SIAM J. NUMER. ANAL.
Vol. 46, No. 4, pp. 1892–1911

ON USING APPROXIMATE FINITE DIFFERENCES IN MATRIX-FREE NEWTON–KRYLOV METHODS*

PETER N. BROWN†, HOMER F. WALKER‡, REBECCA WASYK‡, AND CAROL S. WOODWARD†

Abstract. A Newton–Krylov method is an implementation of Newton's method in which a Krylov subspace method is used to solve approximately the linear systems that characterize steps of Newton's method. Newton–Krylov methods are often implemented in "matrix-free" form, in which the Jacobian-vector products required by the Krylov solver are approximated by finite differences. Here we consider using approximate function values in these finite differences. We first formulate a finite-difference Arnoldi process that uses approximate function values. We then outline a Newton–Krylov method that uses an implementation of the GMRES or Arnoldi method based on this process, and we develop a local convergence analysis for it, giving sufficient conditions on the approximate function values for desirable local convergence properties to hold. We conclude with numerical experiments involving particular function-value approximations suitable for nonlinear diffusion problems. For this case, conditions are given for meeting the convergence assumptions for both lagging and linearizing the nonlinearity in the function evaluation.

Key words. matrix-free Newton–Krylov methods, Newton–GMRES methods, Newton–Arnoldi methods, Newton's method, Krylov subspace methods, GMRES method, Arnoldi method

AMS subject classifications. 65H10, 65F10

DOI. 10.1137/060652749

1. Introduction.

The problem of interest is to determine $u_* \in \mathbb{R}^n$ satisfying a system of nonlinear equations

$$F(u_*) = 0, \tag{1.1}$$

where $F : \mathbb{R}^n \to \mathbb{R}^n$ is continuously differentiable. A classical algorithm for approximately solving (1.1) is *Newton's method*, which, at a current approximate solution u, generates a next approximate solution $u + s$ through a step s that satisfies the linear *Newton equation*

$$F'(u)s = -F(u). \tag{1.2}$$

Our interest here is in *Newton–Krylov methods* (cf. [2], [11], [12], [15]), in which *Krylov subspace methods* are used to solve (1.2) approximately. An extensive introduction to Krylov subspace methods is beyond the scope of this paper; we touch on only the most relevant aspects here and refer the reader to [8], [9], [18], [19] for full treatments.

*Received by the editors February 23, 2006; accepted for publication (in revised form) September 25, 2007; published electronically April 18, 2008.
http://www.siam.org/journals/sinum/46-4/65274.html

†Center for Applied Scientific Computing, Lawrence Livermore National Laboratory, Box 808, L-561, Livermore, CA 94551 (pnbrown@llnl.gov, cswoodward@llnl.gov). The work of these authors was performed under the auspices of the U.S. Department of Energy by University of California, Lawrence Livermore National Laboratory under contract W-7405-ENG-48.

‡Department of Mathematical Sciences, Worcester Polytechnic Institute, Worcester, MA 01609-2280 (walker@wpi.edu, wasykrd@wpi.edu). The work of these authors was supported in part by the Center for Simulation of Accidental Fires and Explosions funded at the University of Utah by the U.S. Department of Energy under contracts LLNL B341493 and B524196.

A Krylov subspace method applied to a general linear system $Ax = b$, $A \in \mathbb{R}^{n\times n}$ invertible, begins with an initial approximate solution x_0 and, at the jth iteration, determines $x_j = x_0 + z_j$ through a correction z_j in the jth *Krylov subspace*

$$\mathcal{K}_j \equiv \text{span}\,\{r_0, Ar_0, \ldots, A^{j-1}r_0\}, \tag{1.3}$$

where $r_0 = b - Ax_0$. Specific methods are determined by the choice of $z_j \in \mathcal{K}_j$.

We focus here on two Krylov subspace methods: the *generalized minimal residual method* (GMRES) [16], in which each z_j is chosen to minimize the residual norm over all corrections in $\mathcal{K}_j$, and the *Arnoldi method* (also known as the *full orthogonalization method*) [17], in which each z_j is chosen to make the residual orthogonal to $\mathcal{K}_j$, if that is possible. (We consider these methods in only their basic forms and do not consider "restarted" or "truncated" variants.) These have in common that their iterates are generated directly from the *Arnoldi process* [1], which we review in section 2.1.

In the Newton–Krylov context, a particular advantage of the most widely used Krylov subspace methods, including the GMRES and Arnoldi methods, is that they require only products of F' with vectors for their implementation. Thus they allow "matrix-free" Newton–Krylov formulations (cf. [12]), in which these products are evaluated or approximated without creating or storing F'. Perhaps the most popular way of approximating these products is with a first-order forward difference

$$F'(u)v \approx \frac{F(u+\sigma v) - F(u)}{\sigma}, \tag{1.4}$$

where σ is an appropriately chosen difference step.

Each approximation (1.4) requires a new F-evaluation. If F-evaluations are expensive or if many iterations of the Krylov solver are required, then the aggregate cost of these evaluations over all Krylov iterations may be undesirably high. In this case, it is natural to consider using an inexpensive approximation of $F(u+\sigma v)$ in (1.4), if one is available.

In the following, we explore this possibility. Our goal is to lay out a framework that will provide guidelines for using such approximations, specifically to outline conditions on such approximations that lead to desirable convergence properties of the resulting Newton–Krylov methods. Although we consider only the GMRES and Arnoldi methods here, it is our expectation that the results will provide useful guidance when other Krylov subspace methods are used as well.

In section 2, we first describe the approximations of F that we allow in difference quotients and formulate a finite-difference Arnoldi process that uses them. We then outline a Newton–Krylov method that uses GMRES or Arnoldi implementations based on this process, and we develop a local convergence analysis for it. In section 3, we discuss numerical experiments involving two illustrative approximations of F that are appropriate for a broad class of nonlinear diffusion problems. We also show how these approximations can be chosen to satisfy the convergence assumptions in section 2. We offer a summary discussion and conclusions in section 4.

This work builds on and extends work of [3]. We assume throughout that F is continuously differentiable on domains of interest and, for convenience, that $\|\cdot\|$ and $\langle\cdot,\cdot\rangle$ are the Euclidean norm and inner-product on $\mathbb{R}^n$, respectively. (All results can be easily extended to the case of an arbitrary inner-product norm on $\mathbb{R}^n$.) For $u \in \mathbb{R}^n$ and $\delta > 0$, we denote $B(u,\delta) \equiv \{v \in \mathbb{R}^n : \|v-u\| < \delta\}$. We implicitly assume that arithmetic is exact and do not consider errors due to rounding in finite-precision arithmetic.

2. The approximate finite-difference framework. We assume that there is a function $\tilde{F} : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^n$ such that $\tilde{F}(u,u) = F(u)$ for all $u \in \mathbb{R}^n$. The presumption is that, in an application of interest, one can determine $\tilde{F}$ so that, for each u and w near u, evaluating $\tilde{F}(u,w)$ is preferred over evaluating $F(w)$. Thus, in lieu of (1.4), we consider approximations of the form

$$F'(u)v \approx \frac{\tilde{F}(u, u+\sigma v) - F(u)}{\sigma}. \tag{2.1}$$

We make further assumptions on $\tilde{F}$ after outlining in section 2.1 a version of the Arnoldi process that uses these approximations.

2.1. The approximate finite-difference Arnoldi process. We begin by recalling the usual Arnoldi process for a general linear system $Ax = b$, $A \in \mathbb{R}^{n\times n}$ invertible, and initial $x_0 \in \mathbb{R}^n$.

ALGORITHM 2.1: ARNOLDI PROCESS

GIVEN x_0, SET $r_0 = b - Ax_0$.

IF $r_0 = 0$, STOP; OTHERWISE SET $v_1 = r_0/\|r_0\|$.

FOR $j = 1, 2, \ldots,$ DO:

SET $w_j = Av_j - \sum_{i=1}^{j} h_{ij} v_i$, WHERE $h_{ij} = \langle Av_j, v_i \rangle$.

IF $h_{j+1,j} \equiv \|w_j\| = 0$, STOP; OTHERWISE, SET

$v_{j+1} = w_j/h_{j+1,j}$ AND CONTINUE.

If $r_0 \neq 0$, then the process terminates at step $j > 0$ if and only if $h_{j+1,j} = 0$, which occurs if and only if $\mathcal{K}_j$ is an invariant subspace of A. Thus the process must terminate for some $j \leq n$. For each j up to termination, $\{v_i\}_{i=1,\ldots,j}$ constitutes an orthonormal basis of $\mathcal{K}_j$. After j steps, particular products of the process are the matrices $V_j \equiv (v_1, \ldots, v_j)$ and upper-Hessenberg

$$H_j \equiv \begin{pmatrix} h_{11} & \cdots & h_{1j} \\ h_{21} & \cdots & h_{2j} \\ & \ddots & \vdots \\ & & h_{j+1,j} \end{pmatrix}.$$

The relationship between V_j and H_j is characterized by

$$AV_j = \begin{cases} V_{j+1}H_j & \text{if } h_{j+1,j} \neq 0, \\ V_j \bar{H}_j & \text{if } h_{j+1,j} = 0, \end{cases}$$

where $\bar{H}_j$ is obtained from H_j by deleting the $(j+1)$st row.

The iterates of the GMRES and Arnoldi methods applied to $Ax = b$ are uniquely determined by the matrices V_j and H_j. In the case of the GMRES method, an approximate solution is defined at each iteration; in the case of the Arnoldi method, an approximate solution may not be defined at every iteration, depending on A and b. However, for both methods, with A invertible, the approximate solution is defined and equal to the exact solution at some step if and only if Algorithm 2.1 terminates at that step. In both methods, the approximate solution is computed only upon termination. During the intermediate iterations, the value of the residual norm is maintained recursively without computing the approximate solution, and termination is based on this value, which we refer to below as the *recursive residual norm*.

To formulate our approximate finite-difference Arnoldi process, we consider a Jacobian system

$$F'(u)x = c \tag{2.2}$$

and outline a process analogous to Algorithm 2.1 that uses approximations of the form (2.1). In this, we use "~" to denote approximate counterparts of quantities in Algorithm 2.1 that are derived from $\tilde{F}$.

ALGORITHM 2.2: APPROXIMATE FINITE-DIFFERENCE ARNOLDI PROCESS

GIVEN x_0, SET $q_0 = \frac{\tilde{F}(u,u+\sigma_0 x_0)-F(u)}{\sigma_0}$ AND $\tilde{r}_0 = c - q_0$.
IF $\tilde{r}_0 = 0$, STOP; OTHERWISE, SET $\tilde{v}_1 = \tilde{r}_0/\|\tilde{r}_0\|$.
FOR $j = 1, 2, \ldots,$ DO:
SET $q_j = \frac{\tilde{F}(u,u+\sigma_j \tilde{v}_j)-F(u)}{\sigma_j}$ AND $\tilde{w}_j = q_j - \sum_{i=1}^{j} \tilde{h}_{ij}\tilde{v}_i$,
WHERE $\tilde{h}_{ij} = \langle q_j, \tilde{v}_i \rangle$.
IF $\tilde{h}_{j+1,j} \equiv \|\tilde{w}_j\| = 0$, STOP; OTHERWISE, SET
$\tilde{v}_{j+1} = \tilde{w}_j/\tilde{h}_{j+1,j}$ AND CONTINUE.

If $\tilde{r}_0 \neq 0$, then the process terminates at step $j > 0$ if and only if $\tilde{h}_{j+1,j} = 0$. For each j up to termination, $\{\tilde{v}_i\}_{i=1,\ldots,j}$ is an orthonormal set; hence, termination must occur for some $j \leq n$.

The GMRES and Arnoldi methods applied to (2.2) can be based on Algorithm 2.2 in an obvious way. In particular, for each j, we define $\tilde{V}_j \equiv (\tilde{v}_1, \ldots, \tilde{v}_j)$ and upper-Hessenberg

$$\tilde{H}_j \equiv \begin{pmatrix} \tilde{h}_{11} & \cdots & \tilde{h}_{1j} \\ \tilde{h}_{21} & \cdots & \tilde{h}_{2j} \\ & \ddots & \vdots \\ & & \tilde{h}_{j+1,j} \end{pmatrix}.$$

Then, like their counterparts V_j and H_j produced by Algorithm 2.1, the matrices $\tilde{V}_j$ and $\tilde{H}_j$ uniquely determine the iterates of the GMRES and Arnoldi methods based on Algorithm 2.2. As before, the GMRES approximate solution is defined at every iteration, but the Arnoldi approximate solution may not be.

We further define

$$\begin{aligned} e_0 &\equiv q_0 - F'(u)x_0, \\ e_i &\equiv q_i - F'(u)\tilde{v}_i, \quad i = 1, \ldots, j. \end{aligned} \tag{2.3}$$

With the orthonormality of $\{\tilde{v}_i\}_{i=1,\ldots,j}$, one easily verifies that

$$\{F'(u) + e_i \tilde{v}_i^T\}\tilde{v}_i = q_i, \quad i = 1, \ldots, j,$$

and, therefore,

$$\{F'(u) + E_j\}\tilde{V}_j = (q_1, \ldots, q_j) = \begin{cases} \tilde{V}_{j+1}\tilde{H}_j & \text{if } \tilde{h}_{j+1,j} \neq 0, \\ \tilde{V}_j \bar{\tilde{H}}_j & \text{if } \tilde{h}_{j+1,j} = 0, \end{cases} \tag{2.4}$$

where

$$E_j \equiv (e_1, \ldots, e_j)\tilde{V}_j^T \tag{2.5}$$

and $\bar{\tilde{H}}_j \in \mathbb{R}^{j \times j}$ is obtained from $\tilde{H}_j$ by deleting the $(j+1)$st row.

In view of (2.3)–(2.5), the following theorem is immediate.

THEOREM 2.3. *Suppose that, for some $j \in \{1, \ldots, n\}$, $\tilde{V}_j$ and $\tilde{H}_j$ have been produced by j steps of Algorithm 2.2 with an initial x_0. Then $\tilde{V}_j$ and $\tilde{H}_j$ are the same*

as, respectively, V_j and H_j produced by j steps of Algorithm 2.1 *with the same initial x_0 and with $A = F'(u) + E_j$ and $b = c - e_0 + E_j x_0$, where E_j and e_0 are defined by* (2.3) *and* (2.5).

Proof. In Algorithm 2.2, one has $\tilde{v}_1 = \tilde{r}_0/\|\tilde{r}_0\|$ and, with (2.3),

$$\tilde{r}_0 = c - q_0 = c - e_0 + E_j x_0 - (F'(u) + E_j)\, x_0.$$

The theorem follows from this and (2.4). ∎

COROLLARY 2.4. *Suppose that, for some $j \in \{1, \ldots, n\}$, the GMRES or Arnoldi method based on Algorithm* 2.2 *has been applied to* (2.2) *for j steps. If $F'(u) + E_j$ is invertible, then the approximate solution determined at step j by the method (if defined) is the same as the approximate solution of the perturbed system*

$$\{F'(u) + E_j\}x = c - e_0 + E_j x_0 \tag{2.6}$$

determined at step j by the same method based on Algorithm 2.1. *The approximate solution at step j is defined and is the exact solution of* (2.6) *if and only if Algorithm* 2.2 *terminates at step j.*

2.2. A local convergence analysis. Newton–Krylov methods are usually implemented as *inexact Newton methods* [5], in which the Newton equation (1.2) is replaced with an *inexact Newton condition*

$$\|F(u) + F'(u)s\| \le \eta\|F(u)\|, \tag{2.7}$$

where $\eta \in [0, 1)$. In a Newton–Krylov implementation, (2.7) normally provides a criterion for terminating the Krylov iterations: one first chooses $\eta \in [0, 1)$ and then applies the Krylov solver to (1.2) until (2.7) holds. Used in this way, η is often called a *forcing term* (cf. [7]). In the seminal paper [5], it is shown that the local convergence of inexact Newton methods is determined by the forcing terms.

In the context of interest here, the Krylov solver is the GMRES or Arnoldi method based on Algorithm 2.2, and the approximate solution it produces is actually that of the perturbed linear system (2.6) with $c = -F(u)$. Thus an accurate evaluation of the linear residual norm $\|F(u) + F'(u)s\|$ is very likely to be undesirably expensive or unavailable. If this is the case, then (2.7) cannot be used directly for terminating the Krylov iterations. A likely course is to replace $\|F(u) + F'(u)s\|$ in (2.7) by the recursive residual norm maintained by the Krylov solver. This leads to the following algorithm.

ALGORITHM 2.5: APPROXIMATE FINITE-DIFFERENCE GMRES/ARNOLDI NEWTON–KRYLOV METHOD

SUPPOSE THAT u_0 IS GIVEN.

FOR $k = 0, 1, \ldots,$ DO:

CHOOSE $\eta_k \in [0, 1)$ AND APPLY THE GMRES OR ARNOLDI METHOD BASED ON ALGORITHM 2.2 TO $F'(u_k)s = -F(u_k)$ TO OBTAIN AN APPROXIMATE SOLUTION s_k FOR WHICH THE RECURSIVE RESIDUAL NORM IS LESS THAN OR EQUAL TO $\eta_k\|F(u_k)\|$.

SET $u_{k+1} = u_k + s_k$.

In the remainder of this subsection, we develop a local convergence analysis for Algorithm 2.5. This culminates in our main result (Theorem 2.12) to the effect that, near a solution of (1.2) at which the Jacobian is invertible, Algorithm 2.5 does not break down and enjoys desirable convergence properties if the difference steps are chosen sufficiently small and the forcing terms are chosen appropriately.

We begin with Lemma 2.6 below, which shows that, if the recursive residual norm satisfies a specified inexact Newton condition, then the actual residual norm satisfies a related inexact Newton condition. This lemma is a perturbation result based on Corollary 2.4. With $c = -F(u)$, the perturbed linear system (2.6) can be viewed as a specific instance of a more general perturbed linear system considered by Catinas in [4]. Perturbation results are given in [4] that are somewhat similar in spirit to Lemma 2.6 but are not as directly suited to our purposes.

For convenience in Lemma 2.6 and that which follows, we regard the initial residual norm as the recursive residual norm at step zero. Throughout the discussion, e_0 and E_j for $j > 0$ are defined by (2.3) and (2.5), with x_0 the initial approximate solution in the Krylov solver and with $c = -F(u)$ in Algorithm 2.2.

LEMMA 2.6. *Suppose that $\eta \in [0, 1)$ is given and that the GMRES or Arnoldi method based on Algorithm* 2.2 *is applied to* (1.2) *until, for some $j \in \{0, \ldots, n\}$, the recursive residual norm is less than or equal to $\eta\|F(u)\|$ at step j. Let s denote the approximate solution of* (1.2) *obtained by the method at step j. If $j = 0$, or if $j > 0$ and $F'(u) + E_j$ is invertible, then*

$$\|F(u) + F'(u)s\| \le \tilde{\eta}\|F(u)\|, \tag{2.8}$$

where

$$\tilde{\eta} \equiv \begin{cases} \eta + \dfrac{\|e_0\|}{\|F(u)\|} & \text{if } j = 0, \\ \eta + (1+\eta)\|E_j(F'(u)+E_j)^{-1}\| & \\ \qquad + \left[1 + \|E_j(F'(u)+E_j)^{-1}\|\right] \dfrac{\|e_0 - E_j\, x_0\|}{\|F(u)\|} & \text{if } j > 0. \end{cases} \tag{2.9}$$

Remark. If the initial approximate solution x_0 in the GMRES or Arnoldi method is zero, as is usually the case in Newton–Krylov methods, then $e_0 = 0$ and (2.9) reduces to

$$\tilde{\eta} \equiv \begin{cases} \eta & \text{if } j = 0, \\ \eta + (1+\eta)\|E_j(F'(u)+E_j)^{-1}\| & \text{if } j > 0. \end{cases} \tag{2.10}$$

Proof. If $j = 0$, then $s = x_0$ in Algorithm 2.2, and, with q_0 as in Algorithm 2.2, we have

$$\begin{aligned} \|F(u) + F'(u)s\| &\le \|F(u) + q_0\| + \|F'(u)s - q_0\| \\ &\le \eta\|F(u)\| + \|e_0\| \\ &= \left(\eta + \frac{\|e_0\|}{\|F(u)\|}\right)\|F(u)\|. \end{aligned}$$

If $j > 0$, then, by using Corollary 2.4 with $c = -F(u)$, we have

$$\begin{aligned} \|F(u) + F'(u)s\| &\le \|F(u) + e_0 - E_j\, x_0 + (F'(u) + E_j)\, s\| \\ &\qquad + \|E_j\, s\| + \|e_0 - E_j\, x_0\| \\ &\le \eta\|F(u)\| + \|E_j(F'(u)+E_j)^{-1}\| \cdot \|(F'(u) + E_j)\, s\| \\ &\qquad + \|e_0 - E_j\, x_0\|. \end{aligned} \tag{2.11}$$

Writing

$$(F'(u)+E_j)s = -F(u) - e_0 + E_j x_0 + r, \quad \|r\| \le \eta\|F(u)\|,$$

one verifies that

$$\begin{aligned}\|(F'(u)+E_j)s\| &\le (1+\eta)\|F(u)\| + \|e_0 - E_j\, x_0\| \\ &= \left(1+\eta+\frac{\|e_0 - E_j\, x_0\|}{\|F(u)\|}\right)\|F(u)\|.\end{aligned}$$

Substituting this in (2.11) and rearranging gives

$$\begin{aligned}\|F(u)+F'(u)s\| &\le \left\{\eta + \|E_j(F'(u)+E_j)^{-1}\|\cdot\left(1+\eta+\frac{\|e_0 - E_j\, x_0\|}{\|F(u)\|}\right)\right. \\ &\qquad \left. + \frac{\|e_0 - E_j\, x_0\|}{\|F(u)\|}\right\}\|F(u)\| \\ &= \left\{\eta + (1+\eta)\|E_j(F'(u)+E_j)^{-1}\|\right. \\ &\qquad \left. + \left[1+\|E_j(F'(u)+E_j)^{-1}\|\right]\frac{\|e_0 - E_j\, x_0\|}{\|F(u)\|}\right\}\|F(u)\|. \qquad \square\end{aligned}$$

We now formulate the first of two additional assumptions on $\tilde{F}$ and establish Lemma 2.8 and Corollary 2.9 below, which will be helpful in developing our local convergence results and also in justifying the assumptions in Lemma 2.6 near a solution of (1.2) at which the Jacobian is invertible.

Assumption 2.7. There exists $\Omega \subseteq \mathbb{R}^n$ such that

$$\omega(\sigma) \equiv \sup_{u\in\Omega,\ \|v\|\le 1}\left\|\frac{\tilde{F}(u,u+\sigma v)-F(u)}{\sigma} - F'(u)v\right\| \tag{2.12}$$

satisfies $\omega_* \equiv \limsup_{\sigma\to 0}\omega(\sigma) < \infty$.

The role of Assumption 2.7 in that which follows is to ensure that approximations of the form (2.1) are uniformly accurate when σ is appropriately small. We note that, if this assumption holds with $\omega_* = 0$, then the Frechet derivative $\tilde{F}_w(u,w) \equiv \frac{\partial \tilde{F}(u,w)}{\partial w}$ exists for $u \in \Omega$ and $w = u$ and, moreover, $\tilde{F}_w(u,u) = F'(u)$ for $u \in \Omega$.

LEMMA 2.8. *Suppose that Assumption* 2.7 *holds and that* $u_* \in \Omega$ *is such that* $F'(u_*)$ *is invertible and*

$$n\|F'(u_*)^{-1}\|\omega_* < 1/2. \tag{2.13}$$

Then, for any ϵ *such that*

$$\frac{n\|F'(u_*)^{-1}\|\omega_*}{1-n\|F'(u_*)^{-1}\|\omega_*} < \epsilon < 1, \tag{2.14}$$

there exist $\delta > 0$ *and* $\sigma_* > 0$ *such that if* $u \in B(u_*,\delta)$ *and, for some* $j \in \{1,\dots,n\}$, *Algorithm* 2.2 *has been applied with* $0 < |\sigma_i| \le \sigma_*$ *for* $i = 1, \dots, j$, *then* $F'(u)+E_j$ *is invertible, and*

$$\|E_j(F'(u)+E_j)^{-1}\| \le \epsilon. \tag{2.15}$$

Moreover, there is a λ independent of j and u for which

$$\|E_j(F'(u)+E_j)^{-1}\| \le \lambda \max_{1\le i\le j}\omega(\sigma_i)\,. \tag{2.16}$$

Proof. Since F is continuously differentiable and $F'(u_*)$ is invertible, there exists $\delta>0$ such that $B(u_*,\delta)\subset\Omega$, $F'(u)$ is invertible whenever $u\in B(u_*,\delta)$, and $\sup_{u\in B(u_*,\delta)}\|F'(u)^{-1}\|<\infty$. Moreover, in view of (2.13) and (2.14), we can choose δ and $\sigma_*>0$ sufficiently small to have $n\|F'(u)^{-1}\|\omega(\sigma)\le 1/2$ and

$$\frac{n\|F'(u)^{-1}\|\omega(\sigma)}{1-n\|F'(u)^{-1}\|\omega(\sigma)}\le\epsilon \tag{2.17}$$

whenever $u\in B(u_*,\delta)$ and $0<|\sigma|\le\sigma_*$. Note that, for $0<|\sigma|\le\sigma_*$, (2.17) implies that

$$\omega(\sigma)\le\frac{\epsilon}{(1+\epsilon)n\|F'(u)^{-1}\|}. \tag{2.18}$$

Suppose that $u\in B(u_*,\delta)$ and that Algorithm 2.2 has been applied with $0<|\sigma_i|\le\sigma_*$ for $i=1,\ldots,j$. Then

$$\begin{aligned}\|E_j\| &= \max_{\|v\|=1}\|E_jv\| = \max_{\|v\|=1}\|(e_1,\ldots,e_j)\tilde V_j^Tv\|\\ &= \max_{\|v\|=1}\left\|\sum_{i=1}^j\langle\tilde v_i,v\rangle\,e_i\right\| \le \max_{\|v\|=1}\sum_{i=1}^j|\,\langle\tilde v_i,v\rangle\,|\,\|e_i\|\\ &\le \textstyle\sum_{i=1}^j\|e_i\|\end{aligned} \tag{2.19}$$

since $|\,\langle\tilde v_i,v\rangle\,|\le\|\tilde v_i\|\,\|v\|=1$. Since

$$\|e_i\| = \|q_i-F'(u)\tilde v_i\| = \left\|\frac{\tilde F(u,u+\sigma_i\tilde v_i)-F(u)}{\sigma_i}-F'(u)\tilde v_i\right\|\le\omega(\sigma_i)$$

for each i, it follows from (2.19) and (2.18) that

$$\|E_j\|\le j\max_{1\le i\le j}\omega(\sigma_i)\le n\max_{1\le i\le j}\omega(\sigma_i)\le\frac{\epsilon}{(1+\epsilon)\|F'(u)^{-1}\|}. \tag{2.20}$$

From (2.20), we have

$$\|E_jF'(u)^{-1}\|\le\|E_j\|\,\|F'(u)^{-1}\|\le\frac{\epsilon}{1+\epsilon}<1, \tag{2.21}$$

and it follows by using well-known arguments (see, e.g., [6, Thm. 3.1.4]) that $I+E_jF'(u)^{-1}$ is invertible and

$$\left\|\left(I+E_jF'(u)^{-1}\right)^{-1}\right\|\le\frac{1}{1-\epsilon/(1+\epsilon)}=1+\epsilon. \tag{2.22}$$

Consequently, $F'(u)+E_j=(I+E_jF'(u)^{-1})F'(u)$ is invertible, and

$$\begin{aligned}\left\|E_j(F'(u)+E_j)^{-1}\right\| &= \left\|E_jF'(u)^{-1}\left(I+E_jF'(u)^{-1}\right)^{-1}\right\|\\ &\le \|E_jF'(u)^{-1}\|\,\|\left(I+E_jF'(u)^{-1}\right)^{-1}\|\\ &\le \frac{\epsilon}{1+\epsilon}\cdot(1+\epsilon)=\epsilon\,.\end{aligned} \tag{2.23}$$

To complete the proof, one easily verifies from the inequalities in (2.18)–(2.23) that

$$\|E_j(F'(u)+E_j)^{-1}\| \le n \max_{1\le i\le j} \omega(\sigma_i)\, \|F'(u)^{-1}\|\, (1+\epsilon)\,.$$

It follows that (2.16) holds with $\lambda = n \sup_{u\in B(u_*,\delta)} \|F'(u)^{-1}\|\,(1+\epsilon)$. ∎

COROLLARY 2.9. *Suppose that Assumption* 2.7 *holds and that* $u_* \in \Omega$ *is such that* $F'(u_*)$ *is invertible and* $F'(u_*)^{-1}$ *satisfies* (2.13). *Then there exist* $\delta > 0$ *and* $\sigma_* > 0$ *such that if* $u \in B(u_*, \delta)$ *and the GMRES or Arnoldi method based on Algorithm* 2.2 *is applied to* (1.2) *with* $0 < |\sigma_j| \le \sigma_*$ *for each* j, *then for some* $j \in \{0, \dots, n\}$ *the approximate solution obtained by the method at step* j *is defined and the recursive residual norm is zero. In particular, given any* $\eta \in [0, 1)$, *the method determines at some step an approximate solution for which the recursive residual norm is less than or equal to* $\eta\|F(u)\|$.

Proof. Suppose that the GMRES or Arnoldi method based on Algorithm 2.2 is applied to (1.2). If $\tilde{r}_0 = 0$, then there is nothing to prove. If $\tilde{r}_0 \neq 0$, then Algorithm 2.2 terminates for some $j \in \{1, \dots, n\}$. By Lemma 2.8, there exist $\delta > 0$ and $\sigma_* > 0$ such that if $u \in B(u_*, \delta)$ and $0 < |\sigma_i| \le \sigma_*$ for $i = 1, \dots, j$, then $F'(u) + E_j$ is invertible. Consequently, assuming $u \in B(u_*, \delta)$ and $0 < |\sigma_i| \le \sigma_*$ for $i = 1, \dots, j$, one has from Corollary 2.4 that the approximate solution determined by the method at step j is defined and is the exact solution of the perturbed system

$$\{F'(u) + E_j\}x = -F(u) - e_0 + E_j x_0. \tag{2.24}$$

Since the recursive residual norm is the norm of the residual of (2.24), it follows that the recursive residual norm is zero. ∎

We use the following somewhat specialized assumption of Hölder continuity of $\tilde{F}_w$. For a general definition of Hölder continuity, see, e.g., [14, section 3.1.9].

Assumption 2.10. With Ω as in Assumption 2.7, there is a $\delta_\Omega > 0$ such that $\tilde{F}_w(u, u+v)$ exists for all $u \in \Omega$ and all v with $\|v\| \le \delta_\Omega$. Moreover, there exist γ and $p \in (0, 1]$ such that

$$\|\tilde{F}_w(u, u+v) - \tilde{F}_w(u, u)\| \le \gamma \|v\|^p$$

for all $u \in \Omega$ and all v with $\|v\| \le \delta_\Omega$.

LEMMA 2.11. *Suppose that Assumption* 2.7 *holds with* $\omega_* = 0$ *and that Assumption* 2.10 *also holds. Then* $\omega(\sigma)$ *given by* (2.12) *satisfies*

$$\omega(\sigma) \le \mu\, |\sigma|^p \tag{2.25}$$

for sufficiently small $|\sigma|$ *and* $\mu = \gamma(\delta_\Omega)^{1+p}/(1+p)$.

Proof. Assume for convenience that $\sigma > 0$. (The proof requires only trivial changes when $\sigma < 0$.) By recalling that $\omega_* = 0$ implies $F'(u) = \tilde{F}_w(u, u)$ and adapting familiar reasoning (see, e.g., [14, section 3.2.12]), we have for $u \in \Omega$, $\|v\| \le \delta_\Omega$, and

sufficiently small $\sigma > 0$ that

$$\begin{aligned}\|\tilde{F}(u, u+\sigma v) - F(u) - \sigma F'(u)v\| &= \left\| \left\{ \int_0^\sigma [\tilde{F}_w(u, u+\tau v) - \tilde{F}_w(u,u)]\, d\tau \right\} v \right\| \\ &\le \left\{ \int_0^\sigma \|\tilde{F}_w(u, u+\tau v) - \tilde{F}_w(u,u)\|\, d\tau \right\} \|v\| \\ &\le \left\{ \int_0^\sigma \gamma \tau^p\, d\tau \right\} \|v\|^{1+p} \\ &= \frac{\gamma}{1+p}\sigma^{1+p}\|v\|^{1+p} \le \frac{\gamma(\delta_\Omega)^{1+p}}{1+p}\sigma^{1+p},\end{aligned}$$

which immediately yields (2.25). ∎

The theorem below is our local convergence result for Algorithm 2.5. It can be viewed as a counterpart of Theorems 2.3 and 3.3 of [5], which are used in its proof. For definitions of the types of convergence referred to in the theorem, see, e.g., [6]. For convenience, we denote by σ_{jk} the jth difference step used by the GMRES or Arnoldi method at the kth iteration of Algorithm 2.5. We also assume that, at each iteration of Algorithm 2.5, the GMRES or Arnoldi method begins with the zero initial approximate solution, and, therefore, the applicable perturbation result from Lemma 2.6 is (2.8) followed by (2.10) rather than (2.9).

THEOREM 2.12. *Suppose that Assumption* 2.7 *holds and that $u_* \in \Omega$ is such that $F(u_*) = 0$, $F'(u_*)$ is invertible, and $n\|F'(u_*)^{-1}\|\omega_* < 1/2$. Then for any t_* and η_* such that*

$$\mu_* \equiv \frac{n\|F'(u_*)^{-1}\|\omega_*}{1 - n\|F'(u_*)^{-1}\|\omega_*} < t_* < 1 \quad \text{and} \quad 0 \le \eta_* < \frac{t_* - \mu_*}{1+\mu_*}, \tag{2.26}$$

there exist $\delta_* > 0$ and $\sigma_* > 0$ such that if $u_0 \in B(u_*, \delta_*)$, $0 \le \eta_k \le \eta_*$ for each k, and $0 < |\sigma_{jk}| \le \sigma_*$ for each j and k, then Algorithm 2.5 *produces $\{u_k\}_{k=0,1,\ldots}$ that converges to u_* with*

$$\|u_{k+1} - u_*\|_* \le t_*\|u_k - u_*\|_* \tag{2.27}$$

for each k, where $\|v\|_* \equiv \|F'(u_*)v\|$ for $v \in \mathbb{R}^n$. Additionally, if $\omega_* = 0$ and if $\eta_k \to 0$ and $\max_j |\sigma_{jk}| \to 0$ as $k \to \infty$, then the convergence is q-superlinear. If Assumption 2.10 *also holds, F' is Hölder continuous with exponent p at u_*, and $\eta_k = O(\|F(u_k)\|^p)$ and $\max_j |\sigma_{jk}| = O(\|F(u_k)\|)$ as $k \to \infty$, then the convergence is of q-order $1+p$.*

Proof. Suppose that t_* and η_* satisfy (2.26). Let ϵ be such that $\mu_* < \epsilon < t_*$ and $\eta_* < (t_* - \epsilon)/(1+\epsilon)$. Note that $\eta_{\max} \equiv \eta_* + (1+\eta_*)\epsilon < t_*$. By [5, Thm. 2.3], there is a $\delta_* > 0$ such that any sequence $\{u_k\}_{k=0,1,\ldots}$ for which $u_0 \in B(u_*, \delta_*)$ and

$$\left.\begin{aligned} u_{k+1} &= u_k + s_k \\ \|F(u_k) + F'(u_k)s_k\| &\le \eta_{\max}\|F(u_k)\| \end{aligned}\right\} \quad k = 0, 1, \ldots$$

satisfies (2.27) for each k and converges to u_*.

Let $\delta > 0$ and $\sigma_* > 0$ be chosen so that the conclusions of Lemma 2.8 and Corollary 2.9 hold. Let M be such that

$$\frac{1}{M}\|v\| \le \|v\|_* \le M\|v\|$$

for all $v \in \mathbb{R}^n$. By taking δ_* smaller if necessary, assume that $M^2\delta_* < \delta$. Suppose that Algorithm 2.5 is applied with $u_0 \in B(u_*, \delta_*)$, $0 \le \eta_k \le \eta_*$ for all k, and $0 < |\sigma_{jk}| \le \sigma_*$ for all j and k. If, for some $k \ge 0$, the algorithm has determined $\{u_j\}_{j=0,\dots,k} \subseteq B(u_*, \delta)$, then, at step k of the algorithm, it follows from Corollary 2.9 that the GMRES or Arnoldi method based on Algorithm 2.2 determines at some iteration an approximate solution s_k for which the recursive residual norm is less than or equal to $\eta_k\|F(u_k)\|$. Then u_{k+1} is defined. Moreover, with Lemmas 2.8 and 2.6, we have, by using (2.8)–(2.10) and (2.15), that

$$\|F(u_k) + F'(u_k)s_k\| \le [\eta_k + (1+\eta_k)\epsilon]\,\|F(u_k)\| \le \eta_{\max}\|F(u_k)\|,$$

and we can apply (2.27) to obtain

$$\begin{aligned}\|u_{k+1} - u_*\| &\le M\|u_{k+1} - u_*\|_* \le Mt_*^{k+1}\|u_0 - u_*\|_* \\ &\le M^2 t_*^{k+1}\|u_0 - u_*\| < M^2\delta_* < \delta.\end{aligned}$$

It follows inductively that Algorithm 2.2 produces $\{u_k\}_{k=0,1,\dots} \subseteq B(u_*, \delta)$ that converges to u_*, with (2.27) holding for each k.

Assuming $u_0 \in B(u_*, \delta_*)$, we also have from (2.8)–(2.10) and (2.16) that

$$\|F(u_k) + F'(u_k)s_k\| \le [\eta_k + (1+\eta_k)\,\lambda \max_j \omega(\sigma_{jk})]\,\|F(u_k)\| \tag{2.28}$$

for each k, where λ is independent of k. Consequently, if $\omega_* \equiv \lim_{\sigma\to 0}\omega(\sigma) = 0$ and if $\eta_k \to 0$ and $\max_j |\sigma_{jk}| \to 0$ as $k \to \infty$, then it follows from [5, Thm. 3.3] that $\{u_k\}_{k=0,1,\dots}$ converges to u_* q-superlinearly. If Assumption 2.10 also holds, then (2.28) and (2.25) yield

$$\|F(u_k) + F'(u_k)s_k\| \le [\eta_k + (1+\eta_k)\,\lambda\,\mu\,(\max_j |\sigma_{jk}|)^p]\,\|F(u_k)\|\,,$$

where μ is independent of k. It follows from [5, Thm. 3.3] that, if $\eta_k = O(\|F(u_k)\|^p)$ and $\max_j |\sigma_{jk}| = O(\|F(u_k)\|)$ as $k \to \infty$, then the convergence is of q-order $1+p$. ∎

Remark. The framework and analysis given above cover the case in which $F = F_1 + F_2$, where F_1 and F_2 are differentiable and such that, for each u and v, $F_1'(u)v$ can be readily evaluated but $F_2'(u)v$ must be approximated with a finite difference.[1] Indeed, setting $\tilde{F}(u, w) \equiv F_1(u) + F_1'(u)(w-u) + F_2(w)$, we obtain

$$\frac{\tilde{F}(u, u+\sigma v) - F(u)}{\sigma} = F_1'(u)v + \frac{F_2(u+\sigma v) - F_2(u)}{\sigma}.$$

Then Assumption 2.7 becomes

$$\omega_* \equiv \sup_{u\in\Omega,\ \|v\|\le 1} \left\| \frac{F_2(u+\sigma v) - F_2(u)}{\sigma} - F_2'(u)v \right\| < \infty,$$

and Assumption 2.10 becomes an assumption of Hölder continuity of $F_2'(u)$. With this understanding, it is straightforward to apply Corollary 2.9 and Theorem 2.12.

3. Numerical experiments. In this section we first describe the test problems considered. We then focus on two possible choices for $\tilde{F}$ and analyze to what degree they satisfy Assumptions 2.7 and 2.10. Finally, we present numerical results obtained with these choices on the test problems.

[1] This case can also be addressed with a modest extension of the results in [3].

TABLE 1
Test cases.

Test case	$D(u)$
1	$\sqrt{u^2+u+1}$
2	$\frac{1}{\sqrt{u^2+u+1}}$
3	$u^{3/5}e^u$
4	$K\sqrt{S(u)}[1-(1-S(u)^{\frac{1}{\mu}})^{\mu}]^2$, where $S(u)=(1+\lvert\alpha u\rvert^{\nu})^{-\mu}$ and K,α,μ,ν are constants

3.1. Test problems. The test problems are of the form

$$\mathcal{F}(u) \equiv \nabla \cdot D(u)\nabla u - g(u) + f = 0. \tag{3.1}$$

This class of nonlinear diffusion problems arises in many applications, such as flow through porous media, radiation transport, phase transition, biochemistry, and dynamics of biological groups. In our experiments, we considered these problems on square domains in $\mathbb{R}^2$ with homogeneous Dirichlet boundary conditions.

We are motivated by circumstances in which g-evaluations are inexpensive, but D-evaluations are relatively costly, and less expensive approximations may be desirable. In our experiments, we considered four choices for D, labeled test cases 1–4 in Table 1. For cases 1–3, we chose $g(u) \equiv u^2$. In case 4, we used $g(u) \equiv 0$. Except in test case 4, we took the domain to be $[0,1] \times [0,1]$ and chose f so that the solution of (3.1) was $u_c(x,y) \equiv cx(1-x)y(1-y)$, with $c = 1$, 2, 5, and 10. Choosing f in this way allowed us to determine success or failure with certainty and also to exercise some control over the magnitude of the solution and its derivatives. Test case 4 was inspired by a formulation of Richards's equation describing fluid flow in unsaturated porous media that appeared in [13]. In this case, the domain was $[0, \frac{1}{2}] \times [0, \frac{1}{2}]$ and f was chosen so that the solution was $u_c(x,y) \equiv \frac{c}{4}x(1-2x)y(1-2y) + \beta$, where $\beta > 0$ is a small constant chosen to avoid singularities ($\beta = \frac{1}{16}$ in the experiments reported here) and again $c = 1$, 2, 5, and 10.

In our tests, (3.1) was discretized by using central differences on an $m \times m$ grid of regularly spaced interior points with grid spacing $h = 1/(m+1)$ in test cases 1–3 and $h = 1/2(m+1)$ in test case 4. The resulting system of equations is

$$\begin{aligned} F_{i,j}(u) \equiv \frac{1}{h^2}\Bigg[& D\left(\frac{u_{i+1,j}+u_{i,j}}{2}\right)(u_{i+1,j}-u_{i,j}) - D\left(\frac{u_{i,j}+u_{i-1,j}}{2}\right)(u_{i,j}-u_{i-1,j}) \\ & + D\left(\frac{u_{i,j+1}+u_{i,j}}{2}\right)(u_{i,j+1}-u_{i,j}) - D\left(\frac{u_{i,j}+u_{i,j-1}}{2}\right)(u_{i,j}-u_{i,j-1})\Bigg] \\ & - g(u_{i,j}) + f(x_i,y_j) = 0 \end{aligned} \tag{3.2}$$

for $1 \le i \le m$ and $1 \le j \le m$, where (x_i, y_j) denotes the ijth grid point and $u_{i,j}$ denotes the approximate solution there. (The boundary conditions give $u_{0,j} = u_{m+1,j} = u_{i,0} = u_{i,m+1} = 0$.) We denote this system by $F(u) = 0$, where, for convenience, u denotes the vector of $u_{i,j}$'s as well as the solution of (3.1).

3.2. Choices of $\tilde{F}$. In our experiments, we considered two illustrative choices of $\tilde{F}$ based on our motivating assumption that D-evaluations are significantly more expensive than g-evaluations. (Other choices of $\tilde{F}$ are possible; see the remark at

the end of this subsection.) We refer to these choices as the "lagged" and "linear" approximations of F and denote them by $\tilde{F}^{\mathrm{Lag}}(u,w)$ and $\tilde{F}^{\mathrm{Lin}}(u,w)$, respectively. These correspond to the respective approximations

$$\tilde{\mathcal{F}}^{\mathrm{Lag}}(u,w) \equiv \nabla \cdot D(u)\nabla w - g(w) + f$$

and

$$\tilde{\mathcal{F}}^{\mathrm{Lin}}(u,w) \equiv \nabla \cdot [D(u) + D'(u)(w-u)]\nabla w - g(w) + f$$

of $\mathcal{F}$ in the continuous problem (3.1). They are defined by respective ijth components

$$\begin{aligned}
&\tilde{F}^{\mathrm{Lag}}_{i,j}(u,w) \\
&\quad \equiv \frac{1}{h^2}\left[D\left(\frac{u_{i+1,j}+u_{i,j}}{2}\right)(w_{i+1,j}-w_{i,j}) - D\left(\frac{u_{i,j}+u_{i-1,j}}{2}\right)(w_{i,j}-w_{i-1,j})\right. \\
&\qquad \left. + D\left(\frac{u_{i,j+1}+u_{i,j}}{2}\right)(w_{i,j+1}-w_{i,j}) - D\left(\frac{u_{i,j}+u_{i,j-1}}{2}\right)(w_{i,j}-w_{i,j-1})\right] \\
&\qquad - g(w_{i,j}) + f(x_i,y_j),
\end{aligned} \tag{3.3}$$

$$\begin{aligned}
&\tilde{F}^{\mathrm{Lin}}_{i,j}(u,w) \\
&\quad \equiv \frac{1}{h^2}\left\{\left[D\left(\frac{u_{i+1,j}+u_{i,j}}{2}\right) + \frac{1}{2}D'\left(\frac{u_{i+1,j}+u_{i,j}}{2}\right)(w_{i+1,j}+w_{i,j}-u_{i+1,j}-u_{i,j})\right]\right. \\
&\qquad \cdot (w_{i+1,j}-w_{i,j}) \\
&\qquad - \left[D\left(\frac{u_{i,j}+u_{i-1,j}}{2}\right) + \frac{1}{2}D'\left(\frac{u_{i,j}+u_{i-1,j}}{2}\right)(w_{i,j}+w_{i-1,j}-u_{i,j}-u_{i-1,j})\right] \\
&\qquad \cdot (w_{i,j}-w_{i-1,j}) \\
&\qquad + \left[D\left(\frac{u_{i,j+1}+u_{i,j}}{2}\right) + \frac{1}{2}D'\left(\frac{u_{i,j+1}+u_{i,j}}{2}\right)(w_{i,j+1}+w_{i,j}-u_{i,j+1}-u_{i,j})\right] \\
&\qquad \cdot (w_{i,j+1}-w_{i,j}) \\
&\qquad - \left[D\left(\frac{u_{i,j}+u_{i,j-1}}{2}\right) + \frac{1}{2}D'\left(\frac{u_{i,j}+u_{i,j-1}}{2}\right)(w_{i,j}+w_{i,j-1}-u_{i,j}-u_{i,j-1})\right] \\
&\qquad \left. \cdot (w_{i,j}-w_{i,j-1})\right\} - g(w_{i,j}) + f(x_i,y_j).
\end{aligned} \tag{3.4}$$

It is clear that both $\tilde{F}^{\mathrm{Lag}}(u,w)$ and $\tilde{F}^{\mathrm{Lin}}(u,w)$ satisfy $\tilde{F}(u,u) = F(u)$. To apply the theoretical results obtained in the previous section, we must determine whether they also satisfy Assumptions 2.7 and 2.10. To facilitate verifying these assumptions, we note that the discrete problem (3.2) is a particular case of the model problem

$$F(u) \equiv A(u)u - b(u) + c = 0, \tag{3.5}$$

where $u \in \mathbb{R}^n$, $A(u) \in \mathbb{R}^{n\times n}$ has continuously differentiable entries $a_{ij}(u)$, and $b(u) \in \mathbb{R}^n$ has continuously differentiable components $b_i(u)$. For this model problem, we have that

$$F'(u)v = A(u)v + [A'(u)v]u - b'(u)v,$$

where $[A'(u)v] \in \mathbb{R}^{n\times n}$ has entries $[A'(u)v]_{ij} = \nabla a_{ij}(u)^T v$ and $\nabla a_{ij}(u)$ denotes the gradient of $a_{ij}(u)$ with respect to u. We define

$$\tilde{F}^{\text{Lag}}(u,w) \equiv A(u)w - b(w) + c, \tag{3.6}$$

$$\tilde{F}^{\text{Lin}}(u,w) \equiv (A(u) + [A'(u)(w-u)])\, w - b(w) + c. \tag{3.7}$$

We show the following results for problem (3.5) and the approximations (3.6)–(3.7).

THEOREM 3.1. *Let F, $\tilde{F}^{\text{Lag}}$, and $\tilde{F}^{\text{Lin}}$ be defined by* (3.5)–(3.7). *Then*

$$\frac{\tilde{F}^{\text{Lag}}(u,u+\sigma v) - F(u)}{\sigma} - F'(u)v = -[A'(u)v]u - \left[\frac{b(u+\sigma v) - b(u)}{\sigma} - b'(u)v\right] \tag{3.8}$$

and

$$\frac{\tilde{F}^{\text{Lin}}(u,u+\sigma v) - F(u)}{\sigma} - F'(u)v = \sigma[A'(u)v]v - \left[\frac{b(u+\sigma v) - b(u)}{\sigma} - b'(u)v\right]. \tag{3.9}$$

If, in addition, there exists an $\Omega \subseteq \mathbb{R}^n$ such that, for all $u \in \Omega$, $|\partial a_{ij}(u)/\partial u_k| \le C_1 < \infty$ and $|\partial b_i(u)/\partial u_j| \le C_2 < \infty$ for all $i,j,k \in \{1,\dots,n\}$ and some constants C_1 and C_2, then both $\tilde{F}^{\text{Lag}}$ and $\tilde{F}^{\text{Lin}}$ satisfy Assumption 2.7 *with $\omega_*^{\text{Lag}} < \infty$ for $\tilde{F}^{\text{Lag}}$ and $\omega_*^{\text{Lin}} = 0$ for $\tilde{F}^{\text{Lin}}$.*

Proof. The proof follows by substituting the definitions of the lagged and linear approximations into the above expressions and then deleting appropriate terms. ∎

THEOREM 3.2. *Let the assumptions regarding $\tilde{F}^{\text{Lin}}$ in Theorem* 3.1 *hold. In addition, assume that there exist γ and $p \in (0,1]$ such that*

$$\|b'(u+v) - b'(u)\| \le \gamma \|v\|^p$$

for all $u \in \Omega$ and all v, with $\|v\| \le \delta_\Omega$. Then Assumption 2.10 *holds for $\tilde{F}^{\text{Lin}}$ with the same value of p.*

Proof. We first show that, for any v,

$$\tilde{F}^{\text{Lin}}_w(u,w)v = A(u)v + [A'(u)(w-u)]v + [A'(u)v]w - b'(w)v.$$

We have

$$\begin{aligned}
&\sigma^{-1}\left[\tilde{F}^{\text{Lin}}(u,w+\sigma v) - \tilde{F}^{\text{Lin}}(u,w)\right] \\
&\quad = \sigma^{-1}\left[(A(u) + [A'(u)(w+\sigma v - u)])(w+\sigma v) - (A(u) + [A'(u)(w-u)])w\right] \\
&\qquad - \sigma^{-1}\left[b(w+\sigma v) - b(w)\right] \\
&\quad = A(u)v + [A'(u)(w-u)]v + [A'(u)v](w+\sigma v) - \sigma^{-1}\left[b(w+\sigma v) - b(w)\right].
\end{aligned}$$

Letting $\sigma \to 0$ gives the result. Next, for any vector $\hat{v}$

$$\begin{aligned}
&\tilde{F}^{\text{Lin}}_w(u,u+v)\hat{v} - F'(u)\hat{v} \\
&\qquad = A(u)\hat{v} + [A'(u)v]\hat{v} + [A'(u)\hat{v}](u+v) - b'(u+v)\hat{v} \\
&\qquad\quad - A(u)\hat{v} - [A'(u)\hat{v}]u + b'(u)\hat{v} \\
&\qquad = [A'(u)v]\hat{v} + [A'(u)\hat{v}]v - (b'(u+v) - b'(u))\hat{v}.
\end{aligned}$$

Letting $\|\cdot\|_F$ denote the Frobenius norm on $\mathbb{R}^{n\times n}$, we have

$$\|[A'(u)v]\|_F = \left(\sum_{i,j=1}^n (\nabla a_{ij}(u)^T v)^2\right)^{1/2} \le \|v\| \left(\sum_{i,j=1}^n \|\nabla a_{ij}(u)\|^2\right)^{1/2} \le K \cdot \|v\|$$

for some constant K since $|\partial a_{ij}(u)/\partial u_k| \le C_1 < \infty$ for all $i, j, k = 1, \ldots, n$. Thus,

$$\begin{aligned}\|\tilde{F}_w^{\mathrm{Lin}}(u, u+v)\hat{v} - F'(u)\hat{v}\| &\le 2K \cdot \|v\| \cdot \|\hat{v}\| + \|b'(u+v) - b'(u)\| \cdot \|\hat{v}\| \\ &\le 2K \cdot \|v\| \cdot \|\hat{v}\| + \gamma\|v\|^p \cdot \|\hat{v}\| \\ &= \left(2K\delta_\Omega^{1-p} + \gamma\right)\|v\|^p \cdot \|\hat{v}\|\end{aligned}$$

for any vector $\hat{v}$. Therefore,

$$\|\tilde{F}_w^{\mathrm{Lin}}(u, u+v) - F'(u)\| \le \left(2K\delta_\Omega^{1-p} + \gamma\right)\|v\|^p$$

for all $u \in \Omega$ and all v with $\|v\| \le \delta_\Omega$, which completes the proof. □

For the discrete problem (3.2), a short computation reveals that $\tilde{F}^{\mathrm{Lag}}$ and $\tilde{F}^{\mathrm{Lin}}$ given by (3.3) and (3.5) both satisfy the assumptions of Theorem 3.1 if there exists a convex set Ω such that $|D'(u)|$ and $|g'(u)|$ are uniformly bounded for $u \in \Omega$. Additionally, $\tilde{F}^{\mathrm{Lin}}$ satisfies the assumptions of Theorem 3.2 if $g'(u)$ is Hölder continuous in Ω. Thus, under these mild assumptions, the local convergence results in section 2.2 can be applied to these two cases.

We note that using $\tilde{F} = \tilde{F}^{\mathrm{Lag}}$ given by (3.3) requires minimal new calculations at each step of the linear solve in Algorithm 2.5. Indeed, each step requires only one g-evaluation at each grid point in addition to a very modest amount of arithmetic, since the necessary D- and f-values are already available. Thus, the computational effort required to implement $\tilde{F}^{\mathrm{Lag}}$ is so small that the time savings may be considerable for problems for which convergence can be achieved with this approximation.

Using $\tilde{F} = \tilde{F}^{\mathrm{Lin}}$ given by (3.5) in Algorithm 2.5 requires evaluating D' at points intermediate to the grid points, which may be problematic if D'-evaluations are expensive or unavailable. However, this needs to be done only once at the outset of each linear solve and, in many applications, may be no more difficult than evaluating D at those points. Once the necessary D'-values have been computed, each step of the linear solve requires the same g-evaluations as $\tilde{F}^{\mathrm{Lag}}$ together with a somewhat greater but still very modest amount of arithmetic.

Remark. One of the referees suggested an alternative to (3.7) in defining $\tilde{F}^{\mathrm{Lin}}$ for the model problem (3.5), viz.,

$$\tilde{F}^{\mathrm{Lin}}(u, w) = A(u)w + [A'(u)(w-u)]u - b(w) + c. \tag{3.10}$$

This has the error term

$$\frac{\tilde{F}^{\mathrm{Lin}}(u, u+\sigma v) - F(u)}{\sigma} - F'(u)v = -\left[\frac{b(u+\sigma v) - b(u)}{\sigma} - b'(u)v\right],$$

which is slightly simpler than (3.9). Implementing this $\tilde{F}^{\mathrm{Lin}}$ on the discrete problem (3.2) has essentially the same cost as implementing $\tilde{F}^{\mathrm{Lin}}$ given by (3.7). With this $\tilde{F}^{\mathrm{Lin}}$, the analysis in section 2.2 can be applied as in the remark at the end of section 2.2 by writing F in (3.5) as $F = F_1 + F_2$, where $F_1(u) \equiv A(u)u$ and $F_2(u) \equiv -b(u) + c$. Then (3.10) becomes

$$\tilde{F}^{\mathrm{Lin}}(u, w) = F_1(u) + F_1'(u)(w-u) + F_2(w),$$

and Assumptions 2.7 and 2.10 become assumptions on F_2 and, hence, on b as in the remark at the end of section 2.2.

TABLE 2
Key to abbreviations.

Abbreviation	Meaning
NEQ	Number of equations (unknowns)
NNI	Number of nonlinear iterations
NLI	Number of linear iterations

3.3. Numerical results. In our numerical experiments, the test problems described above were treated by applying the KINSOL Newton–Krylov code from the SUNDIALS suite [10]; see [10] for details of the algorithm not specified here. KINSOL was applied in matrix-free mode with modifications to allow the approximation (2.1) with $\tilde{F} = \tilde{F}^{\mathrm{Lag}}$ and $\tilde{F} = \tilde{F}^{\mathrm{Lin}}$ to be used in place of (1.4). The linear solver used was GMRES with a banded block-diagonal preconditioner provided by SUNDIALS using finite differences of F-values; see [10]. The linear and nonlinear solver specifications were as follows: the maximum number of GMRES iterations allowed was 100, with no restarts; each forcing term η_k was the constant 10^{-3}; the nonlinear iterations terminated successfully if $\|F(u_k)\| \leq 10^{-8}$. These tight tolerances were chosen to bring out performance differences resulting from the different choices of $\tilde{F}$. Additionally, failure was declared if either the number of nonlinear iterations exceeded 200 or $\|s_k\| \leq 10^{-13}$; however, these failure modes were not observed in our tests.

All runs were done on the ASC Frost machine at Lawrence Livermore National Laboratory, an IBM SP parallel platform running the AIX operating system with 1088 375 MHz processors grouped into 16-processor nodes. In each of our runs, the spatial domain was subdivided into a $p \times p$ array of square subdomains having an equal number of grid points, and the grid points in each subdomain were mapped to one processor.

In test cases 1–3, the domain was divided into 400×400 zones yielding $160,000$ unknowns. In test case 4, a smaller problem was also considered, in which the domain was divided into 200×200 zones yielding 40,000 unknowns. In all cases, the grid was adjusted so that the number of grid points per subdomain (and processor) was always 10,000. Thus, for the 160,000-unknown problems, 16 processors were used, and four were used for the smaller test case 4 problem.

The results of our runs are given in the tables below. A key to the abbreviations used in these tables is provided in Table 2. In Tables 3–6, "None" refers to using no approximation, i.e., using (1.4); "Linear" and "Lagged" refer to using $\tilde{F} = \tilde{F}^{\mathrm{Lag}}$ and $\tilde{F} = \tilde{F}^{\mathrm{Lin}}$, respectively, in (2.1). Run times are given in seconds, and each normalized run time is calculated by dividing the run time by that obtained by using no approximation.

3.3.1. Case 1: $D(u) = \sqrt{u^2 + u + 1}$. This choice for D is the least expensive to evaluate of all of those we experimented with, and so we expected the time savings using the approximations $\tilde{F}^{\mathrm{Lag}}$ and $\tilde{F}^{\mathrm{Lin}}$ to be less significant than in the other test problems. Recall that the problem was posed so that the exact solution was $u_c(x, y) = cx(1-x)y(1-y)$. Here we took $c = 10$ with the initial approximate solution $u_0 \equiv c$. (No significant additional information was gained in this case by considering the other values of c.) The results are given in Table 3. There were no unsuccessful runs in this case.

As expected, since D is not very expensive to evaluate, using the approximations $\tilde{F}^{\mathrm{Lag}}$ and $\tilde{F}^{\mathrm{Lin}}$ did not significantly reduce computational time. In fact, the lagged approximation required more GMRES iterations and, consequently, more run time than

TABLE 3
Case 1: $D(u) = \sqrt{u^2 + u + 1}$, $c = 10$.

$\tilde{F}$ approx.	NNI	NLI	Run time	Norm. run time
None	9	416	38.2	1.00
Linear	9	417	35.3	0.92
Lagged	9	530	43.8	1.15

TABLE 4
Case 2: $D(u) = \frac{1}{\sqrt{u^2+u+1}}$.

c	$\tilde{F}$ approx.	NNI	NLI	Run time	Norm. run time
1	None	7	329	32.7	1.00
	Linear	7	331	30.3	0.93
	Lagged	6	323	29.1	0.89
2	None	12	387	41.1	1.00
	Linear	12	387	38.2	0.93
	Lagged	8	442	37.7	0.92
5	None	7	329	32.1	1.00
	Linear	7	329	29.5	0.92
	Lagged	8	436	37.1	1.16
10	None	7	345	33.7	1.00
	Linear	7	343	30.8	0.91
	Lagged	10	566	47.2	1.40

the two alternatives, apparently because of less compatibility between the approximate Jacobian-vector products obtained with $\tilde{F}^{\mathrm{Lag}}$ and the preconditioner provided by SUNDIALS using exact F-values. Consistent with the theory, there is very close agreement between the numbers of nonlinear and linear iterations obtained by using $\tilde{F}^{\mathrm{Lin}}$ and those obtained by using no approximation.

3.3.2. Case 2: $D(u) = \frac{1}{\sqrt{u^2+u+1}}$. This choice for D is only slightly more expensive to evaluate than the choice in Case 1; however, there are some notable differences in the computational results. For $c = 1$ and $c = 2$, we took $u_0 \equiv c$ as in Case 1. However, for $c = 5$ and $c = 10$, we took $u_0 \equiv 1$; this was necessary in order to obtain convergence to the solution, even when no approximation was used. With these choices of u_0, all runs were successful. The results are shown in Table 4.

This choice for D is interesting because the results with the lagged approximation $\tilde{F}^{\mathrm{Lag}}$ show sensitivity to the value of c. As c gets larger, the lagged approximation becomes less accurate, and, as a result, its performance degrades. The results obtained by using the linear approximation $\tilde{F}^{\mathrm{Lin}}$ are similar to those in Case 1. Again we note very close agreement between the numbers of linear and nonlinear iterations obtained by using $\tilde{F}^{\mathrm{Lin}}$ and those obtained by using no approximation.

3.3.3. Case 3: $D(u) = u^{\frac{3}{5}}e^u$. This choice for D entails the largest values of $D'(u)$ at the solutions $u_c(x, y) = cx(1 - x)y(1 - y)$, and so we expected differences in the effectiveness of $\tilde{F}^{\mathrm{Lag}}$ and $\tilde{F}^{\mathrm{Lin}}$ as approximations in (2.1) to be significant and to become increasingly pronounced as c increases. We took the initial approximate solution to be $u_0 \equiv c$ for each value of c. With this u_0, there were two failures when $\tilde{F}^{\mathrm{Lag}}$ was used. The results are given in Table 5.

TABLE 5
Case 3: $D(u) = u^{\frac{3}{5}} e^u$.

c	$\tilde{F}$ Approx.	NNI	NLI	Run time	Norm. run time
1.0	None	16	600	103.4	1.00
	Linear	16	598	81.5	0.79
	Lagged	33	2374	273.6	2.65
2.0	None	17	645	110.0	1.00
	Linear	17	632	85.1	0.77
	Lagged	35	2503	287.0	2.61
5.0	None	19	771	124.8	1.00
	Linear	19	782	98.0	0.79
	Lagged	-	-	-	-
10.0	None	23	1201	199.2	1.00
	Linear	23	1199	152.4	0.77
	Lagged	-	-	-	-

The results in Table 5 are consistent with our expectations. The lagged approximation resulted in failures of the method and greatly increased run times when it did not fail. (The failures resulted from GMRES failing to produce the requested residual norm reduction within the allowable 100 iterations.) In contrast, the linear approximation led to success in every case and to significant reductions in run times. For all values of c, as seen in the previous cases, there is very close agreement between the numbers of linear and nonlinear iterations obtained by using $\tilde{F}^{\mathrm{Lin}}$ and those obtained by using no approximation.

3.3.4. Case 4: $D(u) = K\sqrt{S(u)}[1 - (1 - S(u)^{\frac{1}{\mu}})^{\mu}]^2$. This problem is the most realistic of all of those considered here, and, of all of the choices for D, this is the most expensive to evaluate. The problem is related to a formulation of Richards's equation, which is often used for modeling fluid flows in unsaturated porous media. Here we chose values of the parameters appropriate for groundwater flow in an unsaturated dune sand (see [13] and the references therein), as follows: $K = 5.040$, $\alpha = 5.470$, $\nu = 4.264$, and $\mu = \frac{\nu-1}{\nu}$. In each test case, we began with $u_0 \equiv \beta = \frac{1}{16}$, for which all runs were successful. Results are given in Table 6.

As in the previous cases, there is very close agreement between the numbers of linear and nonlinear iterations obtained by using the linear approximation and those obtained by using no approximation. For the linear approximation, the greater runtime reductions on the larger problems are notable and likely reflect the benefit of amortizing the cost of evaluating D' over larger numbers of GMRES iterations per inexact Newton step. While not as effective as the linear approximation, the lagged approximation, for all values of c except 10, also resulted in significantly reduced run times, with greater reductions on the larger problems, which required many more GMRES iterations per inexact Newton step than the smaller problems.

4. Summary and conclusions. We have considered Newton–Krylov methods for solving (1.1) in which the matrix-vector products required by the Krylov solver are approximated by finite differences of the form (2.1) involving an approximating function $\tilde{F} \approx F$. The particular Krylov subspace methods considered are the GMRES and Arnoldi methods, which are based on the Arnoldi process (Algorithm 2.1); however, the developments here should also be helpful when other Krylov subspace methods are of interest.

TABLE 6
Case 4: $D(u) = K\sqrt{S(u)}[1 - (1 - S(u)^{\frac{1}{\mu}})^{\mu}]^2$.

NEQ	c	$\tilde{F}$ approx.	NNI	NLI	Run time	Norm. run time
40,000	1.0	None	4	107	98.5	1.00
		Linear	4	105	79.9	0.81
		Lagged	4	104	79.2	0.80
160,000	1.0	None	4	220	141.2	1.00
		Linear	4	220	102.3	0.72
		Lagged	4	223	101.6	0.72
40,000	2.0	None	4	109	99.2	1.00
		Linear	4	104	79.9	0.81
		Lagged	5	135	84.8	0.85
160,000	2.0	None	4	215	139.0	1.00
		Linear	4	215	101.5	0.73
		Lagged	5	280	113.1	0.81
40,000	5.0	None	5	138	110.4	1.00
		Linear	5	139	87.1	0.79
		Lagged	6	163	90.3	0.82
160,000	5.0	None	5	278	163.6	1.00
		Linear	5	278	114.2	0.70
		Lagged	6	330	122.6	0.75
40,000	10.0	None	5	144	112.9	1.00
		Linear	5	147	88.8	0.79
		Lagged	8	232	103.6	0.92
160,000	10.0	None	5	281	162.7	1.00
		Linear	5	289	115.7	0.71
		Lagged	9	514	156.9	0.96

In Algorithm 2.2, we formulate an Arnoldi process that uses approximate finite differences of the form (2.1) on a Jacobian system (2.2), and we consider the GMRES and Arnoldi methods applied to (2.2) that are based on this process. In Theorem 2.3 and Corollary 2.4, it is shown that an approximate solution of (2.2) determined at some step of the GMRES or Arnoldi method based on Algorithm 2.2 is the same as the approximate solution of the perturbed system (2.6) obtained at the same step of the method based on Algorithm 2.1.

In Algorithm 2.5, we outline a Newton–Krylov method that uses the GMRES or Arnoldi method based on Algorithm 2.2. Our main result for Algorithm 2.5 is Theorem 2.12, which can be viewed as a counterpart of results in [5, Thms. 2.3 and 3.3]. This theorem asserts that, under Assumption 2.7 on the approximating function $\tilde{F}$ as well as mild assumptions on F, the iterates produced by Algorithm 2.5 converge to a solution locally and q-linearly in a certain norm, provided the forcing terms are uniformly bounded below one and the difference steps are uniformly sufficiently small. If the forcing terms and difference steps approach zero, then the convergence is q-superlinear. If Assumption 2.10 on $\tilde{F}$ also holds, then the convergence is of q-order $1 + p$, where $p \in (0, 1]$ is the exponent of Hölder continuity in Assumption 2.10.

In section 3, we report on numerical experiments with two illustrative choices of $\tilde{F}$ that are suitable for a broad class of nonlinear diffusion problems. One of these choices, the "linear" approximation $\tilde{F}^{\mathrm{Lin}}$ satisfies the assumptions of Theorem 2.12 under mild conditions and, therefore, has a sound theoretical basis. In our tests, this choice yielded run times that were always less than those obtained by using exact values

of F, with greater run-time reductions observed for problems with more expensive F-evaluations. Additionally, in all test cases, this choice resulted in numbers of linear and nonlinear iterations that were very close to those obtained by using exact values of F. Thus this choice promises to be both robust and helpful in improving efficiency in many nonlinear diffusion problems, especially those in which F-evaluations are expensive. However, its implementation may be problematic in some applications (see section 3.2). Our other choice, the "lagged" approximation $\tilde{F}^{\mathrm{Lag}}$, does not normally satisfy the assumptions of Theorem 2.12; however, it is relatively easy and inexpensive to apply. It improved efficiency in some of our tests, especially in Case 2 (see Table 4); however, it was inefficient in other tests and suffered failures in Case 3 (see Table 5). Thus this choice should be kept in mind for ease of application and potential for improving efficiency but should not be counted on for improvement in all problems.

REFERENCES

[1] W. E. Arnoldi, *The principle of minimized iterations in the solution of the matrix eigenvalue problem*, Quart. Appl. Math., 9 (1951), pp. 17–29.

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

[3] P. N. Brown, *A local convergence theory for combined inexact-Newton/finite-difference projection methods*, SIAM J. Numer. Anal., 24 (1987), pp. 407–434.

[4] E. Catinas, *Inexact perturbed Newton methods and applications to a class of Krylov solvers*, J. Optim. Theory Appl., 108 (2001), pp. 543–571.

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

[6] J. E. Dennis, Jr., and R. B. Schnabel, *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*, Series in Automatic Computation, Prentice-Hall, Englewood Cliffs, NJ, 1983.

[7] S. C. Eisenstat and H. F. Walker, *Choosing the forcing terms in an inexact Newton method*, SIAM J. Sci. Comput., 17 (1996), pp. 16–32.

[8] R. W. Freund, G. H. Golub, and N. M. Nachtigal, *Iterative solution of linear systems*, Acta Numer. 1992, 1 (1992), pp. 57–100.

[9] G. H. Golub and C. Van Loan, *Matrix Computations*, 3rd ed., Johns Hopkins University Press, Baltimore, 1996.

[10] A. C. Hindmarsh, P. N. Brown, K. E. Grant, S. L. Lee, R. Serban, D. E. Shumaker, and C. S. Woodward, *SUNDIALS: Suite of nonlinear and differential/algebraic equation solvers*, ACM Trans. Math. Software, 31 (2005), pp. 363–396.

[11] C. T. Kelley, *Solving Nonlinear Equations with Newton's Method*, Fundam. Algorithms 1, SIAM, Philadelphia, PA, 2003.

[12] D. A. Knoll and D. E. Keyes, *Jacobian-free Newton–Krylov methods: A survey of approaches and applications*, J. Comput. Phys., 193 (2004), pp. 357–397.

[13] C. T. Miller, G. A. Williams, C. T. Kelley, and M. D. Tocci, *Robust solution of Richards' equation for non-uniform porous media*, Water Resources Research, 34 (1998), pp. 2599–2610.

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

[15] R. P. Pawlowski, J. N. Shadid, J. P. Simonis, and H. F. Walker, *Globalization techniques for Newton–Krylov methods and applications to the fully coupled solution of the Navier–Stokes equations*, SIAM Rev., 48 (2006), pp. 700–721.

[16] Y. Saad and M. H. Schultz, GMRES: *A generalized minimal residual algorithm for solving nonsymmetric linear systems*, SIAM J. Sci. Stat. Comput., 7 (1986), pp. 856–869.

[17] Y. Saad, *Krylov subspace methods for solving large unsymmetric linear systems*, Math. Comp., 37 (1981), pp. 105–126.

[18] Y. Saad, *Iterative Methods for Sparse Linear Systems*, PWS Publishing, Boston, 1996.

[19] H. A. van der Vorst, *Iterative Krylov Methods for Large Linear Systems*, Cambridge University Press, Cambridge, 2003.