SIAM J. NUMER. ANAL.
Vol. 27, No. 5, p. 1263–1294 , October 1990

008

LEAST-CHANGE SECANT UPDATES OF NONSQUARE MATRICES*

SAMIH K. BOURJI† AND HOMER F. WALKER‡

Abstract. The notion of least-change secant updates is extended to apply to nonsquare matrices in a way appropriate for quasi-Newton methods used to solve systems of nonlinear equations that depend on parameters. Extensions of the widely used least-change secant updates for square matrices are given. A local convergence analysis for certain paradigm iterations is outlined as motivation for the use of these updates, and numerical experiments involving these iterations are discussed.

Key words. least-change secant updates, quasi-Newton updates, parameter-dependent systems

AMS(MOS) subject classification. 65H10

1. Introduction.

Quasi-Newton methods are very widely used iterative methods for solving systems of nonlinear algebraic equations. The basic form of a quasi-Newton method for solving $F(x) = 0$, $F : \mathbf{R}^n \to \mathbf{R}^n$, is

$$x_{k+1} = x_k - B_k^{-1} F(x_k), \tag{1.1}$$

in which $B_k \approx F'(x_k) \in \mathbf{R}^{n \times n}$, the Jacobian (matrix) of F at x_k. For practical success, it is usually essential to augment this basic form with procedures for modifying the step $-B_k^{-1}F(x_k)$ to ensure progress from bad starting points, but we need not consider such procedures here. For a general reference on all aspects of quasi-Newton methods, see Dennis and Schnabel [11].

The most effective quasi-Newton methods are those in which each successive B_{k+1} is determined as a *least-change secant update* of its predecessor B_k. As the name suggests, B_{k+1} is determined as a least-change secant update of B_k by making the least possible change in B_k (as measured by a suitable matrix norm) which incorporates current secant information (usually expressed in terms of successive x- and F-values) and other available information about the structure of F'. There are also notable updates which, strictly speaking, are least-change *inverse* secant updates obtained in an analogous way by making the least possible change in B_k^{-1}. When speaking generically of least-change secant updates, we intend to include these. In [10], Dennis and Schnabel precisely formalize the notion of a least-change secant update and show how the most widely used updates can be derived as least-change secant updates. In [12], Dennis and Walker show that least-change secant update methods, i.e., quasi-Newton methods which use least-change secant updates, have desirable convergence properties in general.

---

*Received by the editors August 10, 1988; accepted for publication (in revised form) October 5, 1989.

†Department of Mathematics, Weber State College, Ogden, Utah 84408. The work of this author was supported in part by the United States Department of Energy under contract DE-FG02-86ER25018 with Utah State University. Some of the results in this paper first appeared in this author's doctoral dissertation at Utah State University.

‡Department of Mathematics and Statistics, Utah State University, Logan, Utah 84322-3900. The work of this author was supported by United States Department of Energy grant DE-FG02-86ER25018, Department of Defense/Army grant DAAL03-88-K, and National Science Foundation grant DMS-0088995, all with Utah State University.

In this paper we outline general principles and specific formulas which extend least-change secant updating from the traditional square-matrix context to that of nonsquare matrices. Our primary purpose is to provide updates which can be used in quasi-Newton methods for solving nonlinear systems which depend on parameters. We mention two important areas in which such systems occur and which have strongly influenced the work presented here. The first is the solution of a nonlinear system by continuation or homotopy methods, in which the parameter is a continuation parameter used to transform a problem from one which is presumably easy to solve into one for which the solution is desired but which may be difficult to solve ab initio. The second is the solution of ordinary differential equations by implicit methods, in which the parameter is the time variable and a nonlinear system must be solved by a sequence of corrector iterations at each timestep. In this second area, iterative methods used in the corrector iterations typically must carry over Jacobian information from one timestep to the next for the sake of efficiency. As a result, the nonlinear systems arising in this context are appropriately regarded as parametrized by the time variable, rather than as sequences of nonparametric systems, even though time is constant during each set of corrector iterations.

We consider a parameter-dependent nonlinear system in the form

$$F(x,\lambda)=0, \qquad F:\mathbf{R}^n\times\mathbf{R}^m\to\mathbf{R}^n, \tag{1.2}$$

where $x\in\mathbf{R}^n$ is an independent variable and $\lambda\in\mathbf{R}^m$ is a parameter vector. Our interest is in iterative methods that are intended to produce approximate solutions of (1.2) through some range of λ-values. There are many iterative methods that might be appropriate for this, depending in part on how the λ-values are specified (a priori or as the iterations proceed, independent of or in conjunction with the x-values, etc.). We are not concerned here with how successive λ-values might be determined or with the specific forms of the iterations, but we have in mind methods which require approximations at current (x,λ)-values of $F_x\in\mathbf{R}^{n\times n}$, the Jacobian of F with respect to x, or perhaps of the full Jacobian $F'=[F_x,F_\lambda]\in\mathbf{R}^{n\times(n+m)}$, where $F_\lambda\in\mathbf{R}^{n\times m}$ is the Jacobian of F with respect to λ, and which for efficiency require Jacobian information to be retained through at least some changes in λ.

In view of the success of least-change secant updates in maintaining approximate Jacobians in a quasi-Newton iteration (1.1), it is natural to consider their use in maintaining approximations of F_x or $F'=[F_x,F_\lambda]$ in iterations used to solve (1.2). In order to do so, some sort of extension of the usual formulation of these updates is clearly required. Indeed, the usual formulation applies only to square matrices, and F' is nonsquare. Even if we require only an approximation of the square matrix F_x, it is not clear how to incorporate the information in successive F-values in a secant update if λ changes as well as x.

To be more specific about this last point, we suppose that (x,λ) and (x_+,λ_+) are "current" and "next" values and that B is a "current" approximation of F_x. Then the usual formulation for determining a "next" approximate Jacobian B_+ as a least-change secant update of B calls for B_+ to satisfy a *secant equation*

$$B_+(x_+-x)=F(x_+,\lambda_+)-F(x,\lambda) \tag{1.3}$$

as nearly as possible consistent with any structure that may be imposed on B_+. Since

$$F(x_+,\lambda_+)-F(x,\lambda)\approx[F_x(x_+,\lambda_+),F_\lambda(x_+,\lambda_+)]\begin{pmatrix}x_+-x\\ \lambda_+-\lambda\end{pmatrix},$$

this secant equation is clearly inappropriate if $\lambda_+ \neq \lambda$. One modification of the usual formulation which may be useful in some circumstances is to replace the right-hand side of (1.3) with something more appropriate. An ideal replacement would be $F(x_+, \lambda_+) - F(x, \lambda) - F_\lambda(x, \lambda)(\lambda_+ - \lambda)$ or $F(x_+, \lambda_+) - F(x, \lambda) - F_\lambda(x_+, \lambda_+)(\lambda_+ - \lambda)$, provided F_λ is easy to obtain. If F_λ is not easy to obtain, then a finite-difference approximation might be an effective substitute. However, any strategy such as this will necessarily require extra derivative or function evaluations and so may be undesirably expensive in some applications.

In the next section we give extensions to the nonsquare-matrix case of the usual general formulations of square-matrix least-change secant and inverse secant updates. With these, the secant information in successive F-values can be used to determine updates of approximations of $F' = [F_x, F_\lambda]$ according to least-change criteria. These extensions incidentally determine updates of approximations of F_x using available secant information, i.e., without any extra derivative or function evaluations. Having these general extensions, we derive in §3 extensions of the most widely used specific formulas for square-matrix least-change secant updates having various properties. In order to provide some theoretical support for using these updates, we outline in §4 a local convergence analysis for certain paradigm methods which employ them. In §5, we report on some numerical experiments.

The focus here is on least-change secant updates of matrices with more columns than rows; however, our ways of deriving updates using least-change principles readily apply as well to matrices with more rows than columns, such as occur in nonlinear least-squares problems. In this connection, we note that Griewank and Sheng [19] have recently considered Broyden updating for such matrices in the context of nonlinear least-squares problems.

Although the general updating principles, specific update formulas, and local convergence analysis given here are for the most part new, there is other work which is closely related to the developments here. The updates used in the path-following algorithm of Georg [16] and in the augmented Jacobian matrix algorithm in the homotopy method code HOMPACK [32] are really the same as our extension of the first Broyden update (3.1.1) below, although these updates are viewed somewhat differently in [16] and [32]. In work which is complementary to that in this paper, Walker and Watson [31] consider general normal flow and augmented Jacobian algorithms for underdetermined systems which use the general and specific updates given here and give a local q-linear and q-superlinear convergence analysis for these algorithms. (See, e.g., Watson, Billups, and Morgan [32] as well as [31] for a description of the normal flow and augmented Jacobian algorithms. Also, see, e.g., Ortega and Rheinboldt [25] for definitions of the various types of convergence referred to here.) In recent work independent of that here and in [31], Martínez [20] considers Newton-like iterative methods for underdetermined systems which use very general procedures for updating approximate Jacobians, and he develops a general local r-linear and r-superlinear convergence analysis for these methods. He points out as a special case the possibility of maintaining approximate Jacobians in normal flow algorithms with updates which are, in our terms, the Frobenius-norm least-change secant updates developed in §2 and §3.1 below. No specific update formulas are given in [20], although experiments with (sparse) first Broyden updating are discussed; see the additional remarks in §5 below. Finally, we note that in many respects there are strong parallels between the developments here and those in [12]; this is especially so in our general formulations of least-change secant and inverse secant updates in §2 and in the local convergence analysis in §4 and in the Appendix.

Our notational conventions are not strict, but it might be helpful to the reader to note that we use the following guidelines: Unless otherwise indicated, lowercase letters denote vectors and scalars, and capital letters denote matrices and operators. Boldface uppercase letters denote vector spaces, subspaces, and affine subspaces. For convenience, we set $\bar{n} = n + m$. Vectors and matrices with bars are in $\mathbf{R}^{\bar{n}}$ and $\mathbf{R}^{n\times\bar{n}}$, respectively. Without bars, vectors are in $\mathbf{R}^n$ or $\mathbf{R}^m$ and matrices are in $\mathbf{R}^{n\times n}$ or $\mathbf{R}^{n\times m}$, unless otherwise indicated. If $x \in \mathbf{R}^n$ and $\lambda \in \mathbf{R}^m$, then for $\bar{x} = \binom{x}{\lambda} \in \mathbf{R}^{\bar{n}}$ we usually write $\bar{x} = (x, \lambda)$, $F(x, \lambda) = F(\bar{x})$, etc. Vector components and matrix entries are indicated by superscripts in parentheses. Distinguished vectors and matrices are indicated by subscripts, and subscripts on vectors and matrices are inherited by their components and entries. For example, the kth matrix in a sequence in $\mathbf{R}^{n\times n}$ might be denoted by B_k, in which case its ijth entry would be denoted by $B_k^{(ij)}$. We use $|\cdot|$ to denote all vector norms and their induced matrix norms, and we use $\|\cdot\|$ to denote a matrix norm associated with an inner product. A projection onto a subspace or affine subspace which is orthogonal with respect to $\|\cdot\|$ is denoted by P with the subspace or affine subspace appearing as a subscript. If P denotes a projection, then we set $P^\perp = I - P$, where I is the identity operator.

2. General least-change secant updates. To define general least-change secant updates of nonsquare matrices, we suppose that analogues of the usual ingredients in the square-matrix case are given, viz., the following:

(i) $\bar{B} \in \mathbf{R}^{n\times\bar{n}}$;

(ii) $\bar{s} \in \mathbf{R}^{\bar{n}}$ and $y \in \mathbf{R}^n$;

(iii) an inner-product norm $\|\cdot\|$ on $\mathbf{R}^{n\times\bar{n}}$;

(iv) an affine subspace $\mathbf{A} = \bar{A}_N + \mathbf{S} \subseteq \mathbf{R}^{n\times\bar{n}}$, where $\mathbf{S}$ is the parallel subspace and $\bar{A}_N$ is the element of $\mathbf{A}$ normal to $\mathbf{S}$ in the $\|\cdot\|$-inner product.

It is appropriate (but not necessary) to relate these to the equation-solving context outlined in the introduction by regarding $\bar{B} \approx F'(\bar{x})$ for some $\bar{x}$, $\bar{s} = \bar{x}_+ - \bar{x}$ for some $\bar{x}_+$, and $y \approx F'(\bar{x})\bar{s}$, e.g., $y = F(\bar{x}_+) - F(\bar{x})$. The affine subspace $\mathbf{A}$ presumably reflects some structure of F', e.g., a particular pattern of sparsity, which is to be imposed on updates. An appropriate choice of norm $\|\cdot\|$ depends on the problem at hand but seems likely to be the Frobenius norm or a weighted Frobenius norm, as in the square-matrix case.

We define a general least-change secant update via the approach and notation of [12, §2]. We set $\mathbf{N}(\bar{s}) = \{\bar{M} \in \mathbf{R}^{n\times\bar{n}} : \bar{M}\bar{s} = 0\}$ and $\mathbf{Q}(y, \bar{s}) = \{\bar{M} \in \mathbf{R}^{n\times\bar{n}} : \bar{M}\bar{s} = y\}$ and note that $\mathbf{N}(\bar{s})$ is a subspace of $\mathbf{R}^{n\times\bar{n}}$ and $\mathbf{Q}(y, \bar{s})$ is an affine subspace representable as $\mathbf{Q}(y, \bar{s}) = y\bar{s}^T/\bar{s}^T\bar{s} + \mathbf{N}(\bar{s})$. We define $\mathbf{M}(\mathbf{A}, \mathbf{Q}(y, \bar{s}))$ to be the set of elements of $\mathbf{A}$ which are closest to $\mathbf{Q}(y, \bar{s})$ in the norm $\|\cdot\|$ if $\mathbf{Q}(y, \bar{s}) \neq \emptyset$; otherwise, we set $\mathbf{M}(\mathbf{A}, \mathbf{Q}(y, \bar{s})) = \mathbf{A}$. Of course, $\mathbf{M}(\mathbf{A}, \mathbf{Q}(y, \bar{s})) = \mathbf{A} \cap \mathbf{Q}(y, \bar{s})$ if $\mathbf{A} \cap \mathbf{Q}(y, \bar{s}) \neq \emptyset$. As in Theorem 2.3 of [12], it can be shown that $\mathbf{M}(\mathbf{A}, \mathbf{Q}(y, \bar{s}))$ is an affine subspace with parallel subspace $\mathbf{S} \cap \mathbf{N}(\bar{s})$.

Our general definition of a least-change secant update is the following.

DEFINITION 2.1. $\bar{B}_+ \in \mathbf{R}^{n\times\bar{n}}$ is the *least-change secant update* of $\bar{B}$ in $\mathbf{A}$ with respect to $\bar{s}$, y, and the norm $\|\cdot\|$ if $\bar{B}_+$ is the unique solution of

$$\min_{\tilde{\bar{B}} \in \mathbf{M}(\mathbf{A}, \mathbf{Q}(y, \bar{s}))} \left\| \tilde{\bar{B}} - \bar{B} \right\|.$$

We note that $\bar{B}_+ = P_{\mathbf{M}(\mathbf{A}, \mathbf{Q}(y, \bar{s}))}\bar{B}$, the $\|\cdot\|$-orthogonal projection of $\bar{B}$ onto $\mathbf{M}(\mathbf{A}, \mathbf{Q}(y, \bar{s}))$. If $\mathbf{A} \cap \mathbf{Q}(y, \bar{s}) \neq \emptyset$, then $\bar{B}_+$ satisfies the secant equation

$$\bar{B}_+\bar{s} = y. \tag{2.1}$$

It can easily be verified that all of the results of [12, §2] for square-matrix least-change secant updates extend to the present case. Particularly relevant results are the following. We have

$$\begin{aligned}\bar{B}_+ &= \lim_{k\to\infty} \left(P_{\mathbf{A}} P_{\mathbf{Q}(y,\bar{s})}\right)^k P_{\mathbf{A}} \bar{B} \\ &= \left(I - P_{\mathbf{S}} P_{\mathbf{N}(\bar{s})}\right)^{-1} \bar{A}_N + \left(I - P_{\mathbf{S}} P_{\mathbf{N}(\bar{s})}\right)^{-1} P_{\mathbf{S}} P^{\perp}_{\mathbf{N}(\bar{s})} \left(\frac{y\bar{s}^T}{\bar{s}^T\bar{s}}\right) + P_{\mathbf{S}\cap\mathbf{N}(\bar{s})}\bar{B}.\end{aligned} \tag{2.2}$$

In (2.2), $\left(I - P_{\mathbf{S}} P_{\mathbf{N}(\bar{s})}\right)^{-1}$ is well defined as an operator on $(\mathbf{S}\cap\mathbf{N}(\bar{s}))^{\perp}$. In fact,

$$\left(I - P_{\mathbf{S}} P_{\mathbf{N}(\bar{s})}\right)^{-1} \bar{A}_N + \left(I - P_{\mathbf{S}} P_{\mathbf{N}(\bar{s})}\right)^{-1} P_{\mathbf{S}} P_{\mathbf{N}(\bar{s})}{}^{\perp} \left(\frac{y\bar{s}^T}{\bar{s}^T\bar{s}}\right) \in (\mathbf{S}\cap\mathbf{N}(\bar{s}))^{\perp},$$

and it follows that for any $\bar{G} \in \mathbf{M}(\mathbf{A}, \mathbf{Q}(y,\bar{s}))$,

$$\bar{B}_+ = P^{\perp}_{\mathbf{S}\cap\mathbf{N}(\bar{s})}\bar{G} + P_{\mathbf{S}\cap\mathbf{N}(\bar{s})}\bar{B} \tag{2.3}$$

and for any $\bar{M} \in \mathbf{R}^{n\times\bar{n}}$

$$\|\bar{B}_+ - \bar{M}\| \le \|P^{\perp}_{\mathbf{S}\cap\mathbf{N}(\bar{s})}(\bar{G} - \bar{M})\| + \|P_{\mathbf{S}\cap\mathbf{N}(\bar{s})}(\bar{B} - \bar{M})\|. \tag{2.4}$$

We also offer a general definition of a least-change inverse secant update of a nonsquare matrix. Our motivation is the importance of least-change inverse secant updates in the square-matrix case. For example, the Broyden–Fletcher–Goldfarb–Shanno update [6], [7], [13], [17], [29] is such an update (see [10]) and is generally regarded as the most successful update for unconstrained optimization problems. The second Broyden update [5] is another such update (see [10]); although it is not generally as effective as the first Broyden update [5], there is evidence that it may be competitive in stiff ODE applications (see [1], [4]). We hope that the updates prescribed here will be similarly effective, e.g., in optimization problems which depend on parameters or in stiff ODE problems in which the time variable is taken into account in updating.

The idea underlying our definition is to assume that $\bar{B}$ is of full rank n, then to select a set of n columns which constitute a nonsingular matrix, and finally to apply the least-change principle to the n columns of the inverse of this matrix together with another m columns derived from $\bar{B}$. There are several ways in which this idea might be carried out. We choose a way which is not only natural and straightforward but also supported by the local convergence analysis in §4.

For convenience, we assume that the first n columns of $\bar{B}$ constitute a nonsingular matrix, although we stress that any set of n linearly independent columns of $\bar{B}$ can be used instead. We write $\bar{B} = [B, C]$ for nonsingular $B \in \mathbf{R}^{n\times n}$ and for $C \in \mathbf{R}^{n\times m}$ and set $\bar{K} = [K, L]$, where $K = B^{-1}$ and $L = -B^{-1}C$. We define an update $\bar{K}_+$ of $\bar{K}$ with an eye toward an *inverse secant equation*

$$\bar{K}_+\bar{y} = s, \tag{2.5}$$

where we write $\bar{s} = (s,t)$ for $s \in \mathbf{R}^n$ and $t \in \mathbf{R}^m$ and set $\bar{y} = (y,t)$. Our motivation is the following: If we write $\bar{K}_+ = [K_+, L_+]$ and $\bar{B}_+ = [B_+, C_+]$ with $K_+ = {B_+}^{-1}$ and $L_+ = -{B_+}^{-1}C_+$, then $\bar{B}_+$ satisfies the secant equation $\bar{B}_+\bar{s} = y$ if and only if $s = B_+^{-1}y - B_+^{-1}C_+t = \bar{K}_+\bar{y}$. We phrase our definition in terms of $\bar{K}$ and $\bar{K}_+$ as follows.

DEFINITION 2.2. $\bar{K}_+ \in \mathbf{R}^{n\times\bar{n}}$ is the *least-change inverse secant update* of $\bar{K}$ in $\mathbf{A}$ with respect to $\bar{s}$, y, and the norm $\|\cdot\|$ if $\bar{K}_+$ is the unique solution of

$$\min_{\tilde{\bar{K}}\in\mathbf{M}(\mathbf{A},\mathbf{Q}(s,\bar{y}))} \left\|\tilde{\bar{K}} - \bar{K}\right\|,$$

where $\bar{s} = (s,t)$ for $s \in \mathbf{R}^n$ and $t \in \mathbf{R}^m$ and $\bar{y} = (y,t)$.

To provide additional motivation for this definition, we note that Ypma [34] and one of the referees have observed that

$$\begin{bmatrix} B & C \\ 0 & I \end{bmatrix}^{-1} = \begin{bmatrix} B^{-1} & -B^{-1}C \\ 0 & I \end{bmatrix} = \begin{bmatrix} K & L \\ 0 & I \end{bmatrix},$$

and from this it is easy to see that $\bar{K}_+$ of Definition 2.2 can be obtained through a conventional least-change inverse secant update of a square matrix with an appropriate choice of an inner-product norm and affine subspace. We also note that from $\bar{K}_+ = [K_+, L_+]$, we can obtain $\bar{B}_+ = [B_+, C_+]$ by taking $B_+ = K_+^{-1}$ and $C_+ = -K_+{}^{-1}L_+$, although it may be preferred in some applications to maintain $\bar{K}_+$ rather than $\bar{B}_+$. If $\mathbf{A} \cap \mathbf{Q}(s,\bar{y}) \neq \emptyset$, then $\bar{K}_+$ satisfies (2.5) and $\bar{B}_+$ satisfies (2.1). We have in any case that $\bar{K}_+ = P_{\mathbf{M}(\mathbf{A},\mathbf{Q}(s,\bar{y}))}\bar{K}$. Also, inverse analogues of (2.2), (2.3), and (2.4) hold with $\bar{B}_+$, $\bar{B}$, $\bar{s}$, and y in those equations replaced by $\bar{K}_+$, $\bar{K}$, $\bar{y}$, and s, respectively.

3. Some specific updates. We now derive extensions to the nonsquare case of the best-known square-matrix least-change secant updates. These extensions are obtained from the definitions in §2 by making various choices of $\mathbf{A}$ and $\|\cdot\|$ which correspond in each case to the analogous choice made for square matrices. Not surprisingly, many of these updates look very much like their square-matrix counterparts. Our intention is not only to determine updates which are likely to be important but also to demonstrate methods of derivation which can be used to obtain other desirable updates.

We first derive least-change secant updates from Definition 2.1 and then derive least-change inverse secant updates from Definition 2.2. In each case, we refer to the update by the corresponding well-known name in the square-matrix case. For simplicity, we also assume that the matrix to be updated lies in $\mathbf{A}$ in each case. This is typically to be expected in practice. Also, if the matrix were not in $\mathbf{A}$, then we could obtain its least-change secant update in $\mathbf{A}$ by first projecting it onto $\mathbf{A}$ in a straightforward way and then applying the appropriate formula below. For both the symmetry-preserving updates and the inverse updates, it is necessary to assume that a subset of n of the columns of the matrix being updated has some special property. Both for convenience and because it seems most likely to be the case in practice, we assume that the special property resides in the first n columns. We stress that any other subset of n columns will do just as well and that even though an assumption is made about a particular subset of the columns, the updates are derived by applying the least-change principle to the entire matrix.

3.1. Least-change secant updates. We derive below extensions of the first Broyden update, the Powell symmetric Broyden (PSB) update [26], the sparse Broyden (Schubert) update [8], [27], and the Davidon–Fletcher–Powell (DFP) update [9], [14]. At this time we do not have a tractable derivation of an extension of the sparse symmetric update of Marwil [21] and Toint [30], although such an update has been derived by Beattie and Weaver-Smith [2] using other methods.

We suppose that we are given $\bar{B} \in \mathbf{R}^{n\times\bar{n}}$, nonzero $\bar{s} \in \mathbf{R}^{\bar{n}}$, and $y \in \mathbf{R}^n$. When necessary in the following, we write $\bar{B} = [B, C]$ for $B \in \mathbf{R}^{n\times n}$ and $C \in \mathbf{R}^{n\times m}$ and $\bar{s} = (s, t)$ for $s \in \mathbf{R}^n$ and $t \in \mathbf{R}^m$. We note that each update below gives the usual square-matrix update of B when $t = 0$.

The first Broyden update. We take $\mathbf{A} = \mathbf{S} = \mathbf{R}^{n\times\bar{n}}$ and $\|\cdot\| = \|\cdot\|_F$, the Frobenius norm. We have $\mathbf{M}(\mathbf{A}, \mathbf{Q}(y, \bar{s})) = \mathbf{Q}(y, \bar{s})$, and so $\bar{B}_+ = P_{\mathbf{M}(\mathbf{A},\mathbf{Q}(y,\bar{s}))}\bar{B} = P_{\mathbf{Q}(y,\bar{s})}\bar{B}$, where the projection is $\|\cdot\|_F$-orthogonal. Orthogonal projection onto an affine subspace is obtained by adding the normal element of the affine subspace to orthogonal projection onto the parallel subspace. It is easily verified that the normal element of $\mathbf{Q}(y, \bar{s})$ is $y\bar{s}^T/\bar{s}^T\bar{s}$ and that orthogonal projection onto the parallel subspace $\mathbf{N}(\bar{s})$ is given by $P_{\mathbf{N}(\bar{s})}\bar{M} = \bar{M}[I - \bar{s}\bar{s}^T/\bar{s}^T\bar{s}]$ for $\bar{M} \in \mathbf{R}^{n\times\bar{n}}$. It follows that

$$\begin{aligned}\bar{B}_+ &= \bar{B}\left[I - \frac{\bar{s}\bar{s}^T}{\bar{s}^T\bar{s}}\right] + \frac{y\bar{s}^T}{\bar{s}^T\bar{s}} \\ &= \bar{B} + \frac{(y - \bar{B}\bar{s})\bar{s}^T}{\bar{s}^T\bar{s}}.\end{aligned} \tag{3.1.1}$$

The Powell symmetric Broyden update. We now take $\mathbf{A} = \mathbf{S} = \{\bar{M} = [M, N] \in \mathbf{R}^{n\times\bar{n}} : M = M^T \in \mathbf{R}^{n\times n}\}$ and again take $\|\cdot\| = \|\cdot\|_F$. Instead of characterizing $P_{\mathbf{M}(\mathbf{A},\mathbf{Q}(y,\bar{s}))}$ as above, we work with the expression (2.2). With $\bar{A}_N = 0$, (2.2) is

$$\bar{B}_+ = \left(I - P_{\mathbf{S}}P_{\mathbf{N}(\bar{s})}\right)^{-1} P_{\mathbf{S}}P_{\mathbf{N}(\bar{s})}^{\perp}\left(\frac{y\bar{s}^T}{\bar{s}^T\bar{s}}\right) + P_{\mathbf{S}\cap\mathbf{N}(\bar{s})}\bar{B}. \tag{3.1.2}$$

For $\bar{M} = [M, N] \in \mathbf{R}^{n\times\bar{n}}$ with $M \in \mathbf{R}^{n\times n}$ and $N \in \mathbf{R}^{n\times m}$, we easily have $P_{\mathbf{S}}\bar{M} = [(M + M^T)/2, N]$, and $P_{\mathbf{N}(\bar{s})}\bar{M} = \bar{M}[I - \bar{s}\bar{s}^T/\bar{s}^T\bar{s}]$ as before. Assuming $\bar{B} \in \mathbf{A} = \mathbf{S}$, we use these projections to evaluate the right-hand side of (3.1.2).

First, we have $P_{\mathbf{S}\cap\mathbf{N}(\bar{s})}\bar{B} = \lim_{k\to\infty}(P_{\mathbf{S}}P_{\mathbf{N}(\bar{s})})^k\bar{B}$ (cf. [12, §2]). Straightforward calculation gives

$$P_{\mathbf{S}}P_{\mathbf{N}(\bar{s})}\bar{B} = \bar{B} - \frac{1}{2}\bar{B}_1 \quad \text{where} \quad \bar{B}_1 = \frac{\bar{B}\bar{s}\bar{s}^T}{\bar{s}^T\bar{s}} + \left[\frac{s(\bar{B}\bar{s})^T}{\bar{s}^T\bar{s}}, \frac{\bar{B}\bar{s}t^T}{\bar{s}^T\bar{s}}\right],$$

$$P_{\mathbf{S}}P_{\mathbf{N}(\bar{s})}\bar{B}_1 = \frac{1}{2}\frac{s^Ts}{\bar{s}^T\bar{s}}\bar{B}_1 - \bar{B}_2 \quad \text{where} \quad \bar{B}_2 = \frac{s^T(\bar{B}\bar{s})}{\bar{s}^T\bar{s}}\frac{\bar{s}\bar{s}^T}{\bar{s}^T\bar{s}}.$$

Note that $P_{\mathbf{N}(\bar{s})}\bar{B}_2 = 0$, so $P_{\mathbf{S}}P_{\mathbf{N}(\bar{s})}\bar{B}_2 = 0$. It can be shown by induction that for $k = 2, 3, \cdots$,

$$(P_{\mathbf{S}}P_{\mathbf{N}(\bar{s})})^k\bar{B} = \bar{B} - \frac{1}{2}\left\{\sum_{j=0}^{k-1}\left(\frac{s^Ts}{2\bar{s}^T\bar{s}}\right)^j\right\}\bar{B}_1 + \frac{1}{2}\left\{\sum_{j=0}^{k-2}\left(\frac{s^Ts}{2\bar{s}^T\bar{s}}\right)^j\right\}\bar{B}_2,$$

and it follows that

$$\begin{aligned}P_{\mathbf{S}\cap\mathbf{N}(\bar{s})}\bar{B} &= \bar{B} - \frac{\bar{s}^T\bar{s}}{\bar{s}^T\bar{s} + t^Tt}(\bar{B}_1 - \bar{B}_2) \\ &= \bar{B} - \frac{\bar{B}\bar{s}\bar{s}^T + \left[s(\bar{B}\bar{s})^T, \bar{B}\bar{s}t^T\right]}{\bar{s}^T\bar{s} + t^Tt} + \frac{s^T\bar{B}\bar{s}}{\bar{s}^T\bar{s} + t^Tt}\frac{\bar{s}\bar{s}^T}{\bar{s}^T\bar{s}}.\end{aligned} \tag{3.1.3}$$

To evaluate the other term in (3.1.2), we note that $(I - P_{\mathbf{S}}P_{\mathbf{N}(\bar{s})})^{-1} = \sum_{k=0}^{\infty}(P_{\mathbf{S}}P_{\mathbf{N}(\bar{s})})^k$ on $(\mathbf{S} \cap \mathbf{N}(\bar{s}))^{\perp}$ and that

$$P_{\mathbf{S}}P_{\mathbf{N}(\bar{s})}^{\perp}\left(\frac{y\bar{s}^T}{\bar{s}^T\bar{s}}\right) = \frac{1}{2}\left\{\frac{y\bar{s}^T}{\bar{s}^T\bar{s}} + \left[\frac{sy^T}{\bar{s}^T\bar{s}}, \frac{yt^T}{\bar{s}^T\bar{s}}\right]\right\} \tag{3.1.4}$$

and

$$P_{\mathbf{S}}P_{\mathbf{N}(\bar{s})}P_{\mathbf{S}}P_{\mathbf{N}(\bar{s})}^{\perp}\left(\frac{y\bar{s}^T}{\bar{s}^T\bar{s}}\right) = \frac{s^Ts}{2\bar{s}^T\bar{s}}P_{\mathbf{S}}P_{\mathbf{N}(\bar{s})}^{\perp}\left(\frac{y\bar{s}^T}{\bar{s}^T\bar{s}}\right) - \frac{s^Ty}{2\bar{s}^T\bar{s}}\frac{s\bar{s}^T}{\bar{s}^T\bar{s}}.$$

It follows by induction that for $k = 1, 2, \cdots$,

$$\begin{aligned}\left(P_{\mathbf{S}}P_{\mathbf{N}(\bar{s})}\right)^k P_{\mathbf{S}}P_{\mathbf{N}(\bar{s})}^{\perp}\left(\frac{y\bar{s}^T}{\bar{s}^T\bar{s}}\right) = \left(\frac{s^Ts}{2\bar{s}^T\bar{s}}\right)^k P_{\mathbf{S}}P_{\mathbf{N}(\bar{s})}^{\perp}\left(\frac{y\bar{s}^T}{\bar{s}^T\bar{s}}\right) \\ - \left(\frac{s^Ts}{2\bar{s}^T\bar{s}}\right)^{k-1}\frac{s^Ty}{2\bar{s}^T\bar{s}}\frac{s\bar{s}^T}{\bar{s}^T\bar{s}},\end{aligned} \tag{3.1.5}$$

and (3.1.4) and (3.1.5) give

$$\left(I - P_{\mathbf{S}}P_{\mathbf{N}(\bar{s})}\right)^{-1}P_{\mathbf{S}}P_{\mathbf{N}(\bar{s})}^{\perp}\left(\frac{y\bar{s}^T}{\bar{s}^T\bar{s}}\right) = \frac{y\bar{s}^T + [sy^T, yt^T]}{\bar{s}^T\bar{s} + t^Tt} - \frac{s^Ty}{\bar{s}^T\bar{s} + t^Tt}\frac{s\bar{s}^T}{\bar{s}^T\bar{s}}. \tag{3.1.6}$$

Combining (3.1.2), (3.1.3), and (3.1.6) gives our extension of the PSB update:

$$\begin{aligned}\bar{B}_+ = \bar{B} + \frac{(y - \bar{B}\bar{s})\bar{s}^T + [s(y - \bar{B}\bar{s})^T, (y - \bar{B}\bar{s})t^T]}{\bar{s}^T\bar{s} + t^Tt} \\ - \frac{s^T(y - \bar{B}\bar{s})}{\bar{s}^T\bar{s} + t^Tt}\frac{s\bar{s}^T}{\bar{s}^T\bar{s}}.\end{aligned} \tag{3.1.7}$$

The structure of this update may be better revealed if we write $\bar{B}_+ = [B_+, C_+]$ for $B_+ \in \mathbf{R}^{n\times n}$ and $C_+ \in \mathbf{R}^{n\times m}$ and note that (3.1.7) gives

$$\begin{aligned}B_+ &= B + \frac{(y - \bar{B}\bar{s})s^T + s(y - \bar{B}\bar{s})^T}{\bar{s}^T\bar{s} + t^Tt} - \frac{s^T(y - \bar{B}\bar{s})}{\bar{s}^T\bar{s} + t^Tt}\frac{ss^T}{\bar{s}^T\bar{s}}, \\ C_+ &= C + \frac{2(y - \bar{B}\bar{s})t^T}{\bar{s}^T\bar{s} + t^Tt} - \frac{s^T(y - \bar{B}\bar{s})}{\bar{s}^T\bar{s} + t^Tt}\frac{st^T}{\bar{s}^T\bar{s}}.\end{aligned} \tag{3.1.8}$$

The sparse Broyden update. We now take $\mathbf{A}$ and $\mathbf{S}$ to be the set of matrices in $\mathbf{R}^{n\times\bar{n}}$ having a particular pattern of sparsity, and we again take $\|\cdot\| = \|\cdot\|_F$. As in the case of the first Broyden update, we characterize $P_{\mathbf{M}(\mathbf{A},\mathbf{Q}(y,\bar{s}))}$.

For $i = 1, \cdots, n$, we let $\bar{s}_i$ be the vector obtained by imposing on $\bar{s}$ the sparsity pattern of the ith row of matrices in $\mathbf{A}$. We note that if $\bar{M} \in \mathbf{A}$, then $\bar{M} \in \mathbf{M}(\mathbf{A}, \mathbf{Q}(y, \bar{s}))$ if and only if for $i = 1, \cdots, n$, ${\epsilon_i}^T\bar{M}\bar{s} = {\epsilon_i}^Ty$ whenever $\bar{s}_i \neq 0$, where ϵ_i is the ith standard unit basis vector in $\mathbf{R}^n$. Also, $\mathbf{M}(\mathbf{A}, \mathbf{Q}(y, \bar{s}))$ is an affine subspace with parallel subspace $\mathbf{A} \cap \mathbf{N}(\bar{s}) = \mathbf{S} \cap \mathbf{N}(\bar{s})$. Then we have for $\bar{M} \in \mathbf{A}$ that

$$P_{\mathbf{S}\cap\mathbf{N}(\bar{s})}\bar{M} = \sum_{i=1}^{n}\epsilon_i{\epsilon_i}^T\bar{M}\left[I - \left(\bar{s}_i^T\bar{s}_i\right)^+\bar{s}\bar{s}_i^T\right]$$

and that the normal to $\mathbf{M}(\mathbf{A}, \mathbf{Q}(y, \bar{s}))$ is

$$\sum_{i=1}^{n}\left(\bar{s}_i^T\bar{s}_i\right)^+\epsilon_i{\epsilon_i}^Ty\bar{s}_i^T,$$

where "+" denotes pseudo-inverse, i.e., for each i, $(\bar{s}_i^T \bar{s}_i)^+ = (\bar{s}_i^T \bar{s}_i)^{-1}$ if $\bar{s}_i \neq 0$ and $(\bar{s}_i^T \bar{s}_i)^+ = 0$ if $\bar{s}_i = 0$. It follows that if $\bar{B} \in \mathbf{A}$, then

$$\begin{aligned}\bar{B}_+ &= P_{\mathbf{M}(\mathbf{A},\mathbf{Q}(y,\bar{s}))}\bar{B} \\ &= \sum_{i=1}^{n} \epsilon_i \epsilon_i^T \left\{ \left(\bar{s}_i^T \bar{s}_i\right)^+ y \bar{s}_i^T + \bar{B}\left[I - \left(\bar{s}_i^T \bar{s}_i\right)^+ \bar{s}\bar{s}_i^T\right]\right\} \\ &= \bar{B} + \sum_{i=1}^{n} \left(\bar{s}_i^T \bar{s}_i\right)^+ \epsilon_i \epsilon_i^T \left(y - \bar{B}\bar{s}\right) \bar{s}_i^T.\end{aligned}$$

The Davidon–Fletcher–Powell update. We take

$$\mathbf{A} = \mathbf{S} = \left\{\bar{M} = [M, N] \in \mathbf{R}^{n \times \bar{n}} : M = M^T \in \mathbf{R}^{n \times n}\right\}$$

as in the case of the PSB update, and we suppose that $\bar{B} = [B, C] \in \mathbf{A}$. As in the square-matrix case (cf. [10], [12]), we determine an appropriate matrix norm for updating using a *scaling matrix* $W \in \mathbf{R}^{n \times n}$. We ask the reader to bear with our perhaps curious definitions and assumptions pending the justification given following them. We set $\hat{y} = y - Ct$ and assume $s^T \hat{y} > 0$. We let $W \in \mathbf{R}^{n \times n}$ be any positive-definite symmetric matrix such that $Ws = \hat{y}$ and define a norm $\|\cdot\|_W$ on $\mathbf{R}^{n \times \bar{n}}$ as follows: For $\bar{M} = [M, N] \in \mathbf{R}^{n \times \bar{n}}$, we set

$$\|\bar{M}\|_W^2 = \operatorname{tr}\left\{W^{-1} M W^{-1} M^T + W^{-1} N N^T\right\}. \tag{3.1.9}$$

To justify these definitions and assumptions, we note that in the equation-solving context outlined earlier, we presumably want an update of the type sought here in order to reflect positive-definiteness and symmetry of F_x. If $\bar{s} = \bar{x}_+ - \bar{x}$ for nearby $\bar{x}_+$ and $\bar{x}$ and if C is a good approximation of $F_\lambda(\bar{x})$, then $\hat{y} \approx F_x(\bar{x})s$, and so we can expect $s^T \hat{y} > 0$. The tradition is to think of the s, $\hat{y}$ pair as reflecting a natural scaling determined by F_x and approximated by W. In principle, we might define $\|\cdot\|_W$ in any of several ways which extend the definition used in the square-matrix case. We have chosen to define $\|\cdot\|_W$ by (3.1.9) because doing so yields a final update formula in which W does not appear. As in the square-matrix case, this is crucial; see the discussion in [12, pp. 971–972].

We determine $\bar{B}_+$ as a least-change secant update of $\bar{B}$ in $\mathbf{A}$ with respect to $\|\cdot\|_W$. The strategy for doing so is one sometimes used in the square-matrix case (see, e.g., [10]), viz., first to rescale $\bar{B}$ using W, then to update the rescaled matrix using (3.1.7) and (3.1.8), and finally to remove the scale to obtain $\bar{B}_+$. We let $W = JJ^T$ be any factorization of W and for $\bar{B} = [B, C]$ and $\bar{B}_+ = [B_+, C_+]$ we set $\tilde{\bar{B}} = [\tilde{B}, \tilde{C}] = [J^{-1}BJ^{-T}, J^{-1}C]$ and $\tilde{\bar{B}}_+ = [\tilde{B}_+, \tilde{C}_+] = [J^{-1}B_+J^{-T}, J^{-1}C_+]$. We note that $\bar{B}_+\bar{s} = y$ if and only if $\tilde{\bar{B}}_+\tilde{\bar{s}} = \tilde{y}$, where $\tilde{\bar{s}} = (\tilde{s}, t) = (J^T s, t)$ and $\tilde{y} = J^{-1}y$. Furthermore, $\|\bar{B}_+ - \bar{B}\|_W = \|\tilde{\bar{B}}_+ - \tilde{\bar{B}}\|_F$. It follows that $\bar{B}_+$ is the least-change secant update of $\bar{B}$ in $\mathbf{A}$ with respect to $\bar{s}$, y, and $\|\cdot\|_W$ if and only if $\tilde{\bar{B}}_+$ is the least-change secant update of $\tilde{\bar{B}}$ in $\mathbf{A}$ with respect $\tilde{\bar{s}}$, $\tilde{y}$, and $\|\cdot\|_F$. Determining $\tilde{\bar{B}}_+$ by (3.1.7) and (3.1.8) and removing the scale gives our extension of the DFP update, in which $\bar{\hat{y}} = (\hat{y}, t)$:

$$\begin{aligned}\bar{B}_+ = \bar{B} &+ \frac{(y - \bar{B}\bar{s})\bar{\hat{y}}^T + \left[\hat{y}(y - \bar{B}\bar{s})^T, (y - \bar{B}\bar{s})t^T\right]}{\bar{s}^T\bar{\hat{y}} + t^T t} \\ &- \frac{s^T(y - \bar{B}\bar{s})}{\bar{s}^T\bar{\hat{y}} + t^T t} \frac{\hat{y}\bar{\hat{y}}^T}{\bar{s}^T\bar{\hat{y}}}.\end{aligned} \tag{3.1.10}$$

To further clarify the structure of this update, we write $\bar{B}_+ = [B_+, C_+]$ for $B_+ \in \mathbf{R}^{n\times n}$ and $C_+ \in \mathbf{R}^{n\times m}$ and obtain from (3.1.10)

$$B_+ = B + \frac{(y - \bar{B}\bar{s})\hat{y}^T + \hat{y}(y - \bar{B}\bar{s})^T}{\bar{s}^T\hat{y} + t^Tt} - \frac{s^T(y - \bar{B}\bar{s})}{\bar{s}^T\hat{y} + t^Tt}\frac{\hat{y}\hat{y}^T}{\bar{s}^T\hat{y}},$$
$$C_+ = C + \frac{2(y - \bar{B}\bar{s})t^T}{\bar{s}^T\hat{y} + t^Tt} - \frac{s^T(y - \bar{B}\bar{s})}{\bar{s}^T\hat{y} + t^Tt}\frac{\hat{y}t^T}{\bar{s}^T\hat{y}}.$$

3.2. Least-change inverse secant updates. We now give extensions of several square-matrix least-change inverse secant updates, viz., the second Broyden update, a rank-two symmetry preserving update due to Greenstadt [18], which is the inverse-update analogue of the PSB update, and the Broyden–Fletcher–Goldfarb–Shanno (BFGS) update. As we remarked in §2, the BFGS update is generally regarded as the most successful update for unconstrained optimization, and the second Broyden update, while generally not as effective as the first, may have some advantages in stiff ODE applications. In the square-matrix case, the Greenstadt update is generally less successful than the PSB update. We include it here both for completeness and because it may prove to have particular advantages on some special problems. It would be straightforward to prescribe an inverse-update analogue of the sparse Broyden update, but the lack of applications of such an update makes this seem of doubtful worth.

We assume that we are given $\bar{K} \in \mathbf{R}^{n\times\bar{n}}$, nonzero $\bar{s} \in \mathbf{R}^{\bar{n}}$, and $y \in \mathbf{R}^n$. We write $\bar{s} = (s, t)$ for $s \in \mathbf{R}^n$ and $t \in \mathbf{R}^m$ and set $\bar{y} = (y, t)$. We also write $\bar{K} = [K, L]$, where $K \in \mathbf{R}^{n\times n}$ and $L \in \mathbf{R}^{n\times m}$, and denote each updated matrix by $\bar{K}_+ = [K_+, L_+]$. We note that as before each update below gives the usual square-matrix update of K when $t = 0$. Each of these updates is an inverse-update analogue of a direct update in §3.1 and not only can be derived by analogous reasoning but indeed is obtained just by renaming the ingredients in the appropriate formula in §3.1.

The second Broyden update. As in the case of the first Broyden update, we take $\mathbf{A} = \mathbf{S} = \mathbf{R}^{n\times\bar{n}}$ and $\|\cdot\| = \|\cdot\|_F$. Then we obtain from (3.1.1) that

$$\begin{aligned}\bar{K}_+ &= P_{\mathbf{M}(\mathbf{A},\mathbf{Q}(s,\bar{y}))}\bar{K} = P_{\mathbf{Q}(s,\bar{y})}\bar{K}\\ &= \bar{K}\left[I - \frac{\bar{y}\bar{y}^T}{\bar{y}^T\bar{y}}\right] + \frac{s\bar{y}^T}{\bar{y}^T\bar{y}}\\ &= \bar{K} + \frac{(s - \bar{K}\bar{y})\bar{y}^T}{\bar{y}^T\bar{y}}.\end{aligned} \tag{3.2.1}$$

The Greenstadt update. As with the PSB and DFP updates, we take

$$\mathbf{A} = \mathbf{S} = \left\{\bar{M} = [M, N] \in \mathbf{R}^{n\times\bar{n}} : M = M^T \in \mathbf{R}^{n\times n}\right\}$$

and $\|\cdot\| = \|\cdot\|_F$. Then (3.1.7) yields

$$\bar{K}_+ = \bar{K} + \frac{(s - \bar{K}\bar{y})\bar{y}^T + \left[y(s - \bar{K}\bar{y})^T, (s - \bar{K}\bar{y})t^T\right]}{\bar{y}^T\bar{y} + t^Tt} - \frac{y^T(s - \bar{K}\bar{y})}{\bar{y}^T\bar{y} + t^Tt}\frac{y\bar{y}^T}{\bar{y}^T\bar{y}},$$

or

$$K_+ = K + \frac{(s - \bar{K}\bar{y})y^T + y(s - \bar{K}\bar{y})^T}{\bar{y}^T\bar{y} + t^Tt} - \frac{y^T(s - \bar{K}\bar{y})}{\bar{y}^T\bar{y} + t^Tt}\frac{yy^T}{\bar{y}^T\bar{y}},$$
$$L_+ = L + \frac{2(s - \bar{K}\bar{y})t^T}{\bar{y}^T\bar{y} + t^Tt} - \frac{y^T(s - \bar{K}\bar{y})}{\bar{y}^T\bar{y} + t^Tt}\frac{yt^T}{\bar{y}^T\bar{y}}.$$

The Broyden–Fletcher–Goldfarb–Shanno update. We again take

$$\mathbf{A} = \mathbf{S} = \left\{\bar{M} = [M, N] \in \mathbf{R}^{n\times\bar{n}} : M = M^T \in \mathbf{R}^{n\times n}\right\}$$

and define the norm as a weighted Frobenius norm as follows: Set $\hat{s} = s - Lt \approx F_x(\bar{x})^{-1}y$, assume $\hat{s}^T y > 0$, and let $W \in \mathbf{R}^{n\times n}$ be any positive-definite symmetric matrix such that $Wy = \hat{s}$. Then for $\bar{M} = [M, N] \in \mathbf{R}^{n\times\bar{n}}$, define

$$\|\bar{M}\|_W^2 = \operatorname{tr}\left\{W^{-1}MW^{-1}M^T + W^{-1}NN^T\right\}.$$

From (3.1.10) we have that

$$\begin{aligned}\bar{K}_+ = \bar{K} &+ \frac{(s - \bar{K}\bar{y})\bar{\hat{s}}^T + \left[\hat{s}(s - \bar{K}\bar{y})^T, (s - \bar{K}\bar{y})t^T\right]}{\bar{y}^T\bar{\hat{s}} + t^Tt} \\ &- \frac{y^T(s - \bar{K}\bar{y})}{\bar{y}^T\bar{\hat{s}} + t^Tt}\frac{\hat{s}\bar{\hat{s}}^T}{\bar{y}^T\bar{\hat{s}}},\end{aligned} \tag{3.2.2}$$

where $\bar{\hat{s}} = (\hat{s}, t)$, or

$$\begin{aligned}K_+ &= K + \frac{(s - \bar{K}\bar{y})\hat{s}^T + \hat{s}(s - \bar{K}\bar{y})^T}{\bar{y}^T\bar{\hat{s}} + t^Tt} - \frac{y^T(s - \bar{K}\bar{y})}{\bar{y}^T\bar{\hat{s}} + t^Tt}\frac{\hat{s}\hat{s}^T}{\bar{y}^T\bar{\hat{s}}}, \\ L_+ &= L + \frac{2(s - \bar{K}\bar{y})t^T}{\bar{y}^T\bar{\hat{s}} + t^Tt} - \frac{y^T(s - \bar{K}\bar{y})}{\bar{y}^T\bar{\hat{s}} + t^Tt}\frac{\hat{s}t^T}{\bar{y}^T\bar{\hat{s}}}.\end{aligned} \tag{3.2.3}$$

In the square-matrix case, it is often desirable in practice to update approximate Jacobians directly instead of updating their inverses, e.g., in order to maintain and update QR or Cholesky factorizations of approximate Jacobians. Direct update formulas can be obtained in a straightforward way from the expressions above via the Sherman–Morrison–Woodbury formula (see, e.g., [25, p. 50]), and we give such formulas below for the second Broyden update and the BFGS update. Although the direct update formula for the second Broyden update is easy to obtain, deriving the direct formulas for the Greenstadt and BFGS updates is quite tedious. We do not discuss the direct formula for the Greenstadt update here since there appears to be nothing additional to be learned from it and it seems unlikely to be widely applied. Assuming K and K_+ are invertible, we set $\bar{B} = [B, C]$, where $B = K^{-1}$ and $C = -K^{-1}L$, and $\bar{B}_+ = [B_+, C_+]$, where $B_+ = K_+^{-1}$ and $C_+ = -K_+^{-1}L_+$.

For the second Broyden update, an application of the rank-one Sherman–Morrison–Woodbury formula to K_+ given by (3.2.1) yields

$$B_+ = B + \frac{(y - \bar{B}\bar{s})y^TB}{y^T\bar{B}\bar{s} + t^Tt},$$

which gives

$$C_+ = C + \frac{(y - \bar{B}\bar{s})(y^TC + t^T)}{y^T\bar{B}\bar{s} + t^Tt}.$$

For the BFGS update, an application of the rank-two Sherman–Morrison–Woodbury formula to K_+ given by (3.2.2) and (3.2.3) yields (after a very long computation)

$$B_+ = B + \frac{(\bar{B}\bar{s})^TB^{-1}(\bar{B}\bar{s})}{d}yy^T - \frac{c}{d}(\bar{B}\bar{s})(\bar{B}\bar{s})^T + \frac{2t^Tt}{d}\left[y(\bar{B}\bar{s})^T + (\bar{B}\bar{s})y^T\right], \tag{3.2.4}$$

which gives

$$C_+ = B_+B^{-1}\left[C + \frac{2}{d_2}yt^T - \frac{d_2 + y^TB^{-1}y}{d_1d_2}(\bar{B}\bar{s})t^T\right], \tag{3.2.5}$$

where

$$\begin{aligned} d_1 &= y^TB^{-1}\bar{B}\bar{s} + t^Tt, \\ d_2 &= y^TB^{-1}\bar{B}\bar{s} + 2t^Tt, \\ c &= d_2 + \frac{(y^TB^{-1}y + t^Tt)\,d_2}{d_1} - y^TB^{-1}y, \\ d &= 4\,(t^Tt)^2 + (\bar{B}\bar{s})^TB^{-1}(\bar{B}\bar{s})c. \end{aligned} \tag{3.2.6}$$

There are many possible alternatives to these expressions for B_+ and C_+, but these are as appealing to us from both aesthetic and computational points of view as any others which we have explored. The apparent computational difficulty of applying (3.2.5) is mitigated somewhat by the fact that

$$\begin{aligned} B_+B^{-1} = I &+ \frac{(\bar{B}\bar{s})^TB^{-1}(\bar{B}\bar{s})}{d}y(B^{-1}y)^T - \frac{c}{d}(\bar{B}\bar{s})(B^{-1}\bar{B}\bar{s})^T \\ &+ \frac{2t^Tt}{d}\left[y(B^{-1}\bar{B}\bar{s})^T + (\bar{B}\bar{s})(B^{-1}y)^T\right], \end{aligned} \tag{3.2.7}$$

and so by using (3.2.7), the application of (3.2.5) involves mainly forming dot products and linear combinations of vectors and, in particular, no excessive system solving. We note, however, that forming $B^{-1}y$ and $B^{-1}\bar{B}\bar{s}$ requires the solution of two systems involving B, and we have not been able to find any alternative to these expressions that does not require this.

An alternative which might be preferred is to maintain B and L instead of B and C. For example, iteration (4.1) in the next section is naturally implemented by maintaining B in factored form together with L. In any case, C can be recovered from B and L if desired by $C = -BL$, and maintaining B and L offers certain arithmetic advantages. Indeed, from (3.2.3) and (3.2.4) and the identities $\bar{B}\bar{s} = B\hat{s}$ and $s - \bar{K}\bar{y} = \hat{s} - B^{-1}y$, we have

$$\begin{aligned} B_+ &= B + \frac{\hat{s}^TBs}{d}yy^T - \frac{c}{d}(B\hat{s})(B\hat{s})^T + \frac{2t^Tt}{d}\left[y(B\hat{s})^T + (B\hat{s})y^T\right], \\ L_+ &= L + \frac{2(\hat{s} - B^{-1}y)t^T}{d_2} - \frac{y^T(\hat{s} - B^{-1}y)}{d_2}\frac{\hat{s}t^T}{d_1}, \end{aligned} \tag{3.2.8}$$

where

$$\begin{aligned} d_1 &= y^T\hat{s} + t^Tt, \\ d_2 &= y^T\hat{s} + 2t^Tt, \\ c &= d_2 + \frac{(y^TB^{-1}y + t^Tt)\,d_2}{d_1} - y^TB^{-1}y, \\ d &= 4\,(t^Tt)^2 + \hat{s}^TB\hat{s}c. \end{aligned} \tag{3.2.9}$$

The implementation of these expressions requires the solution of only one system involving B, and considerably less work is required to generate L_+ from L using (3.2.8) and (3.2.9) than is required to generate C_+ from C using (3.2.5)–(3.2.7).

4. A local convergence analysis. In this section we outline a general local convergence analysis for certain iterations for solving parameter-dependent systems which use the general least-change secant and inverse secant updates defined in §2. Our objective is to provide some theoretical support for using these general updates and the specific updates in §3 to augment the motivation provided by the success of least-change secant updates in the square-matrix case.

Iterative methods used to solve parameter-dependent systems vary widely in practice and are usually structured with particular applications or contexts in mind. Here we consider certain paradigm iterations which we feel represent in an essential way a significant class of methods that has not been treated elsewhere. We have in mind methods in which some explicit control is exercised over successive parameter values and, in particular, the way in which parameter values approach final or limiting values. For treatments of other iterative methods applicable to solving parameter-dependent systems, especially the normal flow and augmented Jacobian algorithms, see Walker and Watson [31] and Martínez [20].

Our paradigm iterations and local convergence analysis could have taken several forms, and we have tried to offer a good compromise of generality, presentability, etc., which still serves the primary purpose, viz., to show that *if updating is done according to certain principles, then successive approximate Jacobians will approximate the actual Jacobian increasingly well in the directions which count and desirable convergence properties will follow.* The main point of our analysis can be summarized as follows: If the Jacobian has the structure being imposed on updates and if various ancillary hypotheses hold, then locally *the iterates produced by the paradigm iterations converge as fast as the parameter convergence will allow, at least up to q-superlinear convergence.*

In our analysis, parameter values are explicitly allowed to approach a limiting value without necessarily ever reaching it, and the rate at which this limiting value is approached plays an important role. These aspects of our analysis are essential not only to establishing the main point summarized above but also to meaningfully addressing the basic issue, which is whether anything is to be gained by nonsquare-matrix least-change secant updating. If we were to consider only the case in which parameter values attain some final value after a finite number of steps, then the analysis would provide no basis for distinguishing between doing nonsquare-matrix updating at each step and, say, doing no updating at all until the final parameter value is reached and subsequently doing conventional square-matrix updating; indeed, it follows from the analysis in [12] that the latter leads to q-superlinear local convergence. We emphasize that the iterations considered here are intended to serve as paradigms. However, it should be possible to modify our analysis without much difficulty to apply to some other iterations which do not fit within the framework given here; see the discussion in §5 after Algorithm 5.1 for an illustration.

Our analysis closely parallels the general local convergence analysis for square-matrix least-change secant update methods given by Dennis and Walker [12]. It is not fully as general as the analysis of [12]: For one thing, it does not include cases in which part of the Jacobian is computed and part is approximated by a matrix maintained by updating. However, it can easily be extended to include such cases; they have not been considered here only to simplify the exposition. For the same reason, our iterations do not include the option of not updating, although it would be trivial to do so. More seriously, the analysis here does not include cases in which the norm used to define least-change secant updates varies from one iteration to the next (the "iteratively rescaled" cases of [12]). These cases are very important and include

the nonsquare extensions of the DFP and BFGS updates given in §3. It might be considered a serious shortcoming of the analysis given here that these cases are not included; however, we note below that there are still good heuristic arguments based on our analysis which support the use of these updates.

Otherwise, we have retained the generality of [12] in our analysis. In particular, we have allowed for cases in which the Jacobian at the solution of interest is not in the set of allowable approximants in order to show how the speed of convergence is affected in these cases. Since the usual secant equation determined by a difference in F-values is unlikely to be appropriate in these cases, we have also allowed for cases in which there is a variety of admissible secant equations determined by a "choice rule." As in [12], we assume a choice rule is given as a function χ which for each pair $\bar{x}$, $\bar{x}_+ \in \mathbf{R}^{\bar{n}}$ determines a set $\chi(\bar{x}, \bar{x}_+) \subseteq \mathbf{R}^n$ of admissible right-hand sides of secant equations.

Following [12], we have carried out most of the difficult technical work underlying our analysis in an appendix. The results in the Appendix, while used here only to support the analysis in this section, are rather general and may be of interest in their own right.

We assume below that $\mathbf{A} = A_N + \mathbf{S} \subseteq \mathbf{R}^{n \times \bar{n}}$ is a given affine subspace which reflects structure to be imposed on approximate Jacobians and that $\|\cdot\|$ is a given inner-product norm on $\mathbf{R}^{n \times \bar{n}}$. As before, we denote all vector norms and their induced matrix norms by $|\cdot|$, and here we assume that particular norms $|\cdot|$ are given on $\mathbf{R}^n$ and $\mathbf{R}^m$. We also consider various norms on $\mathbf{R}^{\bar{n}}$ which are specified in each instance using the norms on $\mathbf{R}^n$ and $\mathbf{R}^m$.

We consider a nonlinear system (1.2) and assume that a particular $\bar{x}_* = (x_*, \lambda_*)$ is sought satisfying $F(\bar{x}_*) = 0$. Throughout the sequel we assume the following hypothesis.

BASIC HYPOTHESIS. *Let F be differentiable in an open convex neighborhood Ω of $\bar{x}_* = (x_*, \lambda_*) \in \mathbf{R}^{\bar{n}}$ for which $F(\bar{x}_*) = 0$ and suppose that $\gamma \geq 0$ and $p \in (0, 1]$ are such that for $\bar{x} \in \Omega$*

$$|F'(\bar{x}) - F'(\bar{x}_*)| \leq \gamma |\bar{x} - \bar{x}_*|^p ,$$

where $|\bar{x}| = \max\{|x|, |\lambda|\}$ for $\bar{x} = (x, \lambda) \in \mathbf{R}^{\bar{n}}$.

We assume that $\{\lambda_k\}_{k=0,1,\cdots}$ is a sequence prescribed in some way which converges to λ_* and first consider an iteration which begins with some $\bar{x}_0 = (x_0, \lambda_{k_0})$ and $\bar{B}_0 = [B_0, C_0]$ and determines

$$\begin{aligned} x_{k+1} &= x_k - B_k^{-1} \{F(\bar{x}_k) + C_k (\lambda_{k_0+k+1} - \lambda_{k_0+k})\}, \\ \bar{x}_{k+1} &= (x_{k+1}, \lambda_{k_0+k+1}), \qquad y_k \in \chi(\bar{x}_k, \bar{x}_{k+1}), \\ \bar{B}_{k+1} &= [B_{k+1}, C_{k+1}] = (\bar{B}_k)_+ , \end{aligned} \tag{4.1}$$

for $k = 0, 1, \cdots$, where $(\bar{B}_k)_+$ is the least-change secant update of $\bar{B}_k$ in $\mathbf{A}$ with respect to $\bar{s}_k = \bar{x}_{k+1} - \bar{x}_k$, y_k, and the norm $\|\cdot\|$. An analogous iteration which uses least-change inverse secant updates is given by (4.11) below. Iteration (4.1) is modeled after an iteration derived in the manner of Newton's method in which

$$x_{k+1} = x_k - F_x(\bar{x}_k)^{-1} \{F(\bar{x}_k) + F_\lambda(\bar{x}_k) (\lambda_{k_0+k+1} - \lambda_{k_0+k})\} . \tag{4.2}$$

Note that (4.2) reverts to the Newton iteration in x if $\lambda_{k_0+k+1} = \lambda_{k_0+k}$. The index k_0 is used in (4.1) mainly for convenience in our local convergence analysis; we elaborate further on the use of k_0 in the remark following Theorem 4.1 below. We are not

concerned here with exactly how the λ_k's are prescribed. They may be completely specified in some way prior to beginning the iterations. However, in some iterations used in practice, they are determined as the iterations proceed, rather than at the outset, and it happens in some cases that they often remain constant through a range of k-values before changing in some specified way. For a clarifying example, see Algorithm 5.1 and the related discussion in §5.

The assumption in (4.2) that F_x is invertible might seem troubling to those concerned with homotopy methods since it suggests an assumption that a particular set of n columns of F' is of rank n. However, if $F'(\bar{x}_*)$ itself is of rank n, then the variables can be relabeled if necessary so that F_x is invertible near $\bar{x}_*$. We point out that with this understanding, certain corrector iterations used in predictor-corrector homotopy methods, such as the augmented Jacobian algorithm, the normal flow algorithm, and variations of these which use least-change secant and inverse secant updates (see Watson, Billups, and Morgan [32] and Walker and Watson [31]), fall within the framework of iterations (4.1), (4.2), and (4.11) below. However, our *analysis* below fails to apply to those iterations because the main results are conditioned on the rate of q-convergence of the λ_k's to λ_*, and nothing can be said about this q-convergence for those iterations. See Walker and Watson [31] for a local q-linear and q-superlinear convergence analysis which applies to those iterations.

Our local convergence results for iteration (4.1) are given in Theorems 4.1 and 4.2 below, which are analogues of Theorems 3.1 and 3.3 of [12]. Theorem 4.1 addresses the local linear convergence of the iteration; Theorem 4.2 drαws more refined conclusions about the asymptotic speed of convergence. In Theorem 4.1, we have augmented the local q-linear convergence result which the reader might expect with a statement about the local r-linear convergence of the iteration sequence in the case in which $\lambda_k \to \lambda_*$ r-linearly. This has been done for two reasons: First, r-linear convergence of $\{\lambda_k\}$ to λ_* is all that can be expected in many iterations used in practice. Second, the reader might otherwise accuse us of offering only a disguised version of the theory of [12] in cases in which $\lambda_k \to \lambda_*$ q-linearly, but only because $\lambda_k = \lambda_*$ for large k.

Our notation and terminology used in association with q-linear and q-superlinear convergence is that of Ortega and Rheinboldt [25, p. 281]: If $\{z_k\}_{k=0,1,\cdots}$ is a sequence converging to z_* in a space with norm $|\cdot|$, then $Q_1\{z_k\}$, the *linear q-factor of* $\{z_k\}_{k=0,1,\cdots}$, is defined as

$$Q_1\{z_k\} = \begin{cases} 0 & \text{if } z_k = z_*,\ k \geq \text{some } k_0, \\ \overline{\lim}_{k\to\infty} |z_{k+1} - z_*|/|z_k - z_*| & \text{if } z_k \neq z_*,\ k \geq \text{some } k_0, \\ +\infty & \text{otherwise.} \end{cases}$$

We say that $\{z_k\}_{k=0,1,\cdots}$ converges q-linearly to z_* in the norm $|\cdot|$ if and only if $Q_1\{z_k\} < 1$ and that $\{z_k\}_{k=0,1,\cdots}$ converges q-superlinearly to z_* if and only if $Q_1\{z_k\} = 0$. Note that q-superlinear convergence holds in one norm if and only if it holds in every other norm. Also, in saying that a sequence of matrices is uniformly small or uniformly bounded, we mean that in any matrix norm the corresponding sequence of norm values is uniformly small or uniformly bounded.

THEOREM 4.1. *Let F satisfy the basic hypothesis. Let $\mathbf{A}$ and $\bar{B}_* = [B_*, C_*] \equiv P_{\mathbf{A}}(F'(\bar{x}_*))$ be such that $B_* \in \mathbf{R}^{n\times n}$ is invertible and there exists an r_x for which $|I - {B_*}^{-1}F_x(\bar{x}_*)| \leq r_x < 1$. Assume that χ has the property with $\mathbf{A}$ that there exists an $\alpha \geq 0$ such that for any $\bar{x}$, $\bar{x}_+ \in \Omega$ and any $y \in \chi(\bar{x}, \bar{x}_+)$, we have*

$$\left\| P^{\perp}_{\mathbf{S}\cap\mathbf{N}(\bar{s})}(\bar{G} - \bar{B}_*) \right\| \leq \alpha\sigma(\bar{x}, \bar{x}_+)^p \tag{4.3}$$

for every $\bar{G} \in \mathbf{M}(\mathbf{A}, \mathbf{Q}(y, \bar{s}))$, *where* $\bar{s} = \bar{x}_+ - \bar{x}$ *and* $\sigma(\bar{x}, \bar{x}_+) = \max\{|\bar{x} - \bar{x}_*|, |\bar{x}_+ - \bar{x}_*|\}$ *for* $|\bar{x}| = \max\{|x|, |\lambda|\}$ *for* $\bar{x} = (x, \lambda) \in \mathbf{R}^{\bar{n}}$. *Let* $\{\lambda_k\}_{k=0,1,\cdots}$ *be such that one of the following holds:*

(4.4) $|\lambda_k - \lambda_*| \leq \beta r_\lambda^k$, $k = 0, 1, \cdots$, *for some* β *and* $r_\lambda \in (0, 1)$;

(4.5) $\lambda_k \to \lambda_*$ *q-linearly with* $Q_1\{\lambda_k\} = r_\lambda \in [0, 1)$.

Then for any r *such that* $\max\{r_x, r_\lambda\} < r < 1$, *there exist positive constants* ϵ_r, δ_r *and an integer* k_r *such that if* $|x_0 - x_*| < \epsilon_r$, $|\bar{B}_0 - \bar{B}_*| < \delta_r$, *and* $k_0 \geq k_r$, *then iteration* (4.1) *is well defined for* $k = 0, 1, \cdots$, *and converges to* $\bar{x}_*$ *with*

(4.6) $|\bar{x}_k - \bar{x}_*| \leq r^k \epsilon_r$, $k = 0, 1, \cdots$, *if* (4.4) *holds, where* $|\bar{x}| = \max\{|x|, |\lambda|\}$ *for* $\bar{x} = (x, \lambda) \in \mathbf{R}^{\bar{n}}$;

(4.7) $|\bar{x}_{k+1} - \bar{x}_*| \leq r|\bar{x}_k - \bar{x}_*|$, $k = 0, 1, \cdots$, *if* (4.5) *holds, where* $|\bar{x}| = |x| + \Gamma|\lambda|$ *for* $\bar{x} = (x, \lambda) \in \mathbf{R}^{\bar{n}}$ *and for a suitable constant* Γ.

Furthermore, $\{\bar{B}_k - \bar{B}_*\}_{k=0,1,\cdots}$ *and* $\{B_k^{-1} - B_*^{-1}\}_{k=0,1,\cdots}$ *are uniformly small.*

Remark. Our reasons for using the index k_0 in this analysis are somewhat subtle, and we will outline them in the context of iteration (4.1) and Theorem 4.1. Similar remarks are valid in the context of iteration (4.11) and Theorem 4.3 below. The bulk of the technical work underlying Theorem 4.1 is contained in Theorem A.2 in the Appendix, an examination of the proof of which shows that the λ_k's are required, at least eventually, to satisfy one of the following:

(i) an r-linear convergence inequality $|\lambda_k - \lambda_*| \leq \beta r_\lambda^k$, with βr_λ^k small for each k;

(ii) a q-linear convergence inequality $|\lambda_{k+1} - \lambda_*| \leq r''|\lambda_k - \lambda_*|$, with $|\lambda_k - \lambda_*|$ small for each k, where r'' is near but strictly greater than $Q_1\{\lambda_k\}$ (which may be zero).

Since these conditions hold for sufficiently large k under (4.4) and (4.5), respectively, our use of k_0 in conjunction with (4.4) and (4.5) is really an economical way of imposing these conditions.

Proof. Let $N_1 \subseteq \mathbf{R}^{\bar{n}}$ and $N_2 \subseteq \mathbf{R}^{n \times \bar{n}}$ be neighborhoods of $\bar{x}_*$ and $\bar{B}_*$, respectively, such that $N_1 \subseteq \Omega$ and the first n columns of any matrix in N_2 constitute a nonsingular matrix. Set $N = N_1 \times N_1 \times N_2$ and define an *update function* U on N (see the Appendix) by

$$U(\bar{x}, \bar{x}_+, \bar{B}) = \left\{\bar{B}_+ : y \in \chi(\bar{x}, \bar{x}_+)\right\},$$

where $\bar{B}_+$ is the least-change secant update of $\bar{B}$ with respect to $\bar{s} = \bar{x}_+ - \bar{x}$, $y \in \chi(\bar{x}, \bar{x}_+)$, and the norm $\|\cdot\|$. It follows immediately from (4.3) and (2.4) that U satisfies the bounded deterioration hypothesis of the Appendix with $\alpha_1 = 0$ and $\alpha_2 = \alpha$ of (4.3), and the theorem follows from Theorem A.2. □

THEOREM 4.2. *Suppose that the hypotheses of Theorem* 4.1 *hold and that for some* $\bar{x}_0$ *and* $\bar{B}_0$, $\{\bar{x}_k\}_{k=0,1,\cdots}$ *is a sequence generated by* (4.1) *which converges q-linearly to* $\bar{x}_*$ *in some norm with* $\bar{s}_k = \bar{x}_{k+1} - \bar{x}_k \neq 0$ *for all but finitely many* k. *Then*

$$\lim_{k\to\infty} \Big| \left[I - B_*^{-1} F_x(\bar{x}_*)\right](x_k - x_*) - (x_{k+1} - x_*) - B_*^{-1} C_*(\lambda_{k_0+k+1} - \lambda_*) + B_*^{-1}\left[C_* - F_\lambda(\bar{x}_*)\right](\lambda_{k_0+k} - \lambda_*) \Big| \Big/ |\bar{x}_k - \bar{x}_*| = 0 \tag{4.8}$$

for any norms on $\mathbf{R}^n$ *and* $\mathbf{R}^{\bar{n}}$. *In particular, if* $r_x = |I - B_*^{-1} F_x(\bar{x}_*)|$ *and* $r_\lambda = Q_1\{\lambda_k\}$, *then the following hold:*

(4.9) *for any* $\epsilon > 0$, $\overline{\lim}_{k\to\infty} |\bar{x}_{k+1} - \bar{x}_*| / |\bar{x}_k - \bar{x}_*| \leq \max\{r_x, r_\lambda + \epsilon\}$, *where* $|\bar{x}| = |x| + \Gamma|\lambda|$ *for* $\bar{x} = (x, \lambda) \in \mathbf{R}^{\bar{n}}$ *and for a suitable constant* Γ *depending on* ϵ;

(4.10) *if* $r_\lambda = 0$, *then* $\bar{x}_k \to \bar{x}_*$ *q-superlinearly if and only if the norm-independent condition*

$$\overline{\lim_{k\to\infty}} \frac{\left|\left[I - B_*^{-1}F_x(\bar{x}_*)\right](x_k - x_*) + B_*^{-1}\left[C_* - F_\lambda(\bar{x}_*)\right](\lambda_{k_0+k} - \lambda_*)\right|}{|\bar{x}_k - \bar{x}_*|} = 0$$

holds.

In particular, $\bar{x}_k \to \bar{x}_*$ *q-superlinearly if* $\lambda_k \to \lambda_*$ *q-superlinearly and* $F'(\bar{x}_*) \in \mathbf{A}$.

Proof. From Theorem A.3, (4.8) holds if and only if $\lim_{k\to\infty} |(\bar{B}_k - \bar{B}_*)\bar{s}_k|/|\bar{s}_k| = 0$, and this latter condition can be shown to hold by a very minor modification of the proof of Theorem 3.3 of [12]. Proposition A.4 gives (4.9) and (4.10). □

Theorems 4.1 and 4.2 have immediate consequences for an iteration (4.1) which uses the direct, fixed-scale least-change secant updates derived in §3. There is extensive discussion of the square-matrix analogue of condition (4.3) in [12, pp. 962–964], and this discussion is valid with minor appropriate changes in the present context. It follows in general from this discussion that if $F' \in \mathbf{A}$ near $\bar{x}_*$, then not only does $\bar{B}_* = F'(\bar{x}_*)$ but also condition (4.3) is satisfied near $\bar{x}_*$ for some α when $\chi(\bar{x}, \bar{x}_+) = \{F(\bar{x}_+) - F(\bar{x})\}$, the traditional choice which gives $y_k = F(\bar{x}_{k+1}) - F(\bar{x}_k)$ in iteration (4.1). It follows in particular from Theorems 4.1 and 4.2 that *if* $\lambda_k \to \lambda_*$ *q-superlinearly and if* $F'(\bar{x})$ *has the structure imposed on updates for* $\bar{x}$ *near* $\bar{x}_*$ *in each case in* §3, *then iteration* (4.1) *enjoys local q-superlinear convergence when the update is the nonsquare extension of any of the following: the first Broyden update, the Powell symmetric Broyden update, or the sparse Broyden update.*

We now establish counterparts of Theorems 4.1 and 4.2 for an analogue of iteration (4.1) which uses least-change inverse secant updates. We continue assuming that $\{\lambda_k\}_{k=0,1,\cdots}$ is a prescribed sequence which converges to λ_* and consider an iteration which begins with some $\bar{x}_0 = (x_0, \lambda_{k_0})$ and $\bar{K}_0 = [K_0, L_0]$ and determines

$$\begin{aligned} x_{k+1} &= x_k - K_k F(\bar{x}_k) + L_k\left(\lambda_{k_0+k+1} - \lambda_{k_0+k}\right), \\ \bar{x}_{k+1} &= (x_{k+1}, \lambda_{k_0+k+1}), \qquad y_k \in \chi(\bar{x}_k, \bar{x}_{k+1}), \\ \bar{K}_{k+1} &= [K_{k+1}, L_{k+1}] = \left(\bar{K}_k\right)_+, \end{aligned} \tag{4.11}$$

for $k = 0, 1, \cdots$, where $(\bar{K}_k)_+$ is the least-change inverse secant update of $\bar{K}_k$ in $\mathbf{A}$ with respect to $\bar{s}_k = \bar{x}_{k+1} - \bar{x}_k$, y_k, and $\|\cdot\|$. Theorems 4.3 and 4.4 below are analogous to Theorems 5.1 and 5.2 of [12]. Since we are not treating a "computed part" of approximate Jacobians, there is no compelling need to consider a choice rule for determining a vector w_k as well as y_k at each iteration as in [12, §5]. Here we let $s_k = x_{k+1} - x_k$ play the role of w_k in [12, §5] in determining inverse secant equations. The proofs of Theorems 4.3 and 4.4 are minor modifications of those of Theorems 4.1 and 4.2, and so we omit them.

THEOREM 4.3. *Let* F *satisfy the basic hypothesis and assume that* $F_x(\bar{x}_*)$ *is invertible. Let* $\mathbf{A}$ *and* $\bar{K}_* = [K_*, L_*] \equiv P_{\mathbf{A}}([F_x(\bar{x}_*)^{-1}, -F_x(\bar{x}_*)^{-1}F_\lambda(\bar{x}_*)])$ *be such that there exists an* r_x *for which* $|I - K_*F_x(\bar{x}_*)| \le r_x < 1$. *Assume that* χ *has the property with* $\mathbf{A}$ *that there exists an* $\alpha \ge 0$ *such that for any* $\bar{x}$, $\bar{x}_+ \in \Omega$ *and any* $y \in \chi(\bar{x}, \bar{x}_+)$, *we have*

$$\left\| P^{\perp}_{\mathbf{S}\cap\mathbf{N}(\bar{y})}(\bar{G} - \bar{K}_*) \right\| \le \alpha\sigma(\bar{x}, \bar{x}_+)^p$$

for every $\bar{G} \in \mathbf{M}(\mathbf{A}, \mathbf{Q}(s, \bar{y}))$, *where* $\bar{y} = (y, \lambda_+ - \lambda)$ *and* $\sigma(\bar{x}, \bar{x}_+) = \max\{|\bar{x} - \bar{x}_*|, |\bar{x}_+ - \bar{x}_*|\}$ *for* $|\bar{x}| = \max\{|x|, |\lambda|\}$ *for* $\bar{x} = (x, \lambda) \in \mathbf{R}^{\bar{n}}$. *Let* $\{\lambda_k\}_{k=0,1,\cdots}$ *be such that one of* (4.4) *or* (4.5) *holds. Then for any* r *such that* $\max\{r_x, r_\lambda\} < r < 1$, *there exist positive constants* ϵ_r, δ_r *and an integer* k_r *such that if* $|x_0 - x_*| < \epsilon_r$, $|\bar{K}_0 - \bar{K}_*| < \delta_r$, *and* $k_0 \geq k_r$, *then iteration* (4.11) *is well defined for* $k = 0, 1, \cdots$, *and converges to* $\bar{x}_*$ *according to either* (4.6), *if* (4.4) *holds, or* (4.7), *if* (4.5) *holds. Furthermore,* $\{\bar{K}_k - \bar{K}_*\}_{k=0,1,\cdots}$ *and* $\{K_k^{-1} - K_*^{-1}\}_{k=0,1,\cdots}$ *are uniformly small.*

THEOREM 4.4. *Suppose that the hypotheses of Theorem* 4.3 *hold and that for some* $\bar{x}_0$ *and* $\bar{K}_0$, $\{\bar{x}_k\}_{k=0,1,\cdots}$ *is a sequence generated by* (4.11) *which converges q-linearly to* $\bar{x}_*$ *in some norm with* $\bar{s}_k = \bar{x}_{k+1} - \bar{x}_k \neq 0$ *for all but finitely many* k. *Suppose further that* $\{\bar{K}_k\}_{k=0,1,\cdots}$ *and* $\{K_k^{-1}\}_{k=0,1,\cdots}$ *are uniformly bounded and that* $\{y_k\}_{k=0,1,\cdots}$ *satisfies* $|\bar{K}_* \bar{y}_k - s_k| \leq \alpha_k |\bar{s}_k|$ *for each* k, *where* $\bar{y}_k = (y_k, \lambda_{k_0+k+1} - \lambda_{k_0+k})$, $s_k = x_{k+1} - x_k$, *and* $\lim_{k\to\infty} \alpha_k = 0$. *Then*

$$\lim_{k\to\infty} \Big| [I - K_* F_x(\bar{x}_*)](x_k - x_*) - (x_{k+1} - x_*) + L_*(\lambda_{k_0+k+1} - \lambda_*) - [L_* + K_* F_\lambda(\bar{x}_*)](\lambda_{k_0+k} - \lambda_*) \Big| \Big/ |\bar{x}_k - \bar{x}_*| = 0$$

for any norms on $\mathbf{R}^n$ *and* $\mathbf{R}^{\bar{n}}$. *In particular, if* $r_x = |I - K_* F_x(\bar{x}_*)|$ *and* $r_\lambda = Q_1\{\lambda_k\}$, *then* (4.9) *and the following hold:*

if $r_\lambda = 0$, *then* $\bar{x}_k \to \bar{x}_*$ *q-superlinearly if and only if the norm-independent condition*

$$\overline{\lim_{k\to\infty}} \frac{|[I - K_* F_x(\bar{x}_*)](x_k - x_*) - [L_* + K_* F_\lambda(\bar{x}_*)](\lambda_{k_0+k} - \lambda_*)|}{|\bar{x}_k - \bar{x}_*|} = 0$$

holds.

In particular, $\bar{x}_k \to \bar{x}_*$ *q-superlinearly if* $\lambda_k \to \lambda_*$ *q-superlinearly and*

$$[F_x(\bar{x}_*)^{-1}, -F_x(\bar{x}_*)^{-1} F_\lambda(\bar{x}_*)] \in \mathbf{A}.$$

The discussion following Theorem 4.2 remains valid with a few appropriate changes. In particular, Theorems 4.3 and 4.4 imply that *if* $\lambda_k \to \lambda_*$ *q-superlinearly, then iteration* (4.11) *is locally and q-superlinearly convergent when the update is the nonsquare extension of the second Broyden update or, if* $F_x(\bar{x})$ *is symmetric for* $\bar{x}$ *near* $\bar{x}_*$, *the Greenstadt update.*

As we remarked at the beginning of this section, the above analysis does not apply to cases in which the norm used to determine least-change secant updates varies from iteration to iteration. In particular it does not apply to iterations which use the nonsquare extensions of the DFP and BFGS updates given in §3. The counterparts of these iterations in the square-matrix case are the iteratively rescaled least-change secant update methods considered in [12], which include the usual DFP and BFGS methods, and a full local convergence analysis for these methods is given in [12]. However, this local convergence analysis depends critically on Lemma 4.1, p. 969, of that paper, and we cannot adapt that lemma to apply to the present circumstances. To indicate why, we note that in the setting of the extension of the DFP update in §3, the roles of W_* and v in Lemma 4.1 are played by $F_x(\bar{x}_*)$ and $\hat{y} = y - Ct$, respectively, where C is part of a given approximate Jacobian $\bar{B} = [B, C]$ and $\bar{s} = (s, t) = \bar{x}_+ - \bar{x}$

and $y = F(\bar{x}_+) - F(\bar{x})$ for given $\bar{x}$, $\bar{x}_+$. Since $\hat{y}$ depends on C and $C \neq F_\lambda(\bar{x}_*)$ in general, we cannot expect an inequality $|\hat{y} - F_x(\bar{x}_*)s|_2 \leq \kappa\sigma(\bar{x}, \bar{x}_+)^p|s|_2$ to hold for all $\bar{x}$ and $\bar{x}_+$, where $|\cdot|_2$ is the Euclidean norm. This is to say that we cannot expect an inequality of the form (ii) of the hypotheses of Lemma 4.1 to hold in the context of our extension of the DFP update.

We note, however, that we can construct a local convergence analysis along the lines of that given above which, while not applicable to iterations using our extensions of the DFP and BFGS updates, *is* applicable to iterations using certain "nearby" updates. These "nearby" updates may be costly or even impractical in some applications. However, we feel that they are worth noting because they use more computed Jacobian information and therefore may result in more effective Jacobian approximations in some cases. A local convergence analysis for iterations using these "nearby" updates is of value in its own right; furthermore, we feel that it provides some heuristic support for using our extensions of the DFP and BFGS updates. We indicate what these "nearby" updates are and how a local convergence analysis for iterations using them can be developed in the DFP context: The "nearby" update is obtained from our extension of the DFP update by taking $\bar{s} = \bar{x}_+ - \bar{x}$ and $y = F(\bar{x}_+) - F(\bar{x})$ for given $\bar{x}$ and $\bar{x}_+$ and choosing $\hat{y} = y - F_\lambda(\bar{x}_*)t$, $\hat{y} = y - F_\lambda(\bar{x})t$, or $\hat{y} = y - F_\lambda(\bar{x}_+)t$ instead of $\hat{y} = y - Ct$. The first choice is usually impractical, but not always — see, e.g., the homotopy map (5.4) in §5, in which F_λ is constant. The second two choices can be obtained through either analytic evaluation or finite differences; either way will require function evaluation work which may by costly in some cases. Note that the evaluation of $F_\lambda(\bar{x})t$ or $F_\lambda(\bar{x}_+)t$ can be done through finite differences of F in the t-direction and, hence, does not require the evaluation of all of F_λ. (If it is reasonable to evaluate all of F_λ, then we can approximate F_x via traditional square-matrix DFP or BFGS updating in a straightforward way, and this would probably be more effective than using the "nearby" updates discussed here.) For any of these choices of $\hat{y}$, it can easily be shown that an inequality $|\hat{y} - F_x(\bar{x}_*)s|_2 \leq \kappa\sigma(\bar{x}, \bar{x}_+)^p|s|_2$ holds, and Lemma 4.1 of [12] can be readily adapted to apply to the present situation with $W_* = F_x(\bar{x}_*)$ and $v = \hat{y}$. From this point, it is straightforward to adapt the arguments in the proofs of Theorems 4.1 and 4.2 above together with those in the proofs of Theorems 4.2 and 4.3 of [12] to obtain local convergence results analogous to those above for an iteration (4.1) in which $\chi(\bar{x}, \bar{x}_+) = \{F(\bar{x}_+) - F(\bar{x})\}$ and the update is this "nearby" update, i.e., our extension of the DFP update with $\hat{y} = y - Ct$ replaced by one of the choices $\hat{y} = y - F_\lambda(\bar{x}_*)t$, $\hat{y} = y - F_\lambda(\bar{x})t$, or $\hat{y} = y - F_\lambda(\bar{x}_+)t$.

5. Numerical experiments. We report on some simple numerical experiments that are intended to give some insight into the performance of the updates discussed in this paper, especially in the context of iterations of the type considered in §4. We stress that the experimental results given here are by no means intended to provide a complete or thorough study of iterative methods which use these updates, but rather are intended to complement experimental results and other work discussed elsewhere.

For perspective, we briefly review other work involving applications of the updates given here. As noted in the introduction, the first Broyden update (3.1.1) has been used in an algorithm of Georg [16] and in the augmented Jacobian matrix algorithm in the homotopy method code HOMPACK [32]. Both of these are predictor-corrector path-following methods with predictor steps taken in current (approximate) tangent directions. In Georg's algorithm, corrector iterations are of the normal flow type, while the HOMPACK algorithm uses corrector iterations determined by certain augmented Jacobian matrices. (See Walker and Watson [31] as well as [16] and [32] for a

discussion of these iterations.) In Georg's algorithm, first Broyden updating is done on both predictor steps and corrector iterations. In the augmented Jacobian matrix algorithm in HOMPACK, first Broyden updating is done only on the corrector iterations; exact Jacobians are obtained and used at each predictor step. The update used in HOMPACK is described in [32] as a square-matrix first Broyden update of a certain augmented Jacobian matrix, but it is essentially the same as the nonsquare-matrix update (3.1.1).

Martínez [20] augments his local r-linear and r-superlinear local convergence analysis for very general Newton-like methods for underdetermined systems with a discussion of applications of these methods to the problem of finding an interior point of a polytope, which arises in connection with interior point methods for linear programming. In experiments reported in [20], a nonlinear transformation was used to rephrase this problem as one involving an underdetermined nonlinear system, and normal flow algorithms with both exact Jacobians and approximate Jacobians maintained by Broyden updating, together with other related algorithms, were applied to this problem. The Broyden updating was apparently either first Broyden updating or sparse Broyden updating as given in §3.1 here, although the exact nature of the updates used is not made clear in [20].

Walker and Watson [31], in addition to giving a local q-linear and q-superlinear local convergence analysis for general normal flow and augmented Jacobian algorithms which use the updates given in this paper, discuss two sets of numerical experiments involving these algorithms. In the first set, normal flow iterations using the first and second Broyden updates, (3.1.1) and (3.2.1), respectively, were applied to simple two-variable problems; in the second set, the normal flow and augmented Jacobian matrix algorithms in HOMPACK, together with variations which use first and second Broyden updating, were applied to a geometric modeling problem described by Morgan [24].

Other experiments involving use of the updates introduced here in modifications of algorithms in HOMPACK are reported by Bourji [3]. In these experiments, all of the specific updates introduced in §3 except the sparse Broyden and Greenstadt updates were implemented in modifications of the HOMPACK augmented Jacobian matrix algorithm, and the resulting algorithms were tested on five problems obtained by parametrizing problems from the test set of Moré, Garbow, and Hillstrom [23] with simple homotopy maps. The first and second Broyden updates, (3.1.1) and (3.2.1), respectively, and the PSB update (3.1.7) were tested on all five problems.

In the first problem, no subset of n columns of the Jacobian constituted a symmetric matrix; however, the PSB update was still tested on this problem because the $(n-1) \times (n-1)$ principal submatrix of the Jacobian was symmetric. The second through fifth problems amounted to parametrized nonlinear least-squares problems, and the DFP and BFGS updates, (3.1.10) and (3.2.2), respectively, as well as the first and second Broyden and PSB updates were tested on these. We refer the reader to [3] for more details of these experiments and summarize the results here. As measured by numbers of function and Jacobian evaluations, the performances and rankings of the algorithms using the different updates varied considerably from problem to problem. However, the algorithm using the first Broyden update clearly performed best overall; only in one of fifteen trials did another algorithm, the one using the PSB update, take fewer function or Jacobian evaluations. The algorithm which performed second best overall was the one using the PSB update, and the algorithms using the other updates often, but not in every case, performed considerably worse than those using the first Broyden and PSB updates. The algorithm using the DFP update was perhaps third best and outperformed the PSB-update algorithm in one trial. Each algorithm failed

in at least one instance, and there was one trial (involving the Rosenbrock function) in which only the algorithms using the second Broyden, DFP, and BFGS updates succeeded and another trial (involving the extended Rosenbrock function) in which only the algorithms using the DFP and BFGS updates succeeded. However, there were two trials in which only the BFGS-update algorithm failed and three trials in which only the DFP-update algorithm failed. In evaluating these results, it should be kept in mind that there are many things going on in the sophisticated homotopy method codes in HOMPACK, e.g., procedures for selecting stepsizes, determining when to reevaluate Jacobians, etc. The first Broyden update, which performed most successfully in these trials, is the update used in the unmodified augmented Jacobian matrix algorithm in HOMPACK and thus is the one for which the code is "tuned." In view of this, we feel that it would be premature to dismiss the other updates on the basis of these experiments.

As a complement to the above work, we give here the results of experiments involving the application of a simple path-following algorithm to a realistic problem. The object is to explore in some depth the performance of one of the updates given here—the first Broyden update (3.1.1)—in comparison to that of various alternatives involving numerically evaluated Jacobians and traditional square-matrix first Broyden updating.

The problem of interest is the *elastica problem* described by Watson and Wang [33]. In this problem, a large planar deflection of a thin rod, or *elastica*, subject to terminal loads $x^{(1)}$, $x^{(2)}$ and moment $x^{(3)}$ is modeled by

$$EI\frac{d\theta}{ds} = x^{(2)}\xi - x^{(1)}\eta + x^{(3)},$$

where EI is the flexural rigidity, θ is the local angle of inclination at the point (ξ, η), and s is the arc length along the elastica. For simplicity, we take $EI = 1$ and assume that the elastica is of unit length with the left endpoint clamped horizontally at the origin. This gives the initial-value problem

$$\begin{aligned} &\frac{d\theta}{ds} = x^{(2)}\xi - x^{(1)}\eta + x^{(3)}, \qquad \frac{d\xi}{ds} = \cos\theta, \qquad \frac{d\eta}{ds} = \sin\theta, \\ &\xi(0) = \eta(0) = \theta(0) = 0. \end{aligned} \tag{5.1}$$

We denote the solution by $\xi(s,x)$, $\eta(s,x)$, $\theta(s,x)$, where $x = (x^{(1)}, x^{(2)}, x^{(3)})^T \in \mathbf{R}^3$.

The problem is to determine x_* so that the right endpoint of the elastica has a given location and angle of inclination, i.e., so that $x = x_*$ satisfies

$$f(x) \equiv \begin{pmatrix} \xi(1,x) \\ \eta(1,x) \\ \theta(1,x) \end{pmatrix} = a \tag{5.2}$$

for a specified vector a giving the right endpoint location and angle of inclination. Because of the sensitivity of the elastica to end conditions, especially for more complicated shapes which require large forces and torques to achieve, solving (5.2) can be quite challenging for globalized Newton-like methods such as those in MINPACK [22]; see [33]. Homotopy methods seem to fare better. Our strategy here is to choose a homotopy map $F(x,\lambda)$ such that $F(x,0) = 0$ is easy to solve and $F(x,1) \equiv f(x) - a$, and then to track the zero curve of F as λ goes from 0 to 1. For a given F, the simple algorithm for following this curve which we used in our tests is the following.

ALGORITHM 5.1. Given x_0 such that $F(x_0, 0) = 0$, choose *nstep*, the number of λ-increments between 0 and 1, and set $h = 1/nstep$, $x = x_0$, and $\lambda = 0$.

For $i = 1, \cdots, nstep$, do:

1. Predict: Given (x, λ) and $\bar{B} = [B, C] \approx F'(x, \lambda)$, overwrite $x \leftarrow x - B^{-1}\{F(x, \lambda) + Ch\}$ and $\lambda \leftarrow \lambda + h$.
2. Correct: Given $\epsilon > 0$ and initial (x, λ) and $B \approx F_x(x, \lambda)$, overwrite $x \leftarrow x - B^{-1}F(x, \lambda)$ until $|F(x, \lambda)| \leq \epsilon$.

What is unspecified in Algorithm 5.1 is the manner in which successive $\bar{B}$'s and B's are determined. If they are determined by least-change secant or inverse secant updating at each step and if the corrector convergence test is passed after a finite number of iterations for each $i < nstep$, then this algorithm has the form of iteration (4.1) or (4.11), respectively, with an added convergence test when $i = nstep$. With this updating, a reasonable modification of the analysis in §4 and in the Appendix applies to a reasonable modification of Algorithm 5.1 as follows: Suppose $F(x_0, \lambda_0) = 0$ for some (x_0, λ_0) and in Algorithm 5.1 we initialize $h = (1-\lambda_0)/nstep$, $x = x_0$, and $\lambda = \lambda_0$ and take $\bar{B} = F'(x_0, \lambda_0)$ at the initial predictor step. Suppose also that we use some fixed ϵ in the corrector iterations for $i < nstep$ and take $\epsilon = 0$ for $i = nstep$, so that the algorithm proceeds indefinitely once $\lambda = 1$. Suppose finally that for (x, λ) near $(x_*, 1)$, $F'(x, \lambda)$ has the structure (if any) which is imposed on updates and $F_x(x, \lambda)$ is nonsingular. If (x_0, λ_0) is sufficiently near $(x_*, 1)$, then the iterates are well defined, h is small, and furthermore, using bounded deterioration (see the proof of Theorem 4.1 and the Appendix), we can verify that approximate Jacobians remain near their actual values for a given finite number of predictor steps and corrector iterations. It follows that if (x_0, λ_0) is sufficiently near $(x_*, 1)$, then the number of corrector iterations in step 2 is no greater than a prescribed bound for each $i < nstep$, and therefore that the sequence of λ-values used in the predictor steps and corrector iterations satisfies an r-linear convergence inequality (4.4) with $\lambda_* = 1$ and $\beta = |\lambda_0 - \lambda_*|$. With this, an inspection of the analysis in §4 and the Appendix shows that if (x_0, λ_0) is sufficiently near $(x_*, 1)$ and the appropriate ancillary hypotheses of the theorems in §4 are satisfied, then the iterate sequence $\{\bar{x}_k = (x_k, \lambda_k)\}_{k=0,1,\cdots}$ satisfies the r-linear convergence inequality of (4.6). After a finite number of iterations, we have $i = nstep$ and $\lambda = 1$ for all remaining iterations, and since approximate Jacobians are still near their true values if (x_0, λ_0) is sufficiently near $(x_*, 1)$, the convergence is ultimately q-superlinear.

Algorithm 5.1 is quite unsophisticated compared to other homotopy algorithms. For one thing, it uses a preset number of equally spaced λ-increments, rather than determining variable λ-increments in some intelligent way as the algorithm proceeds. Perhaps more seriously, because a prescribed value of λ is used in determining the next value of x at each step, the algorithm is likely to get into trouble if F_x is singular along the zero curve. This lack of sophistication is by design, however. The object of the experiments here is to give insight into the merits of various ways of determining successive $\bar{B}$'s and B's, and the simplicity of Algorithm 5.1 lends itself to this. Also, maintaining constant λ-values through the corrector iterations allows options for traditional square-matrix first Broyden updating in determining successive B's and so is useful for comparative testing.

In our experiments, we tested seven different strategies for determining the successive $\bar{B}$'s and B's in Algorithm 5.1. These are as follows:

Strategy 1. Take $\bar{B} = F'(x_0, 0)$ initially, and then maintain $\bar{B} = [B, C]$ and thereby B through nonsquare first Broyden updating on all subsequent predictor steps

and corrector iterations. See the remarks below.

Strategy 2. Take $\bar{B} = [B, C] = F'(x_0, 0)$ initially, and then use this initial $C = F_\lambda(x_0, 0)$ in all subsequent predictor steps while maintaining B through square-matrix first Broyden updating on all corrector iterations.

Strategy 3. Take $\bar{B} = [B, C] = F'(x_0, 0)$ initially, and then reevaluate $C = F_\lambda(x, \lambda)$ at each predictor step and maintain B through square-matrix first Broyden updating on all predictor steps as well as all corrector iterations, replacing the right-hand side of the secant equation (1.3) with the "ideal replacement" $F(x_+, \lambda_+) - F(x, \lambda) - F_\lambda(x, \lambda)(\lambda_+ - \lambda)$ when updating on the predictor steps. See the remarks below.

Strategy 4. Take $\bar{B} = F'(x, \lambda)$ at each predictor step, and at each set of corrector iterations take $B = F_x(x, \lambda)$ initially and then maintain B through square-matrix first Broyden updating on the subsequent corrector iterations.

Strategy 5. Take $\bar{B} = F'(x, \lambda)$ at each predictor step, and at each set of corrector iterations take $B = F_x(x, \lambda)$ initially and then use this B throughout the subsequent corrector iterations.

Strategy 6. Take $\bar{B} = F'(x, \lambda)$ at each predictor step and $B = F_x(x, \lambda)$ at each corrector iteration.

Strategy 7. Take $\bar{B} = [B, C] = F'(x_0, 0)$ initially, and then use this $\bar{B}$ and B in all subsequent predictor steps and corrector iterations.

Remarks. The ordering of these strategies very roughly reflects an increasing dependence on derivative evaluations and a decreasing dependence on updating. The total chord-method Strategy 7 is included mainly to note that it was quite unsuccessful in the experiments discussed here, as might be expected. The partial chord-method Strategy 5 also resulted in failure in the two particular trials reported below, but it showed some success in other trials while Strategy 7 nearly always led to failure. In Strategy 1, the nonsquare first Broyden updating of course reduces to square-matrix first Broyden updating on the corrector iterations because λ does not change. The updating in Strategy 3 is described as square-matrix first Broyden updating, but it can also be regarded as nonsquare-matrix updating in which the last column of F', i.e., F_λ, is computed while the remainder is approximated through Frobenius-norm least-change secant updating into the subspace of matrices the last column of which is zero. Thus *Strategies* 1 *and* 3 *can be regarded as the strategies which employ nonsquare-matrix updating.* As noted in §4, the discussion in §4 does not explicitly include cases in which part of the Jacobian is computed while the remainder is approximated by updating, but it would be straightforward to extend it to do so.

In the experiments reported here, we made a particular choice of F and in each of a number of trials counted the function evaluations, Jacobian evaluations, and corrector iterations required by the above strategies for the implementation of Algorithm 5.1. Evaluating the function F required the evaluation of f in (5.2). This was done by numerically solving (5.1) using subroutine RKF45 of Shampine and Watts (see Shampine, Watts, and Davenport [28]), which we obtained from the Forsythe, Malcolm, and Moler [15] collection of subroutines. Derivatives of F were obtained by forward-difference approximations, which provided adequate accuracy. The function evaluations required for these derivative approximations were included in the overall function evaluation counts; thus these counts really provide a measure of overall function evaluation work including that required for derivative evaluation. The evaluation of either F' or F_x counted as a Jacobian evaluation; the evaluation of F_λ alone did not. Thus the Jacobian evaluation counts reflect the number of matrix factorizations "from scratch" which were required, as well as the number of function evaluations

which arose from Jacobian evaluations. (As in the square-matrix case, we can update matrix factors or perhaps exercise other options to incorporate the rank-one Broyden update without obtaining a new factorization "from scratch".) In the trials reported here, we chose $a = (0, \pi/2, \pi)^T$, for which the solution of (5.2) is $x_* = (0, 0, \pi)^T$, i.e., the elastica is a semicircle. All computing reported here was done in double precision on a Sun 4/280S using the Sun Microsystems Fortran 1.1 compiler.

Our particular choice of F is

$$F(x, \lambda) = \lambda [f(x) - a] + (1 - \lambda)(x - x_0). \tag{5.3}$$

This choice appears to be not very well behaved. In particular we saw evidence in our testing of bifurcation of the zero curve of F, at least for the values of x_0 which were used in the tests reported here. Although we did not verify bifurcation with certainty, there were at least sharp changes in the direction of the zero curve for values of λ in $[.5, .6]$ and $[.9, 1]$, which resulted in relatively large numbers of corrector iterations for all strategies around the middle and last predictor steps. Still, for our testing we prefer (5.3) to an apparently better behaved alternative suggested in [33], viz.,

$$F(x, \lambda) = f(x) - \lambda a - (1 - \lambda) f(x_0). \tag{5.4}$$

First, the bad behavior of the zero curve of F given by (5.3) creates challenges for the methods being tested; second, if F were given by (5.4), then F_λ would be constant and there would be no need for nonsquare updating.

The results of two typical trials with F given by (5.3) are given in Tables 1 and 2 below. In both trials, we took $x_0 = (-.4, .4, 3)^T$. In the first trial, we also took $nstep = 10$ and used tolerances $\epsilon = 10^{-1}$ for $i < nstep$ and $\epsilon = 10^{-5}$ for $i = nstep$, taking the attitude that an accurate point on the zero curve was of interest only at the final step. In the second trial, we took a greater number of steps and used tighter tolerances, choosing $nstep = 20$ and $\epsilon = 10^{-2}$ for $i < nstep$ and $\epsilon = 10^{-6}$ for $i = nstep$. A maximum number of 20 corrector iterations was allowed before convergence failure was declared.

TABLE 1
Results of trial 1 *for* F *given by* (5.3).

Strategy	1	2	3	4	5	6	7
Corrector Iterations	36	45	22	15	*	7	*
Function Evaluations	51	60	46	78	*	79	*
Jacobian Evaluations	1	1	1	14	*	17	*

* corrector convergence failure at the 6th step

It is evident in these trials that the strategies which use more computed derivative information require (perhaps significantly) fewer corrector iterations, but these strategies may require (perhaps significantly) more function evaluation effort, not to mention matrix arithmetic, and updating, either alone or in conjunction with computed derivative information, offers (perhaps significant) advantages. The success of Strategy 4 and the failure of Strategy 5 show the usefulness of updating on the corrector iterations. The success of Strategies 1–3 shows the effectiveness of updating

TABLE 2
Results of trial 2 for F given by (5.3).

Strategy	1	2	3	4	5	6	7
Corrector Iterations	53	65	38	23	*	18	*
Function Evaluations	78	90	82	151	*	173	*
Jacobian Evaluations	1	1	1	29	*	38	*

* corrector convergence failure at the 11th step

in maintaining Jacobian approximations over the full range of λ-values, even if the updating is only traditional square-matrix first Broyden updating on the corrector iterations. However, comparing the results for Strategy 2 with those for Strategies 1 and 3 shows the importance of updating on the predictor steps as well as on the corrector iterations, i.e., the importance of nonsquare-matrix updating. It happened in a number of other trials not reported here that Strategies 1 and 3 succeeded while Strategy 2 failed. For example, for $x_0 = (-.5, .5, 3)^T$, *nstep* $= 20$, $\epsilon = 10^{-2}$ for $i <$ *nstep* and $\epsilon = 10^{-6}$ for $i =$ *nstep*, Strategy 2 led to corrector convergence failure at the eleventh step while Strategies 1 and 3 succeeded. A comparison of the results for Strategies 1 and 3 reinforces the conventional wisdom that computing part of the Jacobian results in more effective Jacobian approximations. However, this partial computation of the Jacobian requires additional function evaluation work in proportion to the number of predictor steps. Strategy 3 has an overall function-evaluation advantage over Strategy 1 with the ten predictor steps in trial 1, but the advantage is reversed with the twenty predictor steps in trial 2.

Appendix. We now give analogues of the results in the Appendix of [12] which are suitable for application to the local convergence analysis given in §4. We assume as in §4 that $F : \mathbf{R}^{\bar{n}} \to \mathbf{R}^n$ satisfies the basic hypothesis near $\bar{x}_* = (x_*, \lambda_*)$ and that $\{\lambda_k\}_{k=0,1,\cdots}$ is a sequence which converges to λ_*. Our convention regarding norms is as in §4, i.e., $\|\cdot\|$ denotes a particular inner product norm on $\mathbf{R}^{n\times\bar{n}}$ and $|\cdot|$ denotes all of the following: particular norms on $\mathbf{R}^n$ and $\mathbf{R}^m$, various norms on $\mathbf{R}^{\bar{n}}$ which are specified in each instance, and the matrix norms induced by these vector norms.

Our interest is in a very general iteration which begins with some $\bar{x}_0 = (x_0, \lambda_{k_0})$ and $\bar{B}_0 = [B_0, C_0]$ and determines

$$\begin{aligned} x_{k+1} &= x_k - B_k^{-1}\left\{F(\bar{x}_k) + C_k\left(\lambda_{k_0+k+1} - \lambda_{k_0+k}\right)\right\}, \\ \bar{x}_{k+1} &= (x_{k+1}, \lambda_{k_0+k+1}), \\ \bar{B}_{k+1} &= [B_{k+1}, C_{k+1}] \in U\left(\bar{x}_k, \bar{x}_{k+1}, \bar{B}_k\right), \end{aligned} \tag{A.1}$$

for $k = 0, 1, \cdots$. In (A.1), U is an *update function*, the values of which are subsets of $\mathbf{R}^{n\times\bar{n}}$ and which we assume is defined in a neighborhood $N = N_1 \times N_1 \times N_2$ of $(\bar{x}_*, \bar{x}_*, \bar{B}_*) \in \mathbf{R}^{\bar{n}} \times \mathbf{R}^{\bar{n}} \times \mathbf{R}^{n\times\bar{n}}$ for some $\bar{B}_*$, where $N_1 \subseteq \Omega$ and N_2 is such that the first n columns of any matrix in N_2 constitute a nonsingular matrix. Usually, but not always, we also assume that $\bar{B}_*$ is near $F'(\bar{x}_*)$ and that U satisfies the following hypothesis.

BOUNDED DETERIORATION HYPOTHESIS. *Let α_1 and α_2 be nonnegative constants such that for each $(\bar{x}, \bar{x}_+, \bar{B}) \in N$, every $\bar{B}_+ \in U(\bar{x}, \bar{x}_+, \bar{B})$ satisfies*

$$\|\bar{B}_+ - \bar{B}_*\| \le [1 + \alpha_1\sigma(\bar{x}, \bar{x}_+)^p]\,\|\bar{B} - \bar{B}_*\| + \alpha_2\sigma(\bar{x}, \bar{x}_+)^p$$

for $\sigma(\bar{x}, \bar{x}_+) = \max\{|\bar{x}-\bar{x}_*|, |\bar{x}_+-\bar{x}_*|\}$, *where* $|\bar{x}| = \max\{|x|, |\lambda|\}$ *for* $\bar{x} = (x, \lambda) \in \mathbf{R}^{\bar{n}}$.

Proposition A.1 below is a standard elementary technical result which we use in obtaining the local convergence theorems which follow.

PROPOSITION A.1. *Under the basic hypothesis, we have*

$$|F(\bar{x}) - F'(\bar{x}_*)(\bar{x} - \bar{x}_*)| \leq \frac{\gamma}{1+p}|\bar{x} - \bar{x}_*|^{1+p}$$

for $\bar{x} \in \Omega$, *where* $|\bar{x}| = \max\{|x|, |\lambda|\}$ *for* $\bar{x} = (x, \lambda) \in \mathbf{R}^{\bar{n}}$.

Proof. Setting $\bar{x}(t) = \bar{x}_* + t(\bar{x} - \bar{x}_*)$, we have

$$\begin{aligned} |F(\bar{x}) - F'(\bar{x}_*)(\bar{x} - \bar{x}_*)| &\leq \left\{ \int_0^1 |F'(\bar{x}(t)) - F'(\bar{x}_*)|\, dt \right\} |(\bar{x} - \bar{x}_*)| \\ &\leq \frac{\gamma}{1+p}|\bar{x} - \bar{x}_*|^{1+p}. \end{aligned}$$

□

Theorems A.2 and A.3 and Proposition A.4 below establish the local linear convergence of iteration (A.1) and characterize the speed of convergence of the iteration sequence. They are counterparts in our analysis of Theorems A2.1 and A3.1 of [12].

THEOREM A.2. *Let* F *satisfy the basic hypothesis and let* U *satisfy the bounded deterioration hypothesis with respect to* $\bar{B}_* = [B_*, C_*]$, *where* $B_* \in \mathbf{R}^{n\times n}$ *is an invertible matrix with*

$$|I - {B_*}^{-1}F_x(\bar{x}_*)| \leq r_x < 1.$$

Let $\{\lambda_k\}_{k=0,1,\cdots}$ *be such that one of the following holds:*

(A.2) $|\lambda_k - \lambda_*| \leq \beta r_\lambda^k$, $k = 0, 1, \cdots$, *for some* β *and* $r_\lambda \in (0, 1)$;

(A.3) $\lambda_k \to \lambda_*$ *q-linearly with* $Q_1\{\lambda_k\} = r_\lambda \in [0, 1)$.

Then for any r *such that* $\max\{r_x, r_\lambda\} < r < 1$, *there exist positive constants* ϵ_r, δ_r, *and an integer* k_r *such that if* $|x_0 - x_*| < \epsilon_r$, $|\bar{B}_0 - \bar{B}_*| < \delta_r$, *and* $k_0 \geq k_r$, *then iteration* (A.1) *is well defined for* $k = 0, 1, \cdots$, *and converges to* $\bar{x}_*$ *with*

(A.4) $|\bar{x}_k - \bar{x}_*| \leq r^k\epsilon_r$, $k = 0, 1, \cdots$, *if* (A.2) *holds, where* $|\bar{x}| = \max\{|x|, |\lambda|\}$ *for* $\bar{x} = (x, \lambda) \in \mathbf{R}^{\bar{n}}$;

(A.5) $|\bar{x}_{k+1} - \bar{x}_*| \leq r|\bar{x}_k - \bar{x}_*|$, $k = 0, 1, \cdots$, *if* (A.3) *holds, where* $|\bar{x}| = |x| + \Gamma|\lambda|$ *for* $\bar{x} = (x, \lambda) \in \mathbf{R}^{\bar{n}}$ *and for a suitable constant* Γ.

Furthermore, $\{\bar{B}_k - \bar{B}_*\}_{k=0,1,\cdots}$ *and* $\{B_k^{-1} - {B_*}^{-1}\}_{k=0,1,\cdots}$ *are uniformly small.*

Proof. Since norms on $\mathbf{R}^{n\times\bar{n}}$ are equivalent, requiring $|\cdot|$ to be small is equivalent to requiring $\|\cdot\|$ to be small. For convenience, then, we impose smallness requirements on $\|\cdot\|$ rather than on $|\cdot|$. The following may be helpful: If we are given $\bar{x} = (x, \lambda) \in \Omega$, λ_+, and $\bar{B} = [B, C]$ with B invertible, then for $x_+ = x - B^{-1}\{F(\bar{x}) + C(\lambda_+ - \lambda)\}$, we have

$$\begin{aligned} x_+ - x_* = {} & [I - B^{-1}F_x(\bar{x}_*)](x - x_*) + B^{-1}[C - F_\lambda(\bar{x}_*)](\lambda - \lambda_*) \\ & - B^{-1}C(\lambda_+ - \lambda_*) - B^{-1}[F(\bar{x}) - F'(\bar{x}_*)(\bar{x} - \bar{x}_*)]. \end{aligned} \tag{A.6}$$

We first assume that (A.2) holds and establish all conclusions except (A.5). Take $|\bar{x}| = \max\{|x|, |\lambda|\}$ for $\bar{x} = (x, \lambda) \in \mathbf{R}^{\bar{n}}$. Choose positive ϵ_r, δ and an integer k_r such that $(\alpha_1\delta + \alpha_2)\epsilon_r{}^p/(1 - r^p) < \delta$, $\beta r_\lambda^{k_r} \leq \epsilon_r$, and if $\|\bar{B} - \bar{B}_*\| \leq \delta$ and $k \geq k_r$, then B^{-1} exists and

$$|I - B^{-1}F_x(\bar{x}_*)|\epsilon_r + \{|B^{-1}[C - F_\lambda(\bar{x}_*)]| + r_\lambda |B^{-1}C|\}\beta r_\lambda^k + |B^{-1}|\frac{\gamma}{1+p}\epsilon_r{}^{1+p} \leq r\epsilon_r.$$

If necessary, further restrict ϵ_r, δ, and k_r so that if $|x - x_*| < \epsilon_r$, $\|\bar{B} - \bar{B}_*\| \leq \delta$, and $k \geq k_r$, then $(\bar{x}, \bar{x}_+, \bar{B}) \in N$ for $\bar{x} = (x, \lambda_k)$ and

$$\bar{x}_+ = (x - B^{-1}\{F(\bar{x}) + C(\lambda_{k+1} - \lambda_k)\}, \lambda_{k+1}).$$

Choose $\delta_r > 0$ such that $\delta_r + (\alpha_1\delta + \alpha_2)\epsilon_r{}^p/(1 - r^p) < \delta$.

Let $|x_0 - x_*| < \epsilon_r$, $\|B_0 - B_*\| < \delta_r$, and $k_0 \geq k_r$. We have from (A.6) and the other assumptions above and from Proposition A.1 that

$$\begin{aligned} |x_1 - x_*| &\leq \left|I - B_0^{-1}F_x(\bar{x}_*)\right| \epsilon_r + \left\{\left|B_0^{-1}[C_0 - F_\lambda(\bar{x}_*)]\right| + r_\lambda \left|B_0^{-1}C_0\right|\right\} \beta r_\lambda^{k_r} \\ &\quad + \left|B_0^{-1}\right| \frac{\gamma}{1+p} \epsilon_r{}^{1+p} \\ &\leq r\epsilon_r. \end{aligned}$$

Then $|\bar{x}_1 - \bar{x}_*| \leq r\epsilon_r$.

As an inductive hypothesis, assume that for some $k > 0$ and for $j = 0, \cdots, k-1$, $\|\bar{B}_j - \bar{B}_*\| < \delta$ and $|\bar{x}_{j+1} - \bar{x}_*| \leq r^{j+1}\epsilon_r$. Then for $j = 0, \cdots, k-1$, we see that $\sigma(\bar{x}_j, \bar{x}_{j+1}) \leq r^j\epsilon_r$ and by the bounded deterioration hypothesis

$$\|\bar{B}_{j+1} - \bar{B}_*\| - \|\bar{B}_j - \bar{B}_*\| \leq \alpha_1\delta\epsilon_r{}^p r^{pj} + \alpha_2\epsilon_r{}^p r^{pj}.$$

Summing over $j = 0, \cdots, k-1$ gives

$$\|\bar{B}_k - \bar{B}_*\| \leq \|\bar{B}_0 - \bar{B}_*\| + (\alpha_1\delta + \alpha_2)\frac{\epsilon_r{}^p}{1 - r^p} < \delta. \tag{A.7}$$

Also, as in the case $k = 1$,

$$\begin{aligned} |x_{k+1} - x_*| \leq & \left|I - B_k^{-1}F_x(\bar{x}_*)\right| r^k\epsilon_r \\ & + \left\{\left|B_k^{-1}[C_k - F_\lambda(\bar{x}_*)]\right| + r_\lambda \left|B_k^{-1}C_k\right|\right\} \beta r_\lambda^{k_r+k} \\ & + \left|B_k^{-1}\right| \frac{\gamma}{1+p} (r^k\epsilon_r)^{1+p} \\ \leq & r^{k+1}\epsilon_r. \end{aligned}$$

Then $|\bar{x}_{k+1} - \bar{x}_*| \leq r^{k+1}\epsilon_r$, and the induction is complete. It follows from (A.7) and the assumptions on δ that $\{\bar{B}_k - \bar{B}_*\}_{k=0,1,\cdots}$ and $\{B_k^{-1} - B_*{}^{-1}\}_{k=0,1,\cdots}$ are uniformly small, and the theorem is proved in this case.

We now assume that (A.3) holds and establish (A.5). We assume that ϵ_r, δ_r, and k_r have been chosen such that if $|x_0 - x_*| < \epsilon_r$, $\|\bar{B}_0 - \bar{B}_*\| < \delta_r$, and $k_0 \geq k_r$, then iteration (A.1) is well defined for $k = 0, 1, \cdots$, (A.4) holds, and $\{\bar{B}_k - \bar{B}_*\}_{k=0,1,\cdots}$ and $\{B_k^{-1} - B_*{}^{-1}\}_{k=0,1,\cdots}$ are uniformly small. This is allowed, for if (A.3) holds, then (A.2) holds for an appropriate β with r_λ replaced by anything larger. Choose r', r'' such that $\max\{r_x, r_\lambda\} < r'' < r' < r$. If necessary, further restrict ϵ_r, δ_r, and k_r so that if $|x_0 - x_*| < \epsilon_r$, $\|\bar{B}_0 - \bar{B}_*\| < \delta_r$, and $k_0 \geq k_r$, and if $\{\bar{x}_k, \bar{B}_k\}_{k=0,1,\cdots}$ is determined by (A.1), then $|\lambda_{k_0+k+1} - \lambda_*| \leq r''|\lambda_{k_0+k} - \lambda_*|$, $|I - B_k^{-1}F_x(\bar{x}_*)| \leq r'$, and $|B_k^{-1}|(\gamma/(1+p))\max\{|x_k - x_*|, |\lambda_{k_0+k} - \lambda_*|\}^p \leq r - r'$ for $k = 0, 1, \cdots$. Choose $\Gamma \geq 1$ so large that if $|x_0 - x_*| < \epsilon_r$, $\|\bar{B}_0 - \bar{B}_*\| < \delta_r$, and $k_0 \geq k_r$ and if $\{\bar{x}_k, \bar{B}_k\}_{k=0,1,\cdots}$ is determined by (A.1), then $\left|B_k^{-1}[C_k - F_\lambda(\bar{x}_*)]\right| + \left(\left|B_k^{-1}C_k\right| + \Gamma\right) r'' \leq \Gamma r'$ for $k = 0, 1, \cdots$. Take $|\bar{x}| = |x| + \Gamma|\lambda|$ for $\bar{x} = (x, \lambda) \in \mathbf{R}^{\bar{n}}$.

Suppose that $|x_0 - x_*| < \epsilon_r$, $\|\bar{B}_0 - \bar{B}_*\| < \delta_r$, and $k_0 \geq k_r$, and let $\{\bar{x}_k, \bar{B}_k\}_{k=0,1,\cdots}$ be determined by (A.1). We have from (A.6), the other assumptions, and Proposition A.1 that

$$\begin{aligned}|x_{k+1} - x_*| \leq & \left|I - B_k^{-1}F_x(\bar{x}_*)\right| |x_k - x_*| \\ & + \left\{\left|B_k^{-1}\left[C_k - F_\lambda(\bar{x}_*)\right]\right| + r''\left|B_k^{-1}C_k\right|\right\} |\lambda_{k_0+k} - \lambda_*| \\ & + \left|B_k^{-1}\right| \frac{\gamma}{1+p} \max\{|x_k - x_*|, |\lambda_{k_0+k} - \lambda_*|\}^p |\bar{x}_k - \bar{x}_*|,\end{aligned}$$

and so

$$\begin{aligned}|\bar{x}_{k+1} - \bar{x}_*| &\leq r'|x_k - x_*| + \left\{\left|B_k^{-1}\left[C_k - F_\lambda(\bar{x}_*)\right]\right| + r''\left|B_k^{-1}C_k\right| + r''\Gamma\right\} \\ &\qquad \cdot |\lambda_{k_0+k} - \lambda_*| + (r - r')|\bar{x}_k - \bar{x}_*| \\ &\leq r|\bar{x}_k - \bar{x}_*|.\end{aligned}$$

This completes the proof. □

THEOREM A.3. *Suppose that F satisfies the basic hypothesis and that $\{\bar{x}_k\}_{k=0,1,\cdots}$ is a sequence generated by* (A.1) *which converges q-linearly to $\bar{x}_*$ in some norm with $\bar{s}_k = \bar{x}_{k+1} - \bar{x}_k \neq 0$ for all but finitely many k. If $\bar{B}_* = [B_*, C_*]$ is any matrix such that $B_* \in \mathbf{R}^{n\times n}$ is invertible, then the norm-independent condition*

$$\lim_{k\to\infty} \frac{\left|(\bar{B}_k - \bar{B}_*)\,\bar{s}_k\right|}{|\bar{s}_k|} = 0 \tag{A.8}$$

holds if and only if the norm-independent condition

$$\begin{aligned}\lim_{k\to\infty} \Big| & \left[I - {B_*}^{-1}F_x(\bar{x}_*)\right](x_k - x_*) - (x_{k+1} - x_*) - {B_*}^{-1}C_*(\lambda_{k_0+k+1} - \lambda_*) \\ & + {B_*}^{-1}\left[C_* - F_\lambda(\bar{x}_*)\right](\lambda_{k_0+k} - \lambda_*)\Big| \Big/ |\bar{x}_k - \bar{x}_*| = 0\end{aligned} \tag{A.9}$$

holds.

Proof. We have

$$\begin{aligned}(\bar{B}_k - \bar{B}_*)\bar{s}_k = B_*\big\{ & \left[I - {B_*}^{-1}F_x(\bar{x}_*)\right](x_k - x_*) - (x_{k+1} - x_*) \\ & - {B_*}^{-1}C_*(\lambda_{k_0+k+1} - \lambda_*) \\ & + {B_*}^{-1}\left[C_* - F_\lambda(\bar{x}_*)\right](\lambda_{k_0+k} - \lambda_*)\big\} \\ & + F'(\bar{x}_*)(\bar{x}_k - \bar{x}_*) - F(\bar{x}_k)\end{aligned} \tag{A.10}$$

Since $\bar{x}_k \to \bar{x}_*$ q-linearly in some norm, it follows from the equivalence of norms on $\mathbf{R}^{\bar{n}}$ that for any norm $|\cdot|$ on $\mathbf{R}^{\bar{n}}$ there are constants η_1 and η_2 for which

$$\eta_1|\bar{x}_k - \bar{x}_*| \leq |\bar{s}_k| \leq \eta_2|\bar{x}_k - \bar{x}_*|. \tag{A.11}$$

Also, we see from Proposition A.1 that

$$F'(\bar{x}_*)(\bar{x}_k - \bar{x}_*) - F(\bar{x}_k) = o(|\bar{s}_k|) \tag{A.12}$$

in any norm. Since B_* is invertible, it follows immediately from (A.10)–(A.12), and norm equivalence that (A.8) holds in some norms on $\mathbf{R}^n$ and $\mathbf{R}^{\bar{n}}$ if and only if (A.9) holds in some norms on $\mathbf{R}^n$ and $\mathbf{R}^{\bar{n}}$. □

The conditions of Theorem A.3 have particular consequences which are outlined in Proposition A.4 and the remarks below.

PROPOSITION A.4. *If* (A.9) *holds in some norms on* $\mathbf{R}^n$ *and* $\mathbf{R}^{\bar{n}}$ *and if* $r_x = |I - B_*^{-1}F_x(\bar{x}_*)|$ *and* $r_\lambda = Q_1\{\lambda_k\}$, *then the following hold:*

(A.13) *for any* $\epsilon > 0$, $\overline{\lim}_{k\to\infty} |\bar{x}_{k+1} - \bar{x}_*|/|\bar{x}_k - \bar{x}_*| \le \max\{r_x, r_\lambda + \epsilon\}$, *where* $|\bar{x}| = |x| + \Gamma|\lambda|$ *for* $\bar{x} = (x, \lambda) \in \mathbf{R}^{\bar{n}}$ *and for a suitable constant* Γ *depending on* ϵ;

(A.14) *if* $r_\lambda = 0$, *then* $\bar{x}_k \to \bar{x}_*$ *q-superlinearly if and only if the norm-independent condition*

$$\overline{\lim_{k\to\infty}} \frac{\left|\left[I - B_*^{-1}F_x(\bar{x}_*)\right](x_k - x_*) + B_*^{-1}\left[C_* - F_\lambda(\bar{x}_*)\right](\lambda_{k_0+k} - \lambda_*)\right|}{|\bar{x}_k - \bar{x}_*|} = 0$$

holds.

In particular, $\bar{x}_k \to \bar{x}_*$ *q-superlinearly if* $\lambda_k \to \lambda_*$ *q-superlinearly and* $\bar{B}_* = F'(\bar{x}_*)$.

Proof. If (A.9) holds in some norms, then

$$\overline{\lim_{k\to\infty}} \frac{|x_{k+1} - x_*|}{|\bar{x}_k - \bar{x}_*|} \le \overline{\lim_{k\to\infty}} \left\{r_x|x_k - x_*| + \left|B_*^{-1}C_*\right| |\lambda_{k_0+k+1} - \lambda_*| + \left|B_*^{-1}\left[C_* - F_\lambda(\bar{x}_*)\right]\right| |\lambda_{k_0+k} - \lambda_*|\right\} \Big/ |\bar{x}_k - \bar{x}_*|$$

for any norms on $\mathbf{R}^n$, $\mathbf{R}^m$, and $\mathbf{R}^{\bar{n}}$. Consequently,

(A.15)

$$\overline{\lim_{k\to\infty}} \frac{|x_{k+1} - x_*| + \Gamma|\lambda_{k_0+k+1} - \lambda_*|}{|\bar{x}_k - \bar{x}_*|} \le \overline{\lim_{k\to\infty}} \frac{r_x|x_k - x_*| + \left\{\left[\left|B_*^{-1}C_*\right| + \Gamma\right] r_\lambda + \left|B_*^{-1}\left[C_* - F_\lambda(\bar{x}_*)\right]\right|\right\} |\lambda_{k_0+k} - \lambda_*|}{|\bar{x}_k - \bar{x}_*|}$$

for any Γ and any norms on $\mathbf{R}^n$, $\mathbf{R}^m$, and $\mathbf{R}^{\bar{n}}$. From (A.15), we see that (A.13) holds by taking $|\bar{x}| = |x| + \Gamma|\lambda|$ for $\bar{x} = (x, \lambda) \in \mathbf{R}^{\bar{n}}$, where Γ is sufficiently large that $[|B_*^{-1}C_*| + \Gamma]r_\lambda + |B_*^{-1}[C_* - F_\lambda(\bar{x}_*)]| \le \Gamma(r_\lambda + \epsilon)$. If $r_\lambda = 0$, then $\overline{\lim}_{k\to\infty} |\lambda_{k_0+k+1} - \lambda_*|/|\bar{x}_k - \bar{x}_*| = 0$ for any norms on $\mathbf{R}^m$ and $\mathbf{R}^{\bar{n}}$, and it follows from (A.9) that

$$\overline{\lim_{k\to\infty}} \frac{|x_{k+1} - x_*| + \Gamma|\lambda_{k_0+k+1} - \lambda_*|}{|\bar{x}_k - \bar{x}_*|} = \overline{\lim_{k\to\infty}} \frac{\left|\left[I - B_*^{-1}F_x(\bar{x}_*)\right](x_k - x_*) + B_*^{-1}\left[C_* - F_\lambda(\bar{x}_*)\right](\lambda_{k_0+k} - \lambda_*)\right|}{|\bar{x}_k - \bar{x}_*|}$$

for any positive Γ. Then we obtain (A.14) by taking $|\bar{x}| = |x| + \Gamma|\lambda|$ for $\bar{x} = (x, \lambda) \in \mathbf{R}^{\bar{n}}$. □

We now consider an inverse update analogue of iteration (A.1), viz., an iteration which begins with some $\bar{x}_0 = (x_0, \lambda_{k_0})$ and $\bar{K}_0 = [K_0, L_0]$ and determines

$$\begin{aligned} x_{k+1} &= x_k - K_k F(\bar{x}_k) + L_k(\lambda_{k_0+k+1} - \lambda_{k_0+k}), \\ \bar{x}_{k+1} &= (x_{k+1}, \lambda_{k_0+k+1}), \\ \bar{K}_{k+1} &= [K_{k+1}, L_{k+1}] \in U(\bar{x}_k, \bar{x}_{k+1}, \bar{K}_k), \end{aligned} \tag{A.16}$$

for $k = 0, 1, \cdots$, where we assume the update function U is defined in a neighborhood $N = N_1 \times N_1 \times N_2$ of $(\bar{x}_*, \bar{x}_*, \bar{K}_*) \in \mathbf{R}^{\bar{n}} \times \mathbf{R}^{\bar{n}} \times \mathbf{R}^{n \times \bar{n}}$ for some $\bar{K}_*$, where $N_1 \subseteq \Omega$ and N_2 is such that the first n columns of any matrix in N_2 constitute a nonsingular matrix. The appropriate analogue of the bounded deterioration hypothesis is the following hypothesis.

INVERSE BOUNDED DETERIORATION HYPOTHESIS. *Let α_1 and α_2 be nonnegative constants such that for each $(\bar{x}, \bar{x}_+, \bar{K}) \in N$, every $\bar{K}_+ \in U(\bar{x}, \bar{x}_+, \bar{K})$ satisfies*

$$\|\bar{K}_+ - \bar{K}_*\| \leq [1 + \alpha_1 \sigma(\bar{x}, \bar{x}_+)^p] \, \|\bar{K} - \bar{K}_*\| + \alpha_2 \sigma(\bar{x}, \bar{x}_+)^p$$

for $\sigma(\bar{x}, \bar{x}_+) = \max\{|\bar{x} - \bar{x}_*|, |\bar{x}_+ - \bar{x}_*|\}$, where $|\bar{x}| = \max\{|x|, |\lambda|\}$ for $\bar{x} = (x, \lambda) \in \mathbf{R}^{\bar{n}}$.

Theorems A.5 and A.6 below are counterparts for iteration (A.16) of Theorems A.2 and A.3 for iteration (A.1). They are analogous to Theorems A2.2 and A3.2 of [12]. The proof of Theorem A.5 is very similar to the proof of Theorem A.2, and so we omit it.

THEOREM A.5. *Let F satisfy the basic hypothesis and let U satisfy the inverse bounded deterioration hypothesis with respect to $\bar{K}_* = [K_*, L_*]$, where $K_* \in \mathbf{R}^{n \times n}$ is an invertible matrix with*

$$|I - K_* F_x(\bar{x}_*)| \leq r_x < 1.$$

Let $\{\lambda_k\}_{k=0,1,\ldots}$ be such that either (A.2) *or* (A.3) *holds for some r_λ. Then for any r such that $\max\{r_x, r_\lambda\} < r < 1$, there exist positive constants ϵ_r, δ_r and an integer k_r such that if $|x_0 - x_*| < \epsilon_r$, $|\bar{K}_0 - \bar{K}_*| < \delta_r$, and $k_0 \geq k_r$, then iteration* (A.16) *is well defined for $k = 0, 1, \cdots$, and converges to $\bar{x}_*$ according to either* (A.4), *if* (A.2) *holds, or* (A.5), *if* (A.3) *holds. Furthermore, $\{\bar{K}_k - \bar{K}_*\}_{k=0,1,\ldots}$ and $\{K_k^{-1} - K_*^{-1}\}_{k=0,1,\ldots}$ are uniformly small.*

THEOREM A.6. *Suppose that F satisfies the basic hypothesis and that $\{\bar{x}_k\}_{k=0,1,\ldots}$ is a sequence generated by* (A.16) *which converges q-linearly to $\bar{x}_*$ in some norm with $\bar{s}_k = \bar{x}_{k+1} - \bar{x}_k \neq 0$ for all but finitely many k and with $\{\bar{K}_k\}_{k=0,1,\ldots}$ and $\{K_k^{-1}\}_{k=0,1,\ldots}$ uniformly bounded. Let $\bar{K}_* = [K_*, L_*]$ be any matrix such that $K_* \in \mathbf{R}^{n \times n}$ is invertible, and suppose that $\{y_k\}_{k=0,1,\ldots}$ is a sequence satisfying the norm-independent condition*

$$|\bar{K}_* \bar{y}_k - s_k| \leq \alpha_k |\bar{s}_k|, \qquad k = 0, 1, \cdots \tag{A.17}$$

where $\bar{y}_k = (y_k, \lambda_{k_0+k+1} - \lambda_{k_0+k})$, $s_k = x_{k+1} - x_k$, and $\lim_{k\to\infty} \alpha_k = 0$. Then the norm-independent condition

$$\lim_{k\to\infty} \frac{|(\bar{K}_k - \bar{K}_*)\, \bar{y}_k|}{|\bar{y}_k|} = 0 \tag{A.18}$$

holds if and only if the norm-independent condition

$$\lim_{k\to\infty} \Big| [I - K_* F_x(\bar{x}_*)](x_k - x_*) - (x_{k+1} - x_*) + L_*(\lambda_{k_0+k+1} - \lambda_*) - [L_* + K_* F_\lambda(\bar{x}_*)]\,(\lambda_{k_0+k} - \lambda_*) \Big| \Big/ |\bar{x}_k - \bar{x}_*| = 0 \tag{A.19}$$

holds.

Remark. Clearly, (A.19) is the same as (A.9) with $B_* = K_*^{-1}$ and $C_* = -K_*^{-1} L_*$. Consequently if (A.18) and (A.19) hold, then the conclusions of Proposition A.4 hold with $B_* = K_*^{-1}$ and $C_* = -K_*^{-1} L_*$.

Proof. Since (A.19) is the same as (A.9) with $B_* = K_*^{-1}$ and $C_* = -K_*^{-1}L_*$, it suffices to show (A.18) holds if and only if (A.8) holds with $\bar{B}_k = [B_k, C_k]$, $B_k = K_k^{-1}$ and $C_k = -K_k^{-1}L_k$. We have

$$\left(\bar{B}_k - \bar{B}_*\right)\bar{s}_k = \left(B_* - B_k\right)\left(\bar{K}_*\bar{y}_k - s_k\right) - B_k\left(\bar{K}_k - \bar{K}_*\right)\bar{y}_k, \tag{A.20}$$

and it follows from (A.17), (A.20), and the boundedness of $\{\bar{K}_k\}_{k=0,1,\cdots}$ and $\{K_k^{-1}\}_{k=0,1,\cdots}$ that (A.8) holds if and only if

$$\lim_{k\to\infty} \frac{\left|\left(\bar{K}_k - \bar{K}_*\right)\bar{y}_k\right|}{|\bar{s}_k|} = 0 \tag{A.21}$$

holds. But (A.17) implies that there are positive constants η_1 and η_2 for which $\eta_1|\bar{y}_k| \leq |\bar{s}_k| \leq \eta_2|\bar{y}_k|$, $k = 0, 1, \cdots$. It follows that (A.21) is equivalent to (A.18), and the proof is complete. □

Acknowledgments. The authors thank Layne T. Watson for participating in many discussions which were influential in shaping the contents of this paper. We also thank the referees for their thoughtful reviews.

REFERENCES

[1] P. ALFELD, *Two devices for improving the efficiency of stiff* ODE *solvers*, in Proc. 1979 SIGNUM Meeting on Numerical Ordinary Differential Equations, pp. 24-1–24-3, Report 79-1710, Computer Science Dept., University of Illinois, Urbana, IL, 1979.

[2] C. A. BEATTIE AND S. WEAVER-SMITH, *Secant methods for structural model identification*, Report ICAM-TR-88-0601, Interdisciplinary Center for Applied Mathematics, Virginia Polytechnic Institute and State University, Blacksburg, VA, 1988.

[3] S. K. BOURJI, Ph.D. thesis, Utah State University, Logan, UT, 1987.

[4] P. N. BROWN, A. C. HINDMARSH, AND H. F. WALKER, *Experiments with quasi-Newton methods in solving stiff* ODE systems, SIAM J. Sci. Statist. Comput., 6 (1985), pp. 297–313.

[5] C. G. BROYDEN, *A class of methods for solving nonlinear simultaneous equations*, Math. Comp., 19 (1965), pp. 577–593.

[6] ———, *A new double-rank minimization algorithm*, AMS Notices, 16 (1969), p. 670.

[7] ———, *The convergence of a class of double-rank minimization algorithms, Parts* I *and* II, J. Inst. Math. Appl., 6 (1971), pp. 76–90, 222–236.

[8] ———, *The convergence of an algorithm for solving sparse nonlinear systems*, Math. Comp., 25 (1971), pp. 285–294.

[9] W. C. DAVIDON, *Variable metric methods for minimization*, Report ANL-5990 Rev., Argonne National Laboratory, Argonne, IL, November, 1959.

[10] J. E. DENNIS, JR., AND R. B. SCHNABEL, *Least change secant updates for quasi-Newton methods*, SIAM Rev., 21 (1980), pp. 443–459.

[11] ———, *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*, Prentice-Hall Series in Automatic Computation, Englewood Cliffs, NJ, 1983.

[12] J. E. DENNIS, JR., AND H. F. WALKER, *Convergence theorems for least change secant update methods*, SIAM J. Numer. Anal., 18 (1981), pp. 949–987.

[13] R. FLETCHER, *A new approach to variable metric algorithms*, Comput. J., 13 (1970), pp. 317–322.

[14] R. FLETCHER AND M. J. D. POWELL, *A rapidly convergent descent method for minimization*, Comput. J., 6 (1963), pp. 163–168.

[15] G. E. FORSYTHE, M. A. MALCOLM, AND C. B. MOLER, *Computer Methods for Mathematical Computations*, Prentice-Hall Series in Automatic Computation, Englewood Cliffs, NJ, 1977.

[16] K. GEORG, *On tracing an implicitly defined curve by quasi-Newton steps and calculating bifurcation by local perturbations*, SIAM J. Sci. Statist. Comput., 2 (1981), pp. 35–50.

[17] D. GOLDFARB, *A family of variable-metric methods derived by variational means*, Math. Comp., 24 (1970), pp. 23–26.

[18] J. GREENSTADT, *Variations on variable metric methods*, Math. Comp., 24 (1970), pp. 1–18.

[19] A. GRIEWANK AND L. SHENG, *On the Gauss-Broyden method for nonlinear least-squares*, Preprint MCS-PS-0988, Mathematics and Computer Science Div., Argonne National Laboratory, Argonne, IL, September, 1988.

[20] J. M. MARTÍNEZ, *Quasi-Newton methods for solving underdetermined nonlinear simultaneous equations*, Relatório Técnico 21/88, Instituto de Matemática, Estatística, e Ciência da Computação, Universidade Estudual de Campinas, Campinas, Brasil, November, 1988.

[21] E. S. MARWIL, *Exploiting sparsity in Newton-like methods*, Ph.D. thesis, Cornell University, Ithaca, NY, 1978.

[22] J. J. MORÉ, B. S. GARBOW, AND K. E. HILLSTROM, *User Guide for* MINPACK-1, Report ANL-80-74, Applied Mathematics Div., Argonne National Laboratory, Argonne, IL, August, 1980.

[23] ———, *Testing unconstrained optimization software*, ACM Trans. Math. Software, 7 (1981), pp. 17–41.

[24] A. P. MORGAN, *Solving Polynomial Systems Using Continuation for Engineering and Scientific Problems*, Prentice-Hall, Englewood Cliffs, NJ, 1987.

[25] J. M. ORTEGA AND W. C. RHEINBOLDT, *Iterative Solution of Nonlinear Equations in Several Variables*, Academic Press, New York, 1970.

[26] M. J. D. POWELL, *A new algorithm for unconstrained optimization*, in Nonlinear Programming, J. B. Rosen, O. L. Mangasarian, and K. Ritter, eds., Academic Press, New York, 1970.

[27] L. K. SCHUBERT, *Modification of a quasi-Newton method for nonlinear equations with a sparse Jacobian*, Math. Comp., 24 (1978), pp. 27–30.

[28] L. F. SHAMPINE, H. A. WATTS, AND S. M. DAVENPORT, *Solving nonstiff ordinary differential equations — the state of the art*, SIAM Rev., 18 (1976), pp. 376–411.

[29] D. F. SHANNO, *Conditioning of quasi-Newton methods for function minimization*, Math. Comp., 24 (1970), pp. 647–656.

[30] PH. L. TOINT, *On sparse and symmetric matrix updating subject to a linear equation*, Math. Comp., 31 (1977), pp. 954–961.

[31] H. F. WALKER AND L. T. WATSON, *Least-change secant update methods for underdetermined systems*, Res. Report August/88/41, Mathematics and Statistics Dept., Utah State University, Report 88-28, Computer Science Dept., and Report 88-09-03, Interdisciplinary Center for Applied Mathematics, Virginia Polytechnic Institute and State University, August, 1988.

[32] L. T. WATSON, S. C. BILLUPS, AND A. P. MORGAN, HOMPACK: a suite of codes for globally convergent homotopy algorithms, ACM Trans. Math. Software, 13 (1987), pp. 281–310.

[33] L. T. WATSON AND C. Y. WANG, *A homotopy method applied to elastica problems*, Internat. J. Solids and Structures, 17 (1981), pp. 29–37.

[34] T. YPMA, private communication, 1988.