# LOCAL AND LINEAR CONVERGENCE OF AN ALGORITHM FOR SOLVING A SPARSE MINIMIZATION PROBLEM

Earl Marwil

TR77-324

Department of Computer Science
and Center for Applied Mathematics
Cornell University
Ithaca, NY 14853

# LOCAL AND LINEAR CONVERGENCE OF AN ALGORITHM FOR SOLVING A SPARSE MINIMIZATION PROBLEM

Earl Marwil

Department of Computer Science and
Center for Applied Mathematics
Cornell University
Ithaca, N.Y. 14853

ABSTRACT

For an unconstrained minimization problem with a sparse Hessian, a symmetric version of Schubert's update is given which preserves the sparseness structure defined by the Hessian. At each iteration of the algorithm there are two sparse linear systems to be solved. These have the same sparseness structure defined by the Hessian. The differences between succeeding approximations to the Hessian and the Hessian at the solution are related by a careful evaluation of the difference in the Frobenius norm. This relation is used in proving the local and linear convergence of the algorithm.

## 1. INTRODUCTION

Let $f : R^n \to R$ be given. Consider the problem of finding a local minimum of $f$ on some open set $D \subset R^n$. Let $x^* \in D$ be the minimum, so that

$$f(x^*) = \min \{f(x) : x \in D\}.$$

In the following, assume that $f$ is twice continuously differentiable on $D$.

When solving this problem using a Newton-like method to find a zero of $g(x) = \nabla f(x)$, it is usually the case that an approximation to $J_g(x) = \nabla^2 f(x)$, the Jacobian of $g$ or Hessian of $f$, looks as much like $J_g$ as possible. Since $\nabla^2 f$ is symmetric, it is approximated by a symmetric matrix.

Let $S = \{A \in L(R^n) : A = A^T\}$, and for $u,v \in R^n$, let $Q_{v,u} = \{A \in L(R^n) : Au = v\}$. A Newton-like method for the solutio of the minimization problem is: given $x \in D$ and $B \in S$, nonsingular,

(1.1) $$x_+ = x - B^{-1}g(x)$$
$$B_+ \in S \cap Q_{y,s}$$
where $$y = g(x_+) - g(x),$$
and $$s = x_+ - x.$$

Specific methods for choosing a $B_+$ in $S \cap Q_{y,s}$ have the property that near the solution, $B_+$ is nonsingular.

Suppose that $\nabla^2 f$ is sparse. The type of methods considered require the solution of a linear system in (1.1), namely

$$Bs = -g(x).$$

The objective is to define a method of updating B to preserve the sparseness structure of the Hessian. With that in mind, the criterion for deciding when the nonlinear problem is sparse will be the same for deciding when the linear systems are sparse. That is normally [5] if $\nabla^2 f$ has less than 10% nonzeros.

Furthermore assume that $[\nabla^2 f]_{ii} \neq 0$ for $i = 1,\ldots,n$. Otherwise, the $i^{\underline{th}}$ component of $\nabla f$ is linear in $x_i$, so that $x_i$ can be written as a function of the remaining $x_j$, $j = 1,\ldots,i-1,i+1,\ldots, n$, and the dimension of the problem can be reduced.

Let $Z = \{A \in L(R^n) : A_{ij} = 0$ for all $(i,j)$ such that $[\nabla^2 f(x)]_{ij} = 0 \quad \forall x \in D\}$. Schubert's update for sparse nonlinear equations [13], [1], [8] is in $Z \cap Q_{y,s}$, but it is not symmetric. For $B \in S \cap Z$ we want a $B_+ \in S \cap Z \cap Q_{y,s}$.

## 2. NOTATION AND TECHNICAL PRELIMINARIES

Definition 2.1 Let $s \in R^n$ and define the components of s by

$$\rho_i = e_i^T s \qquad \text{for } i = 1,\ldots,n. \tag{2.1}$$

Define $Z_j = \{v \in R^n : e_i^T v = 0$ for all i such that $[\nabla^2 f]_{ji} = 0\}$. In other words, $Z_j$ is the subspace determined by the zero-nonzero structure of the $j^{\underline{th}}$ row of $\nabla^2 f$.

For $j = 1,\ldots,n$ define the $\ell_2$ projection of s onto $Z_j$ by $s_j$. The vector $s_j$ is defined component-wise by

$$e_i^T s_j = \begin{cases} \rho_i & \text{if } [\nabla^2 f]_{ji} \neq 0 \\ 0 & \text{if } [\nabla^2 f]_{ji} = 0 \end{cases} \tag{2.2}$$

Lemma 2.2 Let $s \in R^n$ and suppose $f : R^n \to R$ is $C^2$ and $[\nabla^2 f]_{ii} \neq 0$ for $i = 1,\ldots,n$. Then

(2.3a) $$e_j^T s = e_j^T s_j \,, \qquad \text{and}$$

(2.3b) $$e_j^T s \; e_i^T s_j = e_j^T s_i \; e_i^T s_j$$

(2.3c) $w = \Sigma_{j=1}^n \; e_i^T s_j \; \alpha_j e_j \in Z_i$ for $\alpha_j \in R$, $j = 1,\ldots,n$.

Proof: (a): By the hypothesis, $[\nabla^2 f]_{jj} \neq 0$ implies $e_j^T s_j = \rho_j = e_j^T s$

(b): If $[\nabla^2 f]_{ji} \neq 0$, then by symmetry $[\nabla^2 f]_{ij} \neq 0$ so that $e_j^T s_i = \rho_j = e_j^T s$. Hence (2.3b) holds. If $[\nabla^2 f]_{ji} = 0$, then $e_i^T s_j = 0$ and both sides of (2.3b) are zero.

(c): $e_j^T w = e_i^T s_j \alpha_j = 0$ if $[\nabla^2 f]_{ij} = 0$ $\quad j = 1,\ldots,n$.

Therefore $w \in Z_i$.

Definition 2.3 For a scalar $\alpha \in R$, we use the notation of the generalized inverse and define

(2.4) $$\alpha^+ = \begin{cases} \frac{1}{\alpha} & \text{if } \alpha \neq 0 \\ 0 & \text{if } \alpha = 0. \end{cases}$$

Now, Schubert's update, for $B \in Z$ and $y, s \in R^n$, can be written as

$$B_+ = B + \sum_{\substack{j=1 \\ s_j \neq 0}}^n e_j e_j^T \frac{(y - Bs)}{s_j^T s_j} s_j^T$$

which, using Definition 2.3, is equivalent to

(2.5) $$B_+ = B + \Sigma_{j=1}^n \; (s_j^T s_j)^+ \; e_j^T (y - Bs) \; e_j s_j^T.$$

It is easy to verify that $B_+ \in Z \cap Q_{y,s}$ [8].

<u>Definition</u> 2.4 Let $s \in R^n$. Define $P \in L(R^n)$ by

(2.6) $$P = 1/2\ [I - \Sigma_{j=1}^{n} (s_j^T s_j)^+ e_j^T s\ s_j e_j^T].$$

Define $A, D \in L(R^n)$ by

(2.7) $$\begin{aligned} A &= \Sigma_{j=1}^{n} \rho_j s_j e_j^T \\ &= [\rho_1 s_1 \vdots \rho_2 s_2 \vdots \cdots \vdots \rho_n s_n] \end{aligned}$$

and

(2.8) $$D = \text{diag}\ (s_1^T s_1, s_2^T s_2, \ldots, s_n^T s_n).$$

Observe that A is the Frobenius projection of $ss^T$ onto Z. Also, P can be written as

(2.9) $$P = 1/2\ [I - AD^+].$$

<u>Definition</u> 2.5 For $v \in R^n$ and $M \in L(R^n)$ and an integer $0 < m \le n$, define $M_m \in L(R^m)$ to be the lower right m x m submatrix of $M = M_n$, and $v_m \in R^m$ to be the last m components of v.

The matrix P is a projection operator. This fact is a consequence of the following Lemma on the eigenvalues of P.

<u>Lemma</u> 2.6 The eigenvalues of P are in $[0, \max(1/2, 1 - \min_{\rho_i \neq 0} \frac{\rho_i^2}{s_i^T s_i})]$ and the eigenvalues of $(I - P)$ are in $[\min(1/2, \min_{\rho_i \neq 0} \frac{\rho_i^2}{s_i^T s_i}), 1]$

Proof: Without loss of generality, assume all the zero components of s are ordered first. That is

$$\rho_i = 0 \qquad i = 1,\ldots,\ell \qquad \text{and}$$
$$\rho_i \neq 0 \qquad i = \ell+1,\ldots,n.$$

Then $A_{n-\ell}$ is the nonzero part of A, and $D_{n-\ell}$ is invertible by (2.3a) and definition 2.5. Let $m = n - \ell$.

From (2.9), partioning the matrix,

$$P = 1/2\ [I_n - \begin{pmatrix} O_\ell & \\ & A_m D_m^{-1} \end{pmatrix}] = \begin{pmatrix} 1/2\ I_\ell & \\ & 1/2\ [I_m - A_m D_m^{-1}] \end{pmatrix}. \tag{2.10}$$

It is clear that P has $\ell$ eigenvalues equal to 1/2, and the remaining $n - \ell$ are the eigenvalues of $1/2\ [I_m - A_m D_m^{-1}]$. Similarly, $I - P$ has $\ell$ eigenvalues equal to 1/2, and its other $n - \ell$ are the eigenvalues of $1/2\ [I_m + A_m D_m^{-1}]$.

It is now sufficient to consider P and $I - P$ of dimension m, so we omit the subscripts temporarily. Observe that

$$P = 1/2\ [I - AD^{-1}] \qquad \text{and}$$
$$I - P = 1/2\ [I + AD^{-1}] \qquad \text{are similar}$$

to $1/2\ D^{-1/2}(D - A)D^{-1/2}$ and $1/2\ D^{-1/2}(D + A)D^{-1/2}$ respectively.

Schnabel [12] observed that $D + A$ is the sum of a diagonal positive definite matrix and positive semi-definite rank one matrices. Thus,

(2.11)

$$D + A = 2\begin{bmatrix} \rho_1^2 & & & \\ & \rho_2^2 & & \\ & & \ddots & \\ & & & \rho_n^2 \end{bmatrix} + \sum_{\substack{i<j \\ [\nabla^2 f]_{ij} \neq 0}} \begin{bmatrix} \cdots \rho_j^2 \cdots & \rho_i \rho_j \cdots \\ \vdots & \vdots \\ \cdots \rho_j \rho_i \cdots & \rho_i^2 \cdots \end{bmatrix} .$$

By the Interleaving Eigenvalue Lemma, e.g. Wilkinson [pp. 95-98, 15],

(2.12) $$\min \text{ e.v. } (D + A) \geq \min_i 2\rho_i^2 > 0 .^{(\dagger)}$$

Similarly, D - A can be written as a sum of positive semi-definite rank one matrices

(2.13) $$D - A = \sum_{\substack{i<j \\ [\nabla^2 f]_{ij} \neq 0}} \begin{bmatrix} \cdots \rho_j^2 \cdots & -\rho_i \rho_j \cdots \\ \vdots & \vdots \\ \cdots -\rho_j \rho_i \cdots & \rho_i^2 \cdots \end{bmatrix} .$$

So min e.v. $(D - A) \geq 0$.

Thus $P_m$ is positive semi-definite and

$\min \text{ e.v. } (I - P)_m \geq \min_i \frac{\rho_i^2}{s_i^T s_i} > 0$. So in general, min e.v. $P \geq 0$

and $$\min \text{ e.v. } (I - P) \geq \min \left(1/2, \min_{\rho_i \neq 0} \frac{\rho_i^2}{s_i^T s_i}\right).$$

This gives the lower bounds.

The bounds on the largest eigenvalues follow from

$$x^T x = x^T (I - P) x + x^T P x .$$

---

(†) Toint (14) also showed that D + A is positive definite.

This completes the proof of Lemma 2.6.

Lemma 2.7  Let $u, v, s \in R^n$ and $P$ be given by (2.9). Then

(2.14) $$\left\langle D^+(I - P)^{-1}u, v\right\rangle = \left\langle D^+(I - P)^{-1}v, u\right\rangle.$$

Proof: Suppose that the zero components of s are ordered first. Then

$$I - P = \left(\begin{array}{c|c} 1/2\, I_\ell & \\ \hline & 1/2(I_m + A_m D_m^{-1}) \end{array}\right).$$

Now,

(2.15) $$\left\langle D^+(I - P)^{-1}u, v\right\rangle = \sum_{i=1}^{\ell} [2(s_i^T s_i)^+ e_i^T u] e_i^T v + \left\langle D_m^{-1}(I - P)_m^{-1} u_m, v_m\right\rangle.$$

Note that $(I - P)_m D_m = 1/2(D_m + A_m)$ which is symmetric and positive definite, by (2.12). Its inverse must be symmetric and positive definite also, so that

(2.16) $$\left\langle D_m^{-1}(I - P)_m^{-1} u_m, v_m\right\rangle = \left\langle u_m, D_m^{-1}(I - P)_m^{-1} v_m\right\rangle.$$

From (2.15) and (2.16) we have

(2.17) $$\begin{aligned}\left\langle D^+(I - P)^{-1}u, v\right\rangle &= \sum_{i=1}^{\ell} e_i^T u [2(s_i^T s_i)^+ e_i^T v] + \left\langle u_m, D_m^{-1}(I - P)_m^{-1} v_m\right\rangle \\ &= \left\langle u, D^+(I - P)^{-1} v\right\rangle.\end{aligned}$$

<u>Lemma</u> 2.8 Let $s \in R^n$, and P be defined by (2.9)

(a) If $v \in R^n$ is such that $e_i^T v = 0$

for all i such that $s_i = 0$, then

(2.18) $$-\left\langle D^+(I - P)^{-1}v, v\right\rangle \le -\frac{v^T v}{s^T s}$$

(b) If $u \in R^n$, then

(2.19) $$\left\langle D^+(I - P)^{-1}u, u\right\rangle \le \max\left(2, \max_{\rho_j \neq 0} \frac{s_j^T s_j}{\rho_j^2}\right) \left[\sum_{i=1}^{n} (s_i^T s_i)^+ (e_i^T u)^2\right].$$

<u>Proof</u>: (a): Again assume that the zero components of s are ordered first. Then

(2.20) $$\begin{aligned} -&\left\langle D^+(I - P)^{-1}v, v\right\rangle \\ &= -\sum_{i=1}^{\ell} 2(s_i^T s_i)^+ (e_i^T v)^2 - \left\langle D_m^{-1}(I - P)_m^{-1} v_m, v_m\right\rangle \\ &\le -\frac{2}{s^T s} \sum_{\substack{i=1 \\ s_i \neq 0}}^{\ell} (e_i^T v)^2 - \|(I - P)_m D_m\|_2^{-1} v_m^T v_m \end{aligned}$$

since $s_i^T s_i \le s^T s$ and $\left\langle Mx, x\right\rangle \ge \frac{\left\langle x, x\right\rangle}{\|M\|_2}$ for nonsingular M.

By Lemma 2.6 $\|(I - P)_m\| \le 1$, so

(2.21) $$\begin{aligned} \|(I - P)_m D_m\|_2 &\le \|(I - P)_m\|_2 \, \|D_m\|_2 \\ &\le \|D_m\|_2 \\ &\le \max_{\ell < j \le n} s_j^T s_j \le s^T s . \end{aligned}$$

Therefore, (2.20) and (2.21) yield

(2.22)
$$-\left\langle D^{+}(I-P)^{-1}v, v\right\rangle$$
$$\leq -\frac{1}{s^{T}s}\Big(2\sum_{\substack{i=1\\ s_i\neq 0}}^{\ell}(e_i^{T}v)^2 + v_m^{T}v_m\Big)$$
$$\leq -\frac{1}{s^{T}s}\Big(\sum_{\substack{i=1\\ s_i\neq 0}}^{\ell}(e_i^{T}v)^2 + v_m^{T}v_m\Big)$$
$$= -\frac{v^{T}v}{s^{T}s} \quad \text{since for } s_i = 0,\ e_i^{T}v = 0.$$

(b):

(2.23)
$$\left\langle D^{+}(I-P)^{-1}u, u\right\rangle$$
$$= \sum_{i=1}^{\ell} 2(s_i^{T}s_i)^{+}(e_i^{T}u)^2 + \left\langle D_m^{-1}(I-P)_m^{-1}u_m, u_m\right\rangle$$
$$= \sum_{i=1}^{\ell} 2(s_i^{T}s_i)^{+}(e_i^{T}u)^2 + \left\langle 2(D_m+A_m)^{-1}u_m, u_m\right\rangle$$
$$\leq \sum_{i=1}^{\ell} 2(s_i^{T}s_i)^{+}(e_i^{T}u)^2 + \|2D_m^{1/2}(D_m+A_m)^{-1}D_m^{1/2}\|_2\|D_m^{-1/2}u_m$$
$$= \sum_{i=1}^{\ell} 2(s_i^{T}s_i)^{+}(e_i^{T}u)^2 + \|(I-P)_m^{-1}\|_2\left\langle D_m^{-1}u_m, u_m\right\rangle$$
$$\leq 2\sum_{i=1}^{n}(s_i^{T}s_i)^{+}(e_i^{T}u)^2 + \max_{\rho_j\neq 0}\frac{s_j^{T}s_j}{\rho_j^2}\left\langle D_m^{-1}u_m, u_m\right\rangle$$

by Lemma 2.6 and

$$\|(I-P)_m^{-1}\|_2 = \frac{1}{\min \text{e.v.}(I-P)_m}$$

for positive definite $(I-P)_m$

$$\leq \max\Big(2, \max_{\rho_j\neq 0}\frac{s_j^{T}s_j}{\rho_j^2}\Big)\sum_{i=1}^{n}(s_i^{T}s_i)^{+}(e_i^{T}u)^2.$$

## 3. SYMMETRIC SCHUBERT UPDATE

Powell [10] derived the Powell symmetric Broyden update from the Broyden update,

$$B_+ = B + \frac{(y - Bs)s^T}{s^T s},$$

by iteratively projecting $B_+$ into S and then projecting back to $Q_{y,s}$. Using the same technique, Dennis [3] derived most of the symmetric updates satisfying the linear equation $B_+ s = y$ from the rank 1 updates,

$$B_+ = B + \frac{(y - Bs)c^T}{c^T s},$$

with various choices of $c \in R^n$.

We use the iterative double projection on Schubert's update (2.5), alternately projecting into S and back to $Z \cap Q_{y,s}$. For $B \in S \cap Z$ and $y, s \in R^n$, given, the derivation of the symmetric Schubert update is straightforward, though somewhat tedious, and is omitted here. The update is given by

$$(3.1) \qquad B_+ = B + 1/2 \sum_{j=1}^{n} (s_j^T s_j)^+ e_j^T \lambda (e_j s_j^T + s_j e_j^T)$$

where $\lambda = (\sum_{k=0}^{\infty} P^k)(y - Bs)$ and P is defined by (2.6).

Lemma 2.6 implies that the maximum eigenvalue of P is less than 1, since $\min_{\rho_i \neq 0} (s_i^T s_i)^{-1} \rho_i^2 > 0$. Therefore by the Neumann Lemma [9],

$$\sum_{k=0}^{\infty} P^k = (I - P)^{-1},$$

and $\lambda$ is the solution of

$$(3.2)\qquad (I - P)\lambda = (y - Bs).$$

Observe that the matrix $I - P$ has the same sparseness structure as $\nabla^2 f$. Although $I - P$ is not symmetric we can rewrite the update so that the linear system is symmetric and positive definite. Again assume that the zero components of $s$ are ordered first. Then

$$(3.3)\qquad B_+ = B + \sum_{\substack{j=1 \\ s_j \neq 0}}^{n} (e_j s_j^T + s_j e_j^T) e_j^T \hat{\lambda}$$

where $\hat{\lambda}$ is the solution to

$$(3.4)\qquad \left(\begin{array}{c|c} G_\ell & 0 \\ \hline 0 & D_m + A_m \end{array}\right) \hat{\lambda} = (y - Bs)$$

and $G_\ell = \text{diag}(\gamma_1, \ldots, \gamma_\ell)$

$$\gamma_i = \begin{cases} s_i^T s_i & \text{if } s_i \neq 0 \\ 1 & \text{if } s_i = 0. \end{cases}$$

## 4. PROPERTIES OF THE UPDATE

The first lemma of this section will verify that the update has the desired structure.

Lemma 4.1 Let $B \in S \cap Z$ and $y, s \in R^n$. Then $B_+$ defined by (3.1) satisfies $B_+ \in S \cap Z \cap Q_{y,s}$.

Proof: Obviously $B_+ \in S$. To show $B_+ \in Z$, it is sufficient to check each row.

(4.1) $$e_i^T B_+ = e_i^T B + 1/2(s_i^T s_i)^+ e_i^T \lambda s_i^T + 1/2 \sum_{j=1}^{n} e_i^T s_j (s_j^T s_j)^+ e_j^T \lambda e_j^T.$$

The first and second terms on the right are in $Z_i$. By Lemma 2.2, part c, the third term is in $Z_i$ also. Therefore, $B_+ \in Z$.

To see that $B_+ \in Q_{y,s}$, form the vector $B_+ s$ and check componentwise that $e_i^T B_+ s = e_i^T y$. If $s_i \neq 0$,

(4.2) $$\begin{aligned} e_i^T B_+ s &= e_i^T Bs + 1/2 e_i^T \lambda + 1/2 \sum_{j=1}^{n} e_i^T s_j (s_j^T s_j)^+ e_j^T \lambda e_j^T s \\ &= e_i^T Bs + e_i^T (I - P)\lambda \qquad \text{from (2.6)} \\ &= e_i^T y. \end{aligned}$$

If $s_i = 0$, then

(4.3) $$e_i^T B_+ s = e_i^T B_+ s_i = 0.$$

By the Mean Value Theorem [9], $\exists \bar{t}_i \in (0,1)$ such that

$$e_i^T (g_+ - g) = e_i^T J_g (x + \bar{t}_i s) s$$

Then $$e_i^T y = e_i^T J(x + \bar{t}_i s) s_i = 0 \qquad \text{since } J_g \in Z.$$

Therefore $B_+ \in Q_{y,s}$ and the proof is complete.

The following estimate will be used to prove the convergence properties of the algorithm given in section 5.

<u>Theorem</u> 4.2 Let $B, J \in S \cap Z$ and $y, s \in R^n$, $s \neq 0$ and let $B_+$ be defined by (3.1). Then

(4.4) $$\|B_+ - J\|_F^2 \le \|B - J\|_F^2 - \frac{\|(B - J)s\|_2^2}{\|s\|_2^2} + \max\left(2, \max_{\rho_j \neq 0} \frac{s_j^T s_j}{\rho_j^2}\right) \sum_{i=1}^{n} (s_i^T s_i)^+ [e_i^T (y - Js)]^2$$

Proof:

$$(4.5)\qquad ||B_+ - J||_F^2 = \sum_{i,j=1}^{n} [e_i^T(B_+ - J)e_j]^2$$

$$= \sum_{i,j=1}^{n} [e_i^T(B - J)e_j + 1/2(s_i^Ts_i)^+e_i^T\lambda s_i^Te_j + 1/2e_i^Ts_j(s_j^Ts_j)^+e_j^T\lambda]^2$$

$$= \sum_{i,j=1}^{n} \{[e_i^T(B - J)e_j]^2 + 2e_i^T(B - J)e_j(s_i^Ts_i)^+e_i^T\lambda e_j^Ts_i$$

$$+ 1/2[(s_i^Ts_i)^+(e_i^T\lambda)(e_j^Ts_i)]^2 + 1/2(s_i^Ts_i)^+e_i^T\lambda e_j^Ts_i(s_j^Ts_j)^+e_j^T\lambda e_i^Ts_j\}$$

$$= ||B - J||_F^2 + \sum_{i=1}^{n} (s_i^Ts_i)^+e_i^T\lambda\{2e_i^T(B - J)s$$

$$+ 1/2e_i^T\lambda + 1/2\sum_{j=1}^{n} (s_j^Ts_j)^+e_j^Ts_ie_i^Ts_je_j^T\lambda\}.$$

Now, observe that

$$y - Bs = (I - P)\lambda = 1/2\lambda + 1/2(\sum_{j=1}^{n} (s_j^Ts_j)^+e_j^Tss_je_j^T)\lambda$$

and

$$(4.6)\qquad e_i^T(y - Bs) = 1/2e_i^T\lambda + 1/2\sum_{j=1}^{n} (s_j^Ts_j)^+e_j^Tse_i^Ts_je_j^T\lambda$$

$$= 1/2e_i^T\lambda + 1/2\sum_{j=1}^{n} (s_j^Ts_j)^+e_j^Ts_ie_i^Ts_je_j^T\lambda$$

by (2.3b).

Therefore, (4.6) applied to the last line of (4.5) yields

(4.7)

$$||B_+ - J||_F^2 = ||B - J||_F^2 + \sum_{i=1}^{n} (s_i^Ts_i)^+e_i^T\lambda[2e_i^T(B - J)s + e_i^T(y - Bs)]$$

$$= ||B - J||_F^2 + \sum_{i=1}^{n} (s_i^Ts_i)^+e_i^T\lambda[e_i^T(B - J)s + e_i^T(y - Js)].$$

To complete the proof, we examine the sum on the right hand side of (4.7). Let $u = y - Js$ and $v = (B - J)s$. Then

$$\lambda = (I - P)^{-1}(y - Bs) = (I - P)^{-1}(u - v), \text{ and}$$

(4.8)
$$\begin{aligned}\sum_{i=1}^{n} (s_i^T s_i)^+ e_i^T \lambda [e_i^T (B - J)s + e_i^T (y - Js)] &= \sum_{i=1}^{n} (s_i^T s_i)^+ e_i^T (I - P)^{-1}(u - v)[e_i^T v + e_i^T u] \\ &= \sum_{i=1}^{n} e_i^T D^+ (I - P)^{-1}(u - v) e_i^T (v + u) \\ &= \left\langle D^+(I - P)^{-1}u, v\right\rangle - \left\langle D^+(I - P)^{-1}v, v\right\rangle \\ &\quad + \left\langle D^+(I - P)^{-1}u, u\right\rangle - \left\langle D^+(I - P)^{-1}v, u\right\rangle .\end{aligned}$$

Now, substituting (4.8) into (4.7) gives

(4.9)
$$\begin{aligned}||B_+ - J||_F^2 &= ||B - J||_F^2 + \left\langle D^+(I - P)^{-1}u, v\right\rangle - \left\langle D^+(I - P)^{-1}v, v\right\rangle \\ &\quad + \left\langle D^+(I - P)^{-1}u, u\right\rangle - \left\langle D^+(I - P)^{-1}v, u\right\rangle \\ &= ||B - J||_F^2 - \left\langle D^+(I - P)^{-1}v, v\right\rangle + \left\langle D^+(I - P)^{-1}u, u\right\rangle\end{aligned}$$

by Lemma 2.7

$$\le ||B - J||_F^2 - \frac{v^T v}{s^T s} + \max\left(2, \max_{\rho_j \neq 0} \frac{s_j^T s_j}{\rho_j^2}\right) \sum_{i=1}^{n} (s_i^T s_i)^+ (e_i^T u)^2$$

by Lemma 2.8, and since $s_i = 0$ implies

$$e_i^T v = e_i^T (B - J)s = e_i^T (B - J)s_i = 0.$$

$$= ||B - J||_F^2 - \frac{||(B - J)s||_2^2}{||s||_2^2} + \max\left(2, \max_{\rho_j \neq 0} \frac{s_j^T s_j}{\rho_j^2}\right) \sum_{i=1}^{n} (s_i^T s_i)^+ [e_i^T (y - Js)]^2 .$$

Finally, $B_+$ defined by (3.1) is the solution to $\min\{||\hat{B} - B||_F : \hat{B} \in S \cap Z \cap Q_{y,s}\}$. It is straightforward to apply the variational techniques [7], [6] to this constrained minimization problem. In January 1975, Powell [11] posed the problem in the variational form. The solution [14] is the same as the $B_+$ in (3.3). In fact, the author originally derived the update in this way in March 1975, after the problem and method of attack had been suggested by J.J. Moré in January 1975.

## 5. THE ALGORITHM

Let $x \in R^n$, $B \in S \cap Z$, positive definite be given.

1. Solve $B\Delta x = -g$ for $\Delta x$, the step to be taken; $g = g(x)$.
2. Set $x_+ = x + \Delta x$.
3. Evaluate $g_+ = g(x_+)$; test for convergence.
4. Set $y = g_+ - g$.
5. For each i s.t. $|e_i^T \Delta x| < \dfrac{||\Delta x_i||}{M}$

with $M \geq 2$ fixed, set $\rho_i = 0$; otherwise set $\rho_i = e_i^T \Delta x$. This defines $s = (\rho_1, \rho_2, \ldots, \rho_n)^T$, the step to be used to update B.

6. Compute $B_+$ from B, s, y using (3.36) and (3.37).

The convergence theorem in the next section will show that $B_+$ is positive definite in the region of local convergence. An implementation of the algorithm might avoid singularity of the Hessian approximation by re-evaluating the Hessian, $B_+ = J_g(x_+)$ or by adding a positive diagonal matrix to a singular or nearly singular $B_+$. One possible choice for starting the iteration would be to take $B_0 = I$.

The test in step 2 means that we don't want any component of $\Delta x$ tending to zero faster than its corresponding projection. Also this ensures that the eigenvalues of P given by (2.6) are not tending to 1 as $k \to \infty$, or those of the matrix associated with the linear system (3.4) are not tending to 0 as $k \to \infty$. This is critical for the conditioning of the linear systems involving $\lambda$ or $\hat{\lambda}$, and it is also important in the convergence proof.

The test is about what would happen anyway on the machine. When the size of any component is less than machine precision times the size of the corresponding projection, that component is insignificant. In other words, for $M^{1/2} = \dfrac{1}{n \quad (\text{mach. eps.})}$, the algorithm is roughly the one carried out on the computer.

On the other hand, for small M, say M = 2, it is likely that many $\rho_i$ are set to zero. In that case, the correction to B doesn't take much work. This is close to the idea of keeping the same approximation to the Hessian for several iterations.

In step 6, the form of the update in (3.3) is preferable to (3.1) since the linear system (3.4) is symmetric. This, of course, requires some bookkeeping to do the permutation on the vector s and the corresponding permutations of B and y. However, the important block in that symmetric system has the same spareness structure as the corresponding block of the permuted Hessian. Also, this is a possibly smaller system to solve.

## 6. CONVERGENCE

Theorem 6.1 Let $f : R^n \to R$ be $C^2(D)$ for D open and convex; assume $\exists x^* \in D$ s.t. $\nabla f(x^*) = 0$, $\nabla^2 f(x^*)$ is positive definite and $[\nabla^2 f(x)]_{ii} \neq 0 \; \forall x \in D$. If $\exists k_i > 0$ s.t.

$$||e_i^T(\nabla^2 f(x) - \nabla^2 f(x^*))|| \leq k_i ||x - x^*||$$

for $i = 1,\ldots,n$ and for all $x \in D$, then $\exists \delta, \varepsilon > 0$ such that for $(x_0, B_0)$, $B_0 \in Z \cap S$, which satisfy $||B_0 - \nabla^2 f(x^*)||_2 < \delta$ and $||x_0 - x^*||_2 < \varepsilon$ then the symmetric Schubert method generates $\{B_k\}$ with $B_k$ well-defined and $\{x_k\}$ which converges linearly to $x^*$.

Proof: Choose $M \geq 2$. Set $\kappa = \sum_{i=1}^{n} k_i^2$. From Theorem 3.4.2 with $J = J^* = \nabla^2 f(x^*)$,

(6.1)

$$||B_+ - J^*||_F \leq ||B - J^*||_F^2 + \max\left(2, \max_{\rho_j \neq 0} \frac{s_j^T s_j}{\rho_j^2}\right) \sum_{\substack{i=1 \\ s_i \neq 0}}^{n} \frac{[e_i^T(y - J^*s)]^2}{s_i^T s_i}$$

For $s_i \neq 0$, an application of the Mean Value Theorem [9] gives

$$\frac{|e_i^T(y - J^*s)|^2}{||s_i||^2} \leq k_i^2 \sigma(x, x + s)^2 \tag{6.2}$$

where $\sigma(x, x + s) = \max(||x - x^*||, ||x + s - x^*||)$. Therefore,

$$\max\left(2, \max_{\rho_j \neq 0} \frac{s_j^T s_j}{\rho_j^2}\right) \sum_{\substack{i=1 \\ s_i \neq 0}}^{n} \frac{[e_i^T(y - J^*s)]^2}{s_i^T s_i} \leq M\kappa\sigma(x, x + s)^2. \tag{6.3}$$

Now, (6.3) with (6.2) imply

(6.4) $$\|B_+ - J^*\|_F^2 \le \|B - J^*\|_F^2 + M\kappa\sigma(x, x + s)^2$$

and

(6.5) $$\|B_+ - J^*\|_F \le \|B - J^*\|_F + \alpha\sigma(x, x + s)$$

where $\alpha = (M\kappa)^{\frac{1}{2}}$.

Furthermore,

(6.6)
$$\begin{aligned} \|x + s - x^*\| &\le \|s\| + \|x - x^*\| \\ &\le \sqrt{\tfrac{n}{M}}\, \|\Delta x\| + \|x - x^*\| \\ &\le \sqrt{\tfrac{n}{M}}\, [\|x - x^*\| + \|x_+ - x^*\|] + \|x - x^*\| \\ &\le \sqrt{\tfrac{n}{M}}\, \|x_+ - x^*\| + (1 + \sqrt{\tfrac{n}{M}}\,)\|x - x^*\| \end{aligned}$$

so that

(6.7) $$\begin{aligned} \sigma(x, x + s) &\le \sqrt{\tfrac{n}{M}}\, \|x_+ - x^*\| + (1 + \sqrt{\tfrac{n}{M}}\,)\|x - x^*\| \\ &\le (1 + \sqrt{\tfrac{n}{M}}\,)\sigma(x, x_+). \end{aligned}$$

Therefore, (6.5) and (6.7) imply

(6.8) $$\|B_+ - J^*\|_F \le \|B - J^*\|_F + \alpha_1\sigma(x, x_+)$$

with $\alpha_1 = \alpha(1 + \sqrt{\frac{n}{M}}\,)$. By the Bounded Deterioration Theorem [2], the matrices $B_k$ generated by the symmetric Schubert method are well-defined and the sequence $\{x_k\}$ converges linearly to $x^*$.

## 7. CONCLUSION

The usual benefit of using the sparseness structure is the reduced storage. At each iteration of the algorithm we solve two sparse linear systems. All of these have the same structure. Once a pivoting strategy is chosen [5] for solving the first symmetric sparse linear system we can use that preprocessing for all subsequent sparse linear systems. The usual criterion for sparseness is less than 10% nonzero entries. That would be observed in deciding when a nonlinear problem is sparse since it involves solving sparse linear equations as a subproblem.

## REFERENCES


[1] C.G. Broyden, The convergence of an algorithm for solving sparse nonlinear systems, Math. Comp. 25(1971), pp. 285-294.

[2] C.G. Broyden, J.E. Dennis, Jr., and J.J. Moré, On the local and superlinear convergence of quasi-Newton methods, JIMA 12(1973), pp. 223-246.

[3] J.E. Dennis, Jr., On some methods based on Broyden's secant approximation to the Hessian, in Numerical Methods for Nonlinear Optimization, F.A. Lootsma, ed., AP, London, 1972.

[4] J.E. Dennis, Jr. and J.J. Moré, Quasi-Newton methods, motivation and theory, SIAM Review 19 (1977), pp. 46-89.

[5] I.S. Duff, A survey of sparse matrix research, Computer Science and Systems Division, AERE Harwell report HL 76/485, AERE Harwell, Oxfordshire, England, 1976.

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

[7] J. Greenstadt, Variations on variable-metric methods, Math. Comp. 24(1970), pp. 1-22.

[8] E. Marwil, Local and superlinear convergence of Schubert's method for solving sparse nonlinear equations, submitted for publication.

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

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

[11] M.J.D. Powell, A view of unconstrained optimization, C.S.S. 14, AERE, Harwell, 1975.

[12] R. Schnabel, private communication.

[13] L.K. Schubert, Modification of a quasi-Newton method for nonlinear equations with a sparse Jacobian, Math. Comp. 24(1970), pp. 27-30.

[14] Ph.L. Toint, On sparse and symmetric matrix updating subject to a linear equation, Dept. of Appl. Math. and Theor. Phys. report DAMTP 77/NA1, University of Cambridge, 1977.

[15] J.H. Wilkinson, The Algebraic Eigenvalue Problem, Oxford University Press, London, 1965.