INCOMPLETE ITERATION FOR TIME-STEPPING A GALERKIN METHOD FOR A QUASILINEAR PARABOLIC PROBLEM*  

JIM DOUGLAS, JR.†, TODD DUPONT‡ AND RICHARD E. EWING‡

Abstract. An iterative method is presented and analyzed which is based on using a preconditioned conjugate gradient iteration for approximately solving the linear equations produced at each time step by an extrapolated Crank–Nicolson–Galerkin procedure for time-stepping a quasilinear parabolic problem. Optimal order convergence rates are obtained for the iterative method which is shown to be (asymptotically) computationally more efficient than standard second-order-in-time correct methods.

1. Introduction. We shall consider a numerically efficient modification of an extrapolated Crank–Nicolson–Galerkin method [6], [14] for approximating the solution of the quasilinear parabolic problem given by

\begin{equation}
\tag{1.1a}
c(x, u) \frac{\partial u}{\partial t} - \nabla \cdot (a(x, u) \nabla u + b(x, u)) = f(x, t, u), \quad (x, t) \in Q,
\end{equation}

where $Q = \Omega \times J$, $J = (0, T]$, $\Omega$ is a bounded domain in $\mathbb{R}^d$, $d \leq 3$, with boundary $\partial \Omega$, and $b = (b_1(x, u), \ldots, b_d(x, u))$; the solution $u \in C^1(\overline{Q})$ is subjected to the initial condition

\begin{equation}
\tag{1.1b}
u(x, 0) = u_0(x), \quad x \in \Omega,
\end{equation}

and the Neumann boundary condition

\begin{equation}
\tag{1.2}
a(x, u) \frac{\partial u}{\partial \nu} + \nu \cdot b(x, u) = g(x, t), \quad (x, t) \in \partial \Omega \times J,
\end{equation}

where $\nu$ is outward unit normal to $\partial \Omega$.

Essentially, our modification of the extrapolated Crank–Nicolson–Galerkin method consists of using an iterative method based on a preconditioned conjugate gradient iteration employing a fixed preconditioning matrix to approximately solve the linear (extrapolation produces linear rather than nonlinear) algebraic equations at each time step. We preserve the accuracy inherent in the underlying Crank–Nicolson–Galerkin method, and we obtain very nearly optimal work estimates for the arithmetic required to produce the solution of a second-order-in-time correct method for (1.1)–(1.2).

In § 2 we introduce a finite element space, present the hypotheses on (1.1)–(1.2) and its solution $u$, discuss an elliptic projection of $u$, and recall the extrapolated Crank–Nicolson–Galerkin method. In § 3 we derive our modification of the above method and analyze the effect of the conjugate gradient iteration on a single time step. In § 4 we obtain global error estimates for any iterative method satisfying the local estimates already derived for the conjugate gradient procedure. In § 5 computational requirements are studied for our basic procedure and a number of variants that fall under the analysis of § 4.

2. Preliminaries. Let $(u, v) = \int_\Omega uv \, dx$, $\|u\|^2 = (u, u)$, and $(u, v) = \int_{\partial \Omega} uv \, d\sigma$. Also let $H^s = H^s(\Omega)$ be the Sobolev space of order $s$ over $\Omega$ and denote by $\|u\|_s$ the corresponding norm. Let $\{\mathcal{M}_k\}$ be a family of finite-dimensional subspaces of $H^1(\Omega)$ with the following property:

* Received by the editors December 16, 1977.
† Department of Mathematics, University of Chicago, Chicago, Illinois 60637.
‡ Department of Mathematics, Ohio State University, Columbus, Ohio 43210.
There exist an integer \( r \geq 2 \) and a constant \( K_0 \) such that, for \( 1 \leq p \leq r \) and \( \phi \in H^p(\Omega) \),
\[
\inf_{\chi \in \mathcal{M}_h} \{ \| \phi - \chi \| + h \| \phi - \chi \|_1 \} \leq K_0 \| \phi \|_{H^p}.
\]

Assume that the family \( \{ \mathcal{M}_h \} \) also satisfies the following so-called "inverse hypotheses":

There exists a constant \( K_0 \), independent of \( h \), such that for all \( \phi \in \mathcal{M}_h \),
\[
\begin{align*}
a) & \quad \| \phi \|_1 \leq K_0 h^{-1} \| \phi \|, \\
b) & \quad \| \phi \|_{L^\infty} \leq K_0 h^{-d/2} \| \phi \|.
\end{align*}
\]

We shall make one further restriction on \( \mathcal{M}_h \) below in (2.10).

Restrict \( \Omega \) as follows (with \( S \) denoting the collection of restrictions):

\[
S: \quad 1) \quad \text{The Neumann problem for } -\Delta + I \text{ on } \Omega \text{ is } H^2\text{-regular.}
\]

2) The restricted cone property [1] holds on \( \Omega \); i.e., \( \partial \Omega \) is Lipschitz.

The following regularity assumptions on \( a, b, c, f \) and the solution \( u \) of (1.1)–(1.2) are denoted collectively by \( R \):

\[
R: \quad 1) \quad \text{There exist uniform constants } c_*, c^*, a_*, a^*, \text{ and } K_1 \text{ such that, for all } \]
\[
(x, t) \in \bar{Q} \text{ and } q \in \mathbb{R},
\]
\[
\begin{align*}
a) & \quad 0 < c_* \leq c(x, q) \leq c^*, \\
b) & \quad 0 < a_* \leq a(x, q) \leq a^*, \\
c) & \quad |b_i(x, q)| \leq K_1, \quad i = 1, \ldots, d, \\
d) & \quad |f(x, t, q)| \leq K_1.
\end{align*}
\]

2) The functions \( a = a(x, u), b = b_i(x, u), c = c(x, u), \) and \( f = f(x, t, u) \) are continuously differentiable with respect to \( u \) and have a uniform bound \( K_1 \), for \( (x, t) \in \bar{Q} \) and \( q \in \mathbb{R} \):

\[
|a|, |b_i|, |c|, |f|, \frac{\partial a}{\partial u}, \frac{\partial b_i}{\partial u}, \frac{\partial c}{\partial u}, \frac{\partial f}{\partial t}, \frac{\partial^2 a}{\partial u^2}, \frac{\partial^2 b_i}{\partial u^2}, \frac{\partial^3 a}{\partial u^3} \leq K_1.
\]

3) If
\[
\| \phi \|_{L^p(a,b);X} = \| \phi(\cdot, t) \|_{X(a,b)}, \quad 1 \leq p \leq \infty,
\]

and \( u \) is the solution of (1.1)–(1.2), there exists a constant \( K_2 \) such that
\[
\begin{align*}
\| u \|_{L^\infty(J;H^1)} + \left\| \frac{\partial u}{\partial t} \right\|_{L^2(J;H^{-1})} + \left\| \frac{\partial^2 u}{\partial t^2} \right\|_{L^2(J;H^2)} + \left\| \frac{\partial^3 u}{\partial t^3} \right\|_{L^1(J;H^3)} \leq K_2.
\end{align*}
\]

We note that under the hypotheses of the theorems and corollaries to follow, our approximations converge uniformly to \( u \); thus (2.4) and (2.5) actually need hold only in a neighborhood of the solution.

Let \( \Delta t > 0, N = T/\Delta t \in \mathbb{Z} \), and \( t^\sigma = \sigma \Delta t, \quad \sigma \in \mathbb{R} \). Also, let \( \varphi^n = \varphi(x, t^n), \varphi^{n+1/2} = (\varphi^{n+1} + \varphi^n)/2, \) and \( d_t^\varphi = (\varphi^{n+1} - \varphi^n)/\Delta t \).

The analysis proceeds, following Wheeler [15] and Rachford [14], via an auxiliary elliptic problem. Define \( W \in \mathcal{M}_h \) to be the unique function which, for \( t \in J \), satisfies
\[
(a(\cdot, u(\cdot, t))\nabla [W(\cdot, t) - u(\cdot, t)] + \nabla \chi) + (\nabla \chi) = 0, \quad \chi \in \mathcal{M}_h.
\]
Thus $W$ is a weighted $H^1$-projection of $u$, the solution of (1.1)-(1.2). As in [8], [9], (2.1) and the restrictions $S$ imply the following result.

**Lemma 1.** For some $p$ satisfying $2 \leq p \leq r$, let $u \in L^\infty(J; H^p)$ and $\frac{\partial u}{\partial t} \in L^2(J; H^{p-1})$. Under the assumption $R$, there exists a constant $K_3$, dependent on $\Omega$, $a_*$, $a^*$, $K_0$, $K_1$, and $K_2$ such that if $\eta = u - W$ and $s = 0$ or 1,

\begin{align*}
\text{a) } & \|\eta\|_{L^\infty(J; H^s)} \leq K_3 h^{p-s} \|u\|_{L^\infty(J; H^p)}, \\
\text{b) } & \frac{\partial \eta}{\partial t} \in L^2(J; H^s) \left\{ \|u\|_{L^2(J; H^p)} + \left\| \frac{\partial u}{\partial t} \right\|_{L^2(J; H^p)} \right\}.
\end{align*}

We now make the assumption on $\{\mathcal{M}_h\}$ and $u$ that there exists a constant $K_4$ such that

\begin{equation}
\inf_{\chi \in \mathcal{M}_h} \left\{ \|\varphi - \chi\| + h \|\varphi - \chi\|_1 + h^{d/2} \|\varphi - \chi\|_{L^\infty} + h \|\nabla(\varphi - \chi)\|_{L^\infty} \right\} \leq K_4 h^p(\Omega).
\end{equation}

The hypothesis (2.7) together with (2.9) and (2.11) imply that

\begin{align*}
\text{a) } & \|\eta\|_{L^\infty(J; L^\infty)} \leq C h^{2-(d/2)} \|\mu\|_{L^\infty(J; H^2)}, \\
\text{b) } & \|\nabla \eta\|_{L^\infty(J; L^\infty)} \leq C h^{2-(d/2)} \|u\|_{L^\infty(J; H^3)}, \\
\text{c) } & \frac{\partial \eta}{\partial t} \in L^\infty(J; L^\infty) \left\{ \|u\|_{L^\infty(J; H^2)} + \left\| \frac{\partial u}{\partial t} \right\|_{L^\infty(J; H^2)} \right\}, \\
\text{d) } & \nabla \frac{\partial \eta}{\partial t} \in L^2(J; L^2) \left\{ \|u\|_{L^2(J; H^3)} + \left\| \frac{\partial u}{\partial t} \right\|_{L^2(J; H^3)} \right\}.
\end{align*}

The relations in (2.12), together with (2.7) and the assumption that $\frac{\partial u}{\partial t} \in L^1(J; H^3)$, imply that the terms in (2.10) are bounded. We can also adapt the proofs of corresponding results in [4], [9] to obtain the following lemma.

**Lemma 2.** There exists a constant $K_5$ depending on $K_0$, $K_1$, and $K_2$ such that

\begin{equation}
\left\| \frac{\partial^2 u}{\partial t^2} \right\|_{L^\infty(J; L^1)} + \left\| \frac{\partial^3 u}{\partial t^3} \right\|_{L^2(J; L^2)} + \left\| \frac{\partial^3 u}{\partial t^3} \right\|_{L^1(J; L^1)} \leq K_5.
\end{equation}

Let $u_h : [0, T] \rightarrow \mathcal{M}_h$ be the approximate solution of (1.1)-(1.2) determined by (using the notation $c(u) = c(x, u(x, t))$, etc.)

\begin{equation}
\left( c(u_h) \frac{\partial u_h}{\partial t}, \chi \right) + (a(u_h) \nabla u_h + b(u_h), \nabla \chi) = (f(u_h), \chi) + (g, \chi), \quad \chi \in \mathcal{M}_h, \quad t \in J,
\end{equation}

with $u_h(\cdot, 0) - u_0$ small in a sense to be specified later. It can be shown [15] (in fact it follows from the results in § 4) that

\begin{equation}
\|u - u_h\|_{L^\infty(J; L^2)} + h \|u - u_h\|_{L^\infty(J; H^1)} \leq C(u) h^r,
\end{equation}
provided \( u \) satisfies (2.7). A standard Crank–Nicolson–Galerkin approximation to the solution of (2.14) would result in a time discretization error of the order \((\Delta t)^2\), but it would require the solution of a nonlinear system of algebraic equations at each time step. Rachford [14] has analyzed the following variant of the Crank–Nicolson–Galerkin scheme. Let \( U: \{0 = t_0, t_1, \ldots, t_N = T\} \rightarrow \mathcal{M}_h \) satisfy

\[
\begin{align*}
&\text{a) } (c(EU^n) d_t U^n, \chi) + (a(EU^n) \nabla U^{n+1/2} + b(EU^n), \nabla \chi) \\
&\quad = (f(t^{n+1/2}, EU^n), \chi) + (g(t^{n+1/2}), \chi), \quad \chi \in \mathcal{M}_h, \quad 0 < n < N, \\
&\text{b) } (a(u_0) \nabla [u_0 - U^0], \nabla \chi) = 0, \quad \chi \in \mathcal{M}_h,
\end{align*}
\]

where \( EU^n = \frac{3}{2} U^n - \frac{1}{2} U^{n-1} \). With this definition of \( EU^n \), we see that information is required at two preceding time levels to advance in time. Thus a starting procedure is needed to define \( U^1 \) which will retain the overall accuracy of the method. Such a starting procedure will be discussed in § 3. We note that the method given by (2.16) requires the solution of exactly one linear system of algebraic equations at each time step; however the matrices generated by the linear equations of (2.16) are usually different for each time step. We shall consider a modification of (2.16) which will require the solution of equations associated with one common matrix at all time levels. (Our process actually provides a generalization of and new error estimates for (2.16).)

### 3. Approximate solution of the linear equations by iteration.

In this section we shall present the linear equations to be treated and an iterative method for approximating their solution. We also define a predictor-corrector-corrector starting procedure.

The conjugate gradient procedure presented here provides only one example of the possible modifications of (2.16) that fall under the convergence analysis given in the next section. Any method that provides the norm reduction defined in this section will preserve the results of § 4. Several modifications of the iterative methods presented here are discussed in § 5.

Let \( \{\varphi_i\}_{i=1}^M \) be a basis for \( \mathcal{M}_h \) and denote the solution of (2.16) by

\[
U^m = \sum_{i=1}^M \xi^m_i \varphi_i.
\]

Let

\[
\begin{align*}
&\text{a) } C^m(\theta) = (c_i^m(\theta)) = \left(\left(c \left( E \sum_{i=1}^M \theta_i^m \varphi_i \right), \varphi_i, \varphi_i \right)\right), \\
&\text{b) } A^m(\theta) = (a_i^m(\theta)) = \left(\left(a \left( E \sum_{i=1}^M \theta_i^m \varphi_i \right), \nabla \varphi_i, \nabla \varphi_i \right)\right), \\
&\text{c) } B^m(\theta) = (b_i^m(\theta)) = \left(\left(b \left( E \sum_{i=1}^M \theta_i^m \varphi_i \right), \nabla \varphi_i \right)\right), \\
&\text{d) } F^m(\theta) = (f_i^m(\theta)) = \left(\left(f \left( t^{m+1/2}, E \sum_{i=1}^M \theta_i^m \varphi_i \right), \varphi_i \right)\right), \\
&\text{e) } G^m(\theta) = (G_i^m(\theta)) = \left(\left(g \left( t^{m+1/2}, \varphi_i \right) \right), \right), \\
&\text{f) } C_0 = (c_0 \varphi_i, \varphi_i) \quad \text{and} \quad A_0 = (a_0 \nabla \varphi_i, \nabla \varphi_i),
\end{align*}
\]

for \( i = 1, \ldots, M \) and \( j = 1, \ldots, M \). Here \( a_0 \) and \( c_0 \) can be chosen in a very arbitrary way. A good choice might be \( a_0 = a(x, u_0(x)) \) and \( c_0 = c(x, u_0(x)) \) or if an average value \( \bar{u} \) is more or less known, evaluate \( a \) and \( c \) at \( \bar{u} \).
We can write (2.16a) in the form

\begin{equation}
L''(\xi)(\xi^{n+1} - \xi^n) = \left(C''(\xi) + \frac{\Delta t}{2} A''(\xi)\right)(\xi^{n+1} - \xi^n) = -\Delta tA''(\xi)\xi^n + \Delta t[B''(\xi) + F''(\xi) + G''(\xi)].
\end{equation}

We shall not solve (3.3) exactly; instead, we shall (in this section; see also § 5) use a predetermined number of preconditioned conjugate gradient \cite{2}, \cite{3}, \cite{7}, \cite{10} iterations to advance the solution one time step. The preconditioning matrix will be chosen to be independent of \(n\), and consequently an appropriate “fast Poisson” solution technique, such as nested dissection \cite{11} can be applied to obtain the solution of the linear equations arising below in (3.7c).

Denote by

\begin{equation}
V'' = \sum_{i=1}^{M} \gamma_i \varphi_i
\end{equation}

the approximation to \(U''\), the solution of (2.16). We shall discuss a starting procedure for obtaining \(V^0\) and \(V^1\) later. We now find \(\gamma^{n+1}\) (and thus \(V^{n+1}\)) using a preconditioned conjugate gradient iteration. Let our preconditioner be defined by

\begin{equation}
L_0 = C_0 + \frac{\Delta t}{2} A_0.
\end{equation}

We shall use different initial guesses for \(\xi^{n+1} - \xi^n\) for \(n = 1\) and for \(n > 1\). We shall use linear extrapolation for \(n = 1\) and quadratic extrapolation for \(n > 1\) (defined explicitly in (3.6a)). Specifically, we initialize our iteration as follows:

\begin{align}
\text{a)} & \quad n = 1: x_0 = x_0^2 = y^1 - y^0, \\
\text{a)} & \quad n \geq 2: x_0 = x_0^{n+1} = 2y^n - 3y^{n-1} + y^{n-2}, \\
\text{b)} & \quad n \geq 1: q_0 = q_0^{n+1} = s_0 = s_0^{n+1} = L''(\gamma)x_0 + \Delta tA''(\gamma)y^n - \Delta t[B''(\gamma) + F''(\gamma) + G''(\gamma)].
\end{align}

Then, using the initialization \(x_0, q_0\) and \(s_0\) from (3.6), for \(k = 1, 2, \ldots, \nu - 1\), where the number of iterations \(\nu\) will be chosen later independently of \(n\), set

\begin{align}
\text{a)} & \quad x_{k+1} = x_k + \alpha_k s_k, \quad \text{where} \quad \alpha_k = \frac{-\langle L_0^{-1} q_k, q_k \rangle}{\langle s_k, L''(\gamma) s_k \rangle}, \\
\text{b)} & \quad q_{k+1} = q_k + \alpha_k L''(\gamma) s_k, \\
\text{c)} & \quad s_{k+1} = L_0^{-1} q_{k+1} + \beta_k s_k, \quad \text{where} \quad \beta_k = \frac{\langle L_0^{-1} q_{k+1}, q_{k+1} \rangle}{\langle L_0^{-1} q_k, q_k \rangle},
\end{align}

where \((\cdot, \cdot)_e\) is the Euclidean inner product.

Finally set

\begin{equation}
\gamma^{n+1} = y^n + x_\nu.
\end{equation}

We define \(\tilde{\gamma}^{n+1}\) to be the solution of (3.3) with \(\xi^n\) replaced by \(\gamma^n\); i.e., let \(\tilde{\gamma}^{n+1}\) satisfy

\begin{equation}
L''(\gamma)(\tilde{\gamma}^{n+1} - \gamma^n) = -\Delta tA''(\gamma)y^n + \Delta t[B''(\gamma) + F''(\gamma) + G''(\gamma)].
\end{equation}

It is well known \cite{2}, \cite{3}, \cite{7}, \cite{10} that there exists a constant \(\rho < 1\) such that

\begin{align}
\text{a)} & \quad \|L'(\gamma)^{1/2}(\gamma^n - \gamma^0)\|_e \leq \rho\|L'(\gamma)^{1/2}(\gamma^n - \gamma^0)\|_e, \\
\text{b)} & \quad \|L''(\gamma)^{1/2}(\gamma^{n+1} - \gamma^n)\|_e \leq \rho\|L''(\gamma)^{1/2}(\gamma^{n+1} - 3\gamma^n + 3\gamma^{n-1} - \gamma^{n-2})\|_e, \quad n \geq 2,
\end{align}
where the subscript $e$ indicates the Euclidean norm of the vector. Given $c_0$ and $a_0$ in (3.2f) there exist $\psi_0$ and $\psi_1$ such that, for $n \geq 0$,

$$
0 < \psi_0 \leq \frac{x^T L_0^\alpha x}{x^T L_0 x} \leq \psi_1, \quad 0 \neq x \in \mathbb{R}^M,
$$

where the constants $\psi_0$ and $\psi_1$ are independent of $h$ and depend only on the bounds for the coefficients in (2.4a) and (2.4b). Let

$$
Q = \frac{1 - (\psi_0/\psi_1)^{1/2}}{1 + (\psi_0/\psi_1)^{1/2}}.
$$

Then [2], [3], [7], [10] $\rho < 2Q^\nu$. If $\alpha > 0$ and

$$
\nu \geq \alpha \log \frac{1}{\Delta t} \log \frac{1}{Q},
$$

then

$$
\rho < 2(\Delta t)^\nu.
$$

Note that

$$
V^{n+1} = \sum_{i=1}^M \bar{y}_{i}^{n+1} \varphi_i
$$

satisfies

$$
(c(EV^n) \frac{V^{n+1} - V^n}{\Delta t}, \chi) + \left( \frac{1}{2} a(EV^n) \nabla (V^{n+1} + V^n) + b(EV^n), \nabla \chi \right)
$$

$$
= (f(t^{n+1/2}, EV^n), \chi) + (g(t^{n+1/2}), \chi), \quad \chi \in \mathcal{M}_h.
$$

The following norms play an important role in our analysis:

a) $\|\varphi\|_{2^\nu} = (c(EV^n)\varphi, \varphi)$,

b) $\|\varphi\|_{2^\nu} = (\frac{1}{2} a(EV^n)\nabla \varphi, \nabla \varphi)$.

By (2.4), $\|\cdot\|_{c^\nu}$ is equivalent to $\|\cdot\|$ and $\|\cdot\|_{a^\nu}$ is equivalent to $\|\nabla \cdot\|$ for each $n$. Thus $\|\cdot\|_{2^\nu} + \|\cdot\|_{a^\nu}$ is equivalent to $\|\cdot\|_{2^\nu}$ for each $n$.

For future reference, we note that (3.10), the choice of $\nu$ given by (3.12) with $\alpha = 1$, and the triangle inequality yield the inequalities

$$
\| V^{2} - V^2 \|_{c^\nu} + (\Delta t)^{1/2} \| V^2 - V^2 \|_{a^\nu} \leq C_1 \Delta t \| \delta^2 V^1 \|_{c^\nu} + (\Delta t)^{1/2} \| \delta^2 V^1 \|_{a^\nu},
$$

$$
\| V^{n+1} - V^{n+1} \|_{c^\nu} + (\Delta t)^{1/2} \| V^{n+1} - V^{n+1} \|_{a^\nu} \leq C_1 \Delta t \| \delta^3 V^n \|_{c^\nu} + (\Delta t)^{1/2} \| \delta^3 V^n \|_{a^\nu}, \quad n \geq 2,
$$

where

a) $\delta V^n = V^{n+1} - V^n$,

b) $\delta^2 V^n = V^{n+1} - 2V^n + V^{n-1}$,

c) $\delta^3 V^n = V^{n+1} - 3V^n + 3V^{n-1} - V^{n-2}$.

The convergence results of § 4 depend only on the norm reduction (3.10) which yields analogues of (3.17) for various choices of $\rho$ and not on the particular iterative method used to achieve those norm reductions.
We shall now define a starting procedure which also uses the preconditioned conjugate gradient iteration. For $V^0$ we shall interpolate $u_0$ into $\mathcal{M}_h$ to obtain $Iu_0$. Then for some constant $K_6$ which depends upon $\|u_0\|_r$, we have

$$\|V^0 - W^0\| \leq \|Iu_0 - u_0\| + \|u_0 - W^0\| \leq K_6 h'.$$

We note here that for Corollary 1 and Corollary 3 of §4, in order to obtain the optimal order convergence through the predictor-corrector-corrector method to find $V^1$ to be described below, we must have the estimate

$$\|V^0 - W^0\|_1 \leq K_6 h^r + (\Delta t)^2.$$

The method of computing $V^0$ to satisfy (3.20) would necessarily be more complicated than that described above. For example, if one is willing to factor one additional matrix, a $V^0$ satisfying (3.20) could be obtained by solving the equations generated by equation (2.8) at the initial time.

For $V^1$, we shall obtain an approximate solution of the following predictor-corrector-corrector Crank-Nicolson-Galerkin method by using the same preconditioned conjugate gradient iterative method as before. We first describe the exact equations for the Crank-Nicolson-Galerkin method. Let $U^*$, the prediction for $U^1$, be the unique solution of

$$\left(c(U^*) - U^0\right) \frac{U^* - U^0}{\Delta t} + \left(\frac{1}{2} a(U^0) V(U^* + U^0) + b(U^0), \nabla \chi\right)$$

$$= \langle g(t^{n+1/2}), \chi \rangle + \langle f(t^{n+1/2}, U^0), \chi \rangle, \quad \chi \in \mathcal{M}_h. \tag{3.21}$$

With $U^{1/2} = (U^* + U^0)/2$, let $U^{**}$ be the unique solution of

$$\left(c(U^{1/2}) - U^0\right) \frac{U^{**} - U^0}{\Delta t} + \left(\frac{1}{2} a(U^{1/2}) V(U^{**} + U^0) + b(U^{1/2}), \nabla \chi\right)$$

$$= \langle g(t^{n+1/2}), \chi \rangle + \langle f(t^{n+1/2}, U^{1/2}), \chi \rangle, \quad \chi \in \mathcal{M}_h. \tag{3.22}$$

With $U^{*1/2} = (U^{**} + U^0)/2$, let $U^1$ be the unique solution of

$$\left(c(U^{*1/2}) - U^0\right) \frac{U^1 - U^0}{\Delta t} + \left(\frac{1}{2} a(U^{*1/2}) V(U^1 + U^0) + b(U^{*1/2}), \nabla \chi\right)$$

$$= \langle g(t^{n+1/2}), \chi \rangle + \langle f(t^{n+1/2}, U^{*1/2}), \chi \rangle, \quad \chi \in \mathcal{M}_h. \tag{3.23}$$

Instead of solving (3.21)-(3.23) exactly, we shall approximate their solution by employing a preconditioned conjugate gradient method using the same preconditioning matrix $L_0$ defined in (3.5). If we iterate $\nu$ times, where $\nu$ is given by (3.12) with $\alpha = 1$ for the approximate solutions of (3.21)-(3.23), we obtain optimal order $H^1$ bounds. To obtain optimal order $L^2$ bounds, we choose $\nu$ as in (3.12) with $\alpha = 3/2$ for the approximations of the correctors (3.22) and (3.23). The following lemma can be proved using the arguments of §4.

**Lemma 3.** Assume $S, R, (2.1), (2.2), and (2.10)$ to hold. Let $\nu$ be given by (3.12) for $\alpha = 1$. Then, there exist positive constants $C_2$ and $\tau$ such that, if $\Delta t \leq \tau$,

$$\|V^0 - W^0\| \leq C_2 h', \tag{3.24}$$

$$\|V^1 - W^1\|_1 + (\Delta t)^{1/2} |d_1(V^0 - W^0)| \leq C_2 ((\Delta t)^2 + h^{-1}),$$

where $C_2$ depends on $a_\varepsilon, a^*, c_\varepsilon, c^*, K_0, \ldots, K_5$, and $\|u_0\|_r$. 

Downloaded 11/10/15 to 165.91.112.146. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php
4. **A priori error estimates.** In this section we develop *a priori* bounds for the error $V^n - u^n$ for the procedures defined in § 3. The first result, Theorem 1, states that optimal order $H^1$ estimates can be obtained for $r \geq 3$ (piecewise quadratics or better) if the iterative process reduces the error in the solution of the algebraic problem by a factor proportional to $\Delta t (\rho = O(\Delta t))$ at each time step. The second result, Corollary 1, points out that with additional smoothness restrictions on $\partial u/\partial t$, the choice $\rho = O(\Delta t)$ is also sufficient to obtain optimal order $L^2$ estimates for $r \geq 2$. If $\Delta t = O(h^{r/2})$, then the optimal order rate also applies in $H^1$ for $r \geq 2$. The third result, Corollary 2, says that, if $r \geq 3$ and $\Delta t \leq C h^2$, then, to obtain optimal order $H^1$ bounds, the error at each time step need only be reduced by a fixed (sufficiently small) factor that is independent of $\Delta t$ and $h$. The next result, Corollary 3, extends this result to optimal order $L^2$ bounds for $r \geq 2$ with additional smoothness on $\partial u/\partial t$. Theorem 2 shows that for $\rho = O((\Delta t)^{1/2})$, we can weaken the regularity assumptions used on $u$ in Corollary 3 and still achieve optimal order $L^2$ bounds for $r \geq 3$.

**Theorem 1.** Let $S$ and $R$ and the restrictions on $\{M_n\}$ of § 2 hold. Let $V^n$ satisfy (3.17) and (3.24). Then there exist constants $\tau$ and $C_3$, where $C_3$ is dependent on the constants in $R$, $K_0$, $K_2$, $K_4$, $K_5$ and $C_2$, such that, if $r \geq 3$, $\Delta t \leq r$, and $\Delta t \leq h^{r/3}$,

\[
\sup_{t^n} \|u - V\| \leq C_3((\Delta t)^2 + h^{r-1}).
\]

**Proof.** Letting $\xi^n = V^n - W^n$, we see that

\[
(cEV^n,d\xi^n,\chi) + (aEV^n)\nabla \xi^{n+1/2}, \nabla \chi
\]

\[
= -((cEV^n - c(u(t^{n+1/2})))dW^n,\chi) - (c(u(t^{n+1/2})))\left[dW^n - \frac{\partial u}{\partial t}\right],\chi
\]

\[
+ ([a(u(t^{n+1/2}))W(t^{n+1/2}) - a(EV^n)W^{n+1/2}],\nabla \chi)
\]

\[
- (\eta(t^{n+1/2}),\chi) + ([b(u(t^{n+1/2})) - b(EV^n)],\nabla \chi)
\]

\[
+ ([f(t^{n+1/2},EV^n) - f(t^{n+1/2},u(t^{n+1/2})),\chi)
\]

\[
+ \left(cEV^n \frac{V^{n+1} - \tilde{V}^{n+1}}{\Delta t},\chi\right) + \frac{1}{2}(a(EV^n)\nabla(V^{n+1} - \tilde{V}^{n+1}),\nabla \chi), \chi \in \mathcal{M}_h.
\]

We shall obtain estimates in the $H^1$-norm by using $\chi = \xi^{n+1} - \xi^n = \Delta t d\xi^n$ as a test function. Clearly

\[
(cEV^n)d\xi^n,\xi^{n+1} - \xi^n + (a(EV^n)\nabla \xi^{n+1/2},(\xi^{n+1} - \xi^n))
\]

\[
= \Delta t||d\xi^n||^2 + ||\xi^{n+1}||^2 - ||\xi^n||^2.
\]

We use the assumptions in $R$ of § 2 to obtain estimates for the first two terms on the right hand side of (4.2). Thus,

\[
\Delta t\left|(cEV^n - c(Eu^n) + c(Eu^n) - c(u(t^{n+1/2})))\right|dW^n, \xi^n
\]

\[
+ \Delta t\left|(c(u(t^{n+1/2})))\left[d\eta^n + d\eta^n - \frac{\partial u}{\partial t}(t^{n+1/2})\right], \xi^n\right|
\]

\[
\leq C_4 \Delta t(||\xi^n||^2 + ||\xi^n||^2 + ||\eta^n||^2 + ||\eta^{n-1}||^2 + ||d\eta^n||^2)
\]

\[
+ \epsilon_5 \Delta t||d\xi^n||^2 + C(\Delta t)^{4}\sigma_n.
\]
where

\[ \sigma_n = \int_{s_{n-1}}^{s_{n+1}} \left( \left\| \frac{\partial^3 u}{\partial t^3} (\cdot, s) \right\| + \left\| \frac{\partial^2 u}{\partial t^2} (\cdot, s) \right\| \right) \, ds = \sigma_{1,n} + \sigma_{2,n}. \]

From (2.7), \( \sum_{n=1}^{N-1} \sigma_n < \infty \). We note that \( C_4 \) depends on \( c^*, K_1 \) and the bound \( K_4 \) for \( \| W/\partial t \|_{L^\infty(\Omega, L^\infty)} \). The fourth and sixth terms on the right hand side of (4.2) can be treated similarly. We shall now estimate the last two terms on the right of (4.2) using (3.17). We first note that from (2.13), we have

\begin{align*}
&\text{a)} \quad \| \delta^2 W \|_1 \lesssim C(\Delta t)^2, \\
&\text{b)} \quad \| \delta^3 W \|_1 \lesssim C(\Delta t)^2 \int_{s_{n-2}}^{s_{n+1}} \left\| \frac{\partial^3 W}{\partial t^3} (\cdot, s) \right\|_1 \, ds.
\end{align*}

The constants appearing here will then sum in the proper fashion to achieve the desired results of the theorem. Since different starting procedures were used in the conjugate gradient iteration to obtain \( V^2 \) and \( V^m \) for \( m \geq 3 \), we shall estimate each case separately. From (3.17) we see that for \( n = 1 \),

\[ \left( (c(EV)^1 \frac{V^2 - \bar{V}^2}{\Delta t}, \xi - \bar{\xi}) + \frac{1}{2} (a(EV)^1 \nabla(V^2 - \bar{V}^2), \nabla(\xi^2 - \bar{\xi}^2)) \right) \]

\[ \lesssim \| V^2 - \bar{V}^2 \|_{c_1} \| d_2^1 \|_{c_1} + \| V^2 - \bar{V}^2 \|_{a_1} \| \xi^2 - \bar{\xi}^2 \|_{a_1} \]

\[ \lesssim C(\Delta t)^{1/2} \| \delta^2 V \|_{c_1} + (\Delta t)^{1/2} \| \delta^2 V \|_{a_1} \| \xi^2 \|_{a_1} + \| \xi \|_{a_1} \]

\[ \lesssim C(\Delta t)^{1/2} \| \delta^2 V \|_{c_1} + (\Delta t)^{1/2} \| \delta^2 V \|_{a_1} \| \xi^2 \|_{a_1} + \| \xi \|_{a_1} \]

\[ \lesssim \varepsilon_5 \Delta t \| \delta^2 V \|_{c_1} + \| \delta^2 V \|_{c_1} \nabla(\xi^2 + \| \xi \|_{a_1} + \| \xi \|_{a_1} + \| \xi \|_{a_1}) \]

From (3.17) we obtain the estimate for \( n \geq 2 \) in a similar manner. For \( n \geq 2 \),

\[ \left( (c(EV)^n \frac{V^{n+1} - \bar{V}^{n+1}}{\Delta t}, \xi^{n+1} - \bar{\xi}^n) + \frac{1}{2} (a(EV)^n \nabla(V^{n+1} - \bar{V}^{n+1}), \nabla(\xi^{n+1} - \bar{\xi}^n)) \right) \]

\[ \lesssim \varepsilon_5 \Delta t \| \delta^2 V \|_{c_1} + \| \delta^2 V \|_{c_1} \nabla(\xi^{n+1} + \| \xi \|_{a_1} + \| \xi \|_{a_1} + \| \xi \|_{a_1}) + \| \delta^2 V \|_{c_1} \nabla(\xi^{n+1} + \| \xi \|_{a_1} + \| \xi \|_{a_1}) \]

\[ + C_5 \Delta t \| \xi^{n+1} \|_{a_1} + \| \xi \|_{a_1} + \| \xi \|_{a_1} + \| \xi \|_{a_1} \]

where

\[ \text{a)} \quad \sigma_{3,n} = \int_{s_{n-2}}^{s_{n+1}} \left\| \frac{\partial^3 W}{\partial t^3} (\cdot, s) \right\|_1 \, ds, \\
\text{b)} \quad \sigma_{4,n} = \int_{s_{n-2}}^{s_{n+1}} \left\| \frac{\partial^3 W}{\partial t^3} (\cdot, s) \right\|_2 \, ds.
\]

The third term on the right side of (4.2) can be split in the following manner:

\[ ([a(u(t^{n+1/2})) \nabla W(t^{n+1/2}) - a(EV^n) \nabla W^{n+1/2}], \nabla \chi) = ([a(u(t^{n+1/2})) \nabla W(t^{n+1/2}) - \nabla W^{n+1/2}], \nabla \chi) \]

\[ + (\nabla W^{n+1/2}[a(u(t^{n+1/2})) - a(EV^n)]) \nabla \chi) = A_{1,n} + A_{2,n}. \]
We shall use summation by parts in time to treat $A_{1,n}$. We see that

$$\left| \sum_{n=1}^{l-1} A_{1,n} \right| \leq \left| \sum_{n=2}^{l-1} \left( (a(t^{n+1/2}) - a(t^{n-1/2})) \nabla [W(t^{n+1/2}) - W^{n+1/2}] \right. \right.$$ 

$$\left. - a(t^{n-1/2}) \nabla [W(t^{n-1/2}) - W^{n-1/2}], \nabla \zeta^n \right|$$ 

$$+ \left| (a(u(t^{3/2})) \nabla [W(t^{3/2}) - W^{3/2}], \nabla \zeta^1) \right|$$ 

$$+ \left| (a(u(t^{l-1/2})) \nabla [W(t^{l-1/2}) - W^{l-1/2}], \nabla \zeta^l) \right|$$ 

$$\leq \left| \sum_{n=2}^{l-1} \Delta t \left( \frac{a(u(t^{n+1/2})) - a(u(t^{n-1/2}))}{\Delta t} \right) \nabla [W(t^{n+1/2}) - W^{n+1/2}], \nabla \zeta^n \right|$$ 

$$+ \left| \sum_{n=2}^{l-1} \Delta t \left( a(u(t^{n-1/2})) \nabla \left[ W(t^{n+1/2}) - W^{n+1/2} \right. \right. \right.$$ 

$$\left. \left. - W(t^{n-1/2}) - W^{n-1/2} \right], \nabla \zeta^n \right) \right|$$ 

$$+ C \left\{ \| \zeta^1 \|^2 + (\Delta t)^4 \right\} + \frac{1}{20} \| \zeta^l \|^2_{a, l-1}.$$ 

We note that $C_\delta$ depends on $a^*$, $K_1$, $K_4$, and $K_5$. We next sum by parts to estimate $A_{2,n}$ from (4.10).

$$\left| \sum_{n=1}^{l-1} A_{2,n} \right| \leq \left| \sum_{n=2}^{l-1} \left( (a(u(t^{n+1/2})) - a(EV^n)) \nabla W^{n+1/2} \right. \right.$$ 

$$\left. - (a(u(t^{n-1/2})) - a(EV^{n-1})) \nabla W^{n-1/2}], \nabla \zeta^n \right|$$ 

$$+ \left| (a(u(t^{3/2})) - a(EV^1)) \nabla W^{3/2}], \nabla \zeta^1 \right|$$ 

$$+ \left| (a(u(t^{l-1/2})) - a(EV^l)) \nabla W^{l-1/2}], \nabla \zeta^l \right|$$ 

$$\leq C \gamma \left\{ \| \zeta^1 \|^2 + \| \zeta^l \|^2 + \| \zeta^2 \|^2 + \| \eta^1 \|^2 + \| \eta^2 \|^2 + \| \eta^3 \|^2 + \| \eta^4 \|^2 + \| \eta^5 \|^2 + \| \eta^6 \|^2 + \| \eta^7 \|^2 + \| \eta^8 \|^2 \right\}$$ 

$$+ \frac{1}{20} \| \zeta^l \|^2_{a, l-1} + \left| \sum_{n=2}^{l-1} \left( a(u(t^{n+1/2})) - a(EV^n) \nabla W^{n+1/2} - \nabla W^{n-1/2}], \nabla \zeta^n \right) \right|$$ 

$$+ \left| \sum_{n=2}^{l-1} \nabla W^{n-1/2} [a(u(t^{n+1/2})) - a(EV^n) \nabla W^{n+1/2} - \nabla W^{n-1/2}], \nabla \zeta^n \right|$$ 

We then bound the next to the last term on the right of (4.12) as follows:

$$\left| \sum_{n=2}^{l-1} \left( a(u(t^{n+1/2})) - a(EV^n) \nabla W^{n+1/2} - \nabla W^{n-1/2}], \nabla \zeta^n \right) \right|$$ 

$$\leq C \sum_{n=1}^{l-1} \left\{ \| \zeta^n \|^2 + \| \eta^n \|^2 + (\Delta t)^4 \right\} \sigma_{5,n},$$ 

where

$$\sigma_{5,n} = \int_{t_{n-1}}^{t_n} \left\| \nabla \frac{\partial W}{\partial t} (\cdot, s) \right\|_{L^\infty} ds.$$
Note that $\sum_{n=1}^{l-1} \sigma_{5,n} \equiv 2K_4$ by (2.10). Next define

\begin{align}
\text{a) } a'_{1,n}(x) &= \int_0^1 \frac{\partial a}{\partial u}(x, \theta u(t^{n+1/2}) + (1 - \theta)u(t^{n-1/2})) \, d\theta, \\
\text{b) } a'_{2,n}(x) &= \int_0^1 \frac{\partial a}{\partial u}(x, \theta EU^n + (1 - \theta)EV^{n-1}) \, d\theta.
\end{align}

(4.14)

We can now treat the last term on the right of (4.12) in the following manner:

\begin{align}
&\left| \sum_{n=2}^{l-1} \left( \nabla W^{n-1/2} \{ [a'u(t^{n+1/2}) - a'(u(t^{n-1/2})) - [a(EV^n) - a(EV^{n-1})]], \nabla \xi^n \right) \right| \\
&= \left| \sum_{n=2}^{l-1} \left( \nabla W^{n-1/2} \{ [a'_{1,n}(u(t^{n+1/2}) - u(t^{n-1/2})) - a'_{2,n}(EU + EV^{n-1})]], \nabla \xi^n \right) \right| \\
&\leq \sum_{n=2}^{l-1} \left( \left| \nabla W^{n-1/2} \{ a'_{1,n}-a'_{2,n} \right| \frac{u(t^{n+1/2}) - u(t^{n-1/2})}{\Delta t} \right) \Delta t + \nabla W^{n-1/2} \\
&\leq C \Delta t \sum_{n=0}^{l-1} \left( \|\xi^{n+1}\|_t^2 + \|\eta^n\|_t^2 + \|d_t \eta^n\|_t^2 \right) + \epsilon \Delta t \sum_{n=1}^{l-1} \|d_t \xi^n\|_t^2 \\
&+ \frac{\Delta t}{4} \sum_{n=1}^{l-1} \|d_t \xi^n\|_t^2 + \sum_{n=1}^{l-1} \left( \|\xi^{n+1}\|_t^2 - \|\xi^n\|_t^2 \right) \\
&\leq \frac{3}{4} \|\xi^{l-1}\|_t^2 + C_\gamma \|\xi^{l-1}\|_t^2 + \|d_t \xi^n\|_t^2 + \Delta t \|d_t \xi^n\|_t^2 \right) + C_{\gamma} \|\xi^{l-1}\|_t^2 + \|\xi^{l-2}\|_t^2 \right) \\
&+ C \sum_{n=1}^{l-1} \left( \Delta t + \sigma_{3,n} + \sigma_{5,n} \right) \|\xi^n\|_1^2 \\
&+ C \sum_{n=1}^{l-1} \left( \Delta t + \sigma_{3,n} + \sigma_{5,n} \right) \|\xi^n\|_1^2.
\end{align}

(4.16)

Note that

\begin{align}
\|\xi^{n+1}\|_t^2 - \|\xi^n\|_t^2 &\leq 2 \Delta t (d_t \xi^n, \xi^n) + (\Delta t)^2 \|d_t \xi^n\|_t^2 \leq \epsilon \Delta t \|d_t \xi^n\|_t^2 + C \Delta t \|\xi^n\|_t^2.
\end{align}

(4.17)

Sum this inequality from $n = 1$ to the upper limits $l - 3, l - 2,$ and $l - 1$; then multiply the resulting inequalities by $C_{\gamma} + \frac{1}{2}$ and add them to (4.16), after choosing $\epsilon$ so that

\begin{align}
3\epsilon \left( C_\gamma + \frac{1}{2} \right) \sum_{n=1}^{l-1} \|d_t \xi^n\|_t^2 \leq \frac{1}{4} \sum_{n=1}^{l-1} \|d_t \xi^n\|_t^2.
\end{align}

(4.18)

Thus,

\begin{align}
\frac{\Delta t}{2} \sum_{n=1}^{l-1} \|d_t \xi^n\|_t^2 + \left( C_\gamma + \frac{1}{2} \right) \|\xi^{l-1}\|_t^2 + \|\xi^{l-1}\|_t^2 + \|\xi^{l-2}\|_t^2 - 3 \|\xi^{l-1}\|_t^2 \right) \\
+ \sum_{n=1}^{l-1} \left( \|\xi^{n+1}\|_t^2 - \|\xi^n\|_t^2 \right) \leq \frac{1}{2} \|\xi^{l-1}\|_t^2 + C_\gamma \|\xi^{l-1}\|_t^2 + \|\xi^{l-2}\|_t^2 \right) \\
+ C \sum_{n=1}^{l-1} \left( \Delta t + \sigma_{3,n} + \sigma_{5,n} \right) \|\xi^n\|_1^2 \\
+ C \sum_{n=1}^{l-1} \left( \Delta t + \sigma_{3,n} + \sigma_{5,n} \right) \|\xi^n\|_1^2.
\end{align}

(4.19)
We shall now use ideas of Douglas [5] and Rachford [14] to establish comparability between \( \| \zeta^n \|_{a}^2 \) and \( \| \zeta^n \|_{a}^2 + \| \zeta^n \|_{a}^2 \) to obtain telescoping sums on the left side of (4.19). Note that

\[
\| \zeta^n \|_{a}^2 + \left( \frac{\partial a}{\partial u} E(\zeta^n - \zeta^{n-1}) + \frac{\partial a}{\partial u} E(W^n - W^{n-1}) \right) \nabla \zeta^n, \nabla \zeta^n
\]

\( \leq \| \zeta^n \|_{a}^2 + C \left( \| \delta \zeta^n \|_{L^\infty} + \| \delta \zeta^{n-2} \|_{L^\infty} + \Delta t \| \zeta^n \|_{t}^2. \right) \)

Thus, as in [13],

\[
\sum_{n=2}^{l-1} (\| \zeta^{n+1} \|_{a}^2 - \| \zeta^n \|_{a}^2) \leq C \sum_{n=2}^{l-1} (\| \delta \zeta^{n-1} \|_{L^\infty} + \| \delta \zeta^{n-2} \|_{L^\infty} + \Delta t \| \zeta^n \|_{t}^2. \right)
\]

Next, if \( u \in L^\infty(J; H^r) \) and \( (\partial u/\partial t) \in L^2(J; H^{r-1}) \), then (2.9), (3.24), (4.19), and (4.21) imply that

\[
\| \zeta^l \|_{a}^2 + \| \zeta^{l-1} \|_{a}^2 + \Delta t \sum_{n=1}^{l-1} \| d \zeta^n \|_{t}^2 \leq C_1 \sum_{n=1}^{l-1} (\Delta t + \sigma_3, n + \sigma_5, n) \| \zeta^n \|_{t}^2 + C_1 \sum_{n=2}^{l-1} (\| \delta \zeta^{n-1} \|_{L^\infty} + \| \delta \zeta^{n-2} \|_{L^\infty} + \Delta t \| \zeta^n \|_{t}^2. \right) + C_1 (h^{2r-2} + (\Delta t)^2). \)

We note, for example, that \( C_1 \) depends upon \( a^*, c^*, K_1, K_3, \) and bounds on

\[
\| W_t \|_{L^\infty(J; L^\infty)}, \| \nabla W \|_{L^\infty(J; L^\infty)}, \| \partial^2 u \|_{L^\infty(J; L^2)}, \| \partial^2 u \|_{L^2(J; L^\infty)}, \| \partial^3 u \|_{L^2(J; L^\infty)}, \| \partial^3 u \|_{L^2(J; H^r)}, \| u \|_{L^\infty(J; H^r)}, \text{ and } \| \partial u \|_{L^2(J; H^{r-1})}. \]

In order to apply the discrete Gronwall lemma to (4.22), we wish to show that there exists \( C_{13} > 0 \) such that

\[
\sum_{n=0}^{l-2} \| \delta \zeta^n \|_{L^\infty} < C_{13}. \]

The predictor-corrector-corrector starting method yields

\[
\| \delta \zeta^n \|_{L^\infty} < C_{14}. \]

We shall use an induction argument as in [14] to yield (4.23) with the summation starting at \( n = 1 \). For \( l = 2 \), the inequality (4.22) and the estimate (3.24) imply that

\[
\Delta t \| d \zeta^1 \|_{t}^2 \leq C_9 \Delta t \| \zeta^1 \|_{t}^2 + C_{10} (K_4 + K_5) \| \zeta^1 \|_{t}^2 + C_{12} (h^{2r-2} + (\Delta t)^2) \]

\[
\leq C_{15} (\Delta t)^4 + h^{2r-2}. \]
If
\[ a) \ h \leq (2C_{15})^{-3/d}, \]
\[ b) \ \Delta t \leq h^{d/3}, \]
\[ c) \ r \geq 3 \geq \frac{3}{2}d + 1, \]
then
\[ (4.27) \ \Delta t \|d\xi^{n}\|^2 \leq h^d. \]
Assume the following induction hypothesis:
\[ (4.28) \ \Delta t \sum_{n=1}^{k} \|d\xi^{n}\|^2 \leq h^d \quad \text{for } 1 \leq k \leq l - 2. \]
If we use the inverse hypothesis (2.2c) satisfied by \( M_{\alpha} \), (4.28) and the fact that \( N \Delta t = T \), we see that
\[ (4.29) \ \sum_{n=1}^{l-2} \|\delta\xi^{n}\|_{L^{\infty}} \leq (l-2)^{1/2}\left( \sum_{n=1}^{l-2} \|\delta\xi^{n}\|^{2}_{L^{\infty}} \right)^{1/2} \]
\[ \leq N^{1/2} K_0 h^{-d/2}\left( \sum_{n=1}^{l-2} \|d\xi^{n}\|^{2} \right)^{1/2} \]
\[ \leq (\Delta t)^{-1/2} T^{1/2} K_0 h^{-d/2}(\Delta t)^{1/2}\left( \Delta t \sum_{n=1}^{l-2} \|d\xi^{n}\|^{2} \right)^{1/2} \]
\[ \leq T^{1/2} K_0. \]
Then, with \( C_{13} = T^{1/2} K_0 \), we apply the discrete Gronwall lemma in (4.22) to obtain
\[ (4.30) \ \|\xi^{n}\|^{2} + \Delta t \sum_{n=1}^{l-1} \|d\xi^{n}\|^{2} \leq C_{16}\left\{ (\Delta t)^4 + h^{2r-2} \right\} \]
where
\[ (4.31) \ C_{16} \leq C_{12} \exp \left\{ C_9 T + C_{10}\left[ K_4 + K_5 \right] + 2C_{11} T^{1/2} K_0 \right\}. \]
Then with (4.26a) replaced by \( h \leq (2C_{16})^{-3/d} \) we see that our induction argument is completed. Since (4.30) holds for each \( l \) from \( l = 1 \) to \( l = N \), we have
\[ (4.32) \ \sup_{r^{\ast}} \|V - W\|_{L^{\infty}} \leq C_{17}\{ (\Delta t)^2 + h^{r-1} \}. \]
Then for \( u \in L^{\infty}(J; H'), \) (2.9) and the triangle inequality yield the desired result (4.1).

Note that if \( \partial u/\partial t \in L^2(J; H') \) and if \( V^0 \) and \( V^1 \) are determined such that
\[ (4.33) \ \|V^0 - W^0\|_{L^1} + \|V^1 - W^1\|_{L^1} + (\Delta t)^{1/2}\|d_i(V^0 - W^0)\| \leq C_2\{ (\Delta t)^2 + h^r \}, \]
then by using (2.9), \( h^{2r-2} \) can be replaced by \( h^{2r} \) in the above proof. In this case it suffices to assume that \( r \geq 2 \geq \frac{3}{2}d. \)

**Corollary 1.** Let all the hypotheses of Theorem 1 except the assumption on \( r \) be satisfied. Assume that \( r \geq 2 \), (4.33) is satisfied, and \( (\partial u/\partial t) \in L^2(J; H') \). Then there exist constants \( \tau \) and \( C_{18} \) such that if \( \Delta t \leq \tau \) and \( \Delta t \leq h^{d/3} \), then
\[ \sup_{r^{\ast}} \{ \|u - V\| + h\|u - V\|_{L^1} \} \leq C_{18}\{ (\Delta t)^2 + h^r \}. \]
The constant $C_{18}$ has the same dependencies as $C_3$ with the addition of a bound on $\|\partial u/\partial t\|_{L^\infty(I;W^1)}$.

Next, we show that a further restriction on $\Delta t$ can reduce the number of iterations for $n \geq 2$ from (3.12) to a number independent of $\Delta t$. Assume that
\begin{equation}
\Delta t \leq C_{19}h^2.
\end{equation}

We note that (4.34) is really no restriction if $\mathcal{M}_h$ is a space of piecewise cubic (or higher order) polynomials ($r \geq 4$) since the optimal choice of $\Delta t$ is $O(h^{r/2})$ for such choices of $\mathcal{M}_h$. If the quadratic extrapolation given by (3.6b) is used to initialize the iteration for $n \geq 2$, then (2.2a), (3.10), (3.17), and (3.18) imply that
\begin{equation}
\frac{1}{\Delta t} \left( \frac{c(EV^n)}{\Delta t} \left( \frac{V^{n+1} - \bar{V}^{n+1}}{\Delta t} \right) - \frac{\zeta^{n+1} - \zeta^n}{\Delta t} \right) + \frac{1}{2} \left( a(EV^n) \nabla (V^{n+1} - \bar{V}^{n+1}), \nabla (\zeta^{n+1} - \zeta^n) \right)
\end{equation}
\begin{equation}
\leq \| V^{n+1} - \bar{V}^{n+1} \|_{c,n} \| d_\zeta^n \|_{c,n} + \| V^{n+1} - \bar{V}^{n+1} \|_{a,n} \| d_\zeta^n \|_{a,n} \Delta t
\end{equation}
\begin{equation}
\leq C_{1\rho} \left( \| \delta^3 V^n \|_{c,n} + (\Delta t)^{1/2} \| \delta^3 V^n \|_{a,n} \right) \left( \| d_\zeta^n \|_{c,n} + (\Delta t)^{1/2} \| d_\zeta^n \|_{a,n} \right)
\end{equation}
\begin{equation}
\leq C_{1\rho} \left( 1 + \frac{a*K_0(\Delta t)^{1/2}}{c_*} \right)^2 \{ \| \delta^2 d_\zeta^n \|_{c,n} + \| \delta^3 W^n \|_{c,n} \} \| d_\zeta^n \|_{c,n}
\end{equation}
\begin{equation}
\leq C_{1\rho} \left( 1 + \frac{a*K_0 C_{19}}{c_*} \right)^2 \left( \| \delta^2 d_\zeta^n \|_{c,n} + C \Delta t \sigma_{6,n} \| d_\zeta^n \|_{c,n} \right)
\end{equation}
\begin{equation}
\leq \Delta t C_{1\rho} \left( 1 + \frac{a*K_0 C_{19}}{c_*} \right)^2 \left( \| d_\zeta^n \|_{c,n} + 2 \| d_\zeta^{n-1} \|_{c,n} + 3 \| d_\zeta^{n-2} \|_{c,n} \right)
\end{equation}
\begin{equation}
+ C(\Delta t)^2 \sigma_{6,n} \| d_\zeta^n \|_{c,n},
\end{equation}
where $\sigma_{6,n} = \int_{n \to n+1} \| \delta^3 W/\delta t^3 (\cdot, s) \| ds$. Iterate sufficiently many times such that
\begin{equation}
\rho < \left[ 8 C_{1\rho} \left( 1 + \frac{a*K_0 C_{19}}{c_*} \right) \right]^{-1} \equiv K_8,
\end{equation}
where
\begin{equation}
\rho \equiv \rho_1 = 2Q^*.
\end{equation}

Then for some $\epsilon_7 > 0$,
\begin{equation}
\left( c(EV^n) \frac{V^{n+1} - \bar{V}^{n+1}}{\Delta t}, \zeta^{n+1} - \zeta^n \right) + \frac{1}{2} \left( a(EV^n) \nabla (V^{n+1} - \bar{V}^{n+1}), \nabla (\zeta^{n+1} - \zeta^n) \right)
\end{equation}
\begin{equation}
\leq C(\Delta t)^4 \sigma_{4,n} + \left( \epsilon_7 + \frac{9}{32} \right) \Delta t \| d_\zeta^n \|_{c,n}^2 + \frac{\Delta t}{8} \left( \| d_\zeta^{n-1} \|_{c,n}^2 + \| d_\zeta^{n-2} \|_{c,n}^2 \right).
\end{equation}
Inequality (4.38) will replace inequality (4.8). Note that Lemma 2 yields $\sum_{n=1}^{N-1} \sigma_{4,n} \leq 3K_5$. We recall from (3.10) that the general form of (3.17) is
\begin{equation}
\| \bar{V}^{n+1} - V^{n+1} \|_{c,n} + (\Delta t)^{1/2} \| \bar{V}^{n+1} - V^{n+1} \|_{a,n}
\end{equation}
\begin{equation}
\leq C_{1\rho} \left( \| \delta^3 V^n \|_{c,n} + (\Delta t)^{1/2} \| \delta^3 V^n \|_{a,n} \right), \quad n \geq 2.
\end{equation}
We have thus shown the following corollary.

Corollary 2. Let $S$ and $R$ and the restrictions on $\{\mathcal{M}_h\}$ of § 2 hold. Let $V^0$ and $V^1$ satisfy (3.24) and $V^2$ satisfy (3.17a). For $n \geq 2$, assume that (4.39) is satisfied with $\rho$
satisfying (4.36). Let \( r \geq 3 \) and \( \Delta t \leq C_1 h^2 \). Then, there exist constants \( \tau \) and \( C_{20} \) such that if \( \Delta t \leq \tau \) and \( \Delta t \leq h^{d/3} \),

\[
\sup_{t^n} \| u - V \|_1 \leq C_{20}((\Delta t)^2 + h^{-1} + 1).
\]

\( C_{20} \) has the same dependencies as \( C_3 \).

**Corollary 3.** Let all the hypotheses of Corollary 2 be satisfied except the assumption on \( r \). Assume that \( r \geq 2 \) and that (4.33) is satisfied. If \( \partial u / \partial t \in L^2(J; H^r) \), then there exist constants \( \tau \) and \( C_{21} \) such that, if \( \Delta t \leq C_9 h^2 \), \( \Delta t \leq \tau \) and \( \Delta t \leq h^{d/3} \),

\[
\sup_{t^n} \{ \| u - V \| + h \| u - V \|_1 \} \leq C_{21}((\Delta t)^2 + h^r).
\]

We shall now use Corollary 2 to obtain an optimal order \( L^2 \)-estimate with the smoothness assumptions on \( \partial u / \partial t \) of Theorem 1 if \( \Delta t \leq C_1 h^2 \). Let

\[
\| \varphi \|_1 = \sup \left\{ \int \varphi \phi \, dx : \| \phi \|_1 = 1 \right\}.
\]

**Theorem 2.** Let \( S \) and \( R \) and the restrictions on \( M_t \) of § 2 hold. Assume that the Neumann problem for \(-\Delta + I\) on \( \Omega \) is \( H^3 \)-regular. Let \( r \geq 3 \) and \( \Delta t \leq C_9 h^2 \). Let \( V^0 \) and \( V^1 \) satisfy (3.24), \( V^2 \) satisfy (3.17a), and \( V^{n+1} \) for \( n \geq 2 \) satisfy (3.17b) with \( \Delta t \) replaced by \( (\Delta t)^{1/2} \). Then there exist constants \( \tau \) and \( C_{22} \) such that, if \( \Delta t \leq \tau \) and \( \Delta t \leq h^{d/3} \),

\[
\sup_{t^n} \{ \| u - V \| + h \| u - V \|_1 \} \leq C_{22}((\Delta t)^2 + h^r).
\]

\( C_{22} \) has the same dependencies as \( C_{20} \) or \( C_3 \).

**Proof.** To obtain an \( L^2 \)-estimate, we use the test function \( \chi = \xi^{n+1} + \xi^n = 2 \xi^{n+1/2} \) instead of \( \chi = \xi^{n+1} - \xi^n \) in (4.2). The bounds for most of the terms follow more easily than in the proof of Theorem 1 without using summation by parts in time. The definition (4.42) is used in the bound for the third term on the left hand side of (4.4). After multiplying by \( \Delta t \), we have

\[
|\Delta t(c(EV^n)d_\eta^n, 2\xi^{n+1/2})| \leq C|d_\eta^n|_1|\xi^{n+1/2}|_1 \Delta t \leq \frac{\Delta t}{64}|\xi^{n+1/2}|_1^2 + C \Delta t|d_\eta^n|_1^2.
\]

Then, as noted in [8], since the Neumann problem is \( H^3 \)-regular, for each \( t \in [0, T] \),

\[
|d_\eta^n|_1 \leq h^d|d_\eta^n|_1 \leq C_{23} h^r \left\{ \| u \|_{r-1} + \| \frac{\partial u}{\partial t} \|_{r-1} \right\}.
\]

\( C_{23} \) has the same dependencies as \( K_3 \). We now note that from the proof of Corollary 2 we obtain the estimate

\[
\Delta t \sum_{n=0}^{l-1} |d_\xi^n|^2 \leq C((\Delta t)^5 + h^{2r-2})
\]

which, with the assumption that \( \Delta t \leq C_1 h^2 \), yields

\[
\sum_{n=0}^{l-1} |d_\xi^n|^2 \leq C((\Delta t)^5 + \Delta t h^{2r-2}) \leq C((\Delta t)^5 + h^{2r}).
\]
We use (2.2a), (4.47) and \( \nu \) from (3.12) with \( \alpha = 1/2 \) (a norm reduction of \( O((\Delta t)^{1/2}) \)), to see that for \( n \geq 2 \),

\[
\left| \sum_{n=2}^{l-1} \Delta t \left[ \left( c(EV^n) \left( \frac{V_{n+1}^n - \tilde{V}_{n+1}^n}{\Delta t}, \zeta_{n+1} + \zeta^n \right) \right) + \frac{1}{2} \left( a(EV^n) \nabla (V_{n+1}^n - \tilde{V}_{n+1}^n), \nabla (\zeta_{n+1} + \zeta^n) \right) \right] \right|
\]

\[
\leq C \sum_{n=2}^{l-1} (\Delta t)^{1/2} \left\{ \| \delta^3 V^n \| + (\Delta t)^{1/2} \| \delta^3 V^n \|_1 \right\} \left\{ \| \zeta_{n+1} + \zeta^n \| + (\Delta t)^{1/2} \| \zeta_{n+1} + \zeta^n \|_1 \right\}
\]

\[
(4.48)
\]

\[
\leq C \sum_{n=2}^{l-1} (\Delta t)^{1/2} \left( 1 + \frac{K_0(\Delta t)^{1/2}}{h} \right) \left\{ \| \delta^3 \zeta^n \| + (\Delta t)^2 \sigma_{6,n} \right\} \left\{ \| \zeta_{n+1} \| + \| \zeta^n \| \right\}
\]

The rest of the proof follows in a manner similar to that of Theorem 1. We note in particular that (4.23) holds for \( l = 1, \ldots, N \) without a further induction argument since the hypotheses of Corollary 2 are assumed for this theorem. Thus, a norm-comparability argument for \( \| \cdot \|_c \) similar to that in (4.20) and (4.21) will follow.

5. Computational considerations. In this section we reconsider the preconditioned conjugate gradient (PCG) method of \( \S 3 \). It is computationally wasteful to iterate exactly \( \nu \) times each time step if \( \nu \) is determined using the pessimistic bound (3.12) and the results of \( \S 4 \). In this section we present some additional criteria which are very easy to apply and which can terminate the iteration in fewer than \( \nu \) steps. Next, to illustrate the effect of using incomplete iteration in terms of computational work, we give some rough operation counts for the linearized Crank–Nicolson procedure (3.3) and the PCG method. In one case, one can obtain work estimates for the PCG method which are of optimal order in the sense that the number of operations is proportional to the number of unknowns that define the solution. In another case we modify the basic process by changing the preconditioning matrix each \( (\Delta t)^{-1/2} \) time steps; this gives a norm reduction of \( O(\sqrt{\Delta t}) \) with one iteration and \( O(\Delta t) \) with two.

For \( \varphi \in H^1 \) let

\[
(5.1) \quad \| \varphi \|^2 = \| \varphi \|_c^2 = \| \varphi \|_c^2 + \Delta t \| \varphi \|_{\Delta t}^2.
\]

If we let \( V_{k+1}^n \) correspond to the vector \( x_k \) of (3.7), then, from (3.11),

\[
(5.2) \quad \psi_0 = \| V_{k+1}^n - \tilde{V}_{n+1}^n \|^2 / (L_0^{-1} q_k, q_k) \leq \psi_1.
\]

Since the denominator in (5.2) is computed during the course of the PCG procedure, we can easily estimate the size of the error \( \| V_{k+1}^n - \tilde{V}_{n+1}^n \| \), and, by comparing \( (L_0^{-1} q_0, q_0) \) and \( (L_0^{-1} q_k, q_k) \), we can observe the actual factor by which the norm is reduced. We can use these two quantities to stop the iteration if either is sufficiently small.
Suppose that
\[(5.3) \left\| \nabla^{n+1} - \tilde{\nabla}^{n+1} \right\|^2 \leq \mu. \]
(If the iteration is stopped at the kth step, then \(5.3\) follows if \(\psi_1(L_0^{-1} q_k, q_k)e \leq \mu\)). Then the left hand side of \(4.8\) can be bounded by
\[(5.4) e \left\| d\nabla^n \right\|^2 \Delta t + C \left[ \left\| \nabla^n \right\|^2 + \left\| \nabla^{n+1} \right\|^2 \right] \Delta t + C \mu (\Delta t)^{-2}.
Thus, if at each time step either a norm reduction factor of \(O(\Delta t)\) is achieved or if
\[(5.5) \mu (\Delta t)^{-2} \leq C(h^{2r-2} + (\Delta t)^2) \Delta t,
then the hypotheses of Theorem 1 are satisfied. Hence in the program one could set a parameter \(\kappa = O(h^{2r-2} + (\Delta t)^2)\) and stop iterating if
\[(5.6) (L_0^{-1} q_k, q_k)e \leq \kappa.
Similarly, a parameter \(\rho_1 = O(\Delta t)\) could be defined and the iteration could be halted if
\[(5.7) (L_0^{-1} q_k, q_k)e/(L_0^{-1} q_0, q_0)e \leq \rho_1^2.
With this additional test, a modification of Theorem 1 holds with the same error bounds.

For each result in § 4, there correspond appropriate choices of \(\kappa\) and \(\rho_1\) such that if the iteration is terminated whenever \(5.6\) or \(5.7\) holds, then the error bounds still apply. The following table summarizes these choices:

<table>
<thead>
<tr>
<th>Table 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Result</td>
</tr>
<tr>
<td>Theorem 1</td>
</tr>
<tr>
<td>Corollary 1</td>
</tr>
<tr>
<td>Corollary 2</td>
</tr>
<tr>
<td>Corollary 3</td>
</tr>
<tr>
<td>Theorem 2</td>
</tr>
</tbody>
</table>

As an aside we remark that if one chooses parameters \(\kappa\) and \(\rho_1\) given by the bounds indicated for Theorem 2 and if the solution is actually so smooth that Corollary 3 applies, then, as \(h\) and \(\Delta t\) go to zero, the \(\kappa\) test will almost always stop the iteration before the \(\rho_1\) test will. (By "almost always" we mean that the fraction of the timesteps stopped using the \(\rho_1\) test goes to zero.) This follows from the bounds used in \((4.8)\) and the conclusions of Corollary 3.

We shall now restrict our attention to spaces of piecewise cubic polynomials over quasi-regular meshes and give estimates of number of arithmetic operations needed to compute the extrapolated-Crank–Nicolson and the PCG approximate solutions. The heuristic arguments presented below can currently be made precise only in cases in which the meshes have very special structure, such as a uniform mesh on a square. However, numerical experiments indicate that the assumptions we use appear to be valid more generally.

Since we are using cubic polynomials as our example, \(r = 4\). Restrict \(\Omega\) to be a domain in the plane \((d = 2)\) for the moment. The quasi-regularity of the meshes (all elements in a given mesh are assumed to be about the same size and shape) implies that
\[(5.8) M \approx h^{-2}.
Balancing the $h^4$ and $(\Delta t)^2$ terms in the $L^2$ error bounds leads naturally to

$$
\Delta t \approx h^2;
$$

hence the number of time steps is

$$
N = T/\Delta t \approx M.
$$

We shall assume that the work to factor a matrix with the structure of $L_0$ is

$$
FW \approx M^{3/2}.
$$

In the case of a rectangular mesh on a rectangle this can be achieved by using the nested dissection process of George [11]; the results of Hoffman, Martin, and Rose [12] indicate that it cannot be improved. Experimentally, it has been found that minimal degree orderings (they are not unique) give the same form of work estimate. We shall also assume that the factors of $L_0$ have

$$
SW = M \ln M
$$

nonzero elements. Note that the work to perform one preconditioned conjugate gradient iteration is $O(SW + M) \approx SW$.

Combining these results shows that the work to compute the solution of the linearized-Crank-Nicolson scheme is

$$
N(FW + SW) = O(M^{3/2}).
$$

In the context of Theorem 1, Corollary 1 or Theorem 2 we would need $NI = O(\ln \Delta t) = O(\ln M)$ iterations at each step. Thus the expected work to compute the PCG approximation in these cases is

$$
FW + N \times NI \times SW = O(M^2(\ln M)^2).
$$

The processes analyzed by Corollaries 2 and 3 only require a fixed number of iterations per time step; hence in these cases the work is

$$
FW + N \times NI \times SW = O(M^2 \ln M).
$$

Note that $N \times M = O(M^2)$ parameters are used to define the solution. We see that the work estimates (5.14) and (5.15) are very close to optimal order since any process that deals with each parameter at least once must do at least $N \times M$ operations of some type.

We now indicate how the PCG scheme of § 3 can be modified to give an optimal order work estimate of $O(M^2)$ while still achieving the norm reduction necessary for Corollaries 2 and 3.

Suppose, as is frequently the case, that $C_0$ from (3.2f) is comparable with its diagonal; i.e., assume that there exists $\beta > 0$, independent of $h$, such that, if $D_0 = \text{diag}(C_0)$ and $0 \neq x \in \mathbb{R}^M$, then

$$
\beta^{-1} \leq (C_0 x, x)/(D_0 x, x) \leq \beta.
$$

In the case of Lagrange type cubic elements over a quasi-regular family of meshes this relation follows from a simple homogeneity argument, but it holds for some other element types as well. It follows from (2.2a), the fact that $\Delta t \leq Ch^2$, and (2.4) that there exists $\hat{\beta} > 0$, independent of $h$, such that for $0 \neq x \in \mathbb{R}^M$

$$
\hat{\beta}^{-1} \leq ((C^n + \Delta t A^n)x, x)/(D_0 x, x) \leq \hat{\beta}.
$$
If the iteration (3.7) uses $D_0$ as the preconditioning matrix instead of $L_0$, then (5.17) implies that a fixed number $N_I$ (independent of $h$) of iterations can be used to reduce the error by the factor $K_8$ needed for Corollaries 2 and 3. Thus the work required to compute the corresponding approximate solution is $O(M^3)$, since each iteration necessitates only $O(M)$ operations.

Another interesting modification of the basic process in § 3 is obtained by changing $L_0$ during the course of the computation. If we take $NS$ to be approximately $N^{1/2}$ and, if after each $NS$ steps we set $L_0$ equal to the current $L^n$, then it follows that $\psi_0 = 1 - c\sqrt{\Delta t}$ and $\psi_1 = 1 + c\sqrt{\Delta t}$. For this process we see that the $O(\Delta t)$ norm reduction needed for Theorem 1 and Corollary 1 requires only two iterations per time step, while the $O(\sqrt{\Delta t})$ needed in Theorem 2 can be obtained with one. With the assumptions made above we see that the work for this process is

$$\frac{N}{NS} * FW + N * SW = O(M^2 \ln M).$$

We briefly consider the case of $d = 3$. The best conjectures we know of say that

$$FW = M^2, \quad SW = M^{4/3}.$$  

Since we have $r = 4$, we still use

$$\Delta t = h^2$$

and

$$M = h^{-3},$$

$$N = T/\Delta t = M^{2/3}.$$  

Thus, an optimal order process would require $O(M^{5/3})$ operations and this can be achieved using the procedure described above that utilizes a diagonal preconditioning matrix. Even with the weaker smoothness constraints of Theorem 1 or 2 the expected work is the nearly optimal $O(M^{5/3} \ln M)$. If we use the basic PCG procedure of § 3 we see that the work is $O(M^2 \ln M)$; this is far superior to the $O(M^{8/3})$ operations needed to carry out the extrapolated-Crank–Nicolson process.

REFERENCES


