D. Estep
S. Larsson

The discontinuous Galerkin method for semilinear parabolic problems

Modélisation mathématique et analyse numérique, tome 27, no 1 (1993), p. 35-54

<http://www.numdam.org/item?id=M2AN_1993__27_1_35_0>
THE DISCONTINUOUS GALERKIN METHOD FOR SEMILINEAR PARABOLIC PROBLEMS (*)

by D. ESTEP (1) and S. LARSSON (2)

Communicated by R. TEMAM

Abstract. — We prove a priori error estimates for a space-time finite element method for semilinear parabolic problems. The finite element method has basis functions that are continuous in space and discontinuous in time, and variable spatial meshes and time steps are allowed. The effect of numerical quadrature is emphasized.

Résumé. — Nous montrons des estimations d'erreur a priori pour une méthode des éléments finis en espace et en temps pour des problèmes paraboliques semi-linéaires. La méthode des éléments finis considérée a des fonctions de base continues en espace et discontinues en temps, et admet des maillages spatiaux et des pas de temps variables. L'effet de quadrature numérique est accentué.

1. INTRODUCTION

In this paper we consider the numerical solution of the semilinear parabolic equation

\[ u_t - \Delta u = f(x, t, u), \quad \text{in } \Omega \times (0, t^*), \]

\[ u = 0, \quad \text{on } \partial \Omega \times (0, t^*), \]

\[ u(., 0) = u_0, \quad \text{in } \Omega, \quad (1.1) \]

by using a finite element method with basis functions that are continuous in space and discontinuous in time, which we refer to as the discontinuous Galerkin method. Here \( \Omega \) is a bounded convex polygonal domain in

(*) Received October 1991.
(1) School of Mathematics, Georgia Institute of Technology, Atlanta, GA 30332, USA.
Supported by DARPA and the Swedish National Board for Technical Development (STUF).
(2) Department of Mathematics, Chalmers University of Technology and the University of Göteborg, S-412 96 Göteborg, Sweden.
The discontinuous Galerkin method for linear parabolic equations, i.e., the case of (1.1) when \( f = f(x, t) \), has been analyzed by Eriksson, Johnson and Thomee [3] and Eriksson and Johnson [2]. There are two parts to the analysis in [2]: an a posteriori error analysis, which is used to devise a global error control for an adaptive finite element method, and an a priori error analysis, which guarantees convergence. The a priori error bound is of a specific form related to the a posteriori bound and is proved under conditions which allow variable spatial meshes and time steps.

The purpose of the present work is to discuss some ways of handling the nonlinear term in this context, and to prove a priori error bounds in the style of [2]. In particular, we want to allow the possibility that the spatial mesh will change on each time level.

To avoid some technical difficulties we assume that such refinement (unrefinement) is performed by the addition (removal) of nodes to the current mesh. In fact, it is known that meshes that are changed in an uncontrolled fashion may yield completely false results, see Dupont [1].

The discontinuous Galerkin method involves integrals of the right-hand side of (1.1) with respect to \( x \) and \( t \), which we evaluate by numerical quadrature. Our main result concerns the effect of such numerical integration on the a priori error analysis. Eriksson and Johnson do not consider this in [2], since for linear problems this is less important.

We describe the pertinent results of [2] in Section 2. In Section 3 we formulate our numerical method and state our main result, which is proved in Section 4. We conclude in Section 5 by providing an example showing the implementation of one of our algorithms.

2. THE LINEAR CASE

In this section we briefly review the notation and results of Eriksson and Johnson [2], which we will use in our analysis. For the discretization of (1.1) with respect to \( x = (x_1, x_2) \) we let \( \mathcal{S} \) be the class of all finite element discretizations \( \mathcal{S} = (h, T, S) \) satisfying the following conditions:

1. \( h \in C^1(\Omega) \) is a positive function satisfying
   \[ |\nabla h(x)| \leq \mu \quad \forall x \in \Omega ; \]

2. \( T = \{K\} \) is a partition of \( \Omega \) into triangles \( K \) of diameter \( h_K \) such that
   \[ c_1 h_K^2 \leq \text{area}(K) \quad \forall K \in T , \]

   where the constants \( c_1 \) and \( M^2 \) are positive.
and associated with the function \( h \) through the inequalities

\[
  c_2 h_K \leq h(x) \leq k_K \quad \forall x \in \Omega , \; K \in T ;
\]

(3) \( S \) is the set of all functions which are continuous on \( \bar{\Omega} \), linear on each \( K \in T \) and vanish on \( \partial \Omega \).

For a given domain \( \Omega \) the positive constants \( c_1, c_2 \) and \( \mu \) characterize the family \( \Sigma \) completely.

For the discretization with respect to \( t \) we introduce a partition

\[
  0 = t_0 < t_1 < \ldots < t_{n-1} < t_n < \ldots \quad \text{of} \quad \mathbb{R}^+ \quad \text{into subintervals} \quad I_n = (t_{n-1}, t_n)
\]

of lengths \( k_n = t_n - t_{n-1} \), and we associate with each \( I_n \) a finite element discretization \( \mathcal{S}_n = (h_n, T_n, S_n) \in \Sigma \). For \( q = 0 \) and 1 we define

\[
  \mathcal{V}^{-q}_n = \left\{ V : V(x, t) = \sum_{j=0}^{q} t^j \varphi_j(x), \; \varphi_j \in S_n \right\}.
\]

The discontinuous Galerkin method for (1.1) with \( f = f(x, t) \) consists in computing a function \( U \) such that \( U|_{\Omega \times I_n} \in \mathcal{V}^{-q}_n \) and

\[
  \int_{I_n} ((V_t, X) + (\nabla U, \nabla X)) \, dt + ([V]_{n-1}, X_{n-1}^+) = \int_{I_n} (f(\cdot, t), X(t)) \, dt \quad \forall X \in \mathcal{V}^{-q}_n , \quad (2.1)
\]

for \( n = 1, 2, \ldots \), and \( U_0^- = u_0 \), where \([V]_n = V_n^+ - V_n^-\), \( V_n^\pm = \lim_{s \to 0^\pm} V(t_n + s) \) (\( V_n^- \) should be considered the « nodal value » of \( V \) in \( \mathcal{V}^{-q}_n \)). Here and below \( (\cdot, \cdot) \) denotes the usual inner product in \( L_2 = L_2(\Omega) \) and \( \| \cdot \| \) is the corresponding norm. Hence, we are computing a finite element approximation to \( u \) on space-time « slabs » \( \Omega \times I_n \). By virtue of the discontinuity of \( U \) in time, we can alter the spatial mesh from one time interval to the next.

It turns out that for \( q = 0 \) the scheme (2.1) reduces to the following modification of the backward Euler method:

\[
  \frac{1}{k_n} (U_n^- - U_{n-1}^-, X) + (\nabla U_n^-, \nabla X) = \frac{1}{k_n} \int_{I_n} (f(\cdot, t), X) \, dt \quad \forall X \in S_n .
\]

For \( q = 1 \) in the linear homogeneous case \( f \equiv 0 \) the approximation agrees at the nodal points \( t_n \) with the subdiagonal third order accurate Padé difference approximation. The approximation is second order accurate in the interiors of the intervals \( I_n \) (see [3]).

We quote the following a priori error bound:
THEOREM 2.1 (Eriksson and Johnson [2]): Let \( u \) be the solution of (1.1) with \( f = f(x, t) \) and \( U \) that of (2.1). Suppose that \( \mu \) is sufficiently small and assume that for each \( n \) one of the following assumptions hold:

\[
S_n < S_{n-1} \quad \text{or} \quad \left( \max_{x \in \Omega} h_n(x) \right)^2 \leq \gamma k_n,
\]

with \( \gamma \) sufficiently small and that \( k_n \leq c k_{n+1} \) for all \( n \). Then there exists a constant \( C \) depending only on \( c_x \) and \( c_2 \) such that for \( q = 0, 1 \) and \( n = 1, 2, \ldots \),

\[
\| U - u \|_{L^{\infty}(I_n, L^2)} \leq C L_n \max_{1 \leq j \leq n} E_j^{(q)}(u),
\]
\[
\| U_n^\tau - u(t_n) \| \leq C L_n \max_{1 \leq j \leq n} E_j^{(2q)}(u),
\]

with \( L_n = \frac{1}{4} \sqrt{1 + \log \frac{t_n}{k_n}} \) and

\[
E_j^{(p)}(u) = \| h_j^2 D^2 u \|_{L^{\infty}(I_j, L^2)} + \min_{s \leq p + 1} k_s^p \| u_s^{(1)} \|_{L^{\infty}(I_j, L^2)},
\]

where \( u_1^{(1)} = u_t, u_2^{(2)} = u_{tt}, u_3^{(3)} = \Delta u_{tt} \) and \( D^m u = \left( \sum_{\alpha + m} |D_x^\alpha u|^2 \right)^{1/2} \) (with the usual multi-index notation).

These error bounds imply that the scheme is of second order in \( x \), of order \( q + 1 \) uniformly in \( t \), and of order \( 2q + 1 \) at the nodes \( t_n \).

3. THE SEMILINEAR CASE

In this section we discuss some ways of implementing the scheme (2.1) in the semilinear case, i.e., when \( f = f(x, t, u) \). (In the sequel we sometimes write \( f(u) \) instead of \( f(x, t, u(x, t)) \) for compactness of notation. We also use the notation \( u_t = u(\cdot, t_i), f(u)_t = f(\cdot, t_i, u_t) \) and \( f(V)_t = f(\cdot, t_i, V_t^-) \) for \( V \in \mathcal{V}^t \)).

In this case the right-hand side of (2.1) is an integral of \( f(U)X \) over \( \Omega \times I_n \), which must be evaluated by numerical quadrature if the algorithm is to be employed in a general way. This is the question that we address here. We desire that the completely discrete schemes should retain the order that the discontinuous Galerkin schemes have on linear problems. In particular, for \( q = 1 \), in order to retain the nominal third order accuracy of the discretization of the left-hand side of (2.1), we employ a multistep formula.
based on interpolation of $f(U)$ at the nodes $t_n$, where we expect third order accuracy for $U$.

For the integral with respect to $x$ we choose the following quadrature rule. Let $\mathcal{S} = (h, T, S) \in \Sigma$ be a finite element discretization and define

$$Q_K(f) = \frac{1}{3} \text{area}(K) \sum_{j=1}^{3} f(P_{K,j}) \approx \int_K f(x) \, dx \quad \forall K \in T,$$

where $P_{K,j}$ are the vertices of the triangle $K$. We may then define approximations of the inner product and norm of $L_2$ by

$$(\chi, \psi)_S = \sum_{K \in T} Q_K(\chi \psi), \quad \| \chi \|_S = (\chi, \chi)_S^{1/2}.$$

For the approximation of the integral with respect to $t$ we replace $f(U)$ by an interpolant with respect to $t$ and integrate the resulting polynomials exactly. Two possibilities suggest themselves: one is to use extrapolation of an interpolant computed over previous intervals $I_l$, $l < n$, and the other is to use an interpolant computed over the current and previous intervals $I_l$, $l \leq n$. The former process will yield a set of linear equations for $U$ (semi-implicit method), whereas the latter will produce a nonlinear system (nonlinear implicit method).

We use an interpolant of order $p = 0$ when $q = 0$ and of order $p = 2$ when $q = 1$. Of course, when $p = 2$ we cannot use this kind of interpolant on the first one or two intervals, necessitating the construction of special interpolants there.

Thus our scheme is of the form: find a function $U$ such that $U|_{\Omega \times I_0} \in V_{n}^{q}$, $U_0 = u_0$ and

$$\int_{I_n} \left( (U(t), X) + (\nabla U, \nabla X) \right) \, dt + ([U]_{n-1}, X_{n-1}^+) =$$

$$= \int_{I_n} \langle \Pi f(U)(t), X(t) \rangle \, dt \quad \forall X \in V_{n}^{q}, \quad 0 < t_n < t^*, \quad (3.1)$$

where (except possibly for $n = 1$ or 2) the integrand of the right-hand side is given by

$$\langle \Pi f(U)(t), X(t) \rangle = \langle \Pi_{n-1}^p f(U)(t), X(t) \rangle =$$

$$= \sum_{l = n-1-p}^{n-1} \phi_l(t)(f(\cdot, t_l, U_l^+), X(t))_{S_l \cup S_n}. \quad (3.2)$$

Here $\Pi_{n-1}^p$ is the interpolation operator for polynomials of degree $p = 2q$, computed with respect to the mesh points $\{t_l\}_{l = n-1-p}$ and

vol. 27, n° 1, 1993
\( \phi_i \) are the corresponding Lagrange basis functions. Note that \( i = 0 \) for the nonlinear implicit scheme and \( i = 1 \) for the semi-implicit scheme.

In (3.2) we handle the possibility of variable meshes in the following way: we assume that all spatial meshes that occur are refinements of one common coarse mesh, and that each triangulation \( T_n \) is obtained from its precursor \( T_{n-1} \) by adding some nodes and by removing some other nodes. In this way, the union of the mesh points of two triangulations \( T_i \) and \( T_n \) form a triangulation whose finite element space is equal to \( S_i \cup S_n \). In this situation we define

\[
    h_{i,n}(x) = \min \{ h_i(x), h_n(x) \}.
\]

The discrete \( L_2 \) inner product in (3.2) is thus computed over the totality of all mesh points of \( T_i \) and \( T_n \).

When \( q = 0 \) the equation (3.1) becomes

\[
    \frac{1}{k_n} (U_n^\tau - U_{n-1}^\tau, \chi) + (\nabla U_n^\tau, \nabla \chi) =
\]

\[
    = (f(\cdot, t_{n-i}, U_{n-1}^\tau), \chi)_{S_{n-i} \cup S_n} \quad \forall \chi \in S_n, \quad 0 < t_n \leq t^*,
\]

which is a variant of the backward Euler method. In Section 5 we give details of the implementation of a scheme with \( q = 1 \).

It remains to devise a starting procedure for the case \( q = 1, p = 2 \), i.e., to define the right-hand side \( \int_{I_n} \langle II f(U), X \rangle \, dt \) for \( 1 \leq n \leq 1 + i \). For the nonlinear implicit scheme \((i = 0)\) we simply take

\[
    \langle II f(U), X \rangle = \langle I_1^I f(U), X \rangle \quad \text{on} \quad I_1,
\]

i.e., we use a linear interpolant in (3.2). For the semi-implicit method \((i = 1)\) we similarly define

\[
    \langle II f(U), X \rangle = \langle I_1^I f(U), X \rangle \quad \text{on} \quad I_2.
\]

For \( n = 1 \) we use a prediction-correction procedure in order to obtain the correct accuracy. For the predicted value of \( U_1^\tau \) we define \( \tilde{U} \mid_{\alpha \times I_1} \in \mathcal{V}^q \) such that \( \tilde{U}_0^\tau = u_0 \) and

\[
    \int_{I_1} ((\tilde{U}_t, X) + (\nabla \tilde{U}, \nabla X)) \, dt + ([\tilde{U}]_0, X_0^+) =
\]

\[
    = \int_{I_1} \langle I_0^\tau f(\tilde{U}), X \rangle \, dt \quad \forall X \in \mathcal{V}^{-q},
\]
that is,
\[
\int_{I_1} \left( (\tilde{U}, X) + (\nabla \tilde{U}, \nabla X) \right) dt + ([\tilde{U}]_0, X_0^+) = \int_{I_1} \langle f(\cdot, 0, u_0), X(t) \rangle_{C_0 \cup S_1} dt \quad \forall X \in \mathcal{V}^q.
\]

Then we compute \( U \big|_{\Omega \times I_1} \in \mathcal{V}^q \) such that \( U_0^- = u_0 \) and
\[
\int_{I_1} \left( (U, X) + (\nabla U, \nabla X) \right) dt + ([U]_0, X_0^+) = \int_{I_1} \langle \Pi_1^1 f(\tilde{U}), X \rangle dt \quad \forall X \in \mathcal{V}^q.
\]

For the scheme described above we have the following a priori error bound. We need all the assumptions of this and the previous sections and, in addition, we assume that, for some \( c_3, c_4 > 0 \),
\[
c_3 \leq \frac{k_n}{k_{n-1}} \leq c_4, \quad 0 < t_n \leq t^*.
\] (3.3)

It is also convenient to define
\[
m = \max \{ i + p - 1, 0 \} = \begin{cases} 0, & \text{if } q = 0, \\ 1 + i, & \text{if } q = 1, \end{cases}
\] (3.4)
to distinguish between the general case of (3.2) \( (t_n = t_m + \frac{1}{2}) \) and the special initialization procedure \( (0 < t_n \leq t_m) \) needed when \( q = 1 \).

**Theorem 3.1:** Let \( u \) be the solution of (1.1) and \( U \) be that of (3.1). Then there are constants \( k \) and \( C \) depending on \( M, t^* \) and \( c_i, i = 1, ..., 4 \), such that, if \( k_n \leq k \) for \( 0 < t_n \leq t^* \), then
\[
\| U - u \|_{L^\infty(U_n, L_2)} \leq C L_n \max_{1 \leq j \leq n} E_j^{(q)}(u) + C \max_{1 \leq j \leq n} F_j^{(2,q)}(u), \quad 0 < t_n \leq t^*,
\]
\[
\| U_{n} - u(t_n) \| \leq C L_n \max_{1 \leq j \leq n} E_j^{(2,q)}(u) + C \max_{1 \leq j \leq n} F_j^{(2,q)}(u), \quad 0 < t_n \leq t^*.
\]

For the semi-implicit schemes \((i = 1)\) the condition \( k_n \leq k \) is not needed. Here \( L_n \) and \( E_j^{(p)}(u) \) are as in Theorem 2.1 and
\[
F_j^{(p)}(u) = \sum_{l = j - i - p}^{j-i} \left( \| h_{l,j}^2 D_x u_l \| + \| h_{l,j}^2 D_x^2 f(u)_l \| + \| h_{l,j}^2 D_x f(u)_l \| \right) + \\
+ \min_{s \leq p + 1} (k_j^s \| D_s^f f(u) \|_{L^\infty(U_n, L_2)}), \quad t_{m+1} \leq t \leq t^*.
\]
and for \( q = 1 \) we have, in addition, if \( i = 0 \),

\[
F^{(2)}_1(u) = \sum_{l=0}^{1} \left( \| h_{i,l}^2 D_x^2 u_l \| + \| h_{i,l}^2 D_x^2 f(u)_l \| + \| h_{i,l}^2 D_x f(u)_l \| \right) + \\
+ \min_{s = 2} \left\{ k_{i,l}^{s+1} \| D_s^i f(u) \|_{L^\infty(I_l, I_{l+1})} \right\},
\]

and, if \( i = 1 \),

\[
F^{(2)}_2(u) = \sum_{l=0}^{1} \left( \| h_{i,l}^2 D_x^2 u_l \| + \| h_{i,l}^2 D_x^2 f(u)_l \| + \| h_{i,l}^2 D_x f(u)_l \| \right) + \\
+ \min_{s = 2} \left\{ k_{i,l}^{s+1} \| D_s^i f(u) \|_{L^\infty(I_l, I_{l+1})} \right\} + \min_{s = 1} \left\{ k_{i,l}^{s+2} \| D_s^i f(u) \|_{L^\infty(I_l, I_{l+1})} \right\},
\]

and

\[
F^{(2)}_2(u) = \sum_{l=0}^{1} \left( \| h_{i,l}^2 D_x^2 u_l \| + \| h_{i,l}^2 D_x^2 f(u)_l \| + \| h_{i,l}^2 D_x f(u)_l \| \right) + \\
+ \max_{1 \leq i \leq 2} \min_{s = 2} \left\{ k_{i,l}^{s+1} \| D_s^i f(u) \|_{L^\infty(I_l, I_{l+1})} \right\}.
\]

**Remark**: If one would settle for second order convergence (when \( q = 1 \)), then \( II f(U) \) in (3.2) could be taken to be a linear interpolant, which could even be computed at interior points of the interval \( I_{n-1} \) or \( I_n \), with corresponding simplifications of the analysis and implementation.

4. PROOF OF THEOREM 3.1

For the proof of Theorem 3.1 we shall compare the solution \( U \) of (3.1) with the solution \( V \) of the discrete linear problem (2.1) with \( f(x, t) \) replaced by \( f(x, t, u(x, t)) \). That is, \( V \begin{array}{c} \Omega \times I_n \in \mathcal{V}^{-q}_{n}, 0 < t_n \leq t^* \end{array} \) is defined by \( V_0^- = u_0 \)

\[
\int_{I_n} ((V_t, X) + \nabla V, \nabla X)) \, dt + ([V]_{n-1}, X_{n-1}^+) = \\
= \int_{I_n} (f(u), X) \, dt \quad \forall X \in \mathcal{V}^{-q}_{n}. \quad (4.1)
\]

Theorem 2.1 then shows that

\[
\| V_-^n - u_n \| \leq C L_n \max_{1 \leq j \leq n} E_j^{(2,q)}(u), \quad 0 < t_n \leq t^*, \quad (4.2)
\]

M² AN Modélisation mathématique et Analyse numérique
Mathematical Modelling and Numerical Analysis
and the proof of Theorem 3.1 will be accomplished once we have shown that for \( \theta = U - V \) we have
\[
\| \theta \|_{L^2(I_a; L^2)} \leq C L_n \max_{1 \leq j \leq n} E_j^{(2,q)}(u) + C \max_{1 \leq j \leq n} F_j^{(2,q)}(u), \quad 0 \leq t_n \leq t^* . \tag{4.3}
\]

In order to prove this we shall show below that
\[
\| \theta_n^- \| \leq C L_n \max_{1 \leq j \leq n} E_j^{(2,q)}(u) + C \max_{1 \leq j \leq n} F_j^{(2,q)}(u), \quad 0 \leq t_n \leq t^* . \tag{4.4}
\]

It is convenient to begin by demonstrating that (4.3) follows from (4.4). From (3.1) and (4.1) it follows that \( \theta \big|_{\Omega \times I_a} \in \mathcal{V}^q_n \), \( 0 \leq t_n \leq t^* \), satisfies \( \theta_0 = 0 \) and
\[
\int_{I_a} ((\theta, X) + (\nabla \theta, \nabla X)) \, dt + ([\theta]_{n-1}, X_{n-1}^+ ) = \\
\int_{I_a} (\langle \Pi f(U), X \rangle - (f(u), X)) \, dt \quad \forall X \in \mathcal{V}^q_n . \tag{4.5}
\]

We split the integrand on the right-hand side into three terms,
\[
\langle \Pi f(U), X \rangle - (f(u), X) = \langle \Pi [f(U) - f(u)], X \rangle \\
+ \langle \Pi f(u), X \rangle - (\Pi f(u), X) \\
+ ([\Pi - I] f(u), X) . \tag{4.6}
\]

These are estimated in the following three lemmas. The first term requires a uniform Lipschitz condition on the nonlinearity \( f \), which holds in view of assumption (1.2).

**LEMMA 4.1:** Let \( U \big|_{\Omega \times I_a} \in \mathcal{V}^q_l \) for \( j - i - p \leq l \leq j - i \) and \( X \in \mathcal{V}^q_n \). Then
\[
\left| \langle \Pi_{j-i}^p [f(U) - f(u)](t), X(t) \rangle \right| \leq \\
CM \sum_{l = j - i - p}^l (\| U_l^i - u_l \| + \| h_{l,j} D_x^2 u_l \|) \| X(t) \| , \quad t \in I_j .
\]

**Proof:** Using the uniform boundedness of the Lagrangian basis functions,
\[
| \phi_l(t) | \leq C , \quad t \in I_j , \tag{4.7}
\]

vol. 27, n° 1, 1993
under the assumption (3.3), and employing the Lipschitz condition for $f$, we have

$$\left| \langle \Pi_{j-1}^{p} [f(U) - f(u)](t), X(t) \rangle \right| =$$

$$= \left| \sum_{i=j-i-p}^{i} \phi_i(t) \langle [f(U)_i - f(u)_i], X(t) \rangle_{S_i \cup S_j} \right| \leq$$

$$\leq CM \left( \sum_{i=j-i-p}^{i} \| U_i^+ - u_i \|_{S_i \cup S_j} \right) \| X(t) \|_{S_i \cup S_j}.$$ 

But $\| \cdot \|_{S_i \cup S_j}$ and $\| \cdot \|$ are uniformly equivalent norms on $S_i \cup S_j$, and $X(t) \in S_j \subset S_i \cup S_j$, so that $\| X(t) \|_{S_i \cup S_j} \leq C \| X(t) \|$. Similarly, with $J_{S_i \cup S_j}: C(\bar{\Omega}) \rightarrow S_i \cup S_j$ the Lagrange interpolation operator, we have

$$\| U_i^+ - u_i \|_{S_i \cup S_j} = \| J_{S_i \cup S_j}(U_i^+ - u_i) \|_{S_i \cup S_j} \leq C \| U_i^+ - J_{S_i \cup S_j} u_i \| \leq$$

$$\leq C (\| U_i^+ - u_i \| + \| u_i - J_{S_i \cup S_j} u_i \|),$$ 

since $U_i^+ \in S_i \subset S_i \cup S_j$. In view of a well-known error bound for $J_{S_i \cup S_j}$ this proves the lemma. ■

The second term on the right of (4.6) involves the error in spatial quadrature.

**Lemma 4.2** Let $X \in \mathcal{V}^{1, 1}_{j}$. Then

$$\left| \langle [\Pi_{j-1}^{p} f(u)](t), X(t) \rangle - \langle [\Pi_{j-1}^{p} f(u)](t), X(t) \rangle \right| \leq$$

$$\leq C \sum_{i=j-i-p}^{i} \left( \| h_{i,j}^2 D_x f(u)_i \| + \| h_{i,j}^2 D_x f(u)_i \| \right) \| X(t) \|_{H^1}, \quad t \in I_j.$$ 

**Proof** Using (4.7) we have

$$\left| \langle [\Pi_{j-1}^{p} f(u)](t), X(t) \rangle - \langle [\Pi_{j-1}^{p} f(u)](t), X(t) \rangle \right| =$$

$$= \left| \sum_{i=j-i-p}^{i} \phi_i(t) \epsilon_{i,j}(f(u)_i, X(t)) \right| \leq$$

$$\leq C \sum_{i=j-i-p}^{i} \left| \epsilon_{i,j}(f(u)_i, X(t)) \right|,$$

where $\epsilon_{i,j}(\ldots, \ldots)_{S_i \cup S_j} - (\ldots, \ldots)$ is the quadrature error, for which we have

$$\| \epsilon_{i,j}(f, X) \| \leq C (\| h_{i,j}^2 D_x^2 f \| + \| h_{i,j}^2 D_x f \|) \| X \|_{H^1},$$

$$f \in H^2(\Omega), \quad X \in S_i \cup S_j.$$
This follows by a simple modification of the proof of Lemma 2.3 of [4] or Lemma 1 on p. 170 of [5]. Since $X(t) \in S \subset S_t \cup S_j$ this proves the lemma.

The third term in (4.6) involves the error in interpolation with respect to $t$. The proof of the following lemma is well-known and we omit it. The assumption (3.3) is used here again.

**Lemma 4.3** : Let $m = \max \{i+p-1, 0\}$. We have

$$
\| (\Pi^p_{j-1} - I) f(u) \|_{L^\infty(I_j, L_2)} \leq C \max_{j-m \leq l \leq j} \min_{s = p+1} \left\{ k_l^i \| D^s f (u) \|_{L^\infty(I_j, L_2)} \right\}.
$$

We can now estimate $\| \theta \|_{L^\infty(I_j, L_2)}$ by applying a simple energy argument to (4.5). We formulate this as a lemma.

**Lemma 4.4** : Let $U \in \mathcal{V}_j^q$ for $j - i - p \leq l \leq j - i$ and assume that $\theta \in \mathcal{V}_j^q$ satisfies

$$
\int_{I_j} ((\theta_t, X) + (\nabla \theta, \nabla X)) dt + ([\theta]_{j-1}, X^+_{j-1}) =
$$

$$
= \int_{I_j} \langle \Pi^p_{j-1} f(U), X \rangle - (f(u), X) \rangle dt \text{ for all } X \in \mathcal{V}_j^q.
$$

Then with $m = \max \{i+p-1, 0\}$ we have

$$
\| \theta \|_{L^\infty(I_j, L_2)} \leq C \| \theta^+_{j-1} \| + CMk_j \sum_{l = j-1-p}^{j-i} \left( \| U_l^+ - u_l \| + \| h^2_{l,j} D^2_x u_l \| \right) +
$$

$$
+ C \sqrt{k_j} \sum_{l = j-1-p}^{j-i} \left( \| h^2_{l,j} D^2_x f(u)_l \| + \| h^2_{l,j} D_x f(u)_l \| \right) +
$$

$$
+ C k_j \max_{j-m \leq l \leq j} \min_{s = p+1} (k_l^i \| D^s f (u) \|_{L^\infty(I_j, L_2)}).
$$

**Proof** : Since $\theta(t)$ is a polynomial in $t$ of degree 0 or 1, we have

$$
\| \theta \|_{L^\infty(I_j, L_2)} \leq \max \{ \| \theta^+_{j-1} \|, \| \theta^+_{j-1} \| \},
$$

so it suffices to estimate the two terms on the right. Taking $X = \theta$ in (4.8), we have

$$
\int_{I_j} \langle (\theta_t, \theta) + (\nabla \theta, \nabla \theta) \rangle dt + ([\theta]_{j-1}, \theta^+_{j-1}) =
$$

$$
= \int_{I_j} \langle \Pi^p_{j-1} f(U), \theta \rangle - (f(u), \theta) dt ,
$$

vol. 27, n° 1, 1993

---

**DISCONTINUOUS GALERKIN METHOD**

45
whence,
\[ \| \theta_j^- \|^2 + \| \theta_{j-1}^+ \|^2 + \| \theta \|^2_{L_2(U_j, H')} \leq C \| \theta_{j-1}^- \|^2 + 
+ C \int_{I_j} (\langle \Pi_j^{p-1} f(U), \theta \rangle - (f(u), \theta)) \, dt . \]

In view of Lemmas 4.1, 4.2 and 4.3, we have here
\[ \int_{I_j} (\langle \Pi_j^{p-1} f(U), \theta \rangle - (f(u), \theta)) \, dt = \int_{I_j} (\Pi_j^{p-1} [f(U) - f(u)], \theta) \, dt + 
+ \int_{I_j} (\Pi_j^{p-1} f(u), \theta) - (\Pi_j^{p-1} f(u), \theta)) \, dt + 
+ \int_{I_j} ([\Pi_j^{p-1} - I] f(u), \theta) \, dt \leq 
\leq C k_j R_1 \| \theta \|_{L_2(U_j, L_2)} + C \sqrt{k_j} R_2 \| \theta \|_{L_2(U_j, H')} + C k_j R_3 \| \theta \|_{L_2(U_j, L_2)} \leq 
\leq C \varepsilon^{-1} k_j^2 (R_1^2 + R_2^2) + C \varepsilon^{-1} k_j R_2^2 + \varepsilon \| \theta \|^2_{L_2(U_j, H')} + \varepsilon \| \theta \|^2_{L_2(U_j, H')} , \]
where \( \varepsilon > 0 \) and
\[
R_1 = M \sum_{l=1}^{j-1-p} (\| U_i - u_i \| + \| h^2_{i,j} D_x^2 u_i \| ) , \\
R_2 = \sum_{l=1}^{j-1-p} (\| h^2_{i,j} D_x^2 f(u_i) \| + \| h^2_{i,j} D_x f(u_i) \| ) , \\
R_3 = \max_{j-m \leq l \leq j} \min_{s \leq p+1} (k_j^s \| D_x^s f(u) \|_{L_2(U_j, L_2)} ) .
\]

Hence, choosing \( \varepsilon \) small enough we obtain the desired result. \( \Box \)

We are now in a position to finish the proof that (4.4) implies (4.3). We have to distinguish between the general case and the startup terms. In the general case, i.e., when \( n \geq m + 1 \), where \( m \) is defined in (3.4), we have
\( \Pi = \Pi_{n-1}^{2q} \) in (4.5) and Lemma 4.4, (4.2) and (4.4) show
\[
\| \theta \|_{L_2(U_n, L_2)} \leq C \max_{1 \leq j \leq n} \left( \| \theta_j^- \| + \| V_j^- - u_j \| + \sum_{l=j-1-2q}^{j-1-p} (\| h^2_{i,j} D_x^2 u_i \| + 
+ \| h^2_{i,j} D_x^2 f(u_i) \| + \| h^2_{i,j} D_x f(u_i) \| ) + 
+ \min_{s \leq 2q+1} (k_j^s \| D_x^s f(u) \|_{L_2(U_j, L_2)} ) \right) \leq 
\leq C L_n \max_{1 \leq j \leq n} E_j^{(2q)}(u) + C \max_{1 \leq j \leq n} F_j^{(2q)}(u) , \quad t_m+1 \leq t_n \leq t^* .
\]
This is the desired result in the general case.

If $q = 1$ we also have to consider the startup terms. If $q = 1$, $i = 0$, and $n = 1$, then we have $II = II_1^0$ in (4.5) and Lemma 4.4 shows

$$
\| \theta \|_{L_\infty(I_1, L_2)} \leq C \left( \| \theta_{\tilde{I}} \| + \| V_{\tilde{I}} - u_1 \| + \sum_{l=0}^{1} \left( \| \tilde{h}_{l, 1}^2 D_x^2 u_l \| + 
\| \tilde{h}_{l, 1}^2 D_x^2 f(u_l) \| + \| \tilde{h}_{l, 1}^2 D_x f(u_l) \| \right) + 
\right)

+ C \min_{s = 2} \left\{ k_{s+1}^i \| D_s^i f(u) \|_{L_\infty(I_1, L_2)} \right\} \leq

\leq CL_1 \varepsilon_1^{(2)}(u) + CF_1^{(2)}(u). \quad (4.9)

If $q = 1$, $i = 1$, then we first apply Lemma 4.4 with $\theta$ replaced by $\tilde{\theta} = \tilde{U} - V$ and with $II = II_0^0$. This gives

$$
\| \tilde{\theta} \|_{L_\infty(I_1, L_2)} \leq C \left( \| h_{0, 1}^2 D_x^2 u_0 \| + \| h_{0, 1}^2 D_x^2 f(u)(t_0) \| + \| h_{0, 1}^2 D_x f(u)(t_0) \| + 
\right)

+ \min_{s = 2} \left\{ k_{s+1}^i \| D_s^i f(u) \|_{L_\infty(I_1, L_2)} \right\} .

$$

Next we apply the lemma with $\theta = U - V$ and with $II_1^1 f(\tilde{U}) - f(u)$ on the right-hand side. This gives

$$
\| \theta \|_{L_\infty(I_1, L_2)} \leq C \left( k_1 \| \tilde{\theta} \| + \| V_{\tilde{I}} - u_1 \| + \sum_{l=0}^{1} \left( \| \tilde{h}_{l, 1}^2 D_x^2 u_l \| + 
\| \tilde{h}_{l, 1}^2 D_x^2 f(u_l) \| + \| \tilde{h}_{l, 1}^2 D_x f(u_l) \| \right) + 
\right)

+ C \min_{s = 2} \left\{ k_{s+1}^i \| D_s^i f(u) \|_{L_\infty(I_1, L_2)} \right\} .

Using the above bound for $\tilde{\theta} = \tilde{U} - V$ we conclude

$$
\| \theta \|_{L_\infty(I_1, L_2)} \leq CL_1 \varepsilon_1^{(2)}(u) + CF_1^{(2)}(u) .

$$

We then continue to the interval $I_2$ and use Lemma 4.4 with $\theta = U - V$, $II = II_1^1$:

$$
\| \theta \|_{L_\infty(I_2, L_2)} \leq C \left( \| \theta_{\tilde{I}} \| + \| V_{\tilde{I}} - u_1 \| + \sum_{l=0}^{1} \left( \| h_{l, 1}^2 D_x^2 u_l \| + 
\| h_{l, 1}^2 D_x^2 f(u_l) \| + \| h_{l, 1}^2 D_x f(u_l) \| \right) + 
\right)

+ C \max_{1 \leq i \leq 2} \min_{s = 2} \left\{ k_{s+1}^i \| D_s^i f(u) \|_{L_\infty(I_1, L_2)} \right\} \leq

\leq CL_1 \varepsilon_1^{(2)}(u) + CF_2^{(2)}(u) .

vol. 27, n° 1, 1993
Assumption (3.3) was also employed here. This completes the proof that (4.4) implies (4.3).

We now proceed to prove (4.4). Following [2] we represent $\| \theta_n \|$ by duality. For this purpose we define $\mathcal{V}^q = \{ V : V \in \mathcal{V}^q, 0 < t_n \leq t^* \}$. We then note that, by summation of equation (4.5), we have

$$B(\theta, X) = \sum_{j=1}^{n} \int_{I_j} \langle \Pi f(U), X \rangle \, dt - \int_{0}^{t_n} (f(u), X) \, dt \quad \forall X \in \mathcal{V}^q, \quad (4.10)$$

where

$$B(V, W) =$$

$$\sum_{j=1}^{n} \int_{I_j} ((V_t, W) + (\nabla V, \nabla W)) \, dt + \sum_{j=1}^{n-1} (|V_j, W_j^+| + (V_0^+, W_0^+)) =$$

$$\sum_{j=1}^{n} \int_{I_j} (- (V_t, W) + (\nabla V, \nabla W)) \, dt - \sum_{j=1}^{n-1} (V_j^-, [W_j] + (V_n^-, W_n^-)).$$

Next we consider the discrete analog of the « backward problem »

$$-z_t - \Delta z = 0, \quad 0 < t < t_n; \quad z(t_n) = \theta_n^-.$$

In view of the second form of $B(\ldots)$ above, it is clear that the corresponding discrete problem consists in finding $Z \in \mathcal{V}^q$ such that

$$B(X, Z) = (X_n^-, \theta_n^-) \quad \forall X \in \mathcal{V}^q. \quad (4.11)$$

The following stability bound is proved in [3]:

$$\| Z \|_{L^\infty([0, t_n], L^2)} + \| Z \|_{L^2([0, t_n], H^1)} \leq C \| \theta_n^- \|. \quad (4.12)$$

Taking $X = \theta$ in (4.11) and using (4.10), we obtain

$$\| \theta_n^- \|^2 = B(\theta, Z) = \left( \sum_{j=1}^{m} + \sum_{j=m+1}^{n} \right) \int_{I_j} \left( \langle \Pi f(U), Z \rangle - (f(u), Z) \right) \, dt,$$

where $m$ is as in (3.4) and the first sum is empty if $m = 0$. Splitting the integrand on the right as in (4.6) and using Lemmas 4.1, 4.2 and 4.3, we obtain

$$\left| \sum_{j=m+1}^{n} \int_{I_j} \left( \langle \Pi f(U), Z \rangle - (f(u), Z) \right) \, dt \right| \leq$$

$$\leq C \sum_{j=m+1}^{n} k_j \sum_{l=j+2}^{n} \left( \| \theta_l^- \| + \| V_l^- - u_l \| \right) +$$

$$M^2 \text{ AN Modélisation mathématique et Analyse numérique}$$

Mathematical Modelling and Numerical Analysis
Here we have also used (4.12), (4.2) and the fact that $\delta Q = 0$. The factor $1 - i$ in the first term on the right means that this term vanishes for the lagged schemes ($i = 1$).

If $m = 1$, i.e., $q = 1$, $i = 0$, we have also the term
\[
\left| \int_{I_1} (\langle II_1^1 f(U), Z \rangle - (f(u), Z)) \, dt \right| \leq C k_1 \left( \left\| \theta_1^- \right\| + \left\| V_1^- - u_1 \right\| + \sum_{l=0}^{1} \left\| h_{l_1, 1} D_x^2 u_l \right\| \right) \left\| Z \right\|_{L^\infty(I_1, L_2)} + \\
+ C \sqrt{k_1} \sum_{l=0}^{1} \left( \left\| h_{l, 1} D_x^2 f(u)_l \right\| + \left\| h_{l_1, 1} D_x f(u)_l \right\| \right) \left\| Z \right\|_{L^2(I_1, H^1)} + \\
+ C \min_{s \leq 2} \left( k_{s+1}^2 \left\| D_x^s f(u) \right\|_{L^\infty(I_1, L_2)} \right) \left\| Z \right\|_{L^\infty(I_1, L_2)} \leq \\
\leq (C L_1 E_1^{(2)}(u) + C F_1^{(2)}(u)) \left\| \theta_1^- \right\|. 
\]
where we have used the estimate of $\left\| \theta_1^- \right\|$ from (4.9).

If $m = 2$, i.e., $q = 1$, $i = 1$, we have the terms
\[
\left| \int_{I_1} (\langle II_1^1 f(\bar{U}), Z \rangle - (f(u), Z)) \, dt \right| + \left| \int_{I_2} (\langle II_1^1 f(U), Z \rangle - (f(u), Z)) \, dt \right| \leq \\
\leq C k_1 \left( \left\| \bar{\theta}_1^- \right\| + \left\| V_1^- - u_1 \right\| + \sum_{l=0}^{1} \left\| h_{l_1, 1} D_x^2 u_l \right\| \right) \left\| Z \right\|_{L^\infty(I_1, L_2)} + \\
+ C k_2 \left( \left\| \theta_1^- \right\| + \left\| V_1^- - u_1 \right\| + \sum_{l=0}^{1} \left\| h_{l_1, 1} D_x^2 u_l \right\| \right) \left\| Z \right\|_{L^\infty(I_2, L_2)} + \\
\leq (C L_1 E_1^{(2)}(u) + C F_1^{(2)}(u)) \left\| \theta_1^- \right\|. 
\]
where we have used the already proven bounds for \( \| \tilde{\theta}_1^- \| \) and \( \| \theta_1^- \| \).

Altogether we now have

\[
\| \theta_n^\gamma \| \leq (1 - i) C k_n \| \theta_n^\gamma \| + C \sum_{j=1}^{n-1} k_j \| \theta_j^\gamma \| + C L_n \max_{1 \leq j \leq n} E_j^{(2,q)}(u) + C \max_{1 \leq j \leq n} F_j^{(2,q)}(u),
\]

and (4.4) follows by Gronwall's Lemma. If \( i = 0 \) we first have to eliminate the first term on the right by taking \( k_n \) small. This completes the proof of Theorem 3.1.

5. EXAMPLE

We conclude this paper by providing an example showing the implementation of one of our algorithms in the case \( q = 1 \). We solve the equation \( u_t - \Delta u = 10(u - u^3) \) with homogeneous Dirichlet boundary conditions and \( \Omega = (0, 1) \times (0, 1) \). For the spatial discretization we use the piecewise linear finite element method computed on the fixed mesh shown in figure 1 with mesh spacing \( h = 1/m \). We number the nodes as indicated so that there are \( m = 1/h \) nodes on a side and \( m^2 \) nodes in total.

![Mesh for the sample computation.](image-url)
Associated with this mesh are the finite element space $S$, the mass matrix $A$, $A_{ij} = (\varphi_i, \varphi_j)$ and stiffness matrix $B$, $B_{ij} = (\nabla \varphi_i, \nabla \varphi_j)$, where $\{\varphi_i\}_{i=1}^{m^2}$ is the standard computational basis for $S$. We will also need the lumped mass matrix $\tilde{A}$, $\tilde{A}_{ij} = (\varphi_i, \varphi_j)_S$, see [4] or [5].

We derive the equations for the semi-implicit $q = 1, i = 1$ scheme, which in some sense is the most complicated of our algorithms. We also have experience computing with $q = 0, i = 0, 1$ and $q = 1, i = 0, 1$ in one and two dimensions using fixed spatial meshes, and with $q = 1, i = 0$ in the fully adaptive two dimensional code TRANSI.

Throughout the startup procedure and the rest of the steps the same discrete system results from the left-hand side of (3.1). We let $U \in \mathcal{V}_n^q$ be given by

$$U(t) = U_{n-1}^+ \frac{t - t_n}{k_n} + U_n^- \frac{t - t_{n-1}}{k_n},$$

where $U_{n-1}^+$ and $U_n^-$ now denote vectors of values at the spatial nodes, that is, $U_n^+ = ((U_n^+)_1, (U_n^+)_2, \ldots, (U_n^+)_{m^2})^T$, corresponding to the basis of $S$. (3.1) yields the pair of equations

$$A \frac{U_n^- - U_{n-1}^+}{k_n} \int_{I_n} \frac{t - t_n}{k_n} \, dt + BU_{n-1}^+ \int_{I_n} \frac{(t - t_n)^2}{k_n^2} \, dt +$$

$$+ BU_n^- \int_{I_n} \frac{(t - t_n)(t - t_{n-1})}{-k_n^2} \, dt + AU_{n-1}^- =$$

$$= AU_{n-1}^- + \tilde{A} \int_{I_n} \Pi f(U) \frac{t - t_n}{-k_n} \, dt,$$

and

$$A \frac{U_n^- - U_{n-1}^+}{k_n} \int_{I_n} \frac{t - t_{n-1}}{k_n} \, dt + BU_{n-1}^+ \int_{I_n} \frac{(t - t_n)(t - t_{n-1})}{-k_n^2} \, dt +$$

$$+ BU_n^- \int_{I_n} \frac{(t - t_n)^2}{k_n^2} \, dt = \tilde{A} \int_{I_n} \Pi f(U) \frac{t - t_{n-1}}{k_n} \, dt,$$

where $f(U) = (f((U)_1), f((U)_2), \ldots, f((U)_{m^2}))^T$. This is equivalent to

$$\begin{bmatrix}
\frac{1}{2} A + \frac{1}{3} k_n B & \frac{1}{2} A + \frac{1}{6} k_n B \\
- \frac{1}{2} A + \frac{1}{6} k_n B & \frac{1}{2} A + \frac{1}{3} k_n B
\end{bmatrix}
\begin{bmatrix}
U_{n-1}^+ \\
U_n^-
\end{bmatrix}
= \begin{bmatrix}
0 & A \\
0 & 0
\end{bmatrix}
\begin{bmatrix}
U_{n-2}^+ \\
U_{n-1}^-
\end{bmatrix} +$$

$$+ \begin{bmatrix}
\tilde{A} \int_{I_n} \Pi f(U) \frac{t - t_n}{-k_n} \, dt \\
\tilde{A} \int_{I_n} \Pi f(U) \frac{t - t_{n-1}}{k_n} \, dt
\end{bmatrix},$$

vol. 27, n° 1, 1993
for \( n = 1, 2, \ldots \), with \( U_0^+ = 0 \). We let

\[
\begin{bmatrix}
U_{n-1}^+ \\
U_n^-
\end{bmatrix}
= \begin{bmatrix}
\frac{1}{2}A + \frac{1}{3}k_nB & \frac{1}{2}A + \frac{1}{6}k_nB \\
-\frac{1}{2}A + \frac{1}{6}k_nB & \frac{1}{2}A + \frac{1}{3}k_nB
\end{bmatrix}
\begin{bmatrix}
0 \\
A
\end{bmatrix}.
\]

Then, in order of application, the equations are

\[
B_1 \tilde{U}_1 = AU_0 + \begin{bmatrix}
0 & \frac{1}{2}k_1\bar{A} \\
0 & \frac{1}{2}k_1\bar{A}
\end{bmatrix} \begin{bmatrix}
0 \\
f(U_0^-)
\end{bmatrix},
\]

\[
B_1 U_1 = AU_0 + \begin{bmatrix}
\frac{1}{3}k_1\bar{A} & \frac{1}{6}k_1\bar{A} \\
\frac{1}{6}k_1\bar{A} & \frac{1}{3}k_1\bar{A}
\end{bmatrix} \begin{bmatrix}
f(U_0^-) \\
f(\bar{U}_1^-)
\end{bmatrix},
\]

\[
B_2 U_2 = AU_1 + \begin{bmatrix}
-\frac{k_2^2}{6k_1}\bar{A} & \frac{k_2}{k_1} \left( \frac{k_2}{6} + \frac{k_1}{2} \right) \bar{A} \\
-\frac{k_2^2}{3k_1} \bar{A} & \frac{k_2}{k_1} \left( \frac{k_2}{3} + \frac{k_1}{2} \right) \bar{A}
\end{bmatrix} \begin{bmatrix}
f(U_0^-) \\
f(\bar{U}_1^-)
\end{bmatrix},
\]

and finally, in general,

\[
B_n U_n = AU_{n-1} + \begin{bmatrix}
\frac{k_n^2}{k_n-2(k_n-2+2)} \left( \frac{k_n}{12} + \frac{k_n-1}{6} \right) \bar{A} - \frac{k_n^2}{k_n-2(k_n-2+2)} \left( \frac{k_n}{12} + \frac{k_n-1}{6} + \frac{k_n-2}{6} \right) \bar{A} \\
\frac{k_n}{k_n-2(k_n-2+2)} \left( \frac{k_n}{4} + \frac{k_n-1}{3} \right) \bar{A} - \frac{k_n}{k_n-2(k_n-2+2)} \left( \frac{k_n}{4} + \frac{k_n-1}{3} + \frac{k_n-2}{3} \right) \bar{A}
\end{bmatrix}
\times
\begin{bmatrix}
f(U_{n-2}^-) \\
f(U_{n-2}^-)
\end{bmatrix}.
\]

Implementing these in code is now straightforward. There is one point that should be discussed, namely how to solve the associated systems of equations. In particular, the \( q = 1 \) formulas yield nonsymmetric linear

\[M^2\]
systems which can be expensive to solve. (This difficulty is common in
systems arising from implicit high-order Runge-Kutta schemes, which are
related to these formulas.) Both $B$ and $A$ have a « principal » tridiagonal band
as well as two other bands located $O(1/h)$ above and below the diagonal.
Thus, for the $q = 0$ scheme (backward Euler), using just this band structure
without further rearrangement yields an operations count of $O(1/h^4)$ for
solving the system once. In our case $B_n$ has a block $2 \times 2$ band structure.
Solving this system directly however leads to an operations count of
$O(1/h^6)$ (as if the matrix were full) because of fill-in. Rearranging the matrix
can alleviate this load. For example, it is straightforward to rearrange the
matrix to achieve an operations count of $O(1/h^4)$ once again. First permute
the rows of $B_n$ into the order $1, m^2 + 1, 2, m^2 + 2, \ldots, m^2, 2 m^2$ and then do
the same by columns. The resulting matrix has a band of width 7 down the
diagonal and bands of width 5 located $O(1/h)$ above and below the diagonal.

The following computation was made on a $35 \times 35$ mesh with initial data
equal to 1 inside the disc $\left( x - \frac{1}{2} \right)^2 + \left( y - \frac{1}{2} \right)^2 \leq \frac{1}{16}$ and equal to 0
outside. The solution eventually becomes all 0 at a fairly slow rate. See
figure 2.

\begin{figure}[h]
\centering
\begin{subfigure}{0.4\linewidth}
\centering
\includegraphics[width=\textwidth]{fig1.png}
\caption{Time = 0.0}
\end{subfigure}
\begin{subfigure}{0.4\linewidth}
\centering
\includegraphics[width=\textwidth]{fig2.png}
\caption{Time = 2.2}
\end{subfigure}
\begin{subfigure}{0.4\linewidth}
\centering
\includegraphics[width=\textwidth]{fig3.png}
\caption{Time = 20.8}
\end{subfigure}
\begin{subfigure}{0.4\linewidth}
\centering
\includegraphics[width=\textwidth]{fig4.png}
\caption{Time = 43.3}
\end{subfigure}
\caption{The discontinuous Galerkin $q = 1, i = 0$ approximation for the equation
$u_t = \Delta u = 10 (u - u^3)$.}
\end{figure}

vol. 27, n° 1, 1993
REFERENCES


