ESTIMATING THE ERROR OF NUMERICAL SOLUTIONS OF SYSTEMS OF NONLINEAR REACTION-DIFFUSION EQUATIONS

DONALD J. ESTEP *, MATS G. LARSON †, AND ROY D. WILLIAMS ‡

Abstract. We begin by developing a general theory for the a posteriori estimation of the error of numerical solutions of systems of nonlinear reaction-diffusion equations. We use the results to estimate the error of numerical solutions of nine well known models and determine time scales over which accurate numerical solutions can be computed for each problem. Next we interpret the theory analytically under general assumptions on the differential equation and also discuss some issues that arise when using the a posteriori estimate in practice. Finally, we apply the general theory to the class of problems admitting invariant rectangles in solution space.

Key words. a posteriori error estimates, accumulation of error, error estimates, finite element methods, invariant rectangles, reaction-diffusion equations, residual errors, stability

AMS subject classifications. 65M12, 65M15, 65M20, 65M60, 35B35, 35B50, 35B65, 35K37

Contents.

*School of Mathematics, Georgia Institute of Technology, Atlanta, GA 30332, USA. The research of D. Estep is partially supported by the National Science Foundation, DMS 9506519.

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

‡Center for Advanced Computing Research, California Institute of Technology, Pasadena, CA 91125, USA.

1. Introduction.

The ever increasing activity in the areas of mathematics and science concerned with reaction-diffusion equations marks both their important role in modelling physical phenomena in such diverse fields as biology, chemistry, metallurgy, and combustion, and the beauty and complexity found in their solutions. Numerical analysis of reaction-diffusion equations has become a central tool in their study because of the many barriers that exist for mathematical analysis. It is exactly these situations, when we know little about the true solution, that are particularly needful of accuracy in numerical results. Yet, these same analytic difficulties also give rise to nearly insurmountable barriers to accurate analytic estimation of the error of numerical solutions. In this paper, we investigate a different approach to this problem based on the *computational* estimation of the error of numerical solutions.

Many fundamental models in science take the form

$$\frac{\partial u}{\partial t} - \nabla \cdot \big(\epsilon(u,x,t)\nabla u\big) + \sum_j \beta_j(u,x,t)\frac{\partial u}{\partial x_j} = f(u,x,t) \tag{1.1}$$

for a vector unknown $u \in \mathbf{R}^d$, where ϵ is a diagonal matrix with smooth nonnegative entries, β_j are diagonal matrices with smooth entries that are dominated by the coefficients of ϵ, and $f = (f_i)$ is a smooth vector-valued function. Some well-known examples are:

Example 1: the bistable equation. Also known as the Chafee-Infante problem, the equation has the form (1.1) with $d = 1$, $\epsilon > 0$ constant, $\beta \equiv 0$, and $f(u) = u - u^3$. In one dimension, the bistable equation has been used to model the motion of domain walls in ferromagnetic materials.

Example 2: equations for two species. This is a model for the interaction of two species distributed continuously throughout the region Ω. It has the form (1.1) with $d = 2$, ϵ constant, $\beta \equiv 0$, $f_1 = u_1 M(u_1, u_2)$, and $f_2 = u_2 N(u_1, u_2)$. To model a predator and prey, we assume that $M_{u_2} < 0$ and $N_{u_1} > 0$. To model two competing species, we assume that $M_{u_2} < 0$ and $N_{u_1} < 0$. Finally to model symbiosis, we assume $M_{u_2} > 0$ and $N_{u_1} > 0$.

Example 3: Hodgkin-Huxley equations. These equations model the signal transmission across axons. The system takes the form of (1.1) with $d = 4$, ϵ constant, $\beta \equiv 0$, and

$$f_1 = \alpha_1 u_2^3 u_3(\beta_1 - u_1) + \alpha_2 u_4^4(\beta_2 - u_1) + \alpha_3(\beta_3 - u_1), \quad \beta_1 > \beta_2 > 0 > \beta_3$$
$$f_i = g_i(u_1)(h_i(u_1) - u_i),\ g_i > 0,\ 0 < h_i < 1, \quad 2 \le i \le 4.$$

u_2, u_3, and u_4 represent chemical concentrations and are nonnegative, while u_1 represents electric potential.

Example 4: Fitz-Hugh-Nagumo equations. These equations are a simplified model of the Hodgkin-Huxley equations. They have the form (1.1) with $d = 2$, ϵ constant, $\beta \equiv 0$, and

$$f_1 = -u_1(u_1 - \alpha_1)(u_1 - 1) - u_2, \quad 0 < \alpha_1 < 1/2,$$
$$f_2 = \alpha_2 u_1 - \alpha_3 u_2, \quad \alpha_2, \alpha_3 > 0.$$

Example 5: superconductivity of liquids. These equations are used in the description of superconductivity in liquids. They have the form (1.1) with $d = 2$, ϵ constant, $\beta \equiv 0$, and $f(u) = (1 - |u|^2)u$.

Example 6: Field-Noyes equations. These are used to model the famous Belousov-Zhabotinsky reaction in chemical kinetics. They have the form (1.1) with $d = 3$, ϵ constant, $\beta \equiv 0$, and

$$\begin{aligned} f_1 &= \alpha_1(u_2 - u_1u_2 + u_1 - \alpha_2 u_1^2), \\ f_2 &= \alpha_1^{-1}(\alpha_3 u_3 - u_2 - u_1u_2), \\ f_3 &= \alpha_4(u_1 - u_3), \\ \alpha_1, \alpha_3, \alpha_4 &> 0,\ \alpha_2 \approx 10^{-8}. \end{aligned}$$

The variables represent chemical concentrations and remain nonnegative.

Example 7: model equations for flame propagation. These equations are used in the theory of combustion. They have the form (1.1) with $d = 2$, ϵ constant, $\beta \equiv 0$, and

$$\begin{aligned} f_1 &= -u_1 e^{-\alpha_1/u_2}, \\ f_2 &= \alpha_2 u_1 e^{-\alpha_1/u_2}. \end{aligned}$$

Example 8: model equations for morphogenesis. These equations are used to model morphogenesis of patterns. They have the form (1.1) with $d = 2$, ϵ constant, $\beta \equiv 0$, and

$$\begin{aligned} f_1 &= -u_1u_2^2 + \alpha_1(1 - u_1), \\ f_2 &= u_1u_2^2 - (\alpha_1 + \alpha_2)u_2. \end{aligned}$$

Example 9: model for the spread of rabies in foxes. This equation is a model for the spread of rabies in Europe in the fox population. It has the form (1.1) with $d = 1$, ϵ constant, $\beta \equiv 0$, and

$$\begin{aligned} f_1 &= \alpha_1(1 - u_1 - u_2 - u_3)u_1 - u_3u_1, \\ f_2 &= u_3u_1 - (\alpha_2 + \alpha_3 + \alpha_1u_1 + \alpha_1u_2 + \alpha_1u_3)u_2, \\ f_3 &= \alpha_2u_2 - (\alpha_4 + \alpha_1u_1 + \alpha_1u_2 + \alpha_1u_3)u_3, \end{aligned}$$

where $\alpha_i > 0$ for all i and $0 < \alpha_4 < \left(1 + (\alpha_3 + \alpha_1)/\alpha_2\right)^{-1} - \alpha_1$.

Example 1, which is used as a prototypical example in this paper, was investigated analytically by Chafee [10], Bronsard and Kohn [6], [7], Carr and Pego [9], and Fusco and Hale [41], and numerically by Estep [27]. Descriptions of examples 2–7 can be found conveniently in Smoller [62]. Example 8 is analyzed in Pearson [57]. Example 9 is discussed in Murray [53]. See Fife [37] and Murray [53] for more information and references on the applications of reaction-diffusion equations. It is impossible to give a complete review of the literature on reaction-diffusion equations here. We only note that in addition to the references above, Aronson and Weinberger [3], Brown, Donne, and Gardner [8], Cohen [14], Cooley and Dodge [15], Hastings [42], Hodgkin and Huxley [44], Matano [50], Mimura, Nishiura, and Yamaguti [51], Rauch and Smoller [59], and Troy [64] contain material specifically considered in the preparation of this paper.

The complexity in the solutions arises primarily from the competition between reaction and diffusion and the nonlinear nature of the equations that allows localized behavior in classes of solutions. In particular, it is characteristic for solutions to

encompass behavior on several different scales simultaneously: long time phenomena such as metastability together with rapid transients; localized spatial behavior such as moving layers and blow-up together with global propagation of perturbations and pattern formation. It is also typical for different classes of solutions to exhibit different types of behavior to different extents, making it difficult to perform a meaningful general analysis. Adding to the difficulty is the fact that in many problems, we require information about solutions over moderate to long time intervals.

All of these points cause difficulties for the numerical analysis of reaction-diffusion equations. The upshot is that it is generally easy to produce completely inaccurate numerical solutions and moreover typical for initially accurate numerical solutions to become inaccurate at some point. It is therefore scientifically important to obtain a reasonably accurate and reliable estimate of the error of a numerical computation.

We explain these issues further using the bistable problem

$$\begin{cases} \dfrac{\partial u}{\partial t} - \epsilon \dfrac{\partial^2 u}{\partial x^2} = u - u^3, & 0 < x < 1,\ 0 < t, \\ \dfrac{\partial u}{\partial x}(0,t) = \dfrac{\partial u}{\partial x}(1,t) = 0, & 0 < t, \\ u(x,0) = u_0(x), & 0 < x < 1. \end{cases} \tag{1.2}$$

as an example. The dynamical properties of solutions of (1.2) have generated considerable interest in part because it is one of the simplest problems that produce evolution to equilibrium in the presence of competing stable steady states. The long time behavior of the solutions is now well understood, see Bronsard and Kohn [6], Carr and Pego [9] and Fusco and Hale [41]. When ϵ is sufficiently small, the only stable equilibrium solutions are $u \equiv 1$ and $u \equiv -1$ and all solutions, except unstable equilibrium solutions, converge to one of these two steady-states. However, this convergence may be extremely slow because solutions can exhibit dynamic metastability. Generic initial data forms a pattern of transition layers between the values -1 and 1 during an initial transient, after which the layers coalesce by moving more or less in a horizontal direction. The time scale for substantial motion of the layers is $\exp(Cd/\sqrt{\epsilon})$ where C is a constant and d is the distance between neighboring layers. When two layers become sufficiently close, a rapid transient occurs during which the layers collapse together. The solution then forms a new, simpler metastable pattern and the process begins anew.

We illustrate with a computation made with $\epsilon = .0009$ and

$$u_0(x) = \begin{cases} \tanh((.2-x)/(2\sqrt{\epsilon})), & 0 \le x < .28, \\ \tanh((x-.36)/(2\sqrt{\epsilon})), & .28 \le x < .4865, \\ \tanh((.613-x)/(2\sqrt{\epsilon})), & .4865 \le x < .7065, \\ \tanh((x-.8)/(2\sqrt{\epsilon})), & .7065 \le x \le 1, \end{cases}$$

which produces a function that is very close to a metastable state. We display the evolution of the corresponding numerical solution in Fig. 1.1. The "well" on the left is slightly thinner and collapses first. We estimate the error of this computation and subsequent computations in this section to be less than 10% and explain the reason later on.

Numerical evidence was important in the initial stages of the analysis of the bistable problem. The initial transient and development of the layers, the shape and motion of the layers, the time scales for evolution all were explored initially using

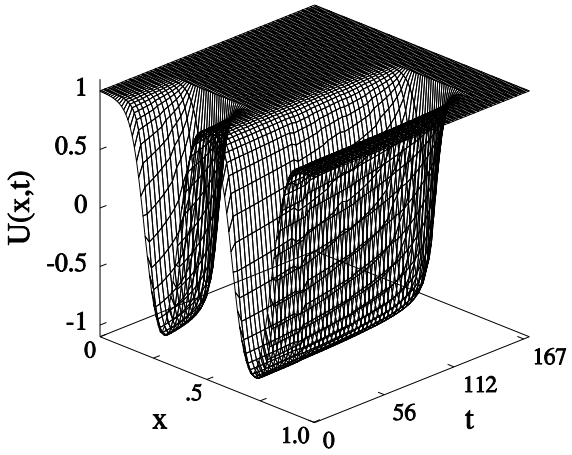

FIG. 1.1. *Evolution of a metastable solution starting with two "wells" and* $\epsilon = .0009$. *The left well is thinner and collapses at* $t \approx 41$ *and the second well collapses at* $t \approx 141$.

numerical solutions and these results provided the basis for the mathematical analysis. Experimentation also revealed that care is needed when computing. In particular, computing without a sufficiently fine time step or space mesh causes "locking" in which a metastable pattern actually becomes stable. (See Elliot and Stuart [18] for a discussion of this phenomena). This does not require gross inaccuracy in the numerical solution and the only indication that a computation is incorrect is the change in time scales. Of course to determine this, the correct time scale must be known a priori.

The fact that little is actually known about the error of individual numerical solutions of reaction-diffusion equations might at first seem surprising. After all, there is a significant amount of literature devoted to a priori error analysis of numerical methods for such problems. The results generally take the form

$$\|e(t)\| \leq e^{Lt}\, C(u)\, (h^p + k^q) \tag{1.3}$$

where $e(t)$ denotes the error at time t, $\|\ \|$ some norm, $L = L(\epsilon, f)$ is a positive constant, h and k are parameters measuring the space and time discretization, i.e. mesh size and time step, p and q are the respective orders of accuracy, and $C(u)$ is a function of u and its derivatives of order depending on p and q. On investigation, the flaws in such an estimate are revealed. For one thing, L is generally quite large. For example, in the computation shown in Fig. 1.1, L is on the order of 1000. This means that the estimate (1.3) is meaningful only over a short initial transient in general. Moreover, the size of $C(u)$ is unknown and very often, we do not even know if u has sufficiently many derivatives for $C(u)$ to be defined. Because of requirements of compatibility between reaction term, the initial data, and the boundary of the domain of the problem, there is generally an upper limit to the number of derivatives of a solution of a reaction-diffusion equation that are defined.

The last two factors on the right-hand side of (1.3) arise from standard interpolation error considerations. Similar quantities would appear in an error estimate for any standard approximation computed by interpolation or projection of the solution, provided it was known. The first factor on the right-hand side of (1.3) is not common to error bounds on interpolants of known functions. It arises because of the possibility of accumulation of errors occurring as the differential equation is solved in time and is

therefore a reflection of the stability properties of the problem. The exponential form of this factor is generally the result of a Gronwall argument that estimates the effect of the reaction term by taking the worst possible rate of growth of perturbations it can induce.

Of course if the error bound (1.3) is accurate, i.e. the error on the left-hand side is more or less the size of the bound on the right-hand side, there is nothing much for it. We can try to find a more accurate way to solve the problem or to use the numerical solution in some other fashion. But in our experience, rapid exponential growth of errors is rarely found outside short transition periods and the error bound (1.3) is generally not accurate past initial transients.

In the bistable problem for example, experimentation shows that the time scale for the collapse of the wells is virtually the same for any numerical solution computed with time steps and space meshes that are finer than some minimum level of discretization determined primarily by ϵ. The time scales for solutions that start with perturbed data are also the same provided the initial positions of the transition layers remain fixed. Furthermore in Estep [27], we prove that approximations of metastable layers computed using schemes that preserve the energy functional that exists for the continuous problem move on the same time scale as the true layer provided the space meshes and time steps are sufficiently fine depending on ϵ. None of these facts amount to an analytic estimate that says that the error of numerical solutions remains small for all time, but they do suggest that the exponential bound in (1.3) is too severe.

Part of the difficulty is caused by the ambition inherent in the goal of a priori error analysis, which is essentially to estimate the error of *any* numerical solution on *any* interval in the possible range of times without using any particular information about a solution. This might be a reasonable approach for many linear problems, for which stability and regularity properties of solutions are often uniform. However, nonlinear problems characteristically allow localized behavior in classes of solutions.

This is easy to demonstrate with the bistable problem. In Fig. 1.2, we show two sets of initial data and the corresponding solutions at time $t \approx 3.33$. The initial values are small oscillations around a constant value: one is centered around the unstable steady-state $u \equiv 0$ and the other around the stable steady-state $u \equiv 1$. Though the

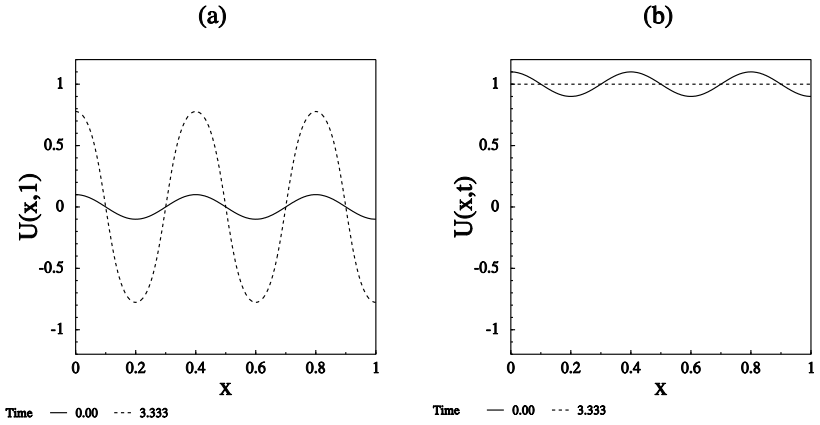

FIG. 1.2. *Plots of solutions of the bistable equation at $t \approx 3.333$ and the corresponding initial data. The data have the same regularity but the solutions at later times have much different regularity properties.*

two initial functions have the same regularity, only a short time later, the corresponding solutions have much different regularity properties. This is example of localized regularity properties of solutions. The ability to distinguish different regularity properties is necessary to estimate the error of numerical solutions accurately. Fortunately, differences in regularity are often obvious, even to the eye. Unfortunately, stability properties of solutions are also generally localized and the effects of stability can be difficult to detect. In Fig. 1.3 (a), we plot the initial data used in Fig. 1.2 centered around $u \equiv 0$ together with a slightly perturbed function. In Fig. 1.3 (b), we show the corresponding solutions at time $t \approx 10.0$. We plot the results of a similar computation using the data in Fig. 1.2 centered around $u \equiv 1$ in Fig. 1.3 (c) and (d). Solutions

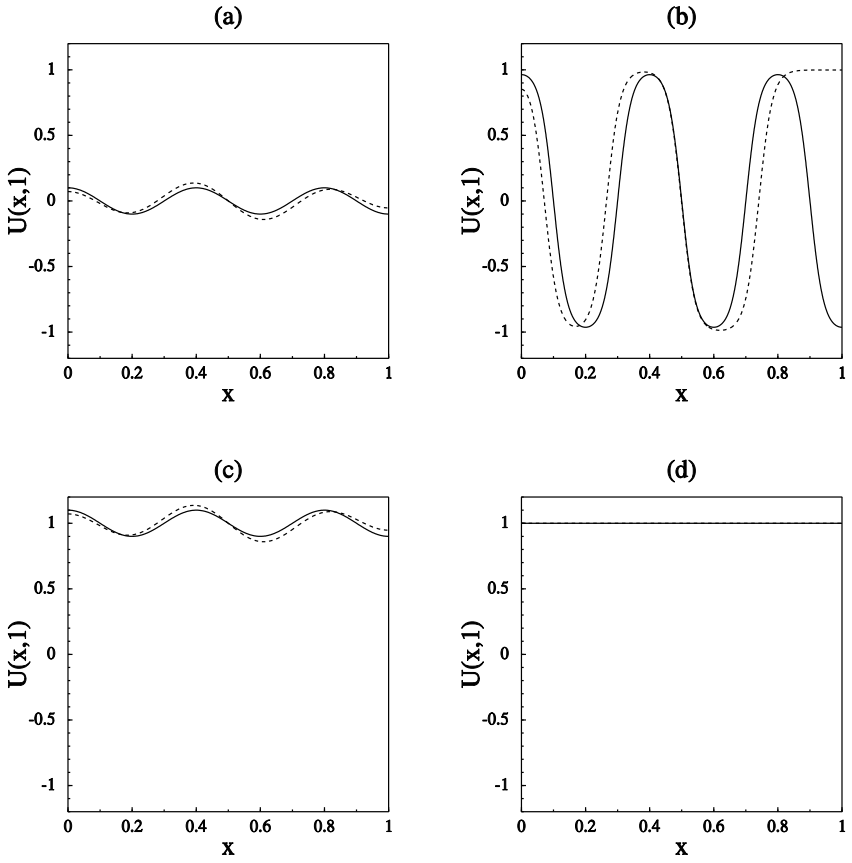

FIG. 1.3. *Solutions of the bistable problem. (a) Plots of the initial data and a slight perturbation and (b) plots of the corresponding solutions at* $t \approx 10$. *(c) Plots of the initial data and a slight perturbation and (d) plots of the corresponding solutions at* $t \approx 3.33$.

that begin near the fixed point $u \equiv 1$ are very stable in the sense that nearby data all converge rapidly to the same function. But the problem is very sensitive to perturbations in data that are near 0 as can clearly be seen. Moreover, note that the two solutions shown in Fig. 1.3 (b) are quite different in the sense of the subsequent evolution of metastable states, yet essentially possess the same regularity properties and on that basis, it would be difficult to distinguish between them.

As we show below, numerical errors do accumulate at an exponential rate in the bistable problem during transient periods like those when the patterns displayed in Fig.'s 1.2 and 1.3 form, as suggested by (1.3). However, the accumulated numerical error actually decreases abruptly after the pattern has formed and the slow-motion regime takes over. Then a new period of accumulation begins.

In other problems, like the Lorenz system discussed below, solutions can undergo dramatic changes in stability without any changes in regularity properties. Thus, it appears necessary to take into account both the regularity and the stability properties of individual numerical solutions in order to obtain an accurate estimate of their error. Attempting to directly estimate the rate of accumulation of error is a feature of the approach to error estimation proposed in this paper that is not commonly encountered in the literature. This aspect is the most problematic part of the proposed theory in terms of costs and analytical difficulties, but practical experience appears to present no alternatives.

Because a general a priori analysis must treat the worst possible case in terms of regularity and stability properties of solutions, it is not surprising that the results tend to overestimation. The solution that we propose is to compute an a posteriori estimate of the error of each specific numerical solution using information obtained from the numerical solution itself. In other words, we propose using computational work to make up for our analytical deficiencies. Accordingly, the mathematical analysis in this paper is directed towards justifying the process used to produce the computational error estimate rather than towards estimating the size of the error itself. There are a priori ingredients in the theory outlined in this paper since, of course, we cannot ignore the classic considerations of well-posedness, consistency, and stability. But because we redirect the analysis away from estimating the size of the error itself, we can use these ingredients in a different, and in some sense more efficient, manner. For example, we employ an a priori error estimate similar in form to (1.3). However, we only use the estimate over one time step.

In this paper, we are concerned with the "pointwise" approximation of individual solutions, which is fundamentally important in applications. Of course, there are different ways to analyze differential equations numerically. For example, we might require the numerical method to accurately approximate some dynamical (long-time) feature of the continuous problem rather than individual trajectories. Another possibility is to get information from numerics using a statistical approach: trying to measure some average quantities from many computations without expecting individual computations to be accurate over a long time. But in our opinion, any analysis that uses information computed from individual trajectories requires some notion of accuracy for the individual trajectories. Simply, we cannot expect to get any useful information from completely inaccurate, non-physical computations. For example, numerical solutions of the bistable equation that have become "locked" will not give the correct time scales for substantial motion of layers. One way to view the work in this paper is as a method for determining the time scale over which accurate trajectories can be computed.

The main contributions contained in this paper are these. We develop a general theory of a posteriori computational error estimation for numerical solutions of systems of nonlinear reaction-diffusion equations that may include coupled ordinary and partial differential equations. The theory is based on computing the residual error of the approximation and estimating the accumulation of errors. We present a new approach to analyze a posteriori error estimates for numerical solutions of differential

equations that divulges both the theoretical and practical meaning of the estimates. The analysis is based on a few reasonable a priori assumptions about the problems and the numerical methods, and in particular, requires minimal regularity of the solutions. We also discuss some important issues that arise in practice when estimating the error of numerical solutions. We use the theory to estimate the error of typical numerical solutions of nine standard reaction-diffusion models and for the first time, make a systematic comparison of the time scale over which accurate numerical solutions can be computed for these problems. Finally, we apply the general theory to the class of problems that admit invariant regions for the solutions, which includes seven of the main examples above, and obtain stronger results in the analysis of the a posteriori error bound.

The work in this paper is part of an on-going collaboration devoted to deriving a posteriori error estimates for general differential equations based on residual errors and Galerkin orthogonality. The review article Eriksson, Estep, Hansbo, and Johnson [20] contains a description of the work and an overview of the literature associated to this project while the text [21] presents the theory for linear ordinary and partial differential equations. This approach was suggested originally for ordinary differential equations by Johnson in [46] and first carried out for linear parabolic equations in Eriksson and Johnson [23]. Estep first applied this approach to nonlinear ordinary differential equations in [28].

The general a posteriori analysis of Section 2 is closely related to Eriksson and Johnson's work on linear and nonlinear parabolic problems in [22]–[26] as well as the analysis for nonlinear ordinary differential equations in Estep [28] and Estep and French [29]. Eriksson and Johnson have concentrated their analysis mainly on strongly parabolic problems for which it is possible to derive accurate a priori bounds on the rate of accumulation of errors. In this paper, we treat systems of coupled parabolic and ordinary differential equations under more general assumptions on the stability properties of the solutions which allows a greater variety of behavior in solutions than occurs for strongly parabolic problems. We have to modify the a posteriori analysis to handle the more general system, especially with regard to the presence of ordinary differential equations, and we have to take a different approach to analyzing the a posteriori error estimate. We also treat systems that admit invariant rectangles for the solutions, showing how to preserve this special stability property in the finite element discretizations and discuss the consequences for the a posteriori error estimate.

Estep [28] and Estep and French [29] considered initial value problems under general assumptions allowing significant growth of errors and systematically developed the idea of computationally estimating the rate of accumulation of errors. Estep and Johnson [31] used this approach to analyze the chaotic behavior of the Lorenz and Duffing systems. Eriksson and Johnson [25], [26] discussed the rate of accumulation of errors in nonlinear parabolic equations analytically using the a posteriori estimate. Estep and Williams [35] discussed the practical issues involved in estimating the error of numerical solutions of large, sparse problems and computationally estimated the rate of accumulation of error in the bistable example (1.2). More recently Sandboge [61] studied the computational estimation of the rate of accumulation of errors in nonlinear parabolic equations. This paper presents a new approach for interpreting and analyzing the a posteriori error estimate including a result that shows that under minimal assumptions the residual error of a numerical solution can always be made small by refinement and a more complete analysis of the estimation of rates of accumulation of errors. This paper also discusses some general issues that arise in the

implementation of the theory in the numerical solution of parabolic partial differential equations.

Other approaches to a posteriori error estimation of numerical solutions of parabolic problems can be found in the work of Adjerid and Flaherty, [1], [2], Bieterman and Babuška [4], [5], Moore [52], and R. Nochetto, M. Paolini, and C. Verdi [54], [55], [56]. Most of this work is not based on the residual error, but instead depends on a comparison between approximations of different accuracies to give some estimate of a "local" error.

Lastly, an important early mathematical paper that discusses the use of adaptive finite element methods for parabolic problems is Dupont [17]. The paper by Hoff [45] discusses the preservation of invariant rectangles for reaction-diffusion equations under discretization by finite difference methods.

The plan of the paper. In Section 2, we develop a general theory of a posteriori error estimation of numerical solutions of (1.1) in one and two space dimensions. We carry out the analysis for two finite element space-time discretizations called the continuous and discontinuous Galerkin methods. With some straightforward modifications, the theory applies to other finite element methods and also to difference schemes that can be written as a Galerkin finite element method with an appropriate choice of quadrature to evaluate the integrals in the variational formulation. Many standard schemes can be written in this way and we consider one example in detail.

The a posteriori theory is based on estimating the error in terms of the residual error of the numerical solution which, roughly speaking, is the remainder resulting from substituting the approximate solution into the differential equation. The residual error is related to the error of the approximate solution through a proportionality factor determined by the stability properties of the problem called a stability factor. The stability factor is something akin to the condition number of a matrix in the relationship between the error and residual error of a computed solution of a linear system of equations. In this case, the stability factor is given by some seminorms on the solution of a linearized adjoint problem to the original differential equation and it is a measure of the sensitivity of numerical solutions of the problem to computational errors.

In Section 3, we analyze the quantities in the a posteriori error estimate with the goal of understanding its implications. In particular, we address two issues: is the residual error defined and what is its size, and is the stability factor defined what is its size? We answer these questions using a a set of a priori assumptions on the continuous problem and the numerical method that are typical of the kind of results that are the goal of classical analysis of reaction-diffusion equations. These assumptions hold for various specific examples and we show that they hold for general systems of the form (1.1) with constant diffusion.

We begin by showing that the residual error on any time step can be made arbitrarily small by refining the space mesh and time step. We also derive precise estimates on the rate that the residual error tends to zero as the discretization is refined. We also discuss the size of the stability factor and conditions that guarantee that it can be approximated computationally. We conclude the section with a *stability factor gallery* that displays the sensitivity to growth of discretization error in numerical solutions of the various examples 1–9 above. We believe that this is the first systematic study of the time scale for accurate numerical solution for these problems.

In Section 4, we discuss issues that arise when the a posteriori error estimate is incorporated into a code that solves reaction-diffusion equations numerically. The

dominate theme is the approximation of the stability factor, since this involves the numerical solution of the linear adjoint problem. We present the results of an experiment that tests the accuracy and reliability of the a posteriori error estimate using the bistable example (1.2). We also discuss an unresolved issue having to do with linearization and present some numerical results regarding this point.

The analysis up to this point is conducted under very mild assumptions on the problem; really no more than necessary to guarantee that the problem is well-posed in a convenient Sobolev space over short time intervals. The results we obtain reflect this. In particular, while the error can be estimated using the a posteriori error estimate, the results do not imply that the error of the numerical solution decreases if the residual error of the numerical solution decreases. The reason is that the stability factor depends on the approximation itself and so it can grow if the discretization is refined. We might expect such behavior in a problem in which solutions "blow-up" at a finite time for example.

In Section 5, we obtain stronger results about the quantities in the a posteriori error estimate by considering systems of reaction-diffusion equations of the form (1.1) that admit invariant rectangles for the solutions under the assumption that there is an invariant rectangle for the numerical method as well. We treat this class of problems because it contains many important models. Examples 1–7 above are problems that admit invariant rectangles.

We show how to preserve invariant rectangles under discretization in two ways. First we show that there is an invariant rectangle for the approximation that is close to an invariant rectangle for the true solution if the residual errors are kept sufficiently small independent of time. This result applies to all of the finite element methods considered in the previous sections. Second, we show that certain finite element methods have the special stability property that any invariant rectangle for the solution of the differential equation is also invariant for the approximate solution. These methods require some modification of the general theory of a posteriori error analysis and we discuss this as well.

The remaining sections contain details of the analysis.

Acknowledgment. The authors gratefully thank Roland Freund, Jack Hale, Theodore Hill, David Hoff, Jeffrey Rauch, and especially Claes Johnson for useful advice.

2. General a posteriori error analysis: ingredients for computational error estimation.

We begin by developing a general theory of a posteriori error analysis of approximate solutions based on residual errors. Roughly speaking, the *residual error* of an approximate solution is obtained by substituting the approximation into the differential equation. The residual error is related to the error of the approximate solution through a factor determined by the stability properties of the problem called a *stability factor*.

2.1. An analogous problem in numerical linear algebra. Our approach is explained easily in the context of the numerical solution of a linear system of equations. The problem there is to estimate the error of a numerical solution $\vec{X}$ of

$$A\vec{x} = \vec{b}.$$

The residual error of $\vec{X}$ is defined simply as

$$\vec{R} = A\vec{X} - \vec{b}$$

and is generally not zero. The point is to relate the unknown error $\vec{e} = \vec{x} - \vec{X}$ to the computable residual error $\vec{R}$.

There are at least two ways to do this. First, we can use the fact that the residual error of the true solution is zero to write

$$A\vec{e} = -\vec{R}.$$

We can then try to obtain an approximation of the error by solving this equation approximately in some fashion. This is not the approach that we use in this paper but it is related to the classic method of estimating the error using high order asymptotic error estimates.

Instead, we settle for the less ambitious goal of obtaining an estimate on the size of a projection of the error. We introduce the dual problem

$$A^\top \vec{\phi} = \vec{\psi},$$

where $\vec{\psi}$ is any unit vector. Computing, we find

$$|(\vec{e}, \vec{\psi})| = |(\vec{e}, A^\top \vec{\phi})| = |(A\vec{e}, \vec{\phi})| \leq \|\vec{\phi}\| \, \|\vec{R}\|.$$

Thus we obtain an estimate on the size of the projection of the error in the direction of the data for the dual problem in terms of the sizes of the solution of the dual problem and the residual error. If we could be so fortunate to choose $\vec{\psi} = \vec{e}/\|\vec{e}\|$ for example, then we would get an estimate on $\|\vec{e}\|$.

We call $\|\vec{\phi}\|$ the stability factor for this problem. It is related to the condition number of A. In fact it follows that

$$\big|\big(\frac{\vec{e}}{\|\vec{x}\,\|}, \psi\big)\big| \leq \mathrm{cond}_\psi(A) \, \frac{\|\vec{R}\|}{\|\vec{b}\|},$$

where $\mathrm{cond}_\psi(A) = \|\vec{\phi}\| \, \|A\| = \|A^{-\top}\psi\| \, \|A\|$. Hence the stability factor is a measure of the sensitivity of numerical solutions of the problem to computational errors.

First, we lay the groundwork for the analysis for differential equations by defining the schemes, residual errors, and stability factors. Next, we derive the a posteriori error estimate.

2.2. The continuous problem and its discretization. We study a system of D reaction-diffusion equations consisting of d, $1 \leq d \leq D$, parabolic equations and $D - d$ ordinary equations for the $\mathbf{R}^D$ valued function $u = (u_i)$:

$$\begin{cases} \dot{u}_i - \nabla \cdot (\epsilon_i(u,x,t)\nabla u_i) = f_i(u,x,t), & (x,t) \in \Omega \times \mathbf{R}^+, \ 1 \leq i \leq D, \\ u_i(x,t) = 0, & (x,t) \in \partial\Omega \times \mathbf{R}^+, \ 1 \leq i \leq d, \\ u(x,0) = u_0(x), & x \in \Omega, \end{cases} \tag{2.1}$$

where Ω is an interval in $\mathbf{R}^1$ and a convex polygonal domain in $\mathbf{R}^2$ with boundary $\partial\Omega$, $\dot{u}_i$ denotes the partial derivative of u_i with respect to time, and there is a constant $\epsilon > 0$ such that

$$\epsilon_i(u,x,t) \geq \epsilon \text{ for } 1 \leq i \leq d \text{ and } \epsilon_i(u,x,t) \equiv 0 \text{ for the rest.}$$

We also assume that $\epsilon = (\epsilon_i)$ and $f = (f_i)$ have smooth second derivatives and for simplicity, we write $\epsilon_i(u, x, t) = \epsilon_i(u)$ and $f(u, x, t) = f(u)$. We use u^p and u^o to denote the parts of u associated to the parabolic and ordinary differential equations respectively. In other words, $u_i^p = u_i$ for $1 \leq i \leq d$ and $u_i^p = 0$ for $d < i \leq D$ and $u^o = u - u^p$.

The presence of ordinary differential equations in the system (2.1) has strong consequences for the regularity properties of solutions. In particular, we can expect parabolic smoothing to occur only for u^p, while the regularity of u^o is generally determined by the regularity of u^p and the initial data since f is smooth. This affects the analysis of the approximation error, for which we try to assume minimal regularity of solutions.

Remark 2.1. It is completely straightforward to extend the a posteriori error estimate derived in this section to systems of equations that include convection terms of the form $\beta_i(u, x, t) \cdot \nabla u_i$ in the i'th equation, $1 \leq i \leq d$, as well as problems with other boundary conditions. We do not give the details to save space.

We consider two finite element space-time discretizations of (2.1) called the continuous and discontinuous Galerkin methods. A finite element approximate solution is a piecewise polynomial function that solves the weak or variational formulation of (2.1) for all test functions in an appropriate finite dimensional space. The variational formulation is obtained by multiplying (2.1) by a test function, integrating over time and space, and using Green's formula on the diffusion term. Both methods use continuous piecewise linear functions in space, yielding nominal second order accuracy. For simplicity, we consider the piecewise constant and piecewise linear discontinuous Galerkin methods, yielding nominal first and third order accuracy in time, and the piecewise linear continuous Galerkin method, yielding nominal second order accuracy in time. The analysis however extends directly to methods with higher order accuracy in time.

With appropriate choice of quadrature, these Galerkin methods yield standard difference schemes. Conversely, many standard first to third order implicit difference schemes for (2.1) can be interpreted as one of these finite element methods implemented with a suitable quadrature. However, the finite element framework is more convenient for a posteriori error analysis.

We partition $[0, \infty)$ as $0 = t_0 < t_1 < t_2 < \cdots < t_n < \ldots$, denoting each time interval by $I_n = (t_{n-1}, t_n]$ and time step by $k_n = t_n - t_{n-1}$. To each interval I_n, we associate a triangulation $\mathcal{T}_n$ of Ω arranged so the union of the elements in $\mathcal{T}_n$ is Ω while the intersection of any two elements is either a common edge, node, or is empty. In order to preserve approximation properties in two space dimensions, we assume that the smallest angle of any triangle in a triangulation is bounded below by a fixed constant, or equivalently that there is a constant λ_0 independent of the triangulation $\mathcal{T}_n$ such that $\text{area}(K) \geq \lambda_0 \, \text{diam}(K)^2$, where $\text{diam}(K)$ is the length of the largest side of K, for any triangle $K \in \mathcal{T}_n$.

Remark 2.2. We use λ_i to denote mesh 'parameters' quantifying the qualities of the mesh and time steps. Constants in the estimates below typically depend on these parameters.

Note that mesh changes can occur across time nodes. To measure the size of the elements of $\mathcal{T}_n$, we use a piecewise constant function h_n, called the *mesh function*, defined so $h_n|_K = \text{diam}(K)$ for $K \in \mathcal{T}_n$, We also use $h_{n,\min} = \min h_n(\cdot)$ and $h_{n,\max} = \max h_n(\cdot)$ and denote the global mesh function by h, where $h|_{I_n} = h_n$. Similarly, we

use k to denote the piecewise constant function that is k_n on I_n. When the time level is clear in the context, we abuse notation by dropping the subscript n.

The approximations are polynomials in time and piecewise polynomials in space on each space-time "slab" $S_n = \Omega \times I_n$. In space, we let $V_n \subset (H_0^1(\Omega))^d \times (H^1(\Omega))^{D-d}$ denote the space of piecewise linear continuous vector-valued functions $v(x) \in \mathbf{R}^D$ defined on $\mathcal{T}_n$, where the first d components of v are zero on $\partial\Omega$. Then on each slab, we define

$$W_n^q = \big\{ w(x,t) : w(x,t) = \sum_{j=0}^{q} t^j v_j(x),\ v_j \in V_n,\ (x,t) \in S_n \big\}.$$

Finally, we let W^q denote the space of functions defined on the space-time domain $\Omega \times \mathbf{R}^+$ such that $v|_{S_n} \in W_n^q$ for $n \geq 1$. Note that functions in W^q are generally discontinuous across the discrete time levels and we denote the jump across t_n by $[w]_n = w_n^+ - w_n^-$ where $w_n^{\pm} = \lim_{s \to t_n \pm} w(s)$. To define the methods, we use the L^2 projection operator P_n onto V_n, i.e. $P_n : L^2(\Omega) \to V_n$ is defined by $(P_n v, w) = (v, w)$ for all $w \in V_n$, where $(\cdot,\cdot)$ denotes the $L_2(\Omega)$ inner product. We use $\| \quad \|$ for the L_2 norm. The global projection operator P is defined by setting $P = P_n$ on S_n.

The *continuous Galerkin* cG(q) approximation $U \in W^q$ satisfies $U_0^- = P_0 u_0$ and for $n \geq 1$, the *Galerkin orthogonality relation*

$$\begin{cases} \displaystyle\int_{t_{n-1}}^{t_n} \big((\dot{U}_i, v_i) + (\epsilon_i(U)\nabla U_i, \nabla v_i) \big)\, dt = \int_{t_{n-1}}^{t_n} (f_i(U), v_i)\, dt \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{for all } v \in W_n^{q-1},\ 1 \leq i \leq D, \\ U_{n-1}^+ = P_n U_{n-1}^-. \end{cases} \tag{2.2}$$

Note that U is continuous across time nodes over which there is no mesh change. In particular, it is usually the case that $U_0^- = U_0^+$. The *discontinuous Galerkin* dG(q) approximation $U \in W^q$ satisfies $U_0^- = P_0 u_0$ and for $n \geq 1$,

$$\int_{t_{n-1}}^{t_n} \big((\dot{U}_i, v_i) + (\epsilon_i(U)\nabla U_i, \nabla v_i) \big)\, dt + \big([U_i]_{n-1}, v_i^+ \big) = \int_{t_{n-1}}^{t_n} (f_i(U), v_i)\, dt$$
$$\text{for all } v \in W_n^q,\ 1 \leq i \leq D. \tag{2.3}$$

See Eriksson, Estep, Hansbo, and Johnson [21] for a general introduction to these methods. Note that the true solution satisfies both (2.2) and (2.3).

To illustrate, we discretize the scalar problem

$$\begin{cases} \dot{u} - \Delta u = f(u), & (x,t) \in \Omega \times \mathbf{R}^+, \\ u(x,t) = 0, & (x,t) \in \partial\Omega \times \mathbf{R}^+, \\ u(x,0) = u_0(x), & x \in \Omega, \end{cases} \tag{2.4}$$

using the dG(0) method. Since U is constant in time on each time interval, we let $\vec{U}_n$ denote the M_n vector of nodal values with respect to the nodal basis $\{\psi_{n,i}\}_{i=1}^{M_n}$ for V_n on I_n. We let B_n : $\big(B_n\big)_{ij} = \big(\psi_{n,i}, \psi_{n,j}\big)$ for $1 \leq i,j \leq M_n$ and $B_{n,n-1}$: $\big(B_{n,n-1}\big)_{ij} = \big(\psi_{n,i}, \psi_{n-1,j}\big)$ for $1 \leq i \leq M_n$, $1 \leq j \leq M_{n-1}$ denote the *mass matrices* and A_n : $\big(A_n\big)_{ij} = \big(\nabla\psi_{n,i}, \nabla\psi_{n,j}\big)$ denote the *stiffness matrix*. Then U_n satisfies

$$\big(B_n - k_n A_n\big)\vec{U}_n - \vec{F}(U_n^-)k_n = B_{n,n-1}\vec{U}_{n-1}, \quad n \geq 1,$$

where $(\vec{F}(U_n^-))_i = (f(U_n^-), \psi_{n,i})$. In practice, these integrals, and often the integrals for the mass matrices B_n and $B_{n,n-1}$, are evaluated using a quadrature formula such as the composite trapezoidal rule or *lumped mass quadrature rule*, see Section 3.3. If M_n is constant and the lumped mass quadrature is used to evaluate the coefficients of B_n and $B_{n,n-1} = B_n$, then the resulting set of equations for the dG(0) approximation is the same as the equations for the nodal values of the backward Euler difference scheme for (2.4). Similarly, the cG(1) method is related to the Crank-Nicolson scheme and the dG(1) method is related to the third order sub-diagonal Padé difference scheme. See Estep [28], Estep and French [29], Estep and Larsson [33], and Estep, Johnson, and Larsson [32] for further details.

2.3. The residual error. The intuition is that the residual error is computed by substituting the approximation into the differential equation. Rigorously however, the approximation does not have sufficient regularity to substituted into (2.1) pointwise. This difficulty is overcome by using the variational formulation of (2.1) and interpreting the residual error in the sense of distributions. The result is that there are two contributions to the total residual error: (1) the remainder left over from substituting U into (2.1) inside elements and time intervals where U is smooth; (2) terms arising from the low order regularity of the approximation across element boundaries and time nodes. Of course, these are the two ways in which the approximate solution is different from the true solution: it does not satisfy the differential equation exactly and it does not have as many derivatives. It turns out to be important to distinguish the contributions to the total residual error from these two sources since these errors accumulate at different rates. Thus, we split the total residual error into several contributing residual errors. The suitability of the following definitions, which are suggested by the analysis of Eriksson and Johnson in [23] and [25], becomes apparent when we derive the a posteriori error estimate and analyze its meaning.

First, we define the residual errors arising from space discretization. It is natural to first divide the residual into two parts corresponding to the parabolic and the ordinary differential equations in (2.1) because of their different regularity properties. These residuals are distinguished by a superscript p or o. Inside an element K, we define the two contributions:

$$R_x^p(U)_i = \dot{U}_i - \nabla \cdot \epsilon_i(U)\nabla U_i - f_i(U), \quad 1 \le i \le d, \tag{2.5}$$

$$R_x^o(U)_i = \dot{U}_i - f_i(U), \quad d < i \le D, \tag{2.6}$$

while the remaining coefficients are set to zero. Here the derivatives are taken only in the interior of the triangle and in particular in the case of constant diffusion, $\nabla \cdot \epsilon_i \nabla U_i \equiv 0$.

There is an additional contribution to the residual for the parabolic equation arising from discontinuity in the first derivative of U across element boundaries. This is defined element-wise for $K \in \mathcal{T}_n$ and $1 \le i \le d$:

$$R_2^p(U)_i = \frac{C_t}{2}\left(h(K)\,\mathrm{area}(K)\right)^{-1/2}\left(\int_{\partial K \setminus \partial\Omega} \left(n_{\partial K} \cdot \epsilon_i(U)[\nabla U_i]_{\partial K}/2\right)^2 \, ds\right)^{1/2}, \tag{2.7}$$

where $[\nabla U]_{\partial K}$ denotes the jump in ∇U across the edge ∂K, $n_{\partial K}$ is the unit outward normal to ∂K, and C_t denotes the constant in the trace inequality applied on an element K (see (6.6). This definition fits the intuition of a "discrete" second derivative in

the case of piecewise linear approximations. For example in the case of constant diffusion and a uniform triangulation of a two dimensional domain under the assumptions on the mesh made above, there is a constant c such that

$$R_2^p(U)_i = c\left(\sum_{\partial K}\left(n_{\partial K}\cdot\frac{[\nabla U_i]_{\partial K}}{h(K)}\right)^2\right)^{1/2},$$

where the sum is over the three sides of K.

Finally, we turn to the residual associated with the time discretization where it is unnecessary to distinguish between the parabolic and ordinary differential equations. We define the $\mathbf{R}^D$ valued time residuals interval-wise on S_n by

$$R_t(U)_i = \dot{U}_i - (\nabla\cdot\epsilon_i(U)\nabla)_h U_i - f_i(U), \quad 1\le i\le D, \tag{2.8}$$

for the cG method and

$$R_t(U)_i = |\dot{U}_i - (\nabla\cdot\epsilon_i(U)\nabla)_h U_i - f_i(U)| + k_n^{-1}|[U_i]_{n-1}|, \quad 1\le i\le D, \tag{2.9}$$

for the dG methods, where the *discrete diffusion operator* $(\nabla\cdot\epsilon_i(U)\nabla)_h$, $1\le i\le D$, is defined on I_n by

$$\big((\nabla\cdot\epsilon_i(U)\nabla)_h V, W\big) = \big(\epsilon_i(U)\nabla V, \nabla W\big) \text{ for all } W\in V_n.$$

There is an additional term in the residual associated to the dG method arising from the discontinuity in the approximation across time nodes.

In the case of the scalar problem (2.4) with constant diffusion, the time residual error becomes

$$R_t(U) = \begin{cases} \dot{U} - \Delta_{h_n} U - f(U), & \text{(cG)} \\ |\dot{U} - \Delta_{h_n} U - f(U)| + k_n^{-1}|[U]_{n-1}|, & \text{(dG)} \end{cases}$$

where the *discrete Laplacian* Δ_{h_n} is the map from V_n into V_n with matrix $B_n^{-1}A_n$, where B_n and A_n are the mass and stiffness matrices respectively. This is the residual of the cG resp. dG time integration schemes applied to the system of ordinary differential equations in t that result from the semi-discretization in space of (2.1) by the piecewise linear Galerkin finite element method.

To illustrate these definitions, we plot the residual errors for the bistable example (1.2) discussed in Section 1 computed with the dG(0) method using a uniform space mesh with 513 elements and the uniform time step .00111. In Fig. 2.1(a), we see that the initial transient and the later transients are clearly indicated by the size of the time residual R_t. Likewise, as expected, the space residuals decrease markedly after the collapse of each well. Note that the non-uniform behavior of the residual errors in both x and t suggest that efficiency could be gained by using non-uniform meshes and time steps.

2.4. The dual problem and a formula for the error. As in the linear algebra example, the next step is to determine a relationship between the residual errors and the error of the approximate solution by introducing a dual problem to the differential equation. However, the argument is complicated by the fact that the differential equation (2.1) is nonlinear, and in order to obtain a linear dual problem, we linearize

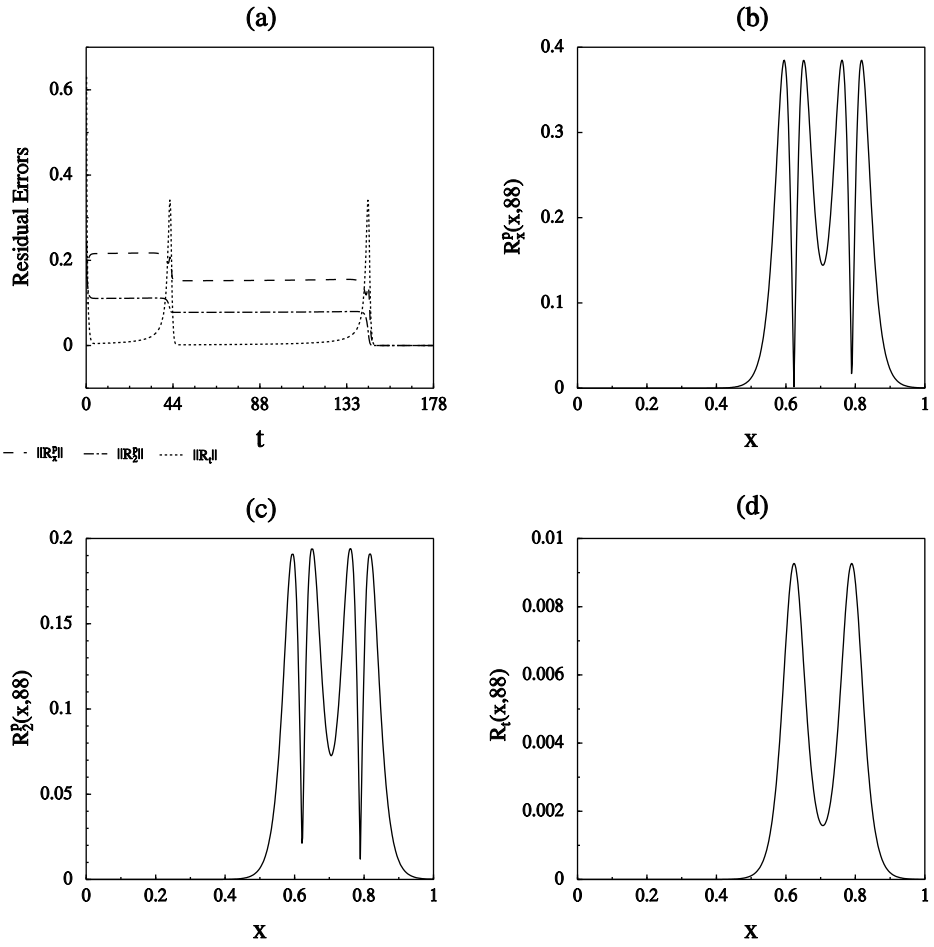

FIG. 2.1. *Residual errors for the bistable example (1.2) computed using the dG(0) method with* $M_n = 512$ *and* $k_n = .00111$ *for all* n. *(a) Plots of the* L_2 *norms of the residuals versus time. (b)-(d) Plots of the residual functions versus* x *at* $t \approx 88$.

(2.1). Classically, a differential equation is linearized either around the true solution trajectory or around the approximate solution trajectory and then the error of the linearization is treated as a high order perturbation to the linearized equation. This approach is problematic when applied to (2.1) however because controlling the perturbation term typically requires bounds on high order derivatives of the solution of (2.1) that may be difficult or impossible to obtain. This approach is also misleading because the accumulation of error is not determined solely by the stability properties of the true or approximate trajectories alone. Rather it is determined by the stability properties of the continuum of trajectories in a neighborhood containing both the true and approximate trajectories.

We define the coefficients for the dual problem by linearizing around an average

of the true and approximate solutions:

$$\begin{aligned}\bar{\epsilon}_i = \bar{\epsilon}_i(u,U) &= \int_0^1 \epsilon_i\big(us + U(1-s)\big)\,ds,\\ \bar{\beta}_{ij} = \bar{\beta}_{ij}(u,U) &= \int_0^1 \frac{\partial \epsilon_j}{\partial u_i}\big(us + U(1-s)\big)\nabla\big(u_j s + U_j(1-s)\big)\,ds,\\ \bar{f}_{ij} = \bar{f}_{ij}(u,U) &= \int_0^1 \frac{\partial f_j}{\partial u_i}\big(us + U(1-s)\big)\,ds.\end{aligned} \tag{2.10}$$

The regularity of u and U typically imply that $\bar{\epsilon}$ and $\bar{f}$ are piecewise continuous with respect to t and continuous, H^1 functions in space while $\bar{\beta}$ is discontinuous in time and space.

To derive the dual equation, we subtract the variational equation satisfied by the approximate solution given by (2.2) resp. (2.3) on I_n from the variational form of (2.1). The linearization we have chosen yields the following pair of equations:

$$\begin{aligned}&\sum_{i=1}^D \left(\epsilon_i(u)\nabla u_i \cdot \nabla\psi_i - \epsilon_i(U)\nabla U_i \cdot \nabla\psi_i\right)\\ &\qquad = \sum_{i=1}^D \left(\bar{\epsilon}_i\nabla(u_i - U_i) + \sum_{j=1}^D \bar{b}_{ij}(u_j - U_j)\right)\cdot\nabla\psi_i\end{aligned} \tag{2.11}$$

and

$$\sum_{i=1}^D \left(f_i(u)\psi_i - f_i(U)\psi_i\right) = \sum_{i,j=1}^D \bar{f}_{ij}(u_i - U_i)\psi_j\,, \tag{2.12}$$

which hold for any test function ψ. For example, (2.11) follows from the computation

$$\begin{aligned}&\sum_{i=1}^D \big(\epsilon_i(u)\nabla u_i - \epsilon_i(U)\nabla U_i\big)\cdot\nabla\psi_i\\ &\qquad = \sum_{i=1}^D \int_0^1 \frac{d}{ds}\epsilon_i(us + U(1-s))\nabla\big(u_i s + U_i(1-s)\big)\,ds\cdot\nabla\psi_i\\ &\qquad = \sum_{i=1}^D \int_0^1 \epsilon_i(us + U(1-s))\,ds\nabla(u_i - U_i)\cdot\nabla\psi_i\\ &\qquad\quad + \sum_{i,j=1}^D (u_j - U_j)\int_0^1 \frac{\partial\epsilon_i}{\partial u_j}(us + U(1-s))\nabla(u_i s + U_i(1-s))\,ds\cdot\nabla\psi_i.\end{aligned}$$

The terms on the left-hand sides of (2.11) and (2.12) occur in the difference of the two variational equations.

Written out pointwise for convenience, the *dual problem* to (2.1) associated to time node t_n is

$$\begin{cases} -\dot{\phi}_i - \nabla\cdot\big(\bar{\epsilon}_i\nabla\phi_i\big) + \sum_{j=1}^D \bar{\beta}_{ij}\cdot\nabla\phi_j = \sum_{j=1}^D \bar{f}_{ij}\phi_j\,, & (x,t)\in\Omega\times(t_n,0],\\ & 1\le i\le D,\\ \phi_i(x,t) = 0, & (x,t)\in\partial\Omega\times(t_n,0],\\ & 1\le i\le d,\\ \phi(x,t_n) = \phi_n(x), & x\in\Omega,\end{cases} \tag{2.13}$$

In the case of the scalar problem (2.4) with constant diffusion, the dual problem is

$$\begin{cases} -\dot{\phi} - \epsilon\Delta\phi = \bar{f}\phi, & (x,t) \in \Omega \times (t_n, 0], \\ \phi(x,t) = 0, & (x,t) \in \partial\Omega \times (t_n, 0], \\ \phi(x,t_n) = \phi_n(x), & x \in \Omega. \end{cases}$$

In the case of one parabolic equation with nonlinear diffusion coupled to one ordinary differential equation, the dual problem is

$$\begin{cases} -\dot{\phi}_1 - \nabla\cdot\bar{\epsilon}_1\nabla\phi_1 + \bar{\beta}_{11}\nabla\phi_1 = \bar{f}_{11}\phi_1 + \bar{f}_{12}\phi_2, & (x,t) \in \Omega \times (t_n, 0], \\ -\dot{\phi}_2 + \bar{\beta}_{21}\nabla\phi_1 = \bar{f}_{21}\phi_1 + \bar{f}_{22}\phi_2, & (x,t) \in \Omega \times (t_n, 0], \\ \phi_1(x,t) = 0, & (x,t) \in \partial\Omega \times (t_n, 0], \\ \phi(x,t_n) = \phi_n(x), & x \in \Omega. \end{cases}$$

The dual problem (2.13) is actually posed in variational form and we require the existence, uniqueness, and boundedness of the solution in appropriate Sobolev spaces. Observe that (2.13) is solved "backwards", i.e. from t_n, where the initial data is given, to 0 and the time derivative term is multiplied by -1 to compensate. Note also that $\bar{b}_{ij} = 0$ when $d < j \leq D$ so that the first d equations of (2.13) are parabolic partial differential equations while the remaining $D - d$ equations are ordinary.

With these definitions, we are in position to derive relationship between the error $e = u - U$ and the residual errors, which we carry out explicitly for the cG method. We use the L^2 projection operator into the piecewise polynomial functions in time, denoted by $\pi_n : L^2(I_n) \to \mathcal{P}^q(I_n)$, where $\mathcal{P}^q(I_n)$ is the space of polynomials of degree q or less defined on I_n. We note that the product $\pi_n P_n : L^2(S_n) \to W^r$ equals the L^2 projection onto W^r and that $\pi_n P_n = P_n \pi_n$. We define the global projection operator π by setting $\pi = \pi_n$ on S_n

Multiplying the dual problem (2.13) by e, integrating over $\Omega \times (0, t_n)$, and then integrating by parts in time, we obtain

$$\begin{aligned} 0 &= \sum_{i=1}^{D} \int_0^{t_n} \big(e_i, -\dot{\phi}_i - \nabla\cdot\bar{\epsilon}_i\nabla\phi_i + \sum_{j=1}^{D} \bar{\beta}_{ij}\cdot\nabla\phi_j - \sum_{j=1}^{D} \bar{f}_{ij}\phi_j\big)\, dt \\ &= \sum_{i=1}^{D} (e_i^+(0), \phi_i(0)) - \sum_{i=1}^{D} (e_i^-(t_n), \phi_i(t_n)) \\ &\quad - \sum_{i=1}^{D} \int_0^{t_n} \big((\dot{e}_i, \phi_i) + (\bar{\epsilon}_i\nabla e_i, \nabla\phi_i) + \sum_{j=1}^{D} (e_i\bar{\beta}_{ij}, \nabla\phi_j) - \sum_{j=1}^{D} (e_i\bar{f}_{ij}, \phi_j)\big)\, dt. \end{aligned}$$

Writing $\sum_{i,j} e_i\bar{\beta}_{ij}\cdot\nabla\phi_j = \sum_{i,j} e_j\bar{\beta}_{ji}\cdot\nabla\phi_i$ and using (2.11) and (2.12), the third term on the right-hand side simplifies to

$$\begin{aligned} &\sum_{i=1}^{D} \int_0^{t_n} \big((\dot{e}_i, \phi_i) + (\bar{\epsilon}_i\nabla e_i + \sum_{j=1}^{D} \bar{\beta}_{ji}e_j, \nabla\phi_i) - \sum_{j=1}^{D} (e_i\bar{f}_{ij}, \phi_j)\big)\, dt \\ &\qquad = \sum_{i=1}^{D} \int_0^{t_n} \Big((\dot{u}_i - \nabla\cdot\epsilon_i(u)\nabla u_i - f_i(u), \phi_i) \\ &\qquad\qquad - \big((\dot{U}_i, \phi_i) + (\epsilon_i(U)\nabla U_i, \nabla\phi_i) - (f_i(U), \phi_i)\big)\Big) dt. \end{aligned}$$

This expression simplifies further because u solves (2.1) and abusing notation to let $(v,w) = \sum_{i=1}^{D}(v_i,w_i)$ for $D-$vector valued functions v and w, we obtain

$$(e^-(t_n),\phi_n) = (e^+(0),\phi(0)) - \int_0^{t_n}\big((\dot{U},\phi) + (\epsilon(U)\nabla U,\nabla\phi) - (f(U),\phi)\big)\,dt. \tag{2.14}$$

U also nearly solves (2.1), as expressed in the Galerkin orthogonality relation (2.2). We can therefore insert the interpolant $\pi P\phi$ of ϕ in W^q to obtain the *error representation formula*:

$$\begin{aligned}(e^-(t_n),\phi_n) &= (e^+(0),\phi(0))\\ &\quad + \int_0^{t_n}\big((\dot{U},\pi P\phi-\phi) + (\epsilon(U)\nabla U,\nabla(\pi P\phi-\phi)) - (f(U),\pi P\phi-\phi)\big)\,dt.\end{aligned} \tag{2.15}$$

A similar analysis for the dG method gives

$$\begin{aligned}(e^-(t_n),\phi_n) &= (e^-(0),\phi(0)) + \sum_{j=1}^{n}\big([U]_{j-1},(\pi P\phi-\phi)^+_{j-1}\big)\\ &\quad + \int_0^{t_n}\big((\dot{U},\pi P\phi-\phi) + (\epsilon(U)\nabla U,\nabla(\pi P\phi-\phi)) - (f(U),\pi P\phi-\phi)\big)\,dt.\end{aligned} \tag{2.16}$$

The first term on the right in these formulas describes the propagation of the initial error $e^-(0)$ to time t_n by scaling the initial error by the size of the dual solution over that interval. The remaining terms on the right describe the cumulative effect of errors made in solving the differential equation approximately. This part of the formulation has the same form as the variational equation defining the finite element method, except that the test function $\pi P\phi - \phi$ is not in the finite element space. The test function is small however when the dual solution has sufficient regularity since it is just the interpolation error of ϕ in the finite element space. Thus, the size of the error is determined both by the size of the residual errors and the Galerkin orthogonality of the approximation.

2.5. The stability factors and the a posteriori error estimate. To derive the a posteriori error estimate, we split the integrals on the right in (2.15) resp. (2.16) to obtain expressions involving the residual errors we defined above, then take norms and estimate. If

$$\phi \in L_\infty((0,t_n);L_2(\Omega)),\ D_t^\alpha\phi \in L_1((0,t_n);L_2(\Omega)),\ \text{and } D^2\phi^p \in L_1((0,t_n);L_2(\Omega)), \tag{2.17}$$

where $0 \le \alpha \le 1$ for the cG(1) and dG(0) methods and $0 \le \alpha \le 2$ for the dG(1) method, then we can take optimal interpolation estimates on $\pi P\phi - \phi$, and this leads to the following definitions of the *stability factors* that scale the various residual errors. First, there is the stability factor associated to the propagation of the initial error:

$$S_0(0,t_n) = \|\phi(0)\|. \tag{2.18}$$

The stability factor associated with time discretization by means of the $cG(q)$ or $dG(q-1)$ method is defined by

$$S_t^\alpha(0,t_n) = C_t^\alpha\int_0^{t_n}\|D_t^\alpha\phi\|\,dt, \quad 0 \le \alpha \le q, \tag{2.19}$$

where C_t^α is the interpolation constant in the L_1 error bound for the L_2 projection into the space of scalar polynomials of degree α (see (6.2)). In order to define the stability factors associated to space discretization, we denote the part of ϕ associated to the parabolic and ordinary differential equations by ϕ^p and ϕ^o respectively. Then,

$$S_x^p(0,t_n) = C_x^p \int_0^{t_n} \|D^2\phi^p\|\,dt \text{ and } S_x^o(0,t_n) = \int_0^{t_n} \|\phi^o\|\,dt, \tag{2.20}$$

where C_x^p is the standard interpolation constant for the L_2 error bound for the L_2 projection into the space of continuous piecewise linear functions V_n (see (6.4)).

The respective values of the stability factors depends on the choice of ϕ_n of course. We discuss the choice of data in Section 4. As a minimum requirement to guarantee that the stability factors are finite, we typically restrict ϕ_n to be a function in $H_0^1(\Omega))^d \times (H^1(\Omega))^{D-d}$.

To illustrate, we compute the stability factors for the heat equation $u_t - u_{xx} = 0$ posed on the interval $[0,1]$ with Dirichlet boundary conditions. Later in the paper, we compute approximate stability factors for various nonlinear problems. The dual problem to the heat equation at t_n is found to be the heat equation itself after the change of variables $t \to t_n - t$. If we set $\phi_n = \sum_{i\geq 1} \phi_{n,i} \sin(i\pi x)$ then

$$S_0(0,t_n) = \left(\sum_{i\geq 1} \frac{1}{2} \phi_{n,i}^2 e^{-2\pi^2 i^2 t_n} \right)^{1/2}$$

$$(C_t^1)^{-1} S_t^1(0,t_n) = (C_x^p)^{-1} S_x^p(0,t_n) = \int_0^{t_n} \left(\sum_{i\geq 1} \frac{1}{2} \pi^4 i^4 \phi_{n,i}^2 e^{-2\pi^2 i^2 s} \right)^{1/2} ds.$$

We plot these functions versus t_n in Fig. 2.2 for a generic choice of ϕ_n in H_0^1. From

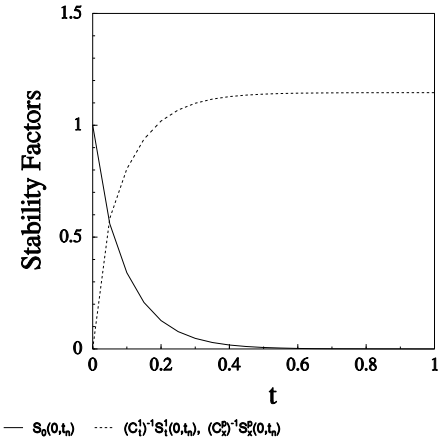

FIG. 2.2. *Plot of $S_0(0,t_n)$ and $(C_t^1)^{-1} S_t^1(0,t_n) = (C_x^p)^{-1} S_x^p(0,t_n)$ versus t_n for the heat equation with a suitable H_0^1 initial function given for the dual problem.*

the plot, we see that $S_0(0,t_n)$ decays exponentially to zero as $t_n \to \infty$, as expected for the heat equation. The other stability factors tend exponentially to a constant value ≈ 1.146, indicating that there is essentially no accumulation of discretization

errors after sufficient time has passed. Again, this is expected of implicit methods for the heat equation.

We prove the following estimate on the error at time nodes in Section 6. We let $\tilde{P}_n$ denote an interpolation operator into V_n and $\tilde{P}$ denote the associated global interpolation operator defined so $\tilde{P} = \tilde{P}_n$ on S_n. For example, $\tilde{P}_n = P_n$ or Q_n, the *nodal interpolation operator*. In the statement of the theorem, $q = 1$ for the cG method and $q = 1$ or 2 for the dG method.

THEOREM 2.1. *For $1 \leq \alpha \leq q$, the error of the $cG(q)$ or $dG(q-1)$ approximation at time t_n, $1 \leq n$, satisfies*

$$\begin{aligned} |(e^-(t_n),\phi_n)| \leq\ & S_0(0,t_n)\|e^-(0)\| + S_t^\alpha(0,t_n)\|k^\alpha R_t(U)\|_{L_\infty(0,t_n)} \\ & + S_x^p(0,t_n)\big(\|h^2 R_x^p(U)\|_{L_\infty(0,t_n)} + \|h^2 R_2^p(U)\|_{L_\infty(0,t_n)}\big) \\ & + S_x^o(0,t_n)\|(I-\tilde{P})R_x^o(U)\|_{L_\infty(0,t_n)}. \end{aligned} \tag{2.21}$$

Remark 2.3. The estimate on the projection of $e^-(t_n)$ in the direction of ϕ_n (2.21) turns into a estimate on $\|e^-(t_n)\|$ if we choose ϕ_n so that $(e^-(t_n),\phi_n) = \|e^-(t_n)\|$. For example, we could choose $\phi_n = e^-(t_n)/\|e^-(t_n)\|$. The difficulty is that the stability factors then depend on the unknown error at time t_n that we are trying to estimate. With this choice, therefore, we need to obtain estimates on the stability factors that are independent of the initial data for the dual problem.

Remark 2.4. Note that the different residual errors are scaled by different stability factors, giving the potential for the different residual errors to accumulate at different rates. This is an important consideration when attempting to estimate the error accurately.

Remark 2.5. We have posed the problem (2.1) beginning at time $t = 0$ and estimated the error over $[0,t_n]$. The results extend to problems beginning at some time t_{n-1} and continuing to some time t_m, $m \geq n$, with the obvious change in notation, e.g. $S_0(0,t_n) \to S_0(t_{n-1},t_m)$, etc.

In case the solution of the dual problem does not have sufficient regularity to admit optimal order interpolation estimates, we use weaker stability factors. For example, we can define

$$S_x^p(0,t_n) = C_x^p \int_0^{t_n} \|\nabla\phi^p\|\,dt,$$

with the appropriate C_x^p, and then a straightforward alteration of the proof of Theorem 2.1 shows

$$\begin{aligned} |(e^-(t_n),\phi_n)| \leq\ & S_0(0,t_n)\|e^-(0)\| + S_t^\alpha(0,t_n)\|k^\alpha R_t(U)\|_{L_\infty(0,t_n)} \\ & + S_x^p(0,t_n)\big(\|hR_x^p(U)\|_{L_\infty(0,t_n)} + \|hR_2^p(U)\|_{L_\infty(0,t_n)}\big) \\ & + S_x^o(0,t_n)\|(I-\tilde{P})R_x^o(U)\|_{L_\infty(0,t_n)}. \end{aligned} \tag{2.22}$$

A weaker stability factor in time can be introduced similarly.

An error estimate at the time nodes for the finite element approximations corresponds to interpreting the methods as finite difference schemes and is usually the main focus of interest. However, we can also prove an estimate for the error inside the intervals with a modification of the proof of Theorem 2.1. The loss of optimality in the term involving the time residual is due to the fact that Galerkin orthogonality

does not hold over intervals that do not end at a time node. We prove the result in Section 6.

THEOREM 2.2. *The error at time $t_{n-1} < t^* \leq t_n$ satisfies*

$$\begin{aligned}|(e(t^*),\phi(t^*))| \leq\ & |(e_n^-,\phi_n)| + S_0(t^*,t_n)k_n\|R_t(U)\|_{L_\infty(t^*,t_n)} \\ & + S_x^p(t^*,t_n)\big(\|h^2R_x^p(U)\|_{L_\infty(0,t_n)} + \|h^2R_2^p(U)\|_{L_\infty(0,t_n)}\big) \\ & + S_x^o(t^*,t_n)\|(I-\tilde{P})R_x^o(U)\|_{L_\infty(0,t_n)}.\end{aligned} \tag{2.23}$$

In the proof of Theorem 2.1, it is clear that different norms can be taken on the residual errors and the stability factors. For example, the estimate can be viewed as a "weighted" norm of the residual errors, with the weights determined by the interpolation error of the solution of the dual problem. Another possibility is to avoid taking norms on the right-hand side of (2.15) altogether. This would allow full reign to the effects of cancellation of error and possibly reduce the tendency to over-estimation. There are some practical difficulties with this approach however. First without clearly defined residual errors, it becomes difficult to decide how to refine the mesh in order to compute an approximation with greater accuracy given an approximation with a large error estimate. Second, it becomes difficult to analyze the a posteriori estimate: for example, to prove that the error decreases as the residual errors decrease, as we show in this paper. Third, there is question of reliability. In practice, we can only approximate the solution of the dual problem for a small set of data. Likewise, we can only compute the residual errors at discrete times and approximate the integrals on the right-hand side of (2.15) using a quadrature. The effect this has on the estimate, and in particular on the tendency to chronic underestimation of the error, is difficult to determine. Our computational experience with reaction-diffusion problems suggests that taking norms inside the integrals on the right-hand side of (2.15) does not lead to severe overestimation of the error most of the time. See Estep and French [29] and Larson [48] for an analysis of a Hamiltonian system in which using norms in the error representation does lead to chronic overestimation.

We can also obtain an estimate for different norms of the error by altering the dual problem. For example, we can obtain a $L_1(L_2)$ estimate by posing the dual problem

$$\begin{cases} -\dot{\phi}_i - \nabla\cdot\left(\bar{\epsilon}_i\nabla\phi_i\right) + \sum_{j=1}^D \bar{\beta}_{ij}\cdot\nabla\phi_j - \sum_{j=1}^D \bar{f}_{ij}\phi_j = \psi_i, & (x,t)\in\Omega\times(t_n,0], \\ & 1\leq i\leq D, \\ \phi_i(x,t) = 0, & (x,t)\in\partial\Omega\times(t_n,0], \\ & 1\leq i\leq d, \\ \phi(x,t_n) = 0, & x\in\Omega, \end{cases}$$

where the coefficients are the same as for (2.13). We derive an error representation as above, obtaining

$$\begin{aligned}\int_0^{t_n}(e,\psi)\,dt =\ & (e^+(0),\phi(0)) \\ & + \int_0^{t_n}\left((\dot{U},\pi P\phi-\phi) + (\epsilon(U)\nabla U,\nabla(\pi P\phi-\phi)) - (f(U),\pi P\phi-\phi)\right)dt.\end{aligned}$$

for the cG method for example. We define

$$S_{t,2}^\alpha(0,t_n) = C_t^\alpha\,\|D_t^\alpha\phi\|_{L_2(0,t_n)},\quad 0\leq\alpha\leq q,$$
$$S_{x,2}^p(0,t_n) = C_x^p\,\|D^2\phi^p\|_{L_2(0,t_n)} \text{ and } S_{x,2}^o(0,t_n) = \|\phi^o\|_{L_2(0,t_n)},$$

for appropriate C_t^α and C_x^p. Estimating as in the proof of Theorem 2.1, we obtain

$$\begin{aligned}\left|\int_0^{t_n}(e^-(t_n),\psi)\,dt\right| &\leq S_0(0,t_n)\|e^-(0)\| + S_{t,2}^\alpha(0,t_n)\|k^\alpha R_t(U)\|_{L_2(0,t_n)} \\ &\quad + S_{x,2}^p(0,t_n)\big(\|h^2R_x^p(U)\|_{L_2(0,t_n)} + \|h^2R_2^p(U)\|_{L_2(0,t_n)}\big) \\ &\quad + S_{x,2}^o(0,t_n)\|(I-\tilde{P})R_x^o(U)\|_{L_2(0,t_n)}.\end{aligned} \tag{2.24}$$

It is often the case that on theoretical and practical grounds, the error must be controlled at all intermediate times even if the results are desired only at some (final) time node t_n. In this context, a disadvantage to the pointwise error estimate (2.21) is that there is a *different* dual problem associated to each time node. Estimating the error at every time node thus requires an estimation of the stability factors associated to each time node. A better approach is to keep control of the error at all times in the L_1 sense via (2.24) and pointwise only at specific interesting times via (2.21). For an example of L_1 control of the error in time, see Estep, Hodges, and Warner [30].

3. Interpreting the a posteriori error bound.

In this section, we discuss the meaning of the Theorem 2.1. This is necessary because without further analysis, the sizes of the residual errors and stability factors in the a posteriori error estimate are unknown. For example, if the approximation method is unstable then the residual errors can grow without bound as the mesh and time steps are refined. Likewise, if the solution of the dual problem (2.13) does not have sufficient regularity, then the stability factors would be infinite. Recall that a classic a priori error bound is derived by establishing the consistency and stability of the numerical scheme after assuming the well-posedness of the differential equation and sufficient regularity of the solution. So far, we have not mentioned any of these properties in the a posteriori analysis.

As a first goal, we show that the a posteriori error estimate is *theoretically* meaningful in the sense that the quantities on the right-hand side of the estimate (2.21) are finite and moreover that the residual errors on any interval can be made arbitrarily small by refining the space mesh and time steps. To do this, we estimate the sizes of the residual errors and stability factors after postulating a set of *local* a priori properties of the continuous problem (2.1), the dual problem (2.13), and the numerical methods.

But this goal is not our main purpose. An important reason is that bounds on the stability factors grow exponentially with time in general, for much the same reason that general a priori error bounds grow exponentially. In practice, this appears to over-estimate the effects of accumulation of error to the point of making the bound uselessly inaccurate after a short time. Therefore, our second goal is to show that the a posteriori error estimate is *practically* meaningful in the sense that the stability factors and residual errors can be computed or estimated computationally.

Coincidentally, both goals are achieved more or less with the same analysis simultaneously. The estimates on the residual errors that indicate the rate they tend to zero as the discretization is refined are used to indicate a strategy for refining a given mesh in order to compute an approximate solution with a desired residual error. Likewise, the analysis that shows the stability factors are bounded also show that the solution of the dual problem can be approximated accurately using a finite element method once the data and coefficients are specified. (There remain some important

issues for computing the stability factors involving the choice of data and coefficients that we discuss in Section 4.)

The a priori properties of the continuous problem and the numerical methods used in the following analysis are typical examples of the kind of results sought after in standard analysis of (2.1) and roughly speaking are the necessary ingredients to show the methods converge. It appears likely to be difficult to establish these properties for the general problem (2.1), and in fact, our goals could be achieved by making different assumptions and estimating the residual errors and stability factors differently. For example in several places, we indicate how weaker assumptions can be used. But, the motivation for the assumptions in this paper is that they can be shown to hold for many specific problems and classes of problems. We discuss an important class that contains examples 1–9 in Section 3.3.

Remark 3.1. We emphasize that an important goal in the subsequent analysis is to use minimal regularity of solutions. As mentioned above, this is partly due to the fact that the regularity of solutions of reaction-diffusion equations is often unknown, and in any case, we can not expect to find globally smooth high order derivatives in general. There is also a practical reason. As much as possible, we want to choose mesh and step sizes based on the criteria of controlling the approximation error rather than choosing mesh and step sizes in order to fulfill requirements of the theory for the estimation of the error. Strong requirements on the mesh and step sizes is a characteristic of some other approaches to error estimation, like those based on high order asymptotic estimates of the error which require refining the space mesh and time steps relative to the size of higher order derivatives of solutions than appear in the actual error estimate.

In this section, we make additional qualitative assumptions on the meshes and time steps. First off, we assume the meshes are quasi-uniform in the sense that there is a constant λ_1 independent of the triangulation $\mathcal{T}_n$ such that $h_{n,\max} \leq \lambda_1 h_{n,\min}$. In addition, we assume that mesh refinement is performed so that the meshes are nested, i.e. T_n is obtained from T_{n-1} either by refinement or coarsening, where in the case of refinement, $V_{n-1} \subset V_n$, and conversely in the case of coarsening, $V_n \subset V_{n-1}$. We discuss the construction of meshes satisfying these assumptions in Section 4. It seems likely that for some problems, the results could be extended to allow "locally" nested meshes under a less stringent quasi-uniformity assumption, i.e. so both refinement and coarsening can be used in each time step, but this would at the least complicate the notation and analysis.

3.1. The size of the residual errors. The basic idea is to estimate the residual errors in terms of the error itself using the fact that the equation for the finite element approximation is an approximation of the continuous differential equation. Quantifying this approximation property is analogous to showing that the method is consistent and stable over one step in the classic a priori analysis. The first result says that the residual errors tend to zero in the limit of discretization if the error of the method over one step tends to zero. In practice, it is also necessary to have a quantitative estimate on the size of the residual errors. For example, we require such information in proving that the a priori assumptions hold for the class of problems considered in Section 3.3 and it is useful for deriving a mesh and time step refinement strategy. The second result contains precise estimates on the residual errors assuming a classic style a priori error bound for the finite element approximation and a set of energy estimates bounding derivatives of a solution of the equation in (2.1) in terms

of derivatives of the data hold over short time intervals.

A serious difficulty arises if we try to estimate the residual errors by comparing the finite element approximation to the solution of the differential equation (2.1) over even moderately long time intervals. Namely, the error may grow rapidly because of accumulation and in any case classic a priori error bounds almost certainly do so. On the other hand, the size of the residual error is a local property in the sense that on a given interval it is determined by the difficulty of approximating nearby solutions on that interval. To exhibit this local nature, we fix the value of the approximate solution U_{n-1}^- at time t_{n-1} and then estimate the residual errors of the finite element solution of (2.1) over the next time step $[t_{n-1}, t_n]$ by comparing U to a local solution $\tilde{u}$ of the differential equation in (2.1) over the interval that begins with "initial" data $\tilde{u}_{n-1}$ at t_{n-1} that is close to U_{n-1}^-. In this way, a priori convergence results are used only over one time step and the problem of the potentially catastrophic loss of accuracy over longer times is avoided.

As initial data, U_{n-1}^- is not sufficiently smooth to expect that the corresponding local solution will have the regularity required in the analysis. So, we form the initial data by "smoothing" U_{n-1}^-. We let $\tilde{u}$ denote the solution of the *local problem*:

$$\begin{cases} \dot{\tilde{u}}_i - \nabla \cdot (\epsilon_i(\tilde{u}, x, t)\nabla \tilde{u}_i) = f_i(\tilde{u}, x, t), & (x,t) \in \Omega \times (t_{n-1}, t_n],\ 1 \le i \le D, \\ \tilde{u}_i(x,t) = 0, & (x,t) \in \partial\Omega \times (t_{n-1}, t_n],\ 1 \le i \le d, \\ \tilde{u}(x, t_{n-1}) = \tilde{u}_{n-1} = T\Delta_{h_n} U_{n-1}^-(x), & x \in \Omega, \end{cases} \tag{3.1}$$

where T denotes the solution operator associated to the Dirichlet problem for the Laplacian on Ω, i.e. Tg solves $-\Delta Tg = g$ on Ω with Dirichlet boundary conditions given by U_{n-1}^- on $\partial\Omega$. We illustrate the smoothing in Fig. 3.1.

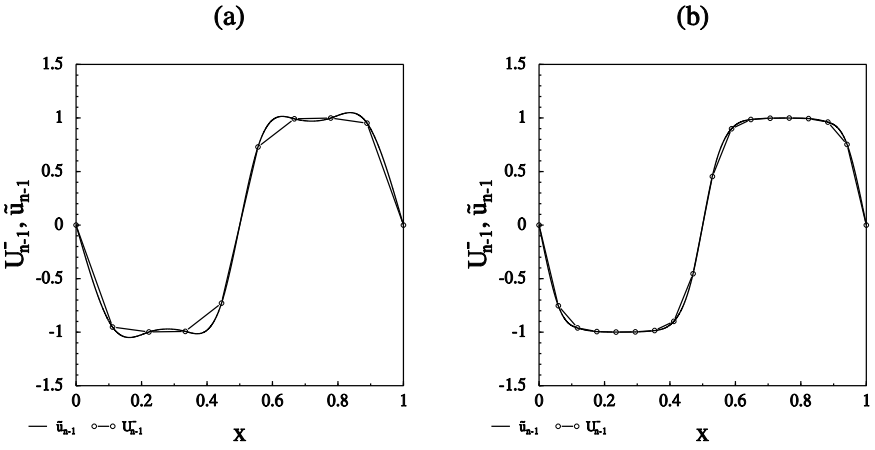

FIG. 3.1. *Plots of $\tilde{u}_{n-1}$ and U_{n-1}^- where U_{n-1}^- approximates a metastable pattern for the bistable equation with $\epsilon = .0009$ using (a) 8 and (b) 16 uniformly spaced mesh points.*

The following lemma, proved in Section 6, implies that $\tilde{u}$ has sufficient regularity for the analysis and gives a bound on the initial error.

LEMMA 3.1. *Under the mesh assumptions above, the initial data $\tilde{u}_{n-1} \in H^2(\Omega) \cap H^1(\bar{\Omega})$ and there is a constant C depending on λ_i such that*

$$\|\tilde{u}_{n-1} - U_{n-1}^-\| \le C\|h_n^2 \Delta_{h_n} U_{n-1}^-\| = \mathbf{O}(h_{n,\max}^{3/2}).$$

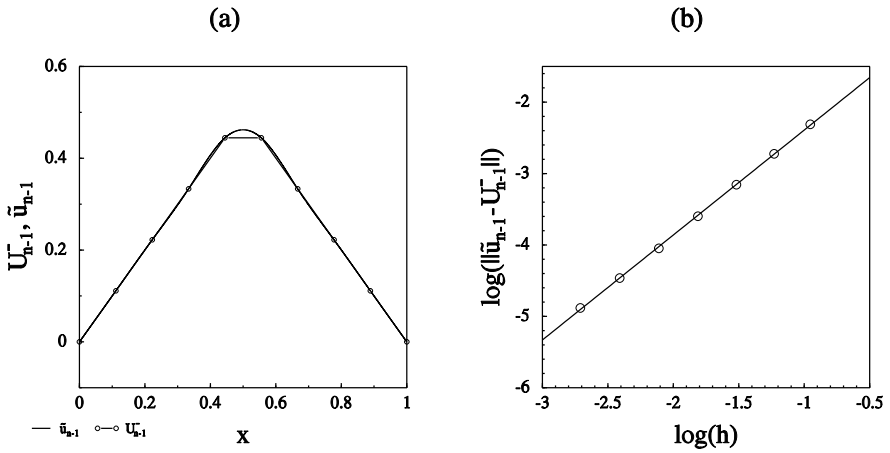

FIG. 3.2. *(a) Plot of $\tilde{u}_{n-1}$ and U^-_{n-1}, where U^-_{n-1} interpolates $g(x)$ which equals x for $0 \le x \le .5$ and $1-x$ for $.5 < x \le 1$ using 8 uniformly spaced mesh points. (b) Plot of a least squares line fit through the data pairs $(\log(h), \log(\|\tilde{u}_{n-1} - U^-_{n-1}\|))$ where U^-_{n-1} interpolates $g(x)$ using uniformly spaced meshes. The least squares line has slope 1.5 with correlation $\rho^2 = .9997$.*

We show an example in which this order of convergence is reached in Fig. 3.2. The implicit constant in the '**O**' in this result and those following is independent of the local solution $\tilde{u}$ and the mesh and time steps except for the λ_i. The assumption of nested meshes is important for this result.

The first result (stated in a form to be valid for both the cG and dG methods) bounds each of the residual errors in terms of the error $\tilde{e} = \tilde{u} - U$ scaled by factors depending on derivatives of $\tilde{u}$ and the size of U.

THEOREM 3.2. *Assume that $\tilde{u} \in L_\infty(I_n; H^1(\Omega))$, $\dot{\tilde{u}} \in L_\infty(I_n; L_2(\Omega))$, and $\tilde{u}^p \in L_\infty(I_n; H^2(\Omega))$. Then there is a constant C depending on λ_i, ϵ, and f such that for $1 \le i \le d$,*

$$\begin{aligned}\|k_n R_t(U)_i\|_{L_\infty(I_n)} \le C\big(&\|\tilde{e}\|_{L_\infty(I_n)} + k_n\|h_n^{-2}\tilde{e}\|_{L_\infty(I_n)} \\ &+ k_n\|h_n^{-1}\tilde{e}\|_{L_\infty(I_n)}\|\nabla\tilde{u}_i\|_{L_\infty(I_n)} + k_n\|\Delta\tilde{u}_i\|_{L_\infty(I_n)} \\ &+ k_n\|\nabla\tilde{u}\|^2_{L_\infty(I_n)} + k_n\|\dot{\tilde{u}}_i\|_{L_\infty(I_n)} + k_n\|\Delta_{h_n}U^-_{n-1,i}\|\big),\end{aligned} \tag{3.2}$$

and for $d < i \le D$,

$$\|k_n R_t(U)_i\|_{L_\infty(I_n)} \le C\big(\|\tilde{e}\|_{L_\infty(I_n)} + k_n\|\dot{\tilde{u}}_i\|_{L_\infty(I_n)} + k_n\|\Delta_{h_n}U^-_{n-1,i}\|\big), \tag{3.3}$$

while

$$\begin{aligned}\|h_n^2 R_x^p(U)\|_{L_\infty(I_n)} \le C\big(&\|\tilde{e}\|_{L_\infty(I_n)}(1 + \|h_n\nabla\tilde{u}\|^2_{L_\infty(I_n)} + \|U\|^2_{L_\infty(I_n)}) \\ &+ \|h_n\nabla\tilde{e}\|_{L_\infty(I_n)}(\|h_n\nabla\tilde{u}\|_{L_\infty(I_n)} + \|U^p\|_{L_\infty(I_n)}) \\ &+ k_n\|\dot{\tilde{u}}^p\|_{L_\infty(I_n)} + \|h_n^2\Delta\tilde{u}^p\|_{L_\infty(I_n)}\big),\end{aligned} \tag{3.4}$$

$$\|h_n^2 R_2^p(U)\|_{L_\infty(I_n)} \le C\big(\|h_n\nabla\tilde{e}^p\|_{L_\infty(I_n)} + \|h_n^2\Delta\tilde{u}^p\|_{L_\infty(I_n)}\big), \tag{3.5}$$

and finally with $\tilde{P} = P$,

$$\|(I - \tilde{P})R_x^o(U)\|_{L_\infty(I_n)} \le C(\|h_n\nabla\tilde{u}\|_{L_\infty(I_n)} + \|\tilde{e}\|_{L_\infty(I_n)}), \tag{3.6}$$

while with $\tilde{P} = Q$,

$$\|(I-\tilde{P})R_x^o(U)\|_{L_\infty(I_n)} \leq C\left(\sum_{K\in\mathcal{T}_n} \|h_n^2(K)D^2f^o(U)\|^2_{L_\infty(I_n;L_2(K))}\right)^{1/2}. \tag{3.7}$$

If in addition $\tilde{u} \in L_\infty(I_n;H^2(\Omega))$ *then*

$$\|(I-\tilde{P})R_x^o(U)\|_{L_\infty(I_n)} \leq C\big(\|h_n\nabla\tilde{u}\|^2_{L_\infty(I_n)} + \|h_n^2\Delta\tilde{u}\|_{L_\infty(I_n)} + \|\tilde{e}\|_{L_\infty(I_n)}\big). \tag{3.8}$$

The proof, given in Section 6, uses straightforward but tedious estimates based on the fact that the residual error of the true solution of (3.1) is zero, a trace inequality, and an inverse estimate. The latter two ingredients are the reasons for the stricter mesh assumptions.

The constant C depends on the Lipschitz constants and the sizes of ϵ, $\nabla_u\epsilon$, f, and $\nabla_u f$ and the size of the second derivatives of f in a region containing both U and $\tilde{u}$ over I_n. If we assume a uniform bound on these quantities then C is truly constant over time. This turns out to be justified in the special case analyzed in Section 5. Otherwise, the value of C could vary from one interval to the next.

This analysis and the estimates on the residuals simplify considerably in the case of constant diffusion:

COROLLARY 3.3. *If* ϵ_i *is constant for* $1 \leq i \leq d$, *then*

$$\begin{aligned}\|k_nR_t(U)_i\|_{L_\infty(I_n)} \leq C\big(&\|\tilde{e}\|_{L_\infty(I_n)} + k_n\|h_n^{-2}\tilde{e}\|_{L_\infty(I_n)} + k_n\|\dot{\tilde{u}}_i\|_{L_\infty(I_n)}\\ &+ k_n\|\Delta\tilde{u}_i\|_{L_\infty(I_n)} + k_n\|\Delta_{h_n}U^-_{n-1,i}\|\big), \quad 1\leq i\leq d,\end{aligned} \tag{3.9}$$

and

$$\|h_n^2R_x^p(U)\|_{L_\infty(I_n)} \leq C\big(\|\tilde{e}\|_{L_\infty(I_n)} + k_n\|\dot{\tilde{u}}^p\|_{L_\infty(I_n)} + \|h_n^2\Delta\tilde{u}^p\|_{L_\infty(I_n)}\big), \tag{3.10}$$

while the other estimates in Theorem 3.2 remain the same.

In particular, Theorem 3.2 implies that the residual errors of an approximate solution of a problem with sufficiently smooth solutions computed using a consistent and stable scheme can be small even when the error is large. Recall that an analogous result holds for numerical solutions of linear algebraic systems. This is a rather startlingly consequence, which we illustrate with the Lorenz system:

$$\begin{cases}\dot{x} = -10x + 10y,\\ \dot{y} = 28x - y - xz,\\ \dot{z} = -\frac{8}{3}z + xy,\end{cases} \tag{3.11}$$

where we have chosen parameters that are believed to lead to chaotic behavior. Chaos is perhaps difficult to define, but certainly for this problem, the error of any numerical solution grows as time passes and eventually any numerical solution becomes inaccurate. We plot two numerical approximations of the same solution in Fig. 3.3. Following Theorem 3.2, we can compute a numerical solution while keeping the residual error R_t over each step below a given residual error tolerance. The a posteriori error estimate implies that as long as the stability factor is bounded on a given interval, we can decrease the error on the interval by decreasing the residual error tolerance. In Fig 3.3

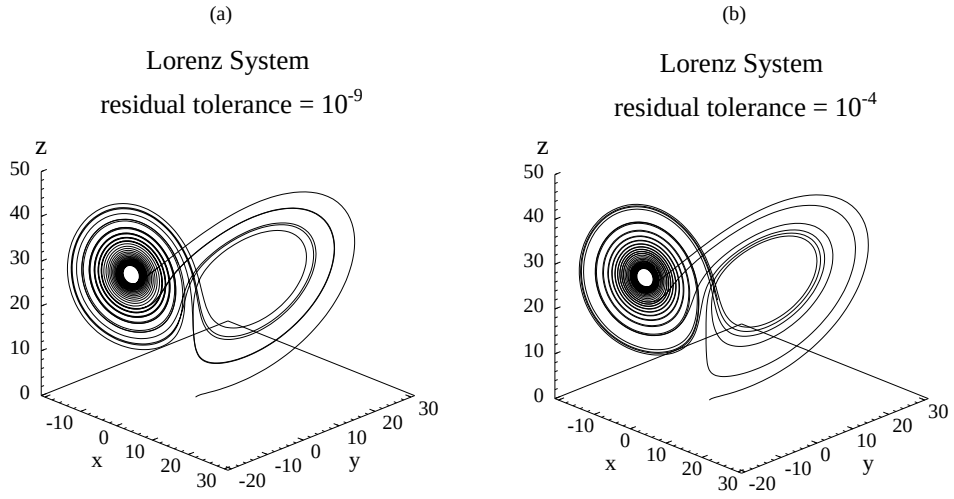

FIG. 3.3. *Two numerical approximations of the trajectory of the Lorenz system corresponding to initial data* $(1,0,0)$. *The numerical solution on the left is computed by keeping the residual errors below* 10^{-9} *on every step and is accurate to within .5%. The numerical solution on the right is computed by keeping the residual errors below* 10^{-4} *on every step and first becomes grossly inaccurate around* $t = 17.8$.

(a), we show a numerical solution computed with residual error tolerance 10^{-9} that is accurate to within 5% on $[0,30]$. In (b), we plot a numerical solution computed with residual error tolerance of 10^{-5} that first becomes grossly inaccurate at $t \approx 17.8$ and remains very inaccurate after that. An interesting question is why exactly does the trajectory in (b) become very inaccurate suddenly? To show that this question has a wider scope, we plot the x component of numerical solutions computed with increasing accuracy, i.e. decreasing residual error tolerances, in Fig. 3.4 (a). As the residual error tolerance decreases, the corresponding numerical solution remains accurate for a longer time as expected. It is interesting to note that all the numerical solutions become grossly inaccurate for the first time in the same region of phase space. This suggests that there is one mechanism behind the sudden decrease in accuracy. This is not due to the residual error becoming large suddenly however. We plot the residual error R_t of the numerical solution shown in Fig. 3.3 versus time in Fig. 3.4 (b). The residual error does not become large near $t \approx 17.8$ even though the error suddenly increases there. We investigate this further below. See Estep and Johnson [31] for further numerical analysis of the Lorenz system.

Because the data for the local solution of (3.1) depends explicitly on the mesh for the n'th interval, we have to determine the dependence of $\tilde{u}$ and its derivatives on h_n in order to show that the residual errors actually tend to zero as the mesh size and time step are refined. We do this in the next theorem by assuming more information about the rate of growth of derivatives of $\tilde{u}$, as measured in terms of energy estimates, and the rate of convergence of the numerical method, as measured by an a priori error bound. These assumptions are suggested by well-known properties of parabolic equations and we prove that they hold for the class of problems considered in Section 3.3. The proof of the following theorem is presented in Section 6.

THEOREM 3.4. *Assume that* $\tilde{u} \in L_\infty(I_n; H^1(\Omega))$, $\dot{\tilde{u}} \in L_\infty(I_n; L_2(\Omega))$, *and* $\tilde{u} \in$

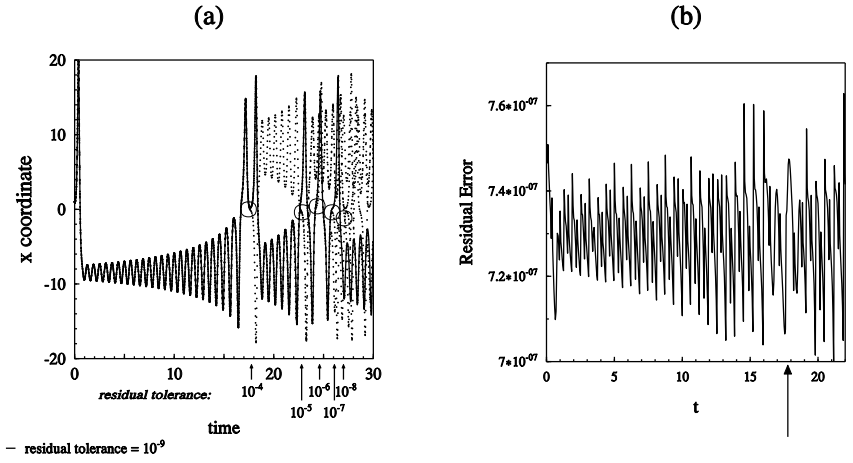

FIG. 3.4. *(a) Plots of numerical approximations of the x component of the Lorenz system computed with decreasing residual error tolerances versus time. As the tolerance decreases, the computations remain accurate for longer times. All the computations become grossly inaccurate for the first time in the same region of phase space near the z axis. (b) Plot of the residual error R_t of the numerical trajectory shown in Fig. 3.3 versus time. The residual error is not particularly large at the point where the solution becomes grossly inaccurate.*

$L_\infty(I_n; H^2(\Omega))$ and that there is a constant C depending on λ_i, ϵ, and f such that

$$\|\nabla\tilde{u}\|_{L\infty(I_n)} \leq Ce^{Ck_n}\big(\|\nabla\tilde{u}_{n-1}\| + 1\big) \tag{3.12}$$

$$\|\Delta\tilde{u}\|_{L\infty(I_n)} \leq Ce^{Ck_n}\big(\|\Delta\tilde{u}_{n-1}\| + \|\nabla\tilde{u}_{n-1}\| + 1\big) \tag{3.13}$$

$$\|\dot{\tilde{u}}^p\|_{L\infty(I_n)} \leq Ce^{Ck_n}\big(\|\Delta\tilde{u}_{n-1}\| + \|\nabla\tilde{u}_{n-1}\| + 1\big) \tag{3.14}$$

$$\|\dot{\tilde{u}}^o\|_{L\infty(I_n)} \leq C. \tag{3.15}$$

In addition, assume that the numerical approximation satisfies the energy estimates

$$\|U^p\|_{L\infty(I_n)} \leq C\big(\|U^{p,-}_{n-1}\| + k_n\|\nabla U^{p,-}_{n-1}\| + k_n\big), \tag{3.16}$$

$$\|U^o\|_{L\infty(I_n)} \leq C\big(\|U^{o,-}_{n-1}\| + k_n\big), \tag{3.17}$$

and the a priori error bound

$$\|\tilde{e}\|_{L\infty(I_n)} \leq C\|h_n^2\Delta\tilde{u}_{n-1}\| + Ce^{Ck_n}\big(k_n\|\dot{\tilde{u}}\|_{L\infty(I_n)} + \|h_n^2\Delta\tilde{u}\|_{L\infty(I_n)}\big). \tag{3.18}$$

Then with $\tilde{P} = P$,

$$\begin{aligned}
\|k_n R_t^p(U)\|_{L_\infty(I_n)} &\le Ce^{Ck_n}\big(\|h_n^2\Delta_{h_n}U_{n-1}^-\| + \|h_n^2\nabla U_{n-1}^-\| + h_{n,\max}^2 \\
&\quad + k_n^2\big(h_{n,\min}^{-2} + \|h_n^{-2}\Delta_{h_n}U_{n-1}^-\| + \|h_n^{-2}\nabla U_{n-1}^-\|\big) \\
&\quad + \big(1 + \|h_n\Delta_{h_n}U_{n-1}^-\| + \|\nabla U_{n-1}^-\|\big) \\
&\quad \cdot k_n\big(\|h_n^{-1}\Delta_{h_n}U_{n-1}^-\| + \|h_n^{-1}\nabla U_{n-1}^-\| + h_{n,\min}^{-1}\big)\big) \\
\|k_n R_t^o(U)\|_{L_\infty(I_n)} &\le Ce^{Ck_n}\mathcal{B}(U_{n-1}^-, h_n, k_n) \\
\|h_n^2 R_x^p(U)\|_{L_\infty(I_n)} &\le Ce^{Ck_n}\big(1 + \|U_{n-1}^-\| + \|h_n^2\Delta_{h_n}U_{n-1}^-\| + \|h_n\nabla U_{n-1}^-\| \\
&\quad + k_n\|\nabla U_{n-1}^-\| + \|U_{n-1}^-\|^2 + \|h_n^2\Delta_{h_n}U_{n-1}^-\|^2 \\
&\quad + \|h_n\nabla U_{n-1}^-\|^2 + k_n^2\|\nabla U_{n-1}^-\|^2\big)\cdot\mathcal{B}(U_{n-1}^-, h_n, k_n) \\
\|h_n^2 R_2^p(U)\|_{L_\infty(I_n)} &\le Ce^{Ck_n}\mathcal{B}(U_{n-1}^-, h_n, k_n) \\
\|(I-\tilde{P})R_x^o(U)\|_{L_\infty(I_n)} &\le Ce^{Ck_n}\big(\|h_n\nabla U_{n-1}^-\|^2 + \|h_n\Delta_{h_n}U_{n-1}^-\|^2 \\
&\quad + \mathcal{B}(U_{n-1}^-, h_n, k_n)\big),
\end{aligned}$$

where

$$\begin{aligned}
\mathcal{B}(U_{n-1}^-, h_n, k_n) = \|h_n^2\Delta_{h_n}U_{n-1}^-\| + \|h_n^2\nabla U_{n-1}^-\| + h_{n,\max}^2 + k_n\|\Delta_{h_n}U_{n-1}^-\| \\
+ k_n\|\nabla U_{n-1}^-\| + k_n.
\end{aligned}$$

Lemma 3.1 implies that $\|k_n R_t^p(U)\|_{L_\infty(I_n)}$ can be bounded as

$$\mathbf{O}\big(h_{n,\max}^{3/2} + k_n^2 h_{n,\min}^{-5/2} + k_n h_{n,\min}^{-1/2}\big)$$

while the rest of the residuals can be bounded as

$$\mathbf{O}\big(h_{n,\max}^{3/2} + k_n h_{n,\min}^{-1/2}\big).$$

Therefore, if we choose $k_n = \mathbf{o}\big(h_{n,\min}^{5/4}\big)$ then the residuals all tend to zero as $h_{n,\max}$ and k_n tend to zero. If we choose the scaling $k_n = Ch_{n,\min}^2$, which is expected in light of (3.18), then all the residuals tend to zero as $\mathbf{O}(h_{n,\max}^{3/2})$ as the mesh size and time steps are refined.

Remark 3.2. Estimates (3.16) and (3.17) on the finite element approximation can typically be proved for the dG and cG methods following the same analysis used to show the energy estimates on the continuous solution. We give an example in Section 3.3. The a priori error estimate (3.18) is closely related to the classic a priori error estimate, but it is simpler to establish since it holds only over one time step. A classic a priori error bound can be derived after (3.18) is established by using a discrete Gronwall argument.

Note that this estimate on the residual is not optimal order in time for the higher order cG(1) and dG(1) methods. An optimal order estimate would require energy estimates on higher order time derivatives of solutions of (3.1). This in turn requires sufficient compatibility between the forcing f and the boundary conditions rarely satisfied in practice. In the case that an energy estimate for $\Delta^2\tilde{u}$ can be proved, the a priori estimate on R_t can be modified to be second order in time for the cG(1) method and third order for the dG(1) method. By the same token, it is possible to

obtain even weaker estimates on the residual errors if necessary using straightforward modifications of the proofs of Theorems 3.2 and 3.4. For example, (3.13) can be replaced by

$$\|\Delta\tilde{u}^p\|_{L\infty(I_n)} \leq Ce^{Ck_n}\left(\|\Delta\tilde{u}^p_{n-1}\| + \|\nabla\tilde{u}^p_{n-1}\| + 1\right)$$

and (3.18) by

$$\|\tilde{e}\|_{L\infty(I_n)} \leq C\|h_n^2\Delta_{h_n}U^-_{n-1}\| + Ce^{Ck_n}\left(k_n\|\dot{\tilde{u}}\|_{L\infty(I_n)} + \|h_n^2\Delta\tilde{u}^p\|_{L\infty(I_n)} + \|h_n\nabla\tilde{u}^o\|_{L\infty(I_n)}\right)$$

and estimates can be derived that show the residuals tend to zero as $\mathbf{O}(h^{1/2}_{n,\max})$.

Remark 3.3. Apropos Remark 2.1, the assumptions in Theorem 3.4 are reasonable for problems with convection terms in the parabolic equations provided the diffusion 'dominates' the convection. Recall that if the diffusion matrix ϵ in (1.1) has diagonal entries ϵ_i while the convection matrices β_j have diagonal entries ϵ_{ji}, the diffusion dominates the convection if there is a constant $C > 0$ such that $\beta_{ji} \leq C\epsilon_i$ for all i and j. Otherwise as mentioned, weaker estimates could be derived.

3.2. The size of the stability factors. The last step in the interpretation of the a posteriori error estimate consists in showing that the dual problem (2.13) has a unique solution satisfying the regularity conditions (2.17). In other words, we show that the stability factors are defined.

We make some observations about the dual problem (2.13). First, note that while the coefficients $\bar{\epsilon}$, $\bar{\beta}$, and $\bar{f}$ are discontinuous in time, the discontinuities in time occur only at the time nodes of the discretization, so we can solve the dual problem on each time interval I_n in succession and on each interval the coefficients are smooth in time. The size of the coefficients, for example whether they are uniformly bounded or not, depends on the stability properties of the original differential equation (2.1) and the numerical method. Since no boundary conditions are imposed on the variables associated to the ordinary differential equations, in general the linear forcing term in (2.13) does not satisfy the boundary conditions. This gives rise to the possibility of boundary layers in ϕ affecting the size of S^p_x and S^1_t.

Note also that when the original problem (2.1) has nonlinear diffusion coefficients, the associated dual problem has convection terms, even if the original problem has no convection. In particular if the diffusion coefficients vary rapidly or otherwise have large derivatives, then the convection terms in the dual problem will be large. This potentially has a strong effect on the stability factors, since it is generic to have both characteristic and boundary layers in solutions of convection-diffusion problems with strong convection.

In the case that the original system (2.1) consists entirely of parabolic equations, i.e. $d = D$, then the techniques described in Ladyženskaja, Solonnikov, and Ural'ceva [47] and Racke [58] can be used to establish the required properties of the solution of the dual problem. The same techniques give the assumed energy estimates once existence is established for a system of coupled parabolic and ordinary differential equations. We show the necessary results for a class of problems coupling ordinary and partial differential equations with constant diffusion in Section 3.3.

Remark 3.4. An a posteriori error estimate is superficially similar to an a posteriori *convergence* result. An example of such a result is the analysis of the forward Euler

difference scheme for an ordinary differential equation in Henrici [43]. By an a posteriori analysis, the scheme is shown to define a Cauchy sequence that converges to a continuously differentiable function which turns out to be the solution. We have shown that under the right conditions, the residual errors of the Galerkin approximations for (2.1) on any interval can be made arbitrarily small and moreover the stability factors on a fixed discretization of a time interval $[0, T]$, where $T = t_n$ for some n, are finite. But without further information, it does not follow from the a posteriori error estimate that the error can be made arbitrarily small by refining the mesh and time steps. This is due to the fact that the stability factors depend on the computed approximation, hence if we refine the mesh and time steps to compute an approximation with a smaller residual error, we also obtain new, and possibly larger, stability factors. In order to obtain a convergence result, we require bounds on the stability factors that are independent of the approximation. One way to obtain such estimates is to show a priori that all approximations computed with sufficiently fine time steps and space meshes are contained in a compact set in $\mathbf{R}^D$. We discuss this further in Section 5.

As we mentioned, the analysis bounding the stability factors also provides the minimum ingredients necessary in order to expect to be able to compute accurate numerical approximation of the solution of the dual problem and the stability factors. This is what we do in practice. To illustrate the potential gain from this effort, we consider the Lorenz system (3.11) and the bistable example (1.2) once more. The technique used to compute the approximate stability factors shown in the following plots is described in Section 4.

Recall that the trajectory of the Lorenz system plotted in Fig. 3.3 (b) first becomes grossly inaccurate at $t \approx 17.8$ but the residual error R_t is not particularly large at this point. In Fig. 3.5(a), we plot an approximation of the stability factor $S_t^1(0, t_n)$ at many time nodes. $S_t^1(0, t_n)$ does not grow exponentially at a steady rate. In general,

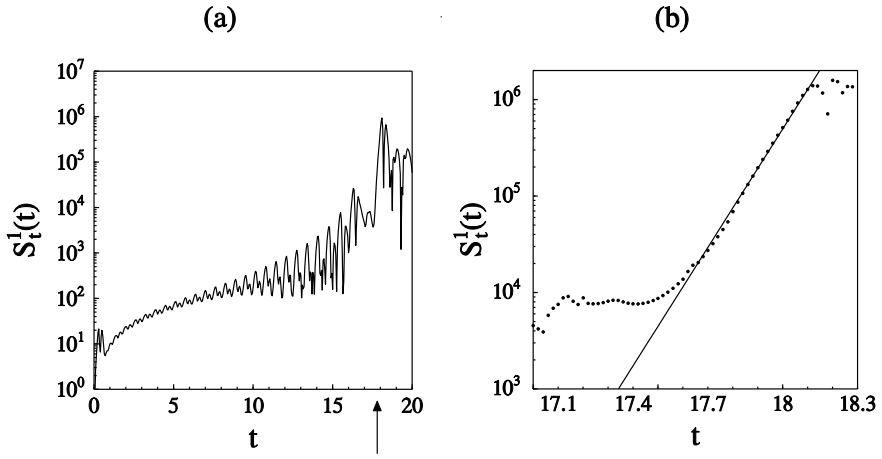

FIG. 3.5. *Log plots of the approximate stability factor $S_t^1(0, t_n)$ versus time for the trajectory plotted in Fig. 3.3(b). (b) shows the values of the $S_t^1(0, t_n)$ during the first period of rapid increase. This coincides with the transition from orbiting around one nonzero fixed point to orbiting around the other nonzero fixed point. A line fitted to the data during the period of greatest increase shows that $S_t^1(t_n)$ is growing exponentially like* $\exp(4.1t)$ *with correlation $\rho^2 = .997$ during this time.*

the parts of trajectories where the solution revolves around one of the nonzero fixed

points are characterized by a polynomial rate of growth of $S_t^1(0,t_n)$ with respect to t_n. But this slower growth is punctuated by short periods of exponential growth while a solution passes from the neighborhood of one fixed point to a neighborhood of the other fixed point. This is clear in Fig. 3.5(b). This exponential growth coincides exactly with the time that the trajectory first becomes grossly inaccurate. In other words, the cause of the sudden decrease in accuracy of numerical solutions of the Lorenz system is that trajectories become strongly unstable in a region of phase space near the z-axis. The instability reflects the fact that trajectories that are very close as they approach the z-axis can end up around different fixed points. This is not reflected in the residual errors of trajectories in the same region.

During the period of exponential growth, $S_t^1(0,t_n)$ is approximately proportional to $\exp(4.1t)$. This is less than the maximum rate of increase suggested by the standard a priori error bound. However, we can find trajectories that pass closer to the z-axis d for which S_t^1 does increase like $\exp(100t)$, see Estep and Johnson [31].

In Fig. 3.6, we plot the stability factors for the numerical solution of the bistable problem plotted in Fig. 1.2 and the corresponding error bound. The stability factors

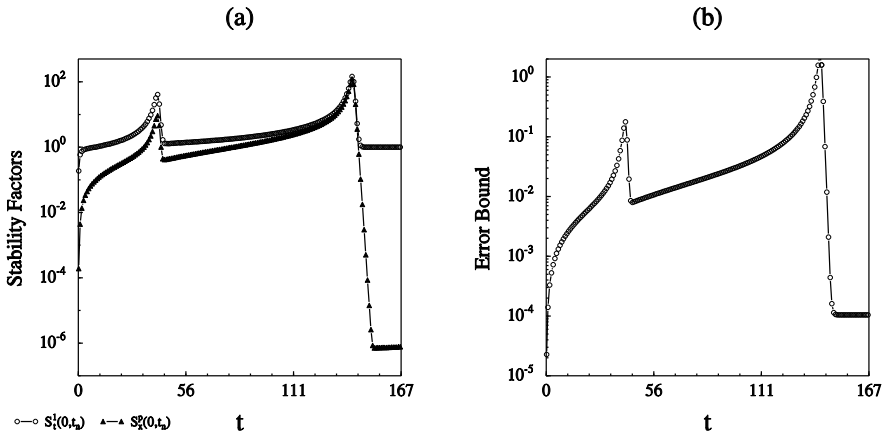

FIG. 3.6. *In (a), we plot the approximate stability factors versus time for the trajectory of the bistable problem plotted in Fig. 1.2. In (b), we plot the corresponding error bound.*

grow super-exponentially as the transients are approached, yet overall remain moderately sized because they decrease extremely rapidly to a value close to one after each transient. This indicates that the trajectory becomes quite stable after each transient. One possible explanation for this behavior is that the Lyapunov spectrum of the system contains just one unstable mode that is initially very close to zero (exponentially close in $\sqrt{\epsilon}$) during the beginning of a metastable phase but then subsequently grows as time passes. In this case, the exponential factor in the a priori error bound clearly overestimates the rate of accumulation of errors also fails to indicate the changes in sensitivity of the solutions to perturbations that is inherent to the metastable phase.

In Section 3.4, we plot approximate stability factors and residual errors for typical trajectories of the applications discussed in Section 1. The technique used to approximate the stability factors is described in Section 4.

3.3. Application of the theory to systems with constant diffusion. In this section, we prove that the a priori assumptions needed for the a posteriori analysis

hold for the system of reaction-diffusion equations with constant diffusion:

$$\begin{cases} \dot{u}_i - \epsilon_i \Delta u_i = f_i(u), & (x,t) \in \Omega \times \mathbf{R}^+,\ 1 \le i \le D, \\ u_i(x,t) = 0, & (x,t) \in \partial\Omega \times \mathbf{R}^+,\ 1 \le i \le d, \\ u(x,0) = u_0(x), & x \in \Omega, \end{cases} \tag{3.19}$$

where $\epsilon_i \ge \epsilon > 0$ for $1 \le i \le d$ and $\epsilon_i = 0$ for the rest. The analysis uses standard energy arguments applied in ways that accommodate the presence of ordinary differential equations coupled together with the parabolic partial differential equations.

Remark 3.5. The analysis in this section also applies to problems with homogeneous Neumann boundary conditions.

The local existence and uniqueness of the solution of a system of semilinear parabolic equations with locally Lipschitz continuous reaction terms is a classical result. Smoller [62] presents a proof based on a fixed point argument in the case when the space domain is all of $\mathbf{R}^i$. A simple variation of this argument can be applied to (3.19) (including ordinary differential equations) if we assume that the reaction f^p associated to the parabolic equations satisfies a *compatibility condition* on the boundary. Namely, if $f^p|_{\partial\Omega} \equiv 0$ then there is a $\delta = \delta(||u_0||_{H^2(\Omega)})$ such that (3.19) has a unique solution

$$u \in C^1\big((0,\delta);\ \prod_{i=1}^{d} H_0^1(\Omega) \cap H^2(\Omega) \times \prod_{i=d+1}^{D} H^2(\Omega)\big)$$

contained in some "ball" $\mathcal{B}(u_0,\rho) \subset \mathbf{R}^D$ centered at the initial value with radius ρ. The compatibility assumption removes technical details having to do with regularity at the boundary of Ω. All of the examples mentioned in the introduction, except the Hodgkin-Huxley equations and the morphogenesis model, satisfy this assumption.

The first result, proved in Section 7, gives the energy estimates used to estimate the residual errors in the a posteriori estimate:

PROPOSITION 3.5. *There is a constant $C = C(\epsilon, L)$, where L is the Lipschitz constant of f on $\mathcal{B}(u_0,\rho)$, such that*

$$\|\nabla u(t)\| \le C\|\nabla u(0)\| e^{Ct}, \tag{3.20}$$

$$\|\Delta u^p(t)\| \le \|\Delta u^p(0)\| + C\|\nabla u(0)\| e^{Ct}, \tag{3.21}$$

$$\|\Delta u^o(t)\| \le C\big(\|\Delta u(0)\| + \|\nabla u(0)\| + \|\nabla u(0)\|^2\big) e^{Ct}. \tag{3.22}$$

Remark 3.6. In the analysis below, we use the local existence of the solution $\tilde{u}$ of (3.19) on an interval $[t_{n-1}, t_{n-1} + \delta_n]$ with initial data $\tilde{u}_{n-1}$ given at t_{n-1} contained in $\mathcal{B}(\tilde{u}_{n-1}, \tilde{\rho}_n)$ and Proposition 3.5 on the same interval with the obvious change in the notation.

Remark 3.7. If f does not satisfy the compatibility condition, it is still possible to show that

$$\|\nabla u^p(t)\| \le \|\nabla u^p(0)\| + Ct^{1/2},$$

$$\|\nabla u^o(t)\| \le C(\|\nabla u(0)\| + M\sqrt{t}) e^{Ct},$$

$$\int_0^t \|\Delta u^p(s)\|\, ds \le \|\nabla u(0)\| + M\sqrt{t},$$

where M is the maximum of $|f|$ on $\mathcal{B}(u_0,\rho)$. This leads to correspondingly weaker estimates on the size of the residual errors.

The next ingredients are the existence, stability, and local a priori convergence properties of the approximation methods. The existence is again classical. We assume that given U_{n-1}^- for some positive integer n, there is a maximum time step K_n such that for all $k_n \le K_n$, U exists uniquely on (t_{n-1},t_n) and is contained in the ball $\mathcal{B}(U_{n-1}^-,\rho_n)$. Such a result can be proved by a fixed point argument similar to the argument used to show existence for the differential equation. See Estep and Stuart [34], French and Jensen [38], and Mascagni [49] for related results.

Recall that we analyze the residual errors on I_n by comparing U to the local solution $\tilde{u}$ of (3.19) on the interval $(t_{n-1},t_n]$ with "initial data" $\tilde{u}_{n-1} = T\Delta_{h_n}U_{n-1}^-$ obtained by smoothing U_{n-1}^-. The following result contains the necessary stability and accuracy estimates on U for the a posteriori theory.

PROPOSITION 3.6. *Let M_n denote the maximum of $|f|$ and L_n the Lipschitz constant of f on the convex hull of $\mathcal{B}(\tilde{u}_{n-1},\tilde{\rho}_n)\cup\mathcal{B}(U_{n-1}^-,\rho_n)$. There is a constant $C = C(M_n)$ such that for all time steps with k_nL_n sufficiently small, the three approximations satisfy*

$$\|U^p\|_{L_\infty(I_n)} \le C\big(\|U_{n-1}^{p,-}\| + k_n\|\nabla U_{n-1}^{p,-}\| + k_n\big), \tag{3.23}$$

$$\|U^o\|_{L_\infty(I_n)} \le C\big(\|U_{n-1}^{o,-}\| + k_n\big). \tag{3.24}$$

In addition, there is a constant $C = C(\epsilon,L_n)$ such that the cG(1) and dG(1) approximations satisfy

$$\|\tilde{u}-U\|_{L_\infty(I_n)} \le \|h_n^2D^2\tilde{u}_{n-1}\| + Ce^{Lk_n}\big(k_n^{1/2}\|h_nD^2\tilde{u}^p\|_{L_\infty(I_n)} + \|h_n^2D^2\tilde{u}\|_{L_\infty(I_n)}\big) + Ce^{Lk_n}\begin{cases} k_n\|\dot{\tilde{u}}\|_{L_\infty(I_n)} \\ k_n^2\|\ddot{\tilde{u}}\|_{L_\infty(I_n)}\end{cases} \tag{3.25}$$

and the dG(0) approximation satisfies

$$\|\tilde{u}-U\|_{L_\infty(I_n)} \le \|h_n^2D^2\tilde{u}_{n-1}\| + Ce^{Lk_n}\big(k_n^{1/2}\|h_nD^2\tilde{u}^p\|_{L_\infty(I_n)} + \|h_n^2D^2\tilde{u}\|_{L_\infty(I_n)}\big) + Ce^{Lk_n}k_n\|\dot{\tilde{u}}\|_{L_\infty(I_n)}. \tag{3.26}$$

Remark 3.8. It is possible to replace the quantities in the middle on the right in (3.25) and (3.26) by

$$k_n^{1/2}\|\nabla\tilde{u}^p\|_{L_\infty(I_n)} + \|h_n\nabla\tilde{u}\|_{L_\infty(I_n)}$$

if weaker estimates are desired.

The results in Propositions 3.5 and 3.6 do not exactly match the assumptions in Theorem 3.4, where the precise estimates on the size of the residual errors are stated. However, the additional terms in (3.21), (3.25), and (3.26) can be handled similarly to the terms in (3.13) and (3.18) and the result stated in Theorem 3.4 holds with

$$\begin{aligned}\mathcal{B}(U_{n-1}^-,h_n,k_n) = {} & \|h_n^2\Delta_{h_n}U_{n-1}^-\| + \|h_n^2\nabla U_{n-1}^-\| + h_{n,\max}^2 + \|h_n^2\Delta_{h_n}U_{n-1}^-\|^2 \\ & + \|h_n^2\nabla U_{n-1}^-\|^2 + k_n^{1/2}\|h_n\Delta_{h_n}U_{n-1}^-\| + k_n^{1/2}\|h_n\nabla U_{n-1}^-\| \\ & + k_n^{1/2}h_{n,\max} + k_n^{1/2}\|h_n^{3/2}\Delta_{h_n}U_{n-1}^-\|^2 + k_n^{1/2}\|h_n^{1/2}\nabla U_{n-1}^-\|^2 \\ & + k_n\|\Delta_{h_n}U_{n-1}^-\| + k_n\|\nabla U_{n-1}^-\| + k_n.\end{aligned}$$

This does not change the overall asymptotic result.

We conclude by estimating the size of the stability factors in the a posteriori estimate. The dual problem for (3.19) reads

$$\begin{cases} -\dot{\phi}_i - \epsilon_i \Delta \phi_i = \sum_{j=1}^{D} \bar{f}_{ji}\, \phi_j, & x \in \Omega, t_n > t \geq 0,\ 1 \leq i \leq d, \\ -\dot{\phi}_i = \sum_{j=1}^{D} \bar{f}_{ji}\, \phi_j, & x \in \Omega, t_n > t \geq 0,\ d < i \leq D, \\ \phi_i(x,t) = 0, & x \in \Omega, t_n > t \geq 0,\ 1 \leq i \leq d, \\ \phi(x,t_n) = \phi_n(x), & x \in \Omega, \end{cases} \tag{3.27}$$

where the coefficients

$$\bar{f}_{ij} = \int_0^1 \frac{\partial f_i}{\partial x_j}(us + U(1-s))\, ds$$

are bounded, piecewise differentiable functions, continuous everywhere except at time nodes t_n. The techniques used in Ladyženskaja, Solonnikov, and Ural'ceva [47] apply directly to give the existence, uniqueness, and regularity of the dual solution required for the definitions of S_0, S_t^1, S_x^p, and S_x^o to make sense. Furthermore, we prove the following proposition in Section 7.

PROPOSITION 3.7. *Let L_n denote the maximum of $|\bar{f}_{ij}|$, $1 \leq i,j \leq D$, on $\overline{\Omega} \times [0,t_n]$. There is a constant $C = C(\epsilon, L_n)$ such that*

$$\|\phi(0)\| \leq e^{Ct_n}\|\phi_n\|$$

$$\int_0^{t_n} \|\phi^o\|\, ds \leq C\big(e^{Ct_n} - 1\big)\|\phi_n\|$$

$$\int_0^{t_n} \|\Delta\phi^p\|\, ds \leq Ct_n^{1/2} e^{Ct_n}\|\phi_n\| + t_n^{1/2}\|\nabla\phi_n^p\|$$

$$\int_0^{t_n} \|\dot{\phi}^p\|\, ds \leq \big(e^{Ct_n} - 1\big)\|\phi_n\| + t_n^{1/2} e^{Lt_n}\|\phi_n\| + t_n^{1/2}\|\nabla\phi_n^p\|$$

$$\int_0^{t_n} \|\dot{\phi}^o\|\, ds \leq \big(e^{Ct_n} - 1\big)\|\phi_n\|$$

We summarize this result by stating

$$\max\big\{S_0(0,t_n), S_t^1(0,t_n), S_x^p(0,t_n), S_x^o(0,t_n)\big\} \leq C(\epsilon, L_n) e^{C(\epsilon,L_n)t_n}\big(\|\phi_n\| + \|\nabla\phi_n^p\|\big). \tag{3.28}$$

3.4. A stability factor gallery. In this section, we apply the a posteriori theory to estimate the error of numerical solutions of the nine example problems listed in Section 1.

Consider first the solution of the bistable problem (1.2) starting with the initial data with two wells plotted in Fig. 1.1. We plotted the evolution of the residual errors in Fig. 2.1 (a) and of the stability factors in Fig. 3.6 (b). Based on the size of the stability factors, we conclude that it is possible to compute accurate numerical solutions over long time intervals when $\epsilon = .0009$. This is born out by the plot of

the error bound in Fig. 3.6 (c) for the numerical solution using 256 uniformly spaced elements through the collapse of the second well and moderately sized time steps and requiring about 15 minutes computation time on a Sparc 10. This is a sharp contrast to the conclusion suggested by classic a priori error analysis.

We now perform a similar analysis on the other examples listed in Section 1. Note that many of these examples allow a great variety of behavior in their solutions and we cannot explore each example fully in this paper. By way of making a touchstone to compare the stability properties of the different problems, we choose the parameters and data to produce solutions that converge to a fixed point if possible.

Example 2: equations for two species. We consider the model of predator-prey interaction analyzed in Smoller [62] with $M = -(u_1 - \alpha_1)(u_1 - 1) - \alpha_2 u_2$ and $N = -\alpha_3 - \alpha_4 u_2 + \alpha_2 u_1$ for values of the parameters that force the existence of a stable fixed point and solve the problem until the components converge. We compute with $\epsilon_1 = \epsilon_2 = 10^{-2}$, $\alpha_1 = .25$, $\alpha_2 = 2$, $\alpha_3 = 1$, $\alpha_4 = 3.4$, and homogeneous Neumann boundary conditions using CARDS with 256 elements and keeping the time residual error below .0001. The evolution of the components is displayed in Fig. 3.7.

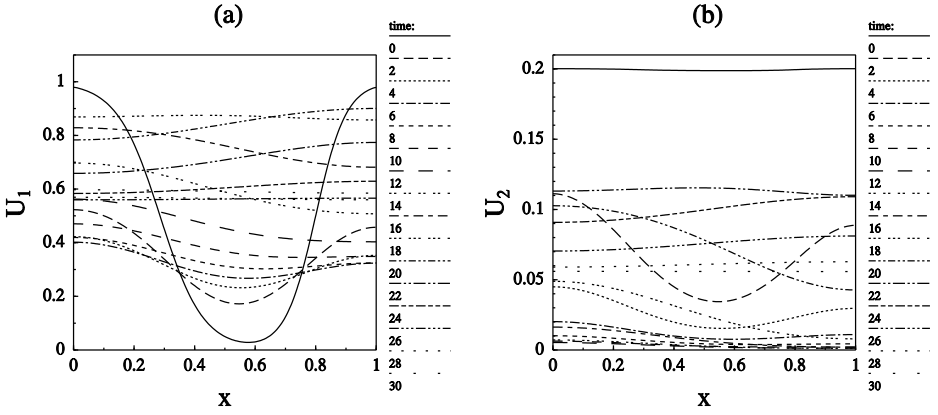

FIG. 3.7. *Equations for two species: plots of $U_1(\cdot, t)$ and $U_2(\cdot, t)$ at the indicated times.*

In Fig. 3.8, we plot the stability factors, residual errors, and error estimate computed by CARDS. Note that the stability factors increase super-exponentially as the components evolve towards a uniform shape, decrease as the components move towards the fixed point, but then increase substantially again. The cause of the second increase in instability is not apparent. A "movie" in the phase space shows the solution spiralling into the fixed point during this period while simultaneously its orientation as a curve in phase space changes. To study this phenomena more closely, we repeated this computation using $\epsilon_1 = \epsilon_2 = 10^{-4}$ over the time interval $[0, 160]$. We plot the stability factors of the resulting solution in Fig. 3.9. The pattern of increasing and decreasing stability factors towards the end is clear.

Example 3: Hodgkin-Huxley equations. Following Cooley and Dodge [15], we use: $\epsilon_1 = .000345$, $\alpha_1 = 120$, $\alpha_2 = 36$, $\alpha_3 = .3$, $\beta_1 = 115$, $\beta_2 = -12$, and $\beta_3 = 10.598$, while for simplicity, we choose homogeneous Neumann boundary conditions. The actual model requires a nonhomogeneous Neumann condition at $x = 0$ that simulates nerve stimulation impulses. We start with non-uniform data in u_1 and constant values for the rest and compute using 512 elements while keeping the time residual

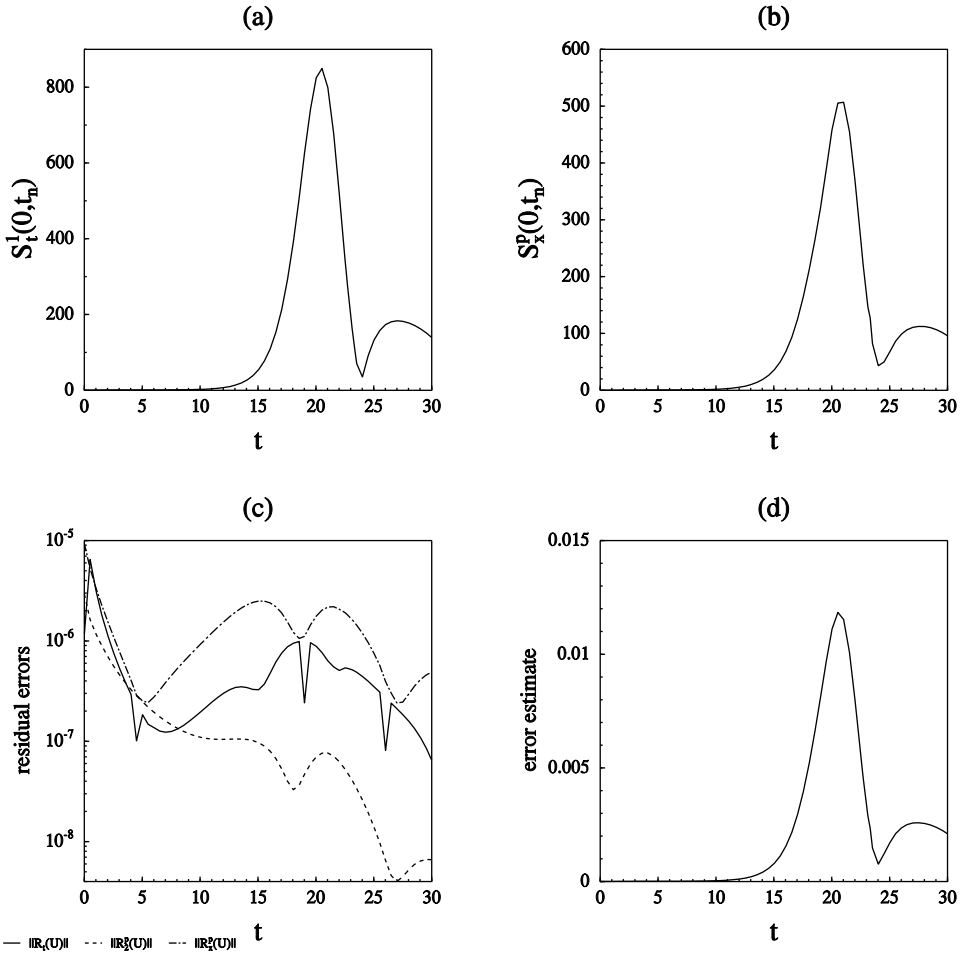

FIG. 3.8. *Equations for two species: in (a) and (b), we plot the stability factors S_t^1 and S_x^p versus time computed every .5 time unit. In (c), we plot the L_2 norms of the residual errors versus time. In (d), we plot the a posteriori error estimate. In this computation, $\epsilon_1 = \epsilon = 10^{-2}$.*

error below .001 The evolution of the components is displayed in Fig. 3.10. The initial signal causes a rapid initial increase in U_1 while by time 1 it is beginning to decrease in value again.

In Fig. 3.11, we plot the stability factors, residual errors, and error estimate computed by CARDS. This problem is difficult to compute and requires a large number of uniform mesh points in order to maintain a specified level of space residual error. Notice that the space stability factor increases initially but decreases again as the first component "flattens". The time stability factor is beginning to increase again as the amplitude of the solution begins to decrease again.

Example 4: Fitz-Hugh-Nagumo equations. These equations were proposed as a simplified model of the Hodgkin-Huxley equations and solutions are supposed to have the same qualitative behavior, see Rauch and Smoller [59]. One natural question about this model is the possibility of a "threshold" phenomena in solutions corresponding

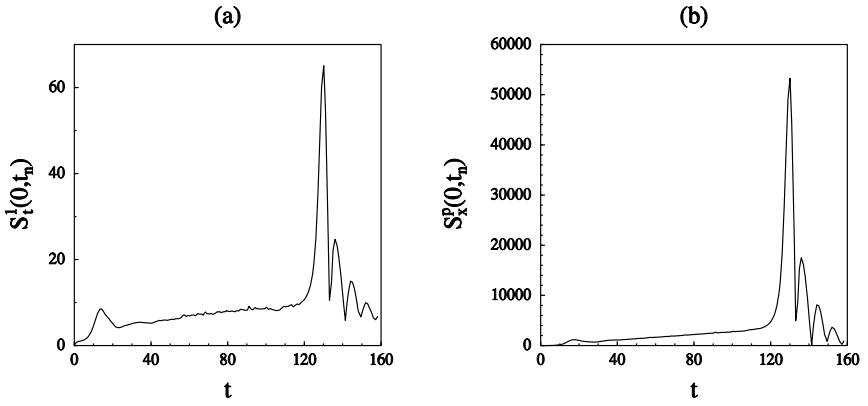

FIG. 3.9. *Equations for two species: in (a) and (b), we plot the stability factors S_t^1 and S_x^p versus time for a computation with $\epsilon_1 = \epsilon = 10^{-4}$.*

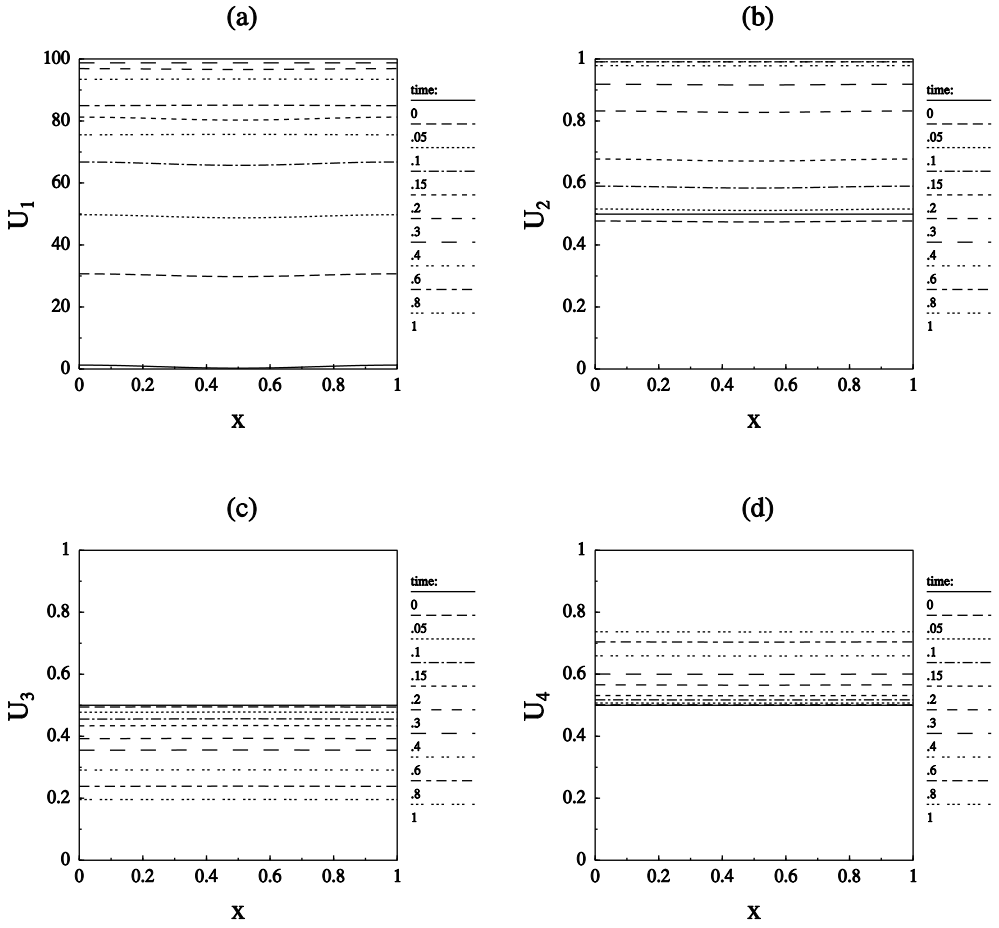

FIG. 3.10. *Hodgkin-Huxley equations: plots of the components of $U(\cdot,t)$ at the indicated times.*

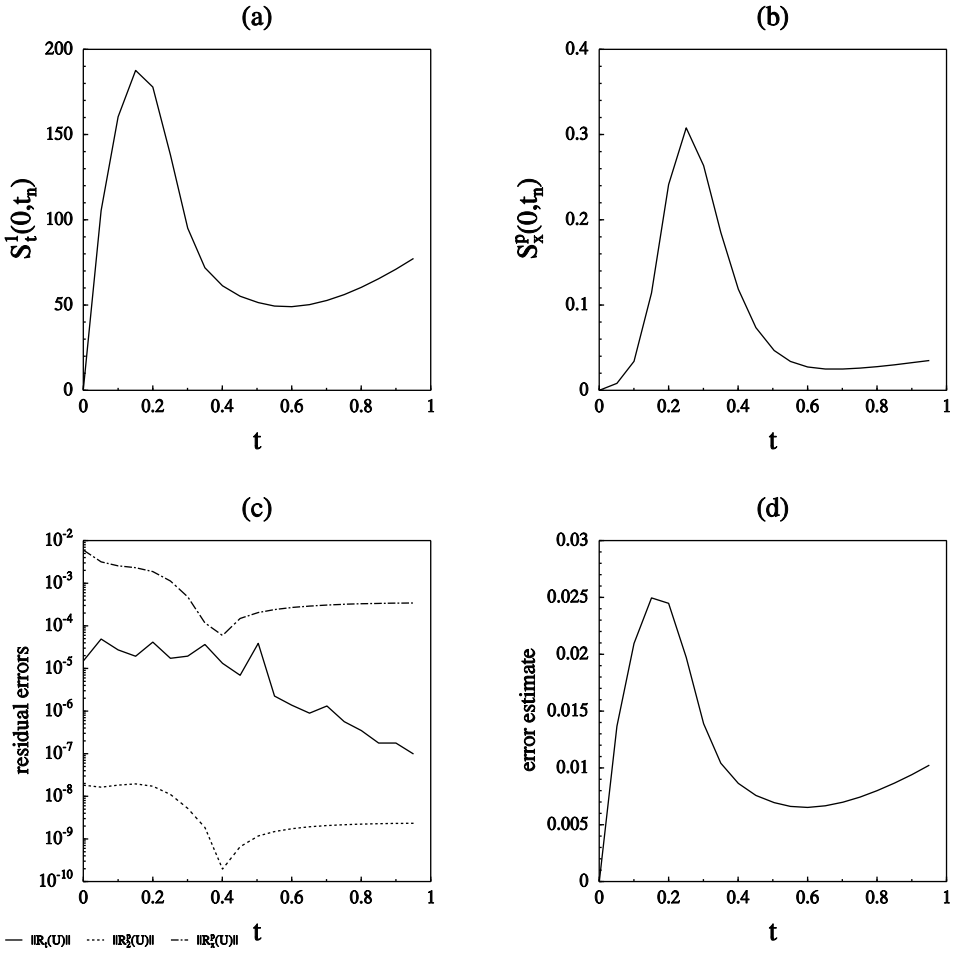

FIG. 3.11. *Hodgkin-Huxley equations: in (a) and (b), we plot the stability factors S_t^1 and S_x^p versus time computed every .05 time unit. In (c), we plot the L_2 norms of the residual errors versus time. In (d), we plot the a posteriori error estimate.*

to the fact that there is a minimum level of stimulus needed to excite a nerve. This translates to determining whether or not "small" initial data converge to zero as time passes. In fact, it is possible to prove that a class of data leads to solutions that decay exponentially quickly to zero. We investigate this numerically for the parameter values $\epsilon_1 = .1$, $\epsilon_2 = 0$, $\alpha_1 = .25$, $\alpha_2 = .1$, and $\alpha_3 = 1$, with Neumann boundary conditions using 64 elements in space and keeping the time residual errors below .001. Notice that this is a coupled parabolic-ordinary system. The evolution of the components is displayed in Fig. 3.12. The evolution towards the origin is clear.

In Fig. 3.13, we plot the stability factors, residual errors, and error estimate computed by CARDS. The time stability factor S_t^1 tends to a constant value reflecting the exponential stability of the evolution towards the fixed point. The space stability factor increases to a relatively large value initially as the components evolve towards becoming uniform but then begins to decay as they move closer to the fixed point.

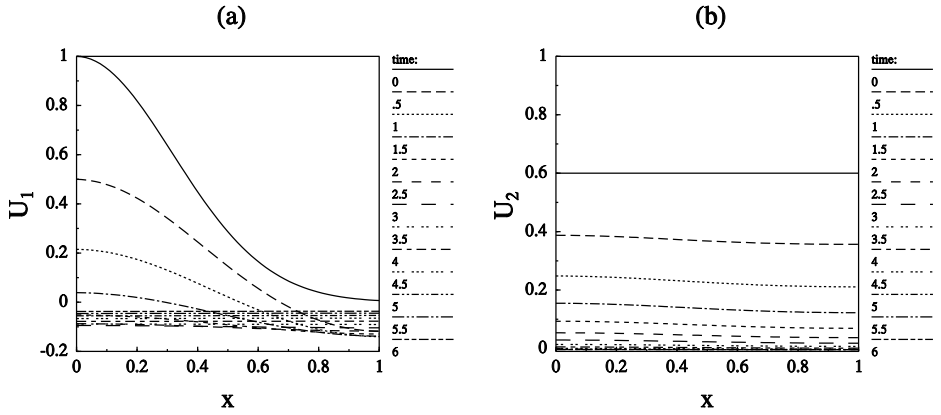

FIG. 3.12. *Fitz-Hugh-Nagumo equations: plots of* $U_1(\cdot,t)$ *and* $U_2(\cdot,t)$ *at the indicated times.*

This is reflected in the error estimate, which decreases slowly towards zero as the residuals become smaller.

Example 5: superconductivity of liquids. This is a model of superconductivity of liquids extensively analyzed in Brown, Donne, and Gardner [8]. With diffusion constants equal to one, it is known that the solutions tend to a continuum of steady-state solutions of the problem and in some cases to an individual steady-state solution. We solve the problem with smaller diffusion $\epsilon_1 = \epsilon_2 = 10^{-3}$ and homogeneous Dirichlet boundary conditions using 64 elements and keeping the time residual error below .001. The evolution of the components is displayed in Fig. 3.14. The evolution towards a steady-state solution is clear.

In Fig. 3.15, we plot the stability factors, residual errors, and error estimate computed by CARDS. As expected when the components converge to a nonconstant steady-state solution, the space residual errors converge to constant values. The error estimate marks a mild transient as the oscillations in the first component dampen, but then tends to a constant value as the solution converges to the steady-state.

Example 6: Field-Noyes equations. This is a model for the celebrated Belousov-Zhabotinsky chemical reaction. It is analyzed briefly in Smoller [62] and in slightly different form in Murray [53]. Following Murray, we choose $\alpha_1 = 2 \cdot 10^4$, $\alpha_2 = 8 \cdot 10^{-4}$, $\alpha_3 = 5 \cdot 10^3$, and $\alpha_4 = 1$ while $\epsilon_1 = \epsilon_2 = \epsilon_3 = 1$. We impose periodic boundary conditions and compute with 2048 elements while keeping the time residual below .001. The rapid evolution of the solution forced very small time steps. This is partly due to the scaling: if we rescale the problem so the diffusions are on the order of 10^{-3}, then the time scale of the results presented below would be on the order of $10--20$. The evolution of the components is displayed in Fig. 3.16. The solution forms steep fronts and this causes large space residual errors.

In Fig. 3.17, we plot the stability factors, residual errors, and error estimate computed by CARDS. Both stability factors increase monotonically, with the rapid change of the first two components in space reflected in S_x^p. The cause of the "spike" in the plot of S_x^p is not clear.

Example 7: model equations for flame propagation. This system arises in the study of combustion and is analyzed briefly in Chueh, Conley, and Smoller [12]. u_1 and u_2

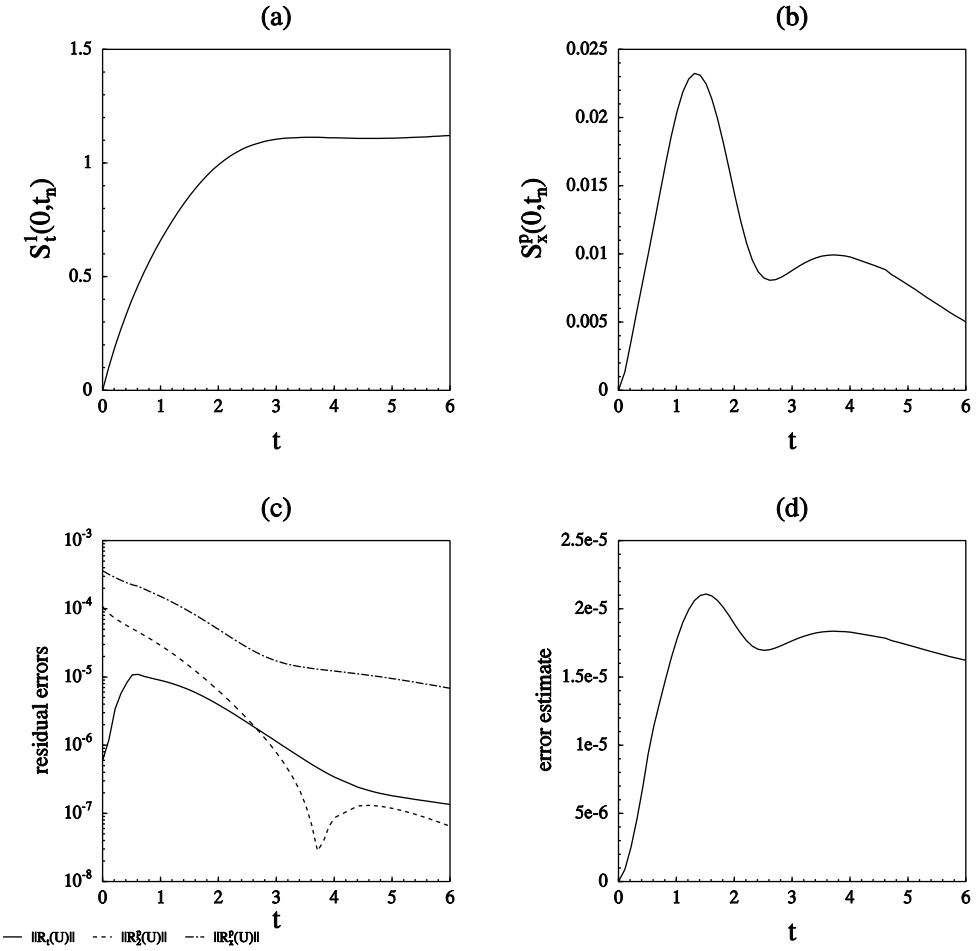

FIG. 3.13. *Fitz-Hugh-Nagumo equations: in (a) and (b), we plot the stability factors* S_t^1 *and* S_x^p *versus time computed every* .1 *time unit. In (c), we plot the* L_2 *norms of the residual errors versus time. In (d), we plot the a posteriori error estimate. In this computation,* $\epsilon_1 = \epsilon = 10^{-2}$.

represent the concentration and temperature of a combustible substance respectively. We solve the problem with $\epsilon_1 = \epsilon_2 = 10^{-3}$, $\alpha_1 = .5$, $\alpha_2 = .5$, and homogeneous Neumann boundary conditions using 64 elements while keeping the time residual error below .001. The evolution of the components is displayed in Fig. 3.18. It is interesting to see the transfer of the pattern in the concentration into the temperature as time passes. The concentration converges to zero while the temperature first evolves towards a pattern centered around a non-zero value and afterwards begins to converge to a uniform state.

In Fig. 3.19, we plot the stability factors, residual errors, and error estimate computed by CARDS. The stability of the problem with respect to discretization in errors in time is reflected in the time stability factor S_t^1. The solution gradually becomes more stable with respect to discretization errors in space as the second component tends to a uniform state. Notice the slow decrease in the space residual errors as the

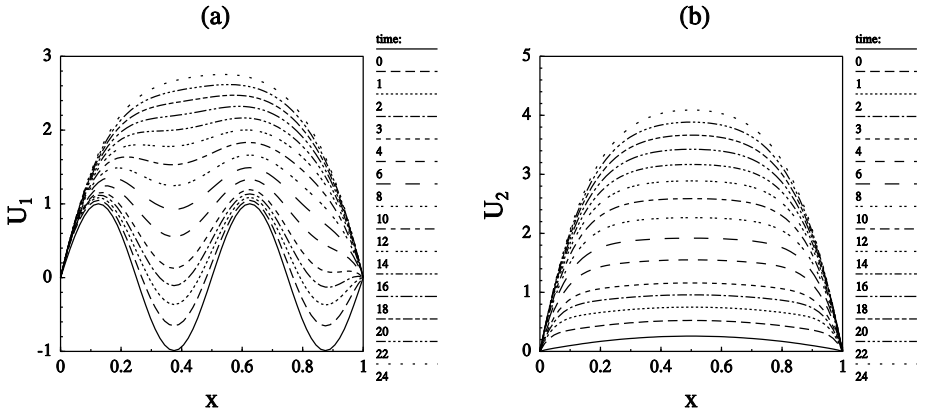

FIG. 3.14. *Model of superconductivity: plots of* $U_1(\cdot,t)$ *and* $U_2(\cdot,t)$ *at the indicated times.*

second component "flattens". We computed the solution to time 150. U_1 is indistinguishable from zero while U_2 has an oscillatory pattern with amplitude less than .06 centered around 1. The error estimate at $t = 150$ is 2.7×10^{-4}.

Example 8: model equations for morphogenesis. This is a version of the autocatalytic Selkov model of glycolysis proposed by Gray and Scott, see Pearson [57], that demonstrates the great number of patterns that can be formed by solutions of a relatively simple parabolic equation. We compute the numerical solution for $\epsilon_1 = 2 \times 10^{-5}$, $\epsilon_2 = 10^{-5}$, $\alpha_1 = .045$, $\alpha_2 = .065$, and periodic boundary conditions using CARDS with 256 elements and keeping the time residual error below .001. The evolution of the components is displayed in Fig. 3.20. The solutions form a pattern of layers after a rapid initial transient which then slowly evolves over a long time until eventually the solutions converge to constant values. Slight changes in the parameters and the data lead to completely different patterns of layers.

In Fig. 3.21, we plot the stability factors, residual errors, and error estimate computed by CARDS. It is interesting to compare these stability factors to the stability factors of the bistable problem displayed in Fig. 3.6. The evolution of the layers in the two problems occur on essentially the same time scale. But the stability properties of the solutions are very different. The strong sensitivity of numerical solutions of these equations to the space discretization is marked by the large values of S_x^p. The limiting factor in determining how long accurate solutions can be computed is the number of mesh points that can be used. The sharp changes in $\|R_t^1(U)\|$ in Fig. 3.21 (c) are due to the decreases in time steps that are required for the QMR iteration to converge.

Example 9: model for the spread of rabies. This model is described in Murray [53] in detail. It is a SIR model in which the fox population is divided into three groups: the susceptible (S), represented by u_1, the infected but not infectious (I), represented by u_2, and the infected, rabid (R), represented by u_3. Diffusion only occurs for the rabid foxes. We use the parameters considered by Murray, $\alpha_1 = .003$, $\alpha_2 = .003$, $\alpha_3 = .08$, and $\alpha_4 = .46$ except that we choose a smaller diffusion, $\epsilon_3 = .001$, to get a more convenient scale in x. We impose homogeneous Neumann boundary conditions on u_3 and give the initial conditions $u_1 \equiv 1$, $u_2 \equiv 0$, and u_3 a small "spike" centered at the midpoint of the interval. We compute with 128 elements keeping the time residual error below .001. The evolution of the components is displayed in Fig. 3.22.

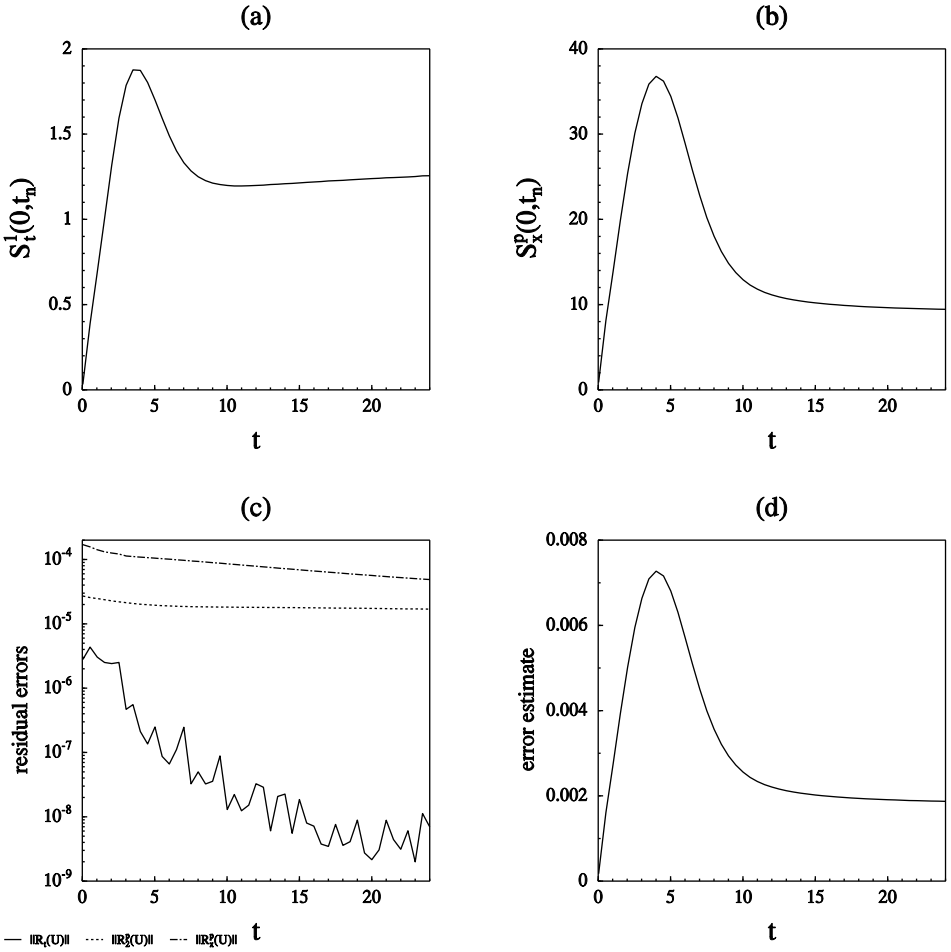

FIG. 3.15. *Model of superconductivity: in (a) and (b), we plot the stability factors* S_t^1 *and* S_x^p *versus time computed every .5 time unit. In (c), we plot the* L_2 *norms of the residual errors versus time. In (d), we plot the a posteriori error estimate.*

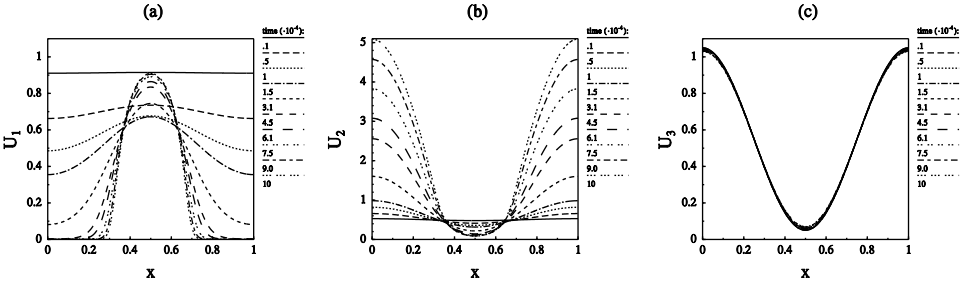

FIG. 3.16. *Field-Noyes equations: plots of* $U_1(\cdot,t)$, $U_2(\cdot,t)$, *and* $U_3(\cdot,t)$ *at the indicated times.*

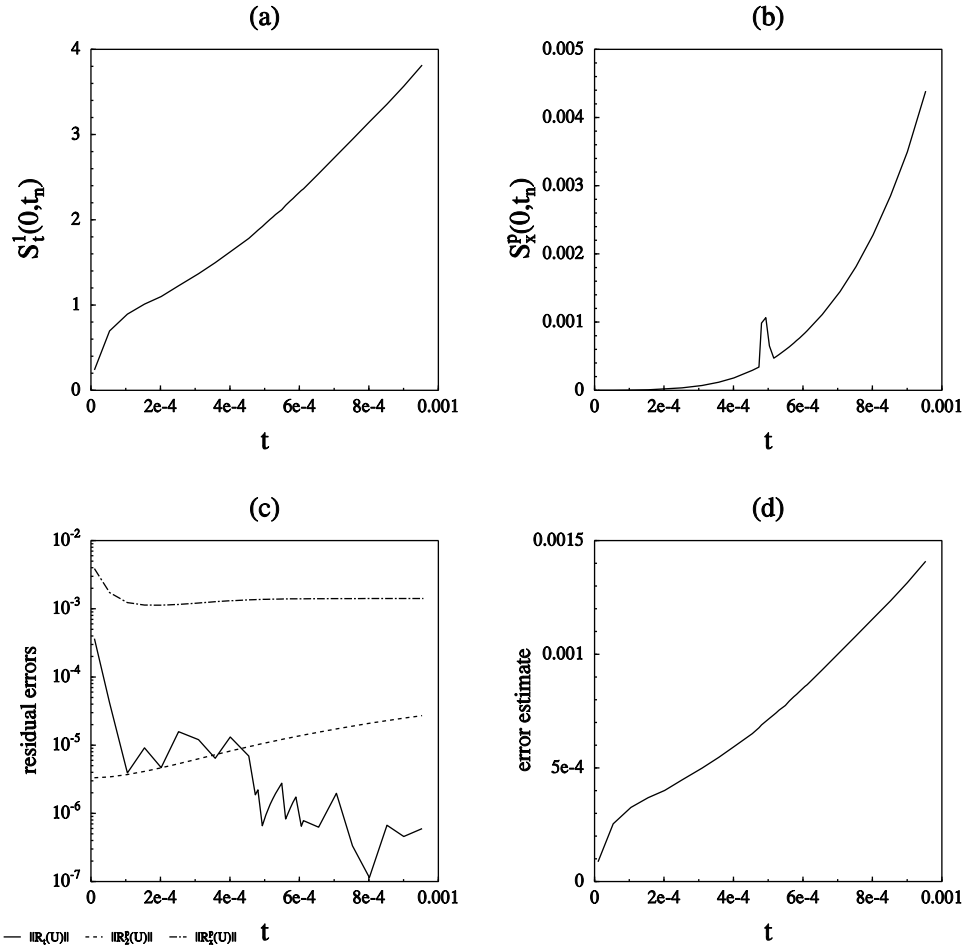

FIG. 3.17. *Field-Noyes equations: in (a) and (b), we plot the stability factors S_t^1 and S_x^p versus time computed every .5 time unit. In (c), we plot the L_2 norms of the residual errors versus time. In (d), we plot the a posteriori error estimate.*

The spread of the infected foxes through the domain and the effect on the susceptible population is clear. However, the rabid foxes appear to die out before a plague can occur. We computed until time $t = 1000$, by which point $u_2 \equiv u_3 \equiv 0$ while u_1 is within .06 of 1 at every x.

In Fig. 3.23, we plot the stability factors, residual errors, and error estimate computed by CARDS. It is interesting to note the linear growth of the space discretization stability factor S_x^p and the error estimate in spite of the apparent stability of the fixed point to which the solution is converging.

4. Practical matters.

In this section, we describe some details of the implementation of the dG and cG finite element methods for reaction-diffusion problems and some practical issues

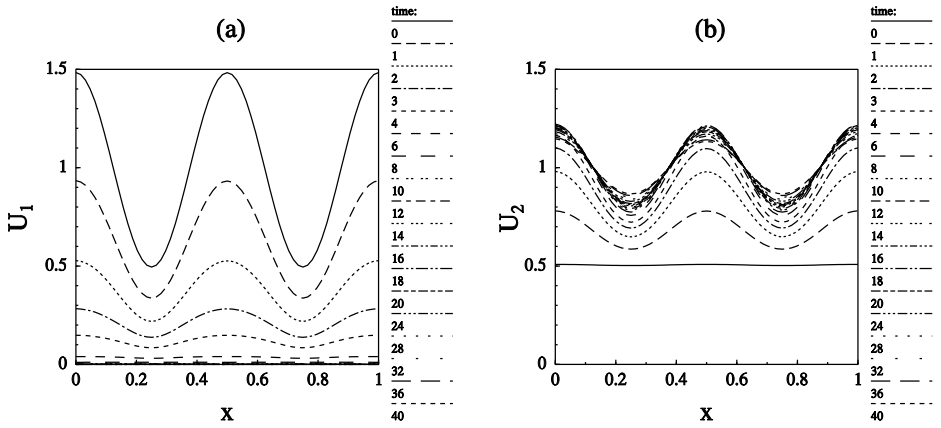

FIG. 3.18. *Model of flame propagation: plots of* $U_1(\cdot,t)$ *and* $U_2(\cdot,t)$ *at the indicated times.*

that arise when using the a posteriori error estimate (2.21) to estimate the error of a computation.

4.1. Some details of implementation. The main purpose of this section is to describe how the a posteriori theory can be implemented into a code that solves a system of reaction-diffusion equations numerically. We do not address issues of efficiency. Rather our purpose is to show that the a posteriori error estimate can be used to estimate the error of numerical solutions of physically interesting problems. Further details can be found in Estep and Williams [35].

4.1.1. A matrix-free implementation. There are several key factors that have to be taken into account when designing a general code to solve reaction-diffusion problems. First, it is commonly necessary to handle some very large vectors and arrays during the course of the solution, for example the approximation and the matrix associated to the discretized Laplacian. Second, concerns about accuracy and efficiency often require that space meshes and time steps vary as time passes, hence there is a practical need for flexibility in the dimensioning of arrays and vectors. Third, there is also a need for flexibility in handling different kinds of boundary conditions and even different numbers and types of equations for practical application. These concerns can be handled by writing the code using a matrix-free approach.

The basis for a matrix-free implementation is object-oriented. Actually, object-oriented thinking has its roots in mathematics and not computer science. As an example, we consider the meaning of the word "vector". To an old-style Fortran programmer, a vector is just an array of floating-point numbers of fixed length and it is created with a command like `DIMENSION V(20)`. This fixes the size of the array at the time the program is compiled, which makes assigning memory to the program and passing the array around the program easy, but does not give flexibility in terms of redefining `V`. In contrast, to a new-style object-oriented computer scientist, a vector is an instance of a class, and the class has functions, called *methods*, that are used to manipulate the vector. Typical methods include multiplication of the vector by a scalar, adding two vectors to produce a third, finding the norm of a vector, and so on. In other words, the paradigm change is to think about what can be done with a data object, rather than its contents.

But this of course is nothing more than the classical mathematical definition of

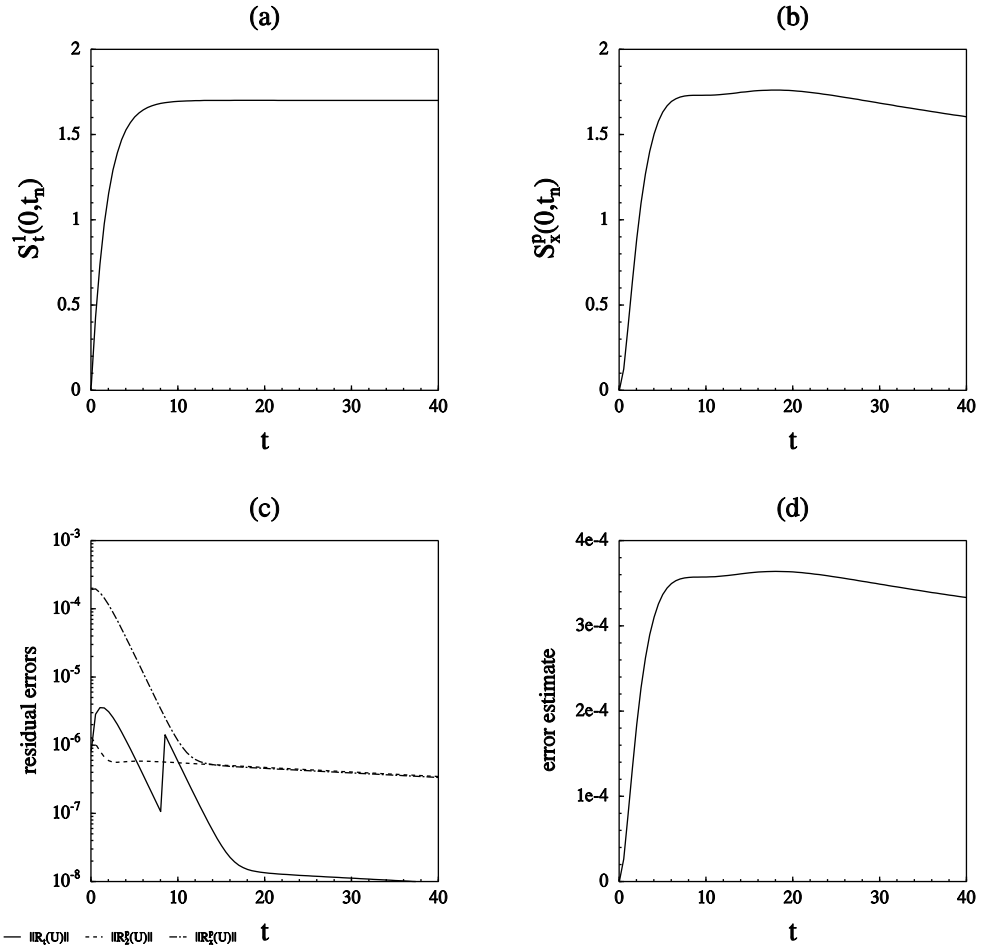

FIG. 3.19. *Model of flame propagation: in (a) and (b), we plot the stability factors S_t^1 and S_x^p versus time computed every .5 time unit. In (c), we plot the L_2 norms of the residual errors versus time. In (d), we plot the a posteriori error estimate.*

vector. Mathematicians usually do not think of a vector in terms of its components, but rather as a member of a set on which certain operations are defined, for example adding vectors to produce another. In this case, the vector space is a *class*, the vector is an *instance* of the class, and the axiomatic operations are the *methods* of the class.

The reason for object-oriented thinking, in mathematics or computer science, is to provide a systematic method for breaking down complex structures into a set of simpler structures that can be handled independently. In this way, theorems and software can be reused in new situations that involve known ideas.

If we introduce object-oriented thinking into the old-fashioned programming style, then the array of numbers `V` above would be considered to be an object on which the only method available is a map whose input is the subscript `i` and whose output is the corresponding array member `V(i)`. The same is true of a matrix that is stored as a two-dimensional array of numbers or of a sparse matrix stored in a clever scheme

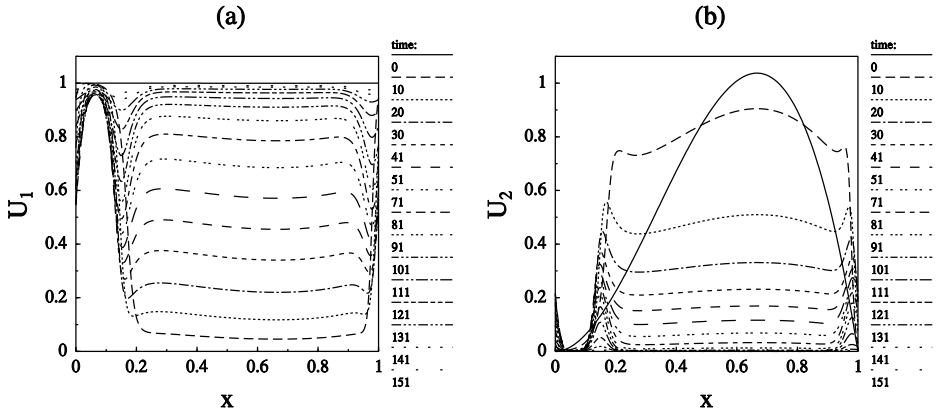

FIG. 3.20. *Model equations for morphogenesis: plots of* $U_1(\cdot,t)$ *and* $U_2(\cdot,t)$ *at the indicated times.*

that reduces storage requirements. In any case, the fundamental method defined on arrays in the classic style programming is accessing an element of the array.

This dichotomy between the abstract mathematical object and the easily implemented array storage of classic programming is the reason that it is very difficult to write flexible, efficient software for solving systems of reaction-diffusion equations. Following modern programming thought, the solution is to use object-oriented thinking in designing the code. The most fundamental example of this "mathematical" approach is that a matrix is treated as a linear operator rather than a collection of matrix elements. This is what is meant by a matrix-free implementation, since matrices are accessed only through operations on vectors and are never explicitly formed. This approach makes the software clean, flexible, and efficient.

The examples in this paper have been computed using the code *CARDS* [36], which uses the method of lines to solve problems of the form (3.19). Codes based on the method of lines can be viewed as consisting of a PDE solver acting on space meshes and an ODE solver acting in time. The mathematical structure of *CARDS* consists of a sequence of vector spaces:

- *Finite element functions* Vectors in this finite dimensional space consist of continuous, piecewise linear functions on a triangulation of the computational domain. An important method on this space is the discretized Laplacian.
- *Spatial approximations* A vector in this space approximates the multiple fields that interact in the system (2.1) in space. This space is made of finite-element functions and an important method on this space is the evaluation of the time-derivative of the differential equation system with the function f in (2.1).
- *Time-space approximations* To solve the system in time, we use an piecewise polynomial approximation of the solution in time whose coefficients in each subinterval consist of spatial approximations associated to a set of nodes appropriately chosen in each subinterval. An important method on this space is the estimation of the error using the a posteriori error bound.

Each of these vector spaces is defined in terms of the previous one in the sequence. A linear operator in one space is defined as a matrix of linear operators on the previous space. Evaluation of a linear operator causes evaluation of linear operators in the previous space, until we get to the base space of finite-element functions, at which

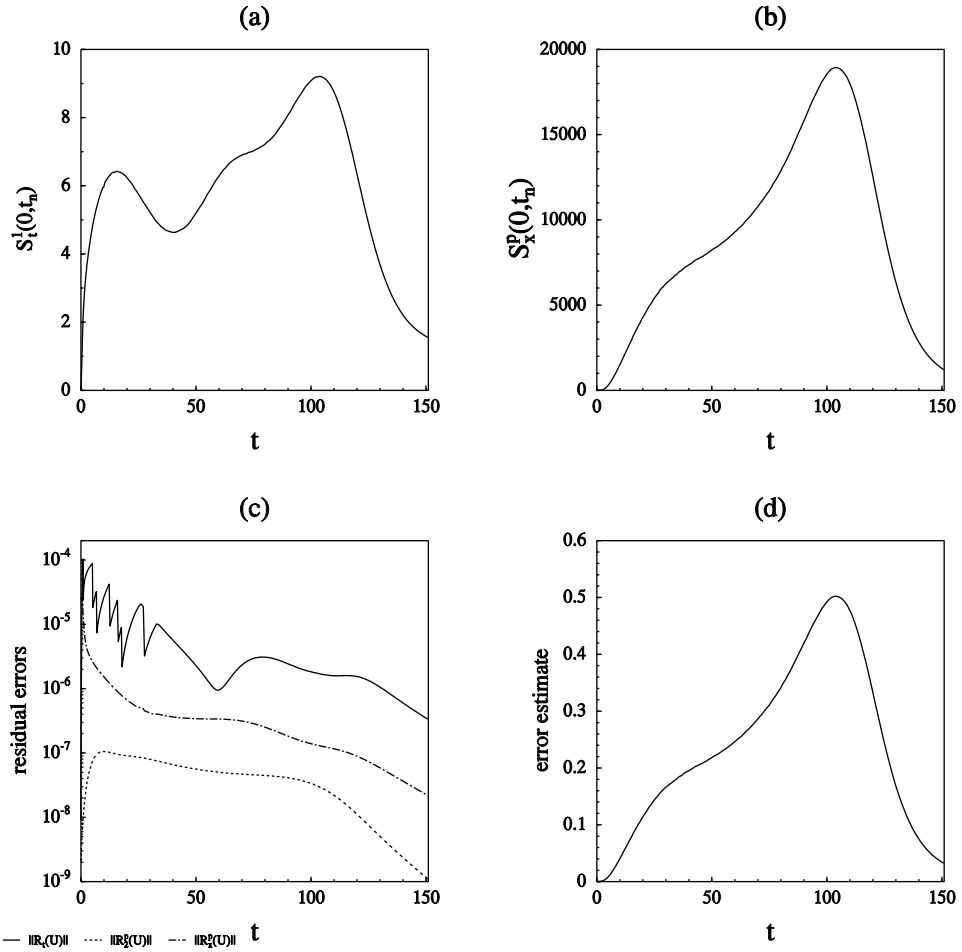

FIG. 3.21. *Model equations for morphogenesis: in (a) and (b), we plot the stability factors* S_t^1 *and* S_x^p *versus time computed every* .05 *time unit until* $t = 4$ *and then every* .5 *after that. In (c), we plot the* L_2 *norms of the residual errors versus time. In (d), we plot the a posteriori error estimate.*

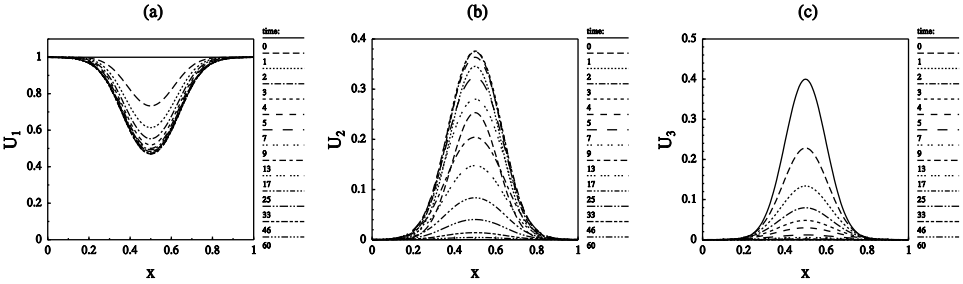

FIG. 3.22. *Model equations for the spread of rabies: plots of* $U_1(\cdot, t)$, $U_2(\cdot, t)$, *and* $U_3(\cdot, t)$ *at the indicated times.*

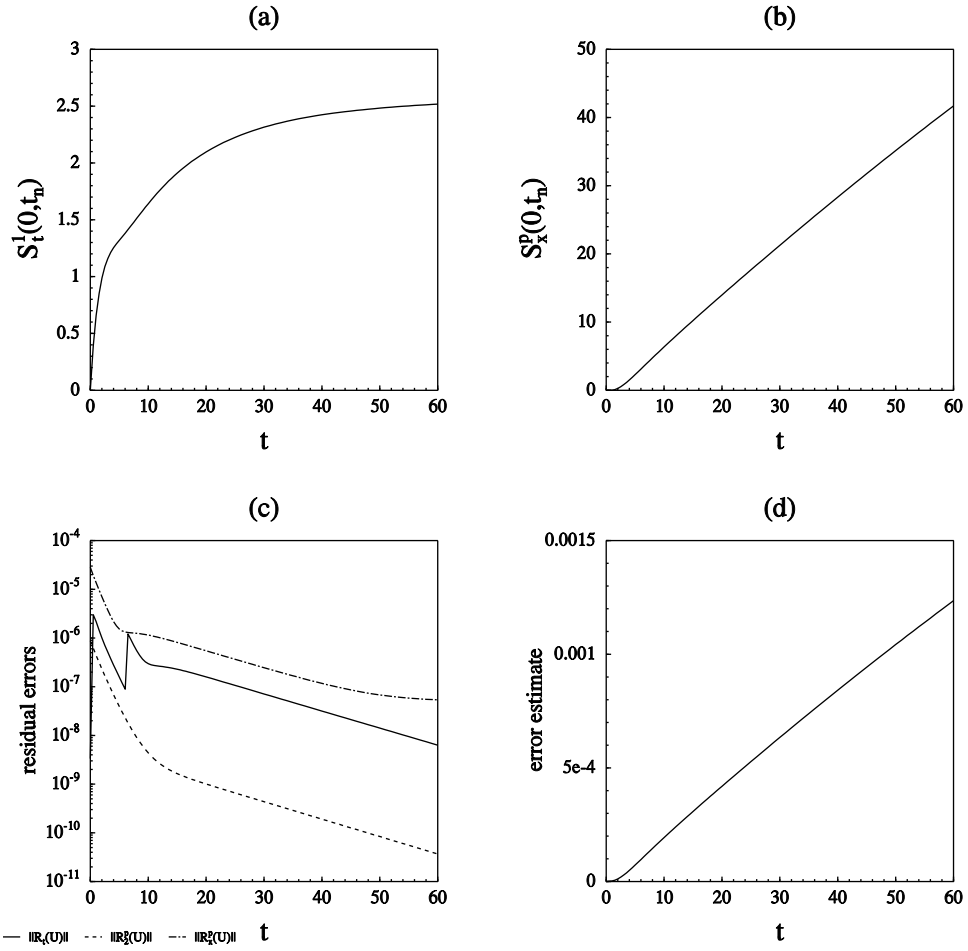

FIG. 3.23. *Model equations for the spread of rabies: in (a) and (b), we plot the stability factors* S_t^1 *and* S_x^p *versus time computed every .5 time unit until* $t = 20$ *and then every* 1 *after that. In (c), we plot the* L_2 *norms of the residual errors versus time. In (d), we plot the a posteriori error estimate.*

point the operator is defined in terms of an array.

To compute a time-step with the ODE solver, we use an operator whose dimensionality is the number of time-nodes according to which Galerkin method is being used. Evaluation of this operator causes a number of evaluations of a PDE solution operator in the spatial approximation space. This in turn requires evaluations in the finite-element space. The advantage of this architecture is that it separates the code into layers. For example, the evaluation of the discrete Laplacian is separated from the description of the PDE, which is separated from the algorithm used for time stepping. The ODE software applies with equal facility to ordinary and partial differential equations, as required by the form of (3.19). Furthermore, each software layer makes its own contribution to the error estimate.

4.1.2. Solving linear systems of equations: the QMR method and preconditioning. An implicit timestep of a discretized partial differential equation involves the repeated solution of a large, sparse linear systems. In consideration of the matrix-free approach, iterative solvers are a natural choice since the only operations involving the matrix of the system are multiplication by the linear operator (and possibly its transpose). The drawback is that iterative methods can converge very slowly and even diverge. Since a code solving reaction-diffusion equations spends most of the time solving linear systems, these issues must be considered carefully.

Iterative methods are divided according to whether the system matrix is symmetric or not and/or positive definite or not. In the case of the dG and cG discretization of (3.19), the associated matrices are both unsymmetric and non-positive definite. Reasons for this include the boundary conditions given in (3.19), the presence of convective terms in the more general form of (1.1), the competition between the reaction and diffusion, and the equations defining the approximation. In such problems for example, the classic Jacobi and Gauss-Seidel methods typically do not converge. Hence, we have employed Krylov subspace methods.

Krylov space methods can be classified according to the need to store previous vectors: The well-known GMRES (Generalized Minimum Residual) method, for example, stores all previous iteration vectors, though in practice the method is restarted to prevent running out of memory. We have chosen the QMR (Quasi-Minimal Residual) method, introduced in Freund and Nachtigal [39], which performs robustly on the systems arising during the solution of (3.19) and which does not require storing the previous iterates. *CARDS* employs both our own version and the package *QMRPACK* written by Freund and Nachtigal [40].

The convergence rate of iterative methods such as QMR is strongly dependent on the condition number of the linear operator, i.e. the ratio of the highest to lowest magnitudes of the eigenvalues of the operator. If we can find another operator that both approximates the inverse of the original system matrix and is easy to invert, then we can use it as a preconditioner to transform the original system into a new system with a smaller condition number. The sense in which a matrix P approximates the inverse of a matrix A is that the condition number of PA should be small compared to the condition number of A. The condition number of $A^{-1}A$ is one of course.

Our analysis of the size of the residual errors depends heavily on the use of nested meshes. Therefore it is natural to consider the hierarchical basis preconditioner for solving the linear systems that arise during the implicit time steps. This preconditioner, analyzed in Yserentant [65] and [66], uses the multi-scaled hierarchical finite element basis associated to nested meshes to construct a preconditioner for possibly non-symmetric elliptic problems. Its use in one and two space dimensions yields a solver with efficiency approaching that of multigrid while requiring much less regularity.

Hierarchical basis preconditioning is less efficient applied to parabolic problems, however speed-up is still possible. In practice, there is a substantial gain in systems in which the reaction term does not dominate. If the reaction term is dominate, then it yields only a small decrease in the number of iterations used by QMR and this savings is overwhelmed by the extra effort of the basis transformations required for its implementation. Thus, we made this preconditioner an option in *CARDS*. Developing better preconditioners is certainly an important area of future research.

4.1.3. Choosing time steps and space meshes. One consequence of the results in this paper is that the a posteriori theory can estimate the error of compu-

tations that are sufficiently accurate, or equivalently, for which the residual errors are sufficiently small. The scale of "sufficiently small" depends on the stability factors, whose approximation we discuss below. But no matter, given a specified *residual error tolerance* there arises the practical issue of selecting the time step and space mesh so that the residual error of the resulting approximation is bounded by the residual tolerance.

The strategy for choosing time steps and space meshes is dictated by the a posteriori nature of the bound (2.21). Thus, after the approximation is computed, the residual errors are computed, and the decision to step forward or to go back to the previous step is made. In either case, we use the size of the current residual errors to "predict" the correct time step and space mesh for the next step. Details can be found in Estep [28] and Estep and Williams [35].

There is another restriction on the choice of time steps in practice that is not reflected in the analysis of this paper, namely the convergence of the nonlinear and linear solvers used to solve the discrete equations for the approximation. Since the matrices involved have the form $I + kA$, where I is the identity, k is the time step, and A is a badly conditioned matrix, the convergence is improved if k is decreased. Likewise, the quality of the initial guess for the Newton iteration, which is computed using the values of the approximation computed on previous intervals, is improved if k is decreased. If either iterative process fails to converge, *CARDS* decreases the time step arbitrarily and tries again.

4.2. Computing the stability factors. In this section, we discuss several important issues concerning the approximation of the dual problem. The *CARDS* code numerically solves the approximate dual problem (2.13) associated to specified time nodes obtained by linearizing around the approximation rather than the average of the approximation and unknown solution as in (2.10). The numerical solution is computed using the same sequence of space meshes and the same or smaller time steps used for the forward computation, taken in reverse order of course. The hope is that meshes that are suitable to approximate the solution of the differential equation are also suitable to approximate the solution of the dual problem. In practice, this seems to work. The dual problem to the dual problem is the forward linearized problem, hence it is reasonable to expect that the time scale over which the dual problem can be solved accurately is roughly the same as the time scale over which the forward problem can be solved accurately.

4.2.1. Archiving the time history of the numerical solution. In order to form and solve the dual problem (2.13) associated to a time node t_n, we must store the approximate solution of the forward problem over the interval $[t_0, t_n]$. *CARDS* uses a dynamic archive to store these values. When each time step has been computed, the solution vector is stored in the archive, together with the time corresponding to the vector and the number of floating-point numbers in the vector. During the backward solve, when a vector is needed at some time, the archive is searched for the times that bracket the requested time, the vectors are extracted, and a linear interpolant is constructed.

The archive object is implemented by allocating large blocks of memory to store many solution vectors, together with an indexing structure that allows fast searching of the archive to find given time values. The implementation is complicated because the dimensionality of the vectors changes. However within each memory block, the dimensionality is constant to allow simple indexing of the block. Whenever the dimensionality changes, a new block is allocated.

4.2.2. Choosing data for the dual problem. The a posteriori error estimate (2.21) is a little unusual in that it is a estimate of a projection of the error e_n^- onto a specific function ϕ_n with norm one. Obviously in many cases, we would prefer to bound a norm of the error. Intuition suggests that if we compute estimates on sufficiently many projections of a function in a given inner product, we should be able to get a good estimate of the size of the function in the corresponding norm. In this section, we show that this idea can be made precise.

One way to address this issue is to answer the question:

Given a constant $\gamma : 0 \leq \gamma \leq 1$, what is the probability that

$$\gamma \|e_n^-\| \leq |(e_n^-, \Phi_n)|$$

for a randomly chosen Φ_n in (a subset of)

$$\left\{\phi \in (H_0^1(\Omega))^d \times (H^1(\Omega))^{D-d} : \|\phi\| = 1\right\}?$$

Without loss of generality, we assume that e_n^- is normalized to have norm one and compute the probability of the equivalent condition

$$|(e_n^-, \Phi_n)| \geq \gamma. \tag{4.1}$$

We first answer the analogous question for the error of a cG or dG discretization of a system of ordinary differential equations, i.e. in a setting when the "space" dimension is finite. The a posteriori error estimate in this case (see Estep [28]) estimates the Euclidean projection $|e_n^- \cdot \phi_n|$ of e_n^- onto the vector ϕ_n that is the data for the dual problem. Given a constant $\gamma : 0 \leq \gamma \leq 1$, we compute the probability that $|e_n^- \cdot \Phi_n| \geq \gamma$ for a random vector $\Phi_n \in \left\{\phi : |\phi| \leq 1,\ \phi \in \mathbf{R}^M\right\}$.

The probability can be computed using a geometric argument. In Fig. 4.1, we plot the unit M-sphere together with e_n^- represented as $(0, 0, ..., 0, 1)^\top$. The set of unit vectors in the upper half space whose projection onto e_n^- is larger than γ touch the sphere in the shaded "cap". The probability of choosing such a vector is therefore

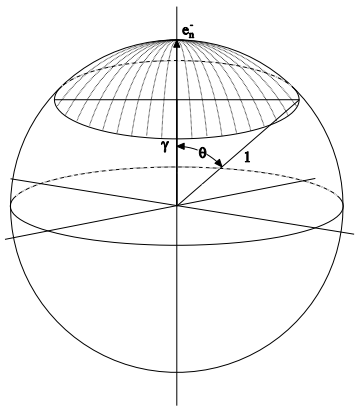

FIG. 4.1. *Drawing of the vector e_n^- and the part of the unit M-sphere containing vectors with projections on e_n^- greater than γ, for some $\gamma \leq 1$.*

$$\mathcal{P}\big(|e_n^- \cdot \Phi_n| \geq \gamma\big) = \frac{\text{area of the "cap"}}{\text{area of the upper hemisphere}} = \frac{\displaystyle\int_0^\theta \big(\sin(u)\big)^{M-1}\,du}{\displaystyle\int_0^{\pi/2} \big(\sin(u)\big)^{M-1}\,du}$$
$$= \varrho(M,\theta),$$

where θ is the angle subtended by the "cap". Thanks to J. Rauch, we have an asymptotic estimate of $\varrho\big(M,\theta\big)$ that shows it decreases like $\sin(\theta)^M$ for large M, i.e. geometrically in the dimension. From the figure, we see that $\theta = \cos^{-1}(\gamma)$. Hence, $\mathcal{P}\big(|e_n^- \cdot \Phi_n| \geq \gamma\big) = \varrho\big(M, \cos^{-1}(\gamma)\big)$ and given r random vectors $\big\{\Phi_{n,i}\big\}$ in $\big\{\phi : |\phi| \leq 1\big\}$,

$$\mathcal{P}\left(\max_{1\leq i\leq r} |e_n^- \cdot \Phi_{n,i}| \geq \gamma\right) = 1 - \big(1 - \varrho\big(M, \cos^{-1}(\gamma)\big)\big)^r. \tag{4.2}$$

We plot this function in Fig. 4.2 with (a) $M = 3$ and (b) $M = 10$.

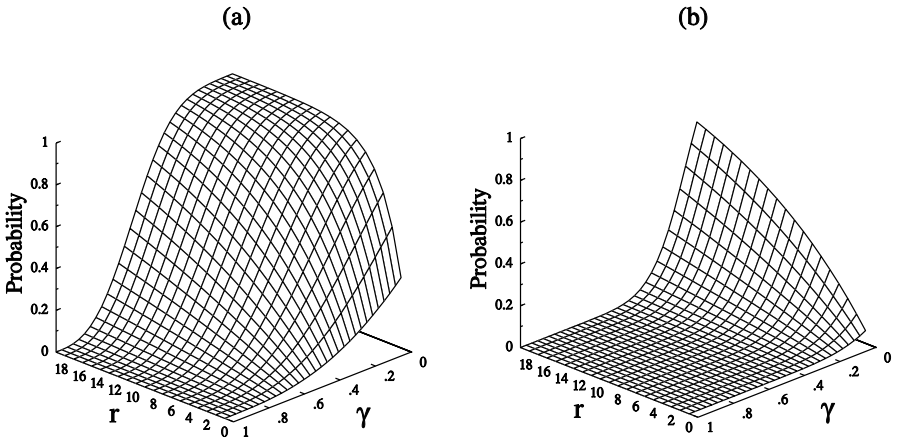

FIG. 4.2. *A plot of the probability function* $\mathcal{P}\left(\max_{1\leq i\leq r} |e_n^- \cdot \Phi_{n,i}| \geq \gamma\right)$ *for (a)* $M = 3$ *and (b)* $M = 10$.

Using (4.1), we obtain a estimate on $|e_n^-|$ from the a posteriori estimate on $|e_n^- \cdot \Phi_n|$. The factor γ^{-1} appearing on the right-hand side of the resulting estimate reflects the amount of "sloppiness" in the estimate. The formula (4.2) implies that for a specified probability, we can improve the quality of the error estimate (i.e. decrease γ^{-1}) by increasing the amount of work as measured by the number of data vectors used for computing dual solutions. Likewise for a fixed level of sloppiness, we can improve the probability that the a posteriori estimate actually bounds the norm of the error by increasing the work. In this context, we can interpret probability as being a natural measure of the reliability of the a posteriori error estimate of the norm of the error.

Returning to the original question, we use this finite dimensional result to estimate the probability that (4.1) holds for any $\gamma < 1$ in the case of a reaction-diffusion equation. For this, we use the fact that the eigenfunctions $\{\psi_i\}$ for the Dirichlet problem for the Laplacian, chosen to be orthonormal with respect to the L_2 inner product, form a basis for $H_0^1(\Omega)$ and write $e_n^- = \sum_i a_i \psi_i$. By assumption, we have $\sum_i a_i^2 = 1$. We now use the projection of e_n^- onto the span of the first M eigenfunctions

to reduce the problem to finite dimensions. We define

$$e^-_{n,M} = \sum_{i=1}^{M} a_i \phi_i$$

and likewise choose the random data Φ_n from the set $\mathcal{S}_M = \text{span}\,\{\psi_1, ..., \psi_M\}$ with $\|\Phi_n\| = 1$.

We fix $\gamma < 1$. Since $e^-_{n,M} \to e^-_n$ in L_2, for any $\epsilon : 0 < \epsilon < 1 - \gamma$, we have

$$\big|\,|(e^-_{n,M}, \Phi_n)| - |(e^-_n, \Phi_n)|\,\big| \leq \epsilon$$

for all M sufficiently large depending on ϵ. Fixing such an M, we conclude that $|(e^-_{n,M}, \Phi_n)| \geq \gamma + \epsilon$ implies that $|(e^-_n, \Phi_n)| \geq \gamma$. Note that if $\Phi_n = \sum_i b_i \psi_i$ then $(e^-_{n,M}, \Phi_n) = \sum_i a_i b_i$. Therefore, using the finite dimensional result for the case where we choose r random vectors $\{\Phi_{n,i}\}$, we conclude that

$$\begin{aligned} \mathcal{P}\left(\max_{1\leq i\leq r} |(e^-_n, \Phi_{n,i})| \geq \gamma\right) &\geq \mathcal{P}\left(\max_{1\leq i\leq r} |(e^-_{n,M}, \Phi_{n,i})| \geq \gamma + \epsilon\right) \\ &= 1 - \big(1 - \varrho\big(M, \cos^{-1}(\gamma + \epsilon)\big)\big)^r . \end{aligned} \tag{4.3}$$

Note that since e^-_n is actually in $H^{2-\epsilon}$ for $\epsilon > 0$, we can expect the coefficients of e^-_n with respect to the basis $\{\psi_i\}$ to decay relatively quickly so that M will not be large in practice.

In Fig. 4.3, we plot the values of approximate stability factors $S^1_t(t_n)$ for the Lorenz (3.11) and bistable (1.2) examples versus the time nodes t_n. In each case, we show stability factors corresponding to different choices of data for the dual problem as well as the plotting the maximum value at each time node using a darker line. This is the value that we use when estimating the error in practice.

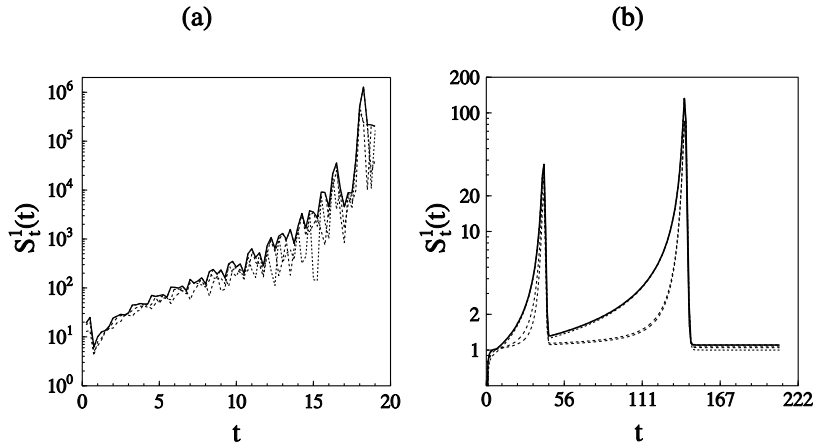

FIG. 4.3. *Plots of approximate stability factors $S^1_t(t_n)$ versus t_n for the (a) Lorenz (3.11) and (b) bistable (1.2) examples. The dark lines show the maximum value obtained at each time node.*

Remark 4.1. Another approach to obtaining estimates on $\|e^-_n\|$ from (2.21) is to try to compute an approximation of the direction of e^-_n that can be used as initial data for the dual problem and thereby getting an estimate on $|(e^-_n, \phi_n)| \approx \|e^-_n\|$. However, we

have experimented with this approach using several different approximations based on heuristic reasoning and in almost every case obtained worse results than those obtained using random initial data for the dual problem. The main problem is almost always a consistent underestimation of the error. Of course, if we do a bad job at estimating the direction of the error but do manage to obtain a partial relationship between ϕ_n and e_n^- then it is not surprising that $|(e_n^-, \phi_n)|$ could be consistently less than $\|e_n^-\|$.

One exception is an application of this a posteriori theory to the computation of optimal ground to space missile trajectories (Estep, Hodges, and Warner [30]). The differential equation describing the missile's position is posed with both initial and final states and hence can be treated as a two point boundary value problem efficiently rather than using a time-marching scheme. In this case, we obtain a good and cheap estimate for the direction of the error using Richardson extrapolation based on the global meshes.

Remark 4.2. Another natural way to address the issue of choosing data for the dual problem is to look for conditions on the problem that guarantee that the stability factors can be bounded *accurately* independent of the choice of data. This is the case for example with strongly parabolic problems in which transient behavior dies out rapidly and the stability factors grow very slowly, see Eriksson and Johnson [23], [25], and [26]. In general, we might expect this to be true for systems with dual problems that admit an exponential dichotomy or otherwise have a well-defined, discrete Lyapunov spectrum. In such problems, roughly speaking, ergodic analysis shows that almost all solutions end up behaving similarly after sufficient time has passed. In practice, this would imply that initially the dual problem should be solved with several initial data in order to use the probability result above but after a transient period, it would suffice to solve the dual problem with one initial value.

4.2.3. Unresolved: linearization and the approximate dual problem. An important issue in the computation of approximate stability factors is the fact that we cannot linearize the differential equation (2.1) around the average of the true and approximate trajectories to obtain the coefficients defined in (2.10) because the true solution is unknown. In practice, we can only solve the "approximate" dual problem with coefficients obtained by linearizing around the approximate solution.

It is possible to use a standard a priori error estimate to guarantee that the approximate and true solutions remain close for at least a short time and from this, we can obtain a priori error bounds on the approximate stability factors computed from the dual problem obtained by linearizing around the approximation. Alternatively when the true solution has sufficient regularity, we can derive the a posteriori theory by linearizing around the approximation and treating the error of linearization as a high order perturbation term. See Estep [28] for a derivation of related results in the case of ordinary differential equations. Both of these approaches lead to the conclusion that the approximate stability factors remain accurate at least over some brief initial time interval.

We don't give the details for these kinds of results because in fact they are not very relevant. First of all, the motivation for computing stability factors is to avoid the kind of overly pessimistic estimates of the growth of perturbations that the analyses described in the previous paragraph, based as they are on Gronwall arguments, yields. Secondly, this is not the correct way to think about the error of the stability factors because we do not require the same kind of accuracy for the stability factors as we

do, say, for the numerical solution itself. An estimate of the stability factors that is accurate just to order-of-magnitude can be used to get a reasonably reliable estimate of the error, by which we mean that the error should be smaller than the estimate most of the time and never significantly larger than the estimate.

We illustrate the idea that the standard of accuracy for the stability factors is different than the standard for the numerical solution using the Lorenz equations (3.11). In Fig. 4.4 (a), we plot the values of stability factors $S_t^1(t)$ versus time computed by linearizing around a variety of trajectories. The trajectories start with initial data

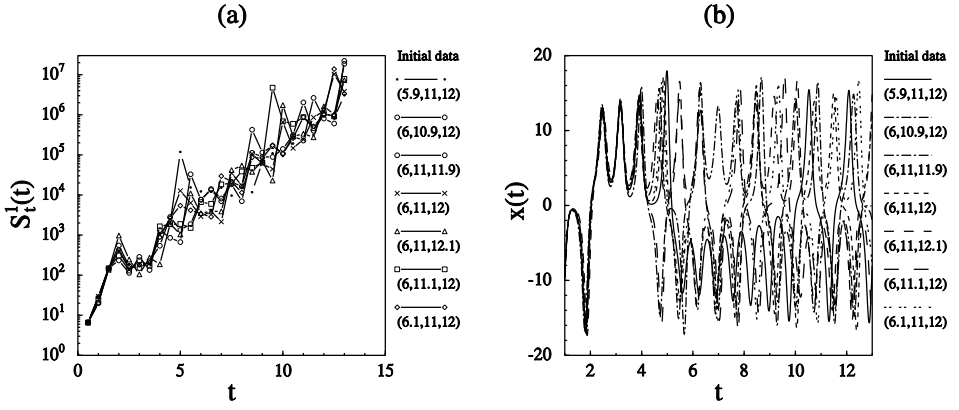

FIG. 4.4. *In (a), we plot the approximate stability factors $S_t^1(t)$ for the Lorenz equation (3.11) computed by linearizing around a variety of trajectories. In (b), we plot the corresponding coordinate $x(t)$. The key listing the initial data is on the right of each plot.*

inside the ball of radius .1 centered at $(6, 11, 12)$ close to the attractor but diverge greatly by $t \approx 5$. The divergence is demonstrated clearly in Fig. 4.4 (b). Even so, the order-of-magnitudes of the stability factors of the different trajectories grow roughly at the same rate. At any given time node, most of the stability factors could be used in the a posteriori error estimate to obtain a reliable estimate of the error of the approximate solution and they all determine the same time scale over which a given level of accuracy can be maintained.

Since the approximate stability factors are accurate initially and the computed stability factors indicate the rate at which the error of the approximate solution grows, the main concern is the rate at which the error in the approximate stability factors grows. If the error in the approximate stability factors grows at the same rate as the stability factors themselves, then we might expect the a posteriori error bound to be reliable. In other words, the degree of reliability is determined by the relative rates of growth of errors in the numerical solution due to discretization and perturbations in the solution of the linear dual problem due to perturbations in the trajectory around which the problem is linearized. Clearly, the scale for comparison depends on the degree of nonlinearity of the problem in the sense that it depends on the extent to which nearby trajectories share the same stability properties.

We illustrate the potential for differences in the sensitivity of a problem to numerical discretization and to linearization using the bistable problem (1.2). We compute approximations using fixed evenly spaced meshes with the number of elements M ranging from $M = 21$ to $M = 351$. To insure that we integrate the resulting systems of ordinary differential equations accurately, we maintain the time residual below the

tolerance .000001 and thereby keep the contribution to error stemming from the discretization in time to less than .0001. For $M \leq 50$, the numerical solutions are subject to "locking" which means that one or more metastable layers actually become stable while the correct behavior is observed for $M \geq 51$. When $M = 21$, the thinner of the two wells collapses (though at a different time than for larger M) while the wider well becomes fixed. We plot the stability factors $S_t^1(t)$ versus time for a sample of computations in Fig. 4.5. The locking phenomena is clearly reflected in the values of the

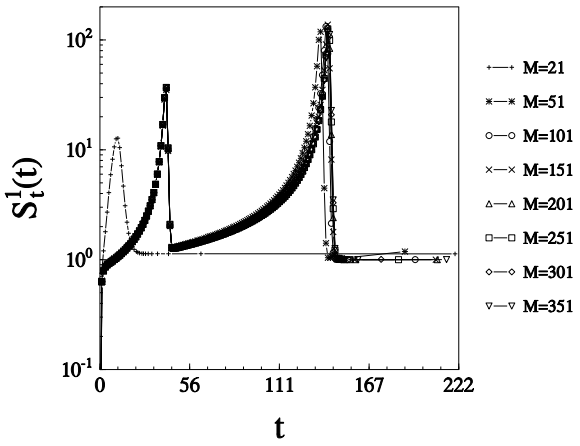

FIG. 4.5. *A plot of approximate stability factors $S_t^1(t)$ for the bistable problem (1.2) computed using trajectories computed with varying accuracy in space. The key between M and the stability factors is listed on the right.*

stability factor for $M = 21$, which remains 1 after the first well collapses indicating that the resulting pattern is stable.

Even though the numerical solutions corresponding to $M = 32$ and $M = 64$ are nearly identical to the eye, the behavior of the two is radically different. In Fig. 4.6, we plot numerical solutions for equally spaced meshes with $M = 32$ and $M = 64$ at $t \approx 5.6$ and again at $t \approx 389$. The two solutions are very close at early times but because the solution on the coarser mesh becomes locked, the numerical solutions end up quite different at later times. The bistable problem is sensitive to linearization in the neighborhood of these two approximate trajectories. While the sensitivity to linearization is not directly reflected in the a posteriori theory developed in this paper, the a posteriori error estimate estimates the error to be 2.23, i.e. more than %200, in the numerical solution with $M = 32$ elements at the time when the first well collapses. It would be difficult to trust the accuracy of the numerical solution after this point. Note that Fig. 4.5 shows that the problem is not sensitive to linearization around numerical trajectories that are sufficiently accurate. $M = 101$, $M = 201$ and $M = 351$ all produce nearly the same behavior and stability factors.

Remark 4.3. There is a plausible explanation for this sensitivity around inaccurate discretizations. It is well known that the eigenvalues of the system of ordinary differential equations arising from discretization space of of a linear constant coefficient parabolic problem by the finite element method only approximate the discrete spectrum of the parabolic problem. The smallest eigenvalue in magnitude is approximated to the same order as the solution itself. As mentioned above, the plot of the stability factors for the bistable example suggest that during a metastable phase there is

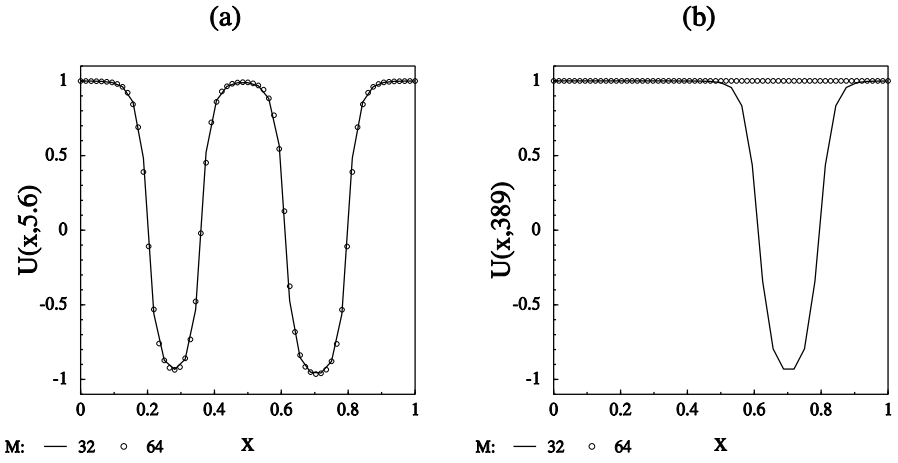

FIG. 4.6. *A plot of numerical solutions computed using equally spaced meshes with* $M = 32$ *and* $M = 64$ *at (a)* $t \approx 5.6$ *and (b)* $t \approx 389$*. Even though the numerical solutions are close initially, they end up being quite different at later times.*

one unstable Lyapunov number for the system that is initially very close to zero and which gradually increases as time passes. If this Lyapunov number is subject to the same kind of error as the eigenvalues in the linear case, then a coarse discretization might introduce enough error to change the sign of the unstable Lyapunov number and make the system stable.

We do not have an analytic method for determining the sensitivity to linearization, and in any case, it seems to be highly problem dependent. For the computations presented in this paper, we computed each example using several residual tolerances and compared the results.

4.3. Testing the accuracy of the a posteriori error estimate. We conclude this section with a numerical experiment designed to test the accuracy and reliability of the a posteriori error estimate applied to the bistable problem (1.2). We do not know the true solution but we compute a very accurate numerical solution $\overline{u}$ using the dG(1) method with $M = 513$ elements and time steps smaller than $K = .00004$ to use as a reference. We next compute an approximation U using the dG(0) method (to get the least accuracy in time) using $M = 129$ elements and time steps greater than .0004. The reference solution $\overline{u}$ is expected to be 16 times more accurate in space and at least 10 times more accurate in time. We approximate the error in U by computing $\|\overline{u} - U\|$. We plot this computed "error" together with the error estimate computed for U versus time in Fig. 4.7 (a). In Fig. 4.7 (b), we plot the ratio of the computed "error" to the estimate versus time.

Through the collapse of the first well, the a posteriori error estimate predicts the size of the error remarkably well, with the ratio of the error to the estimate hovering in the range .3–1. After this transient, the ratio remains constant but with values of size .05–.1. This over-prediction of the size of the error results from the form of the a posteriori error estimate (2.21) which estimates the error over an interval in terms of the maximum residual error in the interval. The residual errors in any interval that contains the collapse of the first well but not the second are dominated by the residual errors from the transient period of the collapse of the first well. Thus, the

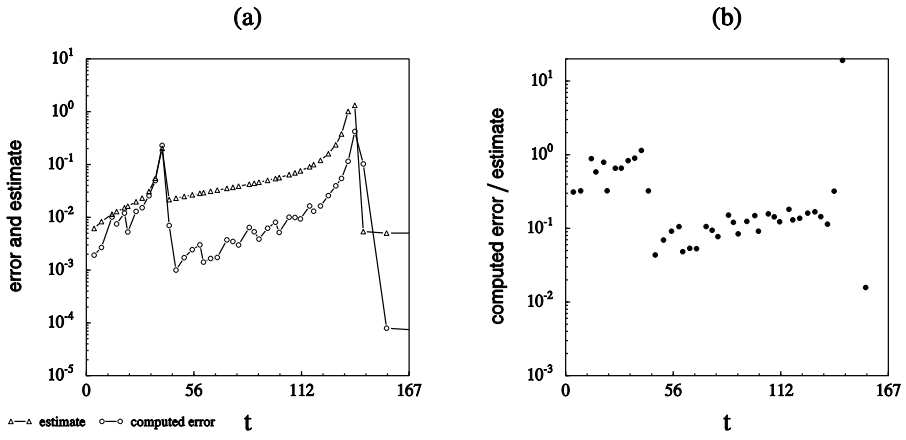

FIG. 4.7. *Results of the accuracy experiment. In (a), we plot the computed error* $||\overline{u} - U||$ *together with the a posteriori error estimate for the dG(0) approximation* U *versus time. In (b), we plot the ratio of the computed "error" to the estimate versus time.*

error estimate over-predicts the size of the error after the collapse of the first well until the last transient. Note however that the ratio still remains quite constant - which verifies that the a posteriori estimate is predicting the rate of accumulation of errors over long time intervals very well.

If we replace the maximum residual error in the a posteriori estimate (2.21) by the local residual error, the ratio of the "error" to the resulting "estimate" remains on the order of one throughout the computation. We plot the "error" together with the estimate obtained by multiplying the residual errors on the time interval $[t_{n-1}, t_n]$ by the corresponding stability factors associated to the time node t_n versus time in Fig. 4.8 (a). In Fig. 4.8 (b), we plot the ratio of the computed "error" to this modified

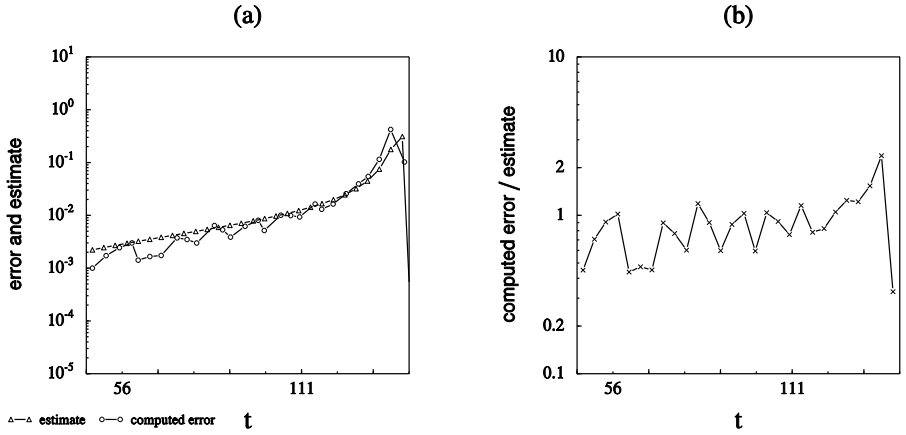

FIG. 4.8. *In (a), we plot the computed error* $||\overline{u} - U||$ *together with the estimate obtained by multiplying the residual errors on the time interval* $[t_{n-1}, t_n]$ *by the corresponding stability factors associated to the time node* t_n *versus time. In (b), we plot the ratio of the computed "error" to this quantity versus time.*

estimate versus time. Practically speaking, we had to use constant step sizes in this computation in order to align the time nodes for the two computations. The result is that the residual errors of the numerical solutions can vary greatly as time passes - as explained above. If the step sizes and space mesh are adjusted to keep the residual errors at every step approximately the same size, the tendency to overestimation due to the form of the a posteriori error estimate is greatly reduced.

Note that the a posteriori error estimate is smaller than the computed "error" at a single time step occurring immediately after the last transient. We suspect that this is due to time step discrepancies between $\overline{u}$ and U. In any case, the discussion in Section 4.2.2 explains that there is some probability that this can occur. We computed stability factors using three different initial guesses for the dual problem and obtained very close values in each case.

5. Improving stability by preserving invariant rectangles under discretization.

The discussion so far has centered on general problems of the form (2.1) under mild assumptions about the stability; really no more than necessary to guarantee that the problem is well-posed in a convenient Sobolev space over a short time interval. The results we have presented reflect this. For one thing, the estimates on the residuals in Theorem 3.4 guarantee that the residual errors on any step can be made small by refinement, but both the constants and the discrete derivatives of U in the bounding quantities can vary with the time interval. Likewise, the stability factors on any given interval are finite, but the sequence of stability factors associated to a progressive sequence of time nodes can grow super-exponentially. This prevents us from concluding, for example, that the error of the approximation becomes smaller if the residual errors are made smaller. Recall Remark 3.4.

But this is no more than can be expected, because the assumptions we have made so far even allow finite time "blow-up". Consider the initial value problem $u_t = u^2$ and $u(0) = 1$ with solution $u(t) = (1+t)^{-1}$ for example. We plot the stability factor $S_t^1(t_n)$ for a sequence of times $.01 \ldots .95$ in Fig. 5.1 together with the sequence of time steps used in the computation. In this example, the sequence of stability factors associated to the time nodes grows super-exponentially and the time steps have to be decreased correspondingly in order to keep the residual errors uniformly bounded.

One way to get stronger results is to consider problems with special stability properties that prevent such wild behavior. For example, there is the class of strongly dissipative, e.g. strongly diffusion-dominated, problems considered by Eriksson and Johnson ([22], [25]). This class is a natural progression from linear problems since they share much of the same behavior. The results for these equations are strong: the constants in the a posteriori estimate are uniform in time and the stability factors for a progressive sequence of time nodes can be shown to grow only logarithmically with time. Consequently for such problems, the error is controlled directly by controlling the residual errors, and it is possible to control the error over very long time intervals. On the other hand, the possible behavior of solutions of such problems is relatively limited and none of the examples discussion in the introduction satisfy the strong assumptions with physically relevant values for the diffusion.

It is therefore interesting to consider parabolic equations under assumptions that prevent unbounded growth in solutions but do not force all the solutions to converge rapidly to a fixed point. We apply the general theory developed in Section 2 to the

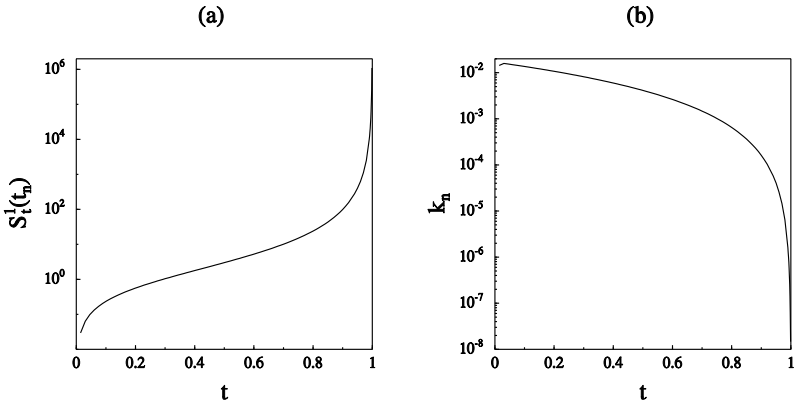

FIG. 5.1. *In (a), we plot $S_t^1(t_n)$ for $u_t = u^2$ and $u(0) = 1$ on a log scale at a sequence of time nodes .01....95. The same values for the stability factor are given by a range of tolerances. In (b), we plot the sequence of time steps on a log scale required to keep the residual errors bounded by a fixed tolerance.*

system of reaction-diffusion with constant diffusion (3.19) under an condition that guarantees the existence of an *invariant rectangle*, which is a generalized rectangle $\mathcal{R} = \prod_i [a_i, b_i]$ in $\mathbf{R}^D$ centered at the point P with sides parallel to the coordinate axes inside of which any solution of (3.19) remains as long as it exists. Namely, we assume that the reaction f satisfies

$$n_{\partial\mathcal{R}}(u) \cdot f(u, \cdot, \cdot) \leq 0 \text{ for } u \in \partial\mathcal{R}, \tag{5.1}$$

where $n_{\partial\mathcal{R}}(u)$ is the outward unit normal to $\partial\mathcal{R}$ at u. See Fig. 5.2. This guarantees that $\mathcal{R}$ is invariant for (3.19), see Chueh, Conway, and Smoller [12] and Smoller [62].

While more specialized than the general system (2.1), systems of the form (3.19) under assumption (5.1) model many interesting physical situations. For example:

1. The bistable equation admits the invariant rectangle $[-\alpha, \alpha]$ for any $\alpha \geq 1$.
2. If we choose $M = -(u_1 - \alpha_1)(u_1 - 1) - \alpha_2 u_2$ and $N = -\alpha_3 - \alpha_4 u_2 + \alpha_2 u_1$ with $0 < \alpha < 1$ and $\alpha_1 \leq \alpha_3/\alpha_2 < 1$ in the predator-prey model, then the equations admit arbitrarily large invariant rectangles.
3. The Hodgkin-Huxley equations admit arbitrarily large invariant rectangles.
4. The Fitz-Hugh-Nagumo equations admit arbitrarily large invariant rectangles.
5. If the ϵ_i are all equal in the equations modelling the superconductivity of liquids, then any rectangle containing the unit circle $|u| = 1$ is invariant.
6. If we choose $r_1 > \max\{1, \alpha_2^{-1}\}$, $r_2 > r_1$, and $r_3 > \alpha_3 r_2$ in the Field-Noyes equations, then the rectangle $\mathcal{R} = \{(u_1, u_2, u_3) : 0 \leq u_1 \leq r_1, 0 \leq u_2 \leq r_2, 0 \leq u_3 \leq r_3\}$ is invariant.
7. If we choose the initial data so that $0 < u_1(x, 0) < r_1$ and $0 < r_2 \leq u_2(x, 0)$ for all x in the flame model in Example 7, then the region $\mathcal{R} = \{(u_1, u_2) : 0 \leq u_1 \leq r_1, r_2 \leq u_2\}$ is invariant.

Perhaps the greatest benefit arising from the existence of an invariant rectangle is the resulting compactness of the set of solutions. First of all, the existence of an invariant rectangle means that the solution (3.19) exists uniquely for all time. This follows from the local existence result by a standard bootstrap argument that uses the

fact that we can replace local bounds on $|f|$ and the Lipschitz constant of f by bounds over $\mathcal{R}$. Below, we let M denote the maximum of f on $\mathcal{R}$, and L the maximum of the first and second order partial derivatives of f on $\mathcal{R}$. An analogous argument also applies to the approximation methods to give global existence of the approximants.

The existence of an invariant rectangle for u also means that ∇u and Δu are continuous functions for smooth initial data in $\mathcal{R}$. This follows by differentiating (3.19) with respect to x to obtain linear equations with smooth, bounded coefficients for the various partial derivatives of u. Also, the constants L_n in the energy estimates on u in Proposition 3.5 can be set to L so the estimates (3.20)– (3.22) hold uniformly in the data. Likewise, if there is an invariant rectangle $\mathcal{R}$ for the approximation U then the constants M_n and L_n in Proposition 3.6 can be set to M and L respectively and the estimates on U in that result hold uniformly in the data. This in turn implies that the constants in the estimates on the residual errors (3.4) can be taken uniform with respect to the time interval.

We can also draw some strong conclusions about the stability factors if there is an invariant rectangle for both u and U. In this case, we can take $L_n \equiv L$ in Proposition 3.7 and get bounds on the stability factors that hold uniformly for all approximations in the invariant rectangle. This means that the stability factors associated to a progressive sequence of time nodes grows at most exponentially quickly, regardless of the approximations used to compute the stability factors. Likewise, we can numerically solve the dual problems associated to different approximations and compute approximate stability factors uniformly well.

One consequence of these results is that it is relatively easy to use the a posteriori estimate on a projection of the error (2.21) to get a estimate on the norm of the error. As mentioned in Remark 2.3, we can obtain such a estimate by making a special choice of data for the dual problem, for example $\phi_n = e_n^-/\|e_n^-\|$ for the dual problem at t_n. But, recalling (3.28), choosing the initial data for the dual problem depending on the error means that the size of stability factors associated to t_n depends on the size of the error and its gradient t_n. Since we are trying to estimate the size of the error at t_n, we therefore require a bound on the stability factors at t_n in the case $\phi_n = e_n^-/\|e_n^-\|$ that is independent of e_n^-.

As long as u and U remain inside some rectangle $\mathcal{R}$, $\|e\|$ remains uniformly bounded. To control $\|\nabla e^p\|$, we use the stability estimates in the following proposition, proved in Section 8.

PROPOSITION 5.1. *Under the assumptions of Propositions 3.5 and 3.6 and assuming in addition that u and U remain in $\mathcal{R}_o$ and only a finite number of mesh coarsenings is allowed on any fixed time interval, there is a constant $C = C(\epsilon, M)$ such that*

$$\|\nabla u^p(t)\| \leq \|\nabla u^p(0)\| + Ct^{1/2}, \qquad t \geq 0, \tag{5.2}$$

$$\|\nabla U_n^{p_i -}\| \leq C\|\nabla U_0^{p_i -}\| + Ct_n^{1/2}, \qquad t_n \geq 0. \tag{5.3}$$

By these estimates and (7.2), we can bound $\|\nabla e^p\|$ at later times in terms of the regularity of U^p at the initial time. More precisely, there is a constant $C = C(\epsilon, M)$ such that

$$\|\nabla e_n^{p_i -}\| \leq C\|h_1 \Delta_{h_1} U_0^{p_i -}\| + \|\nabla U_{n-1}^{p_i -}\| + Ct_n^{1/2}, \quad n \geq 0.$$

We restate (3.28) as there are constants $C_1 = C_1(\epsilon, M, L)$ and $C_2 = C_2(L)$ such that

$$\max\{S_0(0,t_n), S_t^1(0,t_n), S_x^p(0,t_n), S_x^o(0,t_n)\} \leq C_1 e^{C_2 t_n}\left(\|h_1 \Delta_{h_1} U_0^{p,-}\| + \|\nabla U_0^{p,-}\|\right), \quad n \geq 0. \quad (5.4)$$

Remark 5.1. It is useful to distinguish the time dependence of the bounds on the derivatives of the solutions in Proposition 5.1 and Proposition 8.1 below from the time dependence of the stability factors in the a posteriori error estimate (2.21). The stability factors reflect both the possible growth in the size of derivatives of the solution, like (5.2) and (5.3), and the possible accumulation of errors.

We turn now to study conditions which guarantee the existence of an invariant rectangle for the Galerkin approximations. When considering the preservation of an invariant rectangle under discretization, at least two possibilities come to mind: (1) the approximation properties of the numerical method imply that there is an invariant rectangle for the approximation that is close to an invariant rectangle for the true solution; (2) the approximation method has the special stability property that any rectangle on which (5.1) holds is also invariant for the approximation. We obtain results for each situation. In the first result, we derive a condition on the size of the residual error that guarantees an "approximate" invariant region exists when the vector field has the property that it points inwards with a minimum angle on the boundaries of a family of concentric rectangles. In the second result, we modify the Galerkin methods and show that the resulting approximations have the same invariant region as the solution if the time steps are sufficiently small.

Remark 5.2. All the results in this section hold for problems with homogeneous Neumann boundary conditions and nonconstant diffusion and many of the results extend to problems with nonlinear diffusion and even gradient terms provided the equations are 'diffusion-dominated'.

5.1. Preservation of a "fuzzy" invariant rectangle. The rough idea is to use the a posteriori error estimate (2.21) to keep the approximation of a solution starting with data inside an invariant rectangle $\mathcal{R}$ to within a given distance (tolerance) ρ to the solution. As long as this is possible, then the approximation remains inside a rectangle with boundary a distance ρ away from the boundary of $\mathcal{R}$.

There are two difficulties with this approach. First, the estimates on the residual errors that guarantee that residual errors, and hence the error, can be made small by refining the time step and space mesh, depends on comparing the approximation to a local solution of the differential equation started with data obtained by smoothing the approximation. Since the approximation may be outside $\mathcal{R}$ or near its boundary, this local solution may also start with data outside the invariant rectangle. This difficulty can be overcome after observing that all of the examples above admit a *family* of concentric invariant rectangles filling a region of space. By keeping the distance between a solution remaining in a relatively small rectangle and the corresponding approximation sufficiently small, we can hope to keep the approximation and any associated local solutions inside a relatively larger invariant rectangle.

The second difficulty is that the a posteriori estimate on the error may grow exponentially quickly with time, if the bounds in Lemma 3.7 are the only information we have about the size of the stability factors. To keep the distance between a solution remaining inside an invariant rectangle and the corresponding approximation below a uniform bound, we obtain a time-dependent condition on the size of the residual

errors, and moreover, a condition that decreases exponentially quickly as time passes. The first result we present shows that under the appropriate assumptions, there is a *time-independent* condition on the size of the residual errors that guarantees the approximation remains inside an approximate invariant rectangle for all time.

More precisely, we assume the existence of inner and outer rectangles $\mathcal{R}_i \subset \mathcal{R}_o$ with sides parallel to the coordinate axes and centered at the same point P, containing the origin, and with $r := \text{dist}\,(\partial\mathcal{R}_i, \partial\mathcal{R}_o) > 0$. We let l_i denote the length of the smallest side of $\mathcal{R}_i$, $M = \max_{\mathcal{R}_o} |f|$, and L denote the maximum of the first and second order partial derivatives of f on $\mathcal{R}_o$. We also assume that there is a $\gamma > 0$ such that

$$f(v) \cdot n_{\partial\mathcal{R}_\rho}(v) \leq -\gamma M < 0, \quad v \in \partial\mathcal{R}_\rho, \tag{5.5}$$

for all rectangles $\mathcal{R}_i \subset \mathcal{R}_\rho \subset \mathcal{R}_o$ with sides parallel to the coordinate axes and centered at P and $\text{dist}\,(\partial\mathcal{R}_\rho, \partial\mathcal{R}_i) = \rho$, where $n_{\partial\mathcal{R}_\rho}(u)$ denotes the outward pointing unit normal to $\partial\mathcal{R}_\rho$ at the point $u \in \partial\mathcal{R}_\rho$. See Fig. 5.2. Roughly speaking, we prove

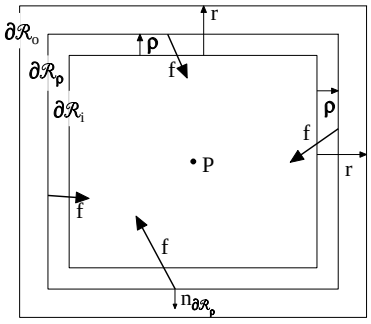

FIG. 5.2. *The invariant rectangles* $\mathcal{R}_i \subset \mathcal{R}_\rho \subset \mathcal{R}_o$.

that U remains in $\mathcal{R}_\rho$ for all ρ sufficiently small provided that U_0^- is in $\mathcal{R}_i$ and U is computed so that the residual errors over each time step are kept smaller than a fixed fraction of ρ.

The proof hinges on the fact that a local solution $\tilde{u}$ of the differential equation in (3.19) for $t > t_{n-1}$ starting with initial data $\tilde{u}_{n-1} = \tilde{u}(t_{n-1})$ in the outer rectangle $\mathcal{R}_o$ must enter the inner rectangle $\mathcal{R}_i$ after a finite time. The reason is conceptually easy to understand in the case of a scalar ordinary differential equation, $\dot{\tilde{u}} = f(\tilde{u})$, where $\mathcal{R}_o = [a_o, b_o]$ and $\mathcal{R}_i = [a_i, b_i]$ with $a_o < a_i < b_i < b_o$. On $[b_i, b_o]$, $f(\cdot) < -\gamma M$, hence integration gives

$$\tilde{u}(t) - b_i \leq \tilde{u}(t_{n-1}) - b_i - \gamma M\,(t - t_{n-1})$$

as long as $\tilde{u}$ remains in $[b_i, b_o]$. We conclude that $\tilde{u}$ enters $\mathcal{R}_i$ for $t - t_{n-1}$ sufficiently large depending on γ, M, and the size of the invariant intervals.

The case of a scalar parabolic partial differential equation, $\dot{\tilde{u}} = \Delta\tilde{u} + f(u)$, is more complicated because the diffusion term does not necessarily have the correct sign to force $\tilde{u}$ to decrease at every point even if $\tilde{u}$ is partly in $\mathcal{R}_o \backslash \mathcal{R}_i$. In other words, $\tilde{u}$ may be increasing in some parts of Ω even while decreasing in other parts. We illustrate in Fig. 5.3. If we consider the point $\bar{x}(t)$ at which $\tilde{u}$ obtains a positive maximum at time t, which must be in the interior of Ω because of the Dirichlet boundary conditions, then

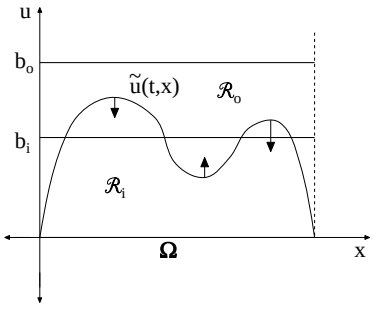

FIG. 5.3. *One difficulty in showing that a solution of a parabolic equation must enter the inner rectangle $\mathcal{R}_i$ when the vector field points inwards in $\mathcal{R}_o \backslash \mathcal{R}_i$.*

$\tilde{u}_{xx}(t, \bar{x}(t)) \leq 0$ and therefore $\dot{\tilde{u}}(t, \bar{x}(t)) \leq f(\tilde{u}(t, \bar{x}(t))$ and apparently the argument used for an ordinary differential equation would apply. Unfortunately, the fact that $\bar{x}(t)$ is not a continuous function of t causes difficulties with this argument. Finally, additional complications ensue when considering systems of equations. To overcome these, we follow Smoller [62] and deal with a functional measuring the size of the smallest rectangle that contains a solution that decreases as time passes.

We first define a norm on $\mathbf{R}^D$ associated to $\mathcal{R}_i$ by

$$|w|_{\mathcal{R}_i} = \inf\{t \geq 0 : w \in t\mathcal{R}_i\}.$$

In other words, $|w|_{\mathcal{R}_i}$ is the smallest multiple of $\mathcal{R}_i$ that contains the point w. We define a continuous functional $F_{\mathcal{R}_i}(\cdot)$ on the bounded continuous functions on Ω by

$$F_{\mathcal{R}_i}(w) = \sup_{x \in \Omega} |w(x)|_{\mathcal{R}_i}.$$

We also recall the definition of the *upper Dini derivative* of a function g,

$$\bar{D}g(t) = \overline{\lim_{k \to 0}} \frac{g(t+k) - g(t)}{k}.$$

The following proposition is a extension of Theorem 14.19 of Smoller [62],

PROPOSITION 5.2. *Suppose that $\tilde{u}$ is a smooth solution of (3.19) with $F_{\mathcal{R}_i}(\tilde{u}(t_{n-1}, \cdot) > 1$ and f satisfies the minimum angle assumption (5.5). Then for any time $t_{n-1} \leq t$ with $F_{\mathcal{R}_i}(\tilde{u}(t, \cdot)) > 1$,*

$$\bar{D}F_{\mathcal{R}_i}(\tilde{u}(t, \cdot)) \leq -\frac{2\gamma M}{l_i}.$$

Smoller gives a proof for systems of equations with one space variable posed on $\mathbf{R}$, but the extension to the present case is straightforward. This result implies that $\tilde{u}$ must enter a smaller rectangle immediately after any time t with $F_{\mathcal{R}_i}(\tilde{u}(t, \cdot)) > 1$. In fact, it is not difficult to show that if

$$\bar{D}F_{\mathcal{R}_i}(\tilde{u}(t, \cdot)) \leq -\frac{2\gamma M}{l_i} \quad \text{for } t_{n-1} \leq t \leq t^*$$

then

$$F_{\mathcal{R}_i}(\tilde{u}(t, \cdot)) \leq F_{\mathcal{R}_i}(\tilde{u}(t_{n-1}, \cdot)) - \frac{2\gamma M}{l_i}(t - t_{n-1}) \quad \text{for } t_{n-1} \leq t \leq t^*. \tag{5.6}$$

This result applies to all of the examples above with invariant regions. For example in the case of the bistable equation, we choose $1 < r_i$ and $2\sqrt{3}/3 \leq r_o$, then any interval $[-\rho, \rho]$ with $r_i \leq \rho \leq r_o$ is invariant and f satisfies (5.5) with

$$\gamma = \left| \frac{r_i - r_i^3}{r_o - r_o^3} \right|.$$

We prove that U preserves an invariant region by using the a posteriori error estimate to control the error between U and local solutions $\tilde{u}$. There are two cases to treat. If U_{n-1}^- is contained in $\mathcal{R}_i$, then we control the residual error on the next step to a sufficient degree to keep U_n^- from going too far outside $\mathcal{R}_i$. If U_{n-1}^- is outside $\mathcal{R}_i$, then we control the residual errors on subsequent steps until the local solution $\tilde{u}$ starting at time t_{n-1} enters $\mathcal{R}_i$, drawing U along with it. These two cases lead to different conditions on the residual errors.

The proof we give uses an energy estimate on the growth of the discrete second derivative $\Delta_h U$. It appears to be technically difficult to obtain such an estimate on the part of U solving an ordinary differential equation, hence we restrict the problem (2.1) to the case $d = D$.

THEOREM 5.3. *Assume that U_0^- is contained in $\mathcal{R}_i$ and $\rho \leq r/5$. Then there is a constant $C = C(\epsilon, M, L)$ such that $U_n^- \in \mathcal{R}_{4\rho}$ for all $n \geq 0$ and $x \in \Omega$ provided that U_n^-, $n \geq 1$, is computed so that the following conditions are met:*

1. *If U_{n-1}^- is contained in $\mathcal{R}_i$, then the mesh and time steps for the next interval should be chosen so that*

$$C\| \log(h_n) h_n^{3/2} \Delta_{h_n} U_{n-1}^- \| \leq \rho \tag{5.7}$$

and in addition

$$\begin{aligned} C\big(1 + \|h_n \Delta_{h_n} U_{n-1}^-\| + \|\nabla U_{n-1}^-\|\big)\big(&\|h_n^{-1/2} k_n R_t(U)\|_{L_\infty(I_n)} \\ &+ \|h_n^{3/2} R_x^p(U)\|_{L_\infty(I_n)} + \|h_n^{3/2} R_2^p(U)\|_{L_\infty(I_n)} \\ &+ \|h_n^{3/2} \Delta_{h_n} U_{n-1}^-\| + \|h_n^{3/2} \nabla U_{n-1}^-\|\big) \leq e^{-CK}\rho. \end{aligned} \tag{5.8}$$

This guarantees that U_n^- is contained in $\mathcal{R}_{2\rho}$.

2. *At any time node t_{n-1} at which U_{n-1}^- is contained in $\mathcal{R}_{2\rho}$, but not contained in $\mathcal{R}_i$, a new mesh and time step should be chosen so that*

$$C\big(\| \log(h_n) h_n^{3/2} \Delta_{h_n} U_{n-1}^- \| + \| \log(h_n) h_n^{3/2} \nabla U_{n-1}^- \|\big) \leq e^{-C(\tau+K)}\rho \tag{5.9}$$

and

$$\begin{aligned} C\big(1 + \|h_n \Delta_{h_n} U_{n-1}^{p,-}\| + \|\nabla U_{n-1}^{p,-}\|\big)\big(&\|(h_n^{3/2} + h_n^{-1/2} k_n + h_n^{-5/2} k_n^2) \Delta_{h_n} U_{n-1}^-\| \\ &+ \|(h_n^{3/2} + h_n^{-1/2} k_n + h_n^{-5/2} k_n^2) \nabla U_{n-1}^-\| + \|h_n^{3/2} + h_n^{-1/2} k_n + h_n^{-5/2} k_n^2\|\big) \\ &\leq e^{-C(\tau+K)}\rho, \end{aligned} \tag{5.10}$$

where

$$\tau = \frac{3 l_i}{2\gamma M}\rho,$$

and U_m^- should be computed using this mesh size and possibly smaller time steps on all subsequent steps $m \geq n$ until U_m^- is contained in $\mathcal{R}_{2\rho}$ once again. The maximum number of steps needed to reach time node t_{m^*} with $U_{m^*}^-$ contained in $\mathcal{R}_{2\rho}$ is finite and depends only on τ, ρ, M, L, and U_{n-1}^-.

The conditions (5.7)–(5.8) and (5.9)–(5.10) respectively are satisfied on all sufficiently refined time steps and space meshes.
We prove this in Section 8.

Remark 5.3. A stronger result would be that there is a maximum time step and mesh size depending on L, M, and ρ, such that U_n^- remains in $\mathcal{R}_{4\rho}$ for all time. But the estimates we use to control the error depend on the discrete derivatives of U in the residual errors and these discrete derivatives of U can grow even when U is confined to an invariant rectangle.

Remark 5.4. The preservation of an invariant rectangle is a pointwise property while the a posteriori theory in Section 2 is developed for $L_\infty(L_2)$. In converting from L_∞ to L_2 in x, there is a loss of order. For example, there is a constant C independent of h such that

$$\big\|\log(h_n)h_n^2\Delta_{h_n}U_{n-1}^-\|_{L_\infty(\Omega)} \le C\big\|\log(h_n)h_n^{3/2}\Delta_{h_n}U_{n-1}^-\|.$$

The loss of order is apparent when comparing quantities in (5.8) and (5.7) to the corresponding expressions in Theorem 2.1.

Remark 5.5. This shows that the nodal values of U remain inside $\mathcal{R}_{4\rho}$ for all time. Using an interior estimate on the error related to (2.23), it follows that U is contained in $\mathcal{R}_{5\rho}$ for all $t \ge 0$ and $x \in \Omega$.

The requirement that the tolerance depend on the "width" of the region on which the reaction term points inwards is natural as we illustrate with the initial value problem

$$\begin{cases} u_t = f(u) = 2\big(1.01 - u^2\big)\big(1-u^2\big)u\,e^{-.5u^2}, & t > 0, \\ u(0) = .5. \end{cases} \tag{5.11}$$

We plot $f(u)$ for $-5 \le u \le 5$ in Fig. 5.4(a). This problem has the invariant region $\big[-\sqrt{1.01}, \sqrt{1.01}\big]$ since f points inwards for $1 < |u| < \sqrt{1.01}$. The minimum value of f in $\big(1, \sqrt{1.01}\big)$ is approximately -3.03×10^{-5} and the width of the interval is approximately .005. For $|u| < 1$ and $|u| > \sqrt{1.01}$, f points outwards. In particular, solutions starting with initial values larger than $\sqrt{1.01}$ tend to infinity. In Fig. 5.4(b), we plot numerical solutions starting with initial value .5 computed while keeping the residual errors below .09 and .001 respectively. The less accurate computation steps outside the invariant region at one point and subsequently grows without bound after that.

5.2. Exact preservation of an invariant rectangle. In the previous section, we used the accuracy of the approximation method to show that it preserves an approximate invariant region under suitable assumptions. Now we consider finite element methods that have the property that (5.1) guarantees that $\mathcal{R}$ is also invariant for the approximation. This is not a universal property and it is related to the issue of whether or not a finite element approximation satisfies a maximum principle when the method is applied to the heat equation. The dG and cG methods using the standard Galerkin finite element discretization in space analyzed above do not, see Thomée [63]. A typical manifestation of this are small oscillations in the approximation near the transition to steep layers.

Consequently, we modify the dG and cG methods by using the lumped mass quadrature to evaluate the space integrals in the variational formulation and show that

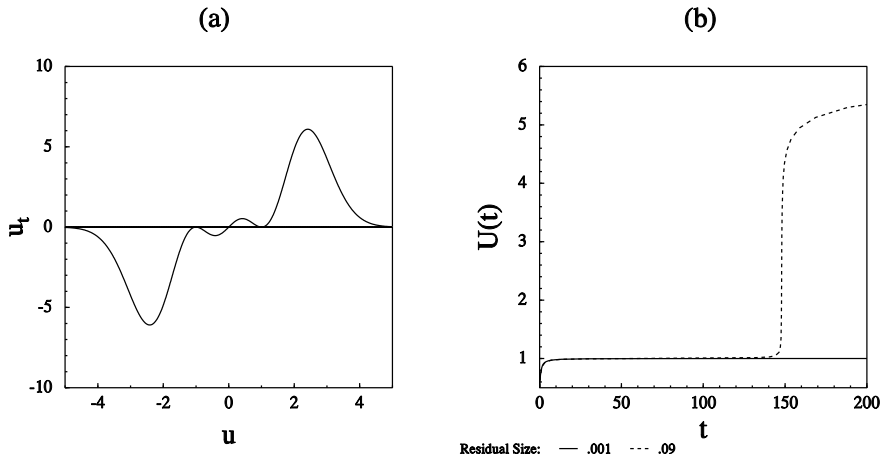

FIG. 5.4. *(a) Plot of $f(u)$ for the problem in (5.11). (b) Plots of numerical solutions of (5.11) starting with initial value .5 computed by keeping the residual errors below .09 and .001 respectively.*

the resulting methods have the desired invariance. We also outline the modifications to the general a posteriori theory needed to cover the new methods, including the analysis of the quadrature error. This introduces some technical difficulties because relatively high order derivatives of the solution are involved and we consider systems consisting only of parabolic equations. This issue is also interesting from the standpoint that the standard Galerkin finite element method using lumped mass quadrature is closely related to the standard five point difference scheme. Further discussion of the lumped mass quadrature can be found in Raviart [60], Chen and Thomée [11], and Thomée [63].

The *lumped mass quadrature formula* is a composite rule computed by using the two dimensional trapezoidal rule on each triangle. For a triangle $K \in \mathcal{T}_n$, we let $a_{K,j}$, $j = 1, 2, 3$, denote the vertices and define the quadrature formula on K as

$$\int_K g(x)\,dx \approx \mathcal{Q}_K(g) = \frac{1}{3}\text{area}\,(K)\sum_{j=1}^{3} g(a_{K,j}).$$

Note that

$$\mathcal{Q}_K(g) = \int_K Q_n(g)\,dx,$$

and the quadrature rule has precision 1. We define the discrete inner product and norm as

$$(g_1, g_2)_{h_n} = \sum_{K\in\mathcal{T}_n} \mathcal{Q}_K(g_1, g_2) \text{ and } \|g\|_{h_n}^2 = (g, g)_{h_n}.$$

The expression "lumped mass" refers to the fact that the mass matrix associated to $(\ ,\ )_h$ is a diagonal matrix with the diagonal entry in each row equal to the sum of the entries in the corresponding row in the standard mass matrix B_n.

The *continuous Galerkin-lumped mass* cGL(q) approximation $U \in W^q$ satisfies

$U_0^- = Q_0 u_0$ and for $n \geq 1$, the *Galerkin orthogonality relation*

$$\begin{cases} \displaystyle\int_{t_{n-1}}^{t_n} \left((\dot{U}_i, V_i)_{h_n} + \epsilon_i(\nabla U_i, \nabla V_i) \right) dt = \int_{t_{n-1}}^{t_n} (f_i(U), V_i)_{h_n}\, dt \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{for all } V \in W_n^{q-1},\ 1 \leq i \leq D \\ U_{n-1}^+ = Q_n U_{n-1}^-, \end{cases} \tag{5.12}$$

and the *discontinuous Galerkin-lumped mass* dGL(q) approximation $U \in W^q$ satisfies $U_0^- = Q_0 u_0$ and for $n \geq 1$,

$$\int_{t_{n-1}}^{t_n} \left((\dot{U}_i, V_i)_{h_n} + \epsilon_i(\nabla U_i, \nabla V_i) \right) dt + \left([U_i]_{n-1}, V_i^+ \right)_{h_n} = \int_{t_{n-1}}^{t_n} (f_i(U), V_i)_{h_n}\, dt$$
$$\text{for all } V \in W_n^q,\ 1 \leq i \leq D. \tag{5.13}$$

We write out the discrete equations in matrix-vector form for the cGL method in (8.10).

The derivation of the error representation formula begins in the same way as for the standard Galerkin methods. With the same definition of the dual problem (2.13), we once again obtain (2.14):

$$(e_i^-(t_n), \phi_{n,i}) = (e_i^+(0), \phi_i(0)) - \int_0^{t_n} \left((\dot{U}_i, \phi_i) + (\epsilon_i(U)\nabla U_i, \nabla \phi_i) - (f_i(U), \phi_i) \right) dt.$$

But the analysis is different from this point because of the difference between the discrete version of Galerkin orthogonality (5.12) and the standard version (2.2) with no quadrature. Now we obtain the error representation:

$$\begin{aligned} (e_i^-(t_n), \phi_{n,i}) &= (e_i^+(0), \phi_i(0)) \\ &+ \int_0^{t_n} \left((\dot{U}_i, \pi P\phi_i - \phi_i) + (\epsilon_i(U)\nabla U_i, \nabla(\pi P\phi_i - \phi_i)) - (f_i(U), \pi P\phi_i - \phi_i) \right) dt \\ &+ \int_0^{t_n} \left(\dot{U}_i - f_i(U), \pi P\phi_i)_h - (\dot{U}_i - f_i(U), \pi P\phi_i) \right) dt. \end{aligned} \tag{5.14}$$

We recognize the last term on the right as reflecting the error due to the use of quadrature. A similar analysis for the dGL method gives

$$\begin{aligned} (e_i^-(t_n), \phi_{n,i}) &= (e_i^-(0), \phi_i(0)) + \sum_{j=1}^{n} \left([U_i]_{j-1}, (\pi P\phi_i - \phi)_{j-1}^+ \right) \\ &+ \int_0^{t_n} \left((\dot{U}_i, \pi P\phi_i - \phi_i) + (\epsilon_i(U)\nabla U_i, \nabla(\pi P\phi_i - \phi_i)) - (f_i(U), \pi P\phi_i - \phi_i) \right) dt \\ &+ \int_0^{t_n} \left(\dot{U}_i - f_i(U), \pi P\phi_i)_h - (\dot{U}_i - f_i(U), \pi P\phi_i) \right) dt \\ &+ \sum_{j=1}^{n} \left(([U_i]_{j-1}, \pi P\phi_{j-1,i}^+)_h - ([U_i]_{j-1}, \pi P\phi_{j-1,i}^+) \right). \end{aligned} \tag{5.15}$$

We define the residuals on I_n componentwise for $1 \leq i \leq D$ as

$$R_Q^1(U)_i = \sum_{K \in \mathcal{T}_n} \|h_n^2 D^2 f_i(U)\|_{L_2(K)}$$

and

$$R_Q^2(U)_i = \begin{cases} \|h_n^2 \nabla(\dot{U}_i - f_i(U))\| & \text{(cGL)} \\ \|h_n^2 \nabla(\dot{U}_i - f_i(U))\| + \|h_n^2 [\nabla U_i]_{n-1}\| k_n^{-1} & \text{(dGL)} \end{cases}$$

as well as the stability factor

$$S_Q(0,t_n) = \int_0^{t_n} \|\nabla\phi\|\, dt.$$

Note that the form of these residual errors is slightly different than the previous residual errors in that the norm is included in the definition. We plot these residual errors and the stability factor for the computation on the bistable problem carried out above in Fig. 5.5.

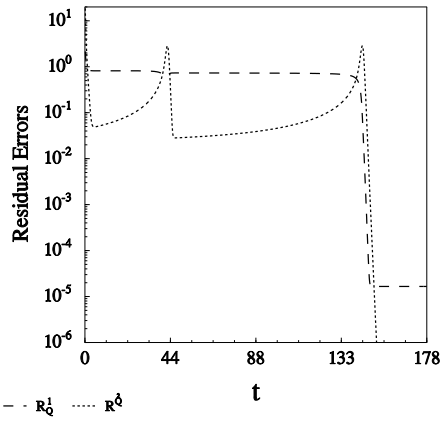

FIG. 5.5. *(a) Plot of the residual errors for the computation on the bistable equation shown in Fig.s 1.1 and 2.1.*

In Section 8, we prove

THEOREM 5.4. *There is a constant C depending on λ_i and q such that for $1 \le \alpha \le q$, the error of the cG(q) or dG(q − 1) approximation at time t_n, $1 \le n$, satisfies*

$$\begin{aligned} \|e^-(t_n)\| \le\ & S_0(0,t_n)\|e^-(0)\| + S_t^\alpha(0,t_n)\|k^\alpha R_t(U)\|_{L_\infty(0,t_n)} \\ & + S_x^p(0,t_n)\left(\|h^2 R_x^p(U)\|_{L_\infty(0,t_n)} + \|h^2 R_2^p(U)\|_{L_\infty(0,t_n)}\right) \\ & + CS_t^0(0,t_n)\max_{[0,t_n]} R_Q^1(U) + CS_Q(0,t_n)\max_{[0,t_n]} R_Q^2(U). \quad (5.16) \end{aligned}$$

Remark 5.6. The residual errors and stability factors present in the first three terms on the right-hand side of (5.16) arise from the approximation of the solution of the differential equation by a piecewise polynomial function via Galerkin's method. The residuals and stability factors in the last two terms occur because we discretize the (variational form of) the differential operator by means of quadrature. We see that these two residual errors accumulate at possibly different rates than the original residual errors and hence it can be important to distinguish the two sources of discretization

error. This point is hidden in the classic analysis of the five point difference scheme, which as we noted is closely related to the finite element method with lumped mass quadrature. See Eriksson, Estep, Hansbo, and Johnson [21] for further discussion of this issue.

We now estimate the size of the residual errors associated to quadrature.

THEOREM 5.5. *Assume that* $\nabla\dot{\tilde{u}} \in L_2(I_n; L_2(\Omega))$ *and* $\tilde{u} \in L_\infty(I_n; H^2(\Omega))$. *Then there is a constant* C *depending on* λ_i, ϵ, *and* f *such that for* $1 \le i \le D$,

$$R_Q^1(U) \le C\|h_n \nabla U\|^2_{L_\infty(I_n)} \tag{5.17}$$

and

$$R_Q^2(U) \le C\big(k_n^{-1}\|h_n\tilde{e}\|_{L_\infty(I_n)} + k_n^{-1}\|h_n^3\Delta\tilde{u}\|_{L_\infty(I_n)} + k_n^{-1/2}\|h_n^2\nabla\dot{\tilde{u}}\|_{L_2(I_n)} + \|h_n^2\nabla U\|_{L_\infty(I_n)}\big) \tag{5.18}$$

Under further assumptions on the solution and approximation, we can get more precise estimates.

THEOREM 5.6. *Assume that* $\tilde{u} \in L_\infty(I_n; H^1(\Omega))$, $\dot{\tilde{u}} \in L_\infty(I_n; L_2(\Omega))$, $\nabla\dot{\tilde{u}} \in L_2(I_n; L_2(\Omega))$, *and* $\tilde{u} \in L_\infty(I_n; H^2(\Omega))$ *and that there is a constant* C *depending on* λ_i, ϵ, *and* f *such that*

$$\begin{aligned}
\|\nabla\tilde{u}\|_{L\infty(I_n)} &\le Ce^{Ck_n}\big(\|\nabla\tilde{u}_{n-1}\| + 1\big)\\
\|\Delta\tilde{u}\|_{L\infty(I_n)} &\le Ce^{Ck_n}\big(\|\Delta\tilde{u}_{n-1}\| + \|\nabla\tilde{u}_{n-1}\| + 1\big)\\
\|\dot{\tilde{u}}\|_{L\infty(I_n)} &\le Ce^{Ck_n}\big(\|\Delta\tilde{u}_{n-1}\| + \|\nabla\tilde{u}_{n-1}\| + 1\big)\\
\|\nabla\dot{\tilde{u}}\|_{L_\infty(I_n)} &\le \|\Delta\tilde{u}_{n-1}\| + C\|\nabla\tilde{u}\|_{L_2(I_n)}.
\end{aligned}$$

In addition, assume that the numerical approximation satisfies the energy estimates

$$\begin{aligned}
\|U\|_{L\infty(I_n)} &\le C\big(\|U^-_{n-1}\| + k_n\|\nabla U^-_{n-1}\| + k_n\big),\\
\|\nabla U\|_{L\infty(I_n)} &\le C\big(k_n^{1/2}\|\Delta_{h_n}U^-_{n-1}\| + \|\nabla U^-_{n-1}\| + k_n^{1/2}\big)
\end{aligned}$$

and the a priori error bound

$$\begin{aligned}
\|\tilde{e}\|_{L\infty(I_n)} \le C\|h_n^2\Delta\tilde{u}_{n-1}\| + Ce^{Ck_n}\big(&k_n\|\dot{\tilde{u}}\|_{L\infty(I_n)} + \|h_n^2\Delta\tilde{u}\|_{L\infty(I_n)}\\
&+ k_n^{1/2}\|h_n\nabla\dot{\tilde{u}}\|_{L_2(I_n)} + k_n^{1/2}\|h_n\dot{\tilde{u}}\|_{L\infty(I_n)}\\
&+ k_n\|h_n\nabla\tilde{u}\|^2_{L\infty(I_n)} + k_n^{1/2}\|h_n\nabla\tilde{u}\|_{L\infty(I_n)}\big).
\end{aligned}$$

Then in addition to the estimates in Theorem 5.6,

$$R_Q^1(U) \le Ce^{Ck_n}\big(k_n\|h_n\Delta_{h_n}U^-_{n-1}\|^2 + \|h_n\nabla U^-_{n-1}\|^2 + h^2_{n,\max}\big)$$

and

$$\begin{aligned}
R_Q^2(U) \le Ce^{Ck_n}\big(&k_n^{-1}\|h_n^3\Delta_{h_n}U^-_{n-1}\| + k_n^{-1}\|h_n^3\nabla U^-_{n-1}\| + k_n^{-1}h^3_{n,\max}\\
&+ \|h_n\Delta_{h_n}U^-_{n-1}\| + \|h_n\nabla U^-_{n-1}\| + h_{n,\max}\\
&+ k_n^{-1/2}\|h_n^2\Delta_{h_n}U^-_{n-1}\| + k_n^{-1/2}\|h_n^2\nabla U^-_{n-1}\| + k_n^{-1/2}h^2_{n,\max}\big).
\end{aligned}$$

We conclude that the bound on all of the residual errors is on the order

$$\mathbf{O}\big(k_n^{-1/2}h_{n,\max}^{3/2}\big) + \mathbf{O}\big(k_n^2 h_{n,\min}^{-5/2}\big).$$

If we choose $k_n \sim h_{n,\min}^{1+\gamma}$ where $1/4 < \gamma < 1$, then all the residuals tend to zero as the time step and mesh size are refined. The proofs of Theorems 5.5 and 5.6 are similar to the proofs of Theorems 3.2 and 3.4 and we do not give the details.

Under the assumption of an invariant rectangle for the solution of (3.19) with $d = D$ and the associated dGL and cGL approximations, the assumptions in Theorems 5.5 and 5.6 can be verified for (3.19). To save space, we do not give the details. Instead, we turn to establishing conditions that guarantee that any rectangle on which (5.1) holds is invariant for the dGL and cGL approximations. This result is closely related to results of Hoff for finite difference schemes for reaction-diffusion equations that admit invariant rectangles in [45] and the analysis of a backward Euler discretization of the Hodgkin-Huxley equations presented in Mascagni [49].

We prove the following theorem in Section 8.

THEOREM 5.7. *Assume that the vector field f satisfies (5.1) and let L denote the maximum of the first partial derivatives of f on $\mathcal{R}$. Also assume that the maximum internal angle of any triangle in a triangulation is less or equal to $\pi/2$. There are constants $\mu_1, \mu_2 > 0$ such that if the time steps satisfy $k_n L \le \mu_1$ and $k_n/h_{n,\min}^2 \le \mu_2$ for all $n \ge 1$ then $\mathcal{R}$ is an invariant region for the approximation.*

6. Details of the analysis in Section 2.

Proof of Theorem 2.1. We present the proof for the cG method. The analysis of the dG methods is similar.

We split the time and space projections of ϕ as $\pi P\phi - \phi = \pi P\phi - P\phi + P\phi - \phi$ in the error representation formula (2.15) to obtain

$$\begin{aligned}(e^-(t_n),\phi_n) &= (e^+(0),\phi(0)) \\ &\quad + \int_0^{t_n} \big((\dot{U},(\pi-I)P\phi) + (\epsilon(U)\nabla U,\nabla((\pi-I)P\phi)) \\ &\qquad\qquad - (f(U),(\pi-I)P\phi)\big)\,dt \\ &\quad + \int_0^{t_N} \big((\dot{U},(P-I)\phi) + (\epsilon(U)\nabla U,\nabla((P-I)\phi)) \\ &\qquad\qquad - (f(U),(P-I)\phi)\big)\,dt \\ &= I + II + III.\end{aligned}$$

We estimate each of the terms I, II, and III, beginning with term I where we use the Cauchy-Schwarz inequality and the boundedness of the L_2 projection to get:

$$|(e^+(0),\phi(0))| \le \|e^-(0)\|\,\|\phi(0)\| = S_0(0,t_n)\|e^-(0)\|. \tag{6.1}$$

Since $P\phi \in V$, we can rewrite II:

$$\begin{aligned}II &= \int_0^{t_n} \big(\dot{U} - (\nabla\cdot\epsilon(U)\nabla)_h U - f(U),(\pi-I)P\phi\big)\,dt \\ &= \int_0^{t_n} \big(R_t(U),(\pi-I)P\phi\big)\,dt.\end{aligned}$$

Multiplying and dividing by k^α, where $0 \le \alpha \le q$ for the cG(q) method or dG(q-1) method and using the Cauchy-Schwarz inequality we get

$$|II| \le \|k^\alpha R_t(U)\|_{L_\infty(0,t_n)} \int_0^{t_n} \|k^{-\alpha}(\pi - I)P\phi\| \, dt.$$

Finally, we use the facts that $\pi P = P\pi$ and $\|P\| \le 1$ and the following standard *interpolation error estimate* for the L_2 projection (see Ciarlet [13]) on each time interval I_n,

$$\int_{I_n} \|k_n^{-\alpha}(I - \pi)v\| \, dt \le C_t^\alpha \int_{I_n} \|D_t^\alpha v\| \, dt, \tag{6.2}$$

where C_t^α depends only on the order α. We obtain

$$\begin{aligned}\int_0^{t_n} \|k^{-\alpha}(\pi - I)P\phi\| \, dt &= \int_0^{t_n} \|k^{-\alpha}P(\pi - I)\phi\| \, dt \le \int_0^{t_n} \|k^{-\alpha}(\pi - I)\phi\| \, dt \\ &\le C_t^\alpha \int_0^{t_n} \|D_t^\alpha \phi\| \, dt,\end{aligned}$$

and thus we conclude

$$|II| \le S_t^\alpha(0,t_n)\|k^\alpha R_t(U)\|_{L_\infty(0,t_n)}. \tag{6.3}$$

To begin the analysis of III, we first rewrite the diffusion term at a given time $t \in I_j$, $1 \le j \le n$, by splitting the integral into a sum of integrals over the elements and using Green's formula on each to get

$$\begin{aligned}\int_\Omega \epsilon_i(U)\nabla U_i \cdot \nabla(P - I)\phi_i \, dx &= \sum_{K \in \mathcal{T}_j} \int_K \epsilon_i(U)\nabla U_i \cdot \nabla(P - I)\phi_i \, dx \\ &= \sum_{K \in \mathcal{T}_j} -\int_K (\nabla \cdot \epsilon_i(U)\nabla U_i)(P - I)\phi_i \, dx \\ &\quad + \sum_{K \in \mathcal{T}_j} \int_{\partial K \setminus \partial\Omega} \epsilon_i(U)[\nabla U_i]_{\partial K}/2 \cdot n_{\partial K}(P - I)\phi_i \, ds,\end{aligned}$$

where as above $n_{\partial K}$ is the unit outward normal of ∂K. With this in mind, III becomes

$$\begin{aligned}III &= \int_0^{t_n} \left(\dot{U} - \nabla \cdot \epsilon(U)\nabla U - f(U), (P - I)\phi\right) dt \\ &\quad + \sum_{j=1}^n \int_{I_j} \sum_{K \in \mathcal{T}_j} \int_{\partial K \setminus \partial\Omega} \epsilon(U)[\nabla U]_{\partial K}/2 \cdot n_{\partial K}(P - I)\phi \, ds \, dt \\ &= IV + V.\end{aligned}$$

Taken in entirety, IV splits naturally into parts associated with the parabolic and ordinary differential equations:

$$IV = \int_0^{t_n} \left(R_x^p(U), (P - I)\phi^p\right) dt + \int_0^{t_n} \left(R_x^o(U), (P - I)\phi^o\right) dt.$$

We estimate the first term by multiplying and dividing by h^2 and using the Cauchy-Schwarz inequality and the standard interpolation error estimate for the L_2 projection in space,

$$\|(h^{-2}(I-P)v\| + \|h^{-1}\nabla(I-P)v\| \leq C_x^p \|D^2 v\|. \tag{6.4}$$

This gives

$$\begin{aligned}\left|\int_0^{t_n} \left(R_x^p(U), (P-I)\phi^p\right) dt\right| &\leq \int_0^{t_n} \|h^2 R_x^p(U)\| \, \|h^{-2}(P-I)\phi^p\| \, dt \\ &\leq \|h^2 R_x^p(U)\|_{L_\infty(0,t_n)} C_x^p \int_0^{t_n} \|D^2\phi^p\| \, dt \\ &= S_x^p(0,t_n)\|h^2 R_x^p(U)\|_{L_\infty(0,t_n)}.\end{aligned}$$

We treat the second term in IV differently because the associated ordinary differential equations do not exhibit elliptic smoothing. We use the orthogonality property of the L_2 projection to insert the interpolant $\tilde{P}$:

$$\begin{aligned}\left|\int_0^{t_n} \left(R_x^o(U), (P-I)\phi^o\right) dt\right| &= \left|\int_0^{t_n} \left((I-\tilde{P})R_x^o(U), (P-I)\phi^o\right) dt\right| \\ &\leq \|(I-\tilde{P})R_x^o(U)\|_{L_\infty(0,t_n)} \int_0^{t_n} \|(P-I)\phi^o\| \, dt \\ &\leq S_x^o(0,t_n)\|(I-\tilde{P})R_x^o(U)\|_{L_\infty(0,t_n)}.\end{aligned}$$

We conclude that

$$|IV| \leq S_x^p(0,t_n)\|h^2 R_x^p(U)\|_{L_\infty(0,t_n)} + S_x^o(0,t_n)\|(I-\tilde{P})R_x^o(U)\|_{L_\infty(0,t_n)}. \tag{6.5}$$

Finally, we estimate V. Note that by the definition of $R_2^p(U)$, it follows that

$$\|h^2 R_2^p(U)\|_{L_2(K)} = \|h^{3/2}\epsilon(U)[\nabla U]_{\partial K}/2\|_{L_2(\partial K)}.$$

Multiplying and dividing with $h^{3/2}$ and estimating using the Cauchy-Schwarz inequality,

$$\begin{aligned}|V| &\leq \sum_{j=1}^n \int_{I_j} \sum_{K\in\mathcal{T}_j} \left\|h^{3/2}\epsilon(U)[\nabla U]_{\partial K}/2 \cdot n_{\partial K}\right\|_{L_2(\partial K)} \left\|h^{-3/2}(P-I)\phi\right\|_{L_2(\partial K)} dt \\ &\leq \sum_{j=1}^n \int_{I_j} \|h^2 R_2^p(U)\| \left(\sum_{K\in\mathcal{T}_j} \left\|h^{-3/2}(P-I)\phi\right\|_{L_2(\partial K)}^2\right)^{1/2} dt.\end{aligned}$$

Now we employ a *trace inequality* on the boundary of the elements:

$$\|v\|_{L_2(\partial K)}^2 \leq C_t \|v\|_{L_2(K)} \|\nabla v\|_{L_2(K)} : \tag{6.6}$$

which implies that

$$\|v\|_{L_2(\partial K)}^2 \leq \frac{C_t}{2}\left(\|h^{-1/2}v\|_{L_2(K)}^2 + \|h^{1/2}\nabla v\|_{L_2(K)}^2\right).$$

Taking $v = h^{-3/2}(P-I)\phi$, we get

$$\|h^{-3/2}(P-I)\phi\|^2_{L_2(\partial K)} \leq \frac{C_t}{2}\left(\|h^{-2}(P-I)\phi\|^2_{L_2(K)} + \|h^{-1}\nabla(P-I)\phi\|^2_{L_2(K)}\right).$$

Summing over all of the elements and applying the interpolation inequality (6.4), we get

$$\left(\sum_{K\in\mathcal{T}_j}\|h^{-3/2}(P-I)\phi\|^2_{L_2(\partial K)}\right)^{1/2} \leq C_x^p\frac{C_t}{2}\|D^2\phi\|.$$

We conclude that

$$|V| \leq C_x^p\frac{C_t}{2}\int_0^{t_n}\|h^2R_2^p(U)\|\,\|D^2\phi\|\,dt \leq \frac{C_t}{2}S_x^p(0,t_n)\|h^2R_2^p(U)\|_{L_\infty(0,t_n)}. \tag{6.7}$$

Collecting the estimates (6.1), (6.3), (6.5), and (6.7), we reach the desired result.

Proof of Theorem 2.2. We begin as in the derivation of the error representation formula (2.15), integrating from 0 to t^* and then splitting the integral into the difference of the integral from 0 to t_n and the integral from t_n to t^*. Writing $\phi = P\phi + (I-P)\phi$, we get

$$\begin{aligned}(e(t^*),\phi(t^*)) &= (e(t_n),\phi(t_n))\\ &\quad+\int_{t_n}^{t^*}\left((\dot{U},P\phi)+(\epsilon(U)\nabla U,\nabla P\phi)-(f(U),P\phi)\right)dt\\ &\quad+\int_{t_n}^{t^*}\left((\dot{U},(I-P)\phi)+(\epsilon(U)\nabla U,\nabla(I-P)\phi)\right.\\ &\qquad\qquad\left.-(f(U),(I-P)\phi)\right)dt\\ &= I+II+III.\end{aligned}$$

Expression III is estimated exactly as the same term in the proof of Theorem 2.1. However, II must be estimated differently since the lack of Galerkin orthogonality of the approximation over (t^*,t_n) means that we cannot insert an interpolant in front of the dual solution ϕ. Therefore, we make a straightforward estimate

$$|II| \leq \int_{t_n}^{t^*}\|R_t(U)\|\,\|\phi\|\,dt \leq k_n\|R_t(U)\|_{L_\infty(t^*,t_n)}\|\phi\|_{L_\infty(t^*,t_n)}.$$

7. Details of the analysis in Section 3.

Proof of Lemma 3.1. The first claim follows from standard elliptic regularity results since $\Delta_{h_n}U^-_{n-1}$ is continuous and the Dirichlet boundary conditions given by U^-_{n-1} are in H^1. As for the second, it follows from the definition of Δ_h and the assumption of nested meshes that U^-_{n-1} is nothing more than the Galerkin finite element approximation of $\tilde{u}_{n-1}$ in V_n and the classic a priori error bound for the finite element method for Laplace's equation implies that

$$\|\tilde{u}_{n-1}-U^-_{n-1}\| + \|h_n\nabla(\tilde{u}_{n-1}-U^-_{n-1})\| \leq C\|h_n^2\Delta_{h_n}U^-_{n-1}\|.$$

We compute

$$\|h_n^{1/2}\Delta_{h_n}U_{n-1}^-\| = \sup_{\substack{\psi\in V_n\\ \|\psi\|=1}} \left(\Delta_{h_n}U_{n-1}^-, h_n^{1/2}\psi\right).$$

Using the definition of Δ_h and Green's theorem and then estimating, we get

$$\begin{aligned}\left(\Delta_{h_n}U_{n-1}^-, h_n^{1/2}\psi\right) &= \left(\nabla U_{n-1}^-, h_n^{1/2}\nabla\psi\right)\\ &= \sum_{K\in T_n}\int_{\partial K\setminus\partial\Omega}\left[n_{\partial K}\cdot\nabla U_{n-1}^-\right]_{\partial K}h_n^{1/2}\psi\,ds\\ &\le \sum_{K\in T_n}\|\left[n_{\partial K}\cdot\nabla U_{n-1}^-\right]_{\partial K}\|_{L_2(\partial K)}\,\|h_n^{1/2}\psi\|_{L_2(\partial K)}.\end{aligned}$$

Using the trace inequality (6.6) and the following *inverse estimate*: there is a constant C depending on λ_i such that for all $W\in V_n$,

$$\|h_n\nabla W\| \le C\|W\| : \tag{7.1}$$

we find

$$\|h_n^{1/2}\psi\|_{L_2(\partial K)}^2 \le C\|\psi\|_{L_2(K)}\,\|h_n\nabla\psi\|_{L_2(K)} \le C\|\psi\|_{L_2(K)}.$$

The Cauchy-Schwarz inequality therefore implies that

$$\sup_{\substack{\psi\in V_n\\ \|\psi\|=1}} \left(\Delta_{h_n}U_{n-1}^-, h_n^{1/2}\psi\right) \le \left(\sum_{K\in T_n}\left\|\left[n_{\partial K}\cdot\nabla U_{n-1}^-\right]_{\partial K}\right\|_{L_2(\partial K)}^2\right)^{1/2}.$$

Since $V_{n-1}\subset V_n$, the expression on the right-hand side is independent of h_n.

Proof of Theorem 3.2.

We first estimate the time residual error R_t for the cG(1) method. From the formula for R_t, we subtract $\dot{\tilde{u}}_i - \nabla\cdot\epsilon_i(\tilde{u})\nabla\tilde{u}_i - f_i(\tilde{u}) \equiv 0$ to get

$$\begin{aligned}\|k_nR_t(U)_i\| &\le k_n\|\dot{U}_i - \dot{\tilde{u}}_i\| + k_n\|(\nabla\cdot\epsilon_i(U)\nabla)_hU_i - \nabla\cdot\epsilon_i(\tilde{u})\nabla\tilde{u}_i\|\\ &\qquad + k_n\|f_i(U) - f_i(\tilde{u})\|\\ &= I + II + III.\end{aligned}$$

I is estimated simply

$$\begin{aligned}I &= \left\|\left(U_{n,i}^- - \tilde{u}_{n,i}\right) - \left(U_{n-1,i}^+ - \tilde{u}_{n-1,i}\right) + \left(\tilde{u}_{n,i} - \tilde{u}_{n-1,i}\right) - \dot{\tilde{u}}_ik_n\right\|\\ &\le 2\|\tilde{e}_i\|_{L_\infty(I_n)} + 2k_n\|\dot{\tilde{u}}_i\|_{L_\infty(I_n)}\end{aligned}$$

Note that II has to be estimated only for the part of U associated to the parabolic part of the differential equation. To do this, we use the weighted elliptic projection R_h into V defined componentwise for $1\le i\le d$ by

$$\left(\epsilon_i(\tilde{u})\nabla R_hv_i, \nabla W_i\right) = \left(\epsilon_i(\tilde{u})\nabla v_i, \nabla W_i\right) \quad \text{for all } W\in V.$$

For this projection, there is a constant C (depending on ϵ) such that

$$\|R_hv_i - v_i\| + \|h_n\nabla(R_hv_i - v_i)\| \le C\|h_n^2D^2v_i\|.$$

See Thomée [63]. Also $\|\nabla R_h v_i\| \le \|\nabla v_i\|$. We write

$$\begin{aligned} II &\le k_n \|(\nabla\cdot\epsilon_i(U)\nabla)_h U_i - (\nabla\cdot\epsilon_i(U)\nabla)_h R_h\tilde{u}_i\| \\ &\qquad + k_n\|(\nabla\cdot\epsilon_i(U)\nabla)_h R_h\tilde{u}_i - \nabla\cdot\epsilon_i(U)\nabla\tilde{u}_i\| \\ &= IIa + IIb. \end{aligned}$$

To estimate IIa for $1 \le i \le d$, we use the definition

$$\begin{aligned} IIa &= k_n \sup_{\substack{\psi\in V\\ \|\psi=1\|}} \big((\nabla\cdot\epsilon_i(U)\nabla)_h U_i - (\nabla\cdot\epsilon_i(U)\nabla)_h R_h\tilde{u}_i, \psi\big) \\ &= -k_n \sup_{\substack{\psi\in V\\ \|\psi=1\|}} \big(\epsilon_i(U)(\nabla U_i - \nabla R_h\tilde{u}_i), \nabla\psi\big) - k_n \sup_{\substack{\psi\in V\\ \|\psi=1\|}} \big((\epsilon_i(U)-\epsilon_i(\tilde{u}))\nabla R_h\tilde{u}_i, \nabla\psi\big) \\ &= IIa1 + IIa2. \end{aligned}$$

Using the mesh assumption, the Cauchy-Schwarz inequality, and the inverse estimate (7.1), we get

$$\begin{aligned} \|IIa1\| &\le C\|\epsilon_i(U)\|\,\|h_n^{-1}\nabla(U_i - R_h\tilde{u}_i)\|\,\|h_n\nabla\psi\|k_n \\ &\le C\|h_n^{-2}(U_i - R_h\tilde{u}_i)\|k_n \le C\big(k_n\|h_n^{-2}\tilde{e}_i\| + k_n\|h_n^{-2}(\tilde{u}_i - R_h\tilde{u}_i)\|\big). \end{aligned}$$

Similarly, we estimate

$$\|IIa2\| \le C\|h_n^{-1}(\epsilon_i(U)-\epsilon_i(\tilde{u}))\|\,\|\nabla R_h\tilde{u}_i\|\,\|h_n\nabla\psi\|k_n \le C\|h_n^{-1}\tilde{e}\|\,\|\nabla\tilde{u}_i\|k_n,$$

and proceed in the same fashion.

To estimate IIb, we start with the definition

$$IIb = k_n \sup_{\substack{\psi\in H_0^1(\Omega)\\ \|\psi\|=1}} \big((\nabla\cdot\epsilon_i(\tilde{u})\nabla)_h R_h\tilde{u}_i - \nabla\cdot\epsilon_i(\tilde{u})\nabla\tilde{u}_i, \psi\big).$$

Now we use the definitions of R_h and P to get

$$\begin{aligned} &\big((\nabla\cdot\epsilon_i(\tilde{u})\nabla)_h R_h\tilde{u}_i - \nabla\cdot\epsilon_i(\tilde{u})\nabla\tilde{u}_i, \psi\big) \\ &\qquad = \big((\nabla\cdot\epsilon_i(\tilde{u})\nabla)_h R_h\tilde{u}_i, \psi\big) - \big(\nabla\cdot\epsilon_i(\tilde{u})\nabla\tilde{u}_i, \psi\big) \\ &\qquad = \big((\nabla\cdot\epsilon_i(\tilde{u})\nabla)_h R_h\tilde{u}_i, P\psi\big) - \big(\nabla\cdot\epsilon_i(\tilde{u})\nabla\tilde{u}_i, \psi\big) \\ &\qquad = -\big(\epsilon_i(\tilde{u})\nabla R_h\tilde{u}_i, \nabla P\psi\big) + \big(\epsilon_i(\tilde{u})\nabla\tilde{u}_i, \nabla\psi\big) \\ &\qquad = -\big(\epsilon_i(\tilde{u})\nabla\tilde{u}_i, \nabla P\psi\big) + \big(\epsilon_i(\tilde{u})\nabla\tilde{u}_i, \nabla\psi\big) \\ &\qquad = \big(\nabla\cdot\epsilon_i(\tilde{u})\nabla\tilde{u}_i, (I-P)\psi\big). \end{aligned}$$

Using the mesh assumption and the stability of P yields

$$IIb \le \|\nabla\cdot\epsilon_i(\tilde{u})\nabla\tilde{u}_i\|k_n \le C\big(\|\nabla\tilde{u}\|^2 + \|\Delta\tilde{u}_i\|\big)k_n.$$

The last term is simple: $III \le Ck_n\|\tilde{e}\|_{L_\infty(I_n)}$.

In the case of the dG methods, we also estimate

$$k_n\|[U]_{n-1,i}\,k_n^{-1}\| \le \|\tilde{e}_i\|_{L_\infty(I_n)} + \|\tilde{e}^-_{n-1,i}\| \le \|\tilde{e}_i\|_{L_\infty(I_n)} + C\|h_n^2\Delta_{h_n}U^-_{n-1,i}\|.$$

Putting together the different contributions in the two cases (ordinary and partial differential equations) yields (3.2) and (3.3).

We begin the estimate on R_x^p by subtracting the differential equation (3.1) from the definition and taking norms as above to get for $1 \le i \le d$,

$$\|h_n^2 R_x^p(U)_i\| \le \sum_{K\in\mathcal{T}_n} \big(\|h_n^2(\dot{U}_i - \dot{\tilde{u}}_i)\|_{L_2(K)} + \|h_n^2(f_i(U) - f_i(\tilde{u}))\|_{L_2(K)} + \|h_n^2(\nabla\cdot\epsilon_i(U)\nabla U_i - \nabla\cdot\epsilon_i(\tilde{u})\nabla\tilde{u}_i)\|_{L_2(K)}\big)$$

The first two terms are treated like the corresponding terms in R_t above, while

$$\begin{aligned}\sum_{K\in\mathcal{T}_n} &\|h_n^2(\nabla\cdot\epsilon_i(U)\nabla U_i - \nabla\cdot\epsilon_i(\tilde{u})\nabla\tilde{u}_i)\|_{L_2(K)} \\ &\le \sum_{K\in\mathcal{T}_n} \|h_n^2\nabla\cdot\epsilon_i(\tilde{u})(\nabla U_i - \nabla\tilde{u}_i)\|_{L_2(K)} \\ &\qquad + \sum_{K\in\mathcal{T}_n} \|h_n^2\nabla\cdot(\epsilon_i(U) - \epsilon_i(\tilde{u}))\nabla U_i\|_{L_2(K)} \\ &= I + II.\end{aligned}$$

Expanding,

$$\begin{aligned} I &= \Big\| \sum_{K\in\mathcal{T}_n} \Big(h_n^2 \sum_{j_1,j_2} \sum_{l=1}^{D} \Big(\Big(\frac{\partial\epsilon_i(U)}{\partial u_l} - \frac{\partial\epsilon_i(\tilde{u})}{\partial u_l} \Big) \frac{\partial\tilde{u}_l}{\partial x_{j_1}} \frac{\partial\tilde{u}_i}{\partial x_{j_2}} \\ &\quad + \frac{\partial\epsilon_i(U)}{\partial u_l} \Big(\frac{\partial U_l}{\partial x_{j_1}} - \frac{\partial\tilde{u}_l}{\partial x_{j_1}} \Big) \frac{\partial U_i}{\partial x_{j_2}} + \frac{\partial\epsilon_i(U)}{\partial u_l} \Big(\frac{\partial U_l}{\partial x_{j_1}} - \frac{\partial\tilde{u}_l}{\partial x_{j_1}} \Big) \frac{\partial\tilde{u}_i}{\partial x_{j_2}} \Big) + \epsilon_i(\tilde{u})\Delta\tilde{u}_i \Big\| \\ &= \|Ia + Ib + Ic + Id\| \end{aligned}$$

Term by term, we estimate

$$\begin{gathered} \|Ia\| \le \|\nabla_u\epsilon_i(\tilde{u}) - \nabla_u\epsilon_i(U)\|\,\|h_n\nabla\tilde{u}\|\,\|h_n\nabla\tilde{u}_i\| \le C\|\tilde{e}\|\|h_n\nabla\tilde{u}\|^2, \\ \|Ib\| \le \|\nabla_u\epsilon_i(U)\|\,\|h_n\nabla\tilde{e}\|\,\|h_n\nabla U_i\| \le C\|h_n\nabla\tilde{e}\|\,\|U_i\|, \\ \|Ic\| \le \|\nabla_u\epsilon_i(U)\|\,\|h_n\nabla\tilde{e}\|\,\|h_n\nabla\tilde{u}_i\| \le C\|h_n\nabla\tilde{e}\|\,\|h_n\nabla\tilde{u}_i\|, \\ \|Id\| \le C\|h_n^2\Delta\tilde{u}_i\|. \end{gathered}$$

Next, we write

$$\begin{aligned} II &= \Big\| \sum_{K\in\mathcal{T}_n} \Big(h_n^2 \sum_{j_1,j_2} \sum_{l=1}^{D} \Big(\Big(\frac{\partial\epsilon_i(U)}{\partial u_l} - \frac{\partial\epsilon_i(\tilde{u})}{\partial u_l} \Big) \frac{\partial U_l}{\partial x_{j_1}} \frac{\partial U_i}{\partial x_{j_2}} \\ &\qquad + \frac{\partial\epsilon_i(\tilde{u})}{\partial u_l} \Big(\frac{\partial U_l}{\partial x_{j_1}} - \frac{\partial\tilde{u}_l}{\partial x_{j_1}} \Big) \frac{\partial U_i}{\partial x_{j_2}} \Big) \Big\| \\ &\le C\big(\|\tilde{e}\|\,\|h_n\nabla U\|^2 + \|h_n\nabla\tilde{e}\|\,\|h_n\nabla U_i\| \big) \le C\big(\|\tilde{e}\|\,\|U\|^2 + \|h_n\nabla\tilde{e}\|\,\|U_i\| \big) \end{aligned}$$

Estimate (3.4) now follows.

To obtain (3.5), we compute using the definition of R_2^p:

$$\|h_n^2 R_2^p(U)_i\|^2 = \sum_{K\in\mathcal{T}_n} \int_{\partial K\setminus\partial\Omega} h_n(K)^3 \big| \epsilon_i(U) n_{\partial K} \cdot [\nabla U_i]_{\partial K}/2 \big|^2 \, ds.$$

Consider the integral along one edge of ∂K for some K. Using ∇U_i^r and ∇U_i^l to denote the values of ∇U_i on the two sides of ∂K, we find that

$$\int_{\partial K} \left|[\nabla U_i]_{\partial K}\right|^2 ds \le 2 \int_{\partial K} \left(\left|\nabla U_i^r - \nabla \tilde{u}_i\right|^2 + \left|\nabla \tilde{u}_i - \nabla U_i^l\right|^2\right) ds$$

Noting that each internal edge occurs twice in the expansion for II, we obtain

$$\|h_n^2 R_2^p(U)_i\|^2 \le C \sum_{K \in \mathcal{T}_n} h_n^3(K) \|\nabla U_i - \nabla \tilde{u}_i\|_{L_2(\partial K)}^2$$

Now, we use the trace inequality to estimate

$$\begin{aligned}\|h_n^2 R_2^p(U)_i\|^2 &\le C \sum_{K \in \mathcal{T}_n} h_n^3(K) \|\nabla \tilde{e}_i\|_{L_2(K)} \|\Delta \tilde{u}_i\|_{L_2(K)} \\ &\le C\left(\|h_n \nabla \tilde{e}_i\|^2 + \|h_n^2 \Delta \tilde{u}_i\|^2\right).\end{aligned}$$

Finally with $\tilde{P} = P$, we insert the ordinary differential equation to get

$$\begin{aligned}\|(I-P)R_x^o(U)_i\| &= \|(I-P)\left((\dot{U}_i - \dot{\tilde{u}}_i) - (f_i(U) - f_i(\tilde{u}))\right)\| \\ &\le \|(I-P)\dot{\tilde{u}}_i\| + \|f_i(U) - f_i(\tilde{u})\| + \|P(f_i(U) - f_i(\tilde{u}))\| \\ &\le \|(I-P)f_i(\tilde{u})\| + 2\|f_i(\tilde{u}) - f_i(U)\| \\ &\le \begin{cases} C\|h_n \nabla f_i(\tilde{u})\| + C\|\tilde{e}\| \\ C\|h_n^2 D^2 f_i(\tilde{u})\| + C\|\tilde{e}\| \end{cases}\end{aligned}$$

and the result follows by differentiating f_i and using elliptic regularity.

Proof of Theorem 3.4.

First we estimate $\|\nabla \tilde{u}_{n-1}\|$ by introducing the discrete solution operator T_h into V_n corresponding to T and using the standard error estimate in the energy norm:

$$\begin{aligned}\|\nabla \tilde{u}_{n-1}\| &\le \|\nabla (T - T_h)\Delta_{h_n} U_{n-1}^-\| + \|\nabla T_h \Delta_{h_n} U_{n-1}^-\| \\ &\le C\|h_n \Delta_{h_n} U_{n-1}^-\| + \|\nabla U_{n-1}^-\|. \end{aligned} \tag{7.2}$$

Applying the energy estimates to the a priori error bound, we find that

$$\begin{aligned}\|\tilde{e}\|_{L_\infty(I_n)} \le Ce^{Ck_n}\big(&\|h_n^2 \Delta_{h_n} U_{n-1}^-\| + \|h_n^2 \nabla U_{n-1}^-\| + h_{n,\max}^2 + k_n \|\Delta_{h_n} U_{n-1}^-\| \\ &+ k_n \|\nabla U_{n-1}^-\| + k_n\big).\end{aligned}$$

To estimate $\|h_n \nabla \tilde{e}\|$, we introduce the weighted elliptic projection and use the inverse estimate (7.1):

$$\|h_n \nabla \tilde{e}\| \le \|h_n \nabla (U - R_h \tilde{u})\| + \|h_n \nabla (R_h \tilde{u} - \tilde{u})\| \le C\|\tilde{e}\| + \|R_h \tilde{u} - \tilde{u}\| + \|h_n \nabla (R_h \tilde{u} - \tilde{u})\|.$$

The result now follows by applying the energy estimates to the coefficients in the a priori bounds (3.2)–(3.6) together with these results.

To simplify the presentation of the proofs of the results in Section 3.3, we analyze the system

$$\begin{cases} \dot{u}_1 - \epsilon_1 \Delta u_1 = f_1(u_1, u_2), & (x,t) \in \Omega \times \mathbf{R}^+, \\ \dot{u}_2 = f_2(u_1, u_2), & (x,t) \in \Omega \times \mathbf{R}^+, \\ u_1(x,t) = 0, & (x,t) \in \Gamma \times \mathbf{R}^+, \\ u(x,0) = u_0(x), & x \in \Omega. \end{cases} \tag{7.3}$$

The proofs extend directly to the more general problem (3.19).

Proof of Proposition 3.5. To obtain the first estimate on the gradient of u, we first take the inner product of the parabolic equation in (7.3) and $-\Delta u_1$ to obtain

$$-(\dot{u}_1, \Delta u_1) + \epsilon_1 \|\Delta u_1\|^2 = -(f_1(u)\Delta u_1).$$

Using Green's formula on the first and last terms yields

$$\frac{1}{2}\frac{d}{dt}\|\nabla u_1\|^2 + \epsilon_1\|\Delta u_1\|^2 = \left(f_{u_{1,1}}\nabla u_1, \nabla u_1\right) + \left(f_{u_{2,1}}\nabla u_2, \nabla u_1\right). \tag{7.4}$$

Differentiating the ordinary differential equation with respect to x_i and taking the inner product with $u_{x_{i,2}}$ gives

$$\frac{1}{2}\frac{d}{dt}\|u_{x_{i,2}}\|^2 = \left(f_{u_{1,2}}u_{x_{i,1}}, u_{x_{i,2}}\right) + \left(f_{u_{2,2}}u_{x_{i,2}}, u_{x_{i,1}}\right). \tag{7.5}$$

Now we add (7.4) with (7.5) with $i = 1, 2$ and use the bound on the partial derivatives of f to obtain

$$\frac{1}{2}\frac{d}{dt}\|\nabla u\|^2 \leq 2L_n\|\nabla u\|^2.$$

Integrating and using Gronwall's lemma gives the desired estimate.

To obtain (3.21), we argue in the same way after taking the inner product of the parabolic equation with $-\Delta\dot{u}_1$ and using Green's formula to get

$$\|\nabla\dot{u}_1\|^2 + \epsilon_1\frac{1}{2}\frac{d}{dt}\|\Delta u_1\|^2 = \left(f_{u_{1,1}}\nabla u_1, \nabla\dot{u}_1\right) + \left(f_{u_{2,1}}\nabla u_2, \nabla\dot{u}_1\right).$$

Estimating with the Cauchy-Schwarz inequality on the right, kicking the terms involving $\|\nabla\dot{u}_1\|$, and using the previous estimate gives the result for $u^p = u_1$. To obtain the estimate for $u^o = u_2$, we begin by differentiating the ordinary differential equation with respect to x_i twice for $i = 1$ and 2, adding the resulting equations, taking the inner product with Δu_2, and estimating on the right-hand side using the Cauchy-Schwarz inequality and the boundedness of the partial derivatives of f. Integrating and using Gronwall's lemma together with the previous results yields the desired estimate.

Proof of Proposition 3.6. We present the proof of the result for the cG(1) method. The proofs for the other methods is similar. The cG(1) equations for (7.3) read

$$\begin{aligned}\int_{I_n}(\dot{U}_1, W_1)\,dt + \epsilon_1\int_{I_n}(\nabla U_1, \nabla W_1)\,dt &= \int_{I_n}(f_1(U), W_1)\,dt,\\ \int_{I_n}(\dot{U}_2, W_2)\,dt &= \int_{I_n}(f_2(U), W_2)\,dt,\end{aligned} \tag{7.6}$$

where $W \in V_n$. We choose $W_2 = U^-_{n,2}$ in the second equation, noting that

$$\int_{I_n}(\dot{U}_2, U^-_{n,2})\,dt = \|U^-_{n,2}\|^2 - (U^+_{n-1,2}, U^-_{n,2}).$$

Using the bound on $|f|$ and the stability of P with respect to the L_2 norm, we conclude

$$\|U^-_{n,2}\| \leq \|U^-_{n-1,2}\| + k_n M_n.$$

Similarly, we choose $W_1 = U^-_{n,1}$ in the first equation to get

$$\|U^-_{n,1}\|^2 + \epsilon_1 \frac{k_n}{2}\|\nabla U^-_{n,1}\|^2 = \big(U^+_{n-1,1}, U^-_{n,1}\big) - \epsilon_1 \frac{k_n}{2}\big(\nabla U^+_{n-1,1}, \nabla U^-_{n-1,1}\big) + \int_{I_n} (f_1(U), U^-_{n,1})\,dt.$$

Estimating on the right using the Cauchy-Schwarz inequality and the stability of P with respect to the energy norm, see Crouzeix and Thomée [16], and kicking the appropriate terms gives the conclusion.

To prove the a priori error bound, we begin by subtracting (7.6) from (7.3) to obtain an equation for the error $\tilde{e} = \tilde{u} - U$,

$$\begin{cases} \displaystyle\int_{I_n} (\dot{\tilde{e}}_1, W_1)\,dt + \epsilon_1 \int_{I_n} (\nabla \tilde{e}_1, \nabla W_1)\,dt = \int_{I_n} (f_1(\tilde{u}) - f_1(U), W_1)\,dt, \\ \displaystyle\int_{I_n} (\dot{\tilde{e}}_2, W_2)\,dt = \int_{I_n} (f_2(\tilde{u}) - f_2(U), W_2)\,dt, \end{cases} \tag{7.7}$$

for all $W \in V_n$. We split the error $\tilde{e} = \rho - \theta$ with $\rho = \tilde{u} - \tilde{U}$ denoting a computable interpolation error and $\theta = U - \tilde{U} \in V_n$. $\tilde{U}$ is defined by

$$\tilde{U} = Q_n \tilde{u}_{n-1} \frac{t - t_n}{-k_n} + Q_n \tilde{u}_n \frac{t - t_{n-1}}{k_n},$$

where Q_n denotes the interpolation operator into V_n. This means that $\|\tilde{e}\|_{L_\infty(I_n)} \le \|\rho\|_{L_\infty(I_n)} + \|\theta\|_{L_\infty(I_n)}$ where

$$\|\rho\|_{L_\infty(I_n)} \le C \begin{cases} k_n \|\dot{\tilde{u}}\|_{L_\infty(I_n)} \\ k_n^2 \|\ddot{\tilde{u}}\|_{L_\infty(I_n)} \end{cases} + C \begin{cases} \|h_n \nabla \tilde{u}\|_{L_\infty(I_n)} \\ \|h^2 D^2 \tilde{u}\|_{L_\infty(I_n)} \end{cases}. \tag{7.8}$$

From (7.7), we obtain an equation for the discrete error:

$$\begin{cases} \displaystyle\int_{I_n} (\dot{\theta}_1, W_1)\,dt + \epsilon_1 \int_{I_n} (\nabla \theta_1, \nabla W_1)\,dt = \int_{I_n} (\dot{\rho}_1, W_1)\,dt - \epsilon_1 \int_{I_n} (\nabla \rho_1, \nabla W_1)\,dt \\ \qquad\qquad + \displaystyle\int_{I_n} (f_1(\tilde{u}) - f_1(\tilde{U}), W_1)\,dt + \int_{I_n} (f_1(\tilde{U}) - f_1(U), W_1)\,dt, \\ \displaystyle\int_{I_n} (\dot{\theta}_2, W_2)\,dt = \int_{I_n} (\dot{\rho}_2, W_2)\,dt \\ \qquad\qquad + \displaystyle\int_{I_n} (f_2(\tilde{u}) - f_2(\tilde{U}), W_2)\,dt + \int_{I_n} (f_2(\tilde{U}) - f_2(U), W_2)\,dt. \end{cases}$$

We choose $W = \theta^-_n$ and estimate on the right, using

$$\int_{I_n} (\dot{\rho}_i, W_i)\,dt = (\rho^-_{n,i} - \rho^+_{n-1,i}, W_i),$$

to get from the first equation

$$\begin{aligned} &\|\theta^-_{n,1}\|^2 + \epsilon_1 \frac{k_n}{2}\|\nabla \theta^-_{n,1}\|^2 \\ &\quad \le \|\theta^+_{n-1,1}\|\|\theta^-_{n,1}\| + \epsilon_1 \frac{k_n}{2}\|\nabla \theta^+_{n-1,1}\|\|\theta^-_{n,1}\| + L_n k_n \|\rho\|_{L_\infty(I_n)} \|\theta^-_{n,1}\| \\ &\qquad + L_n k_n \|\theta\|^2_{L_\infty(I_n)} + 2\|\rho\|_{L_\infty(I_n)}\|\theta^-_{n,1}\| + \epsilon_1 k_n \|\nabla \rho_1\|_{L_\infty(I_n)} \|\nabla \theta^-_{n,1}\|. \end{aligned}$$

Now we use the Cauchy-Schwarz inequality to conclude that there is a constant $C = C(\epsilon)$ such that

$$\|\theta^-_{n,1}\|^2 \leq C\big(\|\theta^+_{n-1,1}\|^2 + k_n\|\nabla\theta^+_{n-1,1}\|^2 + \|\rho\|^2_{L_\infty(I_n)} + k_n\|\nabla\rho_1\|^2_{L_\infty(I_n)}\big) + CL_nk_n\|\theta\|^2_{L_\infty(I_n)}.$$

Adding the analogous estimate for $\|\theta^-_{n,2}\|^2$ and assuming that k_nL_n is sufficiently small, we conclude

$$\|\theta\|_{L_\infty(I_n)} \leq C\big(\|\theta^-_{n-1}\| + k_n^{1/2}\|\nabla\theta^-_{n-1,1}\| + \|\rho\|_{L_\infty(I_n)} + k_n^{1/2}\|\nabla\rho_1\|_{L_\infty(I_n)}\big).$$

The result follows directly.

Proof of Lemma 3.7. We present the proof for the dual problem for (7.3), which we rewrite after changing variable $t \to t_n - t$ as

$$\begin{cases} \dot\phi_1 - \epsilon_1\Delta\phi_1 = \bar f_{11}\,\phi_1 + \bar f_{21}\phi_2, & x \in \Omega, 0 < t \leq t_n, \\ \dot\phi_2 = \bar f_{12}\,\phi_1 + \bar f_{22}\phi_2, & x \in \Omega, 0 < t \leq t_n, \\ \phi_1(x,t) = 0, & x \in \Omega, 0 < t \leq t_n, \\ \phi(x,0) = \phi_n(x), & x \in \Omega. \end{cases} \tag{7.9}$$

Taking the inner product of the differential equations with ϕ then estimating by using the uniform bound on $\bar f_{ji}$ together with the Cauchy-Schwarz inequality, and then concluding by integrating and making a Gronwall argument, yields the first estimate. The second and last estimates follow directly. Next, we take the inner product of the first differential equation in (7.9) with $-\Delta\phi_1$ and use Green's formula on the left and the Cauchy-Schwarz inequality on the right to get

$$\frac{1}{2}\frac{d}{dt}\|\nabla\phi_1\|^2 + \epsilon_1\|\Delta\phi_1\|^2 \leq L_n^2\|\phi_1\|^2 + \frac{1}{4}\|\Delta\phi_1\|^2 + L_n^2\|\phi_2\|^2 + \frac{1}{4}\|\Delta\phi_1\|^2.$$

Kicking the obvious terms, integrating, and using the previous results gives the third estimate. The fourth estimate follows directly.

8. Details of the analysis in Section 5.

Proof of Proposition 5.1.

To simplify the presentation, we once again present the proofs for cG(1) method for the system (7.3) of a coupled parabolic and ordinary differential equation. As before, the proofs extend to all of the methods applied to the general problem (3.19).

To obtain (5.2), we take the inner product of the parabolic equation in (7.3) with $\dot u_1$ and get

$$\|\dot u_1\|^2 + \epsilon_1\frac{1}{2}\frac{d}{dt}\|\nabla u_1\| = (f_1(u), \dot u_1).$$

We estimate using the Cauchy-Schwarz inequality on the right together with the bound on f and then integrate to obtain the desired estimate. To show (5.3), we choose $W_1 = \dot U_1 \in V_n$ in (7.6) and estimate using the Cauchy-Schwarz inequality on the right and the uniform bound on f to get

$$\int_{I_n}\|\dot U_1\|^2\,dt + \epsilon_1\frac{1}{2}\int_{I_n}\frac{d}{dt}\|\nabla U_1\|^2\,dt \leq \frac{1}{2}M^2k_m = \frac{1}{2}\int_{I_n}\|\dot U_1\|^2\,dt.$$

Kicking the last term on the right and computing the second integral on the left, we obtain

$$\int_{I_n} \|\dot{U}_1\|^2\,dt + \epsilon_1\|\nabla U^-_{n,1}\|^2 \le \epsilon_1\|\nabla U^+_{n-1,1}\|^2 + k_n M^2.$$

If there is no mesh change across t_{n-1} or if there was refinement and the meshes are nested, then $U^+_{n-1} = U^-_{n-1}$. Otherwise, there is a constant C such that $\|\nabla P_n U^-_{n-1}\| \le C\|\nabla U^+_{n-1}\|$. Assuming that only a finite number of mesh coarsenings are permitted on any fixed time interval, we can repeat this argument to conclude that

$$\int_0^{t_n} \|\dot{U}_1\|^2\,dt + \epsilon_1\|\nabla U^-_{n,1}\|^2 \le C\|\nabla U^-_{0,1}\|^2 + CM^2 t_n.$$

This give the result.

Proof of Theorem 5.3. Since the proof is based on controlling the error between U and the local solution $\tilde{u}$ starting at time t_{n-1} and using the fact that $\tilde{u}$ remains inside an invariant rectangle, we could prove the theorem directly if the a posteriori error bound was pointwise in space. Instead, we use the equivalence of norms on a finite dimensional space to derive a pointwise estimate on a quantity in the finite element space from an L_2 estimate. To do this, we compare U to the pointwise interpolant $Q\tilde{u}$. Since generalized rectangles are convex, $Q\tilde{u}$ is contained inside any rectangle that contains $\tilde{u}$.

For any $m \ge n$,

$$\begin{aligned}\|U^-_m - Q\tilde{u}_m\|_{L_\infty(\Omega)} &\le \|h_m^{-1/2}(U^-_m - Q\tilde{u}_m)\| \\ &\le \|h_m^{-1/2}(\tilde{u}_m - Q\tilde{u}_m)\| + \|h_m^{-1/2}(U^-_m - \tilde{u}_m)\|.\end{aligned} \tag{8.1}$$

We bound the first term on the right using the standard interpolation estimate

$$\|h_m^{-1/2}(\tilde{u} - Q\tilde{u})\| \le C\|h_m^{3/2}\Delta\tilde{u}\|,$$

where C is an interpolation constant. At first glance, this is an a priori expression. But the regularity estimates in Proposition 3.5 imply that we can bound derivatives of $\tilde{u}$ in terms of derivatives of the initial data $\tilde{u}_{n-1}$, which are determined by U^-_{n-1}, and thus get an a posteriori expression. Assuming that $\tilde{u}$ is contained in $\mathcal{R}_o$, we take the uniform Lipschitz constant L in (3.21) to get

$$\|h_m^{-1/2}(\tilde{u} - Q\tilde{u})\| \le C\big(\|h_m^{3/2}\Delta\tilde{u}_{n-1}\| + \|h_m^{3/2}\nabla\tilde{u}_{n-1}\|\big)e^{2L(t_m - t_{n-1})}.$$

Finally, we use (7.2) and the assumption $h_m = h_n$ to conclude that

$$\|h_m^{-1/2}(\tilde{u} - Q\tilde{u})\| \le C\big(\|h_n^{3/2}\Delta_{h_n}U^-_{n-1}\| + \|h_n^{3/2}\nabla U^-_{n-1}\|\big)e^{2L(t_m - t_{n-1})}. \tag{8.2}$$

We begin estimating the second quantity on the right in (8.1) by using the a posteriori bound (2.21), assuming that $h_n = h_{n+1} = \cdots = h_m$:

$$\begin{aligned}\|h_m^{-1/2}\tilde{e}^-_m\| \le{}& S_0(t_{n-1},t_m)\|h_n^{-1/2}\tilde{e}^-_{n-1}\| + CS^1_t(t_{n-1},t_m)\|h^{-1/2}kR_t(U)\|_{L_\infty(t_{n-1},t_m)} \\ &+ CS^p_x(t_{n-1},t_m)\big(\|h^{3/2}R^p_x(U)\|_{L_\infty(t_{n-1},t_m)} + \|h^{3/2}R^p_2(U)\|_{L_\infty(t_{n-1},t_m)}\big).\end{aligned}$$

Assuming that U is contained in $\mathcal{R}_o$, (5.4) implies that

$$\begin{aligned}\|h_m^{-1/2}\tilde{e}_m^-\| &\leq C_1 e^{C_2(t_m-t_{n-1})}\big(\|h_n\Delta_{h_n}U_{n-1}^{p,-}\| + \|\nabla U_{n-1}^{p,-}\|\big)\\ &\quad\cdot\big(\|h_n^{-1/2}\tilde{e}_{n-1}^-\| + \|h^{-1/2}kR_t(U)\|_{L_\infty(t_{n-1},t_m)}\\ &\qquad + \|h^{3/2}R_x^p(U)\|_{L_\infty(t_{n-1},t_m)} + \|h^{3/2}R_2^p(U)\|_{L_\infty(t_{n-1},t_m)}\big),\end{aligned} \tag{8.3}$$

where $C_1 = C_1(\epsilon, M, L)$ and $C_2 = C_2(L)$ are fixed constants.

Putting (8.3) and (8.2) together with Lemma 3.1, we conclude that as long as U and $\tilde{u}$ are contained in $\mathcal{R}_o$,

$$\begin{aligned}&\|U_m^- - Q\tilde{u}_m\|_{L_\infty(\Omega)}\\ &\qquad\leq C_1 e^{C_2(t_m-t_{n-1})}\big(1 + \|h_n\Delta_{h_n}U_{n-1}^{p,-}\| + \|\nabla U_{n-1}^{p,-}\|\big)\\ &\qquad\quad\cdot\big(\|h^{-1/2}k^\alpha R_t(U)\|_{L_\infty(t_{n-1},t_m)} + \|h^{3/2}R_x^p(U)\|_{L_\infty(t_{n-1},t_m)}\\ &\qquad\quad + \|h^{3/2}R_2^p(U)\|_{L_\infty(t_{n-1},t_m)} + \|h_n^{3/2}\Delta_{h_n}U_{n-1}^-\| + \|h_n^{3/2}\nabla U_{n-1}^-\|\big).\end{aligned} \tag{8.4}$$

Note that the assumption of a uniform mesh size for $m > n$ is critical to the conversion from a pointwise to L_2 bound on the error because the a posteriori error bound on the error at at time node t_m depends on the mesh sizes and time steps on all the previous intervals. The result still holds, with a different constant C_1, if we allow a fixed finite number of mesh refinements in the previous steps, assuming that the ratio of a refined mesh size and the previous mesh size is bounded below uniformly, and any number of mesh coarsenings.

We treat the proof in two cases, first assuming that for some $n \geq 1$, $U_{n-1}^-(x) \in \mathcal{R}_i$ for all $x \in \Omega$ and showing that $U_n^- \in \mathcal{R}_{2\rho}$ for all $x \in \Omega$. Since U_{m-1}^- is the standard continuous piecewise linear finite element approximation of $\tilde{u}_{m-1}$, for any positive integer m, it is possible to prove the pointwise error bound

$$\|U_{m-1}^- - \tilde{u}_{m-1}\|_{L_\infty(\Omega)} \leq C\|\log(h_m)\, h_m^2\Delta_{h_m}U_{m-1}^-\|_{L_\infty(\Omega)}, \tag{8.5}$$

see Eriksson [19] and Eriksson and Johnson [24]. Therefore as long as (5.7) holds, $\tilde{u}_{n-1}$ is contained in $\mathcal{R}_\rho$ and therefore $\tilde{u}$ remains in $\mathcal{R}_\rho$ for all $t \geq t_{n-1}$. The idea of the proof is to control the size of the error to keep U from moving too far away from $\mathcal{R}_i$ over the next step. We control the size of the error by using the a posteriori error bound with the stability factors replaced by a bound.

By the local existence result, U_n^- exists uniquely and is contained in $\mathcal{R}_o$ for k_n sufficiently small depending on the fixed value $\operatorname{dist}(\partial\mathcal{R}_o, \partial\mathcal{R}_i) = r$. Hence, adjusting the value of C_1 in terms of C_1 in (8.3) and C in (5.7), if U_n^- is computed so (5.8) holds then (8.4) guarantees that U_n^- remains within $\mathcal{R}_{2\rho}$. We know that (5.8) can be achieved since $\tilde{u}$ and U are contained in $\mathcal{R}_o$ and we can apply Corollary 3.3 and Theorem 3.4 to bound the residual errors on I_m in terms of quantities that can be made as small as desired by refining the mesh and time step in a suitable way. Namely, there are constants $C_1 = C_1(\epsilon, M, L)$ and $C_2 = C_2(L)$ such that

$$\begin{aligned}&\|h_m{}^{-1/2}k_mR_t(U)\|_{L_\infty(I_m)} + \|h_m^{3/2}R_x^p(U)\|_{L_\infty(I_m)} + \|h_m^{3/2}R_2^p(U)\|_{L_\infty(t_{n-1},t_m)}\\ &\qquad\leq C_1 e^{C_2k_m}\big(\|(h_m^{3/2} + h_m^{-1/2}k_m + h_m^{-5/2}k_m^2)\Delta_{h_m}U_{m-1}^-\|\\ &\qquad\quad + \|(h_m^{3/2} + h_m^{-1/2}k_m + h_m^{-5/2}k_m^2)\nabla U_{m-1}^-\| + \|h_m^{3/2} + h_m^{-1/2}k_m + h_m^{-5/2}k_m^2\|\big).\end{aligned} \tag{8.6}$$

We apply this with $m = n$ to conclude that (5.8) is satisfied for all sufficiently small time steps and mesh size depending on $\|\Delta_{h_n}U_{n-1}^-\|$ and $\|\nabla U_{n-1}^-\|$.

Now consider the case when U_{n-1}^- is contained in $\mathcal{R}_{2\rho}$ but not in $\mathcal{R}_i$. As in the first case, we control the distance between U and $\mathcal{R}_i$ by controlling the error between U and the local solution $\tilde{u}$ beginning at t_{n-1} using the a posteriori error bound until the time that the local solution $\tilde{u}$ enters into $\mathcal{R}_i$.

However, now the argument possibly requires control of the error over several time steps. Theorem 3.4 guarantees that the residual errors on any interval I_m, $m \geq n$, can be made smaller by refinement provided U and the local solutions associated to each time node t_m remain in $\mathcal{R}_o$. By (8.5), this holds with a suitably refined mesh on any given interval I_m as long as U_{m-1}^- is a fixed distance away from the boundary of $\mathcal{R}_o$. Unfortunately, this local requirement for a suitably refined mesh conflicts with the requirement of uniform mesh sizes for the time steps on which we compare U to the local solution $\tilde{u}$ beginning at t_{n-1} arising from the conversion from pointwise to L_2 bounds. The same conflict arises when we impose a maximum value on the residual errors on each interval since the estimate (8.6) guarantees we can satisfy a given tolerance but only with local mesh refinement.

The resolution of this conflict is to derive an energy estimate bounding the discrete second derivative $\Delta_{h_m} U_{m-1}^-$ at t_{m-1} in terms of the discrete second derivative $\Delta_{h_n} U_{n-1}^-$ at the "initial" time t_{n-1} analogous to (3.21). This allows the mesh and time steps to be chosen a priori to satisfy the necessary local requirements, giving (5.9) and (5.10). We used the analogous property of $\tilde{u}$ to prove (8.2). The result is

PROPOSITION 8.1. *Under the assumptions of Propositions 3.5 and 3.6 and assuming in addition that U remains in $\mathcal{R}_o$ and a constant space mesh and a non-increasing sequence of time steps is used on the intervals between t_{n-1} and t_m, there is a constant $C = C(\epsilon, L)$ such that*

$$\|\nabla U_m^-\| \leq C e^{C(t_m - t_{n-1})} \left(\|\nabla U_{n-1}^-\| + k_n \|\Delta_{h_n} U_{n-1}^-\| \right), \tag{8.7}$$

$$\|\Delta_{h_{m+1}} U_m^-\| \leq C \left(1 + (t_m - t_{n-1}) e^{C(t_m - t_{n-1})}\right) \left(\|\nabla U_{n-1}^-\| + \|\Delta_{h_n} U_{n-1}^-\| \right). \tag{8.8}$$

Proof of Proposition 8.1.

To simplify the presentation, we once again present the proofs for cG(1) method for the system (7.3) of a coupled parabolic and ordinary differential equation. As before, the proofs extend to all of the methods applied to the general problem (3.19).

To prove (8.7), we choose $W_1 = -\Delta_{h_m} U_{m,1}^-$ in the first equation in (7.6) (substituting m for n), use Green's formula on the first and last terms in the resulting equation, where the boundary conditions on U_1 and the compatibility conditions on f insure the boundary integrals are zero, and estimate using the uniform Lipschitz constant of f in $\mathcal{R}_o$ to get

$$\begin{aligned}
&\|\nabla U_{m,1}{}^-\|^2 + \frac{\epsilon k_m}{2} \|\Delta_{h_m} U_{m,1}^-\|^2 \\
&\quad \leq \|\nabla U_{m-1,1}^+\| \, \|\nabla U_{m,1}^-\| + \frac{\epsilon k_m}{2} \|\Delta_{h_m} U_{m-1,1}^+\| \, \|\Delta_{h_m} U_{m,1}^-\| \\
&\qquad + L k_m \|\nabla U_{m,1}^-\|^2 + L k_m \|\nabla U_{m-1,1}^+\| \, \|\nabla U_{m,1}^-\| + L k_m \|\nabla U_{m-1,2}^+\| \, \|\nabla U_{m,1}^-\| \\
&\qquad\quad + L k_m \|\nabla U_{m,2}^-\| \, \|\nabla U_{m,1}^-\|.
\end{aligned}$$

Adding the analogous estimate that results from choosing $W_2 = -\Delta_{h_m} U_{m,2}^-$ and using the Cauchy-Schwarz inequality several times and the uniform mesh assumption, we

obtain

$$\|\nabla U_m^-\|^2 + \frac{\epsilon k_m}{2}\|\Delta_{h_m} U_{m,1}^-\|^2 \le \|\nabla U_{m-1}^-\|^2 + 6Lk_m\|\nabla U_m^-\|^2 + 6Lk_m\|\nabla U_{m-1}^-\|^2 + \frac{\epsilon k_m}{2}\|\Delta_{h_m} U_{m-1,1}^-\|^2.$$

Assuming that $6Lk_m < 1$ and using the assumption of non-increasing time steps,

$$\|\nabla U_m^-\|^2 + \frac{\epsilon k_m}{2}\|\Delta_{h_m} U_{m,1}^-\|^2 \le \frac{1+6Lk_m}{1-6Lk_m}\big(\|\nabla U_{m-1}^-\|^2 + \frac{\epsilon k_{m-1}}{2}\|\Delta_{h_{m-1}} U_{m-1,1}^-\|^2\big).$$

A discrete Gronwall argument gives (8.7).

To prove (8.7), we choose $W_1 = -\Delta_{h_m}\dot{U}_{m,1}^-$ in the first equation in (7.6), use Green's formula on the first and last terms in the resulting expression, and estimate using the Cauchy-Schwarz inequality and the uniform Lipschitz constant of f to get

$$\int_{I_m} \|\nabla \dot{U}_1\|^2\,dt + \epsilon\|\Delta_{h_{m+1}} U_m^-\|^2 \le \epsilon\|\Delta_{h_m} U_{m-1}^-\|^2 + L^2\int_{I_n}\|\nabla U\|^2\,dt.$$

Continuing the estimate over the preceding intervals, the result follows from (5.3) and the fact that $\|\nabla U\| \le \|\nabla U_m^-\| + \|\nabla U_{m-1}^+\|$ on I_m.

Remark 8.1. The result holds if we allow only a finite number of time step increases on any fixed interval provided that the ratio of consecutive time steps is bounded above uniformly.

The estimates (8.5) and (5.9) imply that $\tilde{u}_{n-1}$ is contained in $\mathcal{R}_{3\rho}$. Therefore $\tilde{u}$ remains in $\mathcal{R}_{3\rho}$ for $t \ge t_{n-1}$ and furthermore (5.6) implies that for all $t \ge t^*$ with

$$t^* - t_{n-1} = \tau = \frac{l_i 3\rho}{2\gamma M},$$

$\tilde{u}$ is contained in $\mathcal{R}_i$. Note that τ is independent of the particular local solution we use. By induction we show that (5.10) and (8.4) imply that the pointwise distance between U_m and $Q\tilde{u}(t_m)$ remains less than ρ for $n-1 \le m \le m^*$ and therefore U_m^- is contained in $\mathcal{R}_{4\rho}$ for $n-1 \le m < m^*$ and $U_{m^*}^-$ is contained in $\mathcal{R}_\rho$.

If necessary, we decrease the value of k_n determined by (5.10) depending on ρ, L, and M to guarantee that U_m^- exists uniquely and is contained in $\mathcal{R}_o$ whenever U_{m-1}^- is contained in $\mathcal{R}_{4\rho}$ for any $m \ge n$ and use this time step until t_{m^*}. This choice can be made uniformly by the same reasoning that gives the global existence of the approximant from the local existence result under the assumption of an invariant rectangle.

On the first step, since U_n^- and $\tilde{u}$ are both contained in $\mathcal{R}_o$, (8.4), (8.6), and Proposition 8.1 together imply that there are constants C_1 and C_2 as above such that

$$\begin{aligned}\|U_m^- - Q\tilde{u}_m\|_{L_\infty(\Omega)} &\le C_1 e^{C_2(t_m - t_{n-1})}\big(1 + \|h_n\Delta_{h_n} U_{n-1}^{p_i-}\| + \|\nabla U_{n-1}^{p_i-}\|\big)\\ &\cdot \big(\|(h_n^{3/2} + h_n^{-1/2}k_n + h_n^{-5/2}k_n^2)\Delta_{h_n} U_{n-1}^-\|\\ &\quad + \|(h_n^{3/2} + h_n^{-1/2}k_n + h_n^{-5/2}k_n^2)\nabla U_{n-1}^-\| + \|h_n^{3/2} + h_n^{-1/2}k_n + h_n^{-5/2}k_n^2\|\big)\end{aligned} \tag{8.9}$$

with $m = n$. Therefore (5.10) implies that U_n^- remains within $\mathcal{R}_{4\rho}$.

We define t_{m^*} to be the first node with $t_{m^*} \geq t^*$, where $m^* \leq \tau/k_n + 1$. Note that $t_{m^*} \leq t^* + K$. For $n \leq m \leq m^*$ and as long as U_{m-1}^- is contained in $\mathcal{R}_{4\rho}$ but not in $\mathcal{R}_{2\rho}$, we compute U_m^- uniquely contained in $\mathcal{R}_o$. Because of (5.9), (8.8), and (8.5), we know that the data for the local solution beginning at t_{m-1} is contained in $\mathcal{R}_{4\rho}$, so the a posteriori theory applies on this interval. Hence (5.10) and (8.9) imply that U_m^- is contained in $\mathcal{R}_{4\rho}$. By the definition of t_{m^*}, the result follows.

Proof of Theorem 5.4. We analyze the cGL method in detail. The proof for the dGL method is similar. Arguing as in Thomée [63] on a triangle K, there is a constant C such that

$$\left| \mathcal{Q}_K(g) - \int_K g\,dx \right| \leq Ch(K)^2 \sum_{|\alpha|=2} \|D^\alpha g\|_{L_1(K)}.$$

Applying this to $g = (\dot{U}_i - f_i(U))\pi P\phi_i$, noting that $D^2(\dot{U}_i - f_i(U)) = D^2\pi P\phi_i \equiv 0$, and using the stability of P with respect to the energy norm, we get

$$\begin{aligned}\Big| \mathcal{Q}_K\big((\dot{U}_i - f_i(U))\pi P\phi_i\big) - \int_K (\dot{U}_i - f_i(U))\pi P\phi_i\,dx \Big| \\ \leq C\|h^2 D^2 f_i(U)\|_{L_2(K)}\|\phi_i\|\|_{L_2(K)} + C\|h^2\nabla(\dot{U}_i - f_i(U))\|\|_{L_2(K)}\|\nabla\phi_i\|\|_{L_2(K)}.\end{aligned}$$

The result follows by using this estimate on each element in $\mathcal{T}_n$ after writing

$$\begin{aligned}\int_0^{t_n} \big(\dot{U}_i - f_i(U), \pi P\phi_i)_h - (\dot{U}_i - f_i(U), \pi P\phi_i)\big)\,dt \\ = \int_0^{t_n} \sum_{K\in\mathcal{T}_n} \left(\mathcal{Q}_K\big((\dot{U}_i - f_i(U))\pi P\phi_i\big) - \int_K (\dot{U}_i - f_i(U))\pi P\phi_i\,dx \right) dt.\end{aligned}$$

Proof of Theorem 5.7. We only analyze the cGL(1) method, since the analysis of the dGL methods is similar. We let $\vec{U}_{n,i}$ denote the vector of nodal values of $U_{n,i}^-$ associated to nodal basis $\{\psi_{n,i}\}$ for V_n on I_n, $\bar{B}_n : (\bar{B}_n)_{ij} = (\psi_{n,i}, \psi_{n,j})_{h_n}$ denote the *lumped mass matrix*, and A_n denote the *stiffness matrix*. $\bar{B}_n$ is a diagonal matrix and there is a constant μ_1 such that $(\bar{B}_n)_{ii} \geq \mu_1 h_{n,\min}^2$ while A_n has positive elements on the diagonal and non-positive off-diagonal entries and there is a constant μ_2 independent of the mesh such that $\max_i |(A_n)_{ii}| \leq \mu_2$. See Thomée [63]. We abuse notation to denote $f(\vec{U}) = \overrightarrow{Q_n f(U)}$.

The equation for the cG(1) method in matrix form reads

$$\bar{B}_n\big(\vec{U}_{n,i} - \overrightarrow{Q_n U}_{n-1,i}\big) + \frac{k_n}{2}A_n\big(\vec{U}_{n,i} + \overrightarrow{Q_n U}_{n-1,i}\big) = \int_{I_n} \bar{B}_n f_i(\vec{U})\,ds. \tag{8.10}$$

We show that for $1 \leq i \leq D$, $U_{n,i}^- - a_i \geq 0$ provided $U_{n-1}^- \in \mathcal{R}$ for all $x \in \Omega$. (8.10) can be rewritten as

$$\big(\bar{B}_n + \frac{k_n}{2}A_n\big)\big(\vec{U}_{n,i} - \vec{a}_i\big) = \big(\bar{B}_n - \frac{k_n}{2}A_n\big)\overrightarrow{Q_n U}_{n-1,i} - \vec{a}_i\big) + \int_{I_n} \bar{B}_n f_i(\vec{U})\,ds, \tag{8.11}$$

where $\vec{a}_i$ is the obvious vector with coefficients a_i. The first step is to show that the right-hand side of (8.11) is a vector with positive entries. For $1 \leq j \leq D$, we set

$$\tilde{a}_j = \begin{cases} \vec{U}_{n,j}, & j \neq i, \\ \vec{a}_i, & j = i. \end{cases}$$

Note that $\tilde{a} \in \partial\mathcal{R}$. Taylor's formula implies that

$$f_i(\vec{U}) = f_i(\tilde{a}) + \frac{\partial f_i}{\partial u_i}(\xi)(\vec{U}_i - \vec{a}_i),$$

for some $\xi \in \mathcal{R}$. With $\{\phi_n, \phi_{n-1}\}$ denoting the Lagrange basis functions for $\mathcal{P}^1(I_n)$ associated to the endpoints of I_n, $\vec{U}(t) = \vec{U}_n\phi_n(t) + \overrightarrow{Q_nU}_{n-1}\phi_{n-1}(t)$ for $t \in I_n$. We substitute this into (8.11) to obtain

$$\begin{aligned}&\left(\left(I - \int_{I_n} \frac{\partial f_i}{\partial u_i}(\xi)\phi_n\,dt\right)\bar{B}_n + \frac{k_n}{2}A_n\right)(\vec{U}_{n,i} - \vec{a}_i)\\ &= \left(\left(I - \int_{I_n} \frac{\partial f_i}{\partial u_i}(\xi)\phi_{n-1}\,dt\right)\bar{B}_n - \frac{k_n}{2}A_n\right)(\overrightarrow{Q_nU}_{n-1} - \vec{a}_i) + k_n\bar{B}_n f_i(\tilde{a}). \quad (8.12)\end{aligned}$$

Now we note that the coefficients of $f_i(\tilde{a})$ are non-negative since $\tilde{a} \in \partial\mathcal{R}$. Furthermore, the matrix

$$\left(\left(I - \int_{I_n} \frac{\partial f_i}{\partial u_i}(\xi)\phi_{n-1}\,dt\right)\bar{B}_n - \frac{k_n}{2}A_n\right)$$

has positive elements since every diagonal element is larger than $h_n^2(1-\mu_1/2-\mu_2/2) \geq 0$ and the off diagonal elements are positive. Since $\overrightarrow{Q_nU}_{n-1} - \vec{a}_i$ is a vector with positive entries, the right-hand side of (8.12) is a positive vector. On the left-hand side of (8.12), the matrix

$$\left(\left(I - \int_{I_n} \frac{\partial f_i}{\partial u_i}(\xi)\phi_n\,dt\right)\bar{B}_n + \frac{k_n}{2}A_n\right)$$

has positive elements on the diagonal and negative non-diagonal elements, and the following proposition, proved in Thomée [63], applies:

PROPOSITION 8.2. *Let I be the identity matrix and G be a matrix with $G_{ii} \geq 0$ and $G_{ij} \leq 0$ for $i \neq j$. Then the matrix $(I+G)^{-1}$ maps vectors with positive coefficients into vectors with positive coefficients.*

This gives the desired conclusion. Similarly, we can show that $b_i - U_{n,i}^- \geq 0$ for each i.

REFERENCES

[1] S. ADJERID AND J. FLAHERTY, *A local refinement finite element method for two-dimensional parabolic systems*, SIAM J. Sci. Statist. Comput., 9 (1988), pp. 792–810.

[2] ———, *A posteriori error estimation with finite element methods of lines for one-dimensional parabolic systems*, Numer. Math., 65 (1993), pp. 1–21.

[3] D. ARONSON AND H. WEINBERGER, *Nonlinear diffusion in population genetics, combustion, and nerve propagation*, in Proceedings of the Tulane Program in P.D.E.'s, Springer Lecture Notes, **446**, New York, 1975, pp. 5–49.

[4] M. BIETERMAN AND I. BABUŠKA, *The finite element method for parabolic equations, II. A posteriori error estimations and adaptive approach*, Numer. Math., 40 (1982), pp. 373–406.

[5] ———, *An adaptive method of lines with error control for parabolic equations of the reaction–diffusion type*, J. Comput. Phys., 63 (1986), pp. 33–66.

[6] L. BRONSARD AND R. KOHN, *On the slowness of phase-boundary motion in one space dimension*, Comm. Pure Appl. Math., 43 (1990), pp. 983–997.

[7] ———, *Motion by mean-curvature as the singular limit of Ginzburg-Landau dynamics*, J. Diff. Eqns., 90 (1991), pp. 211–237.

[8] K. Brown, P. Donne, and R. Gardner, *A semilinear parabolic system arising in the theory of superconductivity*, J. Diff. Eqns., 40 (1981), pp. 232–252.
[9] J. Carr and R. Pego, *Metastable patterns in $u_t = \epsilon^2 \Delta u - f(u)$*, Comm. Pure Appl. Math., 42 (1989), pp. 523–576.
[10] N. Chafee, *The electric ballast resistor: homogeneous and nonhomogeneous equilibria*, in Nonlinear Differential Equations: Invariance Stability and Bifurcation, P. D. Motoni and L. Salvadori, eds., Academic Press, New York, 1981.
[11] C. Chen and V. Thomée, *The lumped mass finite element method for a parabolic problem*, J. Austral. Math. Soc. Ser. B, 26 (1985), pp. 329–354.
[12] K. Chueh, C. Conley, and J. Smoller, *Positively invariant regions for systems of nonlinear diffusion equations*, Ind. Uni. Math. J., 26 (1977), pp. 373–392.
[13] P. Ciarlet, *The Finite Element Method for Elliptic Problems*, North–Holland, New York, 1978.
[14] H. Cohen, *Nonlinear diffusion problems*, MAA Studies in Math., 7 (1971), pp. 27–64.
[15] J. Cooley and F. Dodge, *Digital computer solutions for excitation and propagation of the nerve impulse*, Biophys. J., 6 (1966), pp. 583–599.
[16] M. Crouzeix and V. Thomée, *The stability in L_p and W_p^1 of the L_2-projection onto finite element spaces*, Math. Comp., 48 (1987), pp. 521–532.
[17] T. Dupont, *Mesh modification for evolution equations*, Math. Comp., 39 (1982), pp. 85–107.
[18] C. Elliot and A. Stuart, *The global dynamics of discrete semilinear parabolic equations*, SIAM J. Numer. Anal., 30 (1993), pp. 1622–1663.
[19] K. Eriksson, *An adaptive finite element method with efficient maximum norm error control for elliptic problems*, Preprint #1993–20, Department of Mathematics, Chalmers University of Technology, Göteborg, 1993.
[20] K. Eriksson, D. Estep, P. Hansbo, and C. Johnson, *Introduction to adaptive methods for differential equations*, Acta Numerica, (1995), pp. 105–158.
[21] ———, *Computational Differential Equations*, Cambridge University Press, New York, 1996.
[22] K. Eriksson and C. Johnson, *Error estimates and automatic time step control for non–linear parabolic problems*, SIAM J. Numer. Anal., 24 (1987), pp. 12–23.
[23] ———, *Adaptive finite element methods for parabolic problems I: A linear model problem*, SIAM J. Numer. Anal., 28 (1991), pp. 43–77.
[24] ———, *Adaptive finite element methods for parabolic problems II: Optimal error estimates in $L_\infty(L_2)$ and $L_\infty(L_\infty)$*, SIAM J. Numer. Anal., 32 (1995), pp. 706–740.
[25] ———, *Adaptive finite element methods for parabolic problems IV: Non–linear problems*, SIAM J. Numer. Anal., 32 (1995), pp. 1729–1749.
[26] ———, *Adaptive finite element methods for parabolic problems V: Long–time integration*, SIAM J. Numer. Anal., 32 (1995), pp. 1750–1763.
[27] D. Estep, *An analysis of numerical approximations of metastable solutions of the bistable equation*, Nonlinearity, 7 (1994), pp. 1445–1462.
[28] ———, *A posteriori error bounds and global error control for approximations of ordinary differential equations*, SIAM J. Numer. Anal., 32 (1995), pp. 1–48.
[29] D. Estep and D. French, *Global error control for the continuous Galerkin finite element method for ordinary differential equations*, RAIRO Modél. Math. Anal. Numér., 28 (1994), pp. 815–852.
[30] D. Estep, D. Hodges, and M. Warner, *Computational error estimation for a finite element solution of missile trajectory optimization problems.* in preparation.
[31] D. Estep and C. Johnson, *The computability of the Lorenz system.* submitted to J. Comput. Phys.
[32] D. Estep, C. Johnson, and M. Larsson, *A new approach to estimating the error of numerical methods for ordinary differential equations.* in preparation.
[33] D. Estep and S. Larsson, *The discontinuous Galerkin method for semilinear parabolic problems*, RAIRO Modél. Math. Anal. Numér., 27 (1993), pp. 35–54.
[34] D. Estep and A. Stuart, *The dynamical behavior of Galerkin finite element methods for ordinary differential equations and related difference schemes.* in preparation.
[35] D. Estep and R. Williams, *The structure of an adaptive differential equation solver*, Math. Models Meth. Appl. Sci., 6 (1996), pp. 535–568.
[36] ———, *Cards: Concurrent Adaptive Reaction-Diffusion Solver, 2.0*, 1997.
[37] P. Fife, *Dynamics of Internal Layers and Diffusive Interfaces*, SIAM, Philadelphia, 1988.
[38] D. French and S. Jensen, *Global dynamics of a discontinuous galerkin approximation to a class of reaction-diffusion equations*, Appl. Numer. Math., 18 (1995), pp. 473–487.
[39] R. Freund and N. Nachtigal, *QMR: A quasi-minimal residual method for non-hermitian linear systems*, Numer. Math., 60 (1991), pp. 315–339.

[40] ———, *QMRPACK: a package of QMR algorithms*, ACM Trans. Math. Software, 22 (1996), pp. 46–77.
[41] G. Fusco and J. Hale, *Slow-motion manifolds, dormant instability, and singular perturbations*, J. Dynam. Differ. Equa., 1 (1989), pp. 75–94.
[42] S. Hastings, *Some mathematical problems from neurobiology*, AMS Monthly, 82 (1975), pp. 881–895.
[43] P. Henrici, *Discrete Variable Methods in Ordinary Differential Equations*, John Wiley and Sons, Inc., New York, 1962.
[44] A. Hodgkin and A. Huxley, *A quantitative description of membrane current and its application to conduction and excitation in nerves*, J. Physiol., 117 (1952), pp. 500–544.
[45] D. Hoff, *Stability and convergence of finite difference methods for systems of nonlinear reaction-diffusion equations*, SIAM J. Numer. Anal., 15 (1978), pp. 1161–1177.
[46] 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), pp. 908–926.
[47] O. A. Ladyženskaja, V. A. Solonnikov, and N. N. Ural'cev, *Linear and Quasi-linear Equations of Parabolic Type*, American Mathematical Society, Providence, 1968.
[48] M. Larson, *Analysis of Adaptive Finite Element Methods*, PhD thesis, Chalmers University of Technology, S-412 96 Göteborg, Sweden, 1997.
[49] M. Mascagni, *The backward euler method for numerical solution of the hodgkin-huxley equations of nerve conduction*, SIAM J. Numer. Anal., 27 (1990), pp. 941–962.
[50] H. Matano, *Asymptotic behavior and stability of solutions of semilinear diffusion equations*, Publ. RIMS, Kyoto Univ., 15 (1979), pp. 401–454.
[51] M. Mimura, Y. Nishiura, and M. Yamaguti, *Some diffusive predator-prey systems and their bifurcation problems*, Ann. New York Acad. Sci., 316 (1979), pp. 490–510.
[52] P. Moore, *A posteriori error estimation with finite element semi- and fully discrete methods for nonlinear parabolic equations in one space dimension*, SIAM J. Numer. Anal., 31 (1994), pp. 149–169.
[53] J. D. Murray, *Mathematical Biology*, Springer-Verlag, New York, 1991.
[54] R. Nochetto, M. Paolini, and C. Verdi, *An adaptive finite-element method for 2-phase Stefan-problems in 2-space dimensions 1: stability and error estimates*, Math. Comp., 57 (1991), pp. 73–108.
[55] ———, *An adaptive finite-element method for 2-phase Stefan-problems in 2-space dimensions 2: implementation and numerical experiments*, SIAM J. Sci. Stat. Comput., 12 (1991), pp. 1207–1244.
[56] ———, *Towards a unified approach for the adaptive solution of evolution phase changes*, in Variational Problems: IMA Series in Mathematics and its Applications, Springer-Verlag, New York, 1993, pp. 1–15.
[57] J. E. Pearson, *Complex patterns in a simple system*, Science, 261 (1993), p. 189.
[58] R. Racke, *Lectures on Nonlinear Evolution Equations*, Max Planck Institüt für Mathematik, Bonn, 1992. Aspects of Mathematics: E, Vol. 19.
[59] J. Rauch and J. Smoller, *Qualitative theory of the Fitz-Hugh-Nagumo equations*, Adv. in Math., 27 (1978), pp. 12–44.
[60] P. Raviart, *The use of numerical integration in finite element methods for solving parabolic equations*, in Topics in Numerical Analysis I, J. Miller, ed., Academic Press, New York, 1973.
[61] R. Sandboge, *Adaptive Finite Element Methods for Reactive Flow Problems*, PhD thesis, Chalmers University of Technology, S-412 96 Göteborg, Sweden, 1996.
[62] J. A. Smoller, *Shock Waves and Reaction-Diffusion Equations*, Springer-Verlag, New York, 1983.
[63] V. Thomée, *Galerkin Finite Element Methods for Parabolic Problems*, Springer-Verlag, New York, 1980.
[64] W. Troy, *A threshold phenomenon in the Field-Noyes model of the Belousov-Zhabotinsky reaction*, J. Math. Anal. Appl., 58 (1977), pp. 233–248.
[65] H. Yserentant, *Hierarchical bases of finite-element spaces in the discretization of nonsymmetric elliptic boundary value problems*, Computing, 35 (1985), pp. 39–49.
[66] ———, *On the multi-level splitting of finite element spaces*, Numer. Math., 49 (1986), pp. 379–412.