Least-Change Secant Update Methods
With Inaccurate Secant Conditions

by

J.E. Dennis, Jr.\textsuperscript{1}

and

Homer F. Walker\textsuperscript{2}

Technical Report 83-26, November 1983

\textsuperscript{1}Mathematical Sciences Department, Rice University, Houston, Texas 77251. Research sponsored by DOE DE-AS05-82ER13016, ARO DAAG-79-C-0124, NSF MCS81-16779. This work was supported in part by the International Business Machine Corporation, Palo Alto Scientific Center, Palo Alto, CA.

\textsuperscript{2}Department of Mathematics, University of Houston, Houston, Texas 77004. Research sponsored by DOE DE-AS05-82ER13016.
Abstract

In this paper, we investigate the role of the secant or quasi-Newton condition in the sparse Broyden or Schubert update method for solving systems of nonlinear equations whose Jacobians are either sparse, or can be approximated acceptably by conveniently sparse matrices. We develop a general theory on perturbations to the secant equation that will still allow a proof of local $q$-linear convergence. To illustrate the theory, we show how to generalize the standard secant condition to the case when the function difference is contaminated by noise.

Key words

quasi-Newton methods, local convergence, sparse nonlinear equations, bounded deterioration, least-change secant methods, Schubert's method, Broyden's method.
1. Introduction.

In earlier work (Dennis and Walker (1983)), we addressed the effects of inaccuracy on the performance of quasi-Newton methods for solving nonlinear algebraic equations. There, the emphasis was on generality: The methods considered were general bounded-deterioration quasi-Newton methods; and we took into account inaccuracy from all sources, including not only finite-precision computer arithmetic but also the differences between an ideal problem, its mathematical model, and the computer implementation of the model. The objective was to determine rates of improvement and limiting accuracies that can be obtained near solutions in the presence of such general inaccuracy.

Here, our interest is again in inaccuracy in quasi-Newton methods; however, the focus is more specific than before. The methods considered are least-change secant update methods (see Dennis and Schnabel (1979)), in particular methods employing the sparse secant updates of Broyden (1971) and Schubert (1970), which include of course the usual Broyden (1965) update. The inaccuracy with which we are concerned is that which resides in the secant conditions by which updates are determined, and we explicitly exclude inaccuracy arising from other sources from consideration here. Our objective is a local convergence analysis which extends that of Dennis and Walker (1981) in the sparse secant update case by relaxing the requirements on secant conditions in the theorems of that paper.

To be more precise, let us specify that the problem of interest is the following:
**Problem:** Given $F: \Omega \subseteq \mathbb{R}^n \to \mathbb{R}^n$, find $x_* \in \Omega$ satisfying

$$F(x_*) = 0. \quad (1.1)$$

By a quasi-Newton method for approximating a solution of (1.1), we mean any iterative method of the general form

$$x_{k+1} = x_k - B_k^{-1} F(x_k), \quad (1.2)$$

in which $B_k$ is regarded as an approximation of the Jacobian matrix $F'(x_k)$. Since our concern here is with local behavior, we consider only iterations in which the full step $-B_k^{-1} F(x_k)$ is appropriate. Each $B_k$ is assumed to have the form

$$B_k = A_k + C(x_k), \quad (1.3)$$

in which $C(x_k)$ is a "computed part" of $F'(x_k)$ determined by a function $C: \Omega \subseteq \mathbb{R}^n \to \mathbb{R}^{n \times n}$ and $A_k$ is an "approximated part" of $F'(x_k)$ maintained by updating.

Since the methods with which we are concerned here are modeled after Newton's method, we assume the following throughout the sequel:

**The standard hypothesis on $F$ and $C$:** Let $F$ be differentiable in an open convex neighborhood $\Omega$ of a point $x_* \in \mathbb{R}^n$ for which $F(x_*) = 0$ and $F'(x_*)$ is nonsingular. Let $\gamma \geq 0$, $\gamma_c \geq 0$, and $\rho \in (0,1]$ be such that for $x \in \Omega$,

$$|F'(x) - F'(x_*)| \leq \gamma |x - x_*|^\rho$$

and

$$|C(x) - C(x_*)| \leq \gamma_c |x - x_*|^\rho.$$
where $|\cdot|$ denotes a norm on $\mathbb{R}^n$ and its subordinate operator norm on $\mathbb{R}^{n \times n}$. The part of this standard hypothesis which applies to $F$ is sufficient to insure that sequences of iterates produced by Newton's method converge locally to $x_*$ with $q$-order $(1 + p)$. See, for example, Dennis and Schnabel (1983) or Ortega and Rheinboldt (1970).

A very successful way of maintaining the $A_k$'s in (1.3) is through least-change secant updates (see Dennis and Schnabel (1979) and Dennis and Walker (1981)). An extensive review of these updates is not in order here; but let us recall that if one is given an inner-product norm $||\cdot||$ on $\mathbb{R}^{n \times n}$ and an affine subspace $\mathbf{A} \subseteq \mathbb{R}^{n \times n}$ which reflects the structure of $[F'(x) - C(x)]$, then $A_{k+1}$ is uniquely determined as a least-change secant update of $A_k$ in $\mathbf{A}$ by a choice of vectors $s_k \neq 0$ and $y_k$ in $\mathbb{R}^n$. Furthermore, one has that

$$A_{k+1}s_k = y_k,$$  

(1.4)

if it is at all possible for this equation to be satisfied by a matrix in $\mathbf{A}$. The purpose of updating $A_k$ is to incorporate in $A_{k+1}$ currently available information about $[F'(x_{k+1}) - C(x_{k+1})]$, and so one usually chooses $s_k = x_{k+1} - x_k$ and $y_k \approx [F'(x_{k+1}) - C(x_{k+1})]s_k$. In fact, in most classical applications, $C(x) = 0$ and one takes $s_k = x_{k+1} - x_k$ and $y_k = F(x_{k+1}) - F(x_k)$. In this case, (1.4) is satisfied if $F'(x) \in \mathbf{A}$ for all $x \in \mathbb{R}^n$.

In light of the usual choice of $s_k$ and $y_k$, we call (1.4) a secant condition on $A_{k+1}$. In the literature, the classical case in which $C(x) = 0$, $s_k = x_{k+1} - x_k$, and $y_k = F(x_{k+1}) - F(x_k)$ is also at times called the quasi-Newton equation. From here on, we assume that $s_k = x_{k+1} - x_k$; and so if $\mathbf{A}$ and a suitable norm on $\mathbb{R}^{n \times n}$ are given, then a choice of $y_k$...
alone determines $A_{k+1}$ as a least-change secant update of $A_k$ in $A$. We shall regard choosing $y_k$ as equivalent to specifying a secant condition (1.4) on $A_{k+1}$, even though it may not always be possible for (1.4) to be satisfied by a matrix in $A$. For convenience, we shall also occasionally refer to $y_k$ itself as a secant condition.

In Dennis and Walker (1981), we established very general requirements on $A$ and the secant conditions (1.4) which guarantee that the resulting least-change secant updates lead to a quasi-Newton method with local q-linear asymptotic convergence which is optimal for a sequence $\{A_k\}$ in $A$. In classical cases in which $[F'(x_*) - C(x_*)] \in A$, these requirements give the standard local and q-superlinear convergence results; and a major feature of the theorems of Dennis and Walker (1981) is that they allow one to relax the condition that $[F'(x_*) - C(x_*)] \in A$. Our requirements are phrased in terms of a choice rule for determining allowable $y_k$'s for each $k$. Such a choice rule is given by a function $\mathbf{x}: \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}^n$; for each $k$, one chooses any $y_k \in \mathbf{x}(x_k, x_{k+1})$, provided $x_k$ and $x_{k+1}$ are in $A$. The utility of prescribing a choice rule in this way is evident when one considers applications in which a multiplicity of workable $y_k$'s naturally present themselves (see the discussion in Dennis and Walker (1981)). In classical cases, the manner of determining an appropriate choice rule is clear; and it is very easy to verify that our requirements are satisfied, at least ideally.

In this report, our concern is with secant conditions which are inaccurate in the sense that they cannot be associated with a choice rule that satisfies the requirements of Dennis and Walker (1981). In the following, we offer a local convergence analysis which extends that of
Dennis and Walker (1981) by relaxing the requirements that secant conditions must satisfy in order to lead to least-change secant update methods which are assured of having desirable local convergence properties. Our analysis is restricted to the case in which the updates are sparse secant updates; the investigation of other cases is left to future work. We are motivated by the fact that in spite of the generality of the requirements of Dennis and Walker (1981), there are circumstances in which secant conditions satisfying them must be regarded as unobservable in practice. We cite two such circumstances below which we feel are particularly important.

The first circumstance is that in which the computed values of \( F \) or \( C \) contain significant inaccuracy. In this case, any choice of \( y_k \) which is determined by computed function values will also exhibit inaccuracy; and so it seems essentially certain that any practical implementation of a choice rule which specifies \( y_k \) in this way will fail to satisfy the conditions of Dennis and Walker (1981), even though the "ideal" choice rule which uses accurate function values may be completely satisfactory. This case is important in practice (see Barrera and Dennis (1979) and Moré, Garbow, and Hillstrom (1980)), and we use it for illustration in the sequel. The analysis given here provides reassurance that in most classical applications, the traditional secant conditions should usually be admissible in practice, even though they are determined by inaccurate computed function values. This reassurance is in the form of a rigorously justified, simple generalization (4.1) of the traditional secant condition in the noisy case. Also, we feel that this analysis might prove to be particularly relevant in problems in which the level
of accuracy in function evaluations can be controlled (up to a point) by the user but increased only at considerable expense. An example is the problem of determining the parameters in a differential equation, in which a function value is computed, say, by numerically solving an initial value problem. A natural strategy to adopt in solving such a problem is to use fairly inaccurate computed function values away from the solution and then to increase function evaluation accuracy as the solution is neared. We hope that this report will help in determining guidelines for the successful implementation of such strategies, although we do not address this matter in depth here.

At this point, perhaps a word of explanation is in order. It was stated at the outset that inaccuracy other than that in secant conditions is explicitly excluded from consideration here. However, our objective is to explore the role of inaccurate secant conditions in the convergence of least-change secant update methods; and so it is not inconsistent to consider inaccuracy in secant conditions which is attributable to inaccurate function evaluations while disregarding inaccuracy in function evaluations elsewhere in our analysis. The direct effects of inaccurate function evaluations on an iteration (1.2) are analyzed in Dennis and Walker (1983) and Ypma (1983). They are likely to be significant in practice only in the limiting behavior of an iteration sequence near $x^*$. Our interests here are more toward the cumulative effects on local convergence of inaccuracy in the matrices $B_k$ than toward the limiting effects of inaccuracy from other sources.

The second circumstance which we cite is that in which it might be advantageous to impose some sort of special structure on each $A_k$, even
though this structure is not fully reflected in $F'$ and $C$. This is to say that it might be desirable to require that each $A_k$ lie in some $A$, despite the fact that $[F'(x) - C(x)]$ fails to belong to $A$ for some $x \in Q$. For example, $A$ might be a subspace of $\mathbb{R}^{n \times n}$ consisting of matrices having a particularly appealing pattern of sparsity; and it might happen that for each $x \in Q$, the part of $[F'(x) - C(x)]$ which lies outside of this pattern of sparsity is so small that one is tempted to discard it and require that $A_k \in A$ for each $k$. In this case, one should choose $y_k = P_A [F'(x_{k+1}) - C(x_{k+1})] s_k$, where $P_A$ denotes the orthogonal projection onto $A$; and even though there may exist a choice rule for $y_k$ satisfying the conditions of Dennis and Walker (1981) (p.964, formula (3.13)), it may be impossible or impractical to implement it. The analysis given here should help to determine whether natural and readily available secant conditions can be safely used for updating in this case; and, if these secant conditions are not suitable, then it should suggest modifications of them which are safe to use and can be easily obtained. We shall explore such applications of this analysis in future work.

For perspective, let us recall that the local convergence analysis of least-change secant update methods in Dennis and Walker (1981) proceeds in the traditional two-stage form. In the first stage, one shows that the updates of interest exhibit a phenomenon known as bounded deterioration; it follows that methods employing these updates enjoy local $q$-linear convergence. In the second stage, one looks more closely at $q$-linearly convergent sequences of iterates produced by these methods and shows that this convergence is optimal in that it is asymptotically
as fast as that of the ideal stationary iteration of the form (1.2) in which \( B_k = C(x_*) + \mathcal{P}_A [F'(x_*) - C(x_*)] \) for each \( k \). The properties of the choice rule for determining secant conditions are fundamental to the analysis at each stage.

The local convergence analysis given in the following is in a similar two-stage form. The first stage is developed in Section 2. There it is assumed that a choice rule satisfying the conditions of Dennis and Walker (1981) is given, and it is first shown that one can make certain generous enlargements of sets of secant conditions specified by this choice rule and still retain bounded deterioration and the associated local q-linear convergence. Then this initial result is used to obtain a corollary in a form more useful in the applications toward which this work is directed. The second stage is developed in Section 3, in which it is shown that q-linearly convergent iteration sequences exhibit asymptotic speeds of convergence which can be regarded as optimal when one considers the inaccuracy in secant conditions used to determine updates. Indeed, the results of Section 3 make visible the extent to which inaccuracy in secant conditions degrades the asymptotic speed of convergence of an iteration sequence and show that this asymptotic speed can be made arbitrarily close to (or even equal to) that of the ideal stationary iteration provided it is feasible to exercise sufficient control over inaccuracy in secant conditions. In Section 4, we conclude with an illustration of how one might make use of the results of Sections 2 and 3.
2. Secant conditions and local linear convergence.

For the remainder of this report, we consider methods of the form (1.2) for solving (1.1) in which the $A_k$'s in (1.3) are maintained through secant updates into a sparse matrix subspace $Z \subseteq \mathbb{R}^{n \times n}$. In this section, our interest is in secant conditions which determine these updates and in the $q$-linear convergence properties of the methods which result.

We begin by establishing some notation and reviewing the basic properties of the sparse secant update. Throughout the sequel, $|| \cdot ||$ denotes both the $l_2$ norm on $\mathbb{R}^n$ and the induced operator norm on $\mathbb{R}^{n \times n}$; and $|| \cdot ||_F$ denotes the Frobenius norm on $\mathbb{R}^{n \times n}$. If $A$ is a subspace or affine subspace of $\mathbb{R}^{n \times n}$, then $P_A$ denotes the projection onto $A$ which is orthogonal with respect to the Frobenius norm and $P_A^T = I - P_A$ denotes the projection orthogonal to $P_A$. Members of the standard unit basis of $\mathbb{R}^n$ are denoted by $e_1, \ldots, e_n$, i.e., for $i = 1, \ldots, n$,

$$e_i = (0_i, 0_i, \ldots, 0_i)^T,$$

where $0_i = 1$ and $0_j = 0$ for $j \neq i$. For $i = 1, \ldots, n$, we use $P_{i}$ to denote the $l_2$ projection that sends a row (column) vector into the nearest row (column) vector having the sparsity of the $i$th row of an element of $Z$, i.e., $P_{i}v = e_i^T P_{i}Z (e_i v)$ for a row vector $v$.

With this notation, we can write down the formula for a sparse secant update as follows: If $A \in Z$ and if $s, y \in \mathbb{R}^n$ with $s \neq 0$, then the sparse secant update $A_+$ of $A$ in $Z$ with respect to $s$ and $y$ is given by

$$A_+ = A + \sum_{i=1}^{n} e_i^T (y - As) e_i P_{i}s^T + e_i^T (s P_{i} s) e_i P_{i} s^T,$$

(2.1)
where "+" denotes pseudo-inverse, i.e.,

\[ c^+ = \begin{cases} c^{-1}, & \text{if } c \neq 0 \\ 0, & \text{if } c = 0 \end{cases} \]

If \( A \) is not in \( Z \), then it is only necessary to alter (2.1) by replacing \( A \) with \( P_{\bar{Z}}A \).

Dennis and Schnabel (1979) have shown that \( A_+ \) given by (2.1) is the least-change secant update of \( A \) in \( Z \) with respect to \( s, y \), and the Frobenius norm \( ||\cdot|| \). This is to say that \( A_+ \) uniquely solves

\[
\min_{A \in M(Z, Q(y, s))} ||\bar{A} - A||,
\]

where \( Q(y, s) = \{ M \in \mathbb{R}^{n \times n} : Ms = y \} \) is the affine subspace of generalized quotients of \( y \) by \( s \) and \( M(Z, Q(y, s)) \) is the affine subspace of elements of \( Z \) for which the distance to \( Q(y, s) \) in the norm \( ||\cdot|| \) is minimal. Note that one has \( A_+ = P_{\bar{A}}A \), where \( M = M(Z, Q(y, s)) \); and if \( Z \cap Q(y, s) \neq \emptyset \), then \( M(Z, Q(y, s)) = Z \cap Q(y, s) \) and \( A_+ = P_{Z \cap Q(y, s)}A \).

Now suppose that \( \mathbf{x}: \Omega \times \Omega \rightarrow \mathbb{R}^n \) is a choice rule for determining secant conditions as considered in Dennis and Walker (1981). Applied to the sparse secant update case, Theorem 3.1 of Dennis and Walker (1981) gives requirements on \( \mathbf{x} \) and \( Z \) which are sufficient to ensure that sparse secant updates exhibit bounded deterioration and therefore yield methods with desirable local q-linear convergence properties. If these requirements are satisfied, then we shall regard \( \mathbf{x} \) and the secant conditions determined by it as accurate. More precisely, we make the following definition incorporating these requirements:
Definition 2.1: The choice rule $X$ and the secant conditions determined by it are \textit{accurate} if $X$ has the property with $Z$ that there exists an $\alpha \geq 0$ such that for any $x,x_+ \in \mathcal{Q}$ and any $y \in \mathcal{X}(x,x_+)$, one has

$$||F^2_{ZnQ}(0,s)(G-A_*)|| \leq \alpha \sigma(x,x_+)P$$

(2.2)

for every $G \in \mathcal{H}(Z,Q(y,s))$, where $s = x_+ - x$, $\sigma(x,x_+) = \max \{ |x - x_*|, |x_+ - x_*| \}$, and

$$A_* = P_z [F'(x_*) - C(x_*)].$$

(2.3)

For a discussion of these requirements, see pp. 962-964 of Dennis and Walker (1981).

Our interest is extending Theorem 3.1 of Dennis and Walker (1981) in the sparse secant update case by enlarging the sets of secant conditions given by an accurate choice rule $X$ without losing bounded deterioration of the updates determined by them. Toward this end, let us define for $v,w \in \mathbb{R}^n$

$$B(v,w) = \{Tv + (I-T)w: T = \text{diag}(t_1, \ldots, t_n), t_i \in [-1,1], i=1,\ldots,n \}. \quad (2.4)$$

One sees that $B(v,w)$ is just a "box" centered at $w$ which has $v$ as a vertex and sides parallel to the coordinate axes. For a set $S \subseteq \mathbb{R}^n$, we also define

$$B_U(v,S) = \bigcup_{w \in S} B(v,w).$$

(2.5)

Our extension of Theorem 3.1 of Dennis and Walker (1981) in the sparse secant update case is the theorem below.

Theorem 2.2. Let the standard hypothesis hold and let $Z$ have the
properties that \( A_* \) given by (2.3) and \( B_* = A_* + C(x_*) \) are such that \( B_* \) is invertible and there exists an \( r_* \) for which

\[
|I - B_*^{-1} F'(x_*)| \leq r_* < 1.
\]

Also assume that the choice rule \( \mathbf{x} \) is accurate in the sense of Definition 2.1. Under these hypotheses, if \( r \in (r_*, 1) \), then there are positive constants \( \epsilon_r, \sigma_r \) such that for \( x_0 \in \Omega \) and \( A_0 \in \mathbb{Z} \) satisfying \(|x_0 - x_*| < \epsilon_r \) and \(|A_0 - A_*| < \sigma_r \), the sequence \( \{x_k\} \) is well defined by \( B_0 = A_0 + C(x_0) \) and

\[
\begin{align*}
x_{k+1} &= x_k - B_k^{-1} F(x_k), \\
S_k &= x_{k+1} - x_k, \quad y_k \in \mathcal{B}_u(A_k s_k, \mathbf{x}(x_k, x_{k+1})), \\
B_{k+1} &\in \{(A_k^+) + C(x_{k+1}), (A_k^+) + C(x_k)\},
\end{align*}
\]

and satisfies \(|x_{k+1} - x_*| \leq r |x_k - x_*| \) for \( k = 0, 1, 2, \ldots \), where \((A_k^+)\) is the sparse secant update of \( A_k \) with respect to \( s_k \) and \( y_k \) given by (2.1). Furthermore, \( \{||B_k||\} \) and \( \{||B_k^{-1}||\} \) are uniformly bounded.

**Proof:** The proof is a straightforward application of Theorem A2.1 of Dennis and Walker (1981). In order to satisfy the hypothesis of that theorem, we begin by defining the update function \( U \) in a neighborhood \( N = N_1 \times N_2 \) of \((x_*, B_*)\). From the above hypotheses, one sees that there exist neighborhoods \( N_1 \) of \( x_* \) and \( N_2 \) of \( B_* \) such that \( N_1 \subseteq \Omega, N_2 \) contains only nonsingular matrices, and \( x_+ = B_*^{-1} F(x) \in \Omega \) for any \((x, B) \in N = N_1 \times N_2 \). We define \( U \) on \( N \) as follows.

**Definition 2.3:** For \((x, B) \in N\), set \( s = x_+ - x, A = P_z[B - C(x)]\), and

\[
U(x, B) = \{A_+ + C(x_+), A_+ + C(x) : y \in \mathcal{B}_u(As, \mathbf{x}(x, x_+))\}.
\]

where \( A_+ \) is the sparse secant update of \( A \) with respect to \( s \) and \( y \) given
by (2.1). Any \( y \in \mathcal{B}_u(\mathcal{A}_u, \mathcal{X}(x,x_+)) \) will be called \textit{admissible} (with respect to \( \mathcal{X} \)).

Lemma 2.4 below shows that \( \mathcal{U} \) so defined has the bounded deterioration property required by Theorem A2.1 of Dennis and Walker (1981), and so the theorem follows from that result.

\textbf{Lemma 2.4:} Let the standard hypothesis hold, and suppose that the choice rule \( \mathcal{X} \) is accurate in the sense of Definition 2.1. Let \( \mathcal{U} \) be defined by (2.7) on \( \mathcal{N} = \mathcal{N}_1 \times \mathcal{N}_2 \) as above. Then the bounded deterioration inequality

\[
||B - B_*|| \leq ||B - B_*|| + 2(\alpha + 2\beta)\gamma_c \sigma(x,x_+)^p
\]

(2.8)

holds for \((x,B) \in \mathcal{N} \) and \( B_+ \in \mathcal{U}(x,B) \), where \( \alpha \) is the constant of inequality (2.2) and \( \beta \) is a constant, e.g. \( \sqrt{n} \), for which \( ||\cdot|| \leq \beta ||\cdot|| \).

\textbf{Proof:} We only prove that (2.8) holds for \( B_+ = A_+ + C(x_+) \), since the proof in the other case is slightly simpler. For convenience, set

\[
P_1 s = s_1, \quad C(x) = C, \quad C(x_+) = C_+, \quad C(x_+) = C_+, \quad \text{and} \quad Q(0,s) = \mathbf{1}.
\]

The choice of \( y \in \mathcal{B}_u(\mathcal{A}_u, \mathcal{X}(x,x_+)) \) which determines \( B_+ \) can be written as \( y = TAs + (I - T)y^X \) for \( y^X \in \mathcal{X}(x,x_+) \) and \( T = \text{diag}(t_1, \ldots, t_n) \) with \( |t_i| \leq 1 \) for \( i = 1, \ldots, n \). Let \( G^X \in \mathcal{M}(\mathbf{Z}, Q(y^X,s)) \) and set \( G = TA + (I - T)G^X \). We first need to show that \( G \in \mathcal{M}(\mathbf{Z}, Q(y,s)) \). Our device for doing this is to show that \( G_+ = G \), where \( G_+ = P_\mathcal{M}G \) is the sparse secant update of \( G \) into \( \mathcal{M}(\mathbf{Z}, Q(y,s)) \). Since \( G^X \in \mathcal{M}(\mathbf{Z}, Q(y^X,s)) \),

\[
0 = \sum_{i=1}^n (s_1^T s_i + \epsilon_i^T (y - G^X) s_i) \epsilon_i s_1^T
\]
and multiplying both sides by the diagonal matrix \((I - T)\) gives

\[
0 = \sum_{i=1}^{n} (s_i T_{s_i})^T (I - T) (y - G s) \epsilon_i s_i^T
\]

\[
= \sum_{i=1}^{n} (s_i T_{s_i})^T (T A s - T A s + (I - T) (y - G s)) \epsilon_i s_i^T
\]

\[
= \sum_{i=1}^{n} (s_i T_{s_i})^T (y - G s) \epsilon_i s_i^T = G_+ - G.
\]

Now we use (2.23) of Dennis and Walker (1981) with \(S = Z\) to get

\[
B_+ - B_* = A_+ + C_+ - B_* \]

\[
= P^T (A + C - B_*) + P^T (G + C - B_*)
\]

\[
= P^T (A + C - B_*) + P^T \left[ T (A + C - B_*) + (I - T) (G + C - B_*) \right]
\]

\[
= P^T (A + C - B_*) + T P^T (A + C - B_*) + (I - T) P^T (G + C - B_*)
\]

The last equality is implied by the following: left-multiplication by the diagonal matrix \(T\) commutes with the projections onto \(Z\) and \(M\), and

\[
P^T (A + C - B_*) \]

\[
= \lim_{j \to \infty} (P^T P M)^j \]

consequently left-multiplication by \(T\) commutes with \(P^T (A + C - B_*)\) and \(P^T (A + C - B_*)\).

To obtain (2.8) from this equation, first note that (2.2) gives

\[
|| (I - T) P^T (G + C - B_*) || \leq 2 || P^T (G - A_+ + C_+ - C_*) ||
\]

\[
\leq 2 \left[ || P^T (G - A_+ || + \beta \gamma_c \sigma(x, x_+) \right]
\]

\[
\leq 2 (\alpha + \beta \gamma_c \sigma(x, x_+) P.
\]

To bound the other two terms, use the Pythagorean Theorem and the form
of $T$ to obtain
\[
\|P_{Zn}(A + C_+ - B_*) + TP_{Zn}^L (A + C_+ - B_*)\|^2 \leq \|P_{Zn}(A + C_+ - B_*)\|^2 + \\
\|TP_{Zn}^L (A + C_+ - B_*)\|^2
\]
\[
\leq \|P_{Zn}(A + C_+ - B_*)\|^2 + \|P_{Zn}^L (A + C_+ - B_*)\|^2
\]
\[
= \|A + C_+ - B_*\|^2.
\]
Consequently,
\[
\|B_+ - B_*\| \leq \|A + C_+ - B_* + \| + 2(\alpha + \beta \gamma_C) \sigma(x, x_*)^P
\]
\[
\leq \|B - B_*\| + 2(\alpha + 2\beta \gamma_C) \sigma(x, x_*)^P,
\]
and the lemma is proved.

The secant conditions admitted in iteration (2.6) are determined explicitly by sets of accurate secant conditions specified by an accurate choice rule. This report is directed principally toward applications in which accurate secant conditions are unobservable, at least in practice; and it may not be clear how to apply Theorem 2.2 in such applications. Consequently, we shall now work toward a local $q$-linear convergence result in which the role of an accurate choice rule in determining admissible secant conditions is more implicit. This result is more directly useful than Theorem 2.2 in many applications of the sort which we have in mind.

We work from the premise that often when accurate secant conditions cannot actually be observed, one can at least determine sets which contain them. Indeed, one might often be able to observe inaccurate secant
conditions together with bounds on the inaccuracy in them and through
these bounds determine sets in which accurate secant conditions lie. As
an illustration, let us consider the classical case with \( C(x) = 0 \) in
which \( y^x_k = F(x^k+1) - F(x^k) \) is an accurate secant condition for each \( k \);
and let us suppose that these accurate secant conditions \( y^x_k \) are not
observable in practice because of inaccuracy in the computed values of
\( F \). If one can bound the inaccuracy in computed \( F \)-values in some way,
then one should be able to specify sets about the inaccurate observed
\( y^x_k \)-values which contain the true \( y^x_k \)-values. To be more specific, let
\( \tilde{F}(x) = F(x) + N(x) \) be the computed value of \( F(x) \) for each \( x \in \Omega \). Then for
each \( k \), the observed value of \( y^x_k \) is \( \tilde{y}^x_k = y^x_k + N(x^k+1) - N(x^k) \). If one
assumes, as in Barrera and Dennis (1979), that \( |N(x^k+1) - N(x^k)| \leq \eta_k \) and,
hence, \( |y^x_k - \tilde{y}^x_k| \leq \eta_k \) for each \( k \), then each set
\[
Y_k = \{ y \in \mathbb{R}^n : |y - \tilde{y}^x_k| \leq \eta_k \}
\] (2.9)
contains an accurate secant condition in this case.

Sets such as \( Y_k \) above are central to our local convergence result
below, for they allow the determination of secant conditions which are
admissible in iteration (2.6). As a device for determining admissible
secant conditions from such sets, let us define for \( v \in \mathbb{R}^n \) and compact
\( S \subseteq \mathbb{R}^n \) a vector \( \hat{\gamma} = \hat{\gamma}(v,S) \) as follows: For \( i = 1, \ldots, n \), set
\[
\delta_1 = \inf \{ \varepsilon^T_S : s \in S \} \text{ and } \delta_2 = \sup \{ \varepsilon^T_S : s \in S \};
\]
then define \( \hat{\gamma} \) by
The vector $\hat{y}$ has the geometrically appealing property that in the $L_2$ norm, it is the vector closest to $v$ in the smallest box containing $S$ and having sides parallel to the coordinate axes. (See Fig. 2.1 below.) It is reasonable to believe that $\hat{y}$ can be easily determined for most sets $S$ and in particular when $S$ is a set such as $Y_k$ which is naturally prescribed as one containing accurate secant conditions. In our local convergence result, it turns out that admissible secant conditions are found in sets of the form $B(v,\hat{y})$ given by (2.4) with $w=\hat{y}$. We shall say more about the role of vectors such as $\hat{y}$ in determining admissible secant conditions following the statement of our local convergence result.

$$\hat{y}^T = \begin{cases} 
1_i, & \text{if } \epsilon_i^T v \leq 1_i \\
\epsilon_i^T v, & \text{if } 1_i \leq \epsilon_i^T v \leq u_i \\
u_i, & \text{if } u_i \leq \epsilon_i^T v 
\end{cases} \tag{2.10}$$

for $i=1,\ldots,n$. The vector $\hat{y}$ has the geometrically appealing property that in the $L_2$ norm, it is the vector closest to $v$ in the smallest box containing $S$ and having sides parallel to the coordinate axes. (See Fig. 2.1 below.) It is reasonable to believe that $\hat{y}$ can be easily determined for most sets $S$ and in particular when $S$ is a set such as $Y_k$ which is naturally prescribed as one containing accurate secant conditions. In our local convergence result, it turns out that admissible secant conditions are found in sets of the form $B(v,\hat{y})$ given by (2.4) with $w=\hat{y}$. We shall say more about the role of vectors such as $\hat{y}$ in determining admissible secant conditions following the statement of our local convergence result.
As an aid in showing that sets of the form $\mathcal{B}(v, \hat{y})$ contain admissible secant conditions, let us also define for $v \in \mathbb{R}^n$ and $S \subseteq \mathbb{R}^n$

$$\mathcal{B}_n(v, S) = \bigcap_{w \in S} \mathcal{B}(v, w),$$

(2.11)

where $\mathcal{B}(v, w)$ is given by (2.4). One sees that $\mathcal{B}_n(v, S)$ is the counterpart for intersection of the union $\mathcal{B}_u(v, S)$ defined by (2.5). We need the following lemma.

**Lemma 2.5:** For any $v \in \mathbb{R}^n$ and compact $S \subseteq \mathbb{R}^n$, one has

$$\mathcal{B}(v, \hat{y}) = \mathcal{B}_n(v, S)$$

(2.12)

where $\hat{y} = \hat{y}(v, S)$ is given by (2.10), $\mathcal{B}(v, \hat{y})$ is given by (2.4) with $w = \hat{y}$, and $\mathcal{B}_n(v, S)$ is given by (2.11).

**Proof:** To show that $\mathcal{B}(v, \hat{y}) \subseteq \mathcal{B}_n(v, S)$, let $y =Tv + (I-T)\hat{y}$ for $T = \text{diag}(t_1, \ldots, t_n)$ with $-1 \leq t_i \leq 1$ for $i = 1, \ldots, n$. Let $w \in S$ and note that one can write $\hat{y} = \hat{y}(v, S)$ for $\hat{T} = \text{diag}(\hat{t}_1, \ldots, \hat{t}_n)$ with $0 \leq \hat{t}_i \leq 1$ for $i = 1, \ldots, n$. Indeed, for $i = 1, \ldots, n$, one has that $1 \leq \epsilon_i T \leq u_i$; and so the appropriate $\hat{c}_i$ is given by

$$\hat{c}_i = \begin{cases} \frac{\epsilon_i T - 1_i}{\epsilon_i (w - v)}, & \text{if } \epsilon_i \hat{y} = 1_i \\ 1, & \text{if } \epsilon_i \hat{y} = \epsilon_i T v \\ \frac{u_i - \epsilon_i T w}{\epsilon_i (v - w)}, & \text{if } \epsilon_i \hat{y} = u_i \end{cases}$$

Now one has

$$y = Tv + (I - T)[\hat{y}v + (I-\hat{T})w]$$
\[ T = T + (I - T) \hat{T}; \]
and it is easily seen that \( T = \text{diag}(\hat{t}_1, \ldots, \hat{t}_n) \) with
\(-1 \leq \hat{t}_i \leq 1 \) for \( i = 1, \ldots, n \). Thus \( y \in B(v, w) \), and one concludes that
\[ B(v, \hat{y}) \subseteq B_n(v, S). \]

To show that \( B_n(v, S) \subseteq B(v, \hat{y}) \), let \( w \in B_n(v, S) \). It must be shown
that for \( i = 1, \ldots, n \), one has

\[ \epsilon_i w = t_i \epsilon_i v + (1 - t_i) \epsilon_i y, \quad (2.13) \]

where \(-1 \leq t_i \leq 1\). Let \( y_1, \ldots, y_n, y^1, \ldots, y^n \) be such that \( l_i = \epsilon_i y_i \) and
\( u_i = \epsilon_i y^1 \) for \( i = 1, \ldots, n \). For some \( i \), suppose that \( \epsilon_i \hat{y} = l_i \). Since
\( w \in B(v, y_i) \), one can write \( \epsilon_i w = t_i \epsilon_i v + (1 - t_i) \epsilon_i y_i \), where \(-1 \leq t_i \leq 1\).

Since \( \epsilon_i \hat{y} = l_i = \epsilon_i y_i \), (2.13) holds with this \( t_i \). Since \( w \) also belongs to
\( B(v, y^1) \), (2.13) holds similarly for an appropriate \( t_i \) if \( \epsilon_i \hat{y} = u_i \). The
remaining case is that in which \( l_i \leq \epsilon_i v = \epsilon_i \hat{y} \leq u_i \). In this case, one
has for some \( t_i, t^i \in [-1, 1] \) that both \( \epsilon_i w = t_i \epsilon_i v + (1 - t_i) \epsilon_i y_i = t_i \epsilon_i v + (1 - t_i) l_i \leq \epsilon_i v \) and, similarly, \( \epsilon_i w = t_i \epsilon_i v + (1 - t_i) u_i \geq \epsilon_i v \). Con­sequently, \( \epsilon_i w = \epsilon_i v \) and (2.13) holds with \( t_i = 1 \). This completes the proof.

With this lemma, we can easily obtain from Theorem 2.2 a local con­vergence result suitable for application in circumstances in which one
can determine for each \( k \) a set \( Y_k \) containing an accurate secant condi­tion.

**Corollary 2.6:** Let the hypotheses of Theorem 2.2 hold. If \( r \in (r_*, 1) \),
then there are positive constants \( \epsilon_r, \sigma_r \) such that for \( x_0 \in \Omega \) and \( A_0 \in \mathbf{Z} \)
satisfying \( |x_0 - x_*| < \epsilon_r \) and \( |A_0 - A_*| < \sigma_r \), the sequence \( \{x_k\} \) is well­defined by \( B_0 = A_0 + C(x_0) \) and
\[ x_{k+1} = x_k - B_k^{-1}F(x_k), \]

\[ Y_k \subseteq \mathbb{R}^n \text{ satisfies } Y_k \cap \mathcal{X}(x_k, x_{k+1}) \neq \emptyset, \tag{2.14} \]

\[ s_k = x_{k+1} - x_k, \quad y_k \in \mathcal{B}(A_k s_k, y_k), \]

\[ B_{k+1} \subseteq \{(A_k)_+, (A_k)_+ + C(x_{k+1})\}, \]

and satisfies \(|x_{k+1} - x_*| \leq r|x_k - x_*| \) for \(k = 0, 1, 2, \ldots\), where \(\hat{\phi}_k = \hat{\phi}(A_k s_k, y_k)\) is determined as in (2.10) with \(v = A_k s_k\) and \(S = Y_k\) and \((A_k)_+\) is the sparse secant update of \(A_k\) with respect to \(s_k\) and \(y_k\) given by (2.1). Furthermore, \(\{||B_k||\}\) and \(\{||B_k^{-1}||\}\) are uniformly bounded.

**Proof:** Since \(\mathcal{X}(x_k, x_{k+1}) \cap Y_k \neq \emptyset\) for each \(k\), one sees that

\[ \mathcal{B}(A_k s_k, y_k) \subseteq \mathcal{B}(A_k s_k, \mathcal{X}(x_k, x_{k+1})) \] provided \(x_k, x_{k+1} \in \Omega\). Since \(\mathcal{B}(A_k s_k, \hat{\phi}_k) = \mathcal{B}(A_k s_k, y_k)\) by Lemma 2.5, the corollary now follows immediately from Theorem 2.2.

This local convergence result merits some discussion. We begin by considering the appropriateness of the vectors \(\hat{\phi}_k = \hat{\phi}(A_k s_k, y_k)\) for specifying secant conditions. These vectors cannot be expected to provide secant conditions which are accurate in the sense understood here. Nevertheless, they have certain very appealing properties, which we now outline.

Let us first note that if \(A \in \mathbb{Z}\) and if \(s, y \in \mathbb{R}^n\) with \(s \neq 0\), then \(A_+\) given by (2.1) satisfies

\[ ||A_+ - A||^2 = \sum_{i=1}^{n} (||P_i s||^+ ||e_i^T(y - As)||)^2. \tag{2.15} \]

Suppose that \(y\) is allowed to range over a subset \(S \subseteq \mathbb{R}^n\). One sees that \(||A_+ - A||^2\) is minimized over \(S\) by any \(y \in S\) for which the weighted
$l_2$-norm on the right-hand side of (2.15) is minimal. In the case of the full Broyden update in which $Z = \mathbb{R}^{n \times n}$ and $P_i = I$ for $i = 1, \ldots, n$, (2.15) becomes

$$||A_+ - A||^2 = \frac{|y - As|^2}{|s|^2}. \quad (2.16)$$

Thus in this case, $||A_+ - A||$ is minimized by any $y \in S$ for which $|y - As|$ is minimal. For general $Z$, it may not be so clear which vectors in $S$ minimize $||A_+ - A||$. However, this norm will certainly be minimized if a $y \in S$ can be found for which each $|\epsilon_i^T(y - As)|$ is minimal for $i = 1, \ldots, n$. Such a $y$ may not exist for some sets $S$, but suppose that $S$ is a box with sides parallel to the coordinate axes. In this case, there is a unique $y \in S$ for which each $|\epsilon_i^T(y - As)|$ is minimal for $i = 1, \ldots, n$, namely $\hat{y} = \hat{y}(As, S)$ determined as in (2.10); and this $\hat{y}$ is also the unique minimizer of $|y - As|$ over $S$.

Returning to the context of iteration (2.14), we denote by $S_k$ the smallest box containing $Y_k$ and having sides parallel to the coordinate axes. One sees from the above discussion that $\hat{y}_k$ is the unique vector which minimizes $||A_k^+ - A_k||$ over $S_k$; and in this sense, it represents the most conservative secant condition among all those determined by vectors in a set, namely $S_k$, which is known to contain a vector associated with an accurate secant condition. We find it attractive that $\hat{y}_k$ thus conservatively imposes on $(A_k^+)$ the presumably useful but less than totally trustworthy secant information in vectors in $S_k$. In one form or another, the principle of making conservative use of current secant information has consistently led to successful updates in the past.
Since $Y_k$ is the subset of $S_k$ in which vectors specifying accurate secant conditions are known to lie, one might wonder at this point if a secant condition better than that associated with $\hat{y}_k$ might be obtained by finding an element of $Y_k$ which minimizes $\| (A_k)^+ - A_k \|$ over $Y_k$. We think not. To support our position, we consider an "accurate" $y_k \in \mathcal{X}(x_k, x_{k+1}) \cap Y_k$ and observe that $|\epsilon^T_i (\hat{y}_k - y_k) | \leq |\epsilon^T_i (A_k s_k - y_k) |$ and, hence,

$$|\epsilon^T_i [(A_k)^+ s_k - y_k] | \leq |\epsilon^T_i (A_k s_k - y_k) | \text{ for } i = 1, \ldots, n, \quad (2.17)$$

where $(A_k)^+$ is the sparse secant update of $A_k$ in $\mathcal{Z}$ determined by $\hat{y}_k$. In other words, if $(A_k)^+$ is determined by $\hat{y}_k$, then each component of $(A_k)^+ s_k$ is at least as close as its counterpart in $A_k s_k$ to the corresponding component of an "accurate" $y_k$. On the other hand, if $(A_k)^+$ is determined by a vector in $Y_k$ such that $\| (A_k)^+ - A_k \|$ is minimal, then (2.17) may not hold; whether or not (2.17) does hold in this case depends on the geometry of $Y_k$. To illustrate this and to lend some perspective to this discussion, let us recall that Barrera and Dennis (1979) consider full Broyden updating with $C(x) = 0$ when the accurate secant condition determined by $y_k = F(x_{k+1}) - F(x_k)$ is not observable because of inaccuracy in computed values of $F$. As we indicated previously, they assume that the inaccuracy in computed $F$-values is bounded so that $y_k$ is contained in $Y_k$ given by (2.9). In Barrera and Dennis (1979), the secant condition studied is that determined by the vector in $Y_k$ which is closest to $A_k s_k$ in the $\| \cdot \|_2$-norm, and which we denote here by $y_k^{BD}$. One sees from (2.16) that $y_k^{BD}$ is the unique minimizer for full Broyden updating of $\| (A_k)^+ - A_k \|$ in $Y_k$. In Fig. 2.2 below, we have shown a circumstance in which (2.17) does not hold for
(A_k)_+ determined by y_k^{BD}, while, of course, it does hold for (A_k)_+ determined by \hat{y}_k. In Fig. 2.2, \tilde{y}_k denotes the observed (inaccurate) value of \hat{y}_k as before.

Barrera and Dennis (1979) report on numerical experiments which suggest that the vectors y_k^{BD} are useful in practice, although they give no supporting convergence analysis. We feel that the analysis given above reinforces their encouraging computational results, even though their methods do not actually fall within its scope. Indeed, the circumstance illustrated in Fig. 2.2 notwithstanding, it seems very unlikely that a vector y_k^{BD} will fail to lie within the box B(A_k, s_k, \hat{y}_k) unless one is very near the solution; and so updates determined by these vectors seem very likely to exhibit bounded deterioration in practice, at least until some stopping criterion is met, and thus to lead to methods with satisfactory local q-linear convergence properties. Simi-
larly, the analysis here strongly suggests that in most traditional applications, the usual secant conditions determined through the use of inaccurate computed function values should be completely adequate.

2. Speed of convergence.

In the preceding section, it was shown that under the hypotheses of Theorem 2.2 sequences of iterates produced by iteration (2.6) exhibit local q-linear convergence. In this section, we analyze the behavior of these iteration sequences in greater depth. Specifically, we consider q-linearly convergent sequences produced by (2.6) with the objective of estimating their asymptotic speeds of convergence.

When accurate secant conditions are used in all iterations of (2.6), the resulting asymptotic speed of convergence has been determined by Dennis and Walker (1981). To recall those results, suppose that the hypotheses of Theorem 2.2 hold and that \( \{x_k\} \) is an iteration sequence produced by (2.6) which converges q-linearly to \( x^* \). One can easily modify Theorem 3.3 of Dennis and Walker (1981) to include the case \( B_{k+1} = (A_k) + C(x_k) \); and from this modification, one sees that if the secant condition is determined by \( y_k \in \pi(x_k, x_{k+1}) \) at each iteration of (2.6), then

\[
\lim_{k \to \infty} \frac{|x_{k+1} - x^*|}{|x_k - x^*|} = \lim_{k \to \infty} |I - B_*^{-1} F'(x_*)| \frac{x_k - x^*}{|x_k - x^*|}.
\] (3.1)

In particular,

\[
\lim_{k \to \infty} \frac{|x_{k+1} - x^*|}{|x_k - x^*|} \leq |I - B_*^{-1} F'(x_*)| \leq r_*.
\] (3.2)

This is to say that if an accurate secant condition is used at each iteration, then the asymptotic q-linear convergence rate constant is the
same as that of the ideal stationary iteration of the form (1.2) which takes each $B_k$ as close as possible to $F'(x_*)$.

We offer two results below concerning the extent to which inaccuracy in secant conditions affects asymptotic speeds of convergence. These results are given as Corollaries 3.2 and 3.3 of Theorem 3.1, which can be regarded as the central technical result of this section. One can certainly obtain other asymptotic convergence results as corollaries of Theorem 3.1, but we feel that the two given here are particularly appealing and useful. Corollary 3.2 can be interpreted as saying that if at least some minimal positive proportion of accurate secant information is imposed on the update at each iteration, then the optimal asymptotic speed of convergence given by (3.1) and (3.2) will result. Corollary 3.3 shows precisely how much given amounts of inaccuracy in secant conditions can cause the asymptotic speed of convergence to deteriorate from this optimal asymptotic speed. It is in the tradition of previous results in Dennis and Moré (1974) and Dennis and Walker (1981) in that it relates the asymptotic speed of convergence to the extent to which the successive $B_k$'s approximate the action of $B_*$ in the directions of the $s_k$'s. From it, one sees that the asymptotic speed of convergence can still be regarded as optimal, given the inaccuracy in the secant conditions used to update the $B_k$'s.

Throughout this section, we assume that $\mathbf{X}:\mathbb{R}^n \times \mathbb{R}^{n+2} \rightarrow \mathbb{R}$ is a choice rule which is accurate in the sense of Definition 2.1. The statements and proofs of our results rely heavily on the characteristic property of each $y_k \in B_u(A_k, s_k, x(x_k, x_{k+1}))$ in (2.6) that it can be written as
We shall use (3.3) and (3.4) freely below with minimal explanation.

Theorem 3.1: Let the hypotheses of Theorem 2.1 hold and assume that for some \( x_0 \in \mathbb{R}^n \) and \( A_0 \in \mathbb{Z} \), \( \{x_k\} \) is a sequence defined by (2.6) which converges \( \varphi \)-linearly to \( x_\star \) with \( x_k \to x_\star \) for all \( k \). Assume further that \( \{ \|B_k\| \} \) is uniformly bounded, and for each \( T_k \) in (3.3) and (3.4), set

\[
T_{ik} = \text{diag}(t_{1k}, \ldots, t_{nk}), \quad t_{ik} \in [-1, 1] \text{ for } i = 1, \ldots, n.
\]

Then

\[
\lim_{k \to \infty} (\lambda - t_{ik})^2 \|P_i s_k\|_{B_k} = 0, \quad i = 1, \ldots, n.
\]

for \( i = 1, \ldots, n \); and it follows that

\[
\lim_{k \to \infty} (\lambda - t_{ik})^2 \|P_i s_k\|_{B_k} = 0, \quad i = 1, \ldots, n.
\]

Proof: For convenience, set \( C(x_k) = C_k \), \( C(x_\star) = C_\star \), \( Q(0, s_k) = M_k \), \( \sigma_k = \max\{ |x_k - x_\star|, |x_{k+1} - x_\star| \} \), and \( E_k = B_k - B_\star \). Note that

\[
y_k = T_k A_k s_k + (I - T_k) y_k^*, \quad (3.3)
\]

for some \( y_k^* \in \mathbb{R}(x_k, x_{k+1}) \) and

\[
T_k = \text{diag}(t_{1k}, \ldots, t_{nk}), \quad t_{ik} \in [-1, 1] \text{ for } i = 1, \ldots, n. \quad (3.4)
\]
and, as in the proof of Lemma 2.4,

\[ E_{k+1} = \mathbf{Z} \mathbf{n}^T_k \mathbf{E}_k + T_k \mathbf{P}^T \mathbf{Z} \mathbf{n}^T_k \mathbf{E}_k + O(\sigma_k^p) \]

\[ = T_k \mathbf{E}_k + (I - T_k) \mathbf{P} \mathbf{Z} \mathbf{n}^T_k \mathbf{E}_k + O(\sigma_k^p) \]

\[ = T_k \mathbf{P} \mathbf{Z} \mathbf{E}_k + (I - T_k) \mathbf{P} \mathbf{Z} \mathbf{n}^T_k \mathbf{E}_k + O(\sigma_k^p). \]

The \( i \)th row of this equation is

\[
\epsilon_i^T \mathbf{E}_{i+1} = \epsilon_i \mathbf{P} \epsilon_i^T \mathbf{E}_i + (1 - \epsilon_i) \{ (\epsilon_i \mathbf{P} \epsilon_i^T \mathbf{E}_i) \mathbf{P} \mathbf{s}_i \mathbf{l}^2 \mathbf{P} \mathbf{s}_i^\top \} + O(\sigma_k^p) \]

\[(3.11)\]

\[
= \mathbf{P} \mathbf{s}_i \mathbf{l}^2 \mathbf{P} \mathbf{s}_i^\top + O(\sigma_k^p). 
\]

We want to bound \( |v_{ik}| \), and direct computation gives

\[ |v_{ik}|^2 = |\epsilon_i \mathbf{P} \epsilon_i^T \mathbf{E}_i|^2 - (1 - \epsilon_i) \{ (\epsilon_i \mathbf{P} \epsilon_i^T \mathbf{E}_i) \mathbf{P} \mathbf{s}_i \mathbf{l}^2 \mathbf{P} \mathbf{s}_i^\top \}^2. \]

Since \((\epsilon_i \mathbf{P} \epsilon_i^T \mathbf{E}_i) \mathbf{P} \mathbf{s}_i \mathbf{l} = (\epsilon_i \mathbf{P} \epsilon_i^T \mathbf{E}_i) \mathbf{P} \mathbf{s}_i \mathbf{l} \mathbf{s}_i \mathbf{l} \mathbf{s}_i^\top \) and \(|\epsilon_i| \leq 1\),

\[ |v_{ik}|^2 \leq |\epsilon_i \mathbf{P} \epsilon_i^T \mathbf{E}_i|^2 - (1 - \epsilon_i) \{ (\epsilon_i \mathbf{P} \epsilon_i^T \mathbf{E}_i) \mathbf{P} \mathbf{s}_i \mathbf{l}^2 \mathbf{P} \mathbf{s}_i^\top \}^2 = \eta_{ik}^2 - \psi_{ik}^2. \]

Now we use the general inequality
\[(a^2 - b^2)^{1/2} \leq a - \frac{b^2}{2a}\]

to obtain

\[|v_{ik}| \leq n_{ik} - \frac{\psi_{ik}^2}{2n_{ik}}\]

It follows from (3.11) that

\[n_{i, k+1} \leq n_{ik} - \frac{\psi_{ik}^2}{2n_{ik}} + O(\sigma_k^p)\].

By a standard argument (see, e.g., Dennis and Walker (1981), p. 966), this implies that \(\sum \psi_{ik}^2 < \infty\) and hence, \(0 = \lim_{k \to \infty} \psi_{ik}\), which is (3.5). To get (3.6), we note that

\[\psi_{ik} \geq \frac{1}{(1 - |\tau_{ik}|^2)/(1 + |\tau_{ik}|^2)}^{1/2} \psi_{ik} \geq 0\].

Finally, we prove (3.8) from (3.6), since (3.7) follows in just the same way from (3.5). For any \(1 \leq i \leq n\), (3.9) implies

\[\frac{e_i^T (I - T_k) E_k s_k}{|s_k|} = \frac{(1 - |\tau_{ik}|) e_i^T \left[ P_k s_k + P_k (C_k - C_k) s_k \right]}{|s_k|}\]

\[= \frac{(1 - |\tau_{ik}|) e_i^T P_k s_k}{|s_k|} + O(\sigma_k^p)\]

\[= \frac{(1 - |\tau_{ik}|) e_i^T P_k s_k}{|s_k|} + O(\sigma_k^p).\]

Since \(|s_k| \geq |P_k s_k|\), (3.8) follows from (3.6).

**Corollary 3.2.** Let \(\{x_k\}\) be an iteration sequence generated under the
hypotheses of Theorem 3.1. If for some \( \tau, 0 \leq \tau < 1 \), \( T_k \) in (3.3) and (3.4) satisfies \( 0 \leq |t_{ik}| \leq \tau \) for \( 1 \leq i \leq n \) and \( k = 0, 1, \ldots \), then

\[
\lim_{k \to \infty} \frac{|x_{k+1} - x^*|}{|x_k - x^*|} = |1 - B_k^{-1} F'(x_k)| \leq \tau^*.
\]  

(3.12)

**Proof:** It is immediate from (3.8) and the existence of the bound \( \tau < 1 \) that

\[
\lim_{k \to \infty} \frac{s_k}{|s_k|} = 0.
\]

Since \( \{x_k\} \) converges q-linearly to \( x_* \),

\[
\lim_{k \to \infty} \frac{|s_k|}{|e_k|} < \infty,
\]

and so the result follows from Theorem A.1 of the appendix.

**Corollary 3.3** Let \( \{x_k\} \) be an iteration sequence generated under the hypotheses of Theorem 3.1. Then for \( y_k \) as in (3.3),

\[
\lim_{k \to \infty} \frac{(B_k - B_*) s_k}{|s_k|} - \bar{D}_k \left( \frac{y_k - y_k \mathbf{1}}{|s_k|} \right) = 0
\]  

(3.13)

and

\[
\lim_{k \to \infty} \frac{(B_k - B_*) s_k}{|x_k - x_*|} - \bar{D}_k \left( \frac{y_k - y_k \mathbf{1}}{|x_k - x_*|} \right) = 0,
\]  

(3.14)

where \( \bar{D}_k = \text{diag}(d_{1k}, \ldots, d_{nk}) \) with \( d_{ik} = t_{ik} |t_{ik}|^+ \) for \( i = 1, \ldots, n \). If

\[
\lim_{k \to \infty} \frac{|B_k^{-1} \bar{D}_k (y_k - y_k \mathbf{1})|}{|x_k - x_*|} \leq \lambda,
\]  

(3.15)

then
\[
\lim_{k \to \infty} \frac{|x_{k+1} - x_*|}{|x_k - x_*|} \leq r_* + \lambda. \tag{3.16}
\]

If
\[
\lim_{k \to \infty} \frac{|B^{-1}D_k(y_k - y_k^*)|}{|s_k|} \leq \mu < 1, \tag{3.17}
\]
then
\[
\lim_{k \to \infty} \frac{|x_{k+1} - x_*|}{|x_k - x_*|} \leq \frac{r_* + \mu}{1 - \mu}. \tag{3.18}
\]

**Proof:** We first establish (3.13) from (3.8). Since \(\{x_k\}\) converges q-linearly to \(x_*\), (3.14) then follows from (3.13). For each \(k\), we define \(E_k\), \(\Pi_k\), and \(\sigma_k\) as in the proof of Theorem 3.1. From (3.8), we have that
\[
\lim_{k \to \infty} \frac{E_k s_k}{|s_k|} = 0.
\]

Since
\[
E_{k+1} s_k = T_k P_k \Pi_k E_k s_k + O(\sigma_k) s_k
= T_k E_k s_k + O(\sigma_k) s_k
\]
by (3.10), it follows that
\[
\lim_{k \to \infty} \frac{E_k s_k}{|s_k|} - \frac{E_{k+1} s_k}{|s_k|} = 0.
\]

But \(E_{k+1} = (A_k) + A_* + O(\sigma_k)\), so
\[
\lim_{k \to \infty} \frac{(A_k) + A_*}{{|s_k|}} \frac{s_k}{|s_k|} = 0.
\]

Let \(C_k \in M(Z,Q(y_k, s_k))\) and \(C_k^\chi \in M(Z,Q(y_k^\chi, s_k^\chi))\). Note that if \(P_i s_k \neq 0\),
then $\varepsilon_{i'k} = \varepsilon G_{ik} s_k$ and $\varepsilon_{i'k} = \varepsilon_{i'k} S_k$. From this and from Dennis and Walker (1981), p. 959, it follows that

\[
\lim_{k \to \infty} B_k^{-1} E_k s_k - \lim_{k \to \infty} B_k^{-1} D_k (y_k - y_k^*) = 0
\]

and (3.13) follows.

It also follows from continuity that

\[
\lim_{k \to \infty} \frac{B_k^{-1} E_k s_k}{|s_k|} - \frac{B_k^{-1} D_k (y_k - y_k^*)}{|s_k|} = 0
\]

and

\[
\lim_{k \to \infty} \frac{B_k^{-1} E_k s_k}{|x_k - x_*|} - \frac{B_k^{-1} D_k (y_k - y_k^*)}{|x_k - x_*|} = 0.
\]

If (3.15) holds, then (3.16) follows immediately from (3.20) and (A.2) of the appendix. If (3.17) holds, then (3.19) implies

\[
\lim_{k \to \infty} \frac{B_k^{-1} E_k s_k}{|s_k|} \leq \mu^*.
\]

We see from (A.2) of the appendix that for

\[
r = \lim_{k \to \infty} \frac{|x_{k+1} - x_*|}{|x_k - x_*|},
\]

we have

31
\[ r \leq r_* + \mu \lim_{k \to \infty} \frac{|s_k|}{|x_k - x_*|} \leq r_* + \mu (1 + \tau); \]

and \( r \leq \frac{(r_* + \mu)}{(1 - \mu)} \) follows if \( \mu < 1 \).

\section{An Application}

We conclude with a somewhat speculative illustration of how one might make use of the local convergence analysis given in this report. For our illustration, let us return to the case considered by Barrera and Dennis (1979) which was touched on in Section 2. For clarity, we review the essential features of this case developed previously:

1. The updates of interest are full Broyden updates with \( C(x) = 0 \), and so \( y_k = F(x_{k+1}) - F(x_k) \) is an accurate secant condition for each \( k \).

2. For each \( x \in \Omega \), one computes an inaccurate value \( \tilde{F}(x) = F(x) + N(x) \); and the observed secant condition is \( \tilde{y}_k = y_k^B + N(x_{k+1}) - N(x_k) \) for each \( k \).

3. For each \( k \), a bound \( |N(x_{k+1}) - N(x_k)| \leq \eta_k \) is known; and so \( y_k \in Y_k \), where \( Y_k \) is given by (2.9).

As we remarked in Section 2, our analysis does not strictly cover the use of the Barrera-Dennis secant conditions \( y_k^{BD} \) illustrated in Fig. 2.2, although it does offer insight into their practical effectiveness. In fact, consideration of circumstances such as that of Fig. 2.2 suggests that these secant conditions may not lend themselves to analysis.

Here, we suggest the use of certain alternative secant conditions in the Barrera-Dennis case and apply our analysis to a method (1.2) which uses sparse secant updates determined by these secant conditions. To emphasize the role of the inaccuracy in these secant conditions most
clearly, we shall maintain our policy of disregarding all inaccuracy except that in the secant conditions. In particular, it is assumed that the accurate value $F(x_k)$ is used explicitly in (1.2) for each $k$. Although this leaves us considering a somewhat artificial situation, to do otherwise would cloud the issues of interest here. We hope that the reader will forgive this and occasional other uses of accurate $F$-values below which might be regarded as unrealistic. In future work, we shall apply the results of this paper together with those of Dennis and Walker (1983) to analyze the effects of inaccuracy from all sources on the performance of methods employing sparse secant updates, including Broyden updates.

On the basis of the analysis given in Section 2, we suggest the secant conditions

$$\bar{y}_k = \arg\min_{y \in B(A_k s_k, \hat{y}_k)} |y - \tilde{y}_k|$$

(4.1)

for $k=0,1,2,\ldots$, where $\hat{y}_k = \hat{y}(A_k s_k, Y_k)$ is determined as in (2.10) with $v = A_k s_k$ and $S = Y_k$. According to Corollary 2.6, a method (1.2) employing sparse secant updates determined by secant conditions (4.1) will enjoy local $q$-linear convergence. These secant conditions are certainly easy to obtain and also have the heuristically attractive property that in most applications, in which $\eta_k$ can be taken very small, one will very likely have $\bar{y}_k = \tilde{y}_k$, except perhaps very near the solution. This is to say that $\bar{y}_k$ will most likely be the traditional secant condition in the great majority of instances. In light of the additional analytical appeal of these secant conditions outlined below, we regard this as evidence in favor of the traditional secant conditions.
Let us show how one might apply the results of Section 3 to determine asymptotic speeds of convergence which can be obtained by using the secant conditions $y_k$. For this demonstration, suppose that there is given a relative bound $|N(x)| \leq \rho |F(x)|$, $x \in \Omega$, on the inaccuracy in computed values of $F$ used in determining secant conditions. Let $\{x_k\}$ be a q-linearly convergent iteration sequence. For each $k$, we have

$$|y_k - y_k| = |N(x_{k+1}) - N(x_k)|$$

$$\leq \rho (|F(x_{k+1})| + |F(x_k)|).$$

Consequently, one can take $\eta_k = \rho (|F(x_{k+1})| + |F(x_k)|)$; and since

$$|y_k - y_k| \leq (1 + \sqrt{n}) \eta_k,$$

one has

$$\lim_{k \to \infty} \frac{|y_k - y_k|}{|x_k - x_*|} \leq (1 + \sqrt{n}) \rho |F'(x_*)|^2 \left(1 + \lim_{k \to \infty} \frac{|x_k - x_*|}{|x_k - x_*|}\right).$$

In the case at hand, $B_* = F'(x_*)$ and $r_* = 0$; and so it follows from (3.15) and (3.16) of Corollary 3.3 that

$$\lim_{k \to \infty} \frac{|x_{k+1} - x_*|}{|x_k - x_*|} \leq \lim_{k \to \infty} \frac{|F'(x_*)^{-1} (y_k - y_k)|}{|x_k - x_*|}$$

$$\leq (1 + \sqrt{n}) \rho K_* \left(1 + \lim_{k \to \infty} \frac{|x_k - x_*|}{|x_k - x_*|}\right),$$

where $K_* = |F'(x_*)^{-1}||F'(x_*)|$ is the condition number of $F'(x_*)$ in the norm $|\cdot|$. If $(1 + \sqrt{n}) \rho K_* < 1$, then one obtains from this the estimate

$$\lim_{k \to \infty} \frac{|x_{k+1} - x_*|}{|x_k - x_*|} \leq \frac{(1 + \sqrt{n}) \rho K_*}{1 - (1 + \sqrt{n}) \rho K_*} \quad (4.2)$$

of the asymptotic speed of convergence.
The estimate (4.2) and the underlying analysis suggest that if one has some control over the amount of inaccuracy in computed values of $F$ used in determining secant conditions, then the results of Section 3 might be used to provide guidelines for exercising this control in order to achieve a reasonable balance of convergence speed and computational expense. To begin with, one sees from (4.2) that under our assumptions, a q-linearly convergent iteration sequence $\{x_k\}$ will exhibit asymptotic q-linear convergence which is as fast as desired, provided the relative inaccuracy in computed function values used to determine secant conditions is kept sufficiently small. However, in most applications, high accuracy in function evaluations can be achieved only at considerable computational cost. In such applications, (4.2) might be helpful in finding an appropriate compromise between fast convergence and less expensive function evaluations.

In some applications, it might be too costly to maintain a uniform level of accuracy in computed function values at all iterations which is sufficiently high for (4.2) to guarantee adequately fast asymptotic convergence. In the introduction, we remarked that a natural strategy in such applications is to use relatively inaccurate computed function values in the early iterations and then to increase function evaluation accuracy in the later stages as the solution is neared. The analysis leading to (4.2) bears out the validity of this strategy in the case under consideration. To see that this is so, suppose that for each $k$, the relative inaccuracy in the computed value of $F(x_k)$ is bounded by $\rho_k$. Then by a derivation similar to that of (4.2) one can show that a q-linearly convergent iteration sequence $\{x_k\}$ satisfies
\[
\lim \frac{|x_{k+1} - x_*|}{|x_k - x_*|} \leq \frac{(1 + \sqrt{n}) \lim_{k \to \infty} \rho_k K_*}{1 - (1 + \sqrt{n}) \lim_{k \to \infty} \rho_k K_*}.
\] (4.3)

Note that (4.3) implies that \( \{x_k\} \) converges q-superlinearly if \( \lim_{k \to \infty} \rho_k = 0 \).

Although (4.3) provides a satisfactory guarantee that adequately fast asymptotic convergence will result whenever \( \lim_{k \to \infty} \rho_k \) is sufficiently small, there remains the question of how one should choose \( \rho_k \) for each \( k \). We conclude this illustration by offering a manner of choosing each \( \rho_k \) which is suggested by Corollary 3.2. Suppose that an initial \( \rho_0 > 0 \) and suitable values \( \tau \in (0,1) \) and \( \rho_{\text{min}} > 0 \) are given at the outset. If one has arrived at the \( k \)th iteration for some \( k > 0 \), then take \( \rho_k = \rho_{k-1} \) as a trial value and set \( \eta_k = \rho_k \{ |F(x_{k+1})| + |F(x_k)| \} \). Even though the matrix \( T_k = \text{diag}(t_{1k}, \ldots, t_{nk}) \) of (3.3) and (3.4) appropriate for \( y_k = y_k \) cannot be observed, if

\[
|\epsilon_{k}^{T} B_{k} s_{k} - \epsilon_{k}^{T} y_{k}^{*} | \geq (1 + \frac{1}{\tau}) \eta_k, \quad 1 = 1, \ldots, n,
\] (4.4)

then one necessarily has \( 0 \leq |t_{ik}| \leq \tau \) for \( i = 1, \ldots, n \), where \( t_{ik} \) is as in Theorem 3.1 and Corollary 3.2; furthermore, \( y_k = y_k \) in this case. If (4.4) holds, then accept \( \rho_k \) and \( y_k = y_k \) and proceed. Otherwise, reduce \( \rho_k \) and redefine \( \eta_k \) until either (4.4) holds or \( \rho_k = \rho_{\text{min}} \). Note that without the restriction \( \rho_k \geq \rho_{\text{min}} \), (4.4) will hold for sufficiently small \( \rho_k > 0 \) except in the unlikely event that \( B_{k} s_{k} = y_k^{*} \). The restriction \( \rho_k \geq \rho_{\text{min}} \) guards against this event and also prevents requiring excessive accuracy in function evaluations.

This manner of choosing each \( \rho_k \) has the appeal that if \( \{x_k\} \) is assumed to converge q-linearly to \( x_* \), then unless or until \( \rho_k = \rho_{\text{min}} \) for
some $k$, one can use the traditional secant conditions and presumably be on the way to enjoying the $q$-superlinear convergence guaranteed by Corollary 3.2 in this case. Actually, it always happens in this case that $\rho_k = \rho_{\text{min}}$ eventually, as we show in Proposition 3.4 below. However, this should be alright, because $\rho_{\text{min}}$ can be chosen so small in most applications that the asymptotic speed of convergence given by (4.3) with $\lim_{k \to \infty} \rho_k = \rho_{\text{min}}$ will be adequately fast.

**Proposition 4.1.** If $\{x_k\}$ converges $q$-linearly to $x_*$ in this case, then $\rho_k = \rho_{\text{min}}$ for sufficiently large $k$.

**Proof:** Suppose the contrary, i.e., that $\rho_k > \rho_{\text{min}}$ for all $k$. Then (4.4) holds for all $k$, and it follows that

$$(1 + \frac{1}{n}) \lim_{k \to \infty} \frac{n_k}{|s_k|} \leq \lim_{k \to \infty} \frac{|e_i B_k s_k - e_i y_k|}{|s_k|}$$

(4.5)

for $i = 1, \ldots, n$. It also follows from Corollary 3.2 that $\{x_k\}$ converges $q$-superlinearly to $x_*$, and so

$$0 = \lim_{k \to \infty} \frac{[B_k - F'(x_*)]s_k}{|s_k|}$$

$$= \lim_{k \to \infty} \frac{(B_k s_k - y_k)}{|s_k|}.$$  

See Dennis and More (1974) or Dennis and Walker (1981). Consequently, for $i = 1, \ldots, n$,

$$\lim_{k \to \infty} \frac{|e_i B_k s_k - e_i y_k|}{|s_k|} \leq \lim_{k \to \infty} \frac{|B_k s_k - y_k|}{|s_k|} + \lim_{k \to \infty} \frac{|e_i y_k - e_i y_k|}{|s_k|}$$

(4.6)

$$\leq \lim_{k \to \infty} \frac{n_k}{|s_k|}.$$
Since \( \{x_k\} \) converges q-linearly to \( x_* \) and the \( p_k \)'s are bounded away from 0, one has that

\[
0 < \lim_{k \to \infty} \frac{n_k}{|s_k|} < \infty.
\]

It follows that (4.5) and (4.6) are contradictory, and the proposition is proved.

**Appendix**

In this appendix, we establish the result referenced in the proofs of Corollaries 3.2 and 3.3. This result is an extension of Theorem A3.1 of Dennis and Walker (1981), and we suspect that it might have uses beyond those in this paper. Our interest here is in a general quasi-Newton iteration (1.2) for solving the problem (1.1) with no presumption about the form of each \( B_k \) or the manner in which it is generated. We suppose that \( F \) satisfies the standard hypothesis given in the introduction, although we have no particular interest in a computed part \( C \) of \( F' \). As in the main body of the paper, we use \( |\cdot| \) to denote both a vector norm on \( \mathbb{R}^n \) and its subordinate operator norm on \( \mathbb{R}^{nxn} \). For an iteration sequence \( \{x_k\} \), we denote \( e_k = x_{k+1} - x_k \) as before; and it is also convenient to set \( e_k = x_k - x_* \) for each \( k \).

**Theorem** A.1. Suppose that \( F \) satisfies the standard hypothesis and that \( \{x_k\} \) is a sequence generated by (1.2) which converges to \( x_* \) with \( x_k \approx x_* \) for all \( k \). If \( B_* \in \mathbb{R}^{nxn} \) is any invertible matrix, then for any norm \( |\cdot| \), one has

\[
1 \lim_{k \to \infty} \frac{|e_{k+1}|}{|e_k|} - [I - B_*^{-1} F'(x_*)] \frac{e_k}{|e_k|} + \frac{B_*^{-1}(B_k - B_*) s_k}{|e_k|} = 0. \tag{A.1}
\]
Setting $r_\ast = |I - B_\ast^{-1}F'(x_\ast)|$, one has in particular that

$$\lim_{k \to \infty} \frac{|e_{k+1}|}{|e_k|} \leq \lim_{k \to \infty} \left| \frac{e_k}{|e_k|} + \frac{B_\ast^{-1}(B_k - B_\ast)s_k}{|e_k|} \right|$$

$$\leq r_\ast + \lim_{k \to \infty} \frac{|B_\ast^{-1}(B_k - B_\ast)s_k|}{|e_k|} \leq r_\ast + \lim_{k \to \infty} \frac{|B_\ast^{-1}(B_k - B_\ast)s_k|}{|s_k|} \lim_{k \to \infty} \frac{|s_k|}{|e_k|}. \tag{A.2}$$

**Proof:** From the standard hypothesis on $F$ and (1.2), one has that

$$B_\ast s_k = -F(x_k) = -F'(x_\ast)e_k + O(e_k^{p+1}).$$

It follows that

$$s_k + B_\ast^{-1}F'(x_\ast)e_k + B_\ast^{-1}(B_k - B_\ast)s_k = O(e_k^{p+1}).$$

With $e_k \neq 0$, one verifies from this equation that

$$\frac{e_{k+1}}{|e_k|} - \left[ I - B_\ast^{-1}F'(x_\ast) \right] \frac{e_k}{|e_k|} + \frac{B_\ast^{-1}(B_k - B_\ast)s_k}{|e_k|} = O(e_k^p),$$

and (A.1) and (A.2) follow immediately.
References.


