SIAM J. NUMER. ANAL.
Vol. 12, No. 3, June 1975

NUMERICAL SOLUTION OF SOBOLEV PARTIAL DIFFERENTIAL EQUATIONS*

RICHARD E. EWING†

Abstract. Finite difference techniques can be applied to the numerical solution of the initial-boundary value problem in S for the semilinear Sobolev or pseudo-parabolic equation

$$\sum_{i=1}^{n}\left[\frac{\partial}{\partial x_i}\left(a_i\frac{\partial}{\partial x_i}u_t\right)+\frac{\partial}{\partial x_i}\left(b_i\frac{\partial}{\partial x_i}u\right)\right]-q=ru_t,$$

where a_i, b_i, q and r are functions of space and time variables, q is a boundedly differentiable function of u, and S is an open, connected domain in $\mathbb{R}^n$. Under suitable smoothness conditions, the solution of a Crank–Nicolson type of difference equation is shown to converge to u in the discrete L^2-norm with an $O((\Delta x)^2+(\Delta t)^2)$ discretization error.

The numerical problem is reduced to the inversion of a certain matrix at each time level. For the problem with constant coefficients in a two- or three-dimensional cube, a two-level iteration scheme with a Picard-type outer iteration and an alternating direction inner iteration is presented. For more general operators and more general regions in $\mathbb{R}^n$ for arbitrary n the same two-level scheme with a successive overrelaxation inner iteration is discussed.

1. Introduction. The purpose of this paper is to consider the numerical solution of certain partial differential equations with one time derivative appearing in the highest order terms. Equations of this type arise in many areas of mathematics and physics. They are used to study consolidation of clay [25], heat conduction [2], homogeneous fluid flow in fissured material [1], shear in second order fluids [3], [18] and other physical models. In connection with nonsteady flow of second order fluids, Ting [26] considers the initial-boundary value problem for the equation

$$\rho v_t=\tfrac{1}{2}av_{xx}+cv_{xxt},\qquad 0\leqq x\leqq h,\quad t\geqq 0,\tag{1.1}$$

with constant coefficients. For a discussion of several physical applications of the nonlinear problem, see [16].

In a Hilbert space setting, this type of equation is of the form

$$u'(t)+Bu'(t)+Au(t)=0,\qquad t>0,\tag{1.2}$$

where A and B are various operators. Yosida used the equation (1.2) with $B=\beta A$ in his proof of the generation theorem for semigroups of operators [28]. The equation (1.2) with $B=\beta A$ has also been used to approximate certain parabolic equations backward in time [13], [23]. In [24], Showalter and Ting discuss the initial-boundary value problem of the type (1.2) using Hilbert space techniques. Davis [6] and Showalter [21], [22] have also considered various mathematical aspects of equations of this type. Ford [14] has considered some numerical aspects of this type of problem.

* Received by the editors January 15, 1974.

† Department of Mathematics, Oakland University, Rochester, Michigan 48063. This work was supported in part by Texas Tech University and by the Center for Numerical Analysis at the University of Texas at Austin.

In this paper we consider linear and semilinear initial-boundary value problems in $S \times (0, T]$ of the form

$$(1.3)\quad \begin{cases} \text{(a)} & \displaystyle\sum_{k=1}^{m}\left\{\frac{\partial}{\partial x_k}\left[a_k(x,t)\frac{\partial}{\partial x_k}u_t\right]+\frac{\partial}{\partial x_k}\left[b_k(x,t)\frac{\partial}{\partial x_k}u\right]\right\}-qu=ru_t+h, \\ & \qquad\qquad (x_1,\cdots,x_m)\in S,\quad 0<t\leqq T, \\ \text{(b)} & u(x_1,\cdots,x_m,0)=f(x_1,\cdots,x_m),\qquad (x_1,\cdots,x_m)\in S, \\ \text{(c)} & u(x_1,\cdots,x_m,t)=g(x_1,\cdots,x_m,t),\qquad (x_1,\cdots,x_m)\in\partial S,\quad 0<t\leqq T, \end{cases}$$

where S is an open connected subset of $\mathbb{R}^m$ and ∂S is the boundary of S. We later describe smoothness assumptions and bounds on u and the coefficients above. Standard problems of this type have $a_i > 0$, $b_i > 0$ and $r > 0$. The backward time problems are given by $a_i > 0$, $b_i < 0$ and $r > 0$. The results of this paper hold for both time cases; however, as we shall see, most of the restrictions on Δt can be dropped when $b_i > 0$.

We note that if we limit the number of time levels in the difference equation approximations to two, the time derivative in the highest order terms necessitates the use of implicit numerical schemes. Thus due to the increased rate of convergence over standard implicit schemes, we use the m-dimensional analogue of the Crank–Nicolson difference equation [5] to replace the differential problem.

In § 2 we establish some special notation and present basic assumptions needed throughout the paper. In § 3 we derive eigenvalue estimates for the problems to be studied in §§ 4 and 5 for fairly general domains and use these to obtain stability. Using the stability analysis and eigenvalue estimates, we derive convergence of the Crank–Nicolson difference schemes for the linear problems of type (1.3) in § 4 and for semilinear problems in § 5. Finally in § 6, we discuss the algebraic problems for the Crank–Nicolson schemes of the previous sections. We reduce the problem of convergence to the inversion of a matrix at each time level. In order to treat the nonlinear difference equations of § 5, we present a pair of two-level iteration schemes. The first method, using an alternating direction inner iteration, applies to equations in a rectangular box with constant coefficients for $m = 2$ or $m = 3$ and requires calculations of the order

$$(1.4)\qquad O((\Delta x)^{-m}\log(\Delta x)^{-1})$$

at each step. The second method, using a successive overrelaxation inner iteration, applies to more general equations in more general regions for arbitrary m and require calculations of the order

$$(1.5)\qquad O((\Delta x)^{-(m+1)})$$

at each step.

2. Preliminaries and notation. We shall require some special notation and assumptions. Let $l_1, l_2, \cdots, l_m$ be a basis of unit coordinate vectors in $\mathbb{R}^m$. To set up a finite difference equation, we fix a point $(0, 0, \cdots, 0)$ and construct a rectangular lattice whose nodes are $x = (x_1, x_2, \cdots, x_m)$ such that

$$(2.1)\qquad x_k = p_k\Delta x_k, \qquad k = 1, 2, \cdots, m,$$

where $p_k = 0, \pm 1, \pm 2, \cdots$, and for each k, Δx_k is the mesh size in the direction l_k. We define the average mesh size by

$$\Delta x = \frac{1}{m} \sum_{k=1}^{m} \Delta x_k. \tag{2.2}$$

Two nodes with coordinates $p_k \Delta x_k$ and $p'_k \Delta x_k$ are *adjacent* if $\sum_{k=1}^{m} (p_k - p'_k)^2 = 1$. The set of nodes in S such that all adjacent nodes belong to $S \cup \partial S$ is the *interior* of S denoted S_h. All other nodes in $S \cup \partial S$ belong to the *boundary* of $S \cup \partial S$ denoted ∂S_h. We assume S is connected. Also we must make the somewhat stringent assumption that $\partial S_h \subset \partial S$.

We consider the vector

$$\alpha = (p_1, p_2, \cdots, p_m) \tag{2.3}$$

and use the notation for any function f,

$$f_\alpha = f(p_1 \Delta x_1, p_2 \Delta x_2, \cdots, p_m \Delta x_m), \tag{2.4a}$$

$$f_{\alpha,n} = f(p_1 \Delta x_1, p_2 \Delta x_2, \cdots, p_m \Delta x_m, t_n), \tag{2.4b}$$

$$f_{\alpha+(1/2)l_k,n} = f(p_1 \Delta x_1, \cdots, p_{k-1} \Delta x_{k-1}, (p_k + \tfrac{1}{2}) \Delta x_k, p_{k+1} \Delta x_{k+1}, \cdots, p_m \Delta x_m, t_n) \tag{2.4c}$$

and similarly for $f_{\alpha-(1/2)l_k,n}$.

We shall not assume that Δt equals Δx_k; however, when we consider S as a cube we assume $\Delta x_1 = \Delta x_2 = \cdots = \Delta x_m$. The standard Landau order notation will be used. If f is a function of several variables,

$$f \in C^\beta \tag{2.5}$$

implies that all partial derivatives of f of order not greater than β are continuous.

Now we present a list of basic difference formulas we shall use. The proofs of these formulas follow from Taylor's theorem.

$$\begin{aligned} &\left(\frac{\partial}{\partial x_k} \left[(a_k)_{n+1/2} \frac{\partial}{\partial x_k} f_n \right] \right)_\alpha \\ &\quad = [(a_k)_{\alpha+(1/2)l_k,n+1/2}(f_{\alpha+l_k,n} - f_{\alpha,n}) \\ &\qquad - (a_k)_{\alpha-(1/2)l_k,n+1/2}(f_{\alpha,n} - f_{\alpha-l_k,n})]/(\Delta x_k)^2 + O((\Delta x)^2), \\ &\qquad\qquad a_k \in C^3, \quad f \in C^4, \end{aligned} \tag{2.6a}$$

$$f_{\alpha,n+1/2} = (f_{\alpha,n+1} + f_{\alpha,n})/2 + O((\Delta t)^2), \qquad f \in C^2. \tag{2.6b}$$

We shall use standard difference notations [9, pp. 2–4] as well as the notation

$$(\Delta_x[a_{n+1/2} \Delta_x f_n])_\alpha = \sum_{k=1}^{n} (\Delta_{x_k}[(a_k)_{n+1/2} \Delta_{x_k} f_n])_\alpha. \tag{2.7}$$

We now make some smoothness assumptions and define some bounds for our problem. For the problem under consideration, we assume:

(2.8)

(i) there exists a unique solution $u \in C^5$ in $\bar{S}$, the closure of S,

(ii) a_k and b_k are three times boundedly differentiable in the kth space variable and in t, and q and r are boundedly differentiable in all space variables and in t,

(iii) the coefficients satisfy the bounds,

(a) $0 < A_* \leqq a_k(x, t) \leqq A^*$ for $k = 1, 2, \cdots, m$,

(b) $0 < R_* \leqq r(x, t) \leqq R^*$,

(c) $B_* \leqq b_k(x, t) \leqq B^*$ for $k = 1, 2, \cdots, m$,

(d) $Q_* \leqq q(x, t) \leqq Q^*$,

where B_* and Q_* may be negative and B^* and Q^* are nonnegative.

3. Stability from eigenvalue estimates. We first consider eigenvalue estimates for a difference equation used to solve differential equations of the type (1.3) where S is a cube, $a_k = a_k(x_1, \cdots, x_m, t)$, $b_k = b_k(x_1, \cdots, x_m, t)$ for $k = 1, 2, \cdots, m$, $q = q(x_1, \cdots, x_m, t)$ and $r = r(x_1, \cdots, x_m, t)$. We shall generalize S later. We make the assumptions described in (2.8).

Using the notation of (2.7), consider the Crank–Nicolson difference equation

$$\frac{(\Delta_x[a_{n+1/2}\Delta_x w_{n+1}])_\alpha - (\Delta_x[a_{n+1/2}\Delta_x w_n])_\alpha}{\Delta t} + \frac{(\Delta_x[b_{n+1/2}\Delta_x w_{n+1}])_\alpha}{2}$$

$$(3.1a) \qquad + \frac{(\Delta_x[b_{n+1/2}\Delta_x w_n])_\alpha}{2} - \frac{q_{\alpha,n+1/2}(w_{\alpha,n+1} + w_{\alpha,n})}{2} = \frac{r_{\alpha,n+1/2}(w_{\alpha,n+1} - w_{\alpha,n})}{\Delta t},$$

$$(x_1, \cdots, x_m) \in S_h,$$

$$(3.1b) \qquad w_{\alpha,0} = u_{\alpha,0}, \qquad (x_1, \cdots, x_m) \in S_h,$$

$$(3.1c) \qquad w_{\alpha,n+1} = u_{\alpha,n+1}, \qquad (x_1, \cdots, x_m) \in \partial S_h.$$

Rearranging (3.1a), we have

$$(3.2) \qquad \begin{aligned} &\frac{r_{\alpha,n+1/2}w_{\alpha,n+1} - (\Delta_x[a_{n+1/2}\Delta_x w_{n+1}])_\alpha}{\Delta t} + \frac{q_{\alpha,n+1/2}w_{\alpha,n+1} - (\Delta_x[b_{n+1/2}\Delta_x w_{n+1}])_\alpha}{2} \\ &= \frac{r_{\alpha,n+1/2}w_{\alpha,n} - (\Delta_x[a_{n+1/2}\Delta_x w_n])_\alpha}{\Delta t} - \frac{q_{\alpha,n+1/2}w_{\alpha,n} - (\Delta_x[b_{n+1/2}\Delta_x w_n])_\alpha}{2}. \end{aligned}$$

Trying separation of variables, we consider a solution of the form

$$(3.3) \qquad w_{\alpha,n} = \rho_n \psi_\alpha.$$

By direct substitution we see that

$$(3.4) \qquad \begin{aligned} \frac{\rho_{n+1}}{\rho_n} &= \frac{[r_{\alpha,n+1/2}\psi_\alpha - (\Delta_x[a_{n+1/2}\Delta_x\psi])_\alpha]/\Delta t - [q_{\alpha,n+1/2}\psi_\alpha - (\Delta_x[b_{n+1/2}\Delta_x\psi])_\alpha]/2}{[r_{\alpha,n+1/2}\psi_\alpha - (\Delta_x[a_{n+1/2}\Delta_x\psi])_\alpha]/\Delta t + [q_{\alpha,n+1/2}\psi_\alpha - (\Delta_x[b_{n+1/2}\Delta_x\psi])_\alpha]/2} \\ &= \nu. \end{aligned}$$

Then for each n we have the eigenvalue problem

$$A_n\phi^{(n)} = \frac{2(1-v)}{\Delta t(1+v)} B_n\phi^{(n)}, \qquad (x_1, \cdots, x_m) \in S_h, \tag{3.5a}$$

$$\phi^{(n)} = 0, \qquad (x_1, \cdots, x_m) \in \partial S_h, \tag{3.5b}$$

where

$$(A_n\phi^{(n)})_\alpha = (q_{n+1/2}\phi^{(n)})_\alpha - (\Delta_x[b_{n+1/2}\Delta_x\phi^{(n)}])_\alpha, \tag{3.6a}$$

$$(B_n\phi^{(n)})_\alpha = (r_{n+1/2}\phi^{(n)})_\alpha - (\Delta_x[a_{n+1/2}\Delta_x\phi^{(n)}])_\alpha. \tag{3.6b}$$

We can see [27] that the matrices A_n and B_n are symmetric.

If N is the number of nodes in S_h, we define the inner product on $\mathbb{R}^N$,

$$(x, y) = (\Delta x)^m \sum_\alpha^N x_\alpha y_\alpha \tag{3.7}$$

and the induced norm

$$\|x\|_2 = \left((\Delta x)^m \sum_\alpha^N x_\alpha^2\right)^{1/2}. \tag{3.8}$$

This norm is the discrete analogue of the integral L^2-norm and will be called the discrete L^2-norm. By direct calculation and as in [8, p. 515] we see that

$$\begin{aligned}(B_ny, y) &= (r_{n+1/2}y, y) - (\Delta_x[a_{n+1/2}\Delta_x y], y)\\ &= (r_{n+1/2}y, y) + \sum_{k=1}^m ((a_k)_{n+1/2}\delta_{x_k}y, \delta_{x_k}y),\end{aligned} \tag{3.9}$$

where

$$\delta_{x_k}y = (y_{\alpha+l_k} - y_\alpha)/\Delta x. \tag{3.10}$$

Thus, due to (2.8) (iii) (a, b), we see that the matrix B_n is positive definite over the real vector space $\mathbb{R}^N$. Since B_n is symmetric and positive definite, we can define a new inner product on $\mathbb{R}^N$ for each n by

$$(x, y)_{B_n} = (B_nx, y), \tag{3.11}$$

with the corresponding norm given by

$$\|x\|_n = (B_nx, x)^{1/2}. \tag{3.12}$$

It is shown in [4, pp. 37–41] that there exists a complete set of eigenvectors of (3.5) for each $n = 1, 2, \cdots$, which are orthogonal with respect to the inner product (3.11). We can now use a variational attack based on the Courant minimax principle [4] to obtain upper and lower bounds on the eigenvalues of (3.5). We shall state the following theorem when B_* and Q_* are negative. Similar but less restrictive theorems hold for other signs of B_* and Q_*.

THEOREM 3.1. *Let $\lambda_1^{(n)}$ be the least eigenvalue of the eigenvalue problem* (3.5) *for a fixed n and $\lambda_N^{(n)}$ the greatest eigenvalue. We have the following bounds which* *are uniform in n.*

$$
\begin{aligned}
&\textit{If } |Q_*| \geqq (R_*|B_*|/A_*), && \lambda_1^{(n)} \geqq Q_*/R_*.\\
&\textit{If } |Q_*| < (R_*|B_*|/A_*), && \lambda_1^{(n)} \geqq B_*/A_*.\\
&\textit{If } \ Q^* \geqq R_*|B^*|/A_*, && \lambda_N^{(n)} \leqq Q^*/R_*.\\
&\textit{If } \ Q^* < R_*|B^*|/A_*, && \lambda_N^{(n)} \leqq B^*/A_*.
\end{aligned}
\tag{3.13}
$$

Proof. First we note that as in (3.9),

$$
(A_n\phi, \phi) = (q_{n+1/2}\phi, \phi) + \sum_{k=1}^{m} ((b_k)_{n+1/2}\delta_{x_k}\phi, \delta_{x_k}\phi),
\tag{3.14a}
$$

and

$$
(B_n\phi, \phi) = (r_{n+1/2}\phi, \phi) + \sum_{k=1}^{m} ((a_k)_{n+1/2}\delta_{x_k}\phi, \delta_{x_k}\phi).
\tag{3.14b}
$$

Therefore for $\phi \neq 0$, we see that $(A_n\phi, \phi)/(B_n\phi, \phi)$ is bounded below by

$$
\left(Q_*\|\phi\|_2^2 + B_* \sum_{k=1}^{m} \|\delta_{x_k}\phi\|_2^2\right) \Big/ \left(R_*\|\phi\|_2^2 + A_* \sum_{k=1}^{m} \|\delta_{x_k}\phi\|_2^2\right)
\tag{3.15}
$$

since Q_* and B_* are both negative. We note here that since the bounds (2.8) hold for all t, the bound (3.15) holds for all n. Thus we can obtain uniform bounds on the eigenvalues by considering the eigenvalue problem

$$
Q_*\phi - B_* \sum_{k=1}^{m} \Delta_{x_k}^2\phi = \mu\left(R_*\phi - A_* \sum_{k=1}^{m} \Delta_{x_k}^2\phi\right),
\tag{3.16a}
$$

$$
\phi = 0 \quad \text{if}\,(x_1, \cdots, x_m) \in \partial S_h.
\tag{3.16b}
$$

Therefore, by the Courant minimax principle and a special case of the minimax theorem [17, p. 181], we know that

$$
\lambda_1^{(n)} \geqq \min_{i=1,\cdots,N} \mu^{(i)},
\tag{3.17}
$$

where $\mu^{(i)}$ are the eigenvalues of (3.16). Consider the following rearrangement of (3.16a):

$$
\sum_{k=1}^{m} \Delta_{x_k}^2\phi = \left(\frac{\mu R_* - Q_*}{\mu A_* - B_*}\right)\phi.
\tag{3.18}
$$

Since we are working on a cube, $\Delta x_k = \Delta x$ for $k = 1, 2, \cdots, m$, we can let the p_k range from 1 to some M for $k = 1, 2, \cdots, m$.

Standard arguments [9], [11] yield

$$
\sum_{k=1}^{m} -\frac{4}{(\Delta x)^2}\sin^2\frac{\pi p_k \Delta x}{2} = \left(\frac{\mu R_* - Q_*}{\mu A_* - B_*}\right).
\tag{3.19}
$$

Solving for μ, we obtain

$$
\mu = \frac{\left[Q_* + B_*(4/(\Delta x)^2) \sum_{k=1}^{m} \sin^2(\pi p_k \Delta x/2)\right]}{\left[R_* + A_*(4/(\Delta x)^2) \sum_{k=1}^{m} \sin^2(\pi p_k \Delta x/2)\right]}.
\tag{3.20}
$$

Next we note that if we consider the function

$$f(x)=\frac{Q_*+B_*x}{R_*+A_*x},\tag{3.21}$$

then $f'(x)\geqq 0$ if

$$(R_*B_*-A_*Q_*)\geqq 0.\tag{3.22}$$

Since $B_*<0$ and $Q_*<0$, this amounts to

$$|Q_*|\geqq (R_*|B_*|/A_*).\tag{3.23}$$

If (3.23) is satisfied, then $f(x)$ is either increasing or a constant function of x which is nonnegative as chosen. We note that no use is made of the size of the eigenvalues and thus there is no restriction on the size of the cube S. We now see that

$$\min_{i=1,2,\cdots,N}\mu^{(i)}\geqq f(0)=Q_*/R_*.\tag{3.24}$$

If (3.23) does not hold, then $f(x)$ is a strictly decreasing function of x and we have

$$\min_{i=1,\cdots,N}\mu^{(i)}\geqq \lim_{x\to\infty}f(x)=B_*/A_*.\tag{3.25}$$

Thus by (3.17), we see that for all n,

$$\text{if } |Q_*|\geqq (R_*|B_*|/A_*),\quad \text{then } \lambda_1^{(n)}\geqq Q_*/R_*,\tag{3.26a}$$

and

$$\text{if } |Q_*|< (R_*|B_*|/A_*),\quad \text{then } \lambda_1^{(n)}\geqq B_*/A_*.\tag{3.26b}$$

The upper bounds in (3.13) follow similarly and the theorem is proved. We remark that if A_* could be zero, we would have the standard parabolic equation and we could not obtain the upper bounds on $\lambda_N^{(n)}$.

We note there are various cases for different signs and magnitudes of B_* and Q_*. The above theorem is the worst case, whereas the least restrictive case is when $B_*>0$ and $Q_*>0$. Clearly, only the lower bounds on the eigenvalues will be affected. Instead of the bound (3.15), for the new choice of B_* and Q_*, we see that $(A_n\phi,\phi)/(B_N\phi,\phi)$ is bounded below by

$$\left(Q_*\|\phi\|_2^2+B_*\sum_{k=1}^{m}\|\delta_{x_k}\phi\|_2^2\right)\Bigg/\left(R^*\|\phi\|_2^2+A^*\sum_{k=1}^{m}\|\delta_{x_k}\phi\|_2^2\right).\tag{3.27}$$

The analysis follows as before and we obtain the following result.

COROLLARY 3.2. *If $B_*>0$ and $Q_*>0$, the uniform lower bounds on the eigenvalues of* (3.5) *are replaced by*

$$\textit{if } Q_*\geqq R^*B_*/A^*,\quad \textit{then } \lambda_1^{(n)}\geqq Q_*/R^*>0,\tag{3.28a}$$

and

$$\textit{if } Q_*< R^*B_*/A^*,\quad \textit{then } \lambda_1^{(n)}\geqq B_*/A^*>0.\tag{3.28b}$$

Now we consider a generalization of the region S. Instead of a cube, we assume S to be as described in § 2. Let Ω be the least cube containing the lattice nodes in S and on ∂S, and $\partial\Omega$ be its boundary. It is well known [20, p. 204] that any matrix corresponding to the operators

$$Q^* - B^* \sum_{k=1}^{m} \Delta_{x_k}^2, \quad Q_* - B_* \sum_{k=1}^{m} \Delta_{x_k}^2, \quad R^* - A^* \sum_{k=1}^{m} \Delta_{x_k}^2, \tag{3.29}$$
$$\text{and} \quad R_* - A_* \sum_{k=1}^{m} \Delta_{x_k}^2$$

as applied to any lattice region, regardless of the ordering of the points, is symmetric. Similarly, by direct substitution as in (3.9) we can see that the matrices corresponding to

$$Q^* - B^* \sum_{k=1}^{m} \Delta_{x_k}^2 \quad \text{and} \quad Q_* - B_* \sum_{k=1}^{m} \Delta_{x_k}^2 \tag{3.30}$$

are positive definite regardless of the ordering of the points. Thus we can order the points in S_h first and then order the other points in Ω. The resulting matrices for the eigenvalue problem (3.16) will still be positive definite and symmetric as required. We can then obtain bounds for the eigenvalues in Ω as outlined above. Then applying a theorem concerning domination of eigenvalues [20, p. 164] successively on the lattice nodes in Ω but not in S, we shall retain the same uniform bounds on the eigenvalues for S as for Ω.

We shall next define the stability of (3.1) with respect to the sequence of norms given in (3.12). First we note that (3.2) is actually of the form

$$(C_n)_\alpha w_{\alpha,n+1} = (D_n)_\alpha w_{\alpha,n}\,, \tag{3.31}$$

where

$$(C_n)_\alpha = \frac{r_{\alpha,n+1/2} - (\Delta_x[a_{n+1/2}\Delta_x \cdot])_\alpha}{\Delta t} + \frac{q_{\alpha,n+1/2} - (\Delta_x[b_{n+1/2}\Delta_x \cdot])_\alpha}{2} \tag{3.32a}$$

and

$$(D_n)_\alpha = \frac{r_{\alpha,n+1/2} - (\Delta_x[a_{n+1/2}\Delta_x \cdot])_\alpha}{\Delta t} - \frac{q_{\alpha,n+1/2} - (\Delta_x[b_{n+1/2}\Delta_x \cdot])_\alpha}{2} \tag{3.32b}$$

and C_n can be shown to be invertible with certain restrictions on Δt. As in [9, p. 42], equation (3.1) will be defined to be *stable* with respect to the sequence of norms given in (3.12) provided

$$\|C_n^{-1}D_n\|_n \leqq (1 + \gamma\Delta t), \qquad n = 0, 1, \cdots, \tag{3.33}$$

for all sufficiently small Δt, where γ is a positive constant independent of Δt and n.

It is well known that for the eigenvalue problem for (3.31) with eigenvalues given by (3.4), since

$$\|C_n^{-1}D_n\|_n \leqq \max_{i=1,2,\cdots,N} |v^{n,i}|, \tag{3.34}$$

where N is the number of nodes in S_h, then (3.33) will be satisfied with γ independent of Δt and n if we can get uniform, in n, estimates of the form

$$\max_{i=1,\cdots,N} |v^{n,i}| \leqq 1 + \gamma\Delta t \tag{3.35}$$

for all sufficiently small Δt. Theorem 3.1 yields the estimate

$$-A_1 \leqq \frac{2(1-v)}{\Delta t(1+v)} \leqq A_2, \tag{3.36}$$

where v is given by (3.4) and

$$A_1 = |B_*|/A_* \quad \text{if } |Q_*| < (R_*|B_*|/A_*), \tag{3.37a}$$

$$A_1 = |Q_*|/R_* \quad \text{if } |Q_*| \geqq (R_*|B_*|/A_*), \tag{3.37b}$$

$$A_2 = B^*/A_* \quad \text{if } Q^* < (R_*B^*/A_*), \tag{3.37c}$$

$$A_2 = Q^*/R_* \quad \text{if } Q^* \geqq (R_*B^*/A_*), \tag{3.37d}$$

when $B_* < 0$ and $Q_* < 0$. We first consider the properties of $(1-v)/(1+v)$.

From properties of $(1-v)/(1+v)$, as in Part II of [12], one can easily see that for

$$\Delta t < 2(1-\varepsilon)/A_1 \tag{3.38}$$

for some $\varepsilon > 0$, where A_1 is given in (3.37), we have

$$-1 < v \leqq 1 + \gamma\Delta t \tag{3.39}$$

where γ is independent of Δt and n. Also, we see that for the least restrictive case, where $B_* > 0$ and $Q_* > 0$, from Corollary 3.2, we have $A_1 > 0$ as a lower bound for (3.36). The resulting restriction on v from (3.36) is just

$$|v| < 1 \tag{3.40}$$

for any choice of $\Delta t > 0$. Therefore, with the restriction (3.38) for the worst case, when $B_* < 0$ and $Q_* < 0$, and no restriction for the best case, (3.39) and (3.40) show (3.35) and thus (3.33) is satisfied. Thus (3.31) is stable with respect to the sequence of norms given in (3.12).

4. Convergence for linear equations. Consider the linear initial-boundary value problem

$$\sum_{k=1}^{m}\left[\frac{\partial}{\partial x_k}\left(a_k\frac{\partial}{\partial x_k}u_t\right) + \frac{\partial}{\partial x_k}\left(b_k\frac{\partial}{\partial x_k}u\right)\right] - qu - h = ru_t, \quad (x_1, \cdots, x_m) \in S, \quad 0 < t \leqq T, \tag{4.1a}$$

$$u(x_1, \cdots, x_m, 0) = f(x_1, \cdots, x_m), \qquad (x_1, \cdots, x_m) \in S, \tag{4.1b}$$

$$u(x_1, \cdots, x_m, t) = g(x_1, \cdots, x_m, t), \qquad (x_1, \cdots, x_m) \in \partial S, \quad 0 < t \leqq T, \tag{4.1c}$$

with

$$a_k = a_k(x_1, \cdots, x_m, t), \quad b_k = b_k(x_1, \cdots, x_m, t),$$

$$q = q(x_1, \cdots, x_m, t), \quad r = r(x_1, \cdots, x_m, t)$$

satisfying (2.8) and S as defined in § 2. Assume u satisfies (2.8). We now shall use the results of § 3 to prove L^2 convergence of the solution of the Crank–Nicolson difference equation (3.2) to the solution u of (4.1).

Using (2.6), (2.7) and (2.8) we see that

$$\begin{aligned}\frac{(\Delta_x[a_{n+1/2}\Delta_x u_{n+1}])_\alpha - (\Delta_x[a_{n+1/2}\Delta_x u_n])_\alpha}{\Delta t} + \frac{(\Delta_x[b_{n+1/2}\Delta_x u])_\alpha}{2} \\ + \frac{(\Delta[b_{n+1/2}\Delta_x U_{n+1}])_\alpha}{2} - \tfrac{1}{2}q_{\alpha,n+1/2}(u_{\alpha,n+1} + u_{\alpha,n}) - h_{\alpha,n+1/2} \\ = \frac{r_{\alpha,n+1/2}(u_{\alpha,n+1} - u_{\alpha,n})}{\Delta t} + \sigma_{\alpha,n},\end{aligned} \tag{4.2}$$

where $\sigma_{\alpha,n}$ is $O((\Delta x)^2 + (\Delta t)^2)$. Then we let w be a solution to the Crank–Nicolson difference equation (3.1). Due to (4.2), (3.1) as defined is consistent.

Let

$$z_{\alpha,n} = u_{\alpha,n} - w_{\alpha,n}. \tag{4.3}$$

By subtracting (3.1) from (4.2), we have the linear difference system

$$(C_n z_{n+1})_\alpha = (D_n z_n)_\alpha + \sigma_{\alpha,n}, \qquad x_\alpha \in S_h, \tag{4.4a}$$

$$z_{\alpha,0} = 0, \qquad x_\alpha \in S_h, \tag{4.4b}$$

$$z_{\alpha,n+1} = 0, \qquad x_\alpha \in \partial S_h, \tag{4.4c}$$

where C_n and D_n are given by (3.22).

Stability has been proved in § 3. From (3.9), it is easily seen that

$$\|x\|_2 \leqq R_*^{-1/2}\|x\|_n \tag{4.5}$$

for all n. The existence of a constant β such that

$$\|x\|_{n+1} \leqq (1 + \beta\Delta t)\|x\|_n \tag{4.6}$$

for all $\Delta t < T$ follows from the smoothness of a_k, $1 \leqq k \leqq m$ and r.

Finally, due to the analysis of Douglas [9, pp. 41–44], all we need to do to prove the convergence of our approximation in $\|\cdot\|_2$ is to show that

$$(\Delta t)^{-1}\|C_n^{-1}\sigma_n\|_n = O((\Delta t)^s) \tag{4.7}$$

for some $s > 0$. First we must show that C_n is actually invertible. If not, there exists a $\phi \in \mathbb{R}^N$, $\phi \neq 0$, such that $C_n\phi = 0$. Then from (3.6) and (3.32) we see that

$$A_n\phi = -(2/\Delta t)B_n\phi \tag{4.8}$$

and $-2/\Delta t$ is an eigenvalue of (4.8) contradicting the restriction (3.38).

Now as in [14] we would like to consider a lemma which gives us a formula for $\|A\|_n$ where A is any $N \times N$ matrix. First note that the spectral radius of A, $\rho(A)$, is just $\rho(A) = \max_i |\lambda_i|$ where λ_i are the eigenvalues of A.

LEMMA 4.1. *Let A and B be $N \times N$ matrices with B symmetric and positive definite. Define the norm $\|x\|_B = (Bx, x)^{1/2}$ where $(\cdot,\cdot)$ is the discrete L^2 inner product from* (3.7). *It follows that*

$$\|A\|_B = (\rho(B^{-1}A^TBA))^{1/2}, \tag{4.9}$$

where A^T is the transpose of A.

We shall now use Lemma 4.1 to estimate $\|C_n^{-1}\sigma_n\|_n^2$. By definition

$$\|C_n^{-1}\sigma_n\|_n^2 = (B_nC_n^{-1}\sigma_n, C_n^{-1}\sigma_n). \tag{4.10}$$

Since $B_nC_n^{-1}$ is not necessarily symmetric, it would be hard to determine $\|B_nC_n^{-1}\|_2$. However, since C_n is symmetric, C_n^{-1} is symmetric,

$$\|C_n^{-1}\sigma_n\|_n^2 = (C_n^{-1}B_nC_n^{-1}\sigma_n, \sigma_n), \tag{4.11}$$

and $C_n^{-1}B_nC_n^{-1}$ is seen to be symmetric. Therefore, since the spectral radius of a matrix is a lower bound for any norm of the matrix [19, p. 13], we have that

$$\begin{aligned}\|C_n^{-1}B_nC_n^{-1}\|_2 &= \rho(C_n^{-1}B_nC_n^{-1})\\ &\leqq \|C_n^{-1}B_nC_n^{-1}\|_n\\ &\leqq \|C_n^{-1}B_n\|_n\|C_n^{-1}\|_n.\end{aligned} \tag{4.12}$$

Using separation of variables to estimate $\|C_n^{-1}B_n\|_n$, since $C_n = B_n/\Delta t + A_n/2$, we see that

$$B_n\phi = \lambda C_n\phi = \lambda(B_n/\Delta t + A_n/2)\phi. \tag{4.13}$$

Rearranging, we arrive at the eigenvalue problem

$$A_n\phi = \frac{2}{\lambda}\left(1 - \frac{\lambda}{\Delta t}\right)B_n\phi. \tag{4.14}$$

Theorem 3.1 gives the inequality from (3.36)

$$-A_1 \leqq \frac{2}{\lambda} - \frac{2}{\Delta t} \leqq A_2 \tag{4.15}$$

or

$$0 < \frac{1}{1/\Delta t + A_2/2} \leqq \lambda \leqq \frac{1}{1/\Delta t - A_1/2} \tag{4.16}$$

for the case where $B_* < 0$ and $Q_* < 0$. Choosing Δt to satisfy (3.38) we see that

$$0 < \lambda \leqq \frac{\Delta t}{\varepsilon} \tag{4.17}$$

and $\lambda = O(\Delta t)$. Thus we see that

$$\|C_n^{-1}B_n\|_n = O(\Delta t). \tag{4.18}$$

Then by Lemma 4.1, since C_n^{-1} is symmetric,

$$\|C_n^{-1}\|_n^2 = \rho(B_n^{-1}C_n^{-1}B_nC_n^{-1}) \leqq \|B_n^{-1}\|_n\|C_n^{-1}B_n\|_n\|C_n^{-1}\|_n \tag{4.19}$$

or

$$\|C_n^{-1}\|_n \leqq \|C_n^{-1}B_n\|_n\|B_n^{-1}\|_n. \tag{4.20}$$

Lemma 4.1 also implies that

$$\|B_n^{-1}\|_n = \rho(B_n^{-1}) = \|B_n^{-1}\|_2. \tag{4.21}$$

Note that

$$\frac{(B_nx, x)}{(x, x)} = \frac{(r_{n+1/2}x, x) + \sum_{k=1}^m (a_{n+1/2}\delta_{x_k}x, \delta_{x_k}x)}{(x, x)} \geqq R_* \tag{4.22}$$

for $x \neq 0$. Then by the minimax principle [17, p. 181], the minimum eigenvalue of B_n is bounded below by R_*. Thus since the eigenvalues of B_n^{-1} are reciprocals of those of B_n, we have the result

$$\|B_n^{-1}\|_n \leqq R_*^{-1}. \tag{4.23}$$

Thus combining (4.12), (4.18), (4.20) and (4.23) we have

$$\begin{aligned} \|C_n^{-1}B_nC_n^{-1}\|_2 &\leqq \|C_n^{-1}B_n\|_n^2\|B_n^{-1}\|_n \\ &= O((\Delta t)^2). \end{aligned} \tag{4.24}$$

Finally, we see that by the Schwarz inequality, (4.2), and (4.24),

$$\begin{aligned} \|C_n^{-1}\sigma_n\|_n^2 &= (C_n^{-1}B_nC_n^{-1}\sigma_n, \sigma_n) \\ &\leqq \|C_n^{-1}B_nC_n^{-1}\|_2\|\sigma_n\|_2^2 \\ &\leqq L(\Delta t)^2((\Delta x)^2 + (\Delta t)^2)^2. \end{aligned} \tag{4.25}$$

Thus,

$$\frac{1}{\Delta t}\|C_n^{-1}\sigma_n\|_n = O((\Delta x)^2 + (\Delta t)^2). \tag{4.26}$$

From (4.26), the analysis of Douglas [9, pp. 41–44] shows that

$$\|z_n\|_n = O((\Delta x)^2 + (\Delta t)^2), \tag{4.27}$$

and we have convergence in the $\|\cdot\|_2$ of order $O((\Delta x)^2 + (\Delta t)^2)$. We summarize this result in the following theorem.

THEOREM 4.2. *If $B_* < 0$ and $Q_* < 0$, under the restrictions on u, a_k, b_k, h, q and r indicated in* (2.8), *and with the restriction on Δt in* (3.38), *then the solution of* (3.1) *converges to the solution of* (4.1) *in $\|\cdot\|_2$. The rate of convergence is* $O((\Delta x)^2 + (\Delta t)^2)$.

COROLLARY 4.3. *If $B_* > 0$ and $Q_* > 0$ the above result holds with no restriction on $\Delta t > 0$.*

Similar results hold for other choices of B_* and Q_*.

We have shown convergence in the discrete L^2-norm. If multilinear interpolation is applied to the solution, the error in the integral L^2-norm,

$$\|u\|_{L^2} = \left(\int_S \int_0^T u^2 \,dt\,dx\right)^{1/2} \tag{4.28}$$

is also $O((\Delta x)^2 + (\Delta t)^2)$ (see [7]).

5. Convergence for semilinear equations. Consider the semilinear initial-boundary value problem given by (4.1) with qu replaced by q, where

$$q = q(x_1, \cdots, x_m, u, t).$$

For this case we need the added assumption that q is boundedly differentiable with respect to u for all $(x_1, \cdots, x_m, t)$ in $S \times (0, T]$, $-\infty < u < \infty$. We thus assume there are $\bar{Q}_*$ and $\bar{Q}^*$ satisfying

$$\bar{Q}_* \leqq \partial q/\partial u \leqq \bar{Q}^*. \tag{5.1}$$

As before, we shall have a consistent approximation if we define the Crank–Nicolson difference approximation to be

$$\begin{aligned}&\frac{(\Delta_x[a_{n+1/2}\Delta_x w_{n+1}])_\alpha - (\Delta_x[a_{n+1/2}\Delta_x w_n])_\alpha}{\Delta t} + \frac{(\Delta_x[b_{n+1/2}\Delta_x w_{n+1}])_\alpha}{2}\\ &+ \frac{(\Delta_x[b_{n+1/2}\Delta_x w_n])_\alpha}{2} - q\left(x_\alpha, \frac{w_{\alpha,n+1} + w_{\alpha,n}}{2}, t_{n+1/2}\right)\\ &= \frac{r_{\alpha,n+1/2}[w_{\alpha,n+1} - w_{\alpha,n}]}{\Delta t}, \qquad x_\alpha \in S_h,\end{aligned} \tag{5.2a}$$

$$w_{\alpha,0} = f_\alpha, \qquad x_\alpha \in S_h, \tag{5.2b}$$

$$w_{\alpha,n+1} = g_{\alpha,n+1}, \qquad x_\alpha \in \partial S_h. \tag{5.2c}$$

After a standard linearization using the mean value theorem, we see that the error equation for the semilinear case is of the same form as that studied in § 4. Just as before, we obtain the following theorem.

THEOREM 5.1. *If $B_* < 0$ and $\bar{Q}_* < 0$, under the restrictions on u, a_k, b_k, h, q and r indicated in* (2.8), *the added restriction that q has a bounded derivative with respect to u for $(x_1, \cdots, x_m, t) \in S \times (0, T]$ and $-\infty < u < \infty$, and the restriction on Δt in* (3.38), *then the solution of the Crank–Nicolson difference approximation* (5.2) *converges to the solution u of* (4.1), *as modified, in the discrete L^2-norm with discretization error of the form $O((\Delta x)^2 + (\Delta t)^2)$.*

COROLLARY 5.2. *If $B_* > 0$ and $\bar{Q}_* > 0$, the above result holds with no restriction on $\Delta t > 0$.*

Similar results hold for other choices of B_* and $\bar{Q}_*$. Also, as in the previous section, the above results can be extended to convergence in the integral L^2-norm (4.28) by multilinear interpolation [7].

6. Algebraic problem. Our original problem was to find a numerical approximation v of the solution u to a problem like (4.1). We must now solve the algebraic problem by showing how to determine a numerical approximation v_n of the solution w_n of (5.2).

The method of solving the algebraic problem for (5.2) will be to reduce the problem to the inversion of a certain matrix at each time level. Thus we will fix a Δt and consider a separate problem at each time level. If $B_* < 0$ and $\bar{Q}_* < 0$, we have an eigenvalue problem which will force restrictions on the fixed Δt to be described later.

In order to treat the nonlinear algebraic problem we shall use a two-level iteration scheme. We present two schemes. The first utilizes a Picard-type outer iteration with an alternating direction inner iteration, while the second replaces the inner iteration by an overrelaxation scheme. The first converges more rapidly, but applies to less general problems [27]. See part II of [12] for a more detailed account of this section.

First, let $m = 3$, let S be a cube in $\mathbb{R}^3$ and consider the Crank–Nicolson difference system (5.2) where $a_k = a$ and $b_k = b$ for $k = 1, 2, 3$ with a and b constants. For each fixed n, we consider $w_{n+1} \equiv w$ a variable. Then for each fixed time level, we have the problem

$$\Delta_3 w_\alpha = Q(x_\alpha, w_\alpha), \qquad (x_1, x_2, x_3) \in S_h, \quad t = t_{n+1}, \tag{6.1a}$$

$$w_\alpha = g_{\alpha,n+1}, \qquad (x_1, x_2, x_3) \in \partial S_h, \tag{6.1b}$$

where

$$Q(x_\alpha, w_\alpha) = \frac{2(r_{n+1/2}w)_\alpha + 2\Delta t q_{n+1/2}(x_\alpha, w_\alpha)}{2a + b\Delta t} + d_{\alpha,n}, \tag{6.2a}$$

$$d_{\alpha,n} = \left(\frac{2a - b\Delta t}{2a + b\Delta t}\right)\Delta_3 w_{\alpha,n} - \frac{2(r_{n+1/2}w_n)_\alpha}{2a + b\Delta t}, \tag{6.2b}$$

and Δ_3 is the 3-dimensional discrete Laplacian. Clearly d is independent of $w \equiv w_{n+1}$. We need upper and lower bounds for $\partial Q/\partial w$. Using (2.8) and (5.1) we obtain

$$p \equiv \frac{2R_* + 2\bar{Q}_*\Delta t}{2a + b\Delta t} < \frac{\partial Q}{\partial w} < \frac{2R^* + 2\bar{Q}^*\Delta t}{2a + b\Delta t} \equiv P. \tag{6.3}$$

We thus consider solving the nonlinear algebraic equations (6.1) with the condition (6.3).

The problem (6.1)–(6.3) is now exactly in the form of the problem discussed by Douglas in [10] and in [11, pp. 59–60]. The outer iteration for the two-level scheme is a Picard-type iteration [10], [11]. The inner, alternating direction, iteration and the parameter sequence needed to use it are given in [11]. The following theorem follows directly from the argument of [10].

THEOREM 6.1. *Given p from* (6.3), *if* $p > -3\pi^2$ (the negative of the least eigenvalue of the Laplace differential operator on S, the unit cube), *the solution of the iteration process defined in* [10], [11] *converges to the solution of* (6.1). *The number of calculations required to reduce the error in the solution by a factor of* ε *is*

$$O((\Delta x)^{-3}\log(\Delta x)^{-1}). \tag{6.4}$$

Now we want to consider the algebraic problem for more general operators from (4.1) as modified in § 5. Replacing the inner iteration by a successive overrelaxation scheme, we are able to treat these more general operators. From (5.2) we consider, for S a cube in $\mathbb{R}^m$,

$$\begin{aligned} &\frac{(\Delta_x[a_{n+1/2}\Delta_x w_{n+1}])_\alpha}{\Delta t} + \frac{(\Delta_x[b_{n+1/2}\Delta_x w_{n+1}])_\alpha}{2} \\ &\quad = \frac{(r_{n+1/2}w_{n+1})_\alpha}{\Delta t} + q\left(x_\alpha, \frac{w_{\alpha,n+1} + w_{\alpha,n}}{2}, t_{n+1/2}\right) + C(x_\alpha, w_{\alpha,n}, t_{n+1/2}), \end{aligned} \tag{6.5}$$

where C is independent of w_{n+1}. For each n, as above, we let $w_{n+1} \equiv w$ and consider a separate problem. Fix n and consider

$$\left\{\left[\frac{(\Delta_x[a_{n+1/2}\Delta_x])/\Delta t + (\Delta_x[b_{n+1/2}\Delta_x])}{2}\right]w\right\}_\alpha = Q(x_\alpha, w_\alpha), \qquad x_\alpha \in S_h, \tag{6.6a}$$

$$w_\alpha = g_{\alpha,n+1}, \qquad x_\alpha \in \partial S_h, \tag{6.6b}$$

where

$$Q(x_\alpha, w_\alpha) = (r_{n+1/2}w)_\alpha/\Delta t + q_{n+1/2}(x_\alpha, w_\alpha) + C_{\alpha,n+1}, \tag{6.7a}$$

$$C_{\alpha,n+1} = \left\{\left[\frac{\Delta_x[a_{n+1/2}\Delta_x]}{\Delta t} - \frac{\Delta_x[b_{n+1/2}\Delta_x]}{2}\right]w_n\right\}_\alpha - (r_{n+1/2}w_n)_\alpha. \tag{6.7b}$$

Clearly C is independent of $w \equiv w_{n+1}$. Thus from (2.8) and (5.1) we see that

$$p \equiv \frac{R_*}{\Delta t} + \bar{Q}_* \leqq \frac{\partial Q}{\partial w} \leqq \frac{R^*}{\Delta t} + \bar{Q}^* \equiv P. \tag{6.8}$$

As before we consider a two-level iteration. The outer iteration is as before, while the inner iteration is a successive overrelaxation iteration [15], [27], [29]. The outer iteration is actually of the form

$$\begin{aligned}&\left\{\left[\frac{\Delta_x[a_{n+1/2}\Delta_x]}{\Delta t} + \frac{\Delta_x[b_{n+1/2}\Delta_x]}{2} - A\right]w^{(k+1)}\right\}_\alpha \\ &\quad = Q(x_\alpha, w_\alpha^{(k)}) - Aw_\alpha^{(k)} + \sigma_\alpha^{(k)}, \qquad x_\alpha \in S_h,\end{aligned} \tag{6.9a}$$

$$w_\alpha^{(k+1)} = g_\alpha, \qquad x_\alpha \in \partial S_h, \tag{6.9b}$$

where $\sigma_\alpha^{(k)}$ is the residual at the end of the inner iteration. As in [10] it is not necessary that the linear equations be solved exactly.

Consider the convergence of $w^{(k)}$ to w. Let

$$z_\alpha^{(k)} = w_\alpha^{(k+1)} - w_\alpha^{(k)}, \qquad k = 1, 2, \cdots. \tag{6.10}$$

Using (6.10) and the mean value theorem,

$$\begin{aligned}&\left\{\left[\frac{\Delta_x[a_{n+1/2}\Delta_x]}{\Delta t} + \frac{\Delta_x[b_{n+1/2}\Delta_x]}{2} - A\right]z^{(k)}\right\}_\alpha \\ &\quad = \frac{\partial Q}{\partial w}(x_\alpha, w_\alpha^*)z_\alpha^{(k-1)} - Az_\alpha^{(k-1)} + \sigma_\alpha^{(k)} - \sigma_\alpha^{(k-1)}, \qquad x_\alpha \in S_h,\end{aligned} \tag{6.11a}$$

$$z_\alpha^{(k)} = 0, \qquad x_\alpha \in \partial S_h. \tag{6.11b}$$

Separating variables, we have the eigenvalue problem

$$E\phi = vF\phi, \qquad x_\alpha \in S_h, \tag{6.12a}$$

$$\phi = 0, \qquad x_\alpha \in \partial S_h, \tag{6.12b}$$

where

$$(E\phi)_\alpha = \left(\frac{\partial Q}{\partial w}(x_\alpha, w_\alpha^*)\phi\right)_\alpha - A\phi_\alpha, \tag{6.13a}$$

$$(F\phi)_\alpha = \left\{\left[\frac{\Delta_x[a_{n+1/2}\Delta_x]}{\Delta t} + \frac{\Delta_x[b_{n+1/2}\Delta_x]}{2} - A\right]\phi\right\}_\alpha. \tag{6.13b}$$

Clearly E and F are symmetric matrices. We can define a consistent ordering and obtain an $N \times N$ matrix representation for F which is diagonally dominant

[30], [32] and positive definite by restricting

$$\Delta t < 2A_*/|B_*| \tag{6.14}$$

if $B_* < 0$. No such restriction is necessary if $B_* \geqq 0$.

Thus from [4, pp. 37–41], we know that (6.12) has a complete set of eigenvectors which are orthogonal with respect to the inner product $(\mathrm{F}\phi, \phi)$ as in (3.11), and we can apply the Courant minimax principle as before. As in part II of [12] we can get bounds on the eigenvalues of (6.12) by considering the eigenvalue problem

$$|\bar{Q} - A|\phi = \lambda\left\{\left[\frac{A_*}{\Delta t} + \frac{B_*}{2}\right]\sum_{k=1}^{m} \Delta_{x_k}^2 \phi - A\phi\right\}, \qquad x_\alpha \in S_h, \tag{6.15a}$$

$$\phi = 0, \qquad x_\alpha \in \partial S_h. \tag{6.15b}$$

As in § 3, by rearranging the above eigenvalue problem and applying the Courant minimax principle and a special case of the minimax theorem [17, p. 181], we have for a cube

$$\max_{i=1,\dots,N} |v^{(i)}| \leqq \frac{|\bar{Q} - A|}{([A_*/\Delta t + B_*/2](m - \delta)\pi^2 + A)}, \tag{6.16}$$

where $|\bar{Q} - A|$ is given by

$$|\bar{Q} - A| = \max\{|\bar{Q}_* - A|, |\bar{Q}^* - A|\}. \tag{6.17}$$

Thus from (6.11), (6.12) and (6.16), we see that

$$\begin{aligned} \|z^{(k)}\|_2 \leqq{}& \frac{|\bar{Q} - A|}{([A_*/\Delta t + B_*/2](m - \delta)\pi^2 + A)}\|z^{(k-1)}\|_2 \\ &+ \left(\left[\frac{A_*}{\Delta t} + \frac{B_*}{2}\right](m - \delta)\pi^2 + A\right)^{-1}(\|\sigma^{(k)}\|_2 + \|\sigma^{(k-1)}\|_2). \end{aligned} \tag{6.18}$$

Thus letting

$$\rho = \frac{|\bar{Q} - A|}{([A_*/\Delta t + B_*/2](m - \delta)\pi^2 + A)} \tag{6.19}$$

and requiring sufficient iterations on the inner iteration so that

$$\|\sigma^{(k)}\|_2 < \frac{[A_*/\Delta t + B_*/2](m - 1)\pi^2}{1 + \rho}\rho^{k+1}, \tag{6.20}$$

we see that

$$\|z^{(k)}\|_2 \leqq \rho\|z^{(k-1)}\|_2 + \rho^k. \tag{6.21}$$

For convergence, it is necessary that

$$|\bar{Q} - A| < \left[\frac{A_*}{\Delta t} + \frac{B_*}{2}\right]m\pi^2 + A. \tag{6.22}$$

One sufficient condition is that

$$p = \frac{R_*}{\Delta t} + \bar{Q}_* > -\left(\frac{A_*}{\Delta t} + \frac{B_*}{2}\right) m\pi^2. \tag{6.23}$$

Since $R_* > 0$, this restriction can be met by taking Δt sufficiently small. We note that (6.23) will hold with no restriction on Δt if $\bar{Q}_* \geqq 0$ and $B_* \geqq 0$.

Now consider the inner, SOR iteration. Each inner iteration consists of approximating the solution of an elliptic difference equation of the form (6.9) by an SOR method. After defining a consistent ordering, as in [30, p. 108], we can obtain a system of N linear equations in N unknowns where for $i, j = 1, 2, \cdots, N$ we have a system of the form

$$\sum_{j=1}^{N} a_{i,j} v_j + d_j = 0. \tag{6.24}$$

As before, we see that $C = (a_{i,j})$ is positive definite under the restriction

$$\Delta t < 2A_*/|B_*| \tag{6.25}$$

with $B_* < 0$ and with no restriction on Δt if $B_* \geqq 0$.

Now we define the following iterative scheme for the elements of C. Using square brackets to distinguish an index of the inner iteration from the outer iteration we have

$$v_i^{[k+1]} = \omega\left\{\sum_{j=1}^{i-1} b_{i,j} v_j^{[k+1]} + \sum_{j=i+1}^{N} b_{i,j} v_j^{[k]} + c_i\right\} - (\omega - 1)v_i^{[k]}, \tag{6.26}$$
$$k \geqq 0, \quad i = 1, 2, \cdots, N,$$

where $v_i^{[0]}$ is arbitrary, $i = 1, 2, \cdots, N$, and where

$$b_{i,j} = \begin{cases} -a_{i,j}/a_{i,i}, & i \neq j, \\ 0, & i = j, \end{cases} \tag{6.27}$$

and

$$c_i = -d_i/a_{i,i}, \qquad i = 1, 2, \cdots, N. \tag{6.28}$$

Equation (6.26) may be written in the form

$$v^{[k+1]} = L_\omega[v^{[k]}] + f, \qquad k \geqq 0, \tag{6.29}$$

where $v^{[k]} = (v_1^{[k]}, v_2^{[k]}, \cdots, v_N^{[k]})$, $f = (f_1, f_2, \cdots, f_N)$, f is fixed, and L_ω denotes a linear operator. Here ω denotes the relaxation factor. Young [32] defines an optimal relaxation factor, ω_b, in terms of the spectral norm $\bar{\mu}$ of $B = (b_{i,j})$ from (6.27). As shown in [32], it is better to overestimate ω_b. By restricting Δt as in (6.25) if necessary, we can obtain, as in [31], a rigorous upper bound for $\bar{\mu}$ and thus a nontrivial upper bound ω_0 for ω_b. With this upper bound ω_0 as a relaxation factor, the rate of convergence $R(L_{\omega_0})$ will be of the same order as $R(L_{\omega_b})$.

As in [30], [32], we see that the number of cycles of SOR iteration sweeps required for each outer iteration is $O((\Delta x)^{-1})$. Thus since the number of calculations per sweep is $O((\Delta x)^{-m})$, we have the following result.

THEOREM 6.2. *By restricting Δt to satisfy* (6.23) *and* (6.25) *when $B_* < 0$ and $\bar{Q}_* < 0$ and if* (6.20) *is satisfied, then the solution of the iteration process defined by* (6.9) *and* (6.26)–(6.28) *converges to the solution of* (6.6). *The number of calculations required to reduce the error in the solution by a factor of ε is*

$$O((\Delta x)^{-(m+1)}). \tag{6.30}$$

COROLLARY 6.3. *The above result holds with no restriction on Δt if $B_* > 0$ and $\bar{Q}_* > 0$.*

Similar results hold for other choices of B_* and $\bar{Q}_*$. We note that since the SOR iteration procedure does not require a rectangular region as does the alternating direction method, the region could be generalized to the arbitrary region described in § 2 as in [30].

REFERENCES

[1] G. BARENBLATT, I. ZHELTOV AND I. KOCHINA, *Basic concepts in the theory of seepage of homogeneous liquids in fissured rocks*, J. Appl. Math. Mech., 24 (1960), pp. 1286–1303.

[2] P. J. CHEN AND M. E. GURTIN, *On a theory of heat conduction involving two temperatures*, Z. Angew. Math. Phys., 19 (1968), pp.614–627.

[3] B. D. COLEMAN AND W. NOLL, *An approximation theorem for functionals, with applications to continuum mechanics*, Arch. Rational Mech. Anal., 6 (1960), pp. 355–370.

[4] R. COURANT AND D. HILBERT, *Methods of Mathematical Physics*, vol. 1, Interscience, New York, 1953.

[5] J. CRANK AND R. NICOLSON, *A practical method for numerical evaluation of solutions of partial differential equations of heat-conduction type*, Proc. Cambridge Philos. Soc., 43 (1947), pp. 50–67.

[6] P. L. DAVIS, *A quasilinear parabolic and a related third order problem*, J. Math. Anal. Appl., 40 (1972), pp. 327–335.

[7] J. DOUGLAS, JR., *On the relation between stability and convergence in the numerical solution of linear parabolic and hyperbolic differential equations*, J. Soc. Indust. Appl. Math., 4 (1956), pp. 20–37.

[8] ———, *The application of stability analysis in the numerical solution of quasi-linear parabolic differential equations*, Trans. Amer. Math. Soc., 89 (1958), pp. 484–518.

[9] ———, *A survey of numerical methods for parabolic differential equations*, Advances in Computers, vol. II, Academic Press, New York, 1961.

[10] ———, *Alternating direction iteration for mildly nonlinear elliptic difference equations*, Numer. Math., 3 (1961), pp. 92–98.

[11] ———, *Alternating direction methods for three space variables*, Ibid., 4 (1962), pp. 41–63.

[12] R. E. EWING, *The numerical approximation of certain parabolic equations backward in time via Sobolev equations*, Doctoral thesis, Univ. of Texas at Austin, Austin, 1974.

[13] ———, *The approximation of certain parabolic equations backward in time by Sobolev equations*, SIAM J. Math. Anal., 6 (1975), pp. 283–294.

[14] W. H. FORD, *Numerical solution of pseudo-parabolic partial differential equations*, Doctoral thesis, Univ. of Illinois at Urbana-Champaign, Urbana, 1972.

[15] S. P. FRANKEL, *Convergence rates of iterative treatments of partial differential equations*, Math. Tables Aids. Comput., 4 (1950), pp. 65–75.

[16] H. GAJEWSKI AND K. ZACHARIAS, *Zur Starken Konvergenz des Galerkinverfahrens bei einer Klasse Pseudoparabolischer Partialler Differentialgleichungen*, Math. Nach., 47 (1970), pp. 365–376.

[17] P. R. HALMOS, *Finite Dimensional Vector Spaces*, Van Nostrand, Princeton, N.J., 1958.

[18] R. HUILGOL, *A second order fluid of the differential type*, Internat. J. Nonlinear Mech., 3 (1968), pp. 471–482.

[19] A. R. MITCHELL, *Computational Methods in Partial Differential Equations*, John Wiley, London, 1969.

[20] W. E. Milne, *Numerical Solutions of Differential Equations*, John Wiley, New York, 1953.

[21] R. E. Showalter, *Weak solutions of nonlinear evolution equations of Sobolev–Galpern type*, J. Differential Equations, 11 (1972), pp. 252–265.

[22] ———, *Partial differential equations of Sobolev–Galpern type*, Pacific J. Math., 31 (1969), pp. 787–793.

[23] ———, *The final-value problem for evolution equations*, J. Math. Anal. Appl., to appear.

[24] R. E. Showalter and T. W. Ting, *Pseudo-parabolic partial differential equations*, SIAM J. Math. Anal., 1 (1970), pp. 1–26.

[25] D. Taylor, *Research on Consolidation of Clays*, Massachusetts Institute of Technology Press, Cambridge, Mass., 1952.

[26] T. W. Ting, *Certain non-steady flows of second order fluids*, Arch. Rational Mech. Anal., 14 (1963), pp. 1–26.

[27] R. S. Varga, *Matrix Iterative Analysis*, Prentice-Hall, Englewood Cliffs, N.J., 1962.

[28] K. Yosida, *Functional Analysis, Die Grundlehren der Mathematischen Wissenschaften in Einzeldarstellungen*, 123, Springer-Verlag, Berlin–Heidelberg–New York, 1965.

[29] David M. Young, *Iterative methods for solving partial difference equations of elliptic type*, Doctoral thesis, Harvard Univ., Cambridge, Mass., 1950.

[30] ———, *Iterative methods for solving partial difference equations of elliptic type*, Trans. Amer. Math. Soc., 76 (1954), pp. 92–111.

[31] ———, *A bound for the optimum relaxation factor for the successive overrelaxation method*, Numer. Math., 16 (1971), pp. 408–413.

[32] David M. Young and Robert T. Gregory, *A Survey of Numerical Mathematics*, vol. II, Addison-Wesley, Reading, Mass., 1973.