academicJournals

Vol. 6(8), pp. 166-176, August, 2013
DOI: 10.5897/AJMCSR11.072
ISSN 2006-9731 ©2013 Academic Journals
http://www.academicjournals.org/AJMCSR

African Journal of Mathematics and Computer Science Research

Full Length Research Paper

A theoretical and experimental study of the Broyden-Fletcher-Goldfarb-Shano (BFGS) update

T. A. Adewale[1] and B. I. Oruh[2]*

[1]Department of Industrial Mathematics, Adekunle Ajasin University, P. M. B. 1, Akungba – Akoko, Nigeria.
[2]Department of Mathematics/Statistics/Computer Science, Michael Okpara University of Agriculture, Umudike, Nigeria.

Accepted 10 May, 2013

This paper discusses theoretically the evolution of a conjugate direction algorithm for minimizing an arbitrary nonlinear, non quadratic function using Broyden-Fletcher-Goldfarb-Shano (BFGS) update in quasi-Newton Method. The updating rule is initialized by a Moore Penrose's generalized inverse. Specifically, an approximation to the inverse Hessian is constructed and the updating rule for this approximation is imbedded in the BFGS update. Numerical experiments show that, using the proposed line search algorithm and the modified quasi-Newton algorithm for unconstrained problems are very competitive. This paper produces a new analysis that demonstrates that the BFGS method with a line search is $(n+1)$ step q-superlinear convergent with assumption of linearly independent iterates. The analysis assumes that the inverse Hessian approximations are positive definite and bounded asymptotically, which from computational experience, are of reasonable assumptions.

Key words: Quasi-Newton method, Moore-Penrose generalized inverse, Broyden-Fletcher-Goldfarb-Shano (BFGS) update, superlinear convergence, conjugate directions, orthogonalization of matrices.

INTRODUCTION

This paper is connected with interactions in the form:

$$\underline{x}_{k+1} = \underline{x}_k + \alpha_k \underline{S}_k - B_k \nabla f(\underline{x}_k), k = 0,1 \dots n-1$$

Where $x_k \in IR^n, \underline{S}_k \in IR^n, B_k \in IR^{nxn}$ and α_k is a scalar and it is the step length parameter chosen under the condition min $f(\underline{x}_k + \alpha \underline{S}_k)$ and is determined by:

$$\alpha_k = \frac{\langle \nabla f(x_0), \underline{S}_k \rangle}{\langle B\underline{S}_k, \underline{S}_k \rangle} = \frac{\langle \nabla f(x_0), \underline{S}_k \rangle}{\langle P_k, \underline{S}_k \rangle}, k = 0,1, \dots n-1,$$

$\underline{S}_k$ is the search direction, $\nabla f(\underline{x}_k)$ is the gradient vector at $\underline{x}_k$. Each B_k is intended to approximate inverse Hessian at $\underline{x}_k$ and $\underline{\alpha}_k$is chosen to prevent divergence of the sequence $\{\underline{x}_k\}$, f is assumed to be at least twice continuously differentiable. Bea-Israel (1996) developed an iterative scheme which can be employed for inversion of the Hessian of the objective function to be minimized, namely;

$$B_{k+1} = B_k(2P_H - HB_k), k = 0, 1 \dots \dots \dots n-1,$$

where H is an n × n matrix, P_H is an orthogonal matrix and B_0 is chosen to be the Moore-Penrose generalized inverse of H and it is derived as follows (Altman, 1960; Bernet 1979, Demidovich 1981, Rao and Mitra 1971, Rao 1973): Let M and N be given positive definite matrices, $\mu_1^2, \dots, \mu_n^2$ be non-zero Eigenvalues of $HN^{-1}H$ with respect to M^{-1} or of H^TMH with respect

*Corresponding author. E-mail:oruhben@yahoo.com.

to N, $\{\Im_r\}_{r=1}^{n}$ be Eigenvectors of $HN^{-1}H^T$ with respect to M^{-1} and $\{\eta_r\}_{r=1}^{n}$ be eigenvectors of H^TMH with respect to N. We can write H in the form:

$$H = M^{-1}\{\mu_1\underline{\Im}_1\eta_1^T + \cdots + \mu_n\Im_n\eta_n^T\}N \quad (1)$$

and $H^+{}_{MN}$ the more Penrose generalized inverse of H as:

$$H^+{}_{MN} = \mu_1^{-1}\eta_1\Im_1^T + \mu_2^{-1}\Im_2^T + \cdots + \mu_n^{-1}\eta_n\Im_n^T \quad (2)$$

We shall take:

$$B_0 = H^+{}_{MN} \quad (3)$$

In a previous paper, P_H was taken to be I_n, the n × n identity matrix since it satisfied the properties of P_H.; namely:

$\|HB_0 - P_H\| < 1, \|.\|$ is any valid matrix norm

$\|H_0B - P_H\| < 1$

We shall chose the Frobenius norm $\|.\|$ defined by:

$$\|H\|_F = \left(\sum h_{ij}^2\right)^{\frac{1}{2}}, h_{ij} \quad i,j, = 1,2,..,n.. \quad (4)$$

Being the entries of H and P_H to be the matrix derived as follows:

Let us take $H(\underline{x}_0) = H_0$ where H_0 non-singular and real is. That is the Hessian matrix at the initial point. Let us represent this by:

$$H_0 = H^{(1)} = \left(h_{ij}{}^{(1)}\right) \qquad i,j = 1,2 \ldots..n, h_{ij} = h_{ji}, \quad (5)$$

And leave unchanged the first two rows, from each i^{th} row, $i \geq 3$, subtract the second row of $H^{(1)}$ multiplied by a scalar λ_{i2} $i = 3,4,\ldots,n$. The new matrix is:

$$H^{(2)} = \left(h_{ij}{}^{(2)}\right), \text{ for i = 1,2.} \quad (6)$$

and

$$h_{ij}{}^{(2)} = h_{ij}{}^{(1)} - \lambda_{12}h_{2\,j,}\, i \geq 3 \quad (7)$$

Observing that the first row of $H^{(2)}$ coincides with the first row of $H^{(1)}$ and all the other rows of $H^{(2)}$ are linear combinations of the rows of $H^{(1)}$ orthogonal to the first row of $H^{(1)}$ and therefore the row of $H^{(2)}$ will also be orthogonal to its first row, we chose λ_{12}, the multiplier so that the row of $H^{(2)}$, from the third onwards are orthogonal to the second row. In summary, this is equivalent to:

$$\sum_{j=1}^{n} h_{ij}{}^{(2)}\, h_{ij}{}^{(2)} = \sum_{j=1}^{n} h_{2\,j}^{(1)}\left(h_{ij}{}^{(1)} - \lambda_{12}h_{2j}{}^{(1)}\right) \quad (8)$$

or

$$\sum_{j-1}^{n} h_{2j}{}^{(2)}\, h_{ij}{}^{(2)} = \sum_{j=1}^{n} h_{2\,j}^{(1)}\, h_{ij}{}^{(1)} - \sum_{j=1}^{n}\left[h_{2j}{}^{(1)}\right]^2 = 0 \quad (9)$$

Whence,

$$\lambda_{12} = \sum_{j=1}^{n} h_{2\,j}^{(1)}\, h_{ij}{}^{(1)} \neq \sum_{j=1}^{n}\left[h_{2j}{}^{(1)}\right]^2, i = 3,4\ldots,n \quad (10)$$

From each ith row of $H = (h_{ij}), i,j = 1,2\ldots,n,$ beginning with the second subtract the first row multiplied by a scalar, $\lambda_{i1}, i = 2,3,\ldots.,n$ dependent on the number of the row we get the transformed matrix $H^{(1)}$ given by:

$$H^{(1)} = \begin{cases}\left(h_{ij}{}^{(1)}\right) = (h_{ij}), \text{for } i = j\\ \left(h_{ij}{}^{(1)}\right) = (h_{ij} - \lambda_{i1}h_{ij}), \text{for } i \geq 2\end{cases} \quad (11)$$

We chose multiplier λ_{ij} such that the first row of matrix $H^{(1)}$ is orthogonal to the other rows of the matrix.
We then have:

$$\sum_{j=i}^{n} h_{ij}{}^{(1)}h_{ij}{}^{(1)} = \sum_{j=i}^{n} h_{ij}\left(h_{ij} - \lambda_{i1}h_{ij}\right) = \sum_{j=i}^{n} h_{ij}h_{ij} - \sum_{j=i}^{n}\lambda_{i1}h_{ij}{}^2 = 0 \quad (12)$$

Whence,

$$\lambda_{i1} = \sum_{j=i}^{n} h_{ij}h_{ij} \Big/ \sum_{j=i}^{n} h_{ij}^2, i = 2,3\ldots,n \quad (13)$$

This process is continued until we get the matrix:

$$H^{(n-1)} = \left(h_{ij}{}^{(n-1)}\right), \quad i,j = 1,2,\ldots,n \quad (14)$$

All the rows are orthogonal in pairs, that is, until:

$$\sum_{j=1}^{n} h_{kj}^{(n-1)} h_{ij}^{(n-1)} = 0 \text{ when } k \neq i \tag{15}$$

The matrix $H^{(n-1)} = \tilde{R}$ has orthogonal row and it is, therefore, not difficult to see that:

$$H^{(n-1)} H^{(n-1)} = \tilde{R}\tilde{R}^T = D = (d_{ij}), i = j \tag{16}$$

That is, a diagonal matrix. Also if H is a matrix with orthogonal columns, then:

$$H^T H = D = (d_{ij}), i = j = 1,2, \dots n \tag{17}$$

Also, if a matrix H has orthogonal row/column it is sufficient to normalize each row/column to orthogonalize it (Barnet 1979; Demidovich 1981)
That is:

$$(\tilde{h}_{ij}) = \left[h_{ij} \Big/ \left[\sum_{k=1}^{n} h_{kj}^2\right]^{1/2} \right], i,j = 1,2, \dots n \tag{18}$$

is an orthogonal matrix, we shall next set:

$$H^{(n-1)} = P_H \tag{19}$$

and the iteration is defined by:

$$B_{k+1} = B_k\left(2H^{(n-1)} - HB_k\right), k = 0,1, \dots, n-1 \tag{20}$$

Since B_{k+1} is an approximation to H^{-1} we intend to improve upon this using the Broyden–Fletcher–Goldfarb (BFGS) update defined by:

$$\hat{B}_{k+1} = \hat{B}_k + \frac{B_k \underline{S}_k (\hat{B}_k \underline{S}_k)}{(\hat{B}_k \underline{S}_k)^T \hat{B}_k \underline{S}_k} + \frac{B \underline{S}_k \underline{S}_k{}^T \hat{B}_k}{\underline{S}_k^T \hat{B}_k \underline{S}_k} \tag{21}$$

That is,

$$\hat{B}_{k+1} = \hat{B}_k + \frac{\hat{B}_k \underline{S}_k{}^T \hat{B}_k}{\underline{S}_k^T \hat{B}_k^T \hat{B}_k \underline{S}_k} + \frac{\hat{B}_k \underline{S}_k \underline{S}_k{}^T \hat{B}_k}{\underline{S}_k^T \hat{B} \underline{S}_k} \tag{22}$$

Where

$$\hat{B}_{k+1} = \hat{B}_k = B_k\left(2H^{(n-1)} - HB_k\right), k = 0,1,2, \dots n-1$$

A major drawback to quasi-Newton method (other than the difficulty of obtaining analytical derivatives) is that the value of the objective function is guaranteed to be improved on each cycle only if the Hessian matrix of the objective function, $H(x) = \nabla^2 f(x)$ is positive definite. $H(x)$ is positive definite for strictly convex functions but for general functions, quasi-Newton method may lead to search directions diverging from the minimum of $f(x)$. Recall that a real symmetric matrix is positive definite if all the Eigenvalues are positive. We shall, therefore need to demonstrate or present schemes for "forcing" positive definiteness on the approximate inverse Hessian.

Certain authors have proposed that the Hessian matrix be forced to be positive definite at each stage of the minimization. Himelblau, (1972) and Rao, (1978) devised a scheme of Eigenvalue analysis that guaranteed that an estimate of the inverse, B_k, would be positive definite. Let B^{-1} be approximate to H(x), scale the matrix B^{-1} as follows:

$$\pi(x) = C^{-1}(x) B_k{}^{-1}(x) C^{-1}(x) \tag{23}$$

Where $C(x)$ is a diagonal matrix whose elements are:

$$C_{ii} = \left(|b_{ii}{}^{(k)-1}|\right)^{\frac{1}{2}} \tag{24}$$

That is, the positive square root of the absolute values of the elements on the main diagonal of $B_k{}^{-1}(x)$, π will have all positive or negative ones on its main diagonal. Because $C^{-1}(x)$ and $B_k{}^{-1}(x)$ are non singular and of order n, the inverse of the product is the product of the inverses in reverse order, or:

$$\pi^{-1}(x) = \left(C^{-1}(x) B_k{}^{-1}(x) C^{-1}(x)\right)^{-1} = \left(C^{-1}(x)\right)^{-1} \left(B_k{}^{-1}(x)\right)^{-1} \left(C^{-1}(x)\right)^{-1}$$

That is:

$$\pi^{-1}(x) = \left(C^{-1}(x)\right)^{-1} \left(B_k{}^{-1}(x)\right)^{-1} \left(C^{-1}(x)\right)^{-1} \tag{25}$$

Then $B_k(x)$ can be calculated from the scaled matrix as:

$$B_k(x) = C^{-1}(x) \pi^{-1}(x) C^{-1}(x) \tag{26}$$

We can express, $\pi^{-1}(x)$ in terms of the Eigenvalues λ_{ii} of $\pi^{-1}(x)$ and the Eigenvalues of the inverse matrix are simply the inverse λ_1^{-1}, of the Eigenvalues of the original matrix. Therefore:

$$\pi^{-1}(x) = \sum_{i=1}^{q} \lambda_i^{-1} \ell_i \ell_i^T \tag{27}$$

Where ℓ_i is the normalized Eigenvector corresponding to the Eigenvalues x_i. Instead of using $\pi^{-1}(x)$, however, $\tilde{\pi}^{-1}(x)$ is used:

$$\tilde{\pi}^{-1}(x) = \sum_{i-1}^{n} |\lambda_i|^{-1} \ell_i \ell_i^T \qquad (28)$$

in which any of λ_i is replaced by a small positive number, so that B_K can now be guaranteed positive definite if computed from:

$$\tilde{B}_k(x) = C^{-1}(x)\tilde{\pi}^{-1}(x)C^{-1}(x) \qquad (29)$$

This scheme described above shall be employed in this presentation. The second scheme that shall be employed in this study was due to Marquardt (1963); Levenberg (1994) and Goldfield et al. (1966). To ensure that the estimate of $H^{-1}(x)$ was positive definite the above named authors suggested the following computation scheme:

$$\tilde{B}_k(x) = C^{-1}(x)(\pi(x) + \mu I)^{-1}C^{-1}(x) = (B_k^{-1}(x) + \mu C^2(x))^{-1} \qquad (30)$$

Where, μ is a positive constant such that $\mu > -min\{\chi_i\}$. Because the Eigenvalues of $(\pi(x) + \mu I)$ are $(\chi_i + \mu)$, Equation (30) guarantees that $B_k(x)$ is positive definite since use of an approximate μ in Equation (30) in effect destroys negative small Eigenvalues of the approximation to the Hessian matrix. Note that with μ sufficiently large, μI can overwhelm $\pi(x)$ and the minimization approach a steepest descent search. A third scheme which is only good for mentioning in this study is due to Zwart (1969) but will not be employed in this investigation.

The main purpose of this paper is to better understand the computational and theoretical properties of the BFGS update in the context of basic line search and quasi-Newton methods for unconstrained optimization for the BFGS method. Ge and Powell (1983) proved, under a different set of assumptions from those of Conn et al. (1988a; 1988b; 1991), that the sequence of general matrices converges, but not necessarily to $\nabla^2 f(x_k)$. We shall demonstrate that under the assumption of uniform linear independence of sequence of steps and boundedness and positive definiteness of the guaranteed matrices a new convergence analysis is possible. We presented computation experience with the BFGS update using a standard line search technique and quasi-Newton algorithms for small to medium size unconstrained optimization problems. Convergence analysis is undertaken and some brief conclusion and comments regarding future research were made.

COMPUTATIONAL RESULTS AND ALGORITHM

In order to test the performance of quasi-Newton (conjugate direction) method for unconstrained optimization using BFGS update we present and discuss some numerical experiments that were conducted. Minimization of the function after orthogonalization of the Hessian matrix using:

i) The Broyden-Fletcher-Goldfarb-Shano (BFGS) Update
ii) The Davidon-Fletcher-Powell (DFP) update,
iii) The symmetric rank one update and,
iv) Minimization of RsenBrock's Banana-shape valley function using Lagrange's reduction of quadratic forms in quasi-Newton (lagroqf q-n) method for comparism.

The line search is based on a cubic modelling of $f(x)$ in the direction of search developed by the authors and the Quasi-Newton is determined using the New Line Search Technique (Rao, 1978; Walsh, 1968) to approximately minimize the function in the experimental set of questions. The frame works of these algorithms are presented below:

Algorithm

Quasi-Newton method (line search)

Step 0: give an initial vector $\underline{x}_0$ of independent variables, an initial positive definite matrix B_0 and $\alpha = \min f(x_1 + \alpha pi)$. Set k (the interaction counter) = 0
Step 1: if a convergence criterion is achieved, then stop
Step 2: Compute a quasi-Newton direction $\underline{S}_k = -(\tilde{B}_k \underline{\nabla} f C(x_k))$ if $\tilde{B}_k$ is safely positive definite, else set $\underline{S}_k = -\left(\tilde{B}_k^{-1} + \mu_k C^2(x)\right)^{-1} \underline{\nabla} f C(x_k)$ where $\mu > 0$ such that $\mu > -\min\{X_i\}$ as defined in Equations 23 to 29 or 30 such that $\tilde{B}_k^{-1} + \mu_k C^2(x)$ is safely positive definite.
Step 3: find an acceptable step length λ_k using algorithm (31) (Adewale, 2003; Demidovick, 1981):
Step 4: Set $x_{k+1} = x_k + \alpha_k \underline{S}_k$
Step 5: Compute the next inverse Hessian approximation $\tilde{B}_{k+1}$ using the BFGS update
Step 6: Set $k = k + 1$ and go to Step 1.

Error in the inverse Hessian approximation and uniform linear independence

Definition

A sequence of vectors $\{\underline{S}_k\}_{k=1}^{n}$ in IR^n is said to be uniformly linearly independent if there exist $\Im > 0$, k_0 and $m \geq n$ such that, for each $m \geq n$, one can choose n distinct indices:

$k \leq k_1 < ... < k_n \leq k + m$ such that the minimum singular value of the matrix

$$S_k = \left[\frac{S_{k_1}}{\|S_{k_1}\|} \cdots \frac{S_{k_n}}{\|S_{k_n}\|} \right] \geq \Im \qquad (31)$$

Also det $S_k = \Delta_k \geq \in$, an arbitrarily small positive number and $\|S_k\| \to 0$ as $k \to \infty$.

Conn et al. (1991) analyzed the Hessian error for the Symmetric-Rank-One update (SR1) and under the assumption of uniform linear independence which is redefined above. Using this definition we shall establish how close the inverse Hessian approximation produced by the BFGS algorithm is to the exact inverse Hessian at the final iterates.

Theorem

Suppose that $f(x)$ is twice continuously differentiable everywhere, that is $[\nabla^2 f(x)]^{-1}$ is bounded and Lipschitz continuous, that is there exist constants M > 0 and $\gamma > 0$ such that $x, y \in IR^n$:

$$\|\nabla^2 f(x)^{-1}\| \leq M \text{ and } \|\nabla^2 f^{-1}(x) - \nabla^2 f^{-1}(y)\| \leq \gamma \|x - y\| \qquad (32)$$

Let $x_{k+1} = x_k + \alpha_k S_k$ where S_k is a uniformly linearly independent sequence of steps, and suppose that $\lim_{k \to \infty} \{x_k\} = x^*$ for some $x^* \in IR^n$. Let $\{\hat{B}_k\}$ be generated by the BFGS formula:

$$\hat{B}_{k+1} = \hat{B}_k + \frac{\hat{B}_k S_k S_k^T \hat{B}_k^T}{S_k^T \hat{B}_k^T \hat{B}_k S_k} + \frac{\hat{B}_k S_k S_k^T \hat{B}_k}{S_k^T \hat{B}_k S_k}, B = B_{k+1} = B_k\left(2H^{(n-1)} - HB_k\right), k = 0,1,2\ldots, n-1, \qquad (33)$$

$H = \nabla^2 f(x), B_0$ is as defined in (3) and suppose that $\Delta_k \geq 0, x_{k+1-i}$ and S_k satisfy:

$$\ell_{k-i} = \nabla f(x_{k+1-i}) - \nabla f(x_{k-i}) \qquad (34)$$

$$S_k = {\alpha_k}^{-1}(x_{k+1}) - x_k \qquad (35)$$

$$\hat{B}_k S_{k-i} = \ell_{k-i}, i = 0,1, \ldots n-1 \qquad (36)$$

$$S_k = \ell_k \in \{X_{k+1}\}$$

Then we have with any $k \geq n-1$

$$\lim_{n \to \infty} \left\|\hat{B}_k - \left[\nabla^2 f(x_k)\right]^{-1}\right\| = 0 \qquad (37)$$

Proof

Using Lagrange's formula for operator (Hessian and inverse Hessian included):

We can write:

$$\nabla f(X_{k+1-i}) - \nabla f(X_{k-i}) = \int_0^1 \left[\nabla^2 f(X_{k-i} + \tau(X_{k-i}))\right]^{-1} S_{k-i} d\tau \qquad (38)$$

$$= \int_0^1 [\nabla^2 f(X_{k-i})]^{-1} S_{k-i} d\tau + \int_0^1 \{[\nabla^2 f(X_{k-i} + \tau S_{k-i})]^{-1}\} S_{k-i} d\tau$$

$$= \left[\nabla^2 f(X_{k-i})\right]^{-1} S_{k-i} d\tau + \int_0^1 \left\{\left[\nabla^2 f(X)_{k-i} + \tau S_{k-i}\right]^{-1} - \left[\nabla^2 f(X_{k-i})\right]^{-1} S_{k-i}\right\} d\tau \qquad (39)$$

Using this expression we have:

$$\left[\hat{B}_k - \left[\nabla^2 f(x_k)\right]^{-1}\right] S_{k-i} = \left[\nabla^2 f(x_k)\right]^{-1} - \left[\nabla^2 f(x_k)\right]^{-1} S_{k-i} +$$

$$\int_0^1 \left\{\left[\nabla^2 f(x_{k-i} + \tau S_{k-i})\right]^{-1}\right\} S_{k-i} d\tau \qquad (40)$$

Introducing the notation:

$$\hat{B}_k - \left[\nabla^2 f(x_{k-i})\right]^{-1} = B_k, \text{ we obtain} \qquad (41)$$

$$\|B_k S_{k-i}\| \leq \left\|\left[\nabla^2 f(x_{k-1})\right]^{-1} - \left[\nabla^2 f(x_i)\right]^{-1}\right\| \|S_{k-1}\| +$$
$$Sup \left\|\nabla^2 f(x_{k-i} + \tau S_{k-i})^{-1} - \left[\nabla^2 f(x_{k-i})\right]^{-1}\right\| \|S_{k-i}\|, 0 \leq t \leq 1$$

Since $\{X_k\}$ is a bounded sequence, with any k we have $X_k \in Q, Q \subset IR^n$ is a closed bounded set.

The function $\left[\nabla^2 f(x)\right]^{-1}$ is uniformly continuous since $\nabla^2 f(x)$ is assumed uniformly continuous in set Q. Consequently:

$$\|[\nabla^2 f(x_{k-i})]^{-1} - [\nabla^2 f(x_k)]^{-1}\| = \eta_{k-i} \to 0 \text{ and}$$
$$\sup_{0 \geq t \leq 1} \left\|\left[\nabla^2 f(x_{k-i} + \tau S_{k-i})\right]^{-1} - \left[\nabla^2 f(x_{k-i})\right]^{-1}\right\| = \mu_{k-i} \to 0 \text{ as } k \to \infty \qquad (42)$$

Thus it follows from (41) that:

$$\|B_k S_{k-i}\| \leq (\eta_{k-i} + \mu_{k-i})\|S_{k-i}\| = h_{k-i}\|S_{k-i}\| \qquad (43)$$

Where, $h_{k-i} \to 0 as$ $k \to \infty$. According to the definition of the operator norm,

$$\|B_k\| = \max_{\|z\|=1} \|B_k Z\|.$$

Let the maximum be attained at element z_k if:

$$Z_k = \delta_k \frac{s_k}{\|s_k\|} + \cdots + \delta_{k-n+1} \frac{s_{k-n+1}}{\|s_{k-n+1}\|}, \quad (44)$$

Then because of the condition:

$\|\Delta_k\| \geq \varepsilon > 0$, Where $\Delta_k = \det S_k$ defined in (31), the coefficients δ_{k-1} will be bounded, that is, $|\delta_k| \leq \varepsilon; i = 0,1 \dots, n-1$. Using (44) we obtain:

$$\|B_k\| = \|B_k Z_k\| = \left\| \sum_{i=0}^{n-1} \delta_{k-1} B_k \frac{S_{k-i}}{S_{k-i}} \right\| \leq \sum_{i=0}^{n} \left\| \delta_{k-i} B_k \frac{S_{k-i}}{\|S_{k-i}\|} \right\| \quad (45)$$

Hence by (43 and the fact that $|\delta_{k-i}|$ is bounded we have:

$$\|B_k\| \leq \sum_{i=0}^{n-1} |\delta_{k-i}| \frac{h_{k-i}\|S_{k-i}\|}{\|S_{k-i}\|} = \sum_{i=0}^{n-1} |\delta_{k-i}| h_{k-i} \to 0 \text{ as } k \to \infty$$

That is:

$\left\| \hat{B}_k - \left[\nabla^2 f(x_k)\right]^{-1} \right\| \to 0 \; as \; k \to \infty$ and the theorem is proved.

Theorem

If $f(x)$ is a continuously differentiable strongly convex function and sequence $\{X_k\}$ is such that $f(X_{k+1}) \leq f(x_k)$ and $\langle \nabla f(x_k), x_{k+} - x_k \rangle \to 0$ as $k \to \infty$, then $\|X_{k+1} - X_k\| \to 0$.

Proof

According to condition $f(x_{k+1}) \leq f(x_k)$ we have: $X_{k+1} \in S_k, S_k = \{x: f(x) \leq f(x_k)\}$ with any k. The set S_k is strongly convex since $f(x)$ is a strongly convex function. Then there is a positive number $\mu > 0 \ni$ any point $\frac{x_{k+1}+x_k}{2} + \Im$, where $\|\Im\| \leq \mu \|x_{k+1} - x_k\|^2$, is an internal point of the set S_k. Let $\frac{x_{k+1}-x_k}{2} = V + W$ where $V \in T_k, T_k$ is a plane tangent to the set S_k and $w \perp T_k$ Then noting that: $\nabla f(x_k) \perp T_k$, we obtain $\frac{1}{2}|\langle \nabla f(x_k), X_{k+1} - X_k \rangle| = |\langle \nabla f(x_k), V + w \rangle| = \|\nabla f(x_k)\| \|w\|$.

But $\|w\| \|\Im\| >$µ since otherwise in addition to point x_k set S_k and plane T_k would have other point in common, which contradicts the strong convexity of S_k.

Therefore:

$$\frac{1}{2} |\langle \nabla f(x_k), X_{k+1} - X_k \rangle| \geq \eta \|\nabla f(x_k)\| \|X_{k+1} - X_k\|^2$$

Hence, if $\|\nabla f(x_k)\| \to 0 \; then \; \|X_{k+1} - X_k\| \to 0$. But if $\|\nabla f(x_k)\| \to 0$, then since $f(x)$ is strongly convex, the maximum diameter of set $S_k d_k \to 0$ which implies that $\|X_{k+1} - X_k\| \to 0$. The theorem is proved.

Theorem

If $f(x)$ is a twice continuously differentiable function for which $m\|y\|^2 \leq \langle \hat{B}_k y, y \rangle \leq m\|y\|^2, m > 0, x, y \in IR''$, are valid, matrix $\hat{B}_k$ with any $k \geq n-1$ is defined by system of equations:

$\hat{B}_k S_{k-i} = \ell_{k-i}, i = 0,1, \dots, n-1$ and satisfies the condition:

$\langle \hat{B}_k \nabla f(x_k), \nabla f(x_k) \rangle < 0 \; and \; \alpha_k$ is determined to be $\min f(x_k + \alpha S_k)$, then whatever the initial point x_0 the following statement stated are valid for sequence:

$$X_{k+1} = X_k - \alpha_k \hat{B}_k \nabla f(x_k), \alpha_k > 0$$

$f(x_{k+1}) < f(x_k)$ and $\|x_k - x^*\| \to 0$ at a superlinear rate of convergence, $\|X_{N+1} - X^*\| \leq q\lambda_N \dots \lambda_{N+1}$

Where $q, N < \infty, \lambda_{N+1} < 1$ with any $\ell \geq 0, \lambda_i \to 0 \; as \; i \to \infty$

COMPUTATIONAL RESULTS

On the basis of analogical heuristics, we shall implement the algorithm on four test problems, three of which are non quadratic functions common with authors of quasi-Newton methods. We shall orthogonalize the constant matrices resulting from the Hessian of the quadratic approximations to the function. They shall be compared under BFGS, SRL and DFP updates.

Problem

A quadratic function

$$f(X_1, X_2) = 2X_1^2 + 2X_1X_2 + X_2^2, \underline{X}_0 = (1,1)^T$$

$$H = \begin{pmatrix} 4 & 2 \\ 2 & 4 \end{pmatrix}, \hat{H} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, z_1 = x_1$$

$$z_2 = (x_1 + x_2), \tilde{f}(\underline{z}) = z_1^2 + z_2^2, \underline{z}_0 = (1,2)$$

Table 1. Minimization of Rosenbrock's function (using the BFGS update).

$\alpha^{(k)}$	η_{EV_k} $\lvert\varepsilon_F\rvert < 10^{-10}$	η_{EV}	η_{it}	ε_F
0	101	104	18	8.6×10^{-20}
10^{-5}	72	74	18	6.8×10^{-20}
10^{-3}	60	61	17	1.6×10^{-16}
10^{-1}	52	53	21	2.6×10^{-16}
0.5	48	49	26	5.1×10^{-18}
0.75	40	42	32	2.8×10^{-20}
0.9	40	42	32	2.8×10^{-20}
1.0	39	40	31	1.3×10^{-7}

$\alpha^{(k)}$ = Steplength parameter; η_{EV} = number of function evaluation; $\lvert\varepsilon_F\rvert$ = error of function value approximation; η_{it} = number of iteration.

Table 2. minimization of Rosenbrock's function using the DFP update.

$a^{(k)}$	η_{EV_k} $\lvert\varepsilon_F\rvert < 10^{-10}$	η_{EV}	η_{it}	ε_F
0	118	119	22	3.7×10^{-20}
10^{-5}	88	89	22	3.4×10^{-20}
10^{-3}	77	78	22	5.8×10^{-20}
10^{-1}	61	63	24	1.3×10^{-20}
0.5	59	60	29	1.1×10^{-17}
0.75	41	42	36	2.1×10^{-19}
0.9	45	46	35	8.2×10^{-20}
1.0	41	42	35	2.7×10^{-18}

$\alpha^{(k)}$ = Steplength parameter; η_{EV} = number of function evaluation; $\lvert\varepsilon_F\rvert$ = error of function value approximation; η_{it} = number of iteration.

Problem

Powell's quattic

$$f(X_1,X_2,X_3,X_4) = (X_1+10X_2)^2 + 5(X_3-X_4)^2 + (X_2-2X_3)^4 + 10(x_1-X_4)^4$$

$$\underline{X}_0 = (3,-1,0,1)^T, H = \begin{pmatrix} 4 & 0 & 0 & 0 \\ 0 & 10 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 20 \end{pmatrix}$$

.

Problem

Woods function

$$f(X_1,X_2,X_3,X_4) = 100(X_1+X_1^2)^2 + (1-X_1)^2 + 90(X_1-X_3^2)^4 + (1-X_1)^2 + 10[(X_2-1)^2 \cdot (X_4-1)^2] + 1.98\,(X_2-1)(X_4-1), \underline{X}_0 = (-3,-1,-3,-1)^T$$

Problem

Rosenbrock's banana – shaped valley function

$$f(X_1,X_2) = (1-X_1)^2 + 100\left(X_2 - X_1^2\right)^2, \underline{X}_0 = (-1.2,1)^T, \quad H = \begin{pmatrix} 2 & 0 \\ 0 & 200 \end{pmatrix}$$

The numerical results are reported in Tables 1-11 including the Steplength parameter($\alpha^{(k)}$), number of function evaluation(η_{EV}), error of function value approximation($\lvert\varepsilon_F\rvert$), number of iteration($\eta_{it}$).

DISCUSSION OF COMPUTATIONAL RESULTS

The BFGS update has appeared to be superior in general application. This finding was corroborated by S.H.C Dutoit when developing computer programs for the analysis of covariance structure arising from nonlinear growth curves and from autoregressive time series with moving average residual (Rao, 1978) In this presentation

Table 3. Minimization of Rosenbrock's function using the symmetric-rank-one update (SR1).

$a^{(k)}$	$\eta_{EV_k} \lvert\varepsilon_F\rvert < 10^{-10}$	ε_F	η_{EV}	η_{it}
0	128	4.9×10^{-17}	130	23
10^{-5}	97	5.2×10^{-17}	99	23
10^{-3}	83	6.0×10^{-17}	84	23
10^{-1}	67	1.9×10^{-21}	69	27
0.5	56	6.2×10^{-16}	56	30
0.75	53	1.8×10^{-14}	54	35
0.9	55	2.1×10^{-15}	56	41
1.0	55	2.1×10^{-20}	57	44

$a^{(k)}$ = Steplength parameter; η_{EV} = number of function evaluation; $\lvert\varepsilon_F\rvert$ = error of function value approximation; η_{it} = number of iteration.

Table 4. Minimization of wood's (using the BFGS update).

$a^{(k)}$	$\eta_{EV_k} \lvert\varepsilon_F\rvert < 10^{-10}$	η_{EV}	η_{it}	ε_F
0	191	194	40	1.3×10^{-16}
10^{-5}	142	144	40	1.6×10^{-16}
10^{-3}	113	116	37	1.7×10^{-21}
10^{-1}	85	86	38	3.8×10^{-20}
0.5	96	98	69	4.0×10^{-17}
0.75	93	95	73	5.4×10^{-17}
0.9	87	89	73	4.6×10^{-16}
1.0	97	98	75	9.0×10^{-15}

$a^{(k)}$ = Steplength parameter; η_{EV} = number of function evaluation; $\lvert\varepsilon_F\rvert$ = error of function value approximation; η_{it} = number of iteration.

Table 5. Minimization of wood's function using DFP update.

$a^{(k)}$	$\eta_{EV_k} \lvert\varepsilon_F\rvert < 10^{-10}$	η_{EV}	η_{it}	ε_F
0	259	261	40	6.7×10^{-17}
10^{-5}	210	213	40	1.5×10^{-16}
10^{-3}	167	168	36	1.9×10^{-21}
10^{-1}	172	173	48	3.2×10^{-19}
0.5	450	452	158	1.9×10^{-21}
0.75	-	>1086	>1032	3.2
0.9	-	>1066	>1044	6.7
1.0	-	>898	>892	7.7

$a^{(k)}$ = Steplength parameter; η_{EV} = number of function evaluation; $\lvert\varepsilon_F\rvert$ = error of function value approximation; η_{it} = number of iteration.

we experiment with four nonlinear functions of many variables and it is discovered that one advantage of the BFGS over DFP update, for instance, is that a search to choose α_k, the step length parameter, is no longer always essential and it is often sufficient to let $\alpha_k = 1$ (Table 5). The DFP update, on the other hand, was first used in the analysis of convergence structure by Joreskog (1967) and has been employed successfully by him in a variety of situations but found that it requires a fairly complicated search on each interaction to choose α_k so as to minimize a discrepancy function. The BFGS

Table 6. Minimization of wood's function using the symmetric-rank-one update (SR1).

$a^{(k)}$	$\eta_{EV_k} \lvert\varepsilon_F\rvert < 10^{-10}$	η_{EV}	η_{It}	ε_F
0	209	212	42	2.5×10^{-17}
10^{-5}	151	152	42	4.3×10^{-17}
10^{-3}	139	142	45	1.5×10^{-17}
10^{-1}	95	96	41	8.8×10^{-18}
0.5	160	161	75	8.7×10^{-18}
0.75	139	141	85	3.0×10^{-20}
0.9	180	181	98	5.4×10^{-19}
1.0	143	144	93	1.8×10^{-18}

$a^{(k)}$ = Steplength parameter; η_{EV} = number of function evaluation; $\lvert\varepsilon_F\rvert$ = error of function value approximation; η_{it} = number of iteration.

Table 7. Minimization of powell's quartic function (using the BFGS update).

$a^{(k)}$	$\eta_{EV_k} \lvert\varepsilon_F\rvert < 10^{-10}$	η_{EV}	η_{It}	ε_F
0	102	132	26	8.0×10^{-14}
10^{-5}	78	106	26	7.7×10^{-14}
10^{-3}	72	97	27	4.3×10^{-15}
10^{-1}	45	51	22	7.7×10^{-14}
0.5	42	50	38	2.5×10^{-14}
0.75	39	46	43	1.5×10^{-11}
0.9	39	46	43	1.5×10^{-11}
1.0	38	41	37	1.3×10^{-11}

$a^{(k)}$ = Steplength parameter; η_{EV} = number of function evaluation; $\lvert\varepsilon_F\rvert$ = error of function value approximation; η_{it} = number of iteration.

Table 8. Minimization of Powell's quartic comparing the uses of the DFP, SR1, and BFGS updates, second order methods.

$a^{(k)}$	$\eta_{EV_k} \lvert\varepsilon_F\rvert < 10^{-10}$			η_{EV}			η_{It}			ε_F		
	BFGS	**SR1**	**DFP**	**BFGS**	**SR1**	**DFP**	**BFGS**	**SR1**	**DFP**	**BFGS**	**SR1**	**DFP**
0	102	99	108	132	112	151	26	21	26	8.0×10^{-13}	2.6×10^{-13}	7.9×10^{-14}
10^{-5}	78	75	81	106	85	119	26	21	26	7.7×10^{-14}	2.6×10^{-13}	8.5×10^{-14}
10^{-3}	72	64	74	97	77	98	27	23	26	4.3×10^{-15}	3.0×10^{-13}	6.1×10^{-15}
10^{-1}	45	36	45	51	52	53	22	24	18	7.7×10^{-14}	3.8×10^{-15}	1.5×10^{-17}
0.5	42	33	40	50	34	66	38	25	37	2.5×10^{-14}	3.4×10^{-10}	8.8×10^{-11}
0.75	39	38	38	46	39	99	43	33	84	1.5×10^{-11}	1.9×10^{-10}	8.9×10^{-13}
0.9	39	37	37	46	47	58	43	36	47	1.5×10^{-11}	9.0×10^{-11}	1.2×10^{-12}
1.0	38	46	194	41	56	214	37	44	200	1.3×10^{-11}	1.3×10^{-10}	4.0×10^{-13}

update has been the most commonly used secant update for many years. It makes a symmetric, rank-two change to the previous Hessian approximation B_0 and if B_0 is positive definite then B_{k+1} is positive definite. The BFGS has been shown to be q-superlinearly convergent provided that the initial Hessian approximation is sufficiently accurate. In this study, the inverse Hessian is initialized by Moor Pencrose's generalized inverse matrices which are not as accurate as required, yet the convergence is q-superlinear. Also for non quadratic functions, convergence of the SR1update is not as well understood as convergence of the BFGS method.

Conclusion

In this study we have attempted to investigate theoretical

Table 9. Minimization of wood's function comparing the uses of DFP, SR1 and the BFGS updates, second-order methods.

$a^{(k)}$	η_{EV_k} $\|\varepsilon_F\| < 10^{-10}$			η_{EV}			η_{It}			ε_F		
	DFP	SR1	BFGS	DFP	SR1	BFGS	DFP	SR1	BFGS	DFP	SR1	BFGS
0	259	209	191	261	212	194	40	42	40	6.7×10^{-17}	2.5×10^{-17}	1.3×10^{-16}
10^{-5}	210	151	142	213	152	144	40	42	40	1.5×10^{-16}	4.3×10^{-17}	1.6×10^{-16}
10^{-3}	167	139	113	168	142	116	36	45	37	1.9×10^{-21}	1.5×10^{-17}	1.7×10^{-21}
10^{-1}	172	95	85	173	96	86	48	41	38	3.2×10^{-19}	8.8×10^{-18}	3.8×10^{-20}
0.5	450	160	96	452	161	98	158	75	69	1.9×10^{-21}	8.7×10^{-18}	4.0×10^{-17}
0.75	-	139	93	>1086	141	95	>1032	85	73	3.2	3.0×10^{-20}	5.4×10^{-17}
0.9	-	180	87	>1066	181	89	>1044	98	73	6.7	5.4×10^{-19}	4.6×10^{-16}
1.0	-	143	97	>898	144	98	>892	93	75	7.7	1.8×10^{-18}	9.0×10^{-15}

Table 10. Minimization of Rosenbrock's banana- shaped valley function using lagroqf q-n

Iterative step	Function value	g^tg	X_1	X_2	Hessian
0	24.2	7.76×10^{3}	2.2	-0.44	Positive DEF
1	6.05×10^{-4}	1.94×10^{3}	1.1	-2.2	Positive DEF
4	9.65×10^{-6}	3.03×10	1.37×10^{-1}	-2.75×10^{-2}	Positive DEF
8	3.69×10^{-6}	1.18×10	8.59×10^{-1}	-1.72×10^{-3}	Positive DEF
12	1.44×10^{-6}	4.63×10^{-6}	5.37×10^{-4}	-1.07×10^{-1}	Positive DEF
16	5.63×10^{-9}	1.80×10^{-6}	3.36×10^{-5}	-6.71×10^{-6}	Positive DEF
20	2.20×10^{-11}	7.06×10^{-9}	2.09×10^{-9}	-4.19×10^{-7}	Positive DEF
24	8.59×10^{-16}	2.76×10^{-11}	1.31×10^{-7}	-2.62×10^{-8}	Positive DEF
28	3.35×10^{-16}	1.07×10^{-13}	8.19×10^{-9}	-1.64×10^{-9}	Positive DEF
32	1.31×10^{-18}	4.21×10^{-16}	5.12×10^{-12}	-1.02×10^{-10}	Positive DEF
36	5.12×10^{-26}	1.64×10^{-21}	3.20×10^{-1}	-6.40×10^{-12}	Positive DEF
40	2.00×10^{-23}	6.64×10^{-21}	2.00×10^{-12}	-4.00×10^{-13}	Positive DEF
44	7.82×10^{-26}	2.50×10^{-23}	1.25×10^{-23}	-1.25×10^{-14}	Positive DEF
48	3.05×10^{-28}	9.97×10^{-26}	7.82×10^{-15}	1.56×10^{-15}	Positive DEF
51	44.77×10^{-30}	1.53×10^{-27}	9.76×10^{-16}	1.95×10^{-16}	Positive DEF
52	0.0	0.0	0.0	0.0	Positive DEF

Table 11. Minimization of a quadratic function.

Iteration step	X_1	X_2	X_3	Function value	NGRAD	Positive definiteness
0	0	0	0	1	3	Positive DEF
1	-7.5×10^{-1}	-5×10^{-1}	1.9×10^{-1}	1.56×10^{-2}	7.5×10^{-1}	Positive DEF
2	-1.125	-7.5×10^{-1}	9.4×10^{-2}	-2.3×10^{-1}	1.88×10^{-1}	Positive DEF
3	-1.3125	-8.75×10^{-1}	1.41×10^{-1}	-2.92×10^{-1}	4.68×10^{-2}	Positive DEF
4	-1.40625	-9.37×10^{-1}	-1.17×10^{-1}	-3.07×10^{-1}	1.17×10^{-2}	Positive DEF
5	-1.453125	-9.69×10^{-1}	-1.29×10^{-1}	-3.11×10^{-1}	2.93×10^{-3}	Positive DEF
6	-1.47656	-9.84×10^{-1}	-1.23×10^{-1}	-3.12×10^{-1}	7.32×10^{-4}	Positive DEF
7	-1.49828	-9.92×10^{-1}	-1.25×10^{-1}	-3.12×10^{-1}	1.83×10^{-4}	Positive DEF
8	-1.49414	-9.96×10^{-1}	-1.25×10^{-1}	-3.12×10^{-1}	4.58×10^{-5}	Positive DEF
9	-1.49707031	-9.98×10^{-1}	-1.25×10^{-1}	-3.12×10^{-1}	1.14×10^{-5}	Positive DEF
10	-1.49853516	-9.99×10^{-1}	-1.25×10^{-1}	-3.12×10^{-1}	2.86×10^{-6}	Positive DEF
11	-1.49926758	-9.99×10^{-1}	-1.25×10^{-1}	-3.12×10^{-1}	7.15×10^{-7}	Positive DEF
12	-1.49963379	-9.99×10^{-1}	-1.25×10^{-1}	-3.12×10^{-1}	1.79×10^{-7}	Positive DEF
13	-1.4998189	-9.99×10^{-1}	-1.25×10^{-1}	-3.12×10^{-1}	4.47×10^{-8}	Positive DEF
14	-1.49990845	-9.99×10^{-1}	-1.25×10^{-1}	-3.12×10^{-1}	1.12×10^{-8}	Positive DEF

Table 11. Contd.

15	-1.49995422	-9.99×10^{-1}	-1.25×10^{-1}	-3.125×10^{-1}	2.79×10^{-9}	Positive DEF
16	-1.49996567	-9.999×10^{-1}	-1.25×10^{-1}	-3.125×10^{-1}	1.11×10^{-9}	Positive DEF
17	-1.499996568	-9.9999×10^{-1}	-1.25×10^{-1}	-3.125×10^{-1}	1.105×10^{-9}	Positive DEF

and numerical aspect of quasi-Newton methods that are based on the BFGS formula for the Hessian approximation. We considered only four functions. The performance of BFGS formula make us feel that the superiority of SR1 over BFGS claimed by Khalfan et al. (1993) needed to be probed further, especially, when combined with line searches. Also further study on the use of trust region strategy and line search techniques need to be undertaken. The reader is referred to the work of Nocedal and Yuan (1998).

REFERENCES

Adewale TA, Aderibigbe FM (2002). A New Line Search Technique, Quastiones Mathematicae, J. South Afr. Math. Soc. 25(4)453-464.

Altman MA (1960). An optimum cubically Convergent Iterative Method of Inverting a linear bounded operator in Hilbert space. Pacific J. Math. 16:7-113, 1107-1113.

Barnet S (1979). Matrix methods for Engineers and scientists McGraw-Hill Book company New York. pp. 139-145.

Conn AR, Gould NIM, Toint PhL (1988a). Global Convergence of a class of trust region algorithms for optimization with simple bounds. SIAM journal on Numerical Analysis, 25(2):433-460.

Conn AR, Gould NIM, Toint PhL (1988b). Testing a class of methods for solving minimization problems with simple bounds on the variables. Mathematics of computation, 50:399-430.

Conn AR, Gould NIM, Toint PhL (1991). Convergence of quasi-Newton matrices generated by the symmetric rank one update. Mathematical Progamming, 50(2):177-196.

Goldfeldt SM, Quandt RE, Trotter HF (1966). Maximization by quadratic hill-climbing. Econometrica, 34:541-551.

Demidovich BP (1981). Computational Mathematics, MIR publishers, Moscow. pp. 410-490.

Himelblau DM (1972). Applied Nonlinear Programming, McGraw Hill Book Company. New York,pp.30-34, 73-96, 190-217

Khalfan HF, Byrd RH, Schnabel RB (1993). A theoretical and experimental Study of the symmetric rank –one update. SIAM J. Optim. 3(1):1-24.

Levenberg K (1944). A method for the solution of certain problems in least squares. Quarterly Journal on Applied Mathematics , 2:164-168.

Marquardt D (1963). An algorithm for least squares estimation of nonlinear parameters. SIAM Journal on Applied Mathematics, 11:431-441.

Murray W (1972). The Relationship Between The Approximate Hessian Matrices Generated by a Class of quasi – Newton methods, NPL Report NAC 12.

Nocedal J, Yaun T(1998). Combining trust region and line search techniques. Advances in Nonlinear Programming, Kluwer Academic Publishers, Dordrecht, the Netherlands, pp. 153-176,

Pscnichey PS (1978). Numerical Methods in Extremal problems, MIR Publishers, Moscow, pp. 69-129.

Rao CR, Mitra SK (1971). Generalized Inverse of Matrices and its Applications, John Wiley and Sons, New York, pp. 7-9, 207-217.

Rao CR (1973). linear Statistical Inference and its Applications, 2nd ed. Willey New York pp. 1-50.

Rao SS (1978). Optimization theory and Applications Wily Eastern Limited. New York, pp. 318-720.

Walsh GR (1968). Methods of Optimization, John Willey and Sons, New York. pp. 97-139.