A Posteriori Bounds in the Numerical Solution of Mildly Nonlinear Parabolic Equations*

By Alfred Carasso

Abstract. We derive a posteriori bounds for \((V - \hat{V})\) and its difference quotient \((V - \hat{V})_\xi\), where \(V\) and \(\hat{V}\) are, respectively, the exact and computed solution of a difference approximation to a mildly nonlinear parabolic initial boundary problem, with a known steady-state solution. It is assumed that the computation is over a long interval of time. The estimates are valid for a class of difference approximations, which includes the Crank-Nicolson method, and are of the same magnitude for both \((V - \hat{V})\) and \((V - \hat{V})_\xi\).

1. Introduction. Let \(\Omega\) be the strip \(\{(x, t) \mid 0 < x < 1, t > 0\}\) in the \((x, t)\) plane and consider the mixed problem

\[
\begin{align*}
\frac{\partial u}{\partial t} &= \left[ a(x, t)\frac{\partial u}{\partial x}\right]_x + b(x, t)\frac{\partial u}{\partial x} - f(x, t, u), \\
u(x, 0) &= \chi(x), \\
u(0, t) &= \phi_1(t), \\
u(1, t) &= \phi_2(t),
\end{align*}
\]

\((1.1)\) \(0 \leq x \leq 1, \quad t > 0.\)

We assume that \(a(x, t), b(x, t)\) are “smooth” bounded functions on \(\Omega\), with \(a(x, t) \geq a_0 > 0\), and that \(f(x, t, w)\) is, at least once, continuously differentiable on \(\Omega \times [-\infty < w < +\infty]\) with \(\partial f/\partial w \geq 0\). Moreover, \(\partial f/\partial w\) is to remain bounded if \(w\) stays bounded. The coefficients, data, and \(f\) are assumed such as to assure the existence and uniqueness of a solution \(u(x, t)\), four times boundedly differentiable in \(\Omega\), and converging to a steady state value \(u^*(x)\), as \(t \to \infty\). We assume \(u^*(x)\) is known and that, by means of asymptotic formulae and the like, one can estimate \(\|u(\cdot, t) - u^*\|_2\) as a function of \(t\), for \(t\) sufficiently large. The analytical theory for such problems is discussed in Friedman [5].

Several finite-difference methods for the numerical computation of (1.1) have been shown to converge; see for example [4], [6], [8], [10], [3] and their references, and especially [9] for the linear case.

Because of round-off error, and the fact that one may need to use iterative methods at each time step to solve the nonlinear difference equations, only an approximation \(\hat{V}\) to the exact solution \(V\) of the difference equations can be computed in general. In [3], a “boundary-value” method for (1.1) was analyzed. This method yields an a posteriori estimate for \(V - \hat{V}\) by simply computing residuals. In the present note we make use of some of the results in [2] and [3] to derive such an estimate for a class of stable “marching” procedures for (1.1). Unlike the situation in [3], however, the estimate will involve bounds on the derivatives of \(u\). It is interesting that the estimate is of the same magnitude for both \((V - \hat{V})\) and its difference quotient \((V - \hat{V})_\xi\).

Received September 18, 1969.

AMS 1969 subject classifications. Primary 6568, 6580.

Key words and phrases. Parabolic equations, Crank-Nicolson method, limiting steady state, computations over long times.

* Supported by NSF Grant GP-13024.

Copyright © 1971, American Mathematical Society

785
2. Notation. Let \( \overline{\Omega} \) be the rectangle \( \{(x, t) \mid 0 < x < 1, 0 < t < T\} \) and let \( M \) and \( N \) be positive integers. Let \( \Delta x = 1/(M + 1), \Delta t = T/(N + 1), \) and introduce a mesh over \( \overline{\Omega} \) by means of the lines \( x = k\Delta x, \ k = 0, 1, \ldots, M + 1, \ t = n\Delta t, \ n = 0, 1, \ldots, N + 1. \) Let \( v^n_k \) denote \( v(k\Delta x, n\Delta t). \) Define \( V^n \) to be the \( M \)-component vector

\[
V^n = \{v^n_1, v^n_2, \ldots, v^n_M\}^T
\]

and let \( V \) be the "block" vector of \( MN \) components

\[
V = \{v^1, v^2, \ldots, v^N\}^T.
\]

Although we will be dealing with real-valued mesh functions, it is convenient to define scalar products and norms for complex vectors. For any two \( M \) vectors \( V^n, W^n \) let

\[
\langle V^n, W^n \rangle = \Delta x \sum_{k=1}^M v^n_k \overline{w^n_k}
\]

and let

\[
||V^n||^2 = \langle V^n, V^n \rangle.
\]

Let

\[
||v^n_k||^2 = \Delta x \sum_{k=0}^M \frac{|v^n_{k+1} - v^n_k|^2}{\Delta x^2}
\]

where \( v^n_0, v^n_{M+1} \) are defined to be zero.

For block vectors \( V, W \) define

\[
\langle V, W \rangle = \Delta t \sum_{n=1}^N \langle V^n, W^n \rangle
\]

and let

\[
||V||^2 = \langle V, V \rangle,
\]

\[
||V_k||^2 = \Delta t \sum_{n=1}^N ||V^n_k||^2.
\]

Finally, for a square matrix \( A \), define \( ||A|| \) in terms of vector norms, i.e., as

\[
||A|| = \sup_{||x||=1} ||Ax||,
\]

the supremum being taken over all complex vectors.

3. Difference Approximations to (1.1). Let \( U^n \) be the \( M \)-vector consisting of the solution to (1.1) evaluated at the interior mesh points of the line \( t = n\Delta t \) and let \( V^n \) be the corresponding exact solution of the difference equations used to approximate (1.1). Define \( E^n = V^n - U^n \). We will consider the class of marching schemes which lead to a priori estimates of the form

\[
||E^n||^2 + ||E^n_0||^2 \leq K(T)(\Delta t^{r+1} + \Delta x^{s+1}), \quad n\Delta t \leq T,
\]

where \( r \) and \( s \) are positive integers and \( K(T) \) is known. An example of a difference
scheme for (1.1) satisfying (3.1) with \( r = s = 1 \), is the Crank-Nicolson version analyzed in [8]. In general \( K(T) \) will involve bounds on \( a, b, f, u \) and their derivatives, as well as a growth factor. The reason for the latter is that, even if the exact solution to (1.1) decays asymptotically to a steady state, the exact solution of a stable, consistent, difference approximation may grow exponentially as \( n\Delta t \to \infty, \Delta t \) fixed. Hence, we cannot expect \( K(T) \) to remain bounded as \( T \to \infty \), in general. We remark, however, that in [7], Kreiss and Widlund have shown how to construct schemes (for linear time-dependent problems with periodic boundary conditions) which preserve the asymptotic behavior of \( u(x, t) \) provided \( |b| \Delta t / \Delta x < 1 \). In the following we will derive bounds for \( ||\psi - V||_2 \) and \( ||\psi_x - V_x||_2 \) for computations of (1.1) up to some “large” but fixed time \( T \). These bounds will depend on \( K(T) \).

We begin by deriving new finite-difference equations for the exact solution \( \{V^n\} \) of a difference scheme used to approximate (1.1). Since \( V^n = U^n + E^n \), we have

\[
\frac{v_k^{n+1} - v_k^{n-1}}{2\Delta t} = \frac{u_k^{n+1} - u_k^{n-1}}{2\Delta t} + \frac{e_k^{n+1} - e_k^{n-1}}{2\Delta t}
\]

(3.2)

\[
= \frac{\partial u^n}{\partial t} + \frac{e_k^{n+1} - e_k^{n-1}}{2\Delta t} + \frac{\Delta t^2}{6} \frac{6}{(u_{i+1})},
\]

where “\( \psi \)” represents a mean value of \( \psi \) called for by Taylor’s theorem. From (1.1) we have

\[
\frac{\partial u^n}{\partial t} + \frac{\Delta t^2}{6} \frac{(u_{i+1})}{(u_{i+1})} = \frac{a_{k+1/2}(u_{k+1} - u_k^n)}{\Delta x^2}
\]

(3.3)

\[
\quad + \frac{b_k^n(u_{k+1} - u_k^n)}{2\Delta x} - f(k\Delta x, n\Delta t, u_k^n) + \tau_k^n,
\]

where

\[
\tau_k^n = \frac{\Delta t^2}{6} \frac{(u_{i+1})}{(u_{i+1})}
\]

(3.4)

\[
- \Delta x^2 \left\{ \frac{(u_x)^2(a_{n+1}^n)}{3} + \frac{(u_{xx})^2(a_{n+1}^n)}{2} + \frac{(u_{xxx})^2(a_{n+1}^n)}{6} + \frac{(a_{n+1}^n u_{xxx})}{12} + b_k^n(u_{xxx}) \right\}
\]

From (3.2) and (3.3) we have

\[
\frac{v_k^{n+1} - v_k^{n-1}}{2\Delta t} = \frac{a_{k+1/2}(v_{k+1}^n - v_k^n)}{\Delta x^2}
\]

(3.5)

\[
+ \frac{b_k^n(v_{k+1}^n - v_{k-1}^n)}{2\Delta x} - f(k\Delta x, n\Delta t, u_k^n)
\]

\[
+ \frac{e_k^{n+1} - e_k^{n-1}}{2\Delta t} - \frac{b_k^n(e_{k+1}^n - e_{k-1}^n)}{2\Delta x}
\]

\[
- \frac{(a_{k+1/2}(e_{k+1}^n - e_k^n))}{\Delta x^2} + \frac{(a_{k-1/2}(e_k^n - e_{k-1}^n))}{\Delta x^2} + \tau_k^n,
\]

\[
k = 1, \ldots, M, \quad n = 1, 2, \ldots,
\]

License or copyright restrictions may apply to redistribution; see http://www.ams.org/journal-terms-of-use
with the initial boundary data

\begin{align}
  v_k^0 &= \chi(k\Delta x), \quad k = 1, \cdots, M, \\
  v_0^n &= \varphi_1(n\Delta t), \quad v_{M+1}^n = \varphi_2(n\Delta t), \quad n = 1, 2, \cdots.
\end{align}

With \( T = (N + 1)\Delta t \) we now consider the system formed by equations (3.5) for \( n = 1, 2, \cdots, N \). It is convenient to write this system in matrix-vector notation.

Let \( L^n \) and \( B^n \) be the tridiagonal \( M \times M \) matrices defined by

\begin{align}
  L^n &= \frac{1}{\Delta x^2} \begin{bmatrix}
    (a_{1,1/2}^n + a_{-1,1/2}^n) & -a_{1,1/2}^n & 0 & \cdots & 0 \\
    -a_{1,1/2}^n & (a_{-1,1/2}^n + a_{1,1/2}^n) & -a_{1,1/2}^n & \cdots & 0 \\
    \vdots & \vdots & \ddots & \ddots & \vdots \\
    \vdots & \vdots & \ddots & \ddots & \ddots \\
    0 & \cdots & \cdots & 0 & (a_{M-1,1/2}^n + a_{M+1,1/2}^n)
  \end{bmatrix}, \\
  B^n &= \frac{1}{2\Delta x} \begin{bmatrix}
    b_2^n & \cdots & 0 \\
    \vdots & \ddots & \vdots \\
    0 & \cdots & b_M^n \\
    0 & \cdots & 0 & -b_{M-1}^n
  \end{bmatrix}
\end{align}

and define the \( M \)-vectors \( \tau^n \), \( F^n(U) \), and \( G^n \) by

\begin{align}
  \tau^n &= \{ \tau_1^n, \tau_2^n, \cdots, \tau_M^n \}^T, \\
  F^n(U) &= \{ f_1^n(u), f_2^n(u), \cdots, f_M^n(u) \}^T,
\end{align}

where

\[ f_k^n(u) = f(k\Delta x, n\Delta t, u_k^n) \]

and

\[ G^n = \frac{1}{\Delta x^2} \left\{ (a_{1,1/2}^n - \frac{1}{2}\Delta x b_1^n)\varphi_1(n\Delta t), 0, 0, \cdots, 0, (a_{M+1,1/2}^n + \frac{1}{2}\Delta x b_M^n)\varphi_2(n\Delta t) \right\}^T. \]

We may then write (3.5), (3.6) as

\begin{align}
  \frac{V^{n+1} - V^{n-1}}{2\Delta t} &= -L^n V^n - B^n V^n - F^n(U) + \tau^n + G^n \\
  &\quad + \frac{E^{n+1} - E^{n-1}}{2\Delta t} + B^n E^n + L^n E^n, \quad n = 1, 2, \cdots, N.
\end{align}

Some further definitions will enable us to write (3.12) in "block" form. Define the \( MN \times MN \) block tridiagonal matrix \( P \) by (with \( \sigma = 1/2\Delta t \))

\begin{align}
  P &= \begin{bmatrix}
    (L^1 + B^1) & \cdots & \sigma I & 0 \\
    -\sigma I & \ddots & \ddots & \ddots \\
    0 & \cdots & -\sigma I & (L^N + B^N) & \sigma I \\
  \end{bmatrix}.
\end{align}
For any real block vector $\xi$ define the $M \times M$ diagonal matrix $C^*(\xi)$ by

$$
C^*(\xi) = \begin{bmatrix}
f_\omega(\Delta x, n\Delta t, \xi_1^1) & \cdots & 0 \\
\vdots & \ddots & \vdots \\
0 & \cdots & f_\omega(\Delta x, n\Delta t, \xi_M^1)
\end{bmatrix}
$$

and let $C(\xi)$ be the block matrix

$$
C(\xi) = \begin{bmatrix}
C^*(\xi) & 0 \\
0 & C^N(\xi)
\end{bmatrix}.
$$

Finally, define the block vectors $F$, $G^*$, $H$, and $\tau$ by

$$
F = \{F_1, F_2, \ldots, F_N\}^T,
$$

$$
G^* = \left\{G^1 + \frac{V^0}{2\Delta t}, G^2, \ldots, G^N - \frac{V^N}{2\Delta t}\right\}^T,
$$

$$
H = \left\{\frac{E^2 - E^0}{2\Delta t} + (L^1 + B^1)E^1, \ldots, \frac{E^{N+1} - E^{N-1}}{2\Delta t} + (L^N + B^N)E^N\right\}^T,
$$

$$
\tau = \{\tau^1, \tau^2, \ldots, \tau^N\}^T.
$$

With this notation we have from (3.12)

$$
P\nu = -F(U) + G^* + \tau + H.
$$

**Lemma 1.** Let $D$ be a diagonal matrix of order $MN$ with nonnegative real entries and let

$$
Q = P + D.
$$

Let $b(x, t)$ in (1.1) satisfy

$$
\left|\frac{\partial b}{\partial x}\right| \leq b_1 < 2a_0\pi^2, \quad \forall (x, t) \in \overline{a}_T.
$$

Fix $\epsilon > 0$ so that $a_0\pi^2 - b_1/2 - \epsilon \geq \omega > 0$ if $\Delta x \leq (12\epsilon/a_0\pi)\frac{1}{2}$, $Q^{-1}$ exists and

$$
\sup_{x \text{ real}, \|X\|_2 \leq 1} \|Q^{-1}X\|_2 \leq \frac{1}{\omega}.
$$

Moreover, if $QW = Z$, where $Z$ is real we have

$$
\|W\|_2 \leq \left(\frac{2\omega + b_1}{2a_0\omega^2}\right)^{1/2} \|Z\|_2.
$$

**Proof.** See [3, Lemma 1].

**Remark.** If

$$
D = \begin{bmatrix}
\Lambda & \circ & \circ \\
\circ & \Lambda & \cdots \\
\circ & \circ & \cdots & \Lambda
\end{bmatrix},
$$
where $\Lambda$ is a diagonal $M \times M$ matrix with nonnegative real entries, and if $a(x, t)$, $b(x, t)$ are independent of $t$, $Q^{-1}$ exists and remains bounded for all sufficiently small $\Delta x$ independently of hypothesis (3.22). This observation is relevant to the case where (1.1) is linear with time independent coefficients, i.e., $a = a(x)$, $b = b(x)$, and $f(x, t, u) = c(x)u + h(x, t)$ with $c(x) \geq 0$. See [1, Lemma 1] and [2, Lemma 4.2].

4. A Posteriori Bounds. For each $n = 1, 2, \cdots, N + 1$, let $\bar{V}^n$ be the computed solution at $t = n\Delta t$, of the difference equations used to approximate (1.1) and consider the block vector

$$
\bar{V} = \{ \bar{V}^1, \bar{V}^2, \cdots, \bar{V}^N \}^T.
$$

Define $\bar{G}^*$ to be the block vector obtained from $G^*$ when $V^{N+1}$ is replaced by $\bar{V}^{N+1}$.

Compute the block vector $R$ given by

$$
R = P\bar{V} + F(\bar{V}) - \bar{G}^*.
$$

Subtracting (4.2) from (3.20) we have

$$
P(V - \bar{V}) = -F(U) + F(\bar{V}) + (G^* - \bar{G}^*) + \tau + H - R
$$

$$
= -F(U) + F(V) + F(\bar{V}) - F(V) + (G^* - \bar{G}^*) + \tau + H - R
$$

$$
= -C(\xi)(U - V) - C(\Psi)(V - \bar{V}) + (G^* - \bar{G}^*) + \tau + H - R,
$$

for some intermediate real block vectors $\xi$ and $\Psi$ on using the mean value theorem. Hence,

$$
[P + C(\Psi)](V - \bar{V}) = \tau + H - R + (G^* - \bar{G}^*) - C(\xi)(U - V).
$$

Since $f_\omega \geq 0$, $C(\Psi)$ is a diagonal matrix with nonnegative real entries. By Lemma 1, we may estimate $||V - \bar{V}||_2$, $||V_n - \bar{V}_n||_2$, provided we can estimate the terms other than $R$ on the right-hand side of (4.4). We will make use of the a priori estimate (3.1).

Let $a^*$, $b^*$ be upper bounds for $a(x, t)$ and $|b(x, t)|$, respectively, in $\bar{a}_r$.

Since

$$
(L^n E^n)_k = \frac{1}{\Delta x^2} a_{k-1/2} (e_k^*-e_{k-1}^*) + \frac{1}{\Delta x^2} a_{k+1/2} (e_k^*-e_{k+1}^*)
$$

we have

$$
||L^n E^n||_2 \leq \frac{a^*}{\Delta x} \left\{ \Delta x \sum_{k=1}^M \left( \frac{e_k^*-e_{k-1}^*}{\Delta x^2} \right)^2 \right\}^{1/2} + \frac{a^*}{\Delta x} \left\{ \Delta x \sum_{k=1}^M \left( \frac{e_k^*-e_{k+1}^*}{\Delta x^2} \right)^2 \right\}^{1/2}
$$

$$
\leq \frac{2a^*}{\Delta x} ||E^n||_2 \leq 2a^* K(T) \left( \frac{\Delta t^{*+1}}{\Delta x} + \Delta x^* \right).
$$

Similarly,

$$
||B^n E^n||_2 \leq b^* K(T)(\Delta x^{*+1} + \Delta t^{*+1})
$$

and we have

$$
\frac{1}{2\Delta t} ||E^{n+1} - E^{n-1}||_2 \leq K(T) \left( \frac{\Delta t^* + \Delta x^{*+1}}{\Delta t} \right).
$$

Hence, we can estimate $||H||_2$ by
We estimate $||G^* - \hat{G}^*||_2$ as follows: First,

$$||G^* - \hat{G}^*||_2 = \frac{1}{2\Delta t^{1/2}} ||\hat{V}^{N+1} - V^{N+1}||_2.$$  

If $U^*$ is the $M$-vector consisting of the steady state solution, we have from (3.1) 

$$||G^* - \hat{G}^*||_2 \leq \frac{1}{2\Delta t^{1/2}} \{ ||\hat{V}^{N+1} - U^*||_2 + ||U^{N+1} - U^*||_2 \}$$

(4.10)

$$+ K(T) \left( \frac{\Delta t^{r+1} + \Delta x^{s+1}}{2\Delta t^{1/2}} \right).$$

Since we assume $U^*$ is known, and that $\{u(x, t) - u^*(x)\}$ can be estimated as a function of $t$, the right-hand side of (4.10) can be estimated.

We may estimate $||C(\xi)||_2$ by using the a priori estimate (3.1), since $\xi$ is an intermediate value, and since $f_w(x, t, w)$ is bounded if $w$ is bounded. This means we can find a constant $K_1$ such that

(4.11) 

$$||C(\xi)(V - U)||_2 \leq K_1 K(T) T^{1/2} (\Delta t^{r+1} + \Delta x^{s+1}).$$

Finally, we assume a bound is known for the derivatives of $u$ occurring in (3.4) so that

(4.12) 

$$||\tau||_2 \leq T^{1/2} K_2 (\Delta t^2 + \Delta x^2),$$

for some constant $K_2$.

Using Lemma 1 and (4.8), (4.10), (4.11) and (4.12) we have

**Theorem.** Let $b(x, t)$ in (1.1) satisfy $|\partial b/\partial x| \leq b_1 < 2a_0\pi^2$ and fix $\epsilon > 0$ so that

$$a_0\pi^2 - b_1 - \epsilon \geq \omega > 0.$$

Let $V = \{V^m\}$ and $\hat{V} = \{\hat{V}^n\}$ be, respectively, the exact and computed solution of a difference approximation for (1.1) satisfying (3.1). Finally, let $R$ be defined by (4.2). Then

$$||V - \hat{V}||_2 \leq \frac{1}{2\omega \Delta t^{1/2}} \{ ||\hat{V}^{N+1} - U^*||_2 + ||U^{N+1} - U^*||_2 \} + \frac{||R||_2}{\omega}$$

(4.14)

$$+ \frac{T^{1/2} K_2 (\Delta t^2 + \Delta x^2)}{\omega} + \frac{T^{1/2} K(T)(\Delta t^{r+1} + \Delta x^{s+1})}{\omega} \left\{ K_1 + \frac{1}{2(T\Delta t)^{1/2}} + b^* + \frac{2a^*}{\Delta x} + \frac{1}{\Delta t} \right\}$$

and

$$||V_x - \hat{V}_x||_2 \leq \left( \frac{2\omega + b_1}{2a_0\omega^2} \right)^{1/2} \frac{1}{2\Delta t^{1/2}} \{ ||\hat{V}^{N+1} - U^*||_2 + ||U^{N+1} - U^*||_2 \}$$

(4.15)

$$+ ||R||_2 + K_2 T^{1/2} (\Delta t^2 + \Delta x^2) + K(T) T^{1/2} (\Delta t^{r+1} + \Delta x^{s+1}) \left( K_1 + \frac{1}{2(T\Delta t)^{1/2}} + b^* + \frac{2a^*}{\Delta x} + \frac{1}{\Delta t} \right).$$
The University of New Mexico
Albuquerque, New Mexico 87106