SIAM J. Numer. Anal.
Vol. 35, No. 2, pp. 508–523, April 1998

CONVERGENCE ANALYSIS OF PSEUDO-TRANSIENT CONTINUATION*

C. T. KELLEY[†] AND DAVID E. KEYES[‡]

Abstract. Pseudo-transient continuation (Ψtc) is a well-known and physically motivated technique for computation of steady state solutions of time-dependent partial differential equations. Standard globalization strategies such as line search or trust region methods often stagnate at local minima. Ψtc succeeds in many of these cases by taking advantage of the underlying PDE structure of the problem. Though widely employed, the convergence of Ψtc is rarely discussed. In this paper we prove convergence for a generic form of Ψtc and illustrate it with two practical strategies.

Key words. pseudo-transient continuation, nonlinear equations, steady state solutions, global convergence

AMS subject classifications. 65H10, 65J15, 65N12, 65N22

PII. S0036142996304796

1. Introduction.

Pseudo-transient continuation (Ψtc) is a method for computation of steady state solutions of partial differential equations. We shall interpret the method in the context of a method-of-lines solution, in which the equation is discretized in space and the resulting finite dimensional system of ordinary differential equations is written as $x' = F(x)$. Here,

$$x' = \frac{\partial x}{\partial t}$$

and the discretized spatial derivatives are contained in the nonlinear term $F(x)$. Marching out the steady state by following the physical transient may be unnecessarily time consuming if the intermediate states are not of interest. On the other hand, Newton's method for $F(x) = 0$ alone will usually not suffice, as initial iterates sufficiently near the root are usually not available. Standard globalization strategies [12, 17, 25], such as line search or trust region methods often stagnate at local minima of $\|F\|$ [20]. This is particularly the case when the solution has complex features such as shocks that are not present in the initial iterate (see [24], for example). Ψtc succeeds in many of these cases by taking advantage of the PDE structure of the problem.

1.1. The basic algorithm.

In the simple form considered in this paper, Ψtc numerically integrates the initial value problem

$$x' = -V^{-1}F(x),\ x(0) = x_0 \tag{1.1}$$

to steady state using a variable timestep scheme that attempts to increase the timestep as $F(x)$ approaches 0. V is a nonsingular matrix used to improve the scaling of

[*] Received by the editors June 5, 1996; accepted for publication (in revised form) November 18, 1996.

http://www.siam.org/journals/sinum/35-2/30479.html

[†] North Carolina State University, Department of Mathematics and Center for Research in Scientific Computation, Box 8205, Raleigh, NC 27695-8205 (Tim_Kelley@ncsu.edu). The research of this author was supported by National Science Foundation grant DMS-9321938.

[‡] Computer Science Department, Old Dominion University, Norfolk, VA 23519-0162 (keyes@cs.odu.edu) and ICASE, MS 132C, NASA Langley Research Center, Hampton, VA 23681-0001. The research of this author was supported by National Science Foundation grant ECS-9527169 and NASA grants NAGI-1692 and NAS1-19480, the latter while the author was in residence at ICASE.

the problem. It is typically diagonal and approximately equilibrates the local CFL number (based on local cell diameter and local wave speed) throughout the domain. In a multicomponent system of PDEs, not already properly nondimensionalized, V might be a block diagonal matrix, with blocksize equal to the number of components.

We can define the Ψtc sequence $\{x_n\}$ by

$$x_{n+1} = x_n - (\delta_n^{-1} V + F'(x_n))^{-1} F(x_n) \tag{1.2}$$

where F' is the Jacobian (or the Fréchet derivative in the infinite-dimensional situation). Algorithmically, we have the following.

ALGORITHM 1.1.

```
1. Set $x = x_0$ and $\delta = \delta_0$. Evaluate $F(x)$.
2. While $\|F(x)\|$ is too large.
    (a) Solve $(\delta^{-1} V + F'(x))s = -F(x)$.
    (b) Set $x = x + s$.
    (c) Evaluate $F(x)$.
    (d) Update $\delta$.
```

The linear equation

$$(\delta^{-1} V + F'(x))s = -F(x) \tag{1.3}$$

for the Newton step that is solved in step 2a is the discretization of a PDE and is usually very large. As such, it is typically solved by an iterative method which terminates on small linear residuals. This results in an inexact method [11] and a convergence theory for Ψtc must account for this; we provide such a theory in section 3. We have not yet explained how δ is updated nor explored the reuse of Jacobian information from previous iterations. These options must be considered to explain the performance of Ψtc as practiced, and we take them up in section 3, too. In order to most clearly explain the ideas, however, we begin our analysis in section 2 with the most simple variant of Ψtc possible, solving (1.3) exactly, and then extend those results to the more general situation.

As a method for integration in time, Ψtc is a Rosenbrock method ([14], p. 223) if δ is fixed. One may also think of this as a predictor-corrector method, where the simple predictor (result from the previous timestep) and a Newton corrector are used. To see this consider the implicit Euler step from x_n with timestep δ_n,

$$z_{n+1} = x_n - \delta_n V^{-1} F(z_{n+1}). \tag{1.4}$$

In this formulation z_{n+1} would be the root of

$$G(\xi) = \xi + \delta_n V^{-1} F(\xi) - x_n.$$

Finding a root of G with Newton's method would lead to the iteration

$$\xi_{k+1} = \xi_k - (I + \delta_n V^{-1} F'(\xi_k))^{-1} (\xi_k + \delta_n V^{-1} F(\xi_k) - x_n).$$

If we take $\xi_0 = x_n$, the first Newton iterate is

$$\xi_1 = x_n - (I + \delta_n V^{-1} F'(x_n))^{-1} \delta_n V^{-1} F(x_n) = x_n - (\delta_n^{-1} V + F'(x_n))^{-1} F(x_n).$$

This leaves open the possibility of taking more corrector iterations, which would lead to a different form of Ψtc than that given by (1.2). This may improve stability for some problems [16].

1.2. Timestep control. We assume that δ_n is computed by some formula like the "switched evolution relaxation" (SER) method, so named in [21], and used in, e.g., [19], [24], and [33]. In its simplest, unprotected form, SER increases the timestep in inverse proportion to the residual reduction.

$$\delta_n = \delta_0 \|F(x_0)\| / \|F(x_n)\|. \tag{1.5}$$

Relation (1.5) implies that, for $n \geq 1$,

$$\delta_n = \delta_{n-1} \frac{\|F(x_{n-1})\|}{\|F(x_n)\|}.$$

In some work [16], δ_n is kept below a large, finite bound δ_{max}. Sometimes δ_n is set to ∞ (called "switchover to steady state form" in [13]) when the computed value of δ_n exceeds δ_{max}. In view of these practices, we will allow for somewhat more generality in the formulation of the sequence $\{\delta_n\}$. We will assume that δ_0 is given and that

$$\delta_n = \phi\left(\delta_{n-1} \frac{\|F(x_{n-1})\|}{\|F(x_n)\|}\right) \tag{1.6}$$

for $n \geq 1$. The choice in [24] and [33] (1.5) is

$$\phi(\xi) = \xi.$$

Other choices could either limit the growth of δ or allow δ to become infinite after finitely many steps. Our formal assumption on ϕ accounts for all of these possibilities.

Assumption 1.1.

1.

$$\phi(\xi) = \left\{ \begin{array}{ll} \xi & \xi \leq \xi_t \\ \delta_{max} & \xi > \xi_t \end{array} \right. . \tag{1.7}$$

2. Either $\xi_t = \delta_{max}$ or $\xi_t < \infty$ and $\delta_{max} = \infty$.

So, if $\xi_t = \infty$ then $\phi(\xi) = \xi$. If $\xi_t = \delta_{max} < \infty$ then

$$\phi(\xi) = \min(\xi, \delta_{max})$$

and the timesteps are held bounded. If $\xi_t < \infty$ and $\delta_{max} = \infty$ then switchover to steady state form is permitted after a finite number of timesteps.

In [16] the timesteps are based not on the norms of the nonlinear residuals $\|F(x_n)\|$ but on the norms of the steps $\|x_n - x_{n-1}\|$. This policy has risks in that small steps need not imply small residuals or accurate results. However, if the Jacobians are uniformly well conditioned, then small steps do imply that the iteration is near a root. Here formulae of the type

$$\delta_n = \phi(\delta_{n-1} \|x_n - x_{n-1}\|^{-1}) \tag{1.8}$$

are used, where ϕ satisfies Assumption 1.1.

1.3. Iteration phases. We divide the Ψtc iteration into three conceptually different and separately addressed phases.

1. *The initial phase.* Here δ is small and x is far from a steady state solution. This phase is analyzed in section 2.3. Success in this phase is governed by stability and accuracy of the temporal integration scheme and proper choice of initial data.

2. *The midrange.* This begins with an accurate solution x and a timestep δ that may be small and produces an accurate x and a large δ. We analyze this in section 2.2. To allow δ to grow without loss of accuracy in x we make a linear stability assumption (part 3 of Assumption 2.1).
3. *The terminal phase.* Here δ is large and x is near a steady state solution. This is a small part of the overall process, usually requiring only a few iterations. Aside from the attention that must be paid to the updating rules for δ, the iteration is a variation of the chord method [25, 17].

We analyze the terminal phase first, as it is the easiest, in section 2.1. Unlike the other two phases, the analysis of the terminal phase does not depend on the dynamics of $x' = -V^{-1}F(x)$. The initial and midrange phases are considered in section 2.3, with the midrange phase considered first to motivate the goals of the initial phase. This decomposition is similar to that proposed for GMRES and related iterations in [22] and is supported by the residual norm plots reported in [24, 10].

2. Exact Newton iteration.

In this section, we analyze the three phases of the solver in reverse order. This ordering makes it clear how the output of an earlier phase must conform to the demands of the later phase.

2.1. Local convergence: Terminal phase.

The terminal phase of the iteration can be analyzed without use of the differential equation at all.

LEMMA 2.1. *Let $\{\delta_n\}$ be given by either* (1.6) *or* (1.8) *and let Assumption* 1.1 *hold. Let $F(x^*) = 0$, $F'(x^*)$ be nonsingular, and F' be Lipschitz continuous with Lipschitz constant γ in a ball of radius ϵ about x^*.*

Then there are $\epsilon_1 > 0$ and Δ_0, such that if $\delta_{max}, \delta_0 > \Delta_0$ and $\|x_0 - x^*\| < \epsilon_1$, then the sequence defined by (1.2) *and* (1.6) *satisfies*

$$\delta_n \to \delta_{max},$$

and $x_n \to x^*$ q-superlinearly if $\delta_{max} = \infty$ and q-linearly if $\delta_{max} < \infty$.

Proof. Let $e = x - x^*$ denote the error. As is standard, [12, 17], we analyze convergence in terms of the transition from a current iterate x_c to a new iterate x_+. We must also keep track of the change in δ and will let δ_c and δ_+ be the current and new pseudotimesteps.

The standard analysis of the chord method ([17], p. 76) implies that there are $\epsilon_1 \le \epsilon$ and K_C such that if $\|e_c\| < \epsilon_1$,

$$\|e_+\| \le K_C(\|e_c\|^2 + \delta_c^{-1}\|e_c\|). \tag{2.1}$$

The constant K_C depends only on ϵ_1, F, and x^* and does not increase if ϵ_1 is reduced.

Now let Δ_0^{-1} and ϵ_1 be small enough to satisfy

$$\epsilon_1 + {\Delta_0}^{-1} \le 1/(2K_C).$$

Then if $\delta_c \ge \Delta_0$, $\|e_+\| \le \|e_c\|/2$ and, in particular, $F(x_+)$ is defined. Reduce ϵ_1 and increase Δ_0 if needed so that

$$\epsilon_1 + {\Delta_0}^{-1} \le 1/(8\kappa(F'(x^*))) \text{ and } \epsilon_1 \le 3, \tag{2.2}$$

where κ denotes condition number. Equations (2.1) and (2.2) imply that

$$\|e_+\| \le \|e_c\|(\epsilon_1 + {\Delta_0}^{-1}) \le \frac{\|e_c\|}{\kappa(F'(x^*))}. \tag{2.3}$$

If $\{\delta_n\}$ is computed with (1.6) we use the following inequality from [17] (p. 72)

$$\frac{\|e_+\|}{4\kappa(F'(x^*))\|e_c\|} \le \frac{\|F(x_+)\|}{\|F(x_c)\|} \le \frac{4\kappa(F'(x^*))\|e_+\|}{\|e_c\|}, \tag{2.4}$$

and (2.3) to obtain

$$\frac{\|F(x_c)\|}{\|F(x_+)\|} \ge \frac{\|e_c\|}{4\kappa(F'(x^*))\|e_+\|} \ge 2.$$

We then have by Assumption 1.1 that

$$\delta_+ = \phi(\delta_c\|F(x_c)\|/\|F(x_+)\|) \ge \phi(2\delta_c) \ge \left\{ \begin{array}{ll} \delta_{max}, & 2\delta_c \ge \xi_t \\ 2\delta_c, & 2\delta_c < \xi_t \end{array} \right.,$$

where ξ_t is from Assumption 1.1.

If $\{\delta_n\}$ is computed with (1.8), we note that

$$\|x_+ - x_c\| \le \|e_+\| + \|e_c\| \le 3\|e_c\|/2 \le 3\epsilon_1/2 \le 1/2$$

and, hence,

$$\delta_+ = \phi(\delta_c/\|x_+ - x_c\|) \ge \phi(2\delta_c) \ge \left\{ \begin{array}{ll} \delta_{max}, & 2\delta_c \ge \xi_t \\ 2\delta_c, & 2\delta_c < \xi_t \end{array} \right.,$$

as before.

In either case, $\delta_+ \ge \delta_c \ge \Delta_0$ and $\|e_+\| \le \|e_c\|/2$. Therefore, we may continue the iteration and conclude that $\delta_n \to \delta_{max}$ and $\|e_n\| \to 0$ at least q-linearly with q-factor of 1/2.

If $\delta_{max} = \infty$, we complete the proof by observing that since $\delta_n \to \infty$ and $x_n \to x^*$ q-superlinear convergence follows from (2.1). □

The following simple corollary of (2.1) applies to the choice $\phi(\xi) = \xi$.

COROLLARY 2.2. *Let the assumptions of Lemma* 2.1 *hold. Assume that* $\phi(\xi) = \xi$. $\|e_0\| \le \epsilon_1$, $\delta_0 \ge \Delta_0$, *and* $\delta_{max} = \infty$. *Then the convergence of* $\{x_n\}$ *to* x^* *is q-quadratic.*

2.2. Increasing the time step: Midrange phase.

Lemma 2.1 states that the second phase of Ψtc should produce both a large δ and an accurate solution. We show how this happens if the initial phase can provide only an accurate solution. This clarifies the role of the second phase in increasing the timestep. We do this by showing that if the steady state problem has a stable solution, then Ψtc behaves well. We now make assumptions that not only involve the nonlinear function but also the initial data and the dynamics of the initial value problem (IVP) (1.1).

Assumption 2.1.

1. F is everywhere defined and Lipschitz continuously Fréchet differentiable, and there is $M > 0$ such that $\|F'(x)\| \le M$ for all x.
2. There is a root x^* of F at which $F'(x^*)$ is nonsingular and $\eta > 0$ such that if $\|z - x_0\| < \eta$ then the solution of the initial value problem

   $$x' = -V^{-1}F(x),\ x(0) = z \tag{2.5}$$

   converges to x^* as $t \to \infty$.
3. There are $\epsilon_2, \beta > 0$ such that if $\|x - x^*\| \le \epsilon_2$ then $\|(I + \delta V^{-1}F'(x))^{-1}\| \le (1 + \beta\delta)^{-1}$ for all $\delta \ge 0$.

The analysis of the midrange uses part 3 of Assumption 2.1 in an important way to guarantee stability. The method for updating δ is not important for this result.

THEOREM 2.3. *Let $\{\delta_n\}$ be given by either* (1.6) *or* (1.8) *and let Assumption* 1.1 *hold. Let Assumption* 2.1 *hold. Let δ_{max} be large enough for the conclusions of Lemma* 2.1 *to hold. Then there is an $\epsilon_3 > 0$ such that if $\|x_0 - x^*\| < \epsilon_3$, and $\delta_0 > 0$, either*

$$\inf_n \delta_n = 0$$

or $x_n \to x^*$ and $\delta_n \to \delta_{max}$.

Proof. Let $\epsilon_3 < \min(\epsilon_1, \epsilon_2)$, where ϵ_1 is from Lemma 2.1 and ϵ_2 is from part 3 of Assumption 2.1. Note that

$$\begin{aligned} e_1 &= e_0 - (I + \delta_0 V^{-1} F'(x_0))^{-1} \delta_0 V^{-1} F(x_0) \\ &= e_0 - (I + \delta_0 V^{-1} F'(x_0))^{-1} \delta_0 V^{-1} F'(x_0) e_0 \\ &\quad + (I + \delta_0 V^{-1} F'(x_0))^{-1} \delta_0 V^{-1} (F'(x_0) e_0 - F(x_0)) \\ &= (I + \delta_0 V^{-1} F'(x_0))^{-1} e_0 + (I + \delta_0 V^{-1} F'(x_0))^{-1} \delta_0 V^{-1} (F'(x_0) e_0 - F(x_0)). \end{aligned}$$

Now there is a $c > 0$ such that

$$\|V^{-1}(F'(x)e - F(x))\| \le c\|e\|^2$$

for all x such that $\|e\| \le \epsilon_1$. Hence, reducing ϵ_3 further if needed so that $\epsilon_3 < \beta/(2c)$, we have

$$\|e_1\| \le \|e_0\| \left(\frac{1 + c\epsilon_3 \delta_0}{1 + \beta\delta_0} \right) \le \|e_0\| \left(\frac{1 + \beta\delta_0/2}{1 + \beta\delta_0} \right) < \|e_0\|. \tag{2.6}$$

If $\inf_n \delta_n = \delta^* > 0$ then

$$\|e_n\| \le \left(\frac{1 + \beta\delta^*/2}{1 + \beta\delta^*} \right) \|e_{n-1}\|$$

for all $n \ge 1$ and hence x_n converges to x^* q-linearly with q-factor $(1+\beta\delta^*/2)(1+\beta\delta^*)$.

This convergence implies that $x_n - x_{n-1} \to 0$ and $F(x_n) \to 0$ so $\delta_n \to \delta_{max}$ if either (1.6) or (1.8) is used. □

This result says that once the iteration has found a sufficiently good solution, either convergence will happen or the the iteration will stagnate with $\inf \delta_n = 0$. This latter failure mode is, of course, easy to detect. Moreover, the radius ϵ_3 of the ball about the root in Theorem 2.3 does not depend on $\inf \delta_n$.

2.3. Integration to steady state: Initial phase.

Theorem 2.3 requires an accurate estimate of x^*, but asks nothing of the timestep. In this section we show that if δ_0 is sufficiently small and (1.6) is used to update the timestep, then the dynamics of (1.1) will be tracked sufficiently well for such an approximate solution to be found. It is not clear how (1.8) allows for this.

THEOREM 2.4. *Let $\{\delta_n\}$ be given by* (1.6) *and let Assumption* 1.1 *hold. Let Assumption* 2.1 *hold. Let $\epsilon > 0$. There is a $\hat{\delta}$ such that if $\delta_0 \le \hat{\delta}$ then there is an n such that $\|e_n\| < \epsilon$.*

Proof. Let $\mathcal{S}$ be the trajectory of the solution to (2.5). By Assumption 2.1 x^* satisfies the assumptions [12, 17, 25], for local quadratic convergence of Newton's method and, therefore, there are ϵ_4 and ϵ_f such that if

$$\|x - y\| < \epsilon_4 \text{ for some } y \in \mathcal{S} \text{ and } \|F(x)\| < \epsilon_f$$

then $\|x - x^*\| < \epsilon_3$, which will suffice for the conclusions of Theorem 2.3 to hold. Let

$$M = \sup_{\substack{y \in \mathcal{S} \\ \|x-y\| < \epsilon_4}} \|F(x)\|.$$

We will show that if δ_0 sufficiently small, then $\|x_k - x(t_k)\| \le \epsilon_4$ until $\|F(x_k)\| < \epsilon_f$. By (1.7), if

$$\delta_0 \|F(x_0)\| / \epsilon_f < \xi_t,$$

then

$$C_L \delta_0 \;\equiv\; \delta_0 \|F(x_0)\|/M \le \delta_k \le \delta_0 \|F(x_0)\|/\epsilon_f \;\equiv\; C_U \delta_0 \tag{2.7}$$

as long as $\|F(x)\| \ge \epsilon_f$ and x_k is within ϵ_4 of the trajectory $\mathcal{S}$.

Let z be the solution of (2.5). Let $0 < T < \infty$ be such that for all $t > T$, $\|F(x)\| < \epsilon_f$ whenever $\|x - z(t)\| < \epsilon_4$. Consider the approximate integration of (2.5) by (1.2). Set

$$t_n = \sum_{l=0}^{n} \delta_l > (nC_L + 1)\delta_0.$$

If $\|x_n - z(t_n)\| < \epsilon_4$ and $\|F(x_n)\| \ge \epsilon_f$ then (2.7) holds. This cannot happen if $n > (T - \delta_0)/(C_L \delta_0)$ (which implies that $t_n > T$). Therefore, the proof will be complete if we can show that

$$\|x_n - z(t_n)\| < \epsilon_4$$

for all

$$n \le T/(C_L \delta_0). \tag{2.8}$$

Note that (1.2) may be written as

$$x_{n+1} = x_n - \delta_n V^{-1} F(x_n) + [I - (I + \delta_n V^{-1} F'(x_n))^{-1}] \delta_n V^{-1} F(x_n). \tag{2.9}$$

There is an m_1 such that the last term in (2.9) satisfies, for δ_n sufficiently small and $\|x_n - z(t_n)\| < \epsilon_4$,

$$\|[I - (I + \delta_n V^{-1} F'(x_n))^{-1}] \delta_n V^{-1} F(x_n)\| \le m_1 \delta_n^2. \tag{2.10}$$

Let $E_n = \|x_n - z(t_n)\|$. Then we have by our assumptions on F that there is an $M_1 > 0$ such that

$$\|x_n - \delta_n V^{-1} F(x_n) - z(t_n) - z'(t_n)\delta_n\| \le (1 + M_1 \delta_n) E_n. \tag{2.11}$$

Finally, there is an m_2 such that for δ_n sufficiently small and $\|x_n - z(t_n)\| < \epsilon_4$,

$$\|z(t_n + \delta_n) - z(t_n) - z'(t_n)\delta_n\| \leq m_2 \delta_n^2.$$

Setting $M_2 = m_1 + m_2$ we have for all $n \geq 1$ (as long as δ_n is sufficiently small and $\|x_n - z(t_n)\| < \epsilon_4$),

$$E_n \leq (1 + M_1 \delta_{n-1}) E_{n-1} + M_2 \delta_{n-1}^2.$$

As long as (2.7) holds, this implies that

$$E_n \leq (1 + M_1 C_U \delta_0) E_{n-1} + M_2 C_U^2 \delta_0^2.$$

Consequently, as is standard [14, 15],

$$E_n \leq \frac{\delta_0 M_2 C_U [\exp(n M_1 C_U \delta_0) - 1]}{M_1},$$

and using (2.8),

$$E_n \leq \delta_0 \frac{M_2 C_U [\exp(C_L M_1 C_U T) - 1]}{M_1}. \tag{2.12}$$

So if

$$\delta_0 < \epsilon_4 \frac{M_1}{M_2 C_U [\exp(C_L M_1 C_U T) - 1]} \tag{2.13}$$

then $\|x_n - z(t_n)\| < \epsilon_4$ for all n until $F(x_n) < \epsilon_f$ or $t_n > T$. This completes the proof. □

The problem with application of this proof to the update given by (1.8) is that bounds on δ like (2.7) do not follow from the update formula.

3. Inexact Newton iteration. In this section we look at Ψtc as implemented in practice. There are two significant differences between the simple version in section 2 and realistic implementations.

1. The Fréchet derivative $\delta_n^{-1} V + F'(x_n)$ is not recomputed with every timestep.
2. The equation for the Newton step is solved only inexactly.

Before showing how the results in section 2 are affected by these differences, we provide some more detail and motivation.

Item 1 is subtle. If one is solving the equation for the Newton step with a direct method, then evaluation and factorization of the Jacobian matrix is not done at every timestep. This is a common feature of many ODE and DAE codes, [30, 26, 27, 3]. Jacobian updating is an issue in continuation methods [31, 28], and implementations of the chord and Shamanskii [29] methods for general nonlinear equations [2, 17, 25]. When the Jacobian is slowly varying as a function of time or the continuation parameter, sporadic updating of the Jacobian leads to significant performance gains. One must decide when to evaluate and factor the Jacobian using iteration statistics and (in the ODE and DAE case) estimates of truncation error. Temporal truncation error is not of interest to us, of course, if we seek only the steady state solution.

In [16], a Jacobian corresponding to a lower-order discretization than that for the residual was used in the early phases of the iteration and in [19], in the context of a matrix-free Newton method, the same was used as a preconditioner.

The risks in the use of inaccurate Jacobian information are that termination decisions for the Newton iteration and the decision to reevaluate and refactor the Jacobian are related and one can be misled by rapidly varying and ill-conditioned Jacobians into premature termination of the nonlinear iteration [30, 32, 18]. In the case of iterative methods, item 1 should be interpreted to mean that preconditioning information (such as an incomplete factorization) is not computed at every timestep.

Item 2 means that the equation for the Newton step is solved inexactly in the sense of [11], so that instead of

$$x_+ = x_c + s, \tag{3.1}$$

where s is given by (1.3), step s satisfies

$$\|(\delta^{-1}V + F'(x))s + F(x)\| \leq \eta\|F(x)\|, \tag{3.2}$$

for some small η, which may change as the iteration progresses. Item 1 can also be viewed as an inexact Newton method with η reflecting the difference between the approximate and exact Jacobians.

The theory in section 2 is not changed much if inexact computation of the step is allowed. The proof of Lemma 2.1 is affected in (2.1), which must be changed to

$$\|e_+\| \leq K_C(\|e_c\|^2 + [\delta_c^{-1} + \eta_c]\|e_c\|). \tag{3.3}$$

This changes the statement of the lemma to the following.

LEMMA 3.1. *Let $\{\delta_n\}$ be given by either* (1.6) *or* (1.8) *and let Assumption* 1.1 *hold. Let $F(x^*) = 0$, $F'(x^*)$ be nonsingular, and F' be Lipschitz continuous with Lipschitz constant γ in a ball of radius ϵ about x^*.*

Then there are $\epsilon_1 > 0$, $\bar{\eta}$ and Δ_0 such that if $\delta_{max}, \delta_0 > \Delta_0$, $\eta_n \geq \bar{\eta}$ for all n, and $\|x_0 - x^*\| < \epsilon_1$, then the sequence defined by (3.1), (3.2), *and* (1.6) *satisfies*

$$\delta_n \rightarrow \delta_{max}$$

and $x_n \rightarrow x^*$ q-superlinearly if $\delta_{max} = \infty$ and $\eta_n \rightarrow 0$ and q-linearly if $\delta_{max} < \infty$.

Corollary 2.2 becomes the following.

COROLLARY 3.2. *Let the assumptions of Lemma* 3.1 *hold. Assume that $\phi(\xi) = \xi$, $\|e_0\| \leq \epsilon_1$, $\delta_0 \geq \Delta_0$, and $\delta_{max} = \infty$. Then the convergence of $\{x_n\}$ to x^* is q-superlinear if $\eta_n \rightarrow 0$ and locally q-quadratic if $\eta_n = O(\|F(x_n)\|)$.*

The analysis of the midrange phase changes in (2.6), where we obtain

$$\|e_1\| \leq \|e_0\| \left(K_\eta \eta_0 + \frac{1 + c\epsilon_3\delta_0}{1 + \beta\delta_0} \right) < \|e_0\| \tag{3.4}$$

for some $K_\eta > 0$. This means that $\bar{\eta}$ must be small enough to maintain the q-linear convergence of $\{x_n\}$ during this phase. The inexact form of Theorem 2.3 is the following.

THEOREM 3.3. *Let $\{x_n\}$ be given by* (3.1) *and* (3.2) *and let $\{\delta_n\}$ be given by either* (1.6) *or* (1.8). *Let Assumption* 1.1 *hold. Let Assumption* 2.1 *hold. Let δ_{max} be large enough for the conclusions of Lemma 2.1 to hold. Then there are $\epsilon_3 > 0$ and $\bar{\eta}$ such that if $\eta_n \leq \bar{\eta}$, $\|x_0 - x^*\| < \epsilon_3$ and $\delta_0 > 0$, either*

$$\inf_n \delta_n = 0$$

or $x_n \rightarrow x^*$ and $\delta_n \rightarrow \delta_{max}$.

Inexact Newton methods, in particular Newton–Krylov solvers, have been applied to ODE/DAE solvers in [1], [5], [4], [6], [7], and [9]. The context here is different in that the nonlinear residual $F(x)$ does not reflect the error in the transient solution but in the steady state solution.

The analysis of the initial phase changes through (2.10). We must now estimate

$$\delta_n V^{-1} F(x_n) + s_n.$$

Set $r = (\delta_n^{-1} V + F'(x_n))s_n + F(x_n)$. Note that

$$\begin{aligned} \delta_n V^{-1} F(x_n) + s_n &= \delta_n V^{-1}(F(x_n) + \delta_n^{-1} V s_n) \\ &= \delta_n V^{-1}(r - F'(x_n)s_n). \end{aligned}$$

Now,

$$s_n = (I + \delta_n V^{-1} F'(x_n))^{-1} \delta_n V^{-1}(r - F(x_n))$$

and hence, assuming that the operators $(I + \delta_n V^{-1} F'(x_n))^{-1}$ are uniformly bounded, there is m_3 such that

$$\|s_n\| \le m_3 \delta_n$$

and hence

$$\|\delta_n V^{-1} F(x_n) + s_n\| \le \delta_n \|V^{-1}\| (\eta_n \|F(x_n)\| + \|F'(x_n)\| m_3 \delta_n). \tag{3.5}$$

We express (3.5) as

$$\|\delta_n V^{-1} F(x_n) + s_n\| \le m_4(\eta_n \delta_n + \delta_n^2). \tag{3.6}$$

Hence, if $\eta_n = O(\delta_0)$ or $\eta_n = O(E_n)$, we still obtain (2.12) and the inexact form of Theorem 2.4.

THEOREM 3.4. *Let $\{x_n\}$ be given by* (3.1) *and* (3.2) *and let $\{\delta_n\}$ be given by* (1.6) *and let Assumption* 1.1 *hold. Let Assumption* 2.1 *hold. Assume that the operators $(I + \delta_n V^{-1} F'(x_n))^{-1}$ are uniformly bounded in n. Let $\epsilon > 0$. There are $\hat{\delta}$ and $\bar{\eta}$ such that if $\delta_0 \le \hat{\delta}$ and $\eta_n \le \bar{\eta}$ then there is an n such that $\|e_n\| < \epsilon$.*

The restrictions on η in Theorem 3.4 seem to be stronger than those on the results on the midrange and terminal phases. This is consistent with the tight defaults on the forcing terms for Newton–Krylov methods when applied in the ODE/DAE context [1, 5, 6, 7, 9].

4. Numerical experiments.

In this section we examine a commonly used Ψtc technique, switched evolution/relaxation (SER) [21], applied to a Newton-like method for inviscid compressible flow over a four-element airfoil in two dimensions. Three phases corresponding roughly to the theoretically motivated iteration phases of section 2 may be identified. We also compare SER with a different Ψtc technique based on bounding temporal truncation error (TTE) [20]. TTE is slightly more aggressive than SER in building up the timestep in this particular problem, but the behavior of the system is qualitatively the same.

The physical problem, its discretization, and its algorithmic treatment in both a nonlinear defect correction iteration and in a Newton–Krylov–Schwarz iteration—as

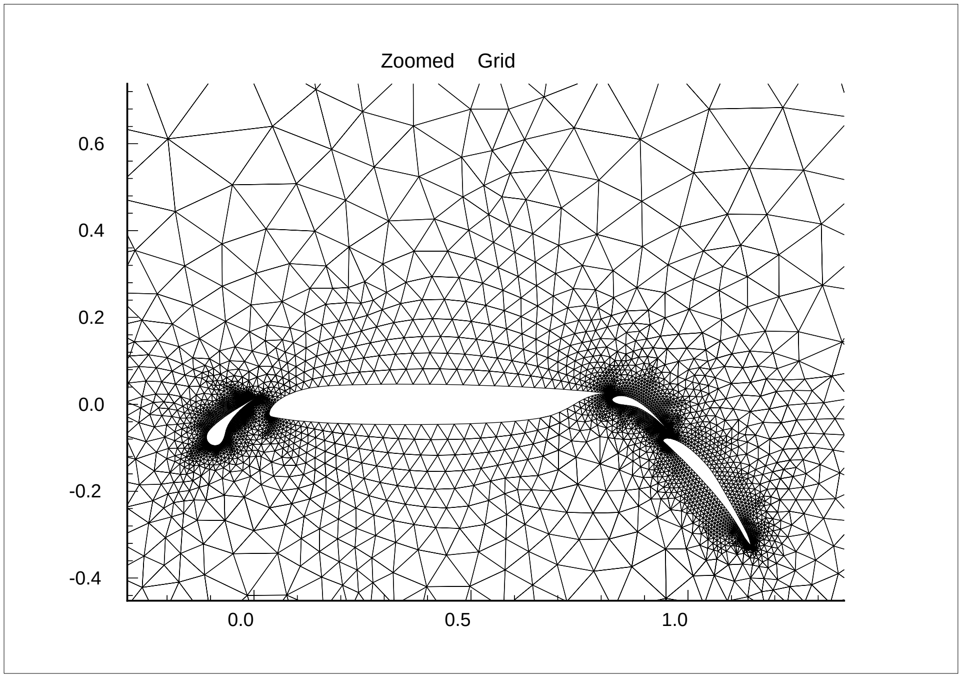

Fig. 4.1. *Unstructured grid around four-element airfoil in landing configuration — near-field view.*

well as its suitability for parallel computation—have been documented in earlier papers, most recently [10], and the references therein. Our description is correspondingly compact.

The unknowns of the problem are nodal values of the fluid density, velocities, and specific total energy, $x = (\ldots, \rho_i, u_i, v_i, e_i, \ldots)^T$ at N vertices in an unstructured grid of triangular cells (see Fig. 4.1). The system $F(x) = 0$ is a discretization of the steady Euler equations:

$$\nabla \cdot (\rho \mathbf{v}) = 0, \tag{4.1}$$

$$\nabla \cdot (\rho \mathbf{v}\mathbf{v} + pI) = 0, \tag{4.2}$$

$$\nabla \cdot ((\rho e + p)\mathbf{v}) = 0, \tag{4.3}$$

where the pressure p is supplied from the ideal gas law, $p = \rho(\gamma - 1)(e - |\mathbf{v}|^2/2)$, and γ is the ratio of specific heats. The discretization is based on a control volume approach, in which the control volumes are the duals of the triangular cells — nonoverlapping polygons surrounding each vertex whose perimeter segments join adjacent cell centers to midpoints of incident cell edges. Integrals of (4.1)–(4.3) over the control volumes are transformed by the divergence theorem to contour integrals of fluxes, which are estimated numerically through an upwind scheme of Roe type. The effective scaling matrix V for the Ψtc term is a diagonal matrix that depends upon the mesh.

The boundary conditions correspond to landing configuration conditions, subsonic (Mach number of 0.2) with a high angle of attack of (5°). The full adaptively clustered unstructured grid contains 6,019 vertices, with four degrees of freedom per vertex (giving 24,076 as the algebraic dimension of the discrete nonlinear problem). Figure 4.1 shows only a near-field zoom on the full grid, whose far-field boundaries are approximately 20 chords away. The initial pseudotimestep, $\delta_0 \approx 3 \times 10^{-4}$, corresponds to a Courant–Friedrichs–Lewy (CFL) number of 20. The pseudotimestep is allowed to grow up to six orders of magnitude over the course of the iterations. It is ultimately bounded at $\delta_{max} = 10^6 \cdot \delta_0$ guaranteeing a modest diagonal contribution that aids the invertibility of $(\delta^{-1}V + F'(x_n))^{-1}$.

The initial iterate is a uniform flow, based on the far-field boundary conditions—constant density and energy, and constant velocity at a given angle of attack.

The solution algorithm is a hybrid between a nonlinear defect correction and a full Newton method, a distinction which requires further discussion of the processes that supply $F(x)$ and $F'(x)$ within the code. The form of the vector-valued function $F(x)$ determines the quality of the solution and is always discretized to required accuracy (second order in this paper). The form of the approximate Jacobian matrix $F'(x)$, together with the scaling matrix V and timestep δ, determines the rate at which the solution is achieved but does not affect the quality of a converged result, and is, therefore, a candidate for replacement with a matrix that is more convenient. In practice, we perform the matrix inversion in (1.2) by Krylov iteration, which requires only the action of $F'(x)$ on a series of Krylov vectors and not the actual elements of $F'(x)$. The Krylov method was restarted GMRES(20) preconditioned with 1-cell overlap additive Schwarz (eight subdomains).

Following [5, 8], we use matrix-free Fréchet approximations of the required action:

$$F'(x)v \approx \frac{1}{h}\left[F(x+hv) - F(x)\right]. \tag{4.4}$$

However, when preconditioning the solution of (1.2), we use a more economical matrix than the Jacobian based on the true $F(x)$, obtained from a first-order discretization of the governing Euler system. This not only decreases the number of elements in the preconditioner, relative to a true Jacobian, but also the computation and (in the parallel context) communication in applying the preconditioner. It also results in a more numerically diffusive and stable matrix, which is desirable for inversion. The price for these advantages is that the preconditioning is inconsistent with the true Jacobian, so more linear subiterations may be required to meet a given linear convergence tolerance. This has an indirect feedback on the nonlinear convergence rate, since we limit the work performed in any linear subiteration in an inexact Newton sense.

In previous work on the Euler and Navier–Stokes equations [10, 23], we have noted that a Ψtc method based on a consistent high-order Jacobian stumbles around with a nonmonotonic steady state residual norm at the outset of the nonlinear iterations for a typical convenient initial iterate far from the steady state solution. On the other hand, a simple defect correction approach, in which $F'(x)$ is based on a regularizing first-order discretization everywhere it appears in the solution of (1.2), not just in the preconditioning, allows the residual to drop smoothly from the outset. In this work, we employ a hybrid strategy, in which defect correction is used until the residual norm has fallen by three orders of magnitude and inexact Newton thereafter. As noted in section 3, inexact iteration based on the true Jacobian and iteration with an inconsistent Jacobian can both be gathered under the η of (3.2), so the theory extends in principal to both.

With this background we turn our attention to Fig. 4.2, in which are plotted on a logarithmic scale against the Ψtc iteration number: the steady state residual norm $||F(x_n)||_2$ at the beginning of each iteration, the norm of the update vector $||x_{n+1} - x_n||_2$, and the pseudotimestep δ_n.

The residual norm falls nearly monotonically, as does the norm of the solution update. Asymptotic convergence cannot be expected to be quadratic or superlinear,

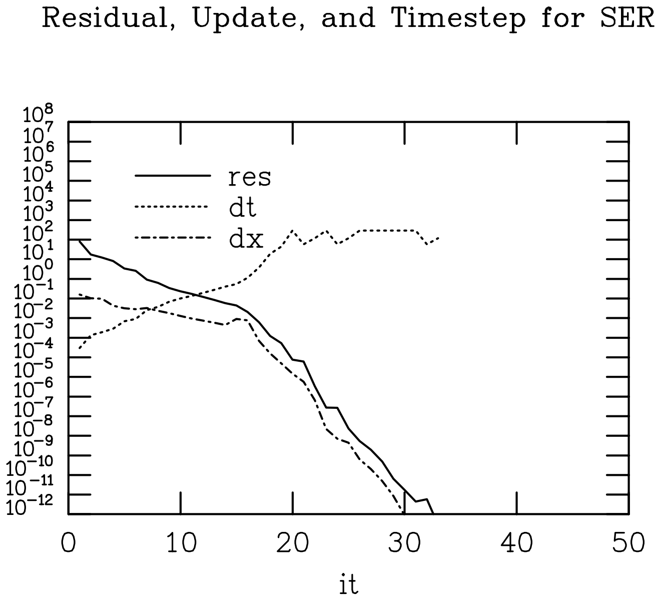

FIG. 4.2. *SER convergence history.*

since we do not enforce $\eta_n \to 0$ in (3.5). However, linear convergence is steep, and our experience shows that overall execution time is increased if too many linear iterations are employed in order to enforce $\eta_n \to 0$ asymptotically. In the results shown in this section, the inner linear convergence tolerance was set at 10^{-2} for the defect correction part of the trajectory, and at 10^{-3} for the Newton part. The work was also limited to a maximum of 12 restart cycles of 20 Krylov vectors each.

Examination of the pseudotimestep history shows monotonic growth that is gradual through the defect correction phase (ending at $n = 14$), then more rapidly growing, and asymptotically at δ_{max} (beginning at $n = 20$). Steps 21, 24, and 32 show momentary retreats from δ_{max} in response to a refinement on the Ψtc strategy that automatically cuts back the pseudotimestep by a fixed factor if a nonlinear residual reduction of less than 3/4 is achieved at the exhaustion of the maximum number of Krylov restarts in the previous step (during the terminal Newton phase). Close examination usually reveals a stagnation plateau in the linear solver, and it is more cost effective to fall back to the physical transient to proceed than to grind on the ill-conditioned linear problem. These glitches in the convergence of $||F(x_n)||_2$ are not of nonlinear origin.

Another timestep policy, common in the ODE literature, is based on controlling TTE estimates. Though we do not need to maintain TTEs at low levels when we are not attempting to follow physical transients, we may maintain them at high levels as a heuristic form of stepsize control. This policy seems rare in external aerodynamic simulations, but is popular in the combustion community and is implemented in [20]. The first neglected term in the Euler discretization of $\partial x/\partial t$ is $1/2x'' \cdot \delta^2$, so a reasonable mixed absolute relative bound on the error in the ith component of x at the nth step is

$$\left| \frac{\delta_n^2}{2(1+|x_i|)} \frac{\partial^2 x_i}{\partial t^2}(t_n) \right| \leq \tau, \tag{4.5}$$

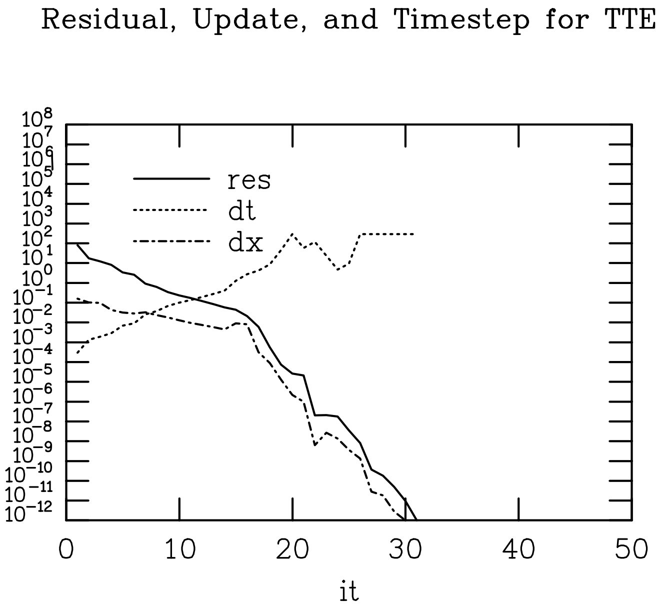

FIG. 4.3. *TTE convergence history.*

where $(x_i'')_n$ can be approximated by

$$\frac{2}{\delta_{n-1}+\delta_{n-2}}\left[\frac{(x_i)_n-(x_i)_{n-1}}{\delta_{n-1}}-\frac{(x_i)_{n-1}-(x_i)_{n-2}}{\delta_{n-2}}\right].$$

Taking τ as 3/4 and implementing this strategy in the Euler code in place of SER yields the results in Fig. 4.3. Arrival at δ_{max} occurs at the same step as for SER, and arrival at the threshold $||F(x_n)|| < 10^{-12}$ occurs one iteration earlier. However, the convergence difficulties after having arrived at δ_{max} are slightly greater.

5. Conclusions. Though the numerical experiments of the previous section do not confirm the theory in detail, in the sense that we do not verify the estimates in the hypotheses, a reassuring similarity exists between the observations of the numerics and the conceptual framework of the theory, which was originally motivated by similar observations in the literature. There is a fairly long induction phase, in which the initial iterate is guided towards the Newton convergence domain by remaining close to the physical transient, with relatively small timesteps. There is a terminal phase which can be made as rapid as the capability of the linear solver permits (which varies from application to application), in which an iterate in the Newton convergence domain is polished. Connecting the two is a phase of moderate length during which the timestep is built up towards the Newton limit of δ_{max}, starting from a reasonably accurate iterate. The division between these phases is not always clear cut, though exogenous experience suggests that it becomes more so when the corrector of section 1 is iterated towards convergence on each timestep. We plan to examine this region of parameter space in conjunction with an extension of the theory to mixed steady/Ψtc systems (analogues of differential-algebraic systems in the ODE context) in the future.

Acknowledgments. This paper began with a conversation at the DOE Workshop on Iterative Methods for Large Scale Nonlinear Problems held in Logan, Utah in the fall of 1995. The authors appreciate the efforts and stimulus of the organizers, Homer Walker and Michael Pernice. They also wish to thank Peter Brown, Rob

Nance, and Dinesh Kaushik for several helpful discussions on this paper. This paper was significantly improved by the comments of a thoughtful and thorough referee.

REFERENCES

[1] K. E. Brenan, S. L. Campbell, and L. R. Petzold, *The Numerical Solution of Initial Value Problems in Differential-Algebraic Equations*, Classics in Applied Mathematics #14, SIAM, Philadelphia, PA, 1996.

[2] R. Brent, *Some efficient algorithms for solving systems of nonlinear equations*, SIAM J. Numer. Anal., 10 (1973), pp. 327–344.

[3] P. N. Brown, G. D. Byrne, and A. C. Hindmarsh, *VODE: A variable coefficient ODE solver*, SIAM J. Sci. Statist. Comput., 10 (1989), pp. 1038–1051.

[4] P. N. Brown and A. C. Hindmarsh, *Matrix-free methods for stiff systems of ODE's*, SIAM J. Numer. Anal., 23 (1986), pp. 610–638.

[5] P. N. Brown and A. C. Hindmarsh, *Reduced storage matrix methods in stiff ODE systems*, J. Appl. Math. Comp., 31 (1989), pp. 40–91.

[6] P. N. Brown, A. C. Hindmarsh, and L. R. Petzold, *Using Krylov methods in the solution of large-scale differential-algebraic systems*, SIAM J. Sci. Comput., 15 (1994), pp. 1467–1488.

[7] P. N. Brown, A. C. Hindmarsh, and L. R. Petzold, *Consistent Initial Condition Calculation for Differential-Algebraic Systems*, Tech. report UCRL-JC-122175, Lawrence Livermore National Laboratory, Livermore, CA, August 1995.

[8] P. N. Brown and Y. Saad, *Hybrid Krylov methods for nonlinear systems of equations*, SIAM J. Sci. Statist. Comput., 11 (1990), pp. 450–481.

[9] G. D. Byrne, *Pragmatic experiments with Krylov methods in the stiff ODE setting*, in Computational Ordinary Differential Equations, J. Cash and I. Duff, eds., Oxford University Press, London, 1992.

[10] X. D. Cai, D. E. Keyes, and V. Venkatakrishnan, *Newton-Krylov-Schwarz: An implicit solver for CFD*, in Proceedings of the 8th International Conference on Domain Decomposition Methods, Domain Decomposition Methods in Sciences and Engineering, R. Glowinski, J. Periaux, Z. Shi, and O. Widlund, eds., Wiley, Chichester, 1996, pp. 387–402.

[11] R. Dembo, S. Eisenstat, and T. Steihaug, *Inexact Newton methods*, SIAM J. Numer. Anal., 19 (1982), pp. 400–408.

[12] J. E. Dennis and R. B. Schnabel, *Numerical Methods for Nonlinear Equations and Unconstrained Optimization*, Prentice–Hall, Englewood Cliffs, NJ, 1983.

[13] A. Ern, V. Giovangigli, D. E. Keyes, and M. D. Smooke, *Towards polyalgorithmic linear system solvers for nonlinear elliptic problems*, SIAM J. Sci. Comput., 15 (1994), pp. 681–703.

[14] C. W. Gear, *Numerical Initial Value Problems in Ordinary Differential Equations*, Prentice–Hall, Englewood Cliffs, NJ, 1971.

[15] E. Isaacson and H. B. Keller, *Analysis of Numerical Methods*, Wiley, New York, 1966.

[16] H. Jiang and P. A. Forsyth, *Robust linear and nonlinear strategies for solution of the transonic Euler equations*, Comput. Fluids, 24 (1995), pp. 753–770.

[17] C. T. Kelley, *Iterative Methods for Linear and Nonlinear Equations*, in Frontiers in Applied Mathematics 16, SIAM, Philadelphia, PA, 1995.

[18] C. T. Kelley, C. T. Miller, and M. D. Tocci, *Termination of Newton/Chord Iterations and the Method of Lines*, Tech. report CRSC-TR96-19, Center for Research in Scientific Computation, North Carolina State University, Raleigh, NC, May 1996; SIAM J. Sci. Comput., to appear.

[19] D. E. Keyes, *Aerodynamic applications of Newton-Krylov-Schwarz solvers*, in Proc. 14th Internat. Conference on Numerical Methods in Fluid Dynamics, M. Deshpande, S. Desai, and R. Narasimha, eds., Springer, 1995, pp. 1–20.

[20] D. E. Keyes and M. D. Smooke, *A parallelized elliptic solver for reacting flows*, in Parallel Computations and Their Impact on Mechanics, A. K. Noor, ed., American Society of Mechanical Engineers, 1987, pp. 375–402.

[21] W. Mulder and B. V. Leer, *Experiments with implicit upwind methods for the Euler equations*, J. Comput. Phys., 59 (1985), pp. 232–246.

[22] O. Nevanlinna, *Convergence of Iterations for Linear Equations*, Birkhäuser, Basel, 1993.

[23] E. J. Nielsen, R. W. Walters, W. K. Anderson, and D. E. Keyes, *Application of Newton-Krylov Methodology to a Three-Dimensional Unstructured Euler Code*, AIAA Paper 95-1733, June 1995.

[24] P. D. Orkwis and D. S. McRae, *A Newton's Method Solver for the Navier-Stokes Equations*, AIAA Paper 90-1524, June 1990.

[25] J. M. Ortega and W. C. Rheinboldt, *Iterative Solution of Nonlinear Equations in Several Variables*, Academic Press, New York, 1970.

[26] L. R. Petzold, *A description of DASSL: A differential/algebraic system solver*, in Scientific Computing, R. S. Stepleman et al., eds., North Holland, Amsterdam, 1983, pp. 65–68.

[27] K. Radhakrishnan and A. C. Hindmarsh, *Description and Use of LSODE, the Livermore Solver for Ordinary Differential Equations*, Tech. report URCL-ID-113855, Lawrence Livermore National Laboratory, December 1993.

[28] R. Schriber and H. B. Keller, *Driven cavity flows by efficient numerical techniques*, J. Comput. Phys., 49 (1983), pp. 310–333.

[29] V. E. Shamanskii, *A modification of Newton's method*, Ukrainn. Mat. Zh., 19 (1967), pp. 133–138 (in Russian).

[30] L. F. Shampine, *Implementation of implicit formulas for the solution of ODEs*, SIAM J. Sci. Statist. Comput., 1 (1980), pp. 103–118.

[31] M. D. Smooke, *An error estimate for the modified Newton method with applications to the solution of nonlinear two-point boundary value problems*, J. Optim. Theory Appl., 39 (1983), pp. 489–511.

[32] M. D. Tocci, C. T. Kelley, and C. T. Miller, *Accurate and economical solution of the pressure head form of Richards' equation by the method of lines*, Adv. Water Resources, 20 (1997), pp. 1–14.

[33] V. Venkatakrishnan, *Newton solution of inviscid and viscous problems*, AIAA J., 27 (1989), pp. 885–891.