le r \le p \le q$. We recall the following *inverse* result from Ciarlet ([1]), there is a $C > 0$ such that for $U \in C^{(q)}$, $1 \le r \le \infty$, $1 \le s \le \infty$, and $0 \le p \le q$,

$$\left(\int_{I_m} |U^{(p)}|^r dt \right)^{1/r} \le C k_m^{-p-(1/s-1/r)} \left(\int_{I_m} |U|^s dt \right)^{1/s}, \tag{2.9}$$

(with the usual interpretation if r or $s = \infty$).

Finally, to simplify the presentation of the proofs, we use a global Lipschitz assumption on f and define the domain $\Omega := \mathbf{R}^d$. More sophisticated assumptions can be used without significantly altering either the results or the proofs.

§3. Error Analysis

We now present the analysis of the error that serve as the basis for error control. We begin with an a priori analysis that reveals the convergence properties of the cG method. The a priori bound measures the error by quantities that depend on the regularity of the solution and the stability properties of the scheme. The first result shows that the cG scheme converges at the optimal order on the interval $[0, T]$.

Theorem 3.1. *Assume that $y \in C_{q+1}([0,T])$ and f is Lipschits continuous with constant L. Then, for $1 \le n \le N$, $q = 1, 2$, and k sufficiently small,*

$$|e|_{[0,t_n]} \le C\big(1 + Lt_n e^{CLt_n}\big)^{1/2} \max_{m \le n} k_m^{q+1} |y^{(q+1)}|_{I_m}, \tag{3.1}$$

and

$$|\dot{e}|_{[0,t_n]} \le C\big(1 + Lt_n e^{CLt_n}\big)^{1/2} \max_{m \le n} k_m^{q} |y^{(q+1)}|_{I_m}. \tag{3.2}$$

The second result shows that the cG method is *superconvergent* at time nodes for $q = 2$, if the ratio of the largest step to the smallest step is bounded. The order of convergence at nodes ($2q$) agrees with the order of convergence of the Runge-Kutta scheme (2.5) for the linear problem in remark 2.1. But, we note that the form of the bound means that there can be an effective loss of order if the problem is stiff, for example. We thank J. Schaeffer [16] for giving us the proof of this result.

Theorem 3.2. *Assume that $q = 2$, that $y \in C_3(I)$, that for δ sufficiently small, the partial derivatives of f of order q are continuous and bounded in norm by L on $\mathcal{N}_\delta$, and that there is a constant $\rho > 0$, independent of k and N, such that for $1 \le n \le N$,*

$$\max_{m \le n} \frac{k_m}{k_n} \le \rho.$$

Then, for $1 \le n \le N$ and k sufficiently small,

$$|e(t_n)| \le Ct_n e^{CLt_n}\big(1 + Lt_n e^{CLt_n}\big)^2 \max_{m \le n} k_m^4 \cdot \max_{\substack{r+s \le 3 \\ r,s \ge 0}} |y^{(r)}|^s (1 + |y^{(r)}|^s), \tag{3.3}$$

where $C = C(L, \rho)$.

The proofs of these results are given in §5. In both cases, the analysis uses Galerkin orthogonality to compare the cG approximation to approximations of the solution in $C^{(q)}$ with known interpolation errors. But, unlike approximations of the solution computed with full global knowledge of the solution, the error in the cG approximation accumulates with time. This is the reason for the exponentially increasing stability factors. The accumulation has to account for the worse possible rate of accumulation in the class of problems under consideration. Under these general assumptions, these large factors are the best possible.

These a priori results are not useful for error control because they involve unknown information about the solution. Next, we present an a posteriori result that bounds the error by computable quantities that reflect the regularity of the approximation and the stability properties of the linearized dual problem to (1.1). As to the latter, we let z denote the solution of

$$\begin{cases} -\dot{z} + A^*(t)z = 0, & t_n > t \ge 0, \\ z(t_n) = e_n/|e_n|, \end{cases} \tag{3.4}$$

and define a quantity that turns out to be the stability factor for the a posteriori bound,

$$S_p(n) := \int_0^{t_n} |z^{(p)}| dt. \tag{3.5}$$

It is convenient to begin the analysis with an a priori bound on S_q.

Proposition 3.3. *Assume that there is an integrable, bounded function $L(t) \geq 0$ such that*

$$|A(t)| \leq L(t), \tag{3.3.1}$$

for $0 \leq t \leq T$. Then,

$$S_1(n) \leq e^{\int_0^{t_n} L(t)dt} - 1. \tag{3.6}$$

If, in addition,

$$(A(t)w, w) \geq 0, \tag{3.3.2}$$

for all $w \in \mathbf{C}^d$ and $t \in [0,T]$, then

$$S_1(n) \leq \int_0^{t_n} L(t)dt. \tag{3.7}$$

If we assume that the partial derivatives of f of order q and less are bounded in norm by $L(t)$ on $\mathcal{N}_\delta$, then

$$S_2(n) \leq \left(e^{\int_0^{t_n} L(t)dt} - 1\right)\left(e^{\int_0^{t_n} L(t)dt} + |L|_{[0,t_n]}\right). \tag{3.8}$$

This result is proved in §6.

Remark 3.1. There are other possibilities for bounds on $S_2(n)$. The case when $A(t)$ is invertible turns out to be computationally important. Then,

$$S_2(n) \leq \left(\left|\frac{\dot{A}^*}{A^*}\right|_{[0,t_n]} + |A|_{[0,t_n]}\right) S_1(n). \tag{3.9}$$

Now, the a posteriori result.

Theorem 3.4. *Assume that the q^{th} order partial derivatives of f are continuous on $\mathcal{N}_\delta$, for some $\delta > 0$, and that there is an integrable, bounded function $L(t) > 0$ such that for all $u, v \in \Omega$, $w, x \in \mathbf{C}^d$, and $t \in [0,T]$,*

$$|f(u,t) - f(v,t)| \leq L(t)|u - v|, \tag{3.4.1}$$

$$|(A(t)w, w)| \leq L(t)|w|^2, \tag{3.4.2}$$

and

$$|((f_y(u,t) - f_y(v,t))w, x)| \leq C \min\{L(t)|u-v||w||x|, |u-v||w||A^*(t)x|\}. \tag{3.4.3}$$

Then, there is a constant $C > 0$ such that for k sufficiently small,

$$|e|_{[0,t_n]} \leq C(S_1(n) + 1) \max_{m \leq n} k_m^{q+1} \left|\frac{d^q}{dt^q} f(Y(t), t)\right|_{I_m}, \tag{3.10}$$

for $1 \le n \le N$, $q = 1, 2$. *If* Y *is computed so that for some* $C > 0$,

$$k_m^{q+1} \left| \frac{d^q}{dt^q} f(Y(t), t) \right|_{I_m} \le C, \tag{3.4.4}$$

for $1 \le m \le n$, *then*

$$\max_{m \le n} |e_m| \le C S_1(n) \max_{m \le n} k_m^{q+1} \left| \frac{d^q}{dt^q} f(Y(t), t) \right|_{I_m}, \tag{3.11}$$

for $1 \le n \le N$, $q = 1, 2$. *Finally, for* $q = 2$,

$$\max_{m \le n} |e_m| \le C S_2(n) \max_{m \le n} k_m^4 \left| \frac{d^2}{dt^2} f(Y(t), t) \right|_{I_m} + C S_1(n) \max_{m \le n} k_m^6 \left| \frac{d^2}{dt^2} f(Y(t), t) \right|_{I_m}^2. \tag{3.12}$$

This result is proved in §6. We discuss the construction of an error control based on theorem 3.4 in §4. We conclude this section with a result that shows that the a posteriori bounds are of optimal order.

Proposition 3.5. *Assume that the hypotheses of theorem 3.1, 3.2, and 3.4 hold. Then, for* k *sufficiently small and* $q = 1, 2$,

$$\left| \frac{d}{dt} f(Y(t), t) - \dot{y} \right|_{I_m} \le C(L, T, y) k^q,$$

while for $q = 2$,

$$\left| \frac{d^2}{dt^2} f(Y(t), t) - y^{(3)} \right|_{I_m} \le C(L, T, \rho, y) k.$$

§4. Adaptive Error Control

We employ adaptive error control in order to achieve the related goals of accuracy and efficiency. In the case that a fixed scheme is used, this means producing approximations of a desired accuracy using the largest possible step-sizes. In this section, we show how the a posteriori error bound can be used as the basis for an adaptive error control and then exhibit properties of the control through a series of experiments.

There are two contributions to the global error in the approximation of an initial value problem (ignoring round-off error). The first is the interpolation error made in approximating a general function by piecewise polynomials. This error is determined by the behavior of certain derivatives of the solution. The first goal of the adaptive error control is to choose a mesh that allows interpolation of the solution with an error that is uniform in an appropriate norm over the interval of computation. The second source of error is due to the cumulative effects of integrating an initial value problem interval by interval. The rate of accumulation is determined

the stability properties of the solution, i.e. the behavior of trajectories that art near the target solution at a given time. The second goal of the adaptive ror control is to choose the mesh size so that the accumulated error is not too rge at specified times.

Examination of the proof of theorem 3.4 makes it clear that the result clearly elineates the two sources of error. For example, we quote (3.11)

$$|e_n| \leq CS_1(n) \max_{m \leq n} k_m^{q+1} \left| \frac{d^q}{dt^q} f(Y(t),t) \right|_{I_m}. \tag{4.1}$$

The quantities inside the maximum taken on the right measure the local approximation properties of the mesh for functions in $C^{(q)}$. The accounting of the accumulation of error is made in $S_1(n)$ which is a semi-norm of the solution of the dual problem over the interval of computation. By controlling the expression on the right in (4.1) at specified times t_n^*, $1 \leq n^* \leq N$, we control the global error at those points as well.

Remark 4.1. We use (3.11) and compute with the $q = 1$ scheme for the purpose of illustration. The other bounds in theorem 3.4 and the higher order scheme can be used in an analogous fashion.

Because the bound uses a maximum of local quantities, it does not seem wise to let the local quantities become large at some points. Hence, we adopt the following strategy: given LTOL > 0, for $1 \leq m \leq n$ compute Y_m on I_m so that

$$k_m^{q+1} \left| \frac{d^q}{dt^q} f(Y(t),t) \right|_{I_m} \leq \text{LTOL}. \tag{4.2}$$

Note that proposition 3.5 implies that (4.2) can be achieved by taking k_m sufficiently small. In practice, k_m is determined iteratively. From a given point t_{m-1}, a step k_m^{pred} is predicted via

$$k_m^{pred} = \left(\frac{\text{LTOL}}{\left| \frac{d^q}{dt^q} f(Y(t),t) \right|_{I_{m-1}}} \right)^{1/(q+1)}$$

and Y_m^{pred} is computed. If (4.2) is satisfied with $Y_m = Y_m^{pred}$, the step is accepted and the computation proceeds. If (4.2) is violated, the iteration is repeated with a new step is predicted via

$$\left(\frac{\text{LTOL}}{\left| \frac{d^q}{dt^q} f(Y^{pred}(t),t) \right|_{I_m}} \right)^{1/(q+1)}.$$

Global error control is achieved by choosing LTOL so that

$$CS_1(n^*)\text{LTOL} \leq \text{GTOL}, \tag{4.3}$$

for certain n^*, $1 \leq n^* \leq N$, where GTOL is the desired global error tolerance. If an a priori bound on $S_1(n)$ is known, then a correct LTOL can be chosen before the

computation begins. Unfortunately, the general bound on $S_1(n)$ given in proposition 3.3 is too large to allow computation past a short transient. Since $S_p(n)$ is specific to the solution of (1.1) that is being approximated, an alternative is to compute $S_p(n)$ for each problem. Of course, this is not directly possible in general precisely because this would require the solution. Instead, consider ζ solving

$$\begin{cases} -\dot{\zeta} + f_y(Y(t),t)^*\zeta = 0, & t_n > t \geq 0, \\ \zeta(t_n) = d_n, & |d_n| = 1. \end{cases} \tag{4.4}$$

We can prove (see [8])

Theorem 4.1. *Assume that f has continuous second partial derivatives and that the assumptions of theorems 3.1 or 3.2 and theorem 3.4 hold. Further assume that there is a constant $\delta > 0$ such that for all $U \in C^{(q)}$ with $\max\{|U - y|_{[0,T]}, \int_0^T |U - y|dt\} \leq \delta$,*

$$\max_{m \leq n} \max_{p \leq q+1} \left\{ \left| \frac{d^p}{dt^p} f(U(t),t) \right|_{I_m} \right\} \leq L.$$

Then, any consistent and stable one step method for (4.4) with $d_n = e_n^-/|e_n^-|$ computed on a mesh that includes $\{t_1, \cdots, t_N\}$ as nodes converges to z as $k \to 0$.

We approximate $S_1(n)$ by using the values of a trapezoidal rule approximation for (4.4) in a Simpson's rule formula for

$$\int_0^{t_n} |f_y(Y(t),t)^*\zeta| dt.$$

Similarly, we can approximate $S_2(n)$ by a quadrature formula for

$$\int_0^{t_n} \left| \frac{d}{dt} f_y(Y(t),t)^*\zeta + f_y(Y(t),t)^{*2}\zeta \right| dt.$$

Since this approximation of $S_p(n)$ requires an approximation of y over $[0, t_n]$, we resort to an iteration to achieve global error control. We begin by assuming that $S_q(n) = 1$ and LTOL = GTOL is chosen. Y is computed so as to satisfy (4.2). Next, $S_p(n^*)$ is approximated using Y and (4.3) is checked at the desired points t_{n^*}. If (4.3) is violated, a new local tolerance is chosen via

$$\text{LTOL}_{new} = \frac{\text{GTOL}}{CS_1(n^*)}.$$

Finally, the computation is restarted with the smallest of the new local tolerances (presuming (4.3) is violated at some point).

Remark 4.2. The constant C in the bounds in theorem 3.4 is determined by the technical details of the analysis and, hence, is somewhat problem dependent. For example, it depends on the choice of norms, and so on the dimension, on the largest possible step size, and so on. To determine C precisely, one would have to be more careful in the analysis than we have been. (For example, use optimal

estimates in each line, etc.) Instead, we compute C for a linear problem (presented in example 4.1) in which the solution is known and use this value in the rest of the computations, which are all low dimension. For many problems, this appears to be a reasonable value, though in some cases, the scale is clearly off.

Remark 4.3. The error control outlined above is robust in the sense that a step is accepted only if (4.3) is satisfied and a computation is accepted only if (4.3) is satisfied. In practice, it is possible that the iterative processes used to achieve (4.2) and (4.3) can produce approximations that are more accurate than requested. Whether this warrants recomputing the step or the computation is uncertain. For the computations below, we did not want computations that are too accurate. We choose η, $0 \leq \eta < 1$, and during the local step control, accepted a step only if

$$\eta \text{LTOL} \leq k_m^{q+1} \left| \frac{d^q}{dt^q} f(Y(t),t) \right|_{I_m} \leq \text{LTOL}, \tag{4.5}$$

while we accept the computation only if

$$\eta \text{GTOL} \leq CS_1(n^*) \leq \text{GTOL}, \tag{4.6}$$

at the desired points t_{n^*}. The modifications to the iterations outlined above are straightforward. We use $\eta = .5$ in the computations below.

Remark 4.4. An important issue in higher dimensions is the choice of the initial direction vector d_n. Note that theorem 4.1 requires the initial vector $e_n/|e_n|$, which is unknown of course. A satisfactory conclusion to this theory would be a result that measures the effect of perturbations in the initial vector in (4.4) on $S_p(n)$ together with an a posteriori estimate of $e_n^-/|e_n^-|$. A rough heuristic argument suggests that if the local interpolation errors in the a priori bound in theorems 3.1 is a good measure of the error on the corresponding intervals, then $\dot{Y}$ should point largely in the direction of the error. In practice, $d_n = (Y_n - Y_{n-1})/|Y_n - Y_{n-1}|$ has proven to be a reliable choice for many problems. It is not clear to us whether this is because the computation of the stability factor is insensitive to the choice of initial direction on many problems or because this is actually a good choice. In the computations below, we compare results computed with $d_n = e_n^-/|e_n^-|$ and $d_n = (Y_n - Y_{n-1})/|Y_n - Y_{n-1}|$.

Next, we present four examples. In each case, we implement the iterative global error control outlined above. For the successful computation, we present a plot of the error-to-bound ratio

$$\frac{|e_n|}{CS_1(n)\text{LTOL}}. \tag{4.7}$$

This is a convenient measure of reliability and efficiency. If the ratio becomes large, then reliability is suspect, while if the ratio becomes small, then the error control is inefficient.

Example 4.1. The problem is

$$\begin{cases} \dot{y}_1 = y_2, & y_1(0) = 0, \\ \dot{y}_2 = y_1, & y_2(0) = 1, \end{cases}$$

with the periodic solution

$$y_1(t) = \sin(t),$$
$$y_2(t) = \cos(t).$$

We use GTOL = .05. The error control iteration halts after two iterations. The first iteration uses LTOL = .045 and the second uses LTOL = .000888. In figure 4.1, we plot the error-to-bound ratio (4.7) versus time; the ratio is nearly constant. The error control yielded a constant stepsize. We plot the stability factor versus time in figure 4.2. The result in proposition 3.3 gives a bound on $S_1(n)$ that grows linearly in time; the computational result suggests that such a bound is not too large.

In this example, there is no discernable difference in the results obtained with the exact and approximate initial data for (4.4).

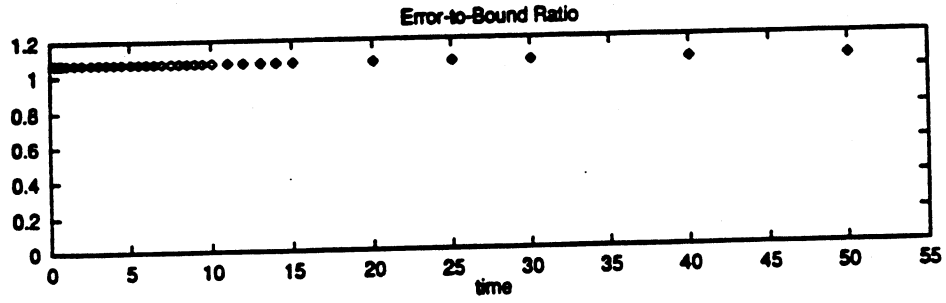

FIGURE 4.1.

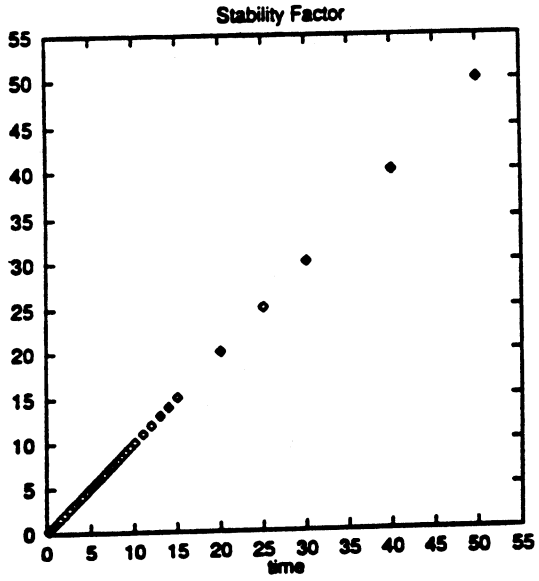

FIGURE 4.2.

Example 4.2. The problem is

$$\begin{cases} \dot{y}_1 = -.01y_1 - .99y_2 + .99y_3, & y_1(0) = 2, \\ \dot{y}_2 = -y_2 - 99y_3, & y_2(0) = 2, \\ \dot{y}_3 = -100y_3, & y_3(0) = 1, \end{cases}$$

with solution

$$\begin{aligned} y_1(t) &= e^{-t} + e^{-t/100}, \\ y_2(t) &= e^{-t} + e^{-100t}, \\ y_3(t) &= e^{-100t}. \end{aligned}$$

This problem was chosen as an example of a stiff computation. There are three time scales in the solution's behavior and the problem becomes stiff when the faster modes have decayed. We use GTOL = .001. The error control iteration halts after two iterations. The first iteration uses LTOL = .0009 and the second uses LTOL = .000304. In figure 4.3, we plot the error-to-bound ratio (4.7) versus time. The ratio tends to a constant value after an initial transient region; stiffness causes no trouble in this sense. However, as discussed in [8], the error in this problem changes direction radically several times in the transient region. These changes correlate to periods when the ratio changes value. In figure 4.4, we plot the step size sequence versus time. Finally, we plot the stability factor versus time in figure 4.5. For this dissipative problem, the stability factor should tend to 3 as time passes, and it clearly does this.

In this problem, the two choices of initial direction for (4.4) yield some differences in the corresponding stability factors. In figure (4.6), we plot the stability factors versus time for the exact direction $e_n/|e_n|$ and for $(Y_n - Y_{n-1})/|Y_n - Y_{n-1}|$. After the transient region, the values become close. In figure (4.7), we plot the error-to-bound ratios versus time for the two choices.

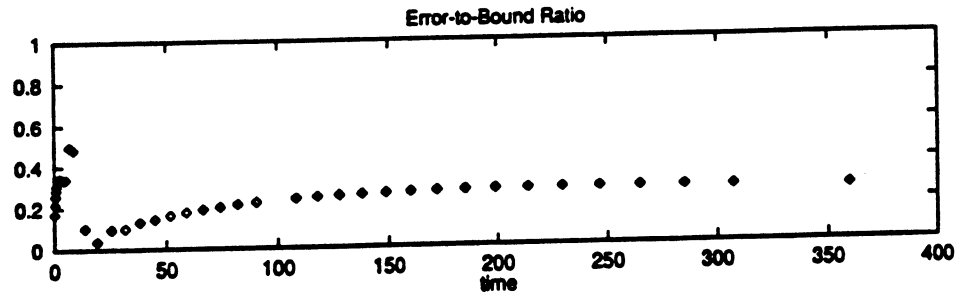

FIGURE 4.3.

Example 4.3. The problem is

$$\begin{cases} \dot{y}_1 = \dfrac{y_1}{2(1+t)} - 2ty_2, & y_1(0) = 1, \\ \dot{y}_2 = 2ty_1 + \dfrac{y_2}{2(1+t)}, & y_2(0) = 0, \end{cases}$$

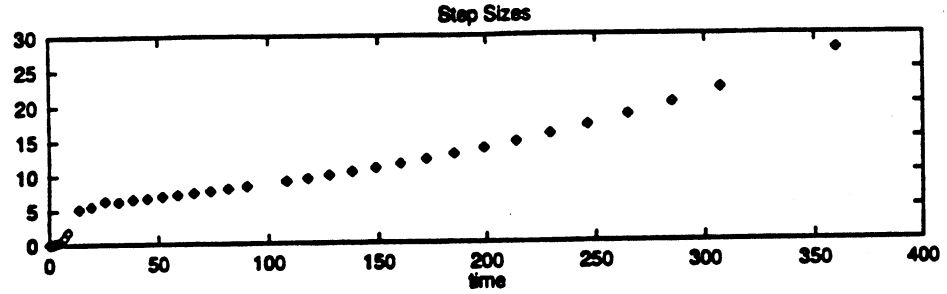

FIGURE 4.4.

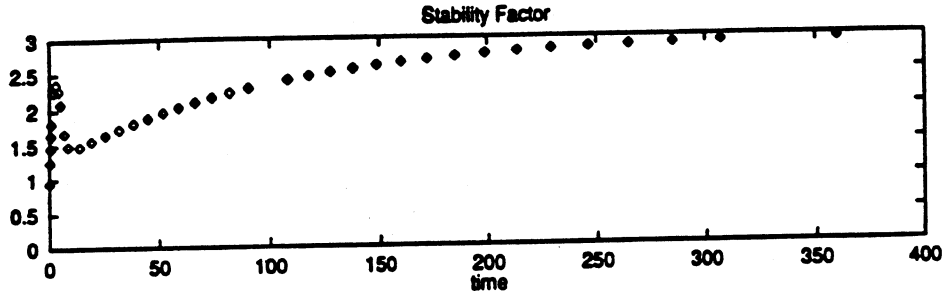

FIGURE 4.5.

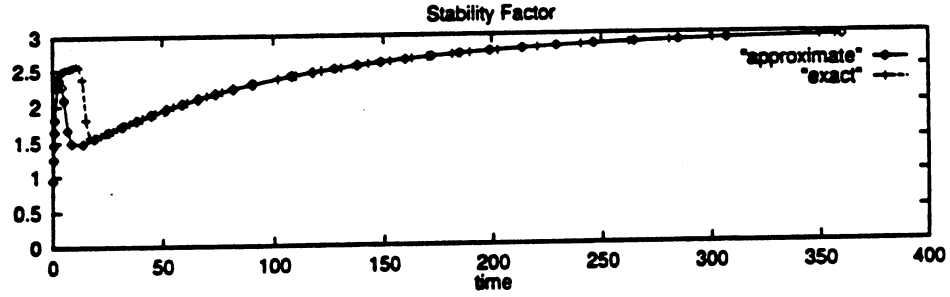

FIGURE 4.6.

with the solution

$$y_1(t) = \sqrt{1+t}\,\cos(t^2),$$
$$y_2(t) = \sqrt{1+t}\,\sin(t^2).$$

The solution is dynamically unstable, so we might expect that the error bounds will be sharp. We use GTOL = .02. The error control iteration halts after two iterations. The first iteration uses LTOL = .018 and the second uses LTOL = .000223. In figure 4.8, we plot the error-to-bound ratio (4.7) versus time and the ratio does remain fairly constant. If figure 4.9, we plot the step size sequence versus time. The solutions oscillate with increasing amplitude, which is reflected in the time steps. Finally, we plot the stability factor versus time in figure 4.10. The

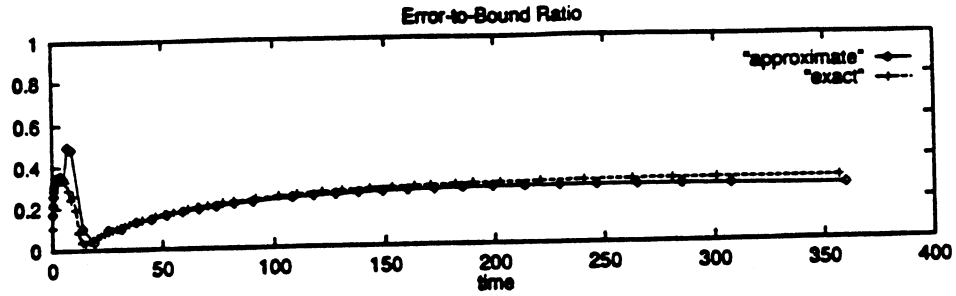

FIGURE 4.7.

stability factor reflects the instability of the solution.

In this example, there is no discernable difference in the results obtained with the exact and approximate initial data for (4.4).

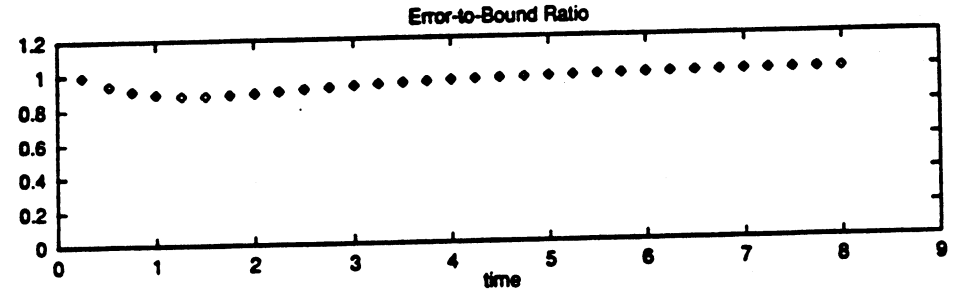

FIGURE 4.8.

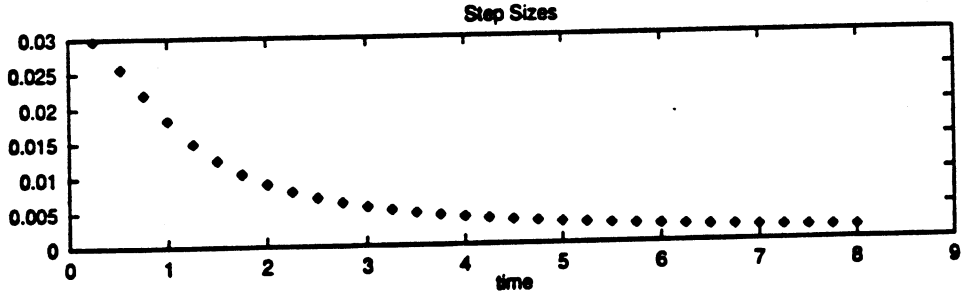

FIGURE 4.9.

Example 4.4. The last example is the two body problem,

$$\begin{cases} \dot{y}_1 = y_3, & y_1(0) = .4, \\ \dot{y}_2 = y_4, & y_2(0) = 0, \\ \dot{y}_3 = -y_1/(y_1^2 + y_2^2)^{3/2}, & y_3(0) = 0, \\ \dot{y}_4 = -y_2/(y_1^2 + y_2^2)^{3/2}, & y_4(0) = 2, \end{cases}$$

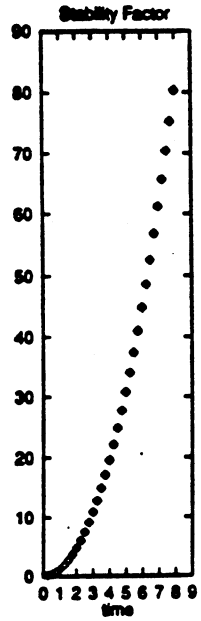

FIGURE 4.10.

with the periodic solution

$$
\begin{aligned}
y_1(t) &= \cos(\tau) - .6, \\
y_2(t) &= .8\sin(\tau), \\
y_3(t) &= -\sin(\tau)/(1 - .6\cos(\tau)), \\
y_4(t) &= .8\cos(\tau)/(1 - .6\cos(\tau)), \\
\tau &: \tau - .6\sin(\tau) = t.
\end{aligned}
$$

This is a well known test problem that is difficult both in terms of performing error control and choosing stability properties of the numerical method. The accumulated error grows rapidly with each successive period and it is not clear that tracking particular trajectories of this problem is meaningful, but it is an interesting test case for this theory. We use GTOL = .01 and compute just past three periods. The error control iteration takes three iterations in this example because the second iteration overpredicts the bound on the error. The first iteration uses LTOL = .009, the second uses LTOL = .000000139, and the third uses LTOL = .000000669. In this problem, very small local tolerances are used to counteract the tremendous rate of accumulation of error. In figure 4.11, we plot the error-to-bound ratio (4.7) versus time. While the ratio remains below one, it is disappointing that it decreases as time passes. The bound is clearly overpredicting the size of the error. On the other hand, the error accumulates at a tremendous rate. We plot the stability factor versus time in figure 4.12. Note the vertical scale. If figure 4.13, we plot the step size sequence versus time. The oscillating behavior of the solutions is reflected in the range of step sizes. Finally, w

There is little difference in the bounds given for the two choices of initial data for (4.4). In figure 4.14, we plot the error-to-bound ratios versus time, where some

difference is notable. But, this is not the reason that the error is overpredicted. It is perhaps not surprising that the stability constants for the two choices of data are so close considering the small local tolerances used in the computation. We show results also for the first iteration with LTOL = .009 for the sake of comparison in figure 4.15. There is more difference between the computations than in the final iteration, but it is still not significant.

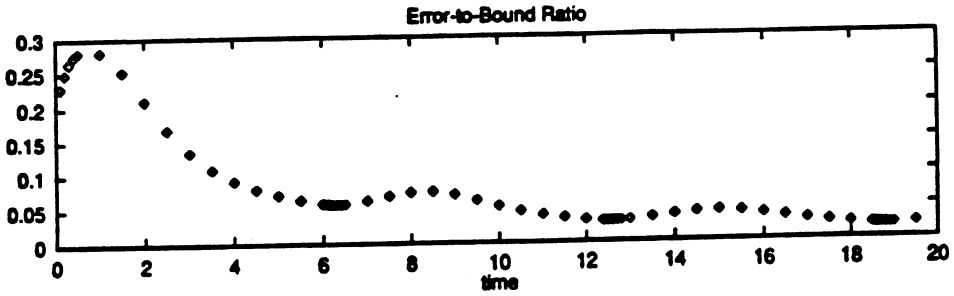

FIGURE 4.11.

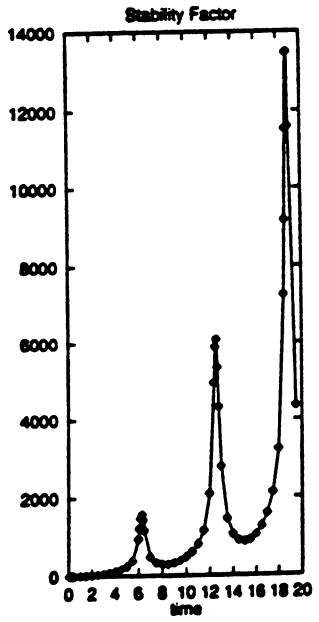

FIGURE 4.12.

Remark 4.5. In an effort to understand the results in example 4.4, we discuss the stability properties of the cG method in more detail. We recall the analysis of the discontinuous Galerkin (dG) method carried out in [8]. The dG method yields stiffly A-stable schemes that are well suited for stiff problems. We make a simple numerical

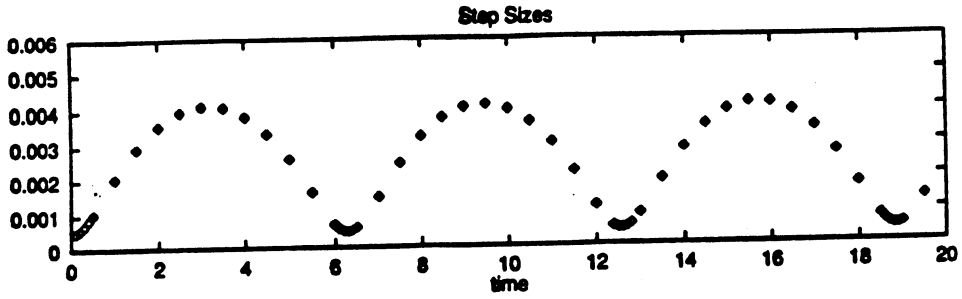

FIGURE 4.13.

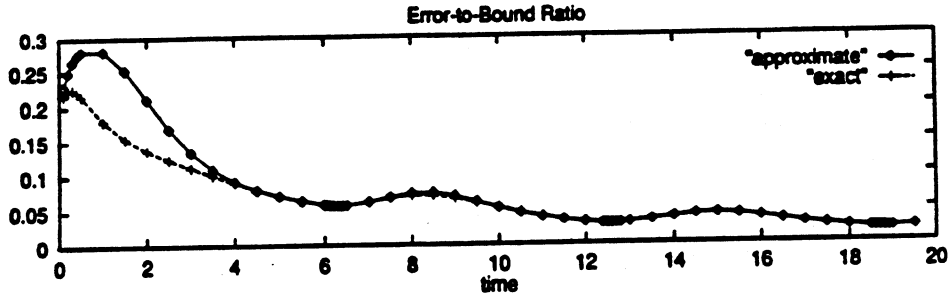

FIGURE 4.14.

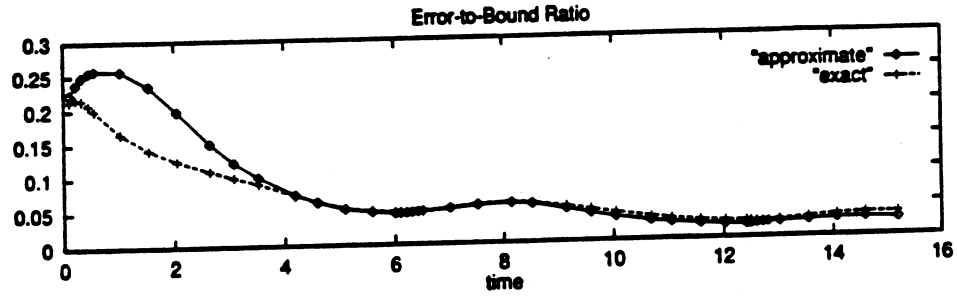

FIGURE 4.15.

comparison of the cG $q = 1$ scheme with the dG $q = 0$ scheme (which is a variation of the backward Euler scheme). The a posteriori error bounds for both methods are closely related and we expect the error control behaves similarly for both methods. In particular, the stability factors for the two schemes is exactly the same. Thus, the theory predicts the same accumulation of error for both schemes applied to a general problem. However, the schemes have different stability properties and we surmise that the error might accumulate more slowly for a particular scheme depending on the stability behavior of the solution. The following computations were made using the same local tolerance LTOL for each scheme and we are interested in the way in which the errors made at each step accumulate.

In figures 4.16 and 4.17, we plot the first component of the approximation and

the solution versus time for the cG and dG schemes for example 4.1 respectively. We note that the dG scheme dissipates the amplitude of the periodic solution and the cG method does not do this. One can show that the dG method must have this behavior. The error-to-bound ratios of both schemes remain almost constant over time, with a little more variation in the dG value.

In figures 4.18 and 4.19, we plot the error versus time for both schemes applied to example 4.2. The oscillations present in the error of the cG scheme are indicative of stiff problems and the amplitude of the oscillations increase with increasing stiffness. The error-to-bound ratios of the schemes behaves roughly the same over time, though there is larger amplitude in the variation of the ratio for the cG method.

In figures 4.20 and 4.21, we plot the first component of the approximation and the solution versus time for the cG and dG schemes applied to example 4.3 respectively. As in example 4.1, the dG scheme introduces dissipation. The error-to-bound ratios of the two schemes again behave similarly. We conjecture that the instability of the solution means that the error of both discretisations increases at the maximum rate.

Finally, in figures 4.22 and 4.23, we plot the first component of the approximation and the solution versus time for the cG and dG schemes applied to example 4.4 respectively. While there is not much decrease in amplitude in the dG approximation, the approximation does "shorten" each successive period. We conjecture that this is due to the dissipative properties of the scheme. In figure 4.24, we plot the error-to-bound ratio for the dG scheme. In contrast to the behavior of the cG scheme, this ratio increases as time passes and it is clear the the error of the dG method is closer to the predicted values. We conjecture that the stability properties of the cG method inhibit the error from growing at the maximum rate.

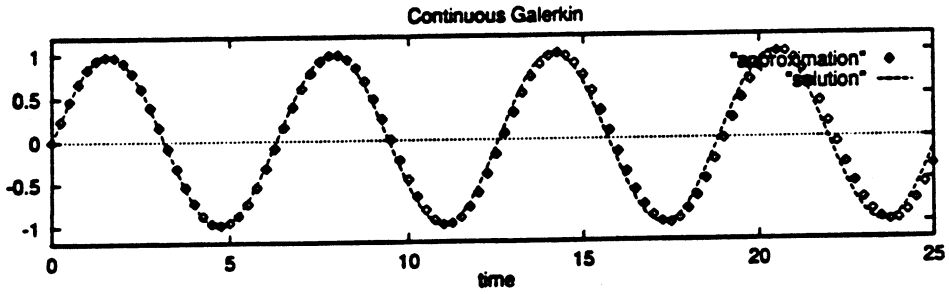

FIGURE 4.16.

§5. Proofs of A Priori Results

Proof of theorem 3.1. Consider

$$\begin{cases} \dot{y}(t) + f(t) = 0, & 0 < t \leq T, \\ y(0) = y_0, \end{cases}$$

and let $X \in C^{(q)}$ denote the cG approximation to y. We choose V with $V|_{I_m} = 1$ in (2.2) and find that $X_m = y_m$, for $1 \leq m \leq n$. When $q = 1$, we conclude that

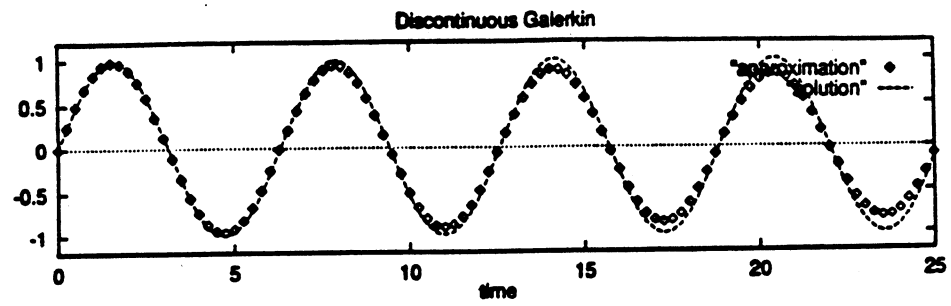

FIGURE 4.17.

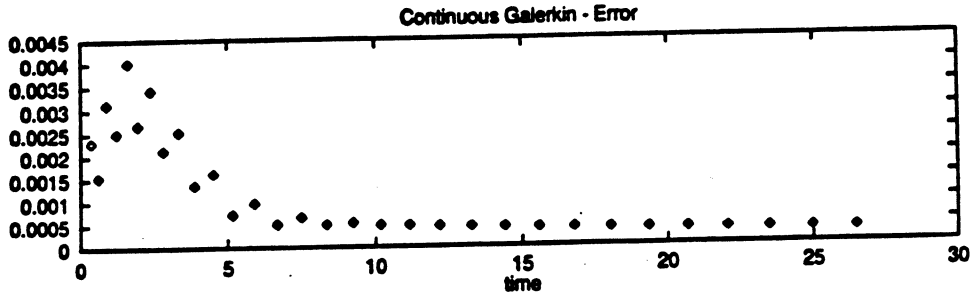

FIGURE 4.18.

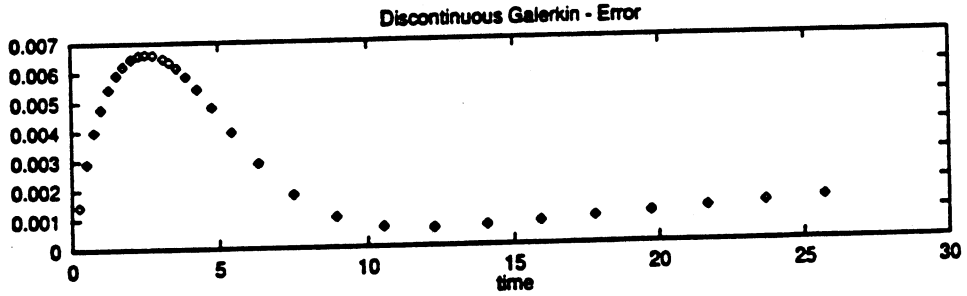

FIGURE 4.19.

there is a $C > 0$ such that

$$|y - X|_{I_m} \leq C k_m^2 |\tilde{y}|_{I_m},$$

for $1 \leq m \leq n$. When $q = 2$, we take V with $V|_{I_m} = (t - t_{m-1})/k_m$ in (2.2) and integrate by parts to obtain

$$X_{m-1/2} = -\frac{1}{4} y_m - \frac{1}{4} y_{m-1} + \frac{3}{2k_m} \int_{t_{m-1}}^{t_m} y(t)dt. \tag{5.1}$$

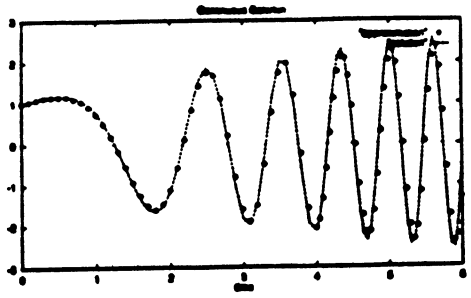

FIGURE 4.20.

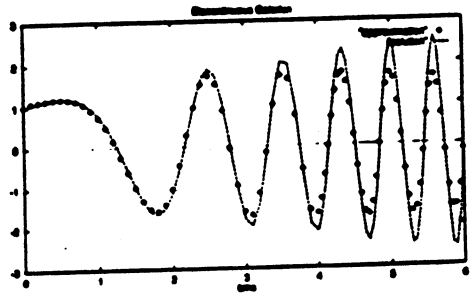

FIGURE 4.21.

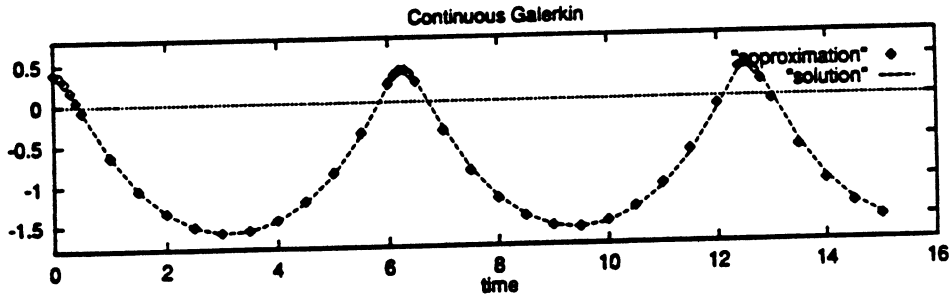

FIGURE 4.22.

We estimate

$$\begin{aligned} |X - y|_{I_m} &\leq |X - \mathcal{I}_q y|_{I_m} + |\mathcal{I}_q y - y|_{I_m} \\ &\leq |X_{m-1/2} - y_{m-1/2}| + |\mathcal{I}_q y - y|_{I_m} \\ &\leq C k_m^3 |y^{(3)}|_{I_m}, \end{aligned}$$

where we use (2.8) and expand in (5.1) around $t_{m-1/2}$ using Taylor's theorem.

Now, we take $f(t) \equiv f(y(t), t)$ and conclude that

$$|y - X|_{I_m} \leq C k_m^{q+1} |y^{(q+1)}|_{I_m}, \tag{5.2}$$

for $1 \leq m \leq n$.

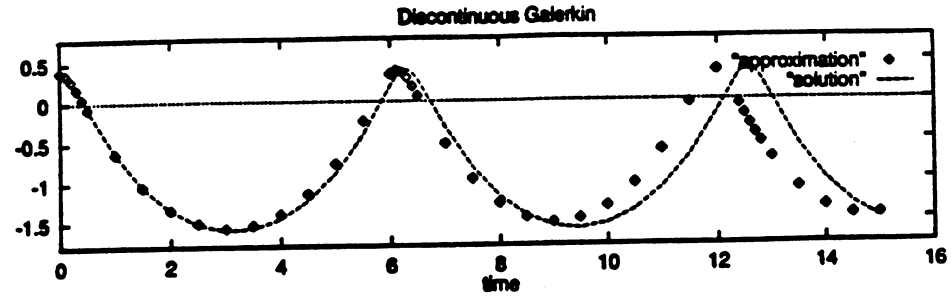

FIGURE 4.23.

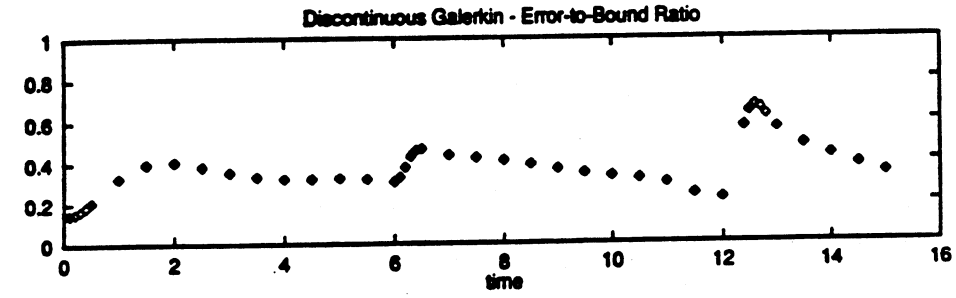

FIGURE 4.24.

Because of (1.1),

$$\int_{I_m} (\dot{y} - \dot{X}, V)dt = 0,$$

for all $V \in \mathcal{D}^{(q-1)}$. Therefore,

$$\int_{I_m} (\dot{X} - \dot{\tilde{Y}}, V)dt = \int_{I_m} (\dot{y} - \dot{\tilde{Y}}, V)dt,$$

for all $V \in \mathcal{D}^{(q-1)}$. We choose $V = \dot{X} - \dot{\tilde{Y}}$ and estimate using Young's inequality and (2.9) to get

$$|\dot{y} - \dot{X}|_{I_m} \leq Ck_m^q |y^{(q+1)}|_{I_m}, \tag{5.3}$$

for $1 \leq m \leq n$.

We split the error $e := \mu - \phi$ with $\mu := y - X$ and $\phi := Y - X \in C^{(q)}$. Starting with (2.2) and using (1.1), we write

$$\int_{I_m} (\dot{\phi}, V)dt + \int_{I_m} (f(Y(t), t) - f(y(t), t), V(t))dt = 0, \tag{5.4}$$

for all $V \in \mathcal{D}^{(q-1)}$. We choose $V = \mathcal{P}_{\mathcal{D}}\phi$ and use the fact that $\dot{\phi} \in \mathcal{D}^{(q-1)}$ to conclude that

$$\frac{1}{2}|\phi_m|^2 - \frac{1}{2}|\phi_{m-1}|^2 + \int_{I_m} (f(Y(t), t) - f(y(t), t), \mathcal{P}_{\mathcal{D}}\phi(t))dt = 0.$$

Therefore,

$$\frac{1}{2}|\phi_m|^2 \leq \frac{1}{2}|\phi_{m-1}|^2 + L\int_{I_m} |Y-y||\mathcal{P}_{\mathcal{D}}\phi|dt.$$

Now, we estimate and use the stability of $\mathcal{P}_{\mathcal{D}}$ to get

$$|\phi_m|^2 \leq |\phi_{m-1}|^2 + L\int_{I_m} |\mu|^2 dt + 3L\int_{I_m} |\phi|^2 dt. \tag{5.5}$$

When $q=1$, $|\phi|^2_{I_m} \leq 2|\phi_{m-1}|^2 + 2|\phi_m|^2$, so

$$|\phi_m|^2 \leq \frac{1+6Lk_m}{1-6Lk_m}|\phi_{m-1}|^2 + \frac{L}{1-6Lk_m}\int_{I_m} |\mu|^2 dt.$$

We continue back to $m=0$, using $\phi_0 = 0$, and get

$$|\phi_n| \leq CLt_n e^{CLt_n}|\mu|^2_{[0,t_n]}, \tag{5.6}$$

for $1 \leq n \leq N$. The bound on $\phi(t)$, $t \in I_n$, is immediate when $q=1$.

When $q=2$, we choose $V = \mathcal{P}_{\mathcal{D}}(t-t_{m-1})\phi(t)$ in (5.4) to get

$$\int_{I_m} (\dot{\phi}(t), (t-t_{m-1})\phi(t))dt + \int_{I_m} \left(f(Y(t),t) - f(y(t),t), \mathcal{P}_{\mathcal{D}}\left((t-t_{m-1})\phi(t)\right)\right)dt = 0.$$

We integrate by parts, use the Lipschitz assumption on f, and the stability of $\mathcal{P}_{\mathcal{D}}$ to find that

$$\int_{I_m} |\phi|^2 dt \leq k_m|\phi_m|^2 + 2L\int_{I_m} |t-t_{m-1}||e(t)||\phi(t)|dt.$$

Now, we assume that $k_m L$ is sufficiently small and estimate

$$\int_{I_m} |\phi|^2 dt \leq 2k_m|\phi_m|^2 + 2Lk_m\int_{I_m} |\mu|^2 dt. \tag{5.7}$$

We combine this with (5.5), taking $k_m L$ even smaller, to get

$$|\phi_m|^2 \leq \frac{1}{1-6Lk_m}|\phi_{m-1}|^2 + \frac{L+6L^2k_m}{1-6Lk_m}\int_{I_m} |\mu|^2 dt,$$

and then undo the recursion,

$$|\phi_m|^2 \leq CLt_n e^{CLt_n}|\mu|^2_{[0,t_n]}.$$

Now, we use (2.9) together with (5.7) to get

$$k_m|\phi|^2_{I_m} \leq Ck_m Lt_n e^{CLt_n}|\mu|^2_{[0,t_n]} + CLk_m^2|\mu|^2_I,$$

or noting that $Lk_m \le C$, for $1 \le m \le n$,

$$|\phi|_{I_m}^2 \le C\big(1 + Lt_n e^{CLt_n}\big)|\mu|_{[0,t_n]}^2. \tag{5.8}$$

To complete the estimate on the error, we use (5.2) and (5.8) to get

$$\begin{aligned}|e|_{[0,t_n]}^2 &\le |\phi|_{[0,t_n]}^2 + |\mu|_{[0,t_n]}^2 \\ &\le C\big(1 + Lt_n e^{CLt_n}\big) \max_{m \le n} k_m^{2(q+1)} |y^{(q+1)}|_{I_m}^2.\end{aligned}$$

Now, for the bound on $|\dot{e}|$, we choose $V = \dot{\phi}$ in (5.4) to get

$$\begin{aligned}\int_{I_m} |\dot{\phi}|^2 dt &\le \int_{I_m} |(f(Y(t),t) - f(y(t),t), \dot{\phi}(t))| dt \\ &\le \frac{L^2}{2} \int_{I_m} |e|^2 dt + \frac{1}{2} \int_{I_m} |\dot{\phi}|^2 dt,\end{aligned}$$

so

$$\int_{I_m} |\dot{\phi}|^2 dt \le L^2 \int_{I_m} |e|^2 dt. \tag{5.9}$$

We use (2.9) on (5.9), (5.3), and the result for $|e|_{[0,t_n]}$ to compute

$$|\dot{e}|_{I_m} \le C\left(1 + kL\big(1 + Lt_n e^{CLt_n}\big)\right) \max_{m \le n} k_m^q |y^{(q+1)}|_{I_m}. \quad \square$$

Proof of theorem 3.2. We start with a regularity result for Y.

Lemma 5.1. *Under the assumptions of theorem 3.2, for k sufficiently small and $1 \le m \le n$,*

$$\begin{aligned}|Y^{(p)}|_{I_m} &\le C|y^{(p)}|_{I_m} + C\big(1 + Lt_m e^{CLt_m}\big) \max_{j \le m} \left(\frac{k_j}{k_m}\right) k_j^{q+1-p} |y^{(q+1)}|_{I_j} \\ &\le C(L, T, \rho),\end{aligned} \tag{5.10}$$

for $0 \le p \le q$.

Proof.

$$|Y^{(p)}|_{I_m} \le C|y^{(p)}|_{I_m} + \left|\frac{d^p}{dt^p}(Y - \mathcal{I}_q y)\right|_{I_m} + \left|\frac{d^p}{dt^p}(\mathcal{I}_q y - y)\right|_{I_m}.$$

By (2.9),

$$\begin{aligned}\left|\frac{d^p}{dt^p}(Y - \mathcal{I}_q y)\right|_{I_m} &\le C k_m^{-p} |Y - \mathcal{I}_q y|_{I_m} \\ &\le C k_m^{-p} \big(|Y - y|_{I_m} + |y - \mathcal{I}_q y|_{I_m}\big).\end{aligned}$$

Now, we combine (3.1) and (2.8) to prove the first equation in (5.10). Under the assumptions of theorem 3.2, the conclusion follows immediately. $\square$

For functions $V(t)$ and $W(t)$, we let

$$M(t,V,W) := \int_0^1 f_y(\rho V(t) + (1-\rho)W(t),t)d\rho,$$

so

$$f(y,t) - f(Y,t) = M(t,y,Y)e.$$

We define on each interval,

$$\tau(t) := \dot{Y}(t) + f(Y(t),t),$$

so e solves the equation

$$\dot{e} + M(t,y,Y)e = -\tau,$$

on each interval. We use $M(t)$ to denote $M(t,y,Y)$ in the following.

We let $\Phi_1 \in C_{q+1} \times C_{q+1}$ denote the fundamental matrix solution on I_1:

$$\begin{cases} \dot{\Phi}_1(t) + M(t)\Phi_1(t) = 0, & 0 < t, \\ \Phi_1(0) = I \text{ (the } d \times d \text{ identity)}. \end{cases}$$

Variation of constants implies

$$e(t_1) = \Phi_1(t_1)\Big(e(0) - \int_0^{t_1} \Phi_1^{-1}(t)\tau(t)dt\Big).$$

Equation (2.2) implies that

$$\int_0^{t_1} (\tau, V)dt = 0,$$

for all $V \in \mathcal{P}^{q-1}(I_1)$. Hence,

$$e(t_1) = \Phi_1(t_1)\Big(e(0) - \int_0^{t_1} (\Phi_1^{-1}(t) - V(t))\tau(t)dt\Big),$$

for all $V \in \mathcal{P}^{q-1}(I_1) \times \mathcal{P}^{q-1}(I_1)$, and therefore,

$$|e(t_1)| \leq |\Phi_1(t_1)|\Big(|e(0)| + Ck_1^{1/2}\Big(\int_0^{t_1} |\tau|^2 dt\Big)^{1/2} |\Phi_1^{-1} - V|_{I_1}\Big).$$

for all $V \in \mathcal{P}^{q-1}(I_1) \times \mathcal{P}^{q-1}(I_1)$, where we use the obvious matrix norm. Using the equivalence of norms on a finite dimensional vector space and (2.8) on each component of Φ_1^{-1}, we know that

$$|\Phi_1^{-1} - \mathcal{I}_{q-1}\Phi_1^{-1}|_{I_1} \leq Ck_1^q \Big|\frac{d^q}{dt^q}\Phi_1^{-1}\Big|_{I_1},$$

where we extend the definition of $\mathcal{I}$ in the obvious way. Next, note that for $V \in \mathcal{P}^{q-1}(I_1)$,

$$\int_{I_1} |\tau - V|^2 dt \geq \int_{I_1} |\tau|^2 dt.$$

Choosing $V = \mathcal{I}_{q-1}\tau$, we get

$$\left(\int_{I_1} |\tau|^2 dt\right)^{1/2} \leq Ck_1^{q+1/2}|\tau^{(q)}|_{I_1}.$$

Thus,

$$|e(t_1)| \leq |\Phi(t_1)|\left(|e(0)| + Ck_1^{2q+1}\left|\frac{d^q}{dt^q}\Phi_1^{-1}\right|_{I_1}|\tau^{(q)}|_{I_1}\right). \tag{5.11}$$

Generalized to the m^{th} interval, $1 \leq m \leq n$,

$$|e(t_m)| \leq |\Phi(t_m)|\left(|e(t_{m-1})| + Ck_m^{2q+1}\left|\frac{d^q}{dt^q}\Phi_m^{-1}\right|_{I_m}|\tau^{(q)}|_{I_m}\right). \tag{5.12}$$

Next, we bound the various quantities in this inequality. Since $\deg(Y|_{I_1}) = q$, $\frac{d^q}{dt^q}\dot{Y} = 0$ and

$$\frac{d^q\tau}{dt^q} = -\frac{d^q}{dt^q}f(Y(t),t).$$

By assumption, the partial derivatives of f of order q and less are smooth and bounded uniformly in $\mathcal{N}_\delta$ while by theorem 3.1, Y is close to y for k small, hence (5.10) implies that

$$|\tau^{(q)}|_{I_1} \leq \mathcal{E}(L,\rho,k,q,t_1,y), \tag{5.13}$$

with

$$\mathcal{E}(L,\rho,k,q,t_1,y) := C\max\{1,L\}\max\{1,\rho^q\}\max\{1,k\}\left(1+Lt_1e^{CLt_1}\right)\max_{\substack{r+s\leq q+1\\ r,s\geq 0}}|y^{(r)}|^s_{[0,t_1]}.$$

Similarly, M is bounded as

$$\left|\frac{d^q}{dt^q}M\right|_{I_1} \leq \begin{cases} cL, & q=0,\\ \mathcal{E}(L,\rho,k,q,t_1,y), & q=1. \end{cases} \tag{5.14}$$

Note, both of these bounds carry over to I_m with $t_1 \to t_m$ in $\mathcal{E}$.

Since,

$$0 = \frac{d}{dt}\left(\Phi_1\Phi_1^{-1}\right) = (-M\Phi_1)\Phi_1^{-1} + \Phi_1\frac{d}{dt}\Phi_1^{-1},$$

on I_1, Φ_1^{-1} solves

$$\begin{cases} \frac{d}{dt}\Phi_1^{-1} = \Phi_1^{-1}M, & t>0,\\ \Phi_1^{(-1)}(0) = I. \end{cases}$$

Thus,

$$\Phi_1^{-1}(t) = I + \int_0^t \Phi_1^{-1}(s)M(s)ds, \tag{5.15}$$

for $0 \le t \le t_1$. We estimate, using the bound on M,

$$|\Phi_1^{-1}(t)| \le 1 + CL\int_0^t |\Phi_1^{-1}(s)|ds,$$

for $0 \le t \le t_1$. Gronwall's inequality implies that

$$|\Phi_1^{-1}(t_1)| \le e^{CLk_1}.$$

By (5.15),

$$\left|\frac{d\Phi_1^{-1}}{dt}(t_1)\right| \le CLe^{CLk_1}.$$

Similarly,

$$\left|\frac{d^2\Phi_1^{-1}}{dt^2}(t_1)\right| \le \left(CL^2 + \mathcal{E}(L,\rho,k,q,t_1,y)\right)e^{CLk_1}.$$

Note that these arguments carry over to I_m, with $\Phi_m^{-1}(t_{m-1}) = I$ and computing the integrals from t_{m-1} to $t : t_{m-1} \le t \le t_m$, hence for $1 \le m \le n$,

$$|\Phi_m^{-1}(t_m)| \le e^{CLk_m}, \tag{5.16}$$

$$\left|\frac{d\Phi_m^{-1}}{dt}(t_m)\right| \le CLe^{CLk_m}, \tag{5.17}$$

and

$$\left|\frac{d^2\Phi_m^{-1}}{dt^2}(t_m)\right| \le \left(CL^2 + \mathcal{E}(L,\rho,k,q,t_m,y)\right)e^{CLk_m}. \tag{5.18}$$

Finally, we estimate $|\Phi_m(t_m)|$ just as we did $|\Phi^{-1}(t_m)|$,

$$|\Phi_m(t_m)| \le e^{CLk_m}, \tag{5.19}$$

for $1 \le m \le n$. We return to (5.11) and use (5.13), (5.14), and (5.16)–(5.19) to find that

$$|e(t_1)| \le Ck_1e^{CLk_1}k_1^{2q}\,\overline{\mathcal{E}}(L,\rho,k,q,t_1,y),$$

with

$$\overline{\mathcal{E}}(L,\rho,k,q,t_m,y) := \mathcal{E}(L,\rho,k,q,t_m,y)(L^2 + \mathcal{E}(L,\rho,k,q,t_m,y)).$$

By (5.12),

$$|e(t_m)| \le e^{C} \quad \left(|e(t_{m-1})| + Ck_m^{2q+1}\,\overline{\mathcal{E}}(L,\rho,k,q,t_m,y)\right).$$

By induction, for $1 \le n \le N$,

$$|e(t_n)| \le Ct_ne^{CLt_n}\max_{m\le n} k_m^{2q}\,\overline{\mathcal{E}}(L,\rho,k,q,t_m,y).$$

The result follows by making straightforward estimates on $\overline{\mathcal{E}}$. □

§6. Proof of A Posteriori Results

Proof of Theorem 3.4. On I_m,

$$\dot{Y} + f(Y,t) = (1 - \mathcal{P}_{\mathcal{D}})f(Y,t), \tag{6.1}$$

taking one-sided limits at the ends of the intervals. By subtracting this from the equation in (1.1), we get

$$\dot{e} + \int_0^1 f_y(\rho y + (1-\rho)Y, t)(y - Y)d\rho = (\mathcal{P}_{\mathcal{D}} - 1)f(Y,t).$$

This motivates the definition

$$\tilde{D}(W,V) := \int_0^{t_n} (\dot{W}, V)dt + \int_0^{t_n} \left(\int_0^1 f_y(\rho y(t) + (1-\rho)Y(t), t)d\rho W(t), V(t) \right) dt,$$

for functions W, V, since

$$\tilde{D}(e,V) = 0, \tag{6.2}$$

for all $V \in \mathcal{D}^{(q-1)}$. Associated to $\tilde{D}$ is the linear form

$$D(W,V) := \int_0^{t_n} \Big((\dot{W}, V) + (A(t)W(t), V(t)) \Big) dt.$$

If V is continuous,

$$D(W,V) = \int_0^{t_n} \big((W(t), -\dot{V}(t)) + (W(t), A^*(t)V(t)) \big) dt + \big((W(t_n), V(t_n)) - (W(0), V(0)) \big).$$

Since $e(0) = 0$,

$$|e(t_n)| = D(e,z) = \tilde{D}(e,z) - (\tilde{D} - D)(e,z), \tag{6.3}$$

where z solves (3.4). We subtract (6.2) and obtain

$$\tilde{D}(e,z) = \tilde{D}(e, z - \mathcal{P}_{\mathcal{D}}z) = -\int_0^{t_n} (f(Y(t),t) - \mathcal{P}_{\mathcal{D}}f(Y(t),t), z(t) - \mathcal{P}_{\mathcal{D}}z(t))dt.$$

We take norms and use (2.7) to get

$$\begin{aligned} |\tilde{D}(e,z)| &\le \sum_{m=1}^{n} k_m \cdot k_m^q \left| \frac{d^q}{dt^q} f(Y(t),t) \right|_{I_m} \cdot \int_{t_{m-1}}^{t_m} |\dot{z}| dt \\ &\le S_1(n) \max_{m \le n} k_m^{q+1} \left| \frac{d^q}{dt^q} f(Y(t),t) \right|_{I_m} . \end{aligned} \tag{6.4}$$

We can write this as follows:

$$|\tilde{D}(e,z)| \leq S_q(n) \max_{m \leq n} k_m^{2q} \left| \frac{d^q}{dt^q} f(Y(t),t) \right|_{I_m}, \tag{6.5}$$

by taking the second derivative of z in the case that $q = 2$. To estimate the second term on the right in (6.3), we compute

$$(D - \tilde{D})(e,z) = \int_0^{t_n} \int_0^1 \Big(\big(f_y(y(t),t) - f_y(\rho y(t) + (1-\rho)Y(t),t)\big)e(t), z(t) \Big) d\rho dt$$

and estimate using (3.4.3),

$$|(D - \tilde{D})(e,z)| \leq C S_1(n) |e|^2_{[0,t_n]}. \tag{6.6}$$

We use (6.3), (6.4), and (6.6) and conclude that

$$|e(t_n)| \leq S_1(n) \max_{m \leq n} k_m^{q+1} \left| \frac{d^q}{dt^q} f(Y(t),t) \right|_{I_m} + C S_1(n) |e|^2_{[0,t_n]}, \tag{6.7}$$

and with (6.5),

$$|e(t_n)| \leq S_q(n) \max_{m \leq n} k_m^{2q} \left| \frac{d^q}{dt^q} f(Y(t),t) \right|_{I_m} + S_1(n) |e|^2_{[0,t_n]}. \tag{6.8}$$

Next, we make a local estimate. Subtracting (6.1) from the equation in (1.1) and taking the inner product with e, we get

$$(\dot{e}, e) + (f(y,t) - f(Y,t), e) = ((\mathcal{P}_D - 1) f(Y,t), e).$$

We integrate from t_{m-1} to t, take norms, and use the fact that e is continuous to conclude that

$$|e|^2_{I_m} \leq |e(t_{m-1})|^2 + \int_{I_m} \big| \big((\mathcal{P}_D - 1) f(Y(t),t), e(t)\big) \big| dt + \int_{I_m} L(t) |e|^2 dt,$$

and so,

$$|e|^2_{I_m} \leq |e(t_{m-1})|^2 + C \left(\int_{I_m} |(\mathcal{P}_D - 1) f(Y(t),t)| dt \right)^2 + \epsilon |e|^2_{I_m} + \int_{I_m} L(t) dt \cdot |e|^2_{I_m},$$

for $\epsilon > 0$ small. We assume that k_m is sufficiently small and use (2.7) to conclude that

$$|e|^2_{I_m} \leq C \left(|e_{m-1}|^2 + k_m^{2q+2} \left| \frac{d^q}{dt^q} f(Y(t),t) \right|^2_{I_m} \right). \tag{6.9}$$

We place (6.7) into (6.9) and use the fact that $S_q(m)$ is monotonically increasing in m to get

$$|e|_{I_n}^2 \le CS_1(n)^2 \max_{m\le n} k_m^{2q+2} \left|\frac{d^q}{dt^q} f(Y(t),t)\right|_{I_m}^2 + CS_1(n)^2 |e|_{[0,t_n]}^4 + C \max_{m\le n} k_m^{2q+2} \left|\frac{d^q}{dt^q} f(Y(t),t)\right|_{I_m}^2.$$

Since the right-hand side is monotonically increasing in n, we have

$$|e|_{[0,t_n]}^2 \le C(S_1(n)+1)^2 \max_{m\le n} k_m^{2q+2} \left|\frac{d^q}{dt^q} f(Y(t),t)\right|_{I_m}^2 + \left(CS_1(n)^2 |e|_{[0,t_n]}^2\right) |e|_{[0,t_n]}^2. \tag{6.10}$$

In the case of the superconvergence result (6.8), we get

$$|e|_{[0,t_n]}^2 \le C(S_q(n)+1)^2 \max_{m\le n} k_m^{4q} \left|\frac{d^q}{dt^q} f(Y(t),t)\right|_{I_m}^2 + \left(CS_1(n)^2 |e|_{[0,t_n]}^2\right) |e|_{[0,t_n]}^2.$$

Next, we prove the bounds on $S_1(n)$ given in proposition 3.3. For simplicity's sake, consider the forward problem which arises from the change of variables $\chi(t) := z(t_n - t)$,

$$\begin{cases} \dot\chi + A^*(t_n - t)\chi = 0, & 0 < t \le t_n, \\ \chi(0) = e_n/|e_n|. \end{cases} \tag{6.11}$$

We take the inner product of the equation in (6.11) with χ and get

$$(\dot\chi, \chi) + (A^*(t_n - t)\chi, \chi) = 0.$$

In the general case, we use (3.3.1) or (3.4.2) and integrate to find that

$$|\chi(t)|^2 \le e^{2\int_0^{t_n} L(s)ds},$$

and so,

$$\int_0^t |\dot\chi(s)|ds \le \int_0^t L(s) e^{\int_0^s L(\tau)d\tau} ds = e^{\int_0^t L(s)ds} - 1.$$

Under (3.3.2), we find that $\frac{d}{dt}|\chi|^2 \le 0$, so $|\chi(t)|^2 \le 1$ and

$$\int_0^t |\dot\chi(s)|ds \le \int_0^t L(s)ds.$$

Returning to the proof of theorem 3.4, in both cases, the a priori result (3.1) implies that we can choose k small enough so that

$$CS_1(n)^2 |e|_{[0,t_n]}^2 \le \frac{1}{2},$$

and therefore, we reach (3.1) via (6.10).

Remark 6.1. This condition on k can be viewed as determining the length of time over which the a posteriori analysis is valid.

To obtain the nodal result, we start with (6.9), which implies that

$$|e|^2_{[0,t_n]} \leq \max_{m\leq n-1} |e_m|^2 + C\max_{m\leq n} k_m^{2q+2}\left|\frac{d^q}{dt^q}f(Y(t),t)\right|^2_{I_m}. \tag{6.12}$$

We put this into (6.7) and use the monotonicity of the right-hand side to get

$$\max_{m\leq n}|e_m| \leq CS_1(n)\max_{m\leq n} k_m^{q+1}\left|\frac{d^q}{dt^q}f(Y(t),t)\right|^2_{I_m} + CS_1(n)\max_{m\leq n} k_m^{2q+2}\left|\frac{d^q}{dt^q}f(Y(t),t)\right|^2_{I_m} + CS_1(n)\max_{m\leq n}|e_m|\cdot\max_{m\leq n}|e_m|.$$

Now, we use (3.4.4) and (3.1) to conclude (3.11) for k sufficiently small. For the superconvergence result, we put (6.12) into (6.8) and obtain

$$\max_{m\leq n}|e_m| \leq CS_q(n)\max_{m\leq n} k_m^{2q}\left|\frac{d^q}{dt^q}f(Y(t),t)\right|^2_{I_m} + CS_1(n)\max_{m\leq n} k_m^{2q+2}\left|\frac{d^q}{dt^q}f(Y(t),t)\right|^2_{I_m} + CS_1(n)\max_{m\leq n}|e_m|\cdot\max_{m\leq n}|e_m|,$$

and finally, (3.12) for k sufficiently small. □

Proof of Proposition 3.5. We give the proof for $d = 1$. The generalization to $d > 1$ is straightforward. First,

$$\begin{aligned}\frac{d}{dt}f(Y(t),t) &= f_y(Y,t)\dot{Y} + f_t(Y,t)\\ &= f_y(y,t)\dot{y} + f_t(y,t) + f_y(Y,t)(\dot{Y}-\dot{y})\\ &\quad + (f_y(Y,t) - f_y(y,t))\dot{y} + f_t(Y,t) - f_t(y,t),\end{aligned}$$

so

$$\left|\frac{d}{dt}f(Y(t),t) - \tilde{y}\right|_{I_m} \leq CL\left(1 + Lt_n e^{CLt_n}\right)\left(1 + |\dot{y}|_{[0,t_n]}\right)|y^{(q+1)}|_{[0,t_n]}k^q.$$

Similarly,

$$\frac{d^2}{dt^2}f(Y(t),t) = f_y(Y,t)\tilde{\dot{Y}} + f_{yy}(Y,t)(\dot{Y})^2 + 2f_{yt}(Y,t)\dot{Y} + f_{tt}(Y,t).$$

We estimate as above and use the fact that

$$\tilde{\dot{Y}} = \frac{4}{k_m^2}\left(Y_{m-1} - 2Y_{m-1/2} + Y_m\right)$$

implies that

$$|\tilde{\dot{Y}} - \tilde{y}|_{I_m} \leq C\left(1 + \rho^2\left(1 + CLt_n e^{CLt_n}\right)^{1/2}\right)k|y^{(3)}|_{[0,t_n]},$$

to prove the result. □

REFERENCES

1. P. Ciarlet, *The Finite Element Method for Elliptic Problems*, North-Holland, New York, 1978.
2. K. Eriksson and C. Johnson, *Error estimates and automatic time step control for nonlinear parabolic problems I*, SIAM J. Numer. Anal. **24** (1987), 12–23.
3. ———, *Adaptive finite element methods for parabolic problems I: a linear model problem*, SIAM J. Numer. Anal. **28** (1991), 43–77.
4. ———, *Adaptive finite element methods for parabolic problems II: optimal error estimates in $L_\infty(L_2)$ and $L_\infty(L_\infty)$*, preprint # 1992-09, Chalmers University of Technology (1992).
5. ———, *Adaptive finite element methods for parabolic problems III: time steps variable in space*, in preparation.
6. ———, *Adaptive finite element methods for parabolic problems IV: nonlinear problems*, preprint #1992-44, Chalmers University of Technology (1992).
7. ———, *Adaptive finite element methods for parabolic problems V: long-time integration*, preprint # 1993-04, Chalmers University of Technology (1993).
8. D. Estep, *A posteriori error bounds and global error control for approximations of ordinary differential equations*, SIAM J. Numer. Anal. (to appear).
9. D. Estep and A. Stuart, *The dynamical behavior of the discontinuous Galerkin method and related difference schemes*, preprint.
10. D. French and S. Jensen, *Long time behaviour of arbitrary order continuous time Galerkin schemes for some one-dimensional phase transition problems*, preprint.
11. ———, *Global dynamics of finite element in time approximations to nonlinear evolution problems*, International Conference on Innovative Methods in Numerical Analysis, Bressanone, Italy, 1992.
12. D. French and J. Schaeffer, *Continuous finite element methods which preserve energy properties for nonlinear problems*, Appl. Math. Comp. **39** (1990), 271–295.
13. J. Hale, *Ordinary Differential Equations*, John Wiley and Sons, Inc., New York, 1980.
14. C. Johnson, *Error estimates and adaptive time-step control for a class of one-step methods for stiff ordinary differential equations*, SIAM J. Numer. Anal. **25** (1988), 908–926.
15. J. Schaeffer, *Personal communication* (1990).
16. A. Stroud, *Numerical Quadrature and Solution of Ordinary Differential Equations*, Applied Mathematical Sciences 10, Springer-Verlag, New York, 1974.

APPLIED MATHEMATICS, CALIFORNIA INSTITUTE OF TECHNOLOGY, PASADENA, CA 91125
Current address: School of Mathematics, Georgia Institute of Technology, Atlanta, GA 30332.
E-mail address: estep@ama.caltech.edu

DEPARTMENT OF MATHEMATICAL SCIENCES, UNIVERSITY OF CINCINNATI, CINCINNATI, OH 45221
E-mail address: french@ucunix.san.uc.edu