Adaptive Computational Methods for Parabolic Problems

K. Eriksson, C. Johnson and A. Logg

Department of Computational Mathematics, Chalmers University of Technology, SE–412 96 Göteborg, Sweden

Abstract

We present a unified methodology for the computational solution of parabolic systems of differential equations with adaptive selection of discretization in space and time, based on a posteriori error estimates involving residuals of computed solutions and stability factors/weights, obtained by solving an associated linearized dual problem. We define parabolicity as boundedness in time (up to logarithmic factors) of a certain strong stability factor measuring the $L_1(L_2)$-norm in time–space of the time derivative of the dual solution with L_2-normalized initial data.

KEY WORDS: parabolic, finite elements, stiff, space-time discretization, Galerkin method, convection-diffusion-reaction, stability factor/weight, dual linearized problem, adaptive error control, time step control, a priori error estimate, a posteriori error estimate

The simpler a hypothesis is, the better it is. (Leibniz)

1. What is a parabolic problem?

A common classification of partial differential equations uses the terms *elliptic*, *parabolic* and *hyperbolic*, with the stationary Poisson equation being a prototype example of an elliptic problem, the time-dependent heat equation that of a parabolic problem, and the time-dependent wave equation being a hyperbolic problem. More generally, parabolic problems are often described vaguely speaking as “diffusion-dominated”, while hyperbolic problems are “convection-dominated” in a setting of systems of convection-diffusion equations. Alternatively, the term “stiff problems” is used to describe parabolic problems, with the term stiff referring to the characteristic presence of a range of time scales, varying from slow to fast with increasing damping.

In the context of computational methods for a general class of systems of time-dependent convection-diffusion-reaction equations, the notion of “parabolicity” or “stiffness” may be given a precise quantitative definition, which will be at the

Encyclopedia of Computational Mechanics. Edited by Erwin Stein, René de Borst and Thomas J.R. Hughes.

focal point of this presentation. We will define a system of convection-diffusion-reaction equations to be *parabolic* if computational solution is possible over long time without error accumulation, or alternatively, if a certain *strong stability factor* $S_c(T)$, measuring error accumulation, is of unit size independent of the length T in time of the simulation. More precisely, the error accumulation concerns the *Galerkin discretization error* in a *discontinuous Galerkin method* dG(q) with piecewise polynomials of degree q of order $2q+1$. (The total discretization error may also contain a *quadrature error*, which typically accumulates at a linear rate in time for a parabolic problem.) This gives parabolicity a precise quantitative meaning with a direct connection to computational methods. A parabolic problem thus exhibits a feature of "loss of memory" for Galerkin errors satisfying an orthogonality condition, which allows long-time integration without error accumulation. As shall be made explicit below, our definition of parabolicity through a certain stability factor is closely related to the definition of an *analytic semigroup*.

For a typical hyperbolic problem the corresponding strong stability factor will grow linearly in time, while for more general initial value problems the growth may be polynomial or even exponential in time.

The solutions of parabolic systems in general vary considerably in space-time and from one component to the other with occasional *transients* where derivatives are large. Efficient computational methods for parabolic problems thus require *adaptive* control of the mesh size in both space and time, or more general *multi-adaptive* control with possibly different resolution in time for different components.

2. Outline

We first consider in Section 4 time-stepping methods for Initial Value Problems (IVPs) for systems of ordinary differential equations. We present an a posteriori error analysis exhibiting the characteristic feature of a parabolic problem of non-accumulation of Galerkin errors in the setting of the backward Euler method (the discontinuous Galerkin method $\mathrm{dG}(0)$), with piecewise constant (polynomial of order 0) approximation in time. The a posteriori error estimate involves the residual of the computed solution and stability factors/weights obtained by solving an associated dual linearized problem expressing in quantitative form the stability features of the IVP being solved. The a posteriori error estimate forms the basis of an adaptive method for time step control with the objective of controlling the Euclidean norm of the error uniformly in time or at selected time levels, or some other output quantity. The form of the a posteriori error estimate expresses the characteristic feature of a parabolic problem that the time step control is independent of the length in time of the simulation.

In Section 5 we compute stability factors for a couple of IVPs modeling chemical reactions and find that the strong stability factor $S_c(T)$ remains of unit size over long time.

Encyclopedia of Computational Mechanics. Edited by Erwin Stein, René de Borst and Thomas J.R. Hughes.

In Section 6 we contrast with an IVP with exponentially growing stability factors: the Lorenz system.

The backward Euler method, or more generally the $\mathrm{dG}(q)$ method, is implicit and requires the solution of a nonlinear system of equations at each time step. In Section 7 we study iterative fixed point-type solution strategies resembling explicit time-stepping methods. However, since explicit time-stepping for stiff problems is unstable unless the time step is smaller than the fastest time scale, which may be unnecessarily restrictive outside fast transients, we include a stabilization technique based on adaptively stabilizing the stiff system by taking a couple of small time steps when needed. We show efficiency gain factors compared to traditional explicit methods with the time step restriction indicated, of the order 10-100 or more depending on the problem. The need for explicit-type methods for parabolic problems avoiding forming Jacobians and solving associated linear systems of equations, is very apparent for the large systems of convection-diffusion-reaction equations arising in the modeling of chemical reactors with many reactants and reactions involved. The need for explicit time-stepping also arises in the setting of multi-adaptive time-stepping with the time step varying in both space and for different reactants, since here the discrete equations may be coupled over several time steps for some of the subdomains (or reactants), leading to very large systems of algebraic equations.

In Section 8 we prove the basic strong stability estimates for an abstract parabolic model problem, and connect to the definition of an analytic semigroup.

In Sections 9–16 we present adaptive space-time Galerkin finite element methods for a model parabolic IVP, the heat equation, including a priori and a posteriori error estimates. The space-time Galerkin discretization method $\mathrm{cG}(p)\mathrm{dG}(q)$ is based on the continuous Galerkin method $\mathrm{cG}(p)$ with piecewise polynomials of degree p in space, and the discontinuous Galerkin method $\mathrm{dG}(q)$ with piecewise polynomials of degree q in time (for $q = 0, 1$). In Section 17 we discuss briefly the extension to convection-diffusion-reaction systems, and present computational results in Section 18.

3. References into the literature

We have decided to give a coherent concise presentation using duality-based space-time Galerkin methods (with some key references), rather than to try to give a survey of the work in the entire field of parabolic or stiff problems (with a massive number of references). However, we believe that our presentation may be viewed as a summary and condensation of much work prior to ours to which specific references can be found through the references given below. This work includes the pioneering work in the early 1970s initiated in Douglas and Dupont, 1970 on Galerkin methods for parabolic problems, and by Thomèe during the 80s and 90s summarized in Thomèe, 2002, including many references to related work by Luskin, Rannacher, Wheeler and many others.

It is of course also highly relevant to point to the world of stiff ordinary differential equations (stiff ode's) discovered by Dahlquist in the 50s, and explored by Deufelhard, Gear, Hairer, Lubich, Petzold, Wanner and many others. As general references into this world, we give the books by Hairer and Wanner, 1996 and Deufelhard and Bornemann, 2002. We give in the concluding Section 19 below our view of the relation between our methods for adaptive control of the global error and the methods for local error control commonly presented in the the ode-world. Concerning Galerkin methods for time discretization we also refer to the early work for ode's by Delfour, Hager and Trochu, 1981.

4. Introduction to adaptive methods for IVPs

We now give a brief introduction to the general topic of *adaptive error control* for numerical time-stepping methods for initial value problems, with special reference to parabolic or stiff problems. In an *adaptive method*, the time steps are chosen automatically with the purpose of controlling the numerical error to stay within a given tolerance level. The adaptive method is based on an *a posteriori error estimate* involving the *residual* of the computed solution and results of auxiliary computations of *stability factors*, or more generally *stability weights*.

We consider an IVP of the form

$$\dot{u}(t) = f(u(t)) \quad \text{for} \quad 0 < t \leq T, \quad u(0) = u^0, \tag{1}$$

where $f : \mathbb{R}^d \to \mathbb{R}^d$ is a given differentiable function, $u^0 \in \mathbb{R}^d$ a given initial value, and $T > 0$ a given final time. For the computational solution of (1), we let $0 = t_0 < t_1 < \ldots < t_{n-1} < t_n < \ldots < t_N = T$ be an increasing sequence of discrete time steps with corresponding time intervals $I_n = (t_{n-1}, t_n]$ and time steps $k_n = t_n - t_{n-1}$, and consider the *backward Euler method*: Find $U(t_n)$ successively for $n = 0, 1, \ldots, N$, according to the formula

$$U(t_n) = U(t_{n-1}) + k_n f(U(t_n)), \tag{2}$$

with $U(0) = u^0$. The backward Euler method is *implicit* in the sense that to compute the value $U(t_n)$ with $U(t_{n-1})$ already computed, we need to solve a system of of equations. We will return to this aspect below.

We associate a *function* $U(t)$ defined on $[0, T]$ to the nodal values $U(t_n)$, $n = 0, 1, \ldots, N$, as follows:

$$U(t) = U(t_n) \quad \text{for } t \in (t_{n-1}, t_n].$$

In other words, $U(t)$ is left-continuous piecewise constant on $[0, T]$ and takes the value $U(t_n)$ on I_n, and thus takes a jump from the limit from the left $U(t_{n-1}^-) = U(t_{n-1})$ to the limit from the right $U(t_{n-1}^+) = U(t_n)$ at the time level $t = t_{n-1}$. We can now write the backward Euler method in the form

$$U(t_n) = U(t_{n-1}) + \int_{t_{n-1}}^{t_n} f(U(t))\, dt,$$

Encyclopedia of Computational Mechanics. Edited by Erwin Stein, René de Borst and Thomas J.R. Hughes.

or equivalently

$$U(t_n) \cdot v = U(t_{n-1}) \cdot v + \int_{t_{n-1}}^{t_n} f(U(t)) \cdot v \, dt, \tag{3}$$

for all $v \in \mathbb{R}^d$ with the dot signifying the scalar product in $\mathbb{R}^d$. This method is also referred to as $\mathrm{dG}(0)$, the *discontinuous Galerkin method of order zero*, corresponding to approximating the exact solution $u(t)$ by a piecewise constant function $U(t)$ satisfying the *Galerkin orthogonality* condition (3).

The general $\mathrm{dG}(q)$ method takes the form (3), with the restriction to each time interval I_n of the solution $U(t)$ and the test function v on each time interval I_n being polynomial of degree q. The $\mathrm{dG}(q)$ method comes also in *multi-adaptive form* with each component and corresponding test function being piecewise polynomial with possibly different sequences of time steps for different components.

We shall now derive an *a posteriori error estimate*, aiming at control of the scalar product of the error $e(T) = (u - U)(T)$ at final time T with a given vector ψ, where we assume that ψ is normalized so that $\|\psi\| = 1$. Here, $\|\cdot\|$ denotes the Euclidean norm on $\mathbb{R}^d$. We introduce the following linearized dual problem running backward in time:

$$-\dot{\phi}(t) = A^\top(t)\phi(t) \quad \text{for } 0 \le t < T, \quad \phi(T) = \psi, \tag{4}$$

with

$$A(t) = \int_0^1 f'(su(t) + (1-s)U(t)) \, ds,$$

where $u(t)$ is the exact solution and $U(t)$ the approximate solution, f' is the Jacobian of f, and $\top$ denotes transpose. We note that $f(u(t)) - f(U(t)) = A(t)(u(t) - U(t))$. We now start from the identity

$$e(T) \cdot \psi = e(T) \cdot \psi + \sum_{n=1}^N \int_{t_{n-1}}^{t_n} e \cdot (-\dot{\phi} - A^\top \phi) \, dt,$$

and integrate by parts on each subinterval (t_{n-1}, t_n) to get the *error representation*:

$$e(T) \cdot \psi = \sum_{n=1}^N \int_{t_{n-1}}^{t_n} (\dot{e} - Ae) \cdot \phi \, dt - \sum_{n=1}^N (U(t_n) - U(t_{n-1})) \cdot \phi(t_{n-1}),$$

where the last term results from the jumps of $U(t)$ at the nodes $t = t_{n-1}$. Since now u solves the differential equation $\dot{u} - f(u) = 0$, and $\dot{U} = 0$ on each time interval (t_{n-1}, t_n), we have

$$\dot{e} - Ae = \dot{u} - f(u) - \dot{U} + f(U) = -\dot{U} + f(U) = f(U) \quad \text{on } (t_{n-1}, t_n).$$

It follows that

$$e(T) \cdot \psi = -\sum_{n=1}^N (U(t_n) - U(t_{n-1})) \cdot \phi(t_{n-1}) + \int_0^T f(U) \cdot \phi \, dt.$$

Encyclopedia of Computational Mechanics. Edited by Erwin Stein, René de Borst and Thomas J.R. Hughes.

Using the Galerkin orthogonality (3) with $v = \bar{\phi}_n$, the mean value of ϕ over I_n, we get

$$e(T)\cdot\psi = -\sum_{n=1}^{N}(U(t_n)-U(t_{n-1}))\cdot(\phi(t_{n-1})-\bar{\phi}_n)+\sum_{n=1}^{N}\int_{t_{n-1}}^{t_n} f(U)\cdot(\phi-\bar{\phi}_n)\,dt.$$

Since now

$$\int_{t_{n-1}}^{t_n} f(U)\cdot(\phi-\bar{\phi}_n)\,dt = 0,$$

because $f(U(t))$ is constant on $(t_{n-1}, t_n]$, the error representation takes the form

$$e(T)\cdot\psi = -\sum_{n=1}^{N}(U(t_n)-U(t_{n-1}))\cdot(\phi(t_{n-1})-\bar{\phi}_n).$$

Finally, from the estimate

$$\|\phi(t_{n-1})-\bar{\phi}_n)\| \le \int_{t_{n-1}}^{t_n}\|\dot{\phi}(t)\|\,dt,$$

we obtain the following *a posteriori error estimate* for the backward Euler or $\mathrm{dG}(0)$ method:

$$|e(T)\cdot\psi| \le S_c(T,\psi)\max_{1\le n\le N}\|U(t_n)-U(t_{n-1})\|, \tag{5}$$

where the *stability factor* $S_c(T,\psi)$, recalling (4), is defined by

$$S_c(T,\psi) = \int_0^T \|\dot{\phi}(t)\|\,dt. \tag{6}$$

Maximizing over ψ with $\|\psi\| = 1$, we obtain a posteriori control of the Euclidean norm of $e(T)$:

$$\|e(T)\| \le S_c(T)\max_{1\le n\le N}\|U(t_n)-U(t_{n-1})\|, \tag{7}$$

with corresponding stability factor

$$S_c(T) = \max_{\|\psi\|=1} S_c(T,\psi). \tag{8}$$

Equivalently, we can write this estimate as

$$\|e(T)\| \le S_c(T)\max_{0\le t\le T}\|k(t)R(U(t))\|, \tag{9}$$

where $k(t) = k_n = t_n - t_{n-1}$ for $t \in (t_{n-1}, t_n]$, and $R(U(t)) = (U(t_n) - U(t_{n-1}))/k_n = f(U(t_n))$ corresponds to the *residual* obtained by inserting the discrete solution into the differential equation (noting that $\dot{U}(t) = 0$ on each time interval).

We can thus also express the a posteriori error estimate (5) in the form

$$|e(T)\cdot\psi| \le \int_0^T k(t)R(U(t))\|\dot{\phi}(t)\|\,dt, \tag{10}$$

Encyclopedia of Computational Mechanics. Edited by Erwin Stein, René de Borst and Thomas J.R. Hughes.

where now the dual solution enters as a *weight* in a time integral involving the residual $R(U(t))$. Maximizing over $k(t)R(U(t))$ and integrating $\|\dot{\phi}(t)\|$ we obtain the original estimate (9).

We now define the IVP (1) to be *parabolic* if (up to possibly logarithmic factors) the stability factor $S_c(T)$ is of unit size for all T. We shall see that another typical feature of a parabolic problem is that the stability factor $S_c(T,\psi)$ varies little with the specific choice of normalized initial data ψ, which means that to compute $S_c(T) = \max_{\|\psi\|=1} S_c(T,\psi)$, we may drastically restrict the variation of ψ and solve the dual problem with only a few different initial data.

If we perturb f to $\hat{f}$ in the discretization with $\mathrm{dG}(q)$, for instance by approximating $f(U(t))$ by a polynomial connecting to *quadrature* in computing $\int_{I_n} f(U(t))\,dt$, we obtain an additional contribution to the a posteriori error estimates of the form

$$S_q(T,\psi)\max_{[0,T]}\|f(U(t)) - \hat{f}(U(t))\|,$$

or $S_q(T)\max_{[0,T]}\|f(U(t)) - \hat{f}(U(t))\|$, with corresponding stability factors defined by

$$S_q(T,\psi) = \int_0^T \|\phi(t)\|\,dt,$$

where ϕ solves the backward dual problem with $\phi(T) = \psi$, and $S_q(T) = \max_{\|\psi\|=1} S_q(T,\psi)$. In a parabolic problem we may have $S_q(T) \sim T$, although $S_c(T) \sim 1$ for all $T > 0$. We note that $S_c(T)$ involves the time derivative $\dot{\phi}$, while $S_q(T)$ involves the dual ϕ itself.

Note that in $\mathrm{dG}(0)$ there is no need for quadrature in the present case of an autonomous IVP, since then $f(U(t))$ is piecewise constant. However, in a corresponding non-autonomous problem of the form $\dot{u} = f(u(t),t)$ with f depending explicitly on t, quadrature may be needed also for $\mathrm{dG}(0)$.

The basic parabolic or stiff problem is a linear constant coefficient IVP of the form $\dot{u}(t) = f(u(t),t) \equiv -Au(t) + f(t)$ for $0 < t \leq T$, $u(0) = u^0$, with A a constant positive semidefinite symmetric matrix with eigenvalues ranging from small to large positive. In this case, $f'(u) = -A$ with eigenvalues $\lambda \geq 0$ and corresponding solution components varying on time scales $1/\lambda$ ranging from very long (slow variation/decay if λ is small positive) to very short (fast variation/decay if λ is large positive). A solution to a typical stiff problem thus has a range of time-scales varying from slow to fast. In this case the dual problem takes the form $-\dot{\phi}(t) = -A\phi(t)$ for $0 \leq t < T$, and the strong stability estimate states that, independent of the distribution of the eigenvalues $\lambda \geq 0$ of A, we have

$$\int_0^T (T-t)\|\dot{\phi}(t)\|^2\,dt \leq \frac{1}{4},$$

where we assume that $\|\phi(T)\| = 1$. From this we may derive that for $0 < \epsilon < T$,

$$\int_0^{T-\epsilon} \|\dot{\phi}(t)\|\,dt \leq \frac{1}{2}(\log(T/\epsilon))^{1/2},$$

Encyclopedia of Computational Mechanics. Edited by Erwin Stein, René de Borst and Thomas J.R. Hughes.

which up to a logarithmic factor states that $S_c(T) \sim 1$ for all $T > 0$. Further, the corresponding (weak) stability estimate states that $\|\phi(t)\| \le \|\psi\|$, from which directly follows that that $S_q(T) \le T$ as indicated. The (simple) proofs of the stability estimates are given below.

The stability factors $S_c(T,\psi)$ and $S_q(T,\psi)$ may be approximately computed a posteriori by replacing $A(t)$ in (4) with $f'(U(t))$, assuming $U(t)$ is sufficiently close to $u(t)$ for all t, and solving the corresponding backward dual problem numerically (e.g. using the $\mathrm{dG}(0)$ method). We may similarly compute approximations of $S_c(T)$ and $S_q(T)$ by varying ψ. By computing the stability factors we get concrete evidence of the parabolicity of the underlying problem, which may be difficult (or impossible) to assess analytically a priori. Of course, there is also a gradual degeneracy of the parabolicity as the stability factor $S_c(T)$ increases.

The a posteriori error estimate (7) can be used as the basis for an adaptive time-stepping algorithm, controlling the size of the Galerkin discretization error, of the form: For $n = 1, 2, \ldots, N$, choose k_n so that

$$\|U(t_n)) - U(t_{n-1}))\| \approx \frac{\mathrm{TOL}}{S_c(T)},$$

for some tolerance $\mathrm{TOL} > 0$. Recalling that the characteristic feature of a parabolic problem is that $S_c(T) \sim 1$ for all $T > 0$, this means that the time step control related to the Galerkin discretization error will be independent of the length of the time interval of the simulation. This means that long-time integration without error accumulation is possible, which may be interpreted as some kind of "parabolic loss of memory". We note again that this concerns the Galerkin error only, which has this special feature as a consequence of the Galerkin orthogonality. However, the quadrature error may accumulate in time typically at a linear rate, and so a long-time simulation may require more accurate quadrature than a simulation over a shorter interval.

Remark 4.1. We now present a very simple parabolic model problem, where we can directly see the basic feature of long-time integration without non-accumulation of dG(0), which we just proved for a general parabolic problem. The model problem is simply $\dot{u}(t) = f(t)$, where f does not depend on $u(t)$. For this problem, which is certainly parabolic with our definition, dG(0) takes the form

$$U(t_n) = U(t_{n-1}) + \int_{t_{n-1}}^{t_n} f(t)\,dt. \tag{11}$$

It follows that dG(0) (with exact quadrature) coincides with the exact solution at the discrete time levels t_n, and thus there is no error accumuation over long time at all. The reason is clearly the mean value property (11) of dG(0) expressing Galerkin orthogonality. On the other hand, if we use quadrature to compute the integral in (11), then we may expact the quadrature error in general to accumulate (at a linear rate).

5. Examples of stiff IVPs

We have stated above that a parabolic or stiff initial value problem $\dot{u}(t) = f(u(t))$ for $0 < t \leq T$, $u(0) = u^0$, may be characterized by the fact that the stability factor $S_c(T)$ is of moderate (unit) size independent of $T > 0$, while the norm of the linearized operator $f'(u(t))$ may be large, corresponding to the presence of large negative eigenvalues. Such initial value problems are common in models of chemical reactions, with reactions on a range of time scales varying from slow to fast. Typical solutions include so-called *transients* where the fast reactions make the solution change quickly over a short (initial) time interval, after which the fast reactions are "burned out" and the slow reactions make the solution change on a longer time scale. We now consider a set of test problems which we solve by the adaptive $\mathrm{dG}(0)$ method, including computation of the strong stability factor $S_c(T)$.

5.1. *Model problem: $\dot{u} + Au(t) = f(t)$ with A positive symmetric semidefinite*

As indicated, the basic example of a parabolic IVP takes the form $\dot{u} + Au(t) = f(t)$ for $0 < t \leq T$, $u(0) = u^0$, where A is a positive semidefinite square matrix. We consider here the case

$$A = \begin{pmatrix} -4.94 & 2.60 & 0.11 & 0.10 & 0.06 \\ 2.60 & -4.83 & 2.69 & 0.17 & 0.10 \\ 0.11 & 2.69 & -4.78 & 2.69 & 0.11 \\ 0.10 & 0.17 & 2.69 & -4.83 & 2.60 \\ 0.06 & 0.10 & 0.11 & 2.60 & -4.94 \end{pmatrix}$$

with eigenvalues $(0, -2.5, -5, -7.5, -9.33)$. In Figure 1, we plot the solution, a dual solution and the stability factor $S_c(T, \psi)$ as a function of T for a collection of different initial values $\phi(T) = \psi$. We note that the variation with ψ is rather small: about a factor 4. We also note the initial transient, both for the solution itself and for the dual problem running backwards in time.

5.2. *The Akzo-Nobel system of chemical reactions*

We next consider the so-called Akzo-Nobel problem, which is a test problem for solvers of stiff ODEs modeling chemical reactions: Find the concentrations $u(t) = (u_1(t), \ldots, u_6(t))$ such that for $0 < t \leq T$,

$$\left\{ \begin{array}{lcl} \dot{u}_1 & = & -2r_1 + r_2 - r_3 - r_4, \\ \dot{u}_2 & = & -0.5r_1 - r_4 - 0.5r_5 + F, \\ \dot{u}_3 & = & r_1 - r_2 + r_3, \\ \dot{u}_4 & = & -r_2 + r_3 - 2r_4, \\ \dot{u}_5 & = & r_2 - r_3 + r_5, \\ \dot{u}_6 & = & -r_5, \end{array} \right. \tag{12}$$

where $F = 3.3 \cdot (0.9/737 - u_2)$ and the reaction rates are given by $r_1 = 18.7 \cdot u_1^4 \sqrt{u_2}$, $r_2 = 0.58 \cdot u_3 u_4$, $r_3 = 0.58/34.4 \cdot u_1 u_5$, $r_4 = 0.09 \cdot u_1 u_4^2$ and $r_5 = 0.42 \cdot u_6^2 \sqrt{u_2}$,

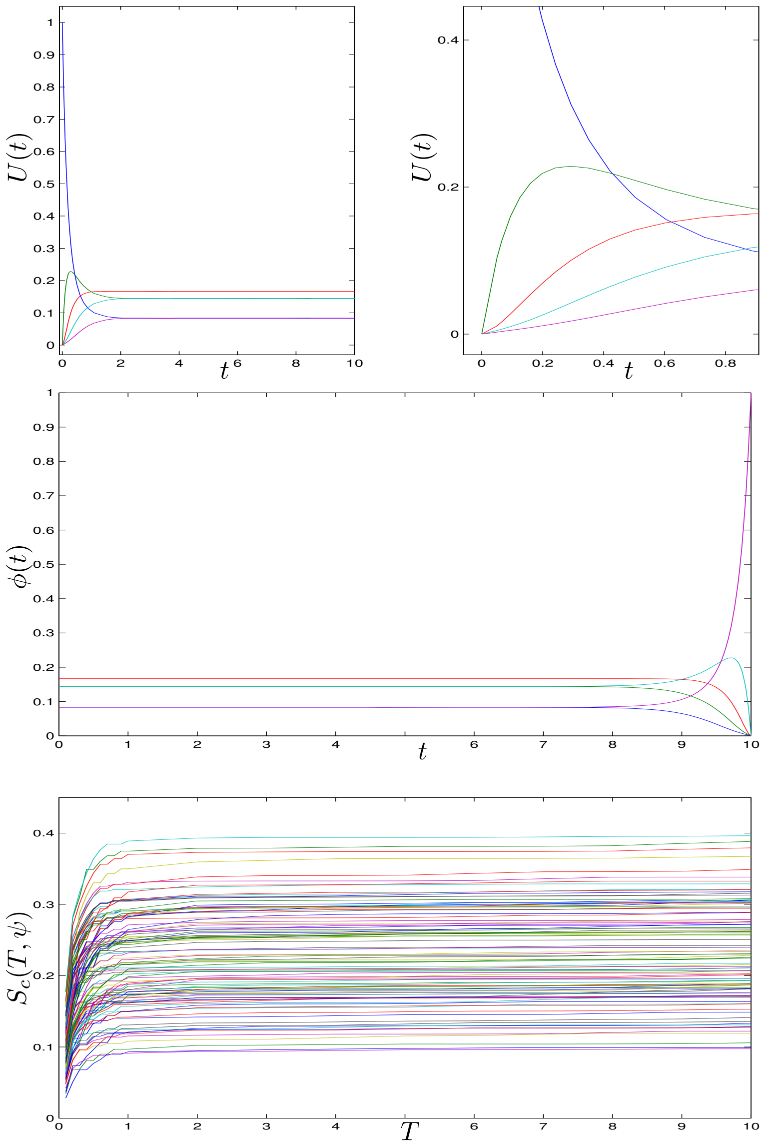

Figure 1. Symmetric IVP: solution, dual solution and stability factors $S_c(T,\psi)$.

Encyclopedia of Computational Mechanics. Edited by Erwin Stein, René de Borst and Thomas J.R. Hughes.

with the initial condition $u^0 = (0.437, 0.00123, 0, 0, 0, 0.367)$. In Figure 2 we plot the solution, a dual solution and the stability factor $S_c(T)$ as a function of T. We note the initial transients in the concentrations and their long-time very slow variation after the active phase of reaction. We also note that $S_c(T)$ initially grows to about 3.5 and then falls back to a value around 2. This is a typical behavior for reactive systems, where momentarily during the active phase of reaction the perturbation growth may be considerable, while over long-time the memory of that phase fades. On the other hand $S_q(T)$ grows consistently, which shows that fading memory requires some mean-value to be zero (Galerkin orthogonality). We present below more examples of this nature exhibiting features of parabolicity.

6. A non-stiff IVP: the Lorenz system

The *Lorenz system* presented 1972 by the meteorologist Edward Lorenz:

$$\begin{cases} \dot{u}_1 = -10u_1 + 10u_2, \\ \dot{u}_2 = 28u_1 - u_2 - u_1u_3, \\ \dot{u}_3 = -\frac{8}{3}u_3 + u_1u_2, \\ u(0) = u^0, \end{cases} \tag{13}$$

is an example of an IVP with exponentially growing stability factors reflecting a strong sensitivity to perturbations. Lorenz chose the model to illustrate perturbation sensitivity in meteorological models, making forecasts of daily weather virtually impossible over a period of more than a week. For the Lorenz system accurate numerical solution using double precision beyond 50 units of time seems impossible. Evidently, the Lorenz system is not parabolic.

The system (13) has three equilibrium points $\bar{u}$ with $f(\bar{u}) = 0$: $\bar{u} = (0, 0, 0)$ and $\bar{u} = (\pm 6\sqrt{2}, \pm 6\sqrt{2}, 27)$. The equilibrium point $\bar{u} = (0, 0, 0)$ is unstable with the corresponding Jacobian $f'(\bar{u})$ having one positive (unstable) eigenvalue and two negative (stable) eigenvalues. The equilibrium points $(\pm 6\sqrt{2}, \pm 6\sqrt{2}, 27)$ are slightly unstable with the corresponding Jacobians having one negative (stable) eigenvalue and two eigenvalues with very small positive real part (slightly unstable) and also an imaginary part. More precisely, the eigenvalues at the two non-zero equilibrium points are $\lambda_1 \approx -13.9$ and $\lambda_{2,3} \approx .0939 \pm 10.1i$.

In Figure 3, we present two views of a solution $u(t)$ that starts at $u(0) = (1, 0, 0)$ computed to time 30 with an error tolerance of $\mathrm{TOL} = 0.5$ using an adaptive IVP-solver of the form presented above. The plotted trajectory is typical: it is kicked away from the unstable point $(0, 0, 0)$ and moves towards one of the non-zero equilibrium points. It then slowly orbits away from that point and at some time decides to cross over towards the other non-zero equilibrium point, again slowly orbiting away from that point and coming back again, orbiting out, crossing over, and so on. This pattern of some orbits around one non-zero equilibrium point followed by a transition to

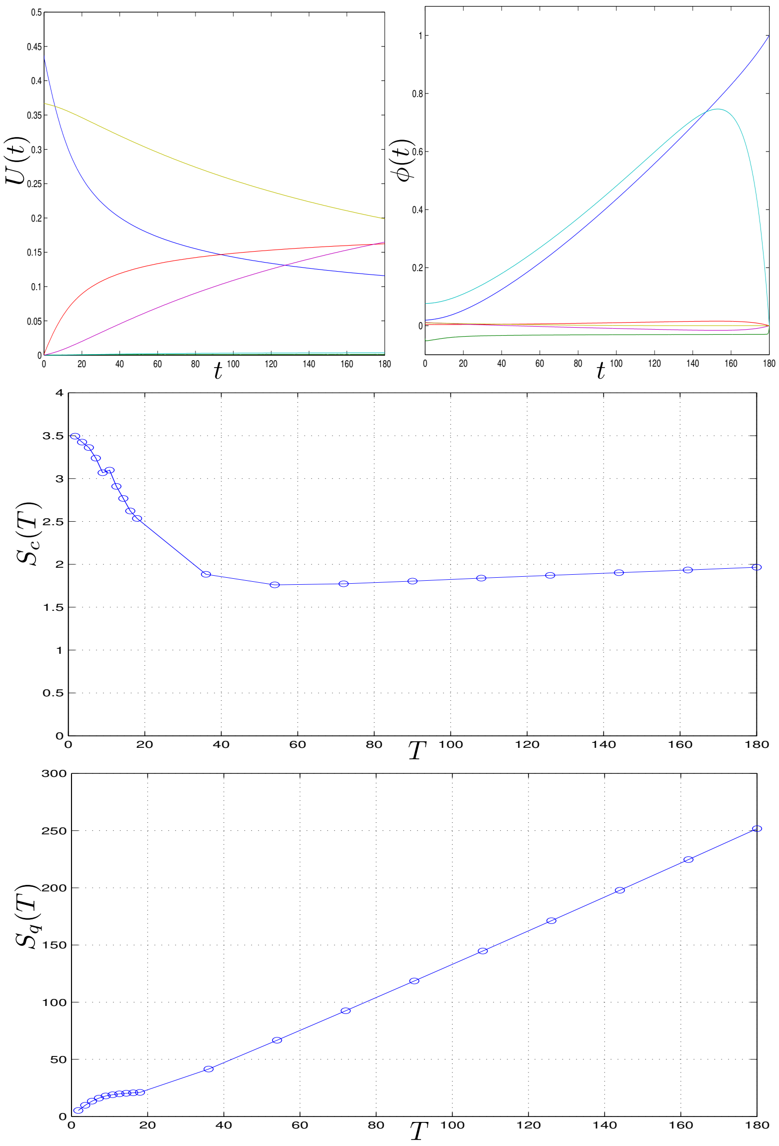

Figure 2. The Akzo-Nobel problem: solution, dual solution, stability factor $S_c(T, \psi)$, and stability factor $S_q(T, \psi)$.

the other non-zero equilibrium point is repeated with a seemingly random number of revolutions around each non-zero equilibrium point.

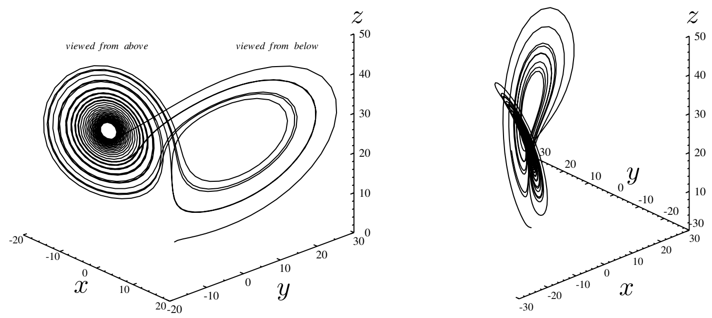

Figure 3. Two views of a numerical trajectory of the Lorenz system over the time interval $[0, 30]$.

In Figure 4, we plot the size of the stability factor $S_q(T)$ connected to quadrature errors as function of final time T. We notice that the stability factor takes an exponential leap every time the trajectory flips, while the growth is slower when the trajectory orbits one of the non-zero equilibrium points. The stability factor grows on the average as $10^{T/3}$ which sets the effective time limit of accurate computation to $T \approx 50$ computing in double precision with say 15 accurate digits.

7. Explicit time-stepping for stiff IVPs

The $\mathrm{dG}(0)$ method for the IVP $\dot{u} = f(u)$ takes the form

$$U(t_n) - k_n f(U(t_n)) = U(t_{n-1}).$$

At each time step we have to solve an equation of the form $v - k_n f(v) = U(t_{n-1})$ with $U(t_{n-1})$ given. To this end we may try a damped fixed point iteration in the form

$$v^{(m)} = (I - \alpha)v^{(m-1)} + \alpha(U(t_{n-1}) + k_n f(v^{(m-1)})),$$

with α some suitable matrix (or constant in the simplest case). Choosing $\alpha = I$ with only one iteration corresponds to the explicit Euler method. Convergence of the fixed point iteration requires that

$$\|I - \alpha + k_n \alpha f'(v)\| < 1,$$

for relevant values of v, which could force α to be small (e.g. in the stiff case with $f'(v)$ having large negative eigenvalues) and result in slow convergence. A simple

Encyclopedia of Computational Mechanics. Edited by Erwin Stein, René de Borst and Thomas J.R. Hughes.

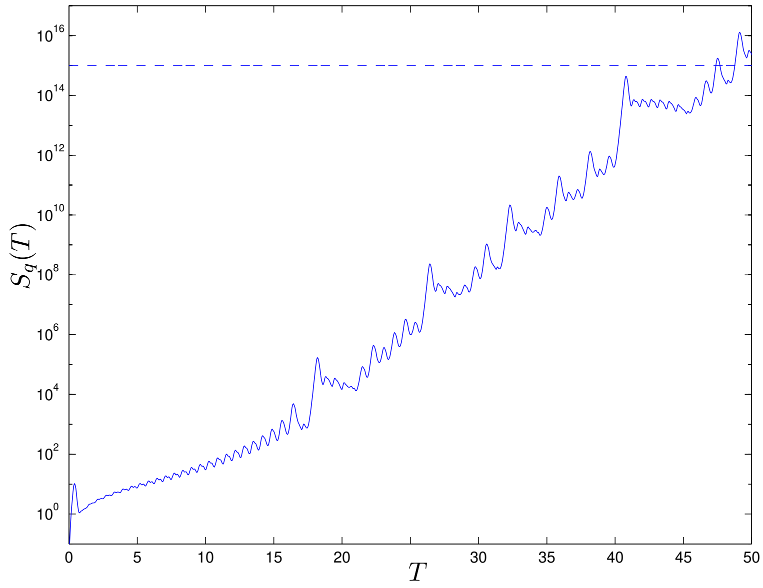

Figure 4. The growth of the stability factor $S_q(T)$ for the Lorenz problem.

choice is to take α to be a diagonal matrix with $\alpha_{ii} = 1/(1 - k_n f'_{ii}(v^{(m-1)}))$, corresponding to a diagonal approximation of Newton's method, with hope that the number of iterations will be small.

We just learned that explicit time-stepping for stiff problems requires small time steps outside transients and thus may be inefficient. We shall now indicate a way to get around this limitation through a process of stabilization, where a large time step is accompanied by a couple of small time steps. The resulting method has similarities with the control system of a modern (unstable) jet fighter like the Swedish JAS Gripen, the flight of which is controlled by quick small flaps of a pair of small extra wings ahead of the main wings, or balancing a stick vertically on the finger tips if we want a more domestic application.

We shall now explain the basic (simple) idea of the stabilization and present some examples, as illustrations of fundamental aspects of adaptive IVP-solvers and stiff problems. Thus to start with, suppose we apply the explicit Euler method to the scalar problem

$$\begin{aligned} \dot{u}(t) + \lambda u(t) &= 0 \quad \text{for } 0 < t \leq T, \\ u(0) &= u^0, \end{aligned} \tag{14}$$

with $\lambda > 0$ taking first a large time step K satisfying $K\lambda > 2$ and then m small time

Encyclopedia of Computational Mechanics. Edited by Erwin Stein, René de Borst and Thomas J.R. Hughes.

steps k satisfying $k\lambda < 2$, to get the method

$$U(t_n) = (1 - k\lambda)^m (1 - K\lambda) U(t_{n-1}), \tag{15}$$

altogether corresponding to a time step of size $k_n = K + mk$. Here K gives a large unstable time step with $|1 - K\lambda| > 1$ and k is a small time step with $|1 - k\lambda| < 1$. Defining the polynomial function $p(x) = (1 - \theta x)^m (1 - x)$, where $\theta = \frac{k}{K}$, we can write the method (15) in the form

$$U(t_n) = p(K\lambda) U(t_{n-1}).$$

For stability, we need

$$|p(K\lambda)| \leq 1, \quad \text{that is } |1 - k\lambda|^m (K\lambda - 1) \leq 1,$$

or

$$m \geq \frac{\log(K\lambda - 1)}{-\log|1 - k\lambda|} \approx 2\log(K\lambda), \tag{16}$$

with $c = k\lambda \approx 1/2$ for definiteness.

We conclude that m may be quite small even if $K\lambda$ is large, since the logarithm grows so slowly, and then only a small fraction of the total time (a small fraction of the time interval $[0, T]$) will be spent on stabilizing time-stepping with the small time steps k.

To measure the efficiency gain we introduce

$$\alpha = \frac{1 + m}{K + km} \in (1/K, 1/k),$$

which is the number of time steps per unit time interval with the stabilized explicit Euler method. By (16) we have

$$\alpha \approx \frac{1 + 2\log(K\lambda)}{K + \log(K\lambda)/\lambda} \approx 2\lambda \frac{\log(K\lambda)}{K\lambda} \ll 2\lambda, \tag{17}$$

for $K\lambda \gg 1$. On the other hand, the number of time steps per unit time interval for the standard explicit Euler method is

$$\alpha_0 = \lambda/2, \tag{18}$$

with the maximum stable time step $k_n = 2/\lambda$.

The cost reduction factor using the stabilized explicit Euler method would thus be

$$\frac{\alpha}{\alpha_0} \approx \frac{4\log(K\lambda)}{K\lambda},$$

which can be quite significant for large values of $K\lambda$. For typical parabolic problems, $\dot{u} + Au(t) = 0$, the eigenvalues of A are distributed on the interval $[0, \lambda_{\max}]$, and for the damping to be efficient we need a slightly modified time step sequence. This is deescribed in more detail in (Eriksson, Johnson and Logg, 2002).

We now present some examples using an adaptive $\mathrm{cG}(1)$ IVP-solver, where explicit fixed point iteration (using only a couple of iterations) on each time interval is combined with stabilizing small time steps, as described for the explicit Euler method. In all problems we note the initial transient, where the solution components change quickly, and the oscillating nature of the time step sequence outside the transient, with large time steps followed by some small stabilizing time steps.

Example. We apply the indicated method to the scalar problem (14) with $u^0 = 1$ and $\lambda = 1000$, and display the result in Figure 5. The cost reduction factor in comparison to a standard explicit method is large: $\alpha/\alpha_0 \approx 1/310$.

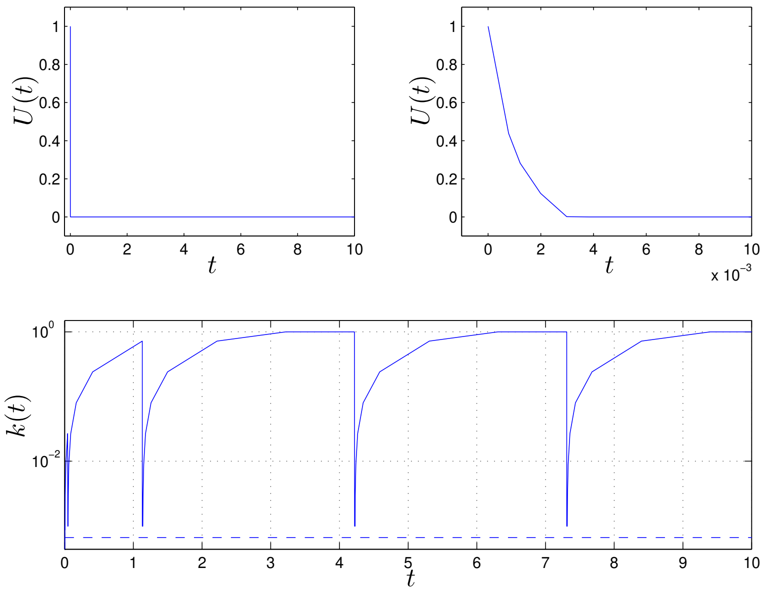

Figure 5. Solution and time step sequence for eq. (14), $\alpha/\alpha_0 \approx 1/310$.

$\square$

Example. We now consider the 2×2 diagonal system

$$\begin{aligned} \dot{u}(t) + \begin{pmatrix} 100 & 0 \\ 0 & 1000 \end{pmatrix} u(t) &= 0 \quad \text{for } 0 < t \leq T, \\ u(0) &= u^0, \end{aligned} \tag{19}$$

with $u^0 = (1, 1)$. There are now two eigenmodes with large eigenvalues that need to be stabilized. The cost reduction factor is $\alpha/\alpha_0 \approx 1/104$.

$\square$

Encyclopedia of Computational Mechanics. Edited by Erwin Stein, René de Borst and Thomas J.R. Hughes.

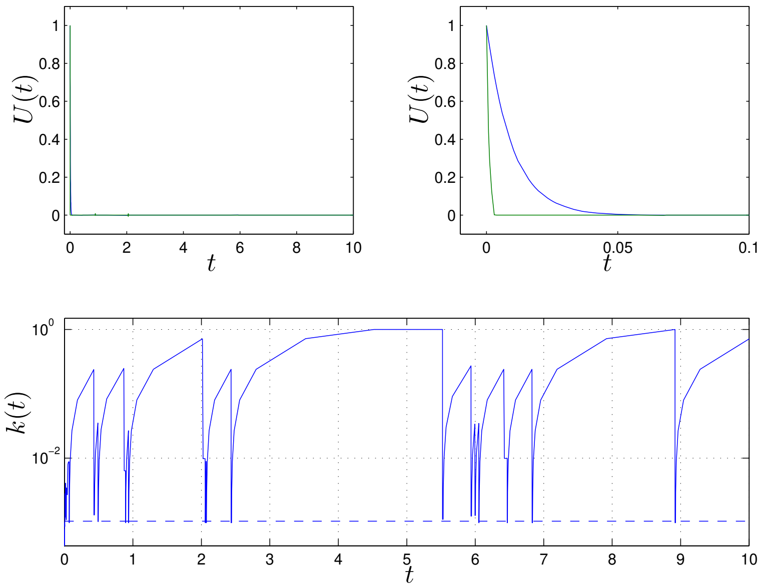

Figure 6. Solution and time step sequence for eq. (19), $\alpha/\alpha_0 \approx 1/104$.

Example. We next consider the so-called HIRES problem ("High Irradiance RESponse") from plant physiology which consists of the following eight equations:

$$\left\{\begin{array}{lcl} \dot u_1 &=& -1.71u_1 + 0.43u_2 + 8.32u_3 + 0.0007,\\ \dot u_2 &=& 1.71u_1 - 8.75u_2,\\ \dot u_3 &=& -10.03u_3 + 0.43u_4 + 0.035u_5,\\ \dot u_4 &=& 8.32u_2 + 1.71u_3 - 1.12u_4,\\ \dot u_5 &=& -1.745u_5 + 0.43u_6 + 0.43u_7,\\ \dot u_6 &=& -280.0u_6u_8 + 0.69u_4 + 1.71u_5 - 0.43u_6 + 0.69u_7,\\ \dot u_7 &=& 280.0u_6u_8 - 1.81u_7,\\ \dot u_8 &=& -280.0u_6u_8 + 1.81u_7, \end{array}\right. \tag{20}$$

together with the initial condition $u^0 = (1.0, 0, 0, 0, 0, 0, 0, 0.0057)$. We present the solution and the time step sequence in Figure 7. The cost is now $\alpha \approx 8$ and the cost reduction factor is $\alpha/\alpha_0 \approx 1/33$.

□

Example. We consider again the Akzo-Nobel problem from above, integrating over the interval $[0, 180]$. We plot the solution and the time step sequence in Figure 8. Allowing a maximum time step of $k_{\max} = 1$ (chosen arbitrarily), the cost is $\alpha \approx 2$ and the cost reduction factor is $\alpha/\alpha_0 \approx 1/9$. The actual gain in a specific situation is determined by the quotient between the large time steps and the small damping time

Encyclopedia of Computational Mechanics. Edited by Erwin Stein, René de Borst and Thomas J.R. Hughes.

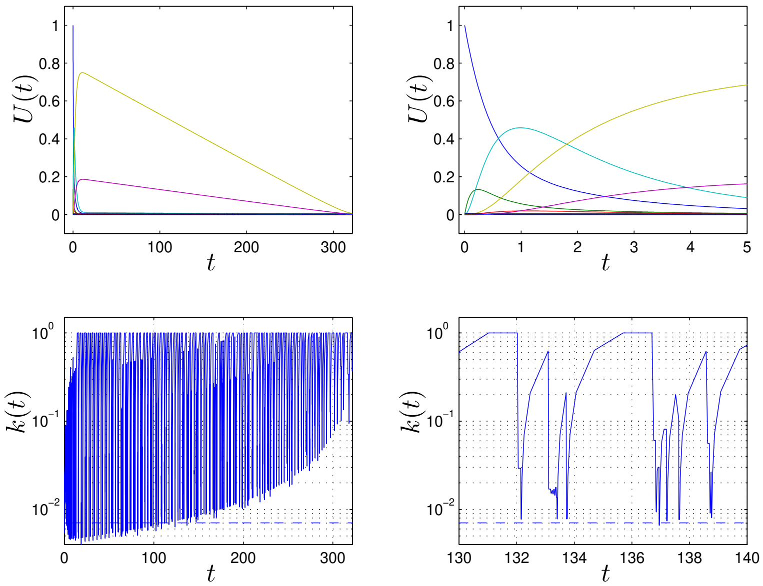

Figure 7. Solution and time step sequence for eq. (20), $\alpha/\alpha_0 \approx 1/33$.

steps, as well as the number of small damping steps that are needed. In this case the number of small damping steps is small, but the large time steps are not very large compared to the small damping steps. The gain is thus determined both by the stiff nature of the problem and the tolerance (or the size of the maximum allowed time step).

□

Example. We consider now Van der Pol's equation:

$$\ddot{u} + \mu(u^2 - 1)\dot{u} + u = 0,$$

which we write as

$$\left\{ \begin{array}{rcl} \dot{u}_1 & = & u_2, \\ \dot{u}_2 & = & -\mu(u_1^2 - 1)u_2 - u_1. \end{array} \right. \tag{21}$$

We take $\mu = 1000$ and solve on the interval $[0, 10]$ with initial condition $u^0 = (2, 0)$. The cost is now $\alpha \approx 140$ and the cost reduction factor is $\alpha/\alpha_0 \approx 1/75$.

□

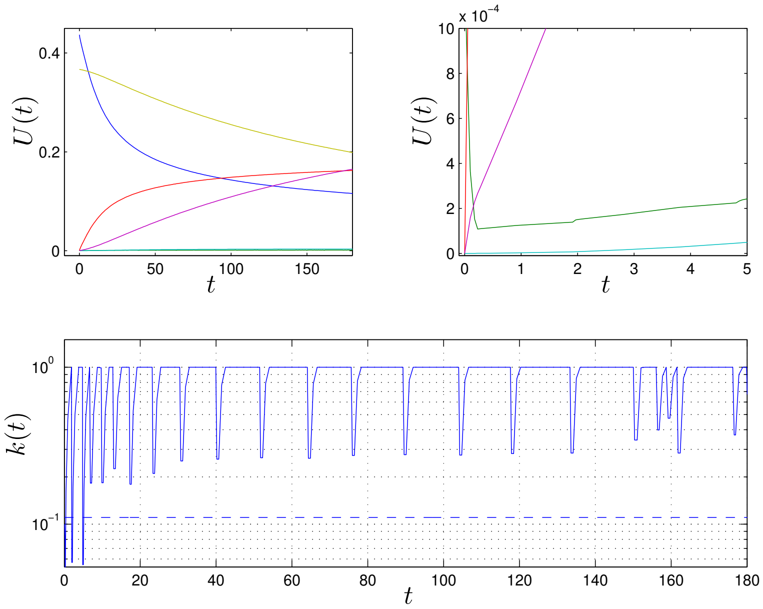

Figure 8. Solution and time step sequence for the Akzo-Nobel problem, $\alpha/\alpha_0 \approx 1/9$.

8. Strong stability estimates for an abstract parabolic model problem

We consider an abstract parabolic model problem of the form: Find $w(t) \in H$ such that

$$\begin{cases} \dot{w}(t) + Aw(t) = 0 \quad \text{for } 0 < t \leq T, \\ w(0) = w^0, \end{cases} \tag{22}$$

where H is a vector space with inner product $(\cdot, \cdot)$ and norm $\| \cdot \|$, A is a positive semi-definite symmetric linear operator defined on a subspace of H, i.e. A is a linear transformation satisfying $(Aw, v) = (w, Av)$ and $(Av, v) \geq 0$ for all v and w in the domain of definition of A, and w^0 is the initial data. In the model problem of Section 4, $H = \mathbb{R}^d$ and A is a positive semi-definite symmetric $d \times d$ matrix. In the case of the heat equation, considered in the next section, $H = L_2(\Omega)$ and $-A = \Delta$ (the Laplacian) with homogeneous Dirichlet boundary conditions.

It is important to notice that we do not require A to be positive definite, only positive semi-definite. In particular A may have very small positive or zero eigenvalues, and our analysis does not depend on exponential decay of all eigenmodes in (22).

We now state and prove the basic strong stability estimates for the parabolic model problem (22), noting that the constants on the right-hand sides of the estimates are independent of the positive semi-definite symmetric operator A. It should be noted

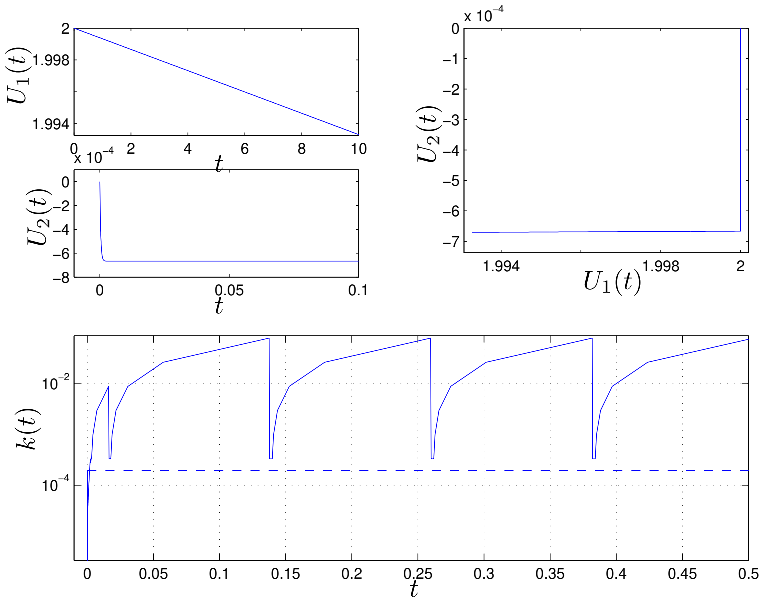

Figure 9. Solution and time step sequence for eq. (21), $\alpha/\alpha_0 \approx 1/75$.

that the dual backward problem of (22), $-\dot{\phi} + A\phi = 0$, takes the form (22) with $w(t) = \phi(T - t)$.

Lemma 8.1. *The solution w of (22) satisfies for $T > 0$,*

$$\|w(T)\|^2 + 2\int_0^T (Aw(t), w(t))\, dt = \|w^0\|^2, \tag{23}$$

$$\int_0^T t\|Aw(t)\|^2\, dt \leq \frac{1}{4}\|w^0\|^2, \tag{24}$$

$$\|Aw(T)\| \leq \frac{1}{\sqrt{2}\,T}\|w^0\|. \tag{25}$$

Proof. Taking the inner product of $\dot{w}(t) + Aw(t) = 0$ with $w(t)$, we obtain

$$\frac{1}{2}\frac{d}{dt}\|w(t)\|^2 + (Aw(t), w(t)) = 0,$$

from which (23) follows.

Next, taking the inner product of $\dot{w}(t) + Aw(t) = 0$ with $tAw(t)$ and using the fact that

$$(\dot{w}(t), tAw(t)) = \frac{1}{2}\frac{d}{dt}\left(t(Aw(t), w(t))\right) - \frac{1}{2}(Aw(t), w(t)),$$

Encyclopedia of Computational Mechanics. Edited by Erwin Stein, René de Borst and Thomas J.R. Hughes.

since A is symmetric, we find after integration that

$$\frac{1}{2}T(Aw(T),w(T))+\int_0^T t\|Aw(t)\|^2\,dt=\frac{1}{2}\int_0^T(Aw(t),w(t))\,dt,$$

from which (24) follows using (23) and the fact that $(Aw,w)\geq 0$.

Finally, taking the inner product in with $t^2A^2w(t)$, we obtain

$$\frac{1}{2}\frac{d}{dt}\left(t^2\|Aw(t)\|^2\right)+t^2(A^2w(t),Aw(t))=t\|Aw(t)\|^2,$$

from which (25) follows after integration and using (24). ∎

The estimates (23)-(25) express in somewhat different ways "parabolic smoothing"; in particular (25) expresses that that the norm of the time derivative $\dot{w}(t)$, or equivalently $Aw(t)$, decreases (increases) like $1/t$ as t increases (decreases), which means that the solution becomes smoother with increasing time. We note a close relation between the two integrals

$$I_1=\int_0^T\|\dot{w}(t)\|\,dt=\int_0^T\|Aw(t)\|\,dt,$$

and

$$I_2=\left(\int_0^T t\|\dot{w}(t)\|^2\,dt\right)^{1/2}=\left(\int_0^T t\|Aw(t)\|^2\,dt\right)^{1/2},$$

both measuring strong stability of (22), with I_1 through Cauchy's inequality being bounded by I_2 up to a logarithm:

$$\int_\epsilon^T\|Aw(t)\|\,dt\leq\left(\int_\epsilon^T\frac{1}{t}\,dt\right)^{1/2}\left(\int_\epsilon^T t\|Aw(t)\|^2\,dt\right)^{1/2}=(\log(T/\epsilon))^{1/2}\,I_2.$$

Remark 8.1. We now give an argument indicating that for the parabolic model problem (22), the stability factor $S_c(T,\psi)$ varies little with the specific choice of data ψ. We do this by noting that the quantity $S(w^0)$ defined by

$$S(w^0)=\left(\int_0^T t\|Aw(t)\|^2\,dt\right)^{1/2},$$

where $w(t)$ solves (22), varies little with the choice of initial data w^0. To see this, we let $\{\chi_i\}$ be an orthonormal basis for H consisting of eigenfunctions of A with corresponding eigenvalues $\{\lambda_j\}$, which allows us to express the solution $w(t)$ in the form $\sum_j\exp(-\lambda_j t)w_j^0\chi_j$ with $w_j^0=(w^0,\chi_j)$. We may then write

$$(S(w^0))^2=\int_0^T t\sum_j\lambda_j^2(w_j^0)^2\exp(-2\lambda_j t)\,dt=\sum_j(w_j^0)^2\int_0^T t\lambda_j^2\exp(-2\lambda_j t)\,dt.$$

Now, the factor

$$\int_0^T t\lambda_j^2 \exp(-2\lambda_j t)\, dt = \int_0^{T\lambda_j} s \exp(-2s)\, ds$$

takes on almost the same value $\int_0^\infty s\exp(-2s)\,ds \approx 1/4$ for all j as soon as $T\lambda_j \geq 1$, that is when λ_j is not very small (since T is typically large). If we randomly choose the initial data w^0, the chance of hitting an eigenfunction corresponding to a very small eigenvalue, must be very small. We conclude that $S(w^0)$ varies little with w^0. As just indicated, $S(w^0)$ is related to the integral $\int_0^T \|\dot{w}(t)\|\,dt$, which is the analog of the stability factor $S_c(T,\psi)$ for the dual problem. The bottom line is that $S_c(T,\psi)$ varies little with the choice of ψ.

Remark 8.2. The solution operator $\{E(t)\}_{t\geq 0}$ of (22), given by $E(t)w^0 = w(t)$, is said to define a uniformly bounded and analytic semigroup if there is a constant S such that the following estimates hold:

$$\begin{aligned} \|w(t)\| &\leq S\|w^0\|, \\ \|Aw(t)\| &\leq \tfrac{S}{t}\|w^0\|, \end{aligned} \tag{26}$$

for $t > 0$. We see that this definition directly couples to the stability estimates of Lemma (8.1), in which case the constant S is of unit size.

9. Adaptive space-time Galerkin methods for the heat equation

We now move on to space-time Galerkin finite element methods for the model parabolic partial differential equation in the form of the *heat equation*: Find $u : \Omega \times I \to \mathbb{R}$ such that

$$\begin{cases} \dot{u} - \Delta u = f & \text{in } \Omega \times I \\ u = 0 & \text{on } \Gamma \times I, \\ u(\cdot,0) = u^0 & \text{in } \Omega, \end{cases} \tag{27}$$

where Ω is a bounded domain in $\mathbb{R}^d$ with boundary Γ, on which we have posed homogeneous Dirichlet boundary conditions, u^0 is a given initial temperature, f a heat source and $I = (0,T]$ a given time interval.

For the discretization of the heat equation in space and time, we use the $\mathrm{cG}(p)\mathrm{dG}(q)$ method based on a tensor product space-time discretization with continuous piecewise polynomial approximation of degree $p \geq 1$ in space and discontinuous piecewise polynomial approximation of degree $q \geq 0$ in time, giving a method which is accurate of order $p+1$ in space and of order $2q+1$ in time. The discontinuous Galerkin $\mathrm{dG}(q)$ method used for the time discretization reduces to the *subdiagonal Padé method* for homogeneous constant coefficient problems and in general, together with quadrature for the evaluation of the integral in time, corresponds to an *implicit Runge-Kutta method*. For the discretization in space we use the standard conforming continuous

Encyclopedia of Computational Mechanics. Edited by Erwin Stein, René de Borst and Thomas J.R. Hughes.

Galerkin $\mathrm{cG}(p)$ method. The $\mathrm{cG}(p)\mathrm{dG}(q)$ method has maximal flexibility and allows the space and time steps to vary in both space and time. We design and analyze reliable and efficient adaptive algorithms for global error control in $L_\infty(L_2(\Omega))$ (maximum in time and L_2 in space), with possible extensions to $L_r(L_s(\Omega))$ with $1 \le r, s \le \infty$.

The $\mathrm{cG}(p)\mathrm{dG}(q)$ method is based on a partition in time $0 = t_0 < t_1 < \cdots < t_N = T$ of the interval $(0, T]$ into time intervals $I_n = (t_{n-1}, t_n]$ of length $k_n = t_n - t_{n-1}$ with associated finite element spaces $S_n \subset H_0^1(\Omega)$ consisting of piecewise polynomials of degree p on a triangulation $\mathcal{T}_n = \{K\}$ of Ω into elements K with local mesh size given by a function $h_n(x)$. We define

$$V_n = \{v : v = \sum_{j=0}^{q} t^j v_j,\, v_j \in S_n\},$$

and

$$V = \{v : v|_{I_n} \in V_n,\, n = 1, .., N\}.$$

We thus define V to be the set of functions $v : \Omega \times I \to \mathbb{R}$ such that the restriction of $v(x, t)$ to each time interval I_n is polynomial in t with coefficients in S_n. The $\mathrm{cG}(p)\mathrm{dG}(q)$ method for (27) now reads: Find $U \in V$ such that for $n = 1, 2, \ldots, N$,

$$\int_{I_n} \{(\dot{U}, v) + (\nabla U, \nabla v)\}\, dt + ([U]_{n-1}, v_{n-1}^+) = \int_{I_n} (f, v)\, dt \quad \forall v \in V_n, \qquad (28)$$

where $[w]_n = w(t_n^+) - w(t_n^-)$, $w_n^{+(-)} = \lim_{s\to 0^{+(-)}} w(t_n + s)$, $U_0^- = u^0$, and $(\cdot, \cdot)$ denotes the $L_2(\Omega)$ or $[L_2(\Omega)]^d$ inner product. Note that we allow the space discretizations to change with time from one *space-time slab* $\Omega \times I_n$ to the next.

For $q = 0$ the scheme (28) reduces to the following variant of the backward Euler scheme:

$$U_n - k_n \Delta_n U_n = P_n U_{n-1} + \int_{I_n} P_n f\, dt, \qquad (29)$$

where $U_n \equiv U|_{I_n}$, $\Delta_n : S_n \to S_n$ is the discrete Laplacian on S_n defined by $(-\Delta_n v, w) = (\nabla v, \nabla w)$ for all $w \in S_n$, and P_n is the L_2-projection onto S_n defined by $(P_n v, w) = (v, w)$ for all $w \in S_n$.

Alternatively, (29) may be written (with $f \equiv 0$) in matrix form as

$$M_n \xi_n + k_n A_n \xi_n = M_n \hat{\xi}_{n-1},$$

where M_n and A_n are mass and stiffness matrices related to a nodal basis for S_n, ξ_n is the corresponding vector of nodal values for U_n, and $\hat{\xi}_{n-1}$ is the vector of nodal values for $P_n U_{n-1}$. Evidently, we have to solve a system of equations with system matrix $M_n + k_n A_n$ to compute ξ_n.

Remark 9.1. Note that in the discretization (28), the space and time steps may vary in time and that the space discretization may be variable also in space, whereas the time steps k_n are kept constant in space. Clearly, optimal mesh design requires the time steps to be variable also in space. Now, it is easy to extend the method (28) to admit time steps which are variable in space simply by defining

$$V_n = \{v : v|_{I_n} = \sum_j c_j(t)v_j,\ v_j \in S_n\},$$

where now the coefficients $c_j(t)$ are piecewise polynomial of degree q in t without continuity requirements, on partitions of I_n which may vary with j. The discrete functions may now be discontinuous in time also inside the space-time slab $\Omega\times I_n$, and the degree q may vary over components and subintervals. The $\mathrm{cG}(p)\mathrm{dG}(q)$ method again takes the form (28), with the term $([U]_{n-1}, v_{n-1}^+)$ replaced by a sum over all jumps in time of U in $\Omega \times [t_{n-1}, t_n)$. Adaptive methods in this generality, so-called *multi-adaptive* methods, are proposed and analyzed in detail for systems of ordinary differential equations in (Logg, 2001a, Logg, 2001b).

10. A priori and a posteriori error estimates for the heat equation

In this section we state a priori and a posteriori error estimates for the $\mathrm{cG}(p)\mathrm{dG}(q)$ method (28) in the case $p = 1$ and $q = 0, 1$, and give the proofs below. A couple of technical assumptions on the space-mesh function $h_n(x)$ and time steps k_n are needed: We assume that each triangulation $\mathcal{T}_n$ with associated mesh size h_n satisfies, with $h_{n,K}$ equal to the diameter and $m_{n,K}$ the volume of $K \in \mathcal{T}_n$,

$$c_1 h_{n,K}^d \le m_{n,K} \quad \forall K \in \mathcal{T}_n, \tag{30}$$

$$c_2 h_{n,K} \le h_n(x) \le h_{n,K} \quad \forall x \in K \quad \forall K \in T_h, \tag{31}$$

$$|\nabla h_n(x)| \le \mu \quad \forall x \in \Omega, \tag{32}$$

for some positive constants c_1, c_2 and μ. The constant μ will be assumed to be small enough in the a priori error estimates (but not in the a posteriori error estimates). We further assume that there are positive constants c_3, c_4 and γ such that for all n we have

$$k_n \le c_3 k_{n+1}, \tag{33}$$

$$c_4 h_n(x) \le h_{n+1}(x) \le \frac{1}{c_4} h_n(x) \quad \forall x \in \Omega, \tag{34}$$

$$\bar{h}_n^2 \le \gamma k_n \quad \text{or} \quad S_n \subset S_{n-1}, \tag{35}$$

where $\bar{h}_n = \max_{x\in\bar{\Omega}} h_n(x)$. Furthermore, we assume for simplicity that Ω is convex, so that the following *elliptic regularity* estimate holds: $\|D^2 v\| \le \|\Delta v\|$ for all functions v vanishing on Γ. Here $(D^2 v)^2 = \sum_{ij} v_{,ij}^2$, where $v_{,ij}$ is the second partial derivative of v with respect to x_i and x_j, and $\|\cdot\|$ denotes the $L_2(\Omega)$-norm. With these assumptions, we have the following a priori error estimates:

Theorem 10.1. *If μ and γ are sufficiently small, then there is a constant C only depending on the constants c_i, $i = 1, 2, 3, 4$, such that for u the solution of (27) and U that of (28), we have for $p = 1$ and $q = 0, 1$,*

$$\|u - U\|_{I_n} \leq CL_n \max_{1 \leq m \leq n} E_{qm}(u), \quad n = 1, \ldots, N, \tag{36}$$

and for $q = 1$,

$$\|u(t_n) - U(t_n)\| \leq CL_n \max_{1 \leq m \leq n} E_{2m}(u), \quad n = 1, \ldots, N, \tag{37}$$

where $L_n = (\log(t_n/k_n) + 1)^{1/2}$, $E_{qm}(u) = \min_{j \leq q+1} k_m^j \|u_t^{(j)}\|_{I_m} + \|h_m^2 D^2 u\|_{I_m}$, $q = 0, 1, 2$ with $u_t^{(1)} = \dot{u}$, $u_t^{(2)} = \ddot{u}$, $u_t^{(3)} = \Delta \ddot{u}$ and $\|w\|_{I_m} = \max_{t \in I_m} \|w(t)\|$.

These estimates state that the discontinuous Galerkin method (28) is of order $q + 1$ globally in time and of order $2q + 1$ at the discrete time levels t_n for $q = 0, 1$, and is second order in space. In particular, the estimate (36) is *optimal* compared to interpolation with piecewise polynomials of order $q = 0, 1$ in time and piecewise linears in space, up to the logarithmic factor L_n. The third order accuracy in time at the discrete time levels for $q = 1$ reflects a *superconvergence* feature of the $\mathrm{dG}(q)$ method.

The a posteriori error estimates for (28) take the form:

Theorem 10.2. *If u is the solution of (27) and U that of (28) with $p = 1$, then we have for $q = 0$,*

$$\|u(t_n) - U(t_n)\| \leq \max_{1 \leq m \leq n} \mathcal{E}_{0m}(U), \quad n = 1, \ldots, N, \tag{38}$$

and for $q = 1$,

$$\|u(t_n) - U(t_n)\| \leq \max_{1 \leq m \leq n} \mathcal{E}_{2m}(U), \quad n = 1, \ldots, N, \tag{39}$$

where

$$\mathcal{E}_{0m}(U) = \gamma_1 \|h_m^2 R(U)\|_{I_m} + \gamma_2 \|h_m^2 [U]_{m-1}/k_m\|^* + \gamma_3 \|k_m R_{0k}(U)\|_{I_m},$$

$$\mathcal{E}_{2m}(U) = \gamma_1 \|h_m^2 R(U)\|_{I_m} + \gamma_2 \|h_m^2 [U]_{m-1}/k_m\|^* + \gamma_4 \|k_m^3 R_{1k}(U)\|_{I_m},$$

and

$$\begin{aligned} R(U) &= |f| + D_m^2 U, \\ R_{0k}(U) &= |f| + |[U]_{m-1}|/k_m, \\ R_{1k}(U) &= |f_{tt}| + |\Delta_m P_m [U]_{m-1}|/k_m^2, \end{aligned}$$

on $\Omega \times I_m$. A star indicates that the corresponding term is present only if S_{m-1} is not a subset of S_m. Further, $\gamma_i = L_N C_i$ where the C_i are constants related to approximation by piecewise constant or linear functions. Finally, $D_m^2 U$ on a space element $K \in \mathcal{T}_m$ is the modulus of the maximal jump in normal derivative of U across an edge of K divided by the diameter of K.

Encyclopedia of Computational Mechanics. Edited by Erwin Stein, René de Borst and Thomas J.R. Hughes.

Remark 10.1. The term $|f|$ in $R(U)$ may a be replaced by $|h_m^2 D^2 f|$. Similarly, the term $|f|$ in R_{0k} may be replaced with $k|\dot{f}|$ and $|\dot{f}|$ in $R_{1k}(U)$ by $k|\Delta\ddot{f}|$.

The a posteriori error estimates are sharp in the sense that the quantities on the right-hand sides can be bounded by the corresponding right-hand sides in the (optimal) a priori error estimates. Therefore, the a posteriori error estimates may be used as a basis for efficient adaptive algorithms, as we indicate below.

11. Adaptive methods/algorithms

An *adaptive method* for the heat equation addresses the following problem: For a given tolerance $\mathrm{TOL} > 0$ find a discretization in space and time $\mathcal{S}_{hk} = \{(\mathcal{T}_n, k_n)\}_{n\geq 1}$, such that

$$\begin{aligned}&(1)\ \|u(t_n) - U(t_n)\| \leq \mathrm{TOL} \ \text{ for } n = 1, 2, \ldots,\\ &(2)\ \mathcal{S}_{hk} \text{ is optimal, in the sense that the number of degrees of freedom is minimal.}\end{aligned} \tag{40}$$

We approach this problem using the a posteriori estimates (38) and (39) in an adaptive method of the form: Find $\mathcal{S}_{hk}$ such that for $n = 1, 2, \ldots,$

$$\begin{aligned}&\mathcal{E}_{on}(U) \leq \mathrm{TOL}, \quad \text{if } q = 0,\\ &\mathcal{E}_{2n}(U) \leq \mathrm{TOL}, \quad \text{if } q = 1,\\ &\text{the number of degrees of freedom of } \mathcal{S}_{hk} \text{ is minimal.}\end{aligned} \tag{41}$$

To solve this problem, we use an *adaptive algorithm* for choosing $\mathcal{S}_{hk}$ based on *equidistribution* of the form: For each $n = 1, 2, \ldots$, with $\mathcal{T}_{n0}$ a given initial space mesh and k_{n0} an initial time step, determine triangulations $\mathcal{T}_{nj}$ with N_j elements of size $h_{nj}(x)$, time steps k_{nj}, and corresponding approximate solutions U_{nj} defined on $I_{nj} = (t_{n-1}, t_{n-1} + k_{nj})$, such that for $j = 0, 1, \ldots, \hat{n} - 1$,

$$\begin{aligned}&\gamma_1 \max_{t\in I_{nj}} \|h_{n,j+1}^2 R(U_{nj})\|_{L_2(K)}\\ &\quad + \gamma_2 \|h_{n,j+1}^2 [U]_{n-1,j}/k_{nj}\|^*_{L_2(K)} = \frac{\theta\mathrm{TOL}}{2\sqrt{N_j}} \quad \forall K \in T_{nj},\\ &k_{n,j+1}\gamma_3 \|R_{0k}(U_{nj})\|_{I_{nj}} = \frac{\mathrm{TOL}}{2}, \quad \text{if } q = 0,\\ &k_{n,j+1}^3 \gamma_4 \|R_{1k}(U_{nj})\|_{I_{nj}} = \frac{\mathrm{TOL}}{2}, \quad \text{if } q = 1,\end{aligned} \tag{42}$$

that is we determine iteratively each new time step $k_n = k_{n\hat{n}}$ and triangulation $\mathcal{T}_n = \mathcal{T}_{n\hat{n}}$. The number of trials $\hat{n}$ is the smallest integer j such that (41) holds with U replaced by U_{nj}, and the parameter $\theta \sim 1$ is chosen so that $\hat{n}$ is small.

Encyclopedia of Computational Mechanics. Edited by Erwin Stein, René de Borst and Thomas J.R. Hughes.

12. Reliability and efficiency.

By the a posteriori estimates (38) and (39) it follows that the adaptive method (41) is *reliable* in the sense that if (41) holds, then the error control (40) is guaranteed. The *efficiency* of (41) follows from the fact that the right-hand sides of the a posteriori error estimates may be bounded by the corresponding right-hand sides in the (optimal) a priori error estimates.

13. Strong stability estimates for the heat equation

We now state the fundamental strong stability results for the continuous and discrete problems to be used in the proofs of the a priori and a posteriori error estimates. Analogous to in Section 8, we consider the problem $\dot{w} - \Delta w = 0$, where $w(t) = \phi(T-t)$ is the backward dual solution with time reversed.

The proof of Lemma 13.1 is similar to that of Lemma 8.1, multiplying by $w(t)$, $-t\Delta w(t)$ and $t^2\Delta^2 w(t)$. The proof of Lemma 13.2 is also analogous: For $q = 0$, we multiply (29) by W_n and $t_n A_n W_n$, noting that if $S_{n-1} \subset S_n$ (corresponding to coarsening in the time direction of the primal problem $\dot{u} - \Delta u = f$), then $P_n W_{n-1} = W_{n-1}$, $(A_n W_n, W_{n-1}) = (W_n, A_n W_{n-1})$ and $(A_n W_{n-1}, W_{n-1}) = (A_{n-1} W_{n-1}, W_{n-1})$. The proof for $q = 1$ i similar.

Lemma 13.1. *Let w be the solution of (27) with $f \equiv 0$. Then for $T > 0$,*

$$\|w(T)\|^2 + 2\int_0^T \|\nabla w(t)\|^2\,dt = \|w^0\|^2, \tag{43}$$

$$\int_0^T t\{\|\dot{w}(t)\|^2 + \|\Delta w(t)\|^2\}\,dt \le \frac{1}{2}\|w^0\|^2, \tag{44}$$

$$\|\Delta w(T)\| \le \frac{1}{\sqrt{2}T}\|w^0\|. \tag{45}$$

Lemma 13.2. *There is a constant C such that if $S_{n-1} \subset S_n$ for $n = 1, 2, \ldots, N$, and W is the solution of (28) with $f \equiv 0$, then for $T = t_N > 0$,*

$$\|W_N^-\|^2 + 2\int_0^T \|\nabla W\|^2\,dt + \sum_{n=1}^N \|[W]_{n-1}\|^2 = \|w^0\|^2, \tag{46}$$

$$\sum_{n=1}^N t_n \int_{I_n} \{\|\dot{W}\|^2 + \|\Delta_n W\|^2\}\,dt + \sum_{n=1}^N t_n \|[W]_{n-1}\|^2/k_n \le C\|w^0\|^2, \tag{47}$$

and

$$\sum_{n=1}^N \int_{I_n} \{\|\dot{W}\| + \|\Delta_n W\|\}\,dt + \sum_{n=1}^N \|[W]_{n-1}\| \le C\left(\log\frac{t_n}{k_1} + 1\right)^{1/2} \|w^0\|. \tag{48}$$

Encyclopedia of Computational Mechanics. Edited by Erwin Stein, René de Borst and Thomas J.R. Hughes.

14. A priori error estimates for the L_2- and elliptic projections

We shall use the following a priori error estimate for the L_2-projection $P_n : L_2(\Omega) \to S_n$ defined by $(w - P_n w, v) = 0$ for all $v \in S_n$. This estimate follows from the fact that P_n is very close to the nodal interpolation operator J_n into S_n, defined by $J_n w = w$ at the nodes of $\mathcal{T}_n$ if w is smooth (and $J_n w = J_n \tilde{w}$ if $w \in H^1(\Omega)$, where $\tilde{w}$ is a locally regularized approximation of w).

Lemma 14.1. *If μ in* (32) *is sufficiently small, then there is a positive constant C such that for all $w \in H_0^1(\Omega) \cap H^2(\Omega)$,*

$$|(f, w - P_n w) - (\nabla U, \nabla(w - P_n w))| \leq C\|h_n^2 R_n(U)\| \, \|D^2 w\|, \tag{49}$$

where $R_n(U) = |f| + D_n^2 U$.

We shall also need the following a priori error estimate for the *elliptic projection* $\pi_n : H_0^1(\Omega) \to S_n$ defined by

$$(\nabla(w - \pi_n w), \nabla v) = 0 \quad \forall v \in S_n. \tag{50}$$

Lemma 14.2. *If μ in* (32) *is sufficiently small, then there is a positive constant C such that for all $w \in H^2(\Omega) \cap H_0^1(\Omega)$,*

$$\|w - \pi_n w\| \leq C\|h_n^2 D^2 w\|. \tag{51}$$

Proof. We shall first prove that with $e = w - \pi_n w$, we have $\|e\| \leq C\|h_n \nabla e\|$. For this purpose, we let ϕ be the solution of the continuous dual problem $-\Delta\phi = e$ in Ω with $\phi = 0$ on Γ, and note that by integration by parts, the Galerkin orthogonality (50), a standard estimate for the interpolation error $u - J_n u$, together with elliptic regularity, we have

$$\begin{aligned} \|e\|^2 &= (e, -\Delta\phi) = (\nabla e, \nabla\phi) = (\nabla e, \nabla(\phi - J_n\phi)) \\ &\leq \|h_n \nabla e\| \, \|h_n^{-1}\nabla(\phi - J_n\phi)\| \leq C\|h_n \nabla e\| \|D^2\phi\| \leq C\|h_n\nabla e\|\|e\|, \end{aligned}$$

which proves the desired estimate. Next to prove that $\|h_n \nabla e\| \leq C\|h_n^2 D^2 w\|$, we note that since $\pi_n J_n u = J_n u$, we have

$$\begin{aligned} \|h_n \nabla e\| &\leq \|h_n \nabla(w - J_n w)\| + \|h_n \nabla \pi_n(w - J_n w)\| \\ &\leq C\|h_n \nabla(w - J_n w)\| \leq C\|h_n^2 D^2 w\|_2, \end{aligned}$$

where we used stability of the elliptic projection π_n in the form

$$\|h_n \nabla \pi_n v\| \leq C\|h_n \nabla v\| \quad \forall v \in H_0^1(\Omega),$$

which is a weighted analog of the basic property of the elliptic projection $\|\nabla \pi_n v\| \leq \|\nabla v\|$ for all $v \in H_0^1(\Omega)$. For the proof of the weighted analog we need the mesh size not to vary too quickly, expressed in the assumption that μ is small. ■

Encyclopedia of Computational Mechanics. Edited by Erwin Stein, René de Borst and Thomas J.R. Hughes.

15. Proof of the a priori error estimates

In this section we give the proof of the a priori estimates, including (36) and (37). For simplicity, we shall assume that $S_n \subset S_{n-1}$, corresponding to a situation where the solution gets smoother with increasing time. The proof is naturally divided into the following steps, indicating the overall structure of the argument:

(a) An error representation formula using duality;
(b) Strong stability of the discrete dual problem;
(c) Choice of interpolant and proof of (36);
(d) Choice of interpolant and proof of (37).

15.1. An error representation formula using duality

Given a discrete time level $t_N > 0$, we write the discrete set of equations (28) determining the discrete solution $U \in V$ up to time t_N in compact form as

$$A(U, v) = (u^0, v_0^+) + (f, v)_I \quad \forall v \in V, \tag{52}$$

where

$$A(w, v) \equiv \sum_{n=1}^{N} \{(\dot{w}, v)_n + (\nabla w, \nabla v)_n\} + (w_0^+, v_0^+) + \sum_{n=2}^{N} ([w]_{n-1}, v_{n-1}^+),$$

$(v, w)_n = \int_{I_n} (v, w)\, dt$ and $I = (0, T]$. The error $e \equiv u - U$ satisfies the *Galerkin orthogonality*

$$A(e, v) = 0 \quad \forall v \in V, \tag{53}$$

which follows from the fact that (52) is satisfied also by the exact solution u of (27). Let now the discrete dual solution $\Phi \in V$ be defined by

$$A(v, \Phi) = (v_N^-, e_N) \quad \forall v \in V, \tag{54}$$

where $e_N = u(t_N) - U(t_N)$ is the error at final time t_N, corresponding to control of the $L_2(\Omega)$-norm of e_N. We note that Φ is a discrete cG(p)dG(q)-solution of the continuous dual problem

$$\begin{aligned} -\dot{\phi} - \Delta\phi &= 0 \quad \text{in } \Omega \times [0, T), \\ \phi &= 0 \quad \text{on } \Gamma \times [0, T), \end{aligned} \tag{55}$$

with initial data $\phi(T) = e_N$. This follows from the fact that the bilinear form $A(\cdot, \cdot)$, after time integration by parts, can also be written as

$$A(w, v) = \sum_{n=1}^{N} \{(w, -\dot{v})_n + (\nabla w, \nabla v)_n\} + \sum_{n=1}^{N-1} (w_n^-, -[v]_n) + (w_N^-, v_N^-). \tag{56}$$

In view of (54) and (53) we have for any $v \in V$,

$$\begin{aligned} \|e_N\|^2 &= (u_N - v_N^-, e_N) + (v_N^- - U_N^-, e_N) \\ &= (u_N - v_N^-, e_N) + A(v - U, \Phi) = (u_N - v_N^-, e_N) + A(v - u, \Phi). \end{aligned} \tag{57}$$

Taking here $v \in V$ to be a suitable interpolant of u, we thus obtain a representation of the error e_N in terms of an interpolation error $u - v$ and the discrete solution Φ of the associated dual problem, combined through the bilinear form $A(\cdot,\cdot)$. To obtain the a priori error estimates, we estimate below the interpolation error $u - v$ in $L_\infty(L_2(\Omega))$ and the time derivative $\dot\Phi$ in $L_1(L_2(\Omega))$ using discrete strong stability.

15.2. Strong stability of the discrete dual problem

We apply Lemma 13.2 to the function $w(t) = \Phi(T - t)$, to obtain the strong stability estimate

$$\|\Phi(t)\|_I + \sum_{n=1}^{N} \int_{I_n} \{\|\dot\Phi\| + \|\Delta_n \Phi\|\}\, dt + \sum_{n=1}^{N} \|[\Phi]_n\| \le C L_N \|e_N\|, \tag{58}$$

with $L_N = (\log(T_N/k_n) + 1)^{1/2}$.

15.3. Proof of the a priori error estimate (36)

In the error representation we take the interpolant to be $v = \tilde u \equiv Q_n \pi_n u$ on I_n, where Q_n is the $L_2(I_n)$-projection onto polynomials of degree q on I_n, and π_n is the elliptic projection defined in Section 14. For $q = 0$, we thus take

$$\tilde u|_{I_n} = k_n^{-1} \int_{I_n} \pi_n u \, ds, \tag{59}$$

and for $q = 1$, we take

$$\tilde u|_{I_n} = k_n^{-1} \int_{I_n} \pi_n u \, ds + \frac{12(t - t_{n-1} - k_n/2)}{k_n^3} \int_{I_n} (s - t_{n-1} - k_n/2) \pi_n u \, ds. \tag{60}$$

With this choice of interpolant, (57) reduces to

$$\begin{aligned} \|e_N\|^2 &= (u_N - \tilde u_N^-, e_N) + \textstyle\sum_{n=1}^N (u - \pi_n u, \dot\Phi)_n \\ &+ \sum_{n=1}^{N-1} (u_n - \tilde u_n^-, [\Phi]_n) - (u_N - \tilde u_N^-, \Phi_N^-), \end{aligned} \tag{61}$$

where we have used (50), (56) and the fact that $(\pi_n u - \tilde u, v)_n = 0$ for all $v \in V_n$, and thus in particular for $v = \dot\Phi$ and $v = \Delta_n \Phi$.

Using Lemma 14.2 and the fact that Q_n is bounded in $\|\cdot\|_{I_n}$, we have

$$\begin{aligned} \|u - \tilde u\|_{I_n} &\le \|u - Q_n u\|_{I_n} + \|Q_n(u - \pi_n u)\|_{I_n} \\ &\le C\left(\min_{j \le q+1} k_n^j \|u_t^{(j)}\|_{I_n} + \|h_n^2 D^2 u\|_{I_n}\right), \end{aligned} \tag{62}$$

where the bound for $u - Q_n u$ follows from the Taylor expansion

$$\begin{aligned} u(t) &= u(t_n) + \int_{t_n}^t \dot u(s)\, ds \\ &= u(t_n) + (t - t_n)\dot u(t_n) + \int_{t_n}^t (t - s)\ddot u(s)\, ds, \end{aligned}$$

noting that Q_n is the identity on the polynomial part of $u(t)$.

From (61) we thus obtain,

$$\|e_N\|^2 \le C \max_{1\le n\le N} \left(\min_{j\le q+1} k_n^j \|u_t^{(j)}\|_{I_n} + \|h_n^2 D^2 u\|_{I_n} \right) \times \left(\|e_N\| + \sum_{n=1}^{N} \int_{I_n} \|\dot\Phi\|\, dt + \sum_{n=1}^{N-1} \|[\Phi]_n\| + \|\Phi_N^-\| \right),$$

and conclude in view of (58) that

$$\|e_N\| \le C L_N \max_{1\le n\le N} \left(\min_{j\le r+1} k_n^j \|u_t^{(j)}\|_{I_n} + \|h_n^2 D^2 u\|_{I_n} \right). \tag{63}$$

By a local analysis this estimate extends to $\|e\|_{I_N}$, completing the proof of (36).

15.4. Proof of the a priori error estimate (37)

In the error representation formula (57) we now choose $v = R_n \pi_n u$, where R_n is the (Radau) projection onto linear functions on I_n, defined by $(R_n\pi_n u)_n^- = \pi_n u_n$ and the condition that $R_n\pi_n u - \pi_n u$ has mean value zero over I_n, that is we take

$$R_n\pi_n u|_{I_n} = \pi_n u_n + (t - t_n)\frac{2}{k_n^2} \int_{I_n} \pi_n(u_n - u)\, ds. \tag{64}$$

With this choice of interpolant, (57) reduces to

$$\begin{aligned} \|e_N\|^2 \;=\;& (u_N - \pi_N u_N, e_N) \\ &+ \textstyle\sum_{n=1}^{N} (u - \pi_n u, \dot\Phi)_n - \sum_{n=1}^{N} (\nabla(\pi_n u - R_n\pi_n u), \nabla\Phi)_n \\ &+ \textstyle\sum_{n=1}^{N-1} (u_n - \pi_n u_n, [\Phi]_n) - (u_N - \pi_N u_N, \Phi_N^-), \end{aligned} \tag{65}$$

where in the first sum we have used the fact that $\pi_n u - R_n\pi_n u$ is orthogonal to $\dot\Phi$ (which is constant in t on I_n), and in the second sum we have used (50). For the latter term, we have

$$\begin{aligned} (\nabla(\pi_n u - R_n\pi_n u), \nabla\Phi)_n \;=\;& (\nabla(\pi_n u - R_n\pi_n u), \nabla\Phi_n^-)_n \\ &+ (\nabla(\pi_n u - R_n\pi_n u), (t - t_n)\nabla\dot\Phi)_n, \end{aligned}$$

so that by our choice of $R_n\pi_n u$,

$$\begin{aligned} |(\nabla(\pi_n u - R_n\pi_n u), \nabla\Phi)_n| \;=\;& \left|(\nabla(\pi_n u - R_n\pi_n u), (t - t_n)\nabla\dot\Phi)_n\right| \\ \le\;& k_n \|\Delta_n(\pi_n u - R_n\pi_n u)\|_{I_n} \textstyle\int_{I_n} \|\dot\Phi\|\, dt. \end{aligned} \tag{66}$$

Using Taylor expansions we easily find that

$$\|\Delta_n(\pi_n u - R_n\pi_n u)\|_{I_n} \le C k_n^2 \|\Delta_n \pi_n u_t^{(2)}\|_{I_n}. \tag{67}$$

Finally, we note that for any $w \in H^2(\Omega) \cap H_0^1(\Omega)$ we have

$$(-\Delta_n \pi_n w, v) = (\nabla\pi_n w, \nabla v) = (\nabla w, \nabla v) = (-\Delta w, v) \quad \forall v \in S_n,$$

from which we deduce by taking $v = -\Delta_n \pi_n w$ that

$$\|\Delta_n \pi_n w\| \le \|\Delta w\| \quad \forall w \in H^2(\Omega) \cap H_0^1(\Omega). \tag{68}$$

It now follows from (65) through (68), together with Lemma 14.2 and strong stability for Φ, that

$$\begin{aligned}
\|e_N\|^2 &\le C \max_{1\le n\le N} (\min_{j\le q+2} k_n^j \|u_t^{(j)}\|_{I_n} + \|h_n^2 D^2 u\|_{I_n}) \\
&\times \quad (\|e_N\| + \sum_{n=1}^{N} \int_{I_n} \|\dot{\Phi}\|\, dt + \sum_{n=1}^{N-1} \|[\Phi]_n\| + \|\Phi_N^-\|) \\
&\le \|e_N\| C L_N \max_{1\le n\le N} (\min_{j\le q+2} k_n^j \|u_t^{(j)}\|_{I_n} + \|h_n^2 D^2 u\|_{I_n}),
\end{aligned}$$

where we have used the notation $u_t^{(3)} = \Delta \ddot{u}$. This completes the proof of (37) and Theorem 10.1.

16. Proof of the a posteriori error estimates

The proof of the a posteriori error estimates is similar to that of the a priori error estimates just presented. The difference is that now the error representation involves the exact solution ϕ of the continuous dual problem (55), together with the residual of the discrete solution U. By the definition of the dual problem and the bilinear form A, we have

$$\|e_N\|^2 = A(e, \phi) = A(u, \phi) - A(U, \phi).$$

Using now that

$$A(u, \phi) = (u^0, \phi_0^+) + (f, \phi)_I,$$

together with the Galerkin orthogonality, we obtain the following error representation in terms of the discrete solution U, the dual solution ϕ and data u^0 and f:

$$\begin{aligned}
\|e_N\|^2 &= A(U, v - \phi) + (u^0, \phi_0^+ - v_0^+) + (f, \phi - v)_I \\
&= \sum_{n=1}^{N} \{(\dot{U}, v - \phi)_n + (\nabla U, \nabla(v - \phi))_n\} + \sum_{n=1}^{N} ([U]_{n-1}, v_{n-1}^+ - \phi_{n-1}^+) \\
&\quad + (f, \phi - v)_I \\
&= I + II + III,
\end{aligned} \tag{69}$$

with obvious notation. This holds for all $v \in V$.

To prove (38), we now choose $v|_{I_n} = \tilde{\phi} \equiv Q_n P_n \phi$ in (69), and note that

$$(\dot{U}, \tilde{\phi} - \phi)_n = 0.$$

Since also

$$(\nabla U, \nabla(\tilde{\phi} - P_n \phi))_n = (-\Delta_n U, \tilde{\phi} - P_n \phi)_n = (-\Delta_n U, (Q_n - I) P_n \phi) = 0,$$

Encyclopedia of Computational Mechanics. Edited by Erwin Stein, René de Borst and Thomas J.R. Hughes.

it follows that

$$I = \sum_{n=1}^{N} (\nabla U, \nabla (P_n - I)\phi)_n = \sum_{n=1}^{N} (\nabla U_n, \nabla (P_n - I) \int_{I_n} \phi \, dt).$$

Using that

$$\Delta \int_{I_n} \phi \, dt = \int_{I_n} \Delta \phi \, dt = \int_{I_n} \dot{\phi} \, dt = \phi(t_n) - \phi(t_{n-1}),$$

together with Lemma 14.1 and elliptic regularity, we get

$$\begin{aligned} |I| &\leq C \sum_{n=1}^{N} \|h_n^2 D_n^2 U\|_{I_n} \|\Delta \int_{I_n} \phi \, dt\| \\ &\leq C \max_{1 \leq n \leq N} \|h_n^2 D_n^2 (U)\|_{I_n} \left(\int_0^{t_{N-1}} \|\dot{\phi}\| \, dt + 2\|\phi\|_{I_N} \right). \end{aligned} \tag{70}$$

To estimate II, we note that by Lemma 14.1 we have

$$\left| ([U]_{n-1}, (P_n - I)\phi_{n-1}^+) \right| \leq C \|h_n^2 [U]_{n-1}\|^* \|\Delta \phi_{n-1}^+\|, \tag{71}$$

noting that the left-hand side is zero if $S_{n-1} \subset S_n$. By obvious stability and approximation properties of the $L_2(I_n)$-projections onto the set of constant functions on I_n, we also have

$$\|\tilde{\phi} - P_n \phi\|_{I_n} \leq \|P_n \phi\|_{I_n} \leq \|\phi\|_{I_n}, \tag{72}$$

and

$$\|\tilde{\phi} - P_n \phi\|_{I_n} \leq \int_{I_n} \|P_n \dot{\phi}\| \, dt \leq \int_{I_n} \|\dot{\phi}\| \, dt. \tag{73}$$

We thus conclude that

$$\begin{aligned} |II| &\leq C \max\nolimits_{1 \leq n \leq N} \|h_n^2 [U]_{n-1}/k_n\|^* \textstyle\sum_{n=1}^{N} k_n \|\Delta \phi_{n-1}^+\| \\ &+ \max_{1 \leq n \leq N} \|[U]_{n-1}\| (\int_o^{t_{N-1}} \|\dot{\phi}\| \, dt + \|\phi\|_{I_N}). \end{aligned} \tag{74}$$

The data term III is estimated similarly. We finally use strong stability of ϕ in the form

$$\begin{aligned} \textstyle\sum_{n=1}^{N-1} k_n \|w_{n-1}^+\| &\leq \textstyle\int_0^{t_{N-1}} \|w\| \, dt \\ &\leq \left(\textstyle\int_0^{t_{N-1}} (t_N - t)^{-1} \, dt \right)^{1/2} \left(\textstyle\int_0^{t_N} (t_N - t) \|w\|^2 \, dt \right)^{1/2} \\ &\leq \tfrac{1}{2} \left(\log \tfrac{t_N}{k_N} \right)^{1/2} \|e_N\|, \end{aligned} \tag{75}$$

for $w = \dot{\phi}$ and $w = \Delta \phi$, together with the estimate $k_N \|\Delta \phi_{N-1}^+\| \leq \exp(-1) \|e_N\|$.

Combining the estimates completes the proof of the posteriori error estimate (38). The proof of the a posteriori error estimate (39) is similar.

17. Extension to systems of convection-diffusion-reaction problems

We may in a natural way directly extend the scope of methods and analysis to systems of convection-diffusion-reaction equations of the form

$$\begin{cases} \dot{u} - \nabla \cdot (a \nabla u) + (b \cdot \nabla) u - f(u) = 0 & \text{in } \Omega \times (0,T], \\ \partial_n u = 0 & \text{on } \Gamma \times (0,T], \\ u(\cdot,) = u^0 & \text{in } \Omega, \end{cases} \tag{76}$$

where $u = (u_1, \ldots, u_d)$ is a vector of concentrations, $a = a(x,t)$ is a diagonal matrix of diffusion coefficients, $b = b(x,t)$ is a given convection velocity, $f(u)$ models reactions, and $\partial_n = n \cdot \nabla$ with n the exterior normal of Γ. Depending on the size of the coefficients a and b and the reaction term, this problem may exhibit more or less parabolic behavior, determined by the size of the strong stability factor coupled to the associated linearized dual problem (here linearized at the exact solution u):

$$\begin{cases} -\dot{\phi} - \nabla \cdot (a \nabla \phi) - \nabla \cdot (\phi b) - (f'(u))^\top \phi = 0 & \text{in } \Omega \times [0,T), \\ \partial_n \phi = 0 & \text{on } \Gamma \times [0,T), \\ \phi(\cdot,T) = \psi & \text{in } \Omega, \end{cases} \tag{77}$$

where $(\nabla \cdot (\phi b))_i \equiv \nabla \cdot (\phi b_i)$ for $i = 1, \ldots, d$.

18. Examples of reaction-diffusion problems

We now present solutions to a selection of reaction-diffusion problems, including solutions of the dual backward problem and computation of stability factors.

18.1. Moving heat source

In Figures 10 and 11 we display mesh and solution at two different times for the adaptive $\mathrm{cG}(1)\mathrm{dG}(0)$ method applied to the heat equation with a moving heat source producing a moving hot spot. We notice that the space mesh adapts to the solution.

18.2. Adaptive time steps for the heat equation

We consider again the heat equation, $\dot{u} - \Delta u = f$ with homogeneous Dirichlet boundary conditions on the unit square $(0,1) \times (0,1)$ over the time interval $[0,100]$. The source $f(x,t) = 2\pi^2 \sin(\pi x_1)\sin(\pi x_2)\left[\sin(2\pi^2 t) + \cos(2\pi^2 t)\right]$ is periodic in time, with corresponding exact solution

$$u(x,t) = \sin(\pi x_1)\sin(\pi x_2)\sin(2\pi^2 t).$$

In Figure 12 we show a computed solution using the $\mathrm{cG}(1)\mathrm{dG}(0)$ method, and we also plot the time evolution of the L_2-error in space together with the sequence of adaptive time steps. We notice that the error does not grow with time, reflecting the

Encyclopedia of Computational Mechanics. Edited by Erwin Stein, René de Borst and Thomas J.R. Hughes.

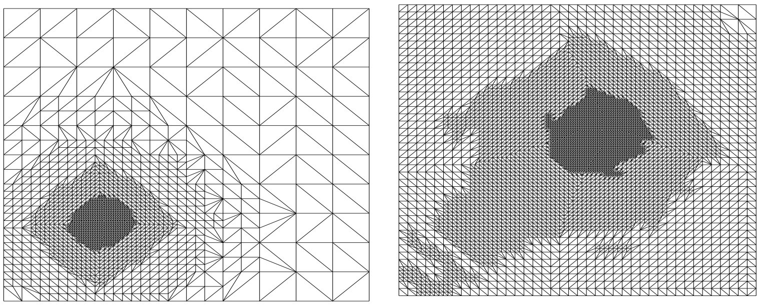

Figure 10. Meshes for moving source problem.

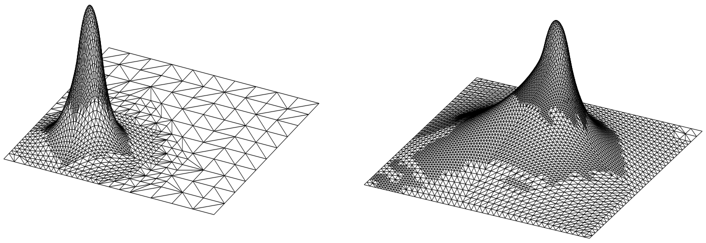

Figure 11. Solution for moving source problem.

parabolic nature of the problem. We also note the periodic time variation of the time steps, reflecting the periodicity of the solution, with larger time steps when the solution amplitude is small.

18.3. Logistics reaction-diffusion

We now consider the heat equation with a non-linear reaction-term, referred to as the *logistics problem*:

$$\begin{cases} \dot{u} - \epsilon\Delta u = u(1-u) & \text{in } \Omega \times (0,T], \\ \partial_n u = 0 & \text{on } \Gamma \times (0,T], \\ u(\cdot,0) = u^0 & \text{in } \Omega, \end{cases} \tag{78}$$

with $\Omega = (0,1)\times(0,1)$, $T = 10$, $\epsilon = 0.01$, and

$$u^0(x) = \begin{cases} 0, & 0 < x_1 < 0.5, \\ 1, & 0.5 \le x_1 < 1. \end{cases} \tag{79}$$

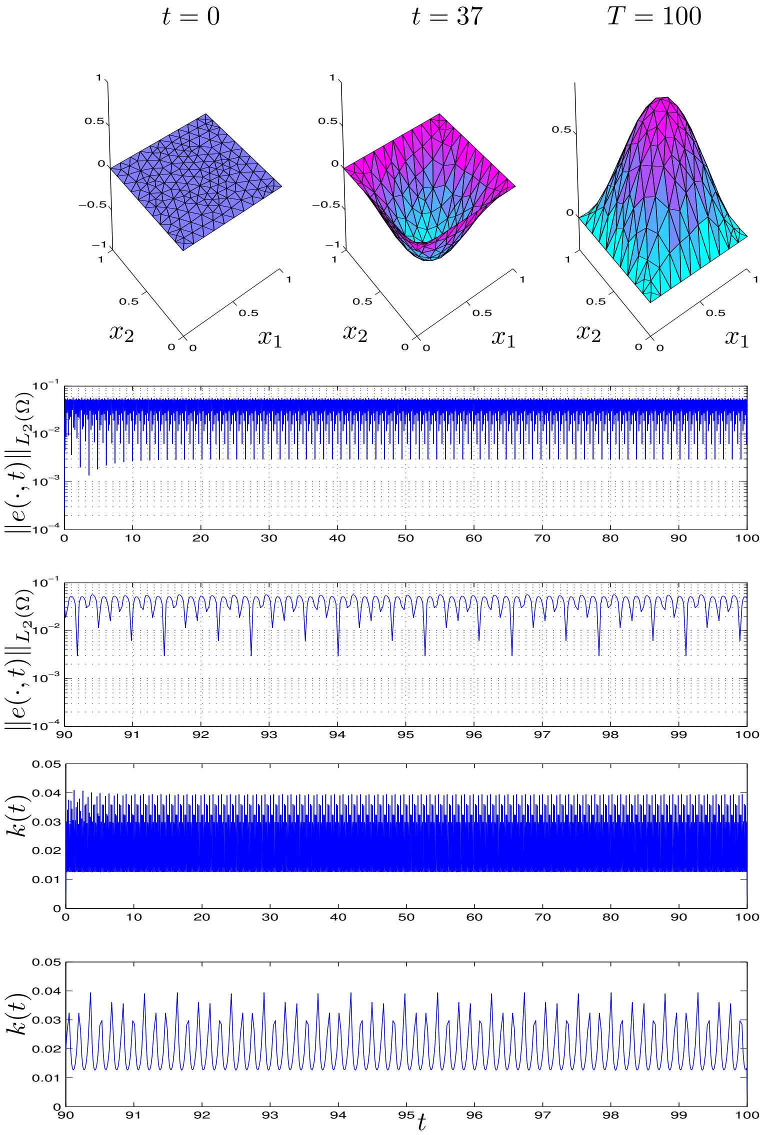

Figure 12. Heat equation: solution, error and adaptive time steps.

Encyclopedia of Computational Mechanics. Edited by Erwin Stein, René de Borst and Thomas J.R. Hughes.

Through the combined action of the diffusion and reaction the solution $u(x,t)$ tends to 1 for all x with increasing time, see Figure 13.We focus interest at final time T to a circle of radius $r = 0.25$ centered at $x = (0.5, 0.5)$. The corresponding dual problem linearized at the exact solution u is given by

$$\begin{cases} -\dot{\phi} - \epsilon\Delta\phi = (1-2u)\phi & \text{in } \Omega \times [0,T), \\ \partial_n \phi = 0 & \text{on } \Gamma \times [0,T), \\ \phi(\cdot,T) = \psi & \text{in } \Omega, \end{cases} \tag{80}$$

where we choose $\psi = 1/\pi r^2$ within the circle and zero outside. In Figure 14 we plot the dual solution $\phi(\cdot,t)$ and also the stability factor $S_c(T,\psi)$ as function of T. As in the Akzo-Nobel problem discussed above, we note that $S_c(T,\psi)$ reaches a maximum for $T \sim 1$, and then decays somewhat for larger T. The decay with larger T can be understood from the sign $(1-2u)$ of the coefficient of the ϕ-term in the dual problem, which is positive when $u(x,t) < 0.5$ and thus is positive for t small and $x_1 < 0.5$ and negative for larger t. The growth phase in $\psi(\cdot,t)$ thus occurs after a longer phase of decay if T is large, and thus $S_c(T,\psi)$ may effectively be smaller for larger T, although the interval of integration is longer for large T.

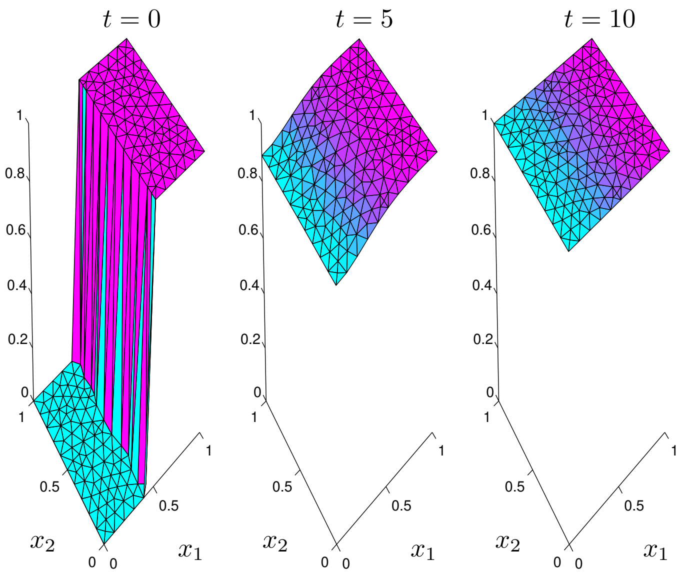

Figure 13. The logistics problem: solution at three different times.

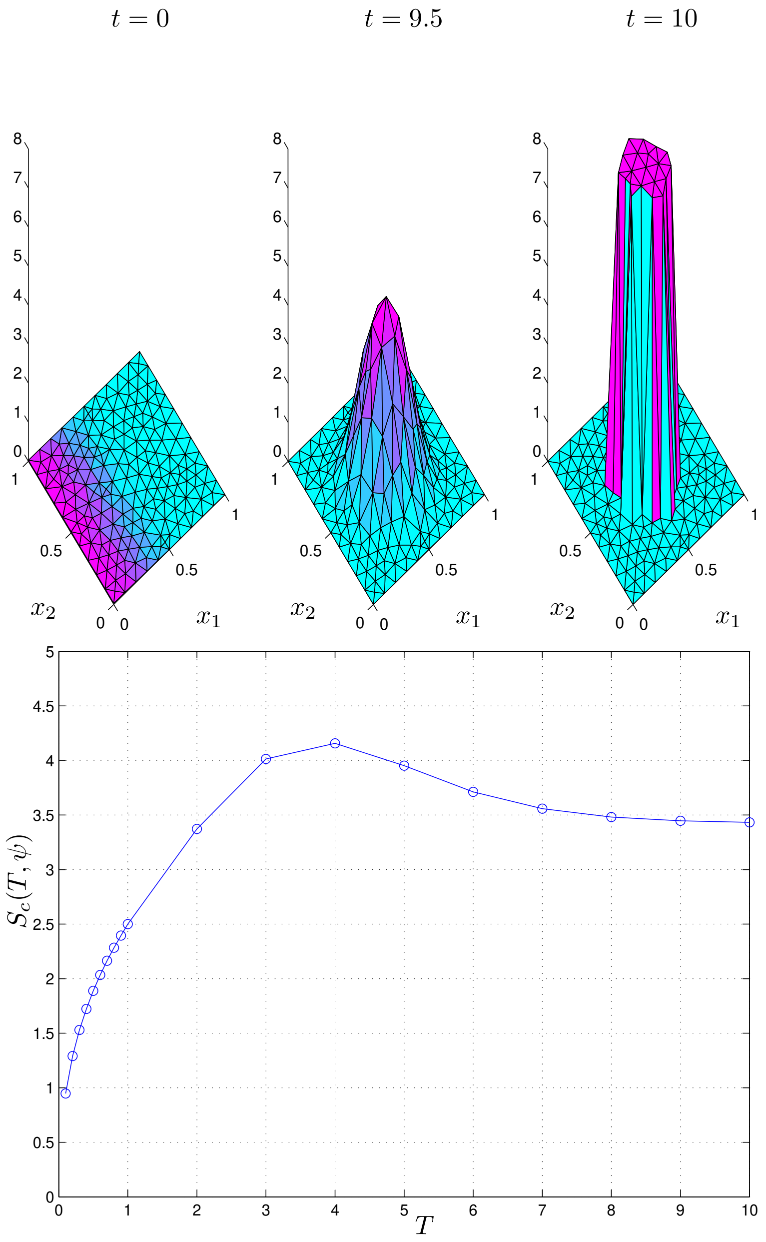

Figure 14. The logistics problem: dual solution and stability factor $S_c(T, \psi)$.

Encyclopedia of Computational Mechanics. Edited by Erwin Stein, René de Borst and Thomas J.R. Hughes.

18.4. Moving reaction front

Next, we consider a system of reaction-diffusion equations, modeling an auto-catalytic reaction where A reacts to form B with B as a catalyst:

$$A + 2B \to B + 2B. \tag{81}$$

With u_1 the concentration of A and u_2 that of B, the system takes the form

$$\begin{cases} \dot{u}_1 - \epsilon\Delta u_1 = -u_1 u_2^2, \\ \dot{u}_2 - \epsilon\Delta u_2 = u_1 u_2^2, \end{cases} \tag{82}$$

on $\Omega \times (0, 100]$ with $\Omega = (0,1) \times (0, 0.25)$, $\epsilon = 0.0001$ and homogeneous Neumann boundary conditions. As initial conditions, we take

$$u_1(x, 0) = \begin{cases} 0, & 0 < x_1 < 0.25, \\ 1, & 0.25 \le x_1 < 1, \end{cases} \tag{83}$$

and $u_2(\cdot, 0) = 1 - u_1(\cdot, 0)$. The solution $u(x,t)$ corresponds to a reaction front, starting at $x_1 = 0.25$ and propagating to the right in the domain until all of A is consumed and the concentration of B is $u_2 = 1$ in all of Ω, see Figure 15.

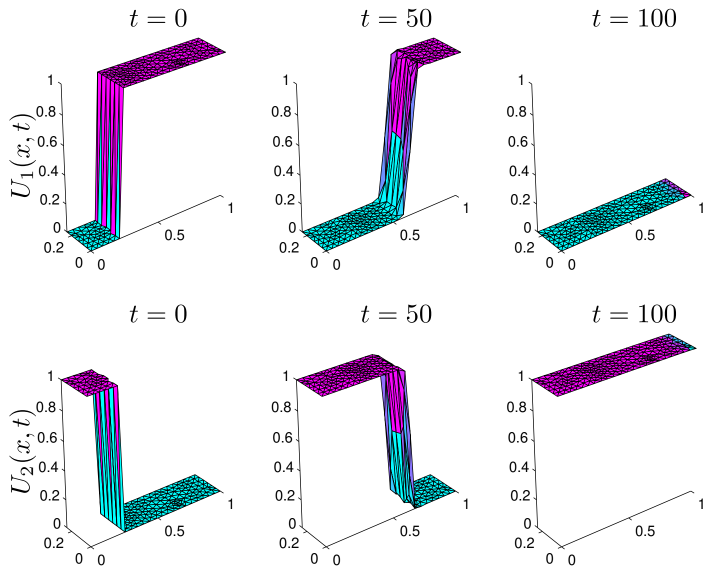

Figure 15. Reaction front problem: solution for the two components at three different times.

The dual problem, linearized at $u = (u_1, u_2)$, is given by

$$\begin{cases} -\dot{\phi}_1 - \epsilon\Delta\phi_1 &= -u_2^2\phi_1 + u_2^2\phi_2, \\ -\dot{\phi}_2 - \epsilon\Delta\phi_2 &= -2u_1u_2\phi_1 + 2u_1u_2\phi_2. \end{cases} \tag{84}$$

As in the previous example, we take the final time data ψ_1 for the first component of the dual to be an approximation of a Dirac delta function centered in the middle of the domain, and $\psi_2 \equiv 0$.

We note that the stability factor peaks at the time of active reaction, and that before and after the reaction front has swept the region of observation the stability factor $S_c(T, \psi)$ is significantly smaller.

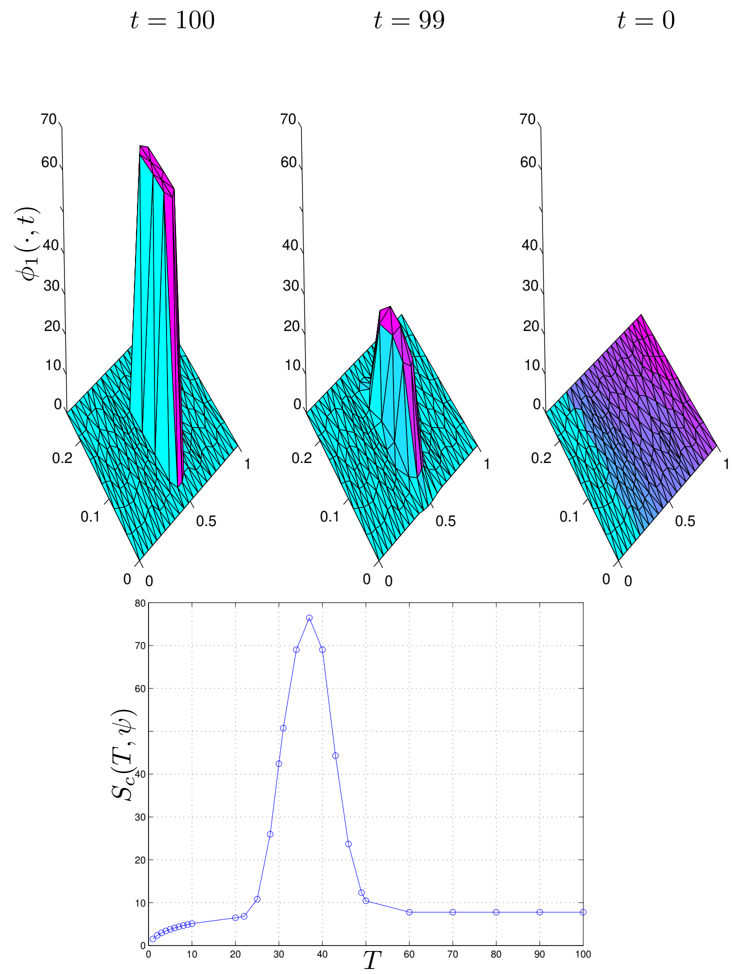

Figure 16. Reaction front problem: dual solution and stability factor $S_c(T)$ as function of T.

19. Comparison with the standard approach to time step control for ode's

We now compare our methods for adaptive error control with the methods for automatic time step control developed within the ode-community as presented e.g. in Hairer and Wanner, 1996 and Deufelhard and Bornemann, 2002, which we refer to as the *standard approach*. The corner-stone of the standard approach is a concept of of *local error*, which is (an estimate of) the error contribution from each time step, and the time-step is then controled so as to keep the local error within a certain *local error tolerance* tol. The relation between the local tolerance tol and the global error after many time steps is then left to the (intelligent) user to figure out.

For the comparison we focus on the backward Euler method and first recall our basic a posteriori error estimate (7) for the global error $e(T)$, that is,

$$\|e(T)\| \leq S_c(T) \max_{1\leq n\leq N} \|U(t_n) - U(t_{n-1})\|, \tag{85}$$

where $S_c(T)$ is the strong stability factor. We thus obtain control of the global error $\|e(T)\| \leq \mathrm{TOL}$ if we control the local time step so that

$$\|U(t_n) - U(t_{n-1})\| \sim TOL$$

assuming $S_c(T) \sim 1$.

Next, the *local error* $\hat{e}_n$ at time step n in the standard approach is defined as the difference between the first order backward Euler solution $U(t_n)$ and an approximate solution obtained after one time step starting from $U(t_{n-1})$ and using a higher order method. With a trapezoidal method as the higher order method, we may obtain

$$\hat{e}_n = \frac{k_n}{2}\|f(U(t_n)) - f(U(t_{n-1}))\|.$$

The standard time step control would then consist in choosing the time step so that

$$\hat{e}_n \approx \mathrm{tol}\,,$$

where tol thus would the local error tolerance. To estimate the global error $e(T)$ following the standard approach it would appear that one would have to sum all the local errors to get

$$\|e(T)\| \leq \sum_{n=1}^{N} \hat{e}_n. \tag{86}$$

It would then seem that we should choose tol so that $\mathrm{TOL} = N\,\mathrm{tol}$, where N is the number of time steps. But as we said, the relation between the local and global error tolerance is left to the user in the standard approach.

To bring out the essential difference between our estimate of the global error (85) with a standard estimate (86), we compare the following a priori analog of the standard estimate (86), obtained using that $\frac{d}{dt}f(u) = \ddot{u}$ (and discarding a factor $\frac{1}{2}$):

$$\|e(T)\| \leq E_1 \equiv \int_0^t k(t)\|\ddot{u}(t)\|\,dt, \tag{87}$$

Encyclopedia of Computational Mechanics. Edited by Erwin Stein, René de Borst and Thomas J.R. Hughes.

and the analog of our estimate (85):

$$\|e(T)\| \leq E_2 \equiv \max_{0\leq t\leq T} \|k(t)\dot{u}(t)\|, \tag{88}$$

where $k(t) = k_n$ on I_n, and we put $S_c(T) = 1$. Now it is clear that by integration $E_2 \leq E_1$ (assuming for simplicity that $\dot{u}(0) = 0$), while we may have $E_2 << E_1$ in particular if T is large. This shows that our estimate may be very much more economical to use than the standard estimate. We note that the standard estimate will look the same independently of the function $f(u)$, and thus does not take any particular advantage of the special properties of stiff or parabolic problems, as our estimate does. This explains the possibly very suboptimal nature of the standard estimate. Notice again that our estimate does not depend on heavy exponential decay of all components: in the model problem $\dot{u} + Au = 0$, we assume that A is positive semi-definite, not strictly positive definite.

In the standard approach it is further common to choose the values of the higher order method rather than the lower order method, because of obvious reasons. With this perspective, the estimate of the local error would then rather come out as the difference after one time step of a higher order method, which is used in the computation and a lower order method, which is used only for the local error estimation. Doing so we would seem to lack an error estimate for the higher order method, since after all we estimated the local error of the lower order method.

Nevertheless, using the values of the higher order method it seems (surprisingly so) that in many problems including stiff problems, the global error tolerance TOL would be of the same size as the local error tolerance tol, as if there were no error accumulation. By using the results of the higher order method it thus seems to be possible to get global error control on the level of the local error tolerance. But no explanation of this "miracle" would be given in the standard approach.

Well, let's see if can explain this mystery using our sharp error analysis. Let us then take as the higher order method the first order backward Euler method, and as the lower order method the zero order (trivial) method: $\hat{U}(t_n) = U(t_{n-1})$. The difference between the higher order and the lower order method at time step n would then simply be $\|U(t_n) - U(t_{n-1})\|$, and the time step would be chosen so that the local error $\|U(t_n) - U(t_{n-1})\| \leq \mathrm{tol}$. But this is the same as our time step control guaranteeing the global error control $\|e(T)\| \leq \mathrm{tol}$ (assuming $S_c(T) = 1$). We have thus proved that for parabolic or stiff problems by using the values of the higher order method, the global error would come out on the level of the local error tolerance, independent of the length of the simulation, and the "miracle" would thus be possible to understand.

20. Software

There is a variety of software available for stiff ode's such as LSODE and RADAU5 using the standard approach to error control. The methods presented in this article have been implemented in a general multi-adaptive framework as a part

Encyclopedia of Computational Mechanics. Edited by Erwin Stein, René de Borst and Thomas J.R. Hughes.

of the DOLFIN project developed by J. Hoffman and A. Logg and presented at http://www.phi.chalmers.se/dolfin/.

REFERENCES

Delfour M., Hager W., Trochu F. Discontinuous Galerkin Metheods for Ordinary Differential Equations. *Math Comp.* 1981; **36**: 453–473.

Deufelhard P. and Bornemann F. Wanner G. Scientific Computing with Ordinary Differential Equations. *Springer Texts in Applied Mathematics 42* 2002;

Douglas J. Jr, Dupont T. and Wheeler M. Galerkin Methods for Parabolic Equations. *SIAM J. Numer. Anal.* 1970; **7**: 575–626.

Eriksson K., Estep D. and Johnson C. Computational Differential Equations. *Springer* 1996.

Eriksson K. and Johnson C. Adaptive Finite Element Methods for Parabolic Problems I: A Linear Model Problem. *SIAM J. Numer. Anal.* 1991; **28**: 43–77.

Eriksson K. and Johnson C. 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.* 1995; **32**: 706–740.

Eriksson K. and Johnson C. Adaptive Finite Element Methods for Parabolic Problems IV: Nonlinear Problems. *SIAM J. Numer. Anal.* 1995; **32**: 1729–1749.

Eriksson K. and Johnson C. Adaptive Finite Element Methods for Parabolic Problems V: Long-time integration. *SIAM J. Numer. Anal.* 1995; **32**: 1750–1763.

Eriksson K., Johnson C. and Larsson S. Adaptive Finite Element Methods for Parabolic Problems VI: Analytic Semigroups. *SIAM J. Numer. Anal.* 1998; **35**: 1315–1325.

Eriksson K., Estep D., Hansbo P. and Johnson C. Introduction to Adaptive Methods for Differential Equations. *Acta Numerica, Cambridge University Press* 1995; 105–158.

Eriksson K., Johnson C. and Thome V. Time Discretization of Parabolic Problems by the Discontinuous Galerkin Method. *RAIRO MAN* 1985; **19**: 611–643.

Eriksson K, Johnson C. and Logg A. Explicit Time-Stepping for Stiff ODEs. *SIAM J. Sci. Comput.* 2003; *25*: 1142–1157.

Johnson C. Error Estimates and Adaptive Time-Step Control for a Class of One-Step Methods for Stiff Ordinary Differential Equations. *SIAM J. Numer. Anal.* 1988; **25**: 908–926.

Johnson C. and Logg A. Applied Mathematics: Body and Soul V: Dynamical Systems. To appear at *Springer* 2005;

Encyclopedia of Computational Mechanics. Edited by Erwin Stein, René de Borst and Thomas J.R. Hughes.

Hairer E., Wanner G. Solving Ordinary Differential Equations II: Stiff and Differential-Algebraic Problems. *Springer Series in Computational Mathematics 14* 1996;

Logg A. Multi-Adaptive Galerkin Methods for ODEs I. *SIAM J. Sci. Comput.* 2003; *24*: 119–1141

Logg A. Multi-Adaptive Galerkin Methods for ODEs II: Implementation & Applications. *SIAM J. Sci. Comput.* 2003; *25*: 1119–1144.

Thomèe V. Galerkin Finite Element Methods for Parabolic Problemss. *Springer Series in Computational Mathematics 25* 1997;

Encyclopedia of Computational Mechanics. Edited by Erwin Stein, René de Borst and Thomas J.R. Hughes.