MATHEMATICS OF COMPUTATION
VOLUME 42, NUMBER 166
APRIL 1984, PAGES 521–533

On the Sparse and Symmetric Least-Change Secant Update*

By Trond Steihaug**

Abstract. To find the sparse and symmetric n by n least-change secant update we have to solve a consistent linear system of n equations in n unknowns, where the coefficient matrix is symmetric and positive semidefinite. We give bounds on the eigenvalues of the coefficient matrix and show that the preconditioned conjugate gradient method is a very efficient method for solving the linear equation. By solving the linear system only approximately, we generate a family of sparse and symmetric updates with a residual in the secant equation. We address the question of how accurate a solution is needed not to impede the convergence of quasi-Newton methods using the approximate least-change update. We show that the quasi-Newton methods are locally and superlinearly convergent after one or more preconditioned conjugate gradient iterations.

1. Introduction. Quasi-Newton methods have proved themselves in dealing with the unconstrained minimization problem: find x_* so that for some $\varepsilon > 0$

$$f(x_*) \leqslant f(x), \quad \forall x: \|x - x_*\| < \varepsilon,$$

where f: $\mathbf{R}^n \to \mathbf{R}$ is a smooth function. Quasi-Newton methods approximate the solution x_* by generating a sequence of iterates $\{x_k\}$ as follows:

(1.1)
Given x_0, and B_0
FOR $k = 0$ STEP 1 UNTIL Convergence DO
  Solve $B_k s_k = -\nabla f(x_k)$
  Set $x_{k+1} = x_k + s_k$
  Update to obtain B_{k+1}.

The basic assumption for quasi-Newton methods is that in a neighborhood of x_*

$$B_k s_k + \nabla f(x_k) \approx \nabla f(x_k + s_k). \tag{1.2}$$

So solving the linear system (1.1) may not be justified when the approximation (1.2) is not accurate, which may occur when x_k is far from the solution x_* or when B_k is an approximation to the Hessian matrix of f at x_k. Instead Steihaug [12] introduces the inexact quasi-Newton method which only approximately solves (1.1) in some unspecified manner. In the inexact quasi-Newton method, we accept an approximate solution s_k of (1.1) if a relative residual is less than a tolerance θ_k that may depend

Received February 24, 1982; revised May 24, 1983.
1980 *Mathematics Subject Classification*. Primary 65K05, 65F35; Secondary 65F15.
* This research was supported by the Norwegian Research Council for Science and the Humanities and Yale University Graduate School.
** *Current address*. Statoil, Stavanger, Norway.

0025-5718/84 $1.00 + $.25 per page

on x_k. The inexact quasi-Newton methods generate the sequence of iterates as follows:

Given x_0, and B_0
FOR $k = 0$ STEP 1 UNTIL Convergence DO
Find *some* s_k so that for $r_k = B_k s_k + \nabla f(x_k)$, then

$$\frac{\|r_k\|}{\|\nabla f(x_k)\|} \leqslant \theta_k \tag{1.3}$$

Set $x_{k+1} = x_k + s_k$
Update to obtain B_{k+1}.

If B_k is the Hessian matrix of f at x_k, then we have an inexact Newton method [4].

Hessian information of f is incorporated in the approximations $\{B_k\}$ by requiring that

$$B_{k+1}s_k = \nabla f(x_k + s_k) - \nabla f(x_k) \equiv y_k, \tag{1.4}$$

which forces the new approximation of the Hessian matrix to satisfy (1.2) with equality, i.e., $B_{k+1}s_k + \nabla f(x_k) = \nabla f(x_k + s_k)$. Updates that satisfy (1.4) are called secant updates and (1.4) is called the secant equation. Since B_k is an approximation of the symmetric Hessian matrix of f, it is natural to require that $\{B_k\}$ are symmetric matrices. In large scale optimization the variables or groups of variables are only weakly connected in the sense that $\partial f/\partial x_i$ depends only on a few variables, i.e.,

$$\frac{\partial^2 f(x)}{\partial x_i \partial x_j} = 0 \quad \text{for all } x \tag{1.5}$$

for most j. A sparsity structure K of the Hessian matrix of f is a set of indices so that if $(i,j) \notin K$, then (1.5) holds for all x. We note that if there exists x so that the Hessian matrix of f at x is positive definite, then $(i,i) \in K$, $i = 1,2,\ldots,n$. We assume in the following that $(i,i) \in K$, $i = 1,2,\ldots,n$. Further, we assume that K preserves the symmetry, i.e., if $(i,j) \in K$, then $(j,i) \in K$. By requiring that B_{k+1} should preserve the symmetric and sparse structure, we hope to reduce the number of arithmetic operations to find s_k so that (1.3) holds, get a better approximation of the Hessian matrix and reduce the computer storage required to store the approximation.

Marwil [9] and Toint [15] derived a sparse and symmetric update of the form

$$(B_+)_{ij} = \begin{cases} B_{ij} + u_i s_j + s_i u_j, & (i,j) \in K, \\ 0, & \text{otherwise}, \end{cases} \tag{1.6}$$

that satisfies (1.4) for some $u \in \mathbf{R}^n$. We have eliminated the subscript k referring to the iteration number, and we let B_+ denote the new update. However, to find the update B_+ we have to solve a consistent linear system of equations

$$Gu = b, \tag{1.7}$$

where G is an n by n symmetric and positive semidefinite matrix with the same sparsity structure as the update. In this paper, we address the question of how accurate a solution of (1.7) is needed in order not to impede the convergence. By

solving (1.7) only approximately and using an update of the form (1.6), we generate the family of updates

$$B(u)_{ij} = \begin{cases} B_{ij} + u_i s_j + s_i u_j, & (i,j) \in K, \\ 0, & \text{otherwise}, \end{cases} \tag{1.8}$$

where u is the approximate solution of (1.7).

In conjunction with the sparsity structure K we also define an operator Z: $\mathbf{R}^{n\times n} \to \mathbf{R}^{n\times n}$ so that

$$Z(m)_{ij} \equiv \begin{cases} M_{ij} & \text{when } (i,j) \in K, \\ 0, & \text{otherwise.} \end{cases} \tag{1.9}$$

For a given vector $s \in \mathbf{R}^n$ and a sparse symmetric matrix $B \in \mathbf{R}^n$, the updates in the family (1.8) are given by

$$B(u) \equiv B + Z(us^T + su^T), \tag{1.10}$$

where we assume that $Z(B) = B$.

In Section 2, we discuss some basic properties of the updates. We show that the preconditioned conjugate gradient method is an efficient method for finding an approximate solution of (1.7).

In Section 3, we discuss local and superlinear convergence results of the inexact quasi-Newton using updates in the family. In the last section we briefly discuss global convergence.

2. Basic Properties of the Updates. If we require that $B(u)$ in (1.10) satisfies the secant equation

$$B(u)s = y, \tag{2.1}$$

where s and y are given vectors, then u has to satisfy

$$0 = y - Bs - Z(us^T)s - Z(su^T)s = y - Bs - Du - Z(ss^T)u,$$

where D is a diagonal matrix with elements

$$D_{ii} \equiv \sum_{j:(i,j)\in K} s_j^2, \qquad i = 1,\ldots,n. \tag{2.2}$$

Put

$$b \equiv y - Bs, \quad \text{and} \quad G \equiv D + Z(ss^T), \tag{2.3}$$

and we have that u has to satisfy the equation (1.7) where b and G are given in (2.3).

If $D_{ii} = 0$, then the ith component of $B(u)s$ is zero and $b_i = y_i$, so if $y_i \neq 0$, we see that no update in the family will satisfy the secant equation (2.1). However, if f is twice continuously differentiable, let $M \in \mathbf{R}^{n\times n}$ be given by

$$M_{ij} = \int_0^1 \frac{\partial^2 f(x+\tau s)}{\partial x_i \partial x_j} d\tau, \qquad i,j = 1,2,\ldots,n. \tag{2.4}$$

Let y be chosen as

$$y \equiv \nabla f(x+s) - \nabla f(x). \tag{2.5}$$

Then we have from Ortega and Rheinboldt [10, 3.2.6] that $y = Ms$ and $y_i = 0$ whenever $D_{ii} = 0$, so the system (1.7) is consistent for this choice of y.

For the given sparsity structure K of the Hessian matrix of f and vector s in $\mathbf{R}^n$, let y be as in (2.5), and define the affine space

$$V \equiv \left\{ B \in \mathbf{R}^{n\times n} : Bs = y,\ B^T = B, \text{ and } B_{ij} = 0\ \forall (i,j) \notin K \right\}. \tag{2.6}$$

From the above discussion, we note that if f is twice continuously differentiable in $\mathbf{R}^n$, then the set V is nonempty. In the following we assume that V is nonempty.

Define the quadratic function

$$q(u) \equiv \tfrac{1}{2}u^T G u - b^T u, \tag{2.7}$$

where b and G are given in (2.3).

LEMMA 2.1. *If $M \in V$, then for all $u \in \mathbf{R}^n$*

$$\|B(u) - M\|_F^2 = \|B - M\|_F^2 + 4q(u). \tag{2.8}$$

Proof. Let $u \in \mathbf{R}^n$. Then

$$\begin{aligned} \|Z(us^T + su^T)\|_F^2 &= \sum_{(i,j)\in K} (s_i u_j + u_i s_j)^2 \\ &= 2u^T D u + 2 \sum_{(i,j)\in K} s_i u_j u_i s_j = 2u^T G u. \end{aligned} \tag{2.9}$$

Let $M \in V$. Hence $Ms = y$ and

$$\begin{aligned} \|B(u) - M\|_F^2 &= \|B - M\|_F^2 + \|Z(us^T + su^T)\|_F^2 + 4 \sum_{(i,j)\in K} u_i (B - M)_{ij} s_j \\ &= \|B - M\|_F^2 + 2u^T G u - 4u^T (y - Bs) \end{aligned}$$

using (2.9) and (2.3). The desired equality (2.8) follows from the defintion of q in (2.7). Q.E.D.

It follows directly from (2.9) that $u^T G u$ is bounded below, hence G is positive semidefinite, and from (2.8) it follows that $q(u)$ is bounded below, so b is in the range space of G. We can thus minimize $q(u)$ to find a $B(u)$ close to the affine space V.

COROLLARY 2.2. *Let $\bar{u}$ be a minimizer of $q(u)$. Then $\bar{B} = B(\bar{u})$ is the least change update*

$$\|\bar{B} - B\|_F = \min\{\|\tilde{B} - B\|_F : \tilde{B} \in V\}. \tag{2.10}$$

Proof. If $\bar{u}$ is a minimizer of $q(u)$, then $\bar{u}$ is a solution of $Gu = b$ and $B(\bar{u}) \in V$. By choosing $M = B(\bar{u})$, we have from (2.8) that $\|B(\bar{u}) - B\|_F^2 = -4q(\bar{u})$. Hence for all $M \in V$ we have, using (2.8),

$$\|B(\bar{u}) - B\|_F^2 = \|B - M\|_F^2 - \|B(\bar{u}) - M\|_F^2 \leqslant \|B - M\|_F^2.$$

Since $B(\bar{u}) \in V$, we have $B(\bar{u})$ is a least-change update. The Frobenius norm is strictly convex, so $B(\bar{u})$ is the unique solution of (2.10). Q.E.D.

The update $B(\bar{u})$ was shown by Toint [15] to be the least-change update, and it follows from the general theory in Dennis and Schnabel [7] that the update is of the form (1.6).

We want to solve the linear system (1.7) with an iterative method. At each iteration we want to make $q(u)$ as small as possible since this will make the update $B(u^{k+1})$ close to V. An appealing iterative method for minimizing $q(u)$ is a

preconditioned conjugate gradient method (see for example Axelsson [1]) since these are optimal over a large class of iterative methods.

Let $(\cdot,\cdot)$ be the standard innerproduct on $\mathbf{R}^n$, i.e., for $d,p \in \mathbf{R}^n$. Then $(d,p) = d^T p$. A preconditioned conjugate gradient method induced by the diagonal matrix (2.2) is

The PCG method:

Let $u^0 = 0$, $r^0 = y - Bs$, and $d^0 = D^+ r^0$

FOR $k = 0$ STEP 1 UNTIL Convergence DO

$$u^{k+1} = u^k + \alpha_k d^k, \qquad \alpha_k = \frac{(r^k, D^+ r^k)}{(d^k, Gd^k)},$$

$$r^{k+1} = r^k + \alpha_k G d^k,$$

$$d^{k+1} = D^+ r^{k+1} + \beta_k d^k, \qquad \beta_k = \frac{(r^{k+1}, D^+ r^{k+1})}{(r^k, D^+ r^k)},$$

where D^+ is the pseudoinverse of D

$$D_{ii}^+ = \begin{cases} \dfrac{1}{D_{ii}} & \text{when } D_{ii} > 0, \\ 0, & \text{otherwise.} \end{cases}$$

From Björk and Elfving [2] we know that when we apply the conjugate gradient (CG) method to

$$D^{+1/2} G D^{+1/2} \tilde{u} = \tilde{b},$$

where $\tilde{u} = D^{1/2} u$ and $\tilde{b} = D^{+1/2} b$ with starting point $\tilde{u}_0 = 0$, the iterates converge to the least-norm solution

$$\tilde{u} = [D^{+1/2} G D^{+1/2}]^+ \tilde{b}.$$

Since v is in the range space of D if and only if v is in the range space of D^+, we can go back to the untransformed variables in the CG method, and the resulting algorithm is the PCG method [1].

The efficiency of the method depends on the ratio of the largest and smallest positive eigenvalue of $D^+ G$ [1], and it follows from the next theorem that this ratio is bounded by the maximum number of nonzero elements in any row of B (or G) and is thus independent of B, s, and y. The maximum eigenvalue of $D^+ G$ is the maximum of $(d,Gd)/(d,Dd)$. From the theorem we have $(d,Dd) = 0$ if and only if $(d,Gd) = 0$, so the minimum positive eigenvalue of $D^+ G$ is the minimum of $(d,Gd)/(d,Dd)$ for d so that $(d,Dd) \neq 0$.

THEOREM 2.3. *Let the number of nonzero elements in each row be $\leqslant m$. Then for all $d \in \mathbf{R}^n$*

$$\frac{2}{m}(d,Dd) \leqslant (d,Gd) \leqslant 2(d,Dd). \tag{2.11}$$

Proof. We first show the upper bound $(d,Gd) \leqslant 2(d,Dd)$. Let e^i be the ith unit vector. Consider

$$D - Z(ss^T) = \frac{1}{2} \sum_{(i,j)\in K} \left(e^i s_j - e^j s_i\right)\left(e^i s_j - e^j s_i\right)^T.$$

Since each term in the sum is a symmetric positive semidefinite rank one matrix, we have

$$\left(d,\left[D - Z(ss^T)\right]d\right) \geqslant 0,$$

and

$$(d,Gd) = (d,Dd) + \left(d,Z(ss^T)d\right) \leqslant 2(d,Dd),$$

which gives the upper bound in (2.11). If A is a diagonal matrix in $\mathbf{R}^{n\times n}$ with diagonal elements $a_1,a_2,\ldots,a_n$, then we write

$$A = \operatorname{diag}(a_i).$$

Consider

$$G = D + Z(ss^T) = 2\operatorname{diag}\left(s_i^2\right) + \frac{1}{2}\sum_{\substack{(i,j)\in K\\ i\neq j}}\left(e^i s_j + e^j s_i\right)\left(e^i s_j + e^j s_i\right)^T \tag{2.12}$$

and

$$Z(ss^T) = \operatorname{diag}\left((2 - m_i)s_i^2\right) + \frac{1}{2}\sum_{\substack{(i,j)\in K\\ i\neq j}}\left(e^i s_i + e^j s_j\right)\left(e^i s_i + e^j s_j\right)^T, \tag{2.13}$$

where m_i, $i = 1,2,\ldots,n$, is the number of nonzero elements in row i (including the diagonal element), i.e.,

$$m_i = \sum_{j:(i,j)\in K} 1.$$

Let m be the maximum number nonzero elements in any row,

$$m = \max\{\, m_i : i = 1,\ldots,n\,\},$$

and assume that $m \geqslant 2$. We show that the matrix $mG - 2D$ is positive semidefinite. From (2.12) and (2.13) we have

$$mG - 2D = (m-2)G + 2Z(ss^T) = 2\operatorname{diag}\left((m - m_i)s_i^2\right) + \frac{m-2}{2}\sum_{\substack{(i,j)\in K\\ i\neq j}}\left(e^i s_j + e^j s_i\right)\left(e^i s_j + e^j s_i\right)^T + \sum_{\substack{(i,j)\in K\\ i\neq j}}\left(e^i s_i + e^j s_j\right)\left(e^i s_i + e^j s_j\right)^T.$$

Since each term in the two sums is a symmetric positive semidefinite rank one matrix, we have

$$\left(d,\left(G - \frac{2}{m}D\right)d\right) \geqslant \frac{2}{m}\sum_{i=1}^{n}(m - m_i)s_i^2 d_i^2 \geqslant 0$$

since $m \geqslant m_i$ for all i. The case $m = 1$ follows immediately. Q.E.D.

If the matrix is full, i.e., $K = \{(i,j): 1 \leqslant i,j \leqslant n\}$, then the lower bound in (2.11) is $(2/n)(d,Dd)$. However, it is easily seen that $(d,Dd) \leqslant (d,Gd)$. We now consider

the case when the minimum number of nonzero elements in each row is larger than $(n/2) + 1$. In this case we can improve the lower bound. Consider

$$Z(ss^T) - ss^T = \frac{1}{2} \sum_{(i,j) \notin K} \left(e^i s_i - e^j s_j\right)\left(e^i s_i - e^j s_j\right)^T - \operatorname{diag}\left((n - m_i)s_i^2\right). \tag{2.14}$$

Let

$$\bar{m} = \max\{ n - m_i + 2 : i = 1,2,\ldots, n\}.$$

We now show that $\bar{m}G - 2(D + ss^T)$ is positive semidefinite. From (2.12) and (2.14) we have

$$\begin{aligned} \bar{m}G - 2(D + ss^T) &= (\bar{m} - 2)G + 2\left(Z(ss^T) - ss^T\right) \\ &= 2(\bar{m} - 2)\operatorname{diag}\left(s_i^2\right) - 2\operatorname{diag}\left((n - m_i)s_i^2\right) \\ &\quad + \frac{\bar{m} - 2}{2} + \sum_{\substack{(i,j) \in K \\ i \neq j}} \left(e^i s_i + e^j s_j\right)\left(e^i s_i + e^j s_j\right)^T \\ &\quad + \sum_{(i,j) \notin K} \left(e^i s_i - e^j s_j\right)\left(e^i s_i - e^j s_j\right)^T. \end{aligned}$$

Since the terms in the sums are symmetric and positive semidefinite matrices, we have

$$\begin{aligned} \left(d, \left(G - \frac{2}{\bar{m}} D\right)d\right) &= \left(d, \left(G - \frac{2}{\bar{m}}(D + ss^T)\right)d\right) + \frac{2}{\bar{m}}(d,s)^2 \\ &\geqslant \frac{2}{\bar{m}} \sum_{i=1}^{n} (\bar{m} - 2 - n + m_i)s_i^2 d_i^2 \geqslant 0, \end{aligned}$$

and we have the lower bound

$$(d, Gd) \geqslant \frac{2}{\bar{m}}(d, Dd). \tag{2.15}$$

We note that if the matrix is full, then $\bar{m} = 2$ and (2.15) is sharp. We now show that the upper bound in (2.11) is sharp. From (2.3) $Gs = 2Ds$, so the upper bound is achieved for $d = s$. We now give an example of a sparsity structure and vector s so that the lower bound is achieved. Consider a tridiagonal matrix with one element in each corner. Then the sparsity structure K is given by

$$\begin{aligned} K = \{ &(1,1), (1,2), (1,n), (n,1), (n,n-1), (n,n), \\ &(i,i-1), (i,i), (i,i+1) : 1 < i < n\} \end{aligned}$$

for even $n \geqslant 4$ and $s^T = (1,\ldots. 1)$. Then $D_{ii} = 3$ for $i = 1,2,\ldots, n$. The eigenvalues of G are

$$\lambda_k = 4 + 2\cos\left(\frac{2k\pi}{n}\right) \quad \text{for } k = 1,2,\ldots, n,$$

with eigenvector v^k, where the jth component, $j = 1,\ldots, n$, is $v_j^k = \sin(2jk\pi/n)$ for $k = 1,\ldots, n-1$ and $v_j^n = 1$. We now have the lower bound

$$(d, Gd) \geqslant \min_{1 \leqslant k \leqslant n} \lambda_k (d,d) = 2(d,d) = 3\frac{2}{m}(d,d).$$

Marwil [9] and Toint [15] have shown that G is positive definite when all null rows and columns are eliminated. Dennis and Schnabel [7] give the bound

$$2 \min_{1 \leq i \leq n} s_i^2(d,d) \leq (d,Gd) \leq 2(s,s)(d,d)$$

if no s_i are zero.

We can conclude from Theorem 2.3 that when we use the preconditioned conjugate gradient method then [1]

$$\|B(u^k) - B(\bar{u})\|_F \leq 2\left(\frac{\sqrt{m}-1}{\sqrt{m}+1}\right)^k \|B - B(\bar{u})\|_F.$$

In the next section, we discuss how accurate a solution is needed to achieve local and superlinear convergence.

3. Local Convergence Results. In this section, we discuss local convergence results for inexact quasi-Newton methods where the new update B_{k+1} is found using updates from Section 2. We first discuss convergence based on the bounded deterioration condition [3]. Let H denote the Hessian matrix of f.

LEMMA 3.1. *Let f be twice continuously differentiable in an open neighborhood Ω of a point x_*, and let $L \geq 0$, $0 < p \leq 1$, be such that for all $x \in \Omega$*

$$\|H(x) - H(x_*)\|_F \leq L\|x - x_*\|^p, \tag{3.1}$$

where $\|\cdot\|$ is a vectornorm. Let x, $x + s \in \Omega$, and y given in (2.5). *If $q(u) \leq 0$, then*

$$\|B(u) - H(x_*)\|_F \leq \|B - H(x_*)\|_F + 2L\sigma(x,x+s)^p, \tag{3.2}$$

where

$$\sigma(x,z) = \max\{\|x - x_*\|, \|z - x_*\|\}. \tag{3.3}$$

Proof. From (2.9) we have that if $q(u) \leq 0$, then

$$\|B(u) - M\|_F \leq \|B - M\|_F. \tag{3.4}$$

Let

$$M = \int_0^1 H(x + \tau s)\, d\tau. \tag{3.5}$$

Then from Ortega and Rheinboldt [10, 3.2.6] we have, using (2.6), that $M \in V$. From (3.1) we have

$$\begin{aligned} \|M - H(x_*)\|_F &\leq \int_0^1 \|H(x+\tau s) - H(x_*)\|_F d\tau \\ &\leq \sup_{0 \leq \tau \leq 1} \|H(x+\tau s) - H(x_*)\|_F \\ &\leq L \sup_{0 \leq \tau \leq 1} \|x + \tau s - x_*\|^p \leq L\sigma(x,x+s)^p \end{aligned} \tag{3.6}$$

using Ortega and Rheinboldt [10, 3.2.11] and the definition of σ in (3.3). Consider

$$\begin{aligned} \|B(u) - H(x_*)\|_F &\leq \|B(u) - M\|_F + \|M - H(x_*)\|_F \\ &\leq \|B - M\|_F + L\sigma(x,x+s)^p \\ &\leq \|B - H(x_*)\|_F + 2L\sigma(x,x+s)^p \end{aligned}$$

using the triangle inequality, (3.4), and (3.6) twice. Q.E.D.

LEMMA 3.2. *Suppose that the hypotheses of Lemma* 3.1 *hold, and let the sequence* $\{x_k\}$ *be in* Ω *and satisfy*

$$\sum_{k=0}^{\infty} \sigma(x_k, x_{k+1})^p < \infty. \tag{3.7}$$

Let $B_{k+1} = B_k(u_k)$ *using* $s_k = x_{k+1} - x_k$ *and* y_k *in* (1.4), *and let* b_k *be defined as in* (2.3). *If* $\beta > 0$ *and*

$$q_k(u_k) \leqslant -\beta \frac{(b_k, b_k)}{(s_k, s_k)}, \tag{3.8}$$

then

$$\lim_{k \to \infty} \frac{\|b_k\|}{\|s_k\|} = 0. \tag{3.9}$$

Proof. The proof follows the technique of Broyden, Dennis and Moré [3] and Dennis and Moré [6]. Let M_k be given by (3.5), and let

$$\beta_k^2 = 4\beta \frac{(b_k, b_k)}{(s_k, s_k)}.$$

Consider

$$\begin{aligned} \|B_{k+1} - M_k\|_F^2 &= \|B_k - M_k\|_F^2 + 4q_k(u_k) \leqslant \|B_k - M_k\|_F^2 - \beta_k^2 \\ &\leqslant (\|B_k - M_{k-1}\|_F + \|M_k - M_{k-1}\|_F)^2 - \beta_k^2 \end{aligned}$$

using (3.8) and the triangle inequality. From (3.6) we have

$$\begin{aligned} \|M_k - M_{k-1}\|_F &\leqslant \|M_k - H(x_*)\|_F + \|H(x_*) - M_{k-1}\|_F \\ &\leqslant L\big(\sigma(x_{k+1}, x_k)^p + \sigma(x_k, x_{k-1})^p\big). \end{aligned}$$

Put $\rho_i = \|B_i - M_{i-1}\|_F$. Then

$$\rho_{k+1}^2 \leqslant \big[\rho_k + L\big(\sigma(x_{k+1}, x_k)^p + \sigma(x_k, x_{k-1})^p\big)\big]^2 - \beta_k^2.$$

In view of the inequality

$$(a^2 - b^2)^{1/2} \leqslant a - \frac{b}{2a} \quad \text{for } 0 < b \leqslant a,$$

we have

$$\begin{aligned} \rho_{k+1} \leqslant \rho_k &+ L\big(\sigma(x_{k+1}, x_k)^p + \sigma(x_k, x_{k-1})^p\big) \\ &- \frac{\beta_k^2}{2\big[\rho_k + L\big(\sigma(x_{k+1}, x_k)^p + \sigma(x_k, x_{k-1})^p\big)\big]}. \end{aligned} \tag{3.10}$$

From (3.2)

$$\|B_k - H(x_*)\|_F \leqslant \|B_0 - H(x_*)\|_F + 2L \sum_{i=0}^{k-1} \sigma(x_{i+1}, x_i)^p,$$

and from (3.7) we have that $\{\|B_k\|\}$ is bounded, hence ρ_k is bounded and we have

$$L\big(\sigma(x_{k+1}, x_k)^p + \sigma(x_k, x_{k-1})^p\big) + \rho_k \leqslant \rho < \infty. \tag{3.11}$$

Rearranging (3.10) and using (3.11), we have

$$\frac{\beta_k^2}{2\rho} \leqslant \rho_k - \rho_{k+1} + L\left(\sigma(x_{k+1},x_k)^p + \sigma(x_k,x_{k-1})^p\right)$$

and

$$\frac{1}{2\rho}\sum_{k\geqslant 0}\beta_k^2 \leqslant \rho_0 + 2L\sum_{k\geqslant 0}\sigma(x_{k+1},x_k)^p < \infty,$$

and we have the desired result $\beta_k \to 0$ as $k \to \infty$. Q.E.D.

We now show that the PCG iterates satisfy (3.8). Eliminate the outer subscript k and consider

$$q(u^i) \leqslant q(u^1) = q(\alpha_0 D^+ b) = -\frac{1}{2}\frac{(b,D^+b)^2}{(b,D^+GD^+b)}, \qquad i \geqslant 1.$$

But from Theorem 2.3 we have

$$(D^+b,GD^+b) \leqslant 2(D^+b,DD^+b) = 2(b,D^+b).$$

From the choice of y, if $D_{ii} = 0$, then $b_i = 0$.

$$(b,D^+b) = \sum_{\substack{i=1\\ D_{ii}\neq 0}}\frac{b_i^2}{D_{ii}} \geqslant \frac{(b,b)}{(s,s)},$$

using that from (2.2) we have $D_{ii} \leqslant (s,s)$. Hence we have

$$q(u^i) \leqslant -\frac{1}{4}(b,D^+b) \leqslant -\frac{1}{4}\frac{(b,b)}{(s,s)}, \qquad i \geqslant 1.$$

The next theorem will show that one or more PCG iterations are sufficient to guarantee local and superlinear convergence for the inexact quasi-Newton method when $1 > \theta_k \to 0$ as $k \to \infty$.

THEOREM 3.3. *Let f be twice continuously differentiable in an open neighborhood Ω of a point x_* for which $\nabla f(x_*) = 0$, $H(x_*)$ is nonsingular, and let $L \geqslant 0$, $0 < p \leqslant 1$, be such that for all $x \in \Omega$,*

$$\|H(x) - H(x_*)\|_F \leqslant L\|x - x_*\|^p.$$

Let the relative residual (1.3) *satisfy $1 > \theta \geqslant \theta_k$, and let $q_k(u_k) \leqslant 0$. For any r which satisfies $\theta < r < 1$ there exist positive constants ε and σ, so that if*

$$\|x_0 - x_*\|_* \leqslant \varepsilon, \qquad \|B_0 - H(x_*)\|_F \leqslant \sigma,$$

where $\|y\|_* = \|H(x_*)y\|$, then for any inexact quasi-Newton method $x_k \to x_*$ as $k \to \infty$, and

$$\|x_{k+1} - x_*\|_* \leqslant r\|x_k - x_*\|_*, \qquad k \geqslant 0. \tag{3.12}$$

Moreover, if u_k satisfies (3.8) *and $\theta_k \to 0$ as $k \to \infty$, then*

$$\lim_{k\to\infty}\frac{\|x_{k+1} - x_*\|}{\|x_k - x_*\|} = 0.$$

Proof. Condition $q_k(u_k) \leqslant 0$ and (3.2) imply that the sequence of approximations $\{B_k\}$ of the Hessian matrix $H(x_*)$ of f is of bounded deterioration. If $\theta_k = 0$, the

local convergence follows from Broyden, Dennis and Moré [3]. The general case $1 > \theta \geqslant \theta_k$ follows from Steihaug [12] and Eisenstat and Steihaug [8].

From (3.12) we have (3.7), so if $\beta > 0$ and

$$q_k(u_k) \leqslant -\beta \frac{(b_k, b_k)}{(s_k, s_k)},$$

where $b_k = y_k - B_k s_k + \nabla f(x_{k+1}) - \nabla f(x_k) - B_k s_k$, then from Lemma 3.2 we have

$$\frac{\|\nabla f(x_{k+1}) - \nabla f(x_k) - B_k s_k\|}{\|s_k\|} \to 0 \quad \text{as } k \to \infty. \tag{3.13}$$

If $\theta_k = 0$, then the superlinear rate of convergence follows from Dennis and Moré [5]. From Steihaug [12] we have that the sequence $\{x_k\}$ is converging superlinearly if and only if

$$\lim_{k \to \infty} \frac{\|r_k\|}{\|\nabla f(x_k)\|} = 0 \tag{3.14}$$

provided (3.13) holds. But (3.14) holds if $\theta_k \to 0$ as $k \to \infty$. Q.E.D.

4. Global Convergence Results. A major problem in globalizing the quasi-Newton methods using the sparse update from Section 2 is that the matrix B_k can be singular. An appealing approach is to replace (1.1) by finding the solution s_k of the trustregion problem

$$\min\{\nabla f(x_k)^T s + \tfrac{1}{2} s^T B_k s : \|s\|_2 \leqslant \Delta_k\} \tag{4.1}$$

for a suitable choice of Δ_k and to replace (1.3) by finding an approximate solution of (4.1). Global algorithms based on trust regions [11], [16], a combination of conjugate gradient methods and trust regions [13], or a backtracking strategy [14] can be shown to be convergent in the sense that

$$\liminf_{k \to \infty} \|\nabla f(x_k)\| = 0$$

for any given x_0 and B_0 under the assumptions that there exist α_1 and α_2 that only depend on x_0 and B_0 so that

$$\|B_k\| \leqslant \alpha_1 + \alpha_2 \sum_{i=0}^{k-1} \|x_{i+1} - x_i\|, \tag{4.2}$$

f is bounded below, f is twice continuously differentiable in $\mathbf{R}^n$ and there exists $L \geqslant 0$ so that for all x and z in $\mathbf{R}^n$

$$\|H(x) - H(z)\|_F \leqslant L\|x - z\|. \tag{4.3}$$

So to establish global convergence results, we have to show that the approximations B_k are not growing too fast.

LEMMA 4.1. *Let f be twice continuously differentiable in $\mathbf{R}^n$, and assume that* (4.3) *holds. Let $x_0, x_1, \ldots, x_k$ be points in $\mathbf{R}^n$, and let $B_{i+1} = B_i(u_i)$ be updated using $s_i = x_{i+1} - x_i$ and y_i in* (1.4). *If $q_i(u_i) \leqslant 0$, $i = 0, 1, \ldots, k-1$, then there exist α_1 and α_2 that only depend on x_0 and B_0, so that* (4.2) *holds.*

Proof. Let M_i be defined as in (3.5) using s_i and y_i. Then

$$\|M_i - H(x_j)\|_F \leq \frac{L}{2}\|x_{i+1} - x_i\|, \qquad j = i, i+1. \tag{4.4}$$

Consider

$$\begin{aligned} \|B_{i+1} - H(x_{i+1})\|_F &\leq \|B_{i+1} - M_i\|_F + \|M_i - H(x_{i+1})\|_F \\ &\leq \|B_i - M_i\|_F + \frac{L}{2}\|x_{i+1} - x_i\| \\ &\leq \|B - H(x_i)\|_F + L\|x_{i+1} - x_i\| \end{aligned} \tag{4.5}$$

using the triangle inequality, (3.4), (4.3), and (4.4). Hence

$$\|B_k - H(x_k)\|_F \leq \|B_0 - H(x_0)\|_F + L\sum_{i=0}^{k-1}\|x_{i+1} - x_k\|. \tag{4.6}$$

But

$$\|H(x_k) - H(x_0)\|_F \leq L\sum_{i=0}^{k-1}\|x_{i+1}\text{-}x_i\|,$$

using the triangle inequality and (4.3), and we have

$$\begin{aligned} \|B_k\|_F &\leq \|B_k - H(x_k)\|_F + \|H(x_k) - H(x_0)\|_F + \|H(x_0)\|_F \\ &\leq \|H(x_0)\|_F + \|B_0 - H(x_0)\|_F + 2L\sum_{i=0}^{k-1}\|x_{i+1} - x_i\|, \end{aligned}$$

using the triangle inequality, (4.5), and (4.6). Q.E.D.

Acknowledgements. This work was performed as a part of the doctoral thesis [12] in the Department of Administrative Sciences at Yale University. I am grateful to Ron Dembo of the Department of Administrative Sciences and Stanley C. Eisenstat of the Computer Science Department for their guidance. The proof of Theorem 2.3 was significantly shortened by the referees. The author is indebted to Robert Schnabel for valuable comments, which improved the current version of the proof of Theorem 2.3 and the lower bound (2.15).

Department of Mathematical Sciences
Rice University
Houston, Texas 77001

1. O. AXELSSON, "Solution of linear systems of equations: iterative methods," *Sparse Matrix Techniques* (V. A. Barker, ed.), Lecture Notes in Math., vol. 572, Springer-Verlag, New York, 1976, pp. 1–51.

2. Å. BJÖRK & T. ELFVING, "Accelerated projection methods for computing pseudoinverse solutions of systems of linear equations," *BIT*, v. 19, 1979, pp. 145–163.

3. C. G. BROYDEN, J. E. DENNIS, JR. & J. J. MORÉ, "On the local and superlinear convergence of quasi-Newton methods," *J. Inst. Math. Appl.*, v. 12, 1973, pp. 223–245.

4. R. S. DEMBO, S. C. EISENSTAT & T. STEIHAUG, "Inexact Newton methods," *SIAM J. Numer. Anal.*, v. 19, 1982, pp. 400–408.

5. J. E. DENNIS, JR. & J. J. MORÉ, "A characterization of superlinear convergence and its application to quasi-Newton methods," *Math. Comp.*, v. 28, 1974, pp. 549–560.

6. J. E. DENNIS, JR. & J. J. MORÉ, "Quasi-Newton methods, motivation and theory," *SIAM Rev.*, v. 19, 1977, pp. 46–89.

7. J. E. DENNIS, JR. & R. B. SCHNABEL, "Least change secant updates for quasi-Newton methods," *SIAM Rev.*, v. 21, 1979, pp. 443–459.

8. S. C. EISENSTAT & T. STEIHAUG, *Local Analysis of Inexact Quasi-Newton Methods*, Department of Mathematical Sciences TR 82-7, Rice University, Houston, 1982.

9. E. S. MARWIL, *Exploiting Sparsity in Newton-Like Methods*, Ph.D. Thesis, Cornell University, Computer Science Research Report TR 78-335, 1978.

10. J. M. ORTEGA & W. C. RHEINBOLDT, *Iterative Solution of Nonlinear Equations in Several Variables*, Academic Press, New York, 1970.

11. M. J. D. POWELL, "Convergence properties of a class of minimization algorithms," *Nonlinear Programming*, Vol. 2 (O. L. Mangasarian, R. R. Meyer, and S. M. Robinson, eds.), Academic Press, New York, 1975, pp. 1–27.

12. T. STEIHAUG, *Quasi-Newton Methods for Large Scale Nonlinear Problems*, Ph.D. Thesis, Yale University, SOM Technical Report #49, 1980.

13. T. STEIHAUG, "The conjugate gradient method and trust regions in large scale optimization," *SIAM J. Numer. Anal.*, v. 20, 1983, pp. 626–637.

14. T. STEIHAUG, *Damped Inexact Quasi-Newton Methods*, Department of Mathematical Sciences TR 81-3, Rice University, Houston, 1981.

15. P. L. TOINT, "On sparse and symmetric matrix updating subject to a linear equation," *Math. Comp.*, v. 31, 1977, pp. 954–961.

16. P. L. TOINT, "On the superlinear convergence of an algorithm for solving a sparse minimization problem," *SIAM J. Numer. Anal.*, v. 16, 1979, pp. 1036–1045.