SIAM J. NUMER. ANAL.
Vol. 26, No. 3, pp. 727–739, June 1989

013

# A TOOL FOR THE ANALYSIS OF QUASI-NEWTON METHODS WITH APPLICATION TO UNCONSTRAINED MINIMIZATION*

RICHARD H. BYRD† AND JORGE NOCEDAL‡

**Abstract.** The BFGS update formula is shown to have an important property that is independent of the algorithmic context of the update, and that is relevant to both constrained and unconstrained optimization. The BFGS method for unconstrained optimization, using a variety of line searches, including backtracking, is shown to be globally and superlinearly convergent on uniformly convex problems. The analysis is particularly simple due to the use of some new tools introduced in this paper.

**Key words.** quasi-Newton methods, minimization, nonlinear optimization

**AMS(MOS) subject classifications.** 65, 49

**1. Introduction.** The object of this paper is to present a very general result about the BFGS update formula and to introduce some tools that can be very useful in the convergence analysis of quasi-Newton methods. We will focus our attention here on the unconstrained optimization problem

$$\min_{x \in \mathbf{R}^n} f(x). \tag{1.1}$$

However, as we will discuss later, the results presented here also provide the foundation for our study of reduced Hessian methods for constrained optimization (Byrd and Nocedal [1]).

Quasi-Newton methods for solving (1.1) are iterative methods of the form

$$x_{k+1} = x_k + \alpha_k d_k, \tag{1.2}$$

where $\alpha_k$ is a steplength, and $d_k$ is a search direction of the form

$$d_k = -B_k^{-1} g_k. \tag{1.3}$$

The matrix $B_k$ is updated at every step by means of a quasi-Newton update formula, and $g_k$ is the gradient of $f$ at $x_k$. In particular, the BFGS update formula is given by

$$B_{k+1} = B_k - \frac{B_k s_k s_k^T B_k}{s_k^T B_k s_k} + \frac{y_k y_k^T}{y_k^T s_k}, \tag{1.4}$$

where $y_k = g_{k+1} - g_k$ and $s_k = x_{k+1} - x_k$.

Global convergence of the BFGS method on uniformly convex functions has been shown by Powell [10] and by Werner [15], for a large class of line search methods.

* Received by the editors January 15, 1988; accepted for publication (in revised form) June 21, 1988.

† Department of Computer Science, University of Colorado, Boulder, Colorado 80309. This author was supported, in part, by National Science Foundation grants DCR-8403483 and CCR-8702403 and by Air Force Office of Scientific Research grant AFOSR-85-0251.

‡ Department of Electrical Engineering and Computer Science, Northwestern University, Evanston Illinois 60208. This author was supported by the Applied Mathematical Sciences subprogram of the Office of Energy Research, U.S. Department of Energy, under contracts W-31-109-Eng-38 and DE-FG02-87ER25047, and by National Science Foundation grant DCR-86-02071. This work was performed while this author was at Argonne National Laboratory, Argonne, Illinois 60439.

However, these analyses do not cover backtracking line searches. In this paper we introduce some new tools that make the global analysis of the BFGS method quite simple and that allow backtracking strategies to be covered. These tools can also be very useful in the study of other quasi-Newton methods for optimization.

In the next section we derive useful properties of the BFGS update in isolation. Then, in §3 these results are applied to the BFGS method for unconstrained optimization to give an improved global convergence result and a new proof of superlinear convergence. The details of backtracking algorithms are covered in §4. Throughout the paper $\|\cdot\|$ denotes the Euclidean vector norm.

**2. A basic property of the BFGS update.** In this section we show that, under very mild conditions, the BFGS update formula possesses a property that is very useful for establishing convergence results.

Let $B_1$ be symmetric and positive definite, and let $\{y_k\}$ and $\{s_k\}$ be two sequences of $n$-vectors such that $y_k^T s_k > 0$ for all $k$. The BFGS update formula (1.4) generates a sequence of positive definite matrices $\{B_k\}$ that are used in generating the search directions. We consider two useful quantities that depend on $B_k$. The first is the angle between $s_k$ and $B_k s_k$, defined by

$$\cos\theta_k = \frac{s_k^T B_k s_k}{\|s_k\| \, \|B_k s_k\|}. \tag{2.1}$$

Note, from (1.3), that $\theta_k$ is the angle between $d_k$ and $-g_k$, and as is well known, the convergence properties of unconstrained optimization algorithms can be studied by monitoring $\cos\theta_k$. The other important quantity is the Rayleigh quotient

$$q_k = \frac{s_k^T B_k s_k}{s_k^T s_k}. \tag{2.2}$$

As was first shown by Powell [10], the global convergence of the BFGS method can be studied by measuring the trace and determinant of $B_k$. From (1.4) we have

$$\mathrm{tr}(B_{k+1}) = \mathrm{tr}(B_k) - \frac{\|B_k s_k\|^2}{s_k^T B_k s_k} + \frac{\|y_k\|^2}{y_k^T s_k}, \tag{2.3}$$

where $\mathrm{tr}(B_k)$ denotes the trace of $B_k$. There is also a closed-form expression for the determinant (see Pearson [9])

$$\det(B_{k+1}) = \det(B_k)\frac{y_k^T s_k}{s_k^T B_k s_k}. \tag{2.4}$$

Bounds on the growth of the trace and the determinant of $B_k$ are derived by Powell [10], who subsequently combines them by means of the geometric/arithmetic mean inequality to prove global convergence. A similar procedure is employed by Werner [15], Ritter [11], [12], Toint [13], and Byrd, Nocedal, and Yuan [3]. One can simplify these proofs by working *simultaneously* with the trace and determinant of $B_k$. For this purpose we define, for any positive definite matrix $B$, the function

$$\psi(B) = \mathrm{tr}(B) - \ln(\det(B)), \tag{2.5}$$

where ln denotes the natural logarithm. Note that $\psi(B) > 0$ since

$$\psi(B) = \sum_{i=1}^{n}(\lambda_i - \ln\lambda_i), \tag{2.6}$$

where $0 < \lambda_1 \leq \cdots \leq \lambda_n$ are the eigenvalues of $B$. We also note that the function

$$w(t) = t - \ln t, \quad t > 0$$

is strictly convex and has the minimum value of 1 at $t = 1$. Therefore $\psi(B) \geq n$, and $\psi(B)$ can be considered as a measure of closeness between $B$ and the identity matrix, for which $\psi(I) = n$. Note also that since $w(t) > \ln t$, (2.6) gives

$$\begin{aligned} \psi(B) &> \ln \lambda_n - \ln \lambda_1 \\ &= \ln\left(\frac{\lambda_n}{\lambda_1}\right) \\ &= \ln \operatorname{cond}(B). \end{aligned} \tag{2.7}$$

We will now derive a closed-form expression for $\psi(B_k)$. From (2.3), (2.4), and (2.5) we have

$$\begin{aligned} \psi(B_{k+1}) &= \psi(B_k) - \frac{\|B_k s_k\|^2}{s_k^T B_k s_k} + \frac{\|y_k\|^2}{y_k^T s_k} - \ln\left(\frac{y_k^T s_k}{s_k^T B_k s_k}\right) \\ &= \psi(B_k) - \left[\frac{\|B_k s_k\| \|s_k\|}{s_k^T B_k s_k}\right]^2 \frac{s_k^T B_k s_k}{s_k^T s_k} + \frac{\|y_k\|^2}{y_k^T s_k} - \ln\left(\frac{y_k^T s_k}{s_k^T s_k} \frac{s_k^T s_k}{s_k^T B_k s_k}\right). \end{aligned}$$

Using the definitions (2.1) and (2.2) we have

$$\psi(B_{k+1}) = \psi(B_k) + \frac{\|y_k\|^2}{y_k^T s_k} - \ln \frac{y_k^T s_k}{s_k^T s_k} - \frac{q_k}{\cos^2 \theta_k} + \ln q_k \tag{2.8}$$

$$\begin{aligned} &= \psi(B_k) + \frac{\|y_k\|^2}{y_k^T s_k} - 1 - \ln \frac{y_k^T s_k}{s_k^T s_k} + [\ln \cos^2 \theta_k] \\ &\quad + \left[1 - \frac{q_k}{\cos^2 \theta_k} + \ln \frac{q_k}{\cos^2 \theta_k}\right]. \end{aligned} \tag{2.9}$$

Note that this recursion is analogous to the key recursion used by Dennis and Moré [5, Eq. (3.8)] to prove superlinear convergence of quasi-Newton methods. Here the function $\Psi$ is used instead of the weighted Frobenius norm to measure the badness of the Hessian approximation. The nonpositive quantities in square brackets can force a decrease in $\Psi$ when the direction is almost orthogonal to the gradient ($\cos \theta_k$ very small), or the step length is too short, and thus their role is similar to that of the term in the Dennis–Moré recursion involving $(B_k - F'(x_*))s_k$. The other terms in (2.9) tend to increase $\Psi$ (deteriorate $B_k$) but can be bounded, although the bounds are weaker than the corresponding ones for the Dennis–Moré recursion.

The following theorem uses the recursion (2.9) for $\psi$ to show that, provided certain derivatives of $f$ are appropriately bounded, a fraction $p$ of the search directions satisfy some useful boundedness properties. In the next section these properties will be shown to imply that, with a good line search, the objective function will be significantly decreased along those directions.

THEOREM 2.1. *Let $\{B_k\}$ be generated by the* BFGS *formula* (1.4), *where $B_1$ is symmetric and positive definite and where, for all $k \geq 1$, $y_k$ and $s_k$ satisfy*

$$\frac{y_k^T s_k}{s_k^T s_k} \geq m > 0, \tag{2.10}$$

$$\frac{\|y_k\|^2}{y_k^T s_k} \leq M. \tag{2.11}$$

*Then for any* $p \in (0,1)$ *there exist constants* $\beta_1, \beta_2, \beta_3 > 0$ *such that, for any* $k > 1$, *the relations*

$$\cos\theta_j \geq \beta_1, \tag{2.12}$$

$$\beta_2 \leq q_j \leq \beta_3, \tag{2.13}$$

$$\beta_2 \leq \frac{\|B_j s_j\|}{\|s_j\|} \leq \frac{\beta_3}{\beta_1} \tag{2.14}$$

*hold for at least* $\lceil pk \rceil$ *values of* $j \in [1,k]$.

*Proof.* From (2.9), (2.10), and (2.11)

$$\psi(B_{k+1}) \leq \psi(B_1) + (M - 1 - \ln m)k + \sum_{j=1}^{k} \left( \ln\cos^2\theta_j + 1 - \frac{q_j}{\cos^2\theta_j} + \ln\frac{q_j}{\cos^2\theta_j} \right).$$

Let us define $\eta_j \geq 0$ by

$$\eta_j = -\ln\cos^2\theta_j - \left[ 1 - \frac{q_j}{\cos^2\theta_j} + \ln\frac{q_j}{\cos^2\theta_j} \right]. \tag{2.15}$$

Since $\psi(B_{k+1}) > 0$ we have

$$\frac{1}{k}\sum_{j=1}^{k} \eta_j < \frac{\psi(B_1)}{k} + (M - 1 - \ln m). \tag{2.16}$$

Let us now define $J_k$ to be a set consisting of the $\lceil pk \rceil$ indices corresponding to the $\lceil pk \rceil$ smallest values of $\eta_j$, for $j \leq k$, and let $\eta_{m_k}$ denote the largest of the $\eta_j$ for $j \in J_k$. Then

$$\frac{1}{k}\sum_{j=1}^{k} \eta_j \quad \geq \quad \frac{1}{k}\left[ \eta_{m_k} + \sum_{\substack{j=1 \\ j \notin J_k}}^{k} \eta_j \right] \geq \eta_{m_k}(1-p).$$

Thus from (2.16) we have that, for all $j \in J_k$,

$$\eta_j < \frac{1}{1-p}\left[ \psi(B_1) + M - 1 - \ln m \right] \equiv \beta_0. \tag{2.17}$$

Since the term inside the brackets in (2.15) is less than or equal to zero, we conclude from (2.17) and (2.15) that for all $j \in J_k$

$$-\ln\cos^2\theta_j < \beta_0.$$

Therefore

$$\cos\theta_j > e^{-\beta_0/2} \equiv \beta_1,$$

which proves (2.12). Likewise, from (2.17) and (2.15) we have that, for all $j \in J_k$,

$$1 - \frac{q_j}{\cos^2 \theta_j} + \ln \frac{q_j}{\cos^2 \theta_j} > -\beta_0.$$

Note that the function

$$u(t) = 1 - t + \ln t, \tag{2.18}$$

is nonpositive for all $t > 0$, achieves its maximum value at $t = 1$, and satisfies $u(t) \to -\infty$ both as $t \to 0$ and as $t \to \infty$. Therefore it follows that for all $j \in J_k$

$$0 < \tilde{\beta}_2 \leq \frac{q_j}{\cos^2 \theta_j} \leq \beta_3,$$

for some constants $\tilde{\beta}_2$ and $\beta_3$. Therefore using (2.12) we obtain

$$\beta_2 \equiv \beta_1^2 \tilde{\beta}_2 \leq q_j \leq \beta_3.$$

Finally, since

$$\frac{\|B_j s_j\|}{\|s_j\|} = \frac{q_j}{\cos \theta_j},$$

we have for $j \in J_k$

$$\beta_2 \leq \frac{\|B_j s_j\|}{\|s_j\|} \leq \frac{\beta_3}{\beta_1}. \qquad \square$$

Since the conditions on $\{s_k\}$ and $\{y_k\}$ are not very restrictive, this theorem is relevant to both constrained and unconstrained optimization. Theorem 2.1 is fundamental for the analysis presented in the following section, as well as for our study of reduced Hessian methods for constrained optimization [1]. In the latter, $y_k$ represents the difference of reduced gradients of the Lagrangian, and $s_k$ is the component of the displacement lying in the tangent space of the constraints.

The set of iterates for which (2.12)–(2.14) hold will be called the "set of good iterates." Theorem 2.1 states that if the update conditions (2.10)–(2.11) are satisfied then a fraction $p$ of the iterates are good iterates. Since $p$ can be chosen to be close to 1, we can assume that most of the iterates are good iterates.

**3. Convergence of the BFGS method.** We will now use the tools and results of §2 to study the convergence of quasi-Newton methods with line searches. To facilitate the application of the results of this section in other contexts, we present them in a very general form, and then apply them to the BFGS method. We first introduce some notation and state the assumptions we make about the objective function $f$ and the steplength $\alpha_k$.

The matrix of second derivatives of $f$ will be denoted by $G$. The starting point for the algorithm is $x_1$, and we define the level set $D = \{x \in \mathbf{R}^n : f(x) \leq f(x_1)\}$.

*Assumptions* 3.1. (1) The objective function $f$ is twice continuously differentiable. (2) The level set $D$ is convex, and there exist positive constants $m$ and $M$ such that

$$m\|z\|^2 \leq z^T G(x) z \leq M\|z\|^2 \tag{3.1}$$

for all $z \in \mathbf{R}^n$ and all $x \in D$. Note that this implies that $f$ has a unique minimizer $x_*$ in $D$.

We want our analysis to cover a large class of line search strategies. Therefore we will assume that the procedure for choosing $\alpha_k$ in (1.2) is such that at each iteration either

$$f(x_k + \alpha_k d_k) - f(x_k) \leq -\eta_1 \frac{(g_k^T d_k)^2}{\|d_k\|^2}, \tag{3.2}$$

or

$$f(x_k + \alpha_k d_k) - f(x_k) \leq \eta_2 g_k^T d_k, \tag{3.3}$$

where $\eta_1$ and $\eta_2$ are positive constants. Condition (3.2) is fairly well known. It is a special case of what Ortega and Rheinboldt [8] call a *principle of sufficient decrease*; Warth and Werner [14] call a line search algorithm satisfying (3.2) an *efficient line search*. In these works and in Werner [15], it is shown that several well-known line search conditions, including the Curry-Altman condition, the Goldstein condition and the Wolfe condition, imply (3.2), if the gradient is Lipschitz continuous.

However, condition (3.2) is not always satisfied by backtracking methods such as the Goldstein-Armijo procedure. By a backtracking method we mean a method that uses $\alpha_k = 1$, if that satisfies

$$f(x_k + \alpha_k d_k) - f(x_k) \leq \eta \alpha_k g_k^T d_k, \tag{3.4}$$

for some constant $\eta$, and otherwise chooses repeatedly a smaller value of $\alpha_k$ by some procedure such as interpolation, until (3.4) is satisfied. In §4 we discuss backtracking and prove that, if the ratio between successive trial values of $\alpha$ is bounded away from zero, then the new iterate produced by a backtracking line search will satisfy either (3.2) or (3.3).

We will now show that if a fraction of the iterates are "good," in the sense of §2, and if the line search satisfies (3.2) or (3.3), then any quasi-Newton method is R-linearly convergent.

THEOREM 3.1. *Let $x_1$ be a starting point for which $f$ satisfies Assumptions* 3.1, *and suppose $\{x_k\}$ is generated by* (1.2)–(1.3), *where $\alpha_k$ is chosen so that either* (3.2) *or* (3.3) *is satisfied at each iterate. Suppose in addition that the matrices $B_k$ are positive definite and that there exist $p \in (0,1)$ and $\beta, \beta' > 0$, such that for any $k \geq 1$, the inequalities*

$$\cos\theta_j \geq \beta, \tag{3.5}$$

$$\frac{\|B_j s_j\|}{\|s_j\|} \leq \beta' \tag{3.6}$$

*hold for at least $\lceil pk \rceil$ values of $j \in [1,k]$. Then $\{x_k\} \to x_*$; moreover*

$$\sum_{k=1}^{\infty} \|x_k - x_*\| < \infty, \tag{3.7}$$

*and there is a constant $0 \leq r < 1$ such that*

$$f_{k+1} - f_* \leq r^k [f_1 - f_*] \tag{3.8}$$

*holds for all* $k$.

*Proof.* We define $J$ to be the set of indices for which (3.5) and (3.6) hold. Consider an iterate $x_j$ with $j \in J$. From (3.2)–(3.3) and from (3.5) and (3.6) we have that

$$f(x_j) - f(x_j + \alpha_j d_j) \geq \eta \|g_j\|^2, \tag{3.9}$$

where $\eta = \eta_1 \beta^2$ if the line search condition (3.2) holds, or $\eta = \eta_2 \beta / \beta'$ if (3.3) holds. Now from Assumptions 3.1 it is easy to see that, for all $k \geq 1$,

$$\frac{1}{2} m \|x_k - x_*\|^2 \leq f_k - f_* \leq \frac{1}{m} \|g_k\|^2. \tag{3.10}$$

(The lower bound follows immediately from (3.1); the upper bound is derived, for example, by Byrd, Nocedal, and Yuan [3, p. 1175]). From (3.9) and the upper bound in (3.10) we have that, for all $j \in J$,

$$f_{j+1} - f_* \leq r^{1/p}(f_j - f_*),$$

where $r^{1/p} \equiv (1 - \eta m)$. (Note that $1 - \eta m \geq 0$ since $\{f_k\}$ is a decreasing sequence.) Since $J \cap [1, k]$ has at least $\lceil pk \rceil$ elements, and since $\{f_k\}$ is decreasing

$$f_{k+1} - f_* \leq r^k (f_1 - f_*).$$

Using (3.10),

$$\begin{aligned} \sum_{k=1}^{\infty} \|x_k - x_*\| &\leq [2/m]^{1/2} \sum_{k=1}^{\infty} [f_k - f_*]^{1/2} \\ &\leq [2(f_1 - f_*)/m]^{1/2} \sum_{k=0}^{\infty} (r^{1/2})^k \\ &< \infty. \end{aligned}$$

□

Note that by (3.10), the R-linear convergence of $f$ in (3.8) implies that $\|x_{k+1} - x_*\| \leq r^{k/2}[2(f_1 - f_*)/m]^{1/2}$ so that the sequence $\{x_k\}$ is R-linearly convergent also.

Now we combine Theorem 2.1 and Theorem 3.1 to prove R-linear convergence for the BFGS update.

COROLLARY 3.1. *Let* $x_1$ *be a starting point for which* $f$ *satisfies Assumptions* 3.1. *Then for any positive definite* $B_1$, *the* BFGS *algorithm,* (1.2)–(1.4), *with a line search satisfying* (3.2) *or* (3.3) *at each step, generates iterates that converge to* $x_*$. *Moreover*

$$\sum_{k=1}^{\infty} \|x_k - x_*\| < \infty,$$

*and there is a constant* $0 \leq r < 1$ *such that*

$$f_{k+1} - f_* \leq r^k [f_1 - f_*]$$

*holds for all* $k$.

*Proof.* We just need to show that the hypotheses of Theorem 2.1 are satisfied. If we define

$$\overline{G}_k = \int_0^1 G(x_k + \tau s_k) d\tau, \tag{3.11}$$

then

$$y_k = \overline{G}_k s_k. \tag{3.12}$$

Thus (3.1) and (3.12) give

$$m\|s_k\|^2 \le y_k^T s_k \le M\|s_k\|^2, \tag{3.13}$$

and

$$\frac{\|y_k\|^2}{y_k^T s_k} \le \frac{s_k^T \overline{G}_k^2 s_k}{s_k^T \overline{G}_k s_k} \le M. \tag{3.14}$$

Then, by Theorem 2.1 we have that the matrices $B_k$ satisfy the hypotheses of Theorem 3.1 and the result follows. □

Powell [10] was the first to prove a linear convergence result of this type for the BFGS method, using a line search satisfying the Wolfe conditions. Werner [15] extended Powell's result to any line search satisfying (3.2). Corollary 3.1 is therefore an extension of Werner's result to a larger class of line search methods which, as we will see in §4, includes backtracking.

If the iterates $\{x_k\}$ are converging to a point $x_*$ it is to be expected that $y_k$ be approximately equal to $G(x_*)s_k$. In the next theorem we show that if $y_k$ and $s_k$ are so related then the Dennis–Moré condition for superlinear convergence is satisfied.

THEOREM 3.2. *Let $\{B_k\}$ be generated by the* BFGS *formula* (1.4), *where $B_1$ is symmetric and positive definite, and where $y_k^T s_k > 0$ for all $k$. Furthermore assume that $\{s_k\}$ and $\{y_k\}$ are such that*

$$\frac{\|y_k - G_* s_k\|}{\|s_k\|} \le \epsilon_k, \tag{3.15}$$

*for some symmetric and positive definite matrix $G_*$, and for some sequence $\{\epsilon_k\}$ with the property $\sum_{k=1}^{\infty} \epsilon_k < \infty$. Then*

$$\lim_{k\to\infty} \frac{\|(B_k - G_*)s_k\|}{\|s_k\|} = 0, \tag{3.16}$$

*and the sequences $\{\|B_k\|\}$, $\{\|B_k^{-1}\|\}$ are bounded.*

*Proof.* Let us define

$$\tilde{s}_k = G_*^{1/2} s_k, \quad \tilde{y}_k = G_*^{-1/2} y_k, \tag{3.17}$$

$$\tilde{B}_k = G_*^{-1/2} B_k G_*^{-1/2}, \tag{3.18}$$

$$\cos\tilde{\theta}_k = \frac{\tilde{s}_k^T \tilde{B}_k \tilde{s}_k}{\|\tilde{B}_k \tilde{s}_k\| \, \|\tilde{s}_k\|}, \tag{3.19}$$

and

$$\tilde{q}_k = \frac{\tilde{s}_k^T \tilde{B}_k \tilde{s}_k}{\tilde{s}_k^T \tilde{s}_k}. \tag{3.20}$$

From (1.4), (3.17) and (3.18)

$$\tilde{B}_{k+1} = \tilde{B}_k - \frac{\tilde{B}_k \tilde{s}_k \tilde{s}_k^T \tilde{B}_k}{\tilde{s}_k^T \tilde{B}_k \tilde{s}_k} + \frac{\tilde{y}_k \tilde{y}_k^T}{\tilde{y}_k^T \tilde{s}_k}. \tag{3.21}$$

Thus we obtain, just as in (2.9),

$$\begin{aligned} \psi(\tilde{B}_{k+1}) &= \psi(\tilde{B}_k) + \frac{\|\tilde{y}_k\|^2}{\tilde{y}_k^T \tilde{s}_k} - 1 - \ln \frac{\tilde{y}_k^T \tilde{s}_k}{\tilde{s}_k^T \tilde{s}_k} + \ln \cos^2 \tilde{\theta}_k \\ &\quad + \left[1 - \frac{\tilde{q}_k}{\cos^2 \tilde{\theta}_k} + \ln \frac{\tilde{q}_k}{\cos^2 \tilde{\theta}_k}\right]. \end{aligned} \tag{3.22}$$

From the condition (3.15)

$$\frac{\|\tilde{y}_k - \tilde{s}_k\|}{\|\tilde{s}_k\|} \le \bar{c}\epsilon_k, \tag{3.23}$$

where $\bar{c} = \|G_*^{-1/2}\|^2$. Using the triangle inequality we obtain

$$(1 - \bar{c}\epsilon_k)\|\tilde{s}_k\| \le \|\tilde{y}_k\| \le (1 + \bar{c}\epsilon_k)\|\tilde{s}_k\|. \tag{3.24}$$

From (3.23) and (3.24)

$$(1 - \bar{c}\epsilon_k)^2\|\tilde{s}_k\|^2 - 2\tilde{y}_k^T \tilde{s}_k + \|\tilde{s}_k\|^2 \le \|\tilde{y}_k\|^2 - 2\tilde{y}_k^T \tilde{s}_k + \|\tilde{s}_k\|^2 \le \bar{c}^2\epsilon_k^2\|\tilde{s}_k\|^2,$$

from which we obtain

$$\frac{\tilde{y}_k^T \tilde{s}_k}{\tilde{s}_k^T \tilde{s}_k} \ge 1 - \bar{c}\epsilon_k. \tag{3.25}$$

Using (3.24),

$$\frac{\|\tilde{y}_k\|^2}{\tilde{y}_k^T \tilde{s}_k} \le (1 + \bar{c}\epsilon_k)^2 \frac{\|\tilde{s}_k\|^2}{\tilde{y}_k^T \tilde{s}_k},$$

and therefore from (3.25) we have that for $k$ sufficiently large there is a constant $c > \bar{c}$ such that

$$\frac{\|\tilde{y}_k\|^2}{\tilde{y}_k^T \tilde{s}_k} \le 1 + c\epsilon_k.$$

If we assume that $c\epsilon_k < \frac{1}{2}$ then $\ln(1 - \bar{c}\epsilon_k) \ge -2c\epsilon_k$. Thus from (3.22) we have, for $k$ large enough,

$$\psi(\tilde{B}_{k+1}) \le \psi(\tilde{B}_k) + 3c\epsilon_k + \ln \cos^2 \tilde{\theta}_k + \left[1 - \frac{\tilde{q}_k}{\cos^2 \tilde{\theta}_k} + \ln \frac{\tilde{q}_k}{\cos^2 \tilde{\theta}_k}\right].$$

Summing over the iterates we obtain

$$0 < \psi(\tilde{B}_{k+1}) \le \psi(\tilde{B}_1) + \sum_{j=1}^{k} \left(3c\epsilon_j + \ln \cos^2 \tilde{\theta}_j + \left[1 - \frac{\tilde{q}_j}{\cos^2 \tilde{\theta}_j} + \ln \frac{\tilde{q}_j}{\cos^2 \tilde{\theta}_j}\right]\right) + \hat{c},$$

for some positive constant $\hat{c}$. From the condition $\sum_{k=1}^{\infty} \epsilon_k < \infty$, we see that $\{\psi(\tilde{B}_k)\}$ is bounded, and that

$$\ln \cos^2 \tilde{\theta}_k \to 0,$$

and

$$\left[1 - \frac{\tilde{q}_k}{\cos^2 \tilde{\theta}_k} + \ln \frac{\tilde{q}_k}{\cos^2 \tilde{\theta}_k}\right] \to 0,$$

since both these quantities are nonpositive for all $k$. These relations, and the comments following (2.18), imply that

$$\lim_{k\to\infty} \cos \tilde{\theta}_k = \lim_{k\to\infty} \tilde{q}_k = 1. \tag{3.26}$$

Now

$$\begin{aligned} \frac{\|G_*^{-1/2}(B_k - G_*)s_k\|^2}{\|G_*^{1/2}s_k\|^2} &= \frac{\|(\tilde{B}_k - I)\tilde{s}_k\|^2}{\|\tilde{s}_k\|^2} \\ &= \frac{\|\tilde{B}_k\tilde{s}_k\|^2 - 2\tilde{s}_k^T\tilde{B}_k\tilde{s}_k + \tilde{s}_k^T\tilde{s}_k}{\tilde{s}_k^T\tilde{s}_k} \\ &= \frac{\tilde{q}_k^2}{\cos \tilde{\theta}_k^2} - 2\tilde{q}_k + 1. \end{aligned} \tag{3.27}$$

It is clear from (3.26) that this quantity converges to zero, which implies (3.16). Since $\{\psi(\tilde{B}_k)\}$ is bounded, (2.6) implies that $\{\|B_k\|\}$ and $\{\|B_k^{-1}\|\}$ are bounded. □

We introduced $\tilde{s}_k$, $\tilde{y}_k$, and $\tilde{B}_k$ to present the proof of this theorem in its most general form. However, due to the scale invariance properties of the BFGS method, these new quantities can be avoided by rescaling the problem so that $G_*$ becomes the identity matrix. We note also, from (3.27), that condition (3.26) is equivalent to the Dennis–Moré condition (3.16).

To apply this result to the case when $\{y_k\}$ and $\{s_k\}$ are given by the BFGS method, we make the further assumption that the Hessian matrix $G$ is Lipschitz continuous at $x_*$, i.e., that there exists a positive constant $L$ such that

$$\|G(x) - G(x_*)\| \le L\|x - x_*\|, \tag{3.28}$$

for all $x$ in a neighborhood of $x_*$. (The weaker assumption that $G$ be Hölder continuous at $x_*$ would suffice; we have assumed Lipschitz continuity for simplicity.) With (3.12) this assumption implies that

$$\frac{\|y_k - G(x_*)s_k\|}{\|s_k\|} \le L \max\{\|x_{k+1} - x_*\|, \|x_k - x_*\|\}. \tag{3.29}$$

We can now apply Theorem 3.2 with $\epsilon_k = L\max\{\|x_{k+1} - x_*\|, \|x_k - x_*\|\}$ and $G_* = G(x_*)$, since $\sum_{k=1}^{\infty} \epsilon_k < \infty$ by Corollary 3.1. This establishes the following result for the BFGS method.

COROLLARY 3.2. *Let $x_1$ be a starting point for which $f$ satisfies Assumptions* 3.1, *and assume that the Hessian matrix $G(x)$ is Lipschitz continuous at $x_*$. Then for any* *positive definite* $B_1$*, the* BFGS *algorithm,* (1.2)–(1.4)*, with a line search satisfying* (3.2) *or* (3.3) *at each step, gives the Dennis-Moré condition*

$$\lim_{k\to\infty} \frac{\|(B_k - G_*)s_k\|}{\|s_k\|} = 0.$$

From this corollary and from the well-known characterization result of Dennis and Moré [5], we conclude that the rate of convergence is Q-superlinear if the line search algorithm sets $\alpha_k = 1$ for all sufficiently large $k$. Several practical line searches will satisfy this condition if the unit steplength is always tried first. Indeed, Dennis and Moré [6] show that when $\|(B_k - G_*)s_k\|/\|s_k\|$ and $\|x_k - x_*\|$ are sufficiently small then the steplength $\alpha_k = 1$ satisfies the Wolfe conditions. Similar arguments show that this is true also for the Goldstein conditions, and that a backtracking line search will set $\alpha_k = 1$ under these conditions.

Theorem 3.2 is essentially the same as Theorem 8.8 of [6]. However, it is interesting that this result can be proved easily with the same tools that were used for the global convergence analysis.

**4. Backtracking line search.** In this section we show that a backtracking line search satisfies conditions (3.2) or (3.3), and therefore the convergence results of §3 apply when this type of line search is used. By a backtracking line search we mean an algorithm of the following form for computing $\alpha$.

ALGORITHM 4.1. The constants $\eta \in (0,1)$ and $\tau, \tau'$, with $0 < \tau < \tau' < 1$, are given.

(1) Set $\alpha = 1$.

(2) Test the relation

$$f(x_k + \alpha d_k) \le f(x_k) + \eta \alpha g_k^T d_k. \tag{4.1}$$

(3) If (4.1) is not satisfied, choose a new $\alpha$ in $[\tau\alpha, \tau'\alpha]$ and go to (2). If (4.1) is satisfied, set $\alpha_k = \alpha$ and $x_{k+1} = x_k + \alpha_k d_k$.

Several procedures have been used to choose a new trial value of $\alpha$ in step (3). The classical Goldstein-Armijo method is to simply multiply the old value of $\alpha$ by $\frac{1}{2}$ or some other constant in (0,1). An alternative is to choose the new $\alpha$ as the minimizer of a polynomial interpolating already computed function and derivative values, subject to being in the interval $[\tau\alpha, \tau'\alpha]$.

LEMMA 4.1. *Under Assumptions* 3.1 *there exist positive constants* $\eta_1$ *and* $\eta_2$ *such that, for any* $x_k$ *and any* $d_k$ *with* $g_k^T d_k < 0$*, the steplength* $\alpha_k$ *produced by Algorithm* 4.1 *will satisfy either*

$$f(x_k + \alpha_k d_k) - f(x_k) \le -\eta_1 \frac{(g_k^T d_k)^2}{\|d_k\|^2}, \tag{4.2}$$

*or*

$$f(x_k + \alpha_k d_k) - f(x_k) \le \eta_2 g_k^T d_k. \tag{4.3}$$

*Proof.* If (4.1) is satisfied by $\alpha_k = 1$ then (4.3) holds with $\eta_2 \equiv \eta$. Suppose that $\alpha_k < 1$, which means that (4.1) failed for a steplength $\alpha' \leq \alpha_k/\tau$:

$$f(x_k + \alpha' d_k) - f(x_k) > \eta \alpha' g_k^T d_k.$$

Then, using the mean value theorem, we obtain

$$g(x_k + \theta \alpha' d_k)^T \alpha' d_k > \eta \alpha' g_k^T d_k,$$

where $\theta \in (0,1)$. Thus from (3.1)

$$(\eta - 1)\alpha' g_k^T d_k < \alpha'[g(x_k + \theta\alpha' d_k) - g_k]^T d_k \leq M(\alpha' \|d_k\|)^2,$$

which implies that

$$\alpha_k \geq \tau\alpha' > \tau(1-\eta)\frac{-g_k^T d_k}{M\|d_k\|^2}.$$

Substituting this into (4.1) we have

$$f(x_k + \alpha_k d_k) - f(x_k) \leq -\frac{\tau\eta(1-\eta)}{M}\frac{(g_k^T d_k)^2}{\|d_k\|^2},$$

which gives (4.2). □

**5. Final remarks.** We have shown in this paper that the BFGS method with any of a wide variety of line searches is globally and superlinearly convergent on uniformly convex objective functions.

These results go beyond those of Powell [10] and Werner [15] in that the class of allowable line searches includes backtracking strategies. In addition, this paper has introduced some tools that we believe will prove to be quite useful for the analysis of quasi-Newton methods.

The function $\psi$, defined by (2.5), and in its scaled form, $\psi(G_*^{1/2} B G_*^{1/2})$, is a useful measure of the ability of the matrices $B_k$ to generate good search directions, and it obeys a simple and useful recurrence under BFGS updating. Since it enables one to consider simultaneously the effect of both small and large eigenvalues of $B$, it allows us to put the global analysis of Powell into a framework similar to the one employed by Dennis and Moré in their superlinear convergence analysis, with $\psi$ playing the role of the weighted Frobenius norm (see the comments following the recursion (2.9)).

Theorem 2.1 and the function $\psi$ are also useful in other contexts. Griewank [7] uses $\psi$ to analyze the partitioned BFGS algorithm for partially separable optimization, and Byrd, Liu, and Nocedal [2] use it to show that global and superlinear convergence can be obtained if the scalar parameter in Broyden's class is allowed to vary in the interval [-c,1), where $c > 0$ is a constant that depends on the objective function.

Theorem 2.1 is invoked by Byrd, Schnabel, and Shultz [4] in their analysis of unconstrained optimization algorithms intended for parallel computation in which extra BFGS updates are made at each iteration. Finally, the results of this paper are also applicable in the context of constrained optimization. Byrd and Nocedal [1] use Theorem 2.1, as well as Theorem 3.2, to show convergence for reduced Hessian methods for constrained nonlinear optimization.

**Acknowledgments.** We would like to acknowledge several helpful conversations with John Dennis. We also thank Rice University and Argonne National Laboratory for providing us with excellent research environments.

## REFERENCES

[1] R. H. BYRD AND J. NOCEDAL, *An analysis of reduced Hessian methods for constrained optimization*, Technical Report CU-CS-398-88, Computer Science Dept., University of Colorado, Boulder, CO, 1988.

[2] R. H. BYRD, D. C. LIU, AND J. NOCEDAL, *On the behavior of Broyden's class of quasi-Newton methods*, Technical Report NAM-01, Dept. of Electrical Engineering and Computer Science, Northwestern University, Evanston, Illinois, 1989.

[3] R. H. BYRD, J. NOCEDAL, AND Y. YUAN, *Global convergence of a class of quasi-Newton methods on convex problems*, SIAM J. Numer. Anal., 24 (1987), pp. 1171-1189.

[4] R. H. BYRD, R. B. SCHNABEL, AND G. A. SHULTZ, *Parallel quasi-Newton methods for unconstrained optimization*, Technical Report CU-CS-396-88, Computer Science Dept., University of Colorado, Boulder, CO, 1988.

[5] J. E. DENNIS AND J. J. MORÉ, *A characterization of superlinear convergence and its application to quasi-Newton methods*, Math. Comput., 28 (1974), pp. 549-560.

[6] ——, *Quasi-Newton methods, motivation and theory*, SIAM Rev., 19 (1977), pp. 46-89.

[7] A. GRIEWANK, *The global convergence of partitioned BFGS on semi-smooth problems with convex decompositions*, ANL/MCS-TM-105, Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, IL, 1987.

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

[9] J. D. PEARSON, *Variable metric methods of minimization*, Computer J., 12 (1969), pp. 171-178.

[10] M. J. D. POWELL, *Some global convergence properties of a variable metric algorithm for minimization without exact line searches*, in Nonlinear Programming, SIAM-AMS Proceedings, Vol. IX, R.W. Cottle and C.E. Lemke, eds., American Mathematical Society, Providence, RI, 1976.

[11] K. RITTER, *Local and superlinear convergence of a class of variable metric methods*, Computing, 23 (1979), pp. 287-297.

[12] ——, *Global and superlinear convergence of a class of variable metric methods*, Math. Programming Stud., 14 (1981), pp. 178-205.

[13] Ph. L. TOINT, *Global convergence of the partitioned BFGS algorithm for convex partially separable optimization*, Math. Programming, 36 (1986), pp. 290-306.

[14] W. WARTH AND J. WERNER, *Effiziente Schrittweitenfunktionen bei unrestringierten Optimierungsaufgaben*, Computing, 19 (1977), pp. 59-72.

[15] J. WERNER, *Über die globale konvergenz von Variable-Metric Verfahren mit nichtexakter Schrittweitenbestimmung*, Numer. Math., 31 (1978), pp. 321-334.