Mathematical Programming 61 (1993) 19–37
North-Holland

Analysis of a self-scaling quasi-Newton method

Jorge Nocedal

Department of Electrical Engineering and Computer Science, Northwestern University, Evanston, IL, USA

Ya-xiang Yuan*

Computing Center, Academia Sinica, Beijing, China

Received 25 September 1991
Revised manuscript received 21 February 1992

We study the self-scaling BFGS method of Oren and Luenberger (1974) for solving unconstrained optimization problems. For general convex functions, we prove that the method is globally convergent with inexact line searches. We also show that the directions generated by the self-scaling BFGS method approach Newton's direction asymptotically. This would ensure superlinear convergence if, in addition, the search directions were well-scaled, but we show that this is not always the case. We find that the method has a major drawback: to achieve superlinear convergence it may be necessary to evaluate the function twice per iteration, even very near the solution. An example is constructed to show that the step-sizes required to achieve a superlinear rate converge to 2 and 0.5 alternately.

Key words: Self-scaling, BFGS method, quasi-Newton method, optimization.

1. Introduction

We analyze the convergence properties of a self-scaling BFGS method for solving the unconstrained optimization problem

$$\min f(x), \tag{1.1}$$

where f is a smooth function of n variables. At the kth iteration of the self-scaling method, a symmetric and positive definite matrix B_k is given, and a search direction is computed by

$$d_k = -B_k^{-1} g_k, \tag{1.2}$$

where g_k is the gradient of f evaluated at the current iterate x_k. One then computes the next iterate by

$$x_{k+1} = x_k + \alpha_k d_k, \tag{1.3}$$

Correspondence to: Prof. Jorge Nocedal, Department of Electrical Engineering and Computer Science, Northwestern University, Evanston, IL 60208, USA.
This work was supported by National Science Foundation Grant CCR-9101359, and by the Department of Energy Grant DE-FG02-87ER25047.
* This work was performed while the author was visiting Northwestern University.

where the stepsize α_k satisfies the Wolfe conditions:

$$f(x_k+\alpha_k d_k) \leq f(x_k)+\delta_1 \alpha_k d_k^{\mathrm{T}} g_k, \tag{1.4}$$

and

$$g(x_k+\alpha_k d_k)^{\mathrm{T}} d_k \geq \delta_2 d_k^{\mathrm{T}} g_k, \tag{1.5}$$

where $0<\delta_1<\frac{1}{2}$ and $\delta_1<\delta_2<1$. The Hessian approximation is then updated by

$$B_{k+1}=\phi_k\left[B_k-\frac{B_k s_k s_k^{\mathrm{T}} B_k}{s_k^{\mathrm{T}} B_k s_k}\right]+\frac{y_k y_k^{\mathrm{T}}}{y_k^{\mathrm{T}} s_k}, \tag{1.6}$$

where

$$\phi_k=\frac{y_k^{\mathrm{T}} s_k}{s_k^{\mathrm{T}} B_k s_k}, \tag{1.8}$$

which is one of the self-scaling quasi-Newton methods proposed by Oren and Luenberger (1974). If $\phi_k=1$, this method reduces to the standard BFGS method. Other choices for ϕ_k are given in Luenberger (1984) and Oren (1982), and in the references therein. In this paper we only study (1.8) because it simplifies the analysis and because we believe that other choices of ϕ_k do not possess stronger convergence properties than (1.8).

Self-scaling methods were derived, motivated and analyzed in the context of unconstrained minimization of quadratic functions. They also arise when deriving secant methods by variational means (Dennis and Wolkowicz, 1991). In this paper we consider the behavior of one of these methods on general convex problems and show that it can be implemented so as to be globally and superlinearly convergent, but we find that the superlinear rate can normally be obtained only at an additional computational expense. We construct an example that illustrates this.

2. Global convergence analysis

In this section we show that the self-scaling BFGS method with an inexact line search is globally convergent on general convex functions. The analysis is based on the study of the trace and determinant expressions for the matrices B_k — an approach due to Powell (1976).

The analysis is greatly simplified by defining the auxiliary sequence $\{\bar{B}_k\}$ by

$$\bar{B}_1=B_1, \tag{2.1}$$

$$\bar{B}_{k+1}=\frac{s_k^{\mathrm{T}} \bar{B}_k s_k}{y_k^{\mathrm{T}} s_k} B_{k+1}, \quad k \geq 1. \tag{2.2}$$

Since $\bar{B}_k$ and B_k are related by a constant factor, we can express the step s_k in terms of $\bar{B}_k$. Indeed, using (1.2), (1.3), (1.7) and (2.2), we obtain, for $k>1$,

$$s_k=-\bar{\alpha}_k \bar{B}_k^{-1} g_k, \tag{2.3}$$

where

$$\bar{\alpha}_k = \alpha_k \frac{s_{k-1}^T \bar{B}_{k-1} s_{k-1}}{y_{k-1}^T s_{k-1}}. \tag{2.4}$$

We now find a recursion for $\bar{B}_k$. From (2.2), (1.6) and (1.8) we obtain

$$\bar{B}_{k+1} = \frac{s_k^T \bar{B}_k s_k}{y_k^T s_k} \frac{y_k^T s_k}{s_k^T B_k s_k} \left[B_k - \frac{B_k s_k s_k^T B_k}{s_k^T B_k s_k} \right] + \frac{s_k^T \bar{B}_k s_k}{y_k^T s_k} \frac{y_k y_k^T}{y_k^T s_k}, \tag{2.5}$$

and using (2.2) gives

$$\bar{B}_{k+1} = \left(\frac{s_k^T B_k s_k}{y_k^T s_k} \frac{y_k^T s_k}{s_k^T B_k s_k} \right) \frac{s_{k-1}^T \bar{B}_{k-1} s_{k-1}}{y_{k-1}^T s_{k-1}} \left[B_k - \frac{B_k s_k s_k^T B_k}{s_k^T B_k s_k} \right]$$
$$+ \frac{s_k^T \bar{B}_k s_k}{y_k^T s_k} \frac{y_k y_k^T}{y_k^T s_k}.$$

Thus

$$\bar{B}_{k+1} = \bar{B}_k - \frac{\bar{B}_k s_k s_k^T \bar{B}_k}{s_k^T \bar{B} s_k} + \frac{s_k^T \bar{B}_k s_k}{y_k^T s_k} \frac{y_k y_k^T}{y_k^T s_k}. \tag{2.6}$$

From the determinant relation for the BFGS formula (see Pearson (1969)) we obtain

$$\det(\bar{B}_{k+1}) = \det(\bar{B}_k) \frac{s_k^T \bar{B}_k s_k}{y_k^T s_k} \frac{y_k^T s_k}{s_k^T \bar{B}_k s_k}.$$
$$= \det(\bar{B}_k).$$

We have thus found that the determinant of the matrices $\bar{B}_k$ stays constant,

$$\det(\bar{B}_{k+1}) = \det(\bar{B}_1). \tag{2.7}$$

If B_1 is positive definite and $s_k^T y_k > 0$ for all k, we can easily see from (1.6) and (1.8) that all the matrices B_k generated by the self-scaling BFGS method are positive definite, and consequently, from relations (2.1)–(2.2), that all $\bar{B}_k$ are also positive definite.

Next we study the trace relation. To do this we define the scalars

$$q_k = \frac{s_k^T \bar{B}_k s_k}{s_k^T s_k}, \qquad \cos \theta_k = \frac{s_k^T \bar{B}_k s_k}{\|\bar{B}_k s_k\| \, \|s_k\|}, \tag{2.8}$$

$$m_k = \frac{y_k^T s_k}{s_k^T s_k}, \qquad M_k = \frac{\|y_k\|^2}{y_k^T s_k}, \tag{2.9}$$

where here, and for the rest of the paper, $\|\cdot\|$ denotes the ℓ_2 vector or matrix norm. Since B_k and $\bar{B}_k$ differ only by a constant factor, it is clear that $\cos \theta_k$ remains unchanged if we replace $\bar{B}_k$ by B_k. Thus θ_k is the angle between the steepest descent direction $-g_k$ and the search direction d_k of the self-scaling BFGS method, and the convergence properties of the method can be deduced directly from the behavior of $\cos\theta_k$.

From (2.6) we obtain the trace relation

$$\mathrm{tr}(\bar{B}_{k+1}) = \mathrm{tr}(\bar{B}_k) - \frac{\|\bar{B}_k s_k\|^2}{s_k^T\bar{B}_k s_k} + \frac{s_k^T\bar{B}_k s_k}{y_s^T s_k}\frac{\|y_k\|^2}{y_k^T s_k} \tag{2.10}$$

$$= \mathrm{tr}(\bar{B}_k) - \frac{q_k}{\cos^2\theta_k} + \frac{q_k}{m_k}M_k. \tag{2.11}$$

We are now ready to establish a global convergence result, under the following assumptions. The Hessian matrix of f is denoted by G.

Assumptions 2.1. (1) The objective function f is convex, bounded below, and twice continuously differentiable in $\mathbb{R}^n$. (2) The Hessian matrix $G(x)$ is bounded above in norm for all $x\in D=\{x\in\mathbb{R}^n: f(x)\leqslant f(x_1)\}$.

Theorem 2.1. *Let x_1 be a starting point for which Assumptions* 2.1 *are satisfied. Then for any positive definite starting matrix B_1, the self-scaling BFGS method* (1.2)-(1.8) *is globally convergent, i.e.*

$$\liminf_{k\to\infty}\|g_k\| = 0. \tag{2.12}$$

Proof. It is easy to see (Powell, 1976, Lemma 1) that Assumptions 2.1 imply that $\{M_k\}$ is bounded above by some positive constant M. The line search condition (1.5) implies that

$$s_k^T y_k \geqslant -(1-\delta_2)s_k^T g_k. \tag{2.13}$$

Therefore, from (2.10), (2.13) and (2.3), and defining $\delta_3 = M(1-\delta_2)^{-1}$, we have

$$\begin{aligned}\mathrm{tr}(\bar{B}_{k+1}) &\leqslant \mathrm{tr}(\bar{B}_k) + M\frac{s_k^T\bar{B}_k s_k}{-(1-\delta_2)s_k^T g_k}\\ &= \mathrm{tr}(\bar{B}_k) + \delta_3\bar{\alpha}_k\\ &= \mathrm{tr}(\bar{B}_k)[1+\delta_3\bar{\alpha}_k/\mathrm{tr}(\bar{B}_k)].\end{aligned} \tag{2.14}$$

Therefore

$$\mathrm{tr}(\bar{B}_{k+1}) \leqslant \mathrm{tr}(B_1)\prod_{i=1}^{k}[1+\delta_3\bar{\alpha}_i/\mathrm{tr}(\bar{B}_i)]. \tag{2.15}$$

Now, from (1.4),

$$f_{k+1} - f_1 \leqslant \delta_1\sum_{i=1}^{k} s_i^T g_i,$$

and since f is bounded below, we obtain

$$-\sum_{k=1}^{\infty} s_k^{\mathrm{T}} g_k < \infty. \tag{2.16}$$

We now proceed by contradiction and assume that

$$\liminf_{k\to\infty} \|g_k\| \neq 0. \tag{2.17}$$

Then there exists a constant $\delta_4 > 0$ such that

$$\|g_k\| \geq \delta_4, \tag{2.18}$$

for all k. Then from (2.3) and (2.18) we have

$$-\sum_{k=1}^{\infty} s_k^{\mathrm{T}} g_k = \sum_{k=1}^{\infty} \bar{\alpha}_k g_k^{\mathrm{T}} \bar{B}_k^{-1} g_k \geq \delta_4^2 \sum_{k=1}^{\infty} \frac{\bar{\alpha}_k}{\mathrm{tr}(\bar{B}_k)},$$

since $\|\bar{B}_k\| \leq \mathrm{tr}(\bar{B}_k)$. Thus by (2.16),

$$\sum_{k=1}^{\infty} \frac{\bar{\alpha}_k}{\mathrm{tr}(\bar{B}_k)} < \infty. \tag{2.19}$$

A basic result of infinite products (see for example Apostol, 1957) states that if $\{\gamma_k\}$ is a sequence of positive numbers, then

$$\sum_{k=1}^{\infty} \gamma_k < \infty \Rightarrow \prod_{k=1}^{\infty} (1+\gamma_k) < \infty. \tag{2.20}$$

Using this result, (2.19) and (2.15) we see that the trace of $\bar{B}_k$ is bounded. Since we have also shown that the determinant of $\bar{B}_k$ is bounded away from zero, we conclude that both $\{\bar{B}_k\}$ and $\{\bar{B}_k^{-1}\}$ are uniformly bounded above. Moreover, by (2.8), we conclude that $\cos\theta_k$ is bounded away from zero. This and assumption (2.18) give a contradiction, since Zoutendijk's result (see Zoutendijk, 1970, or Wolfe, 1969, 1971) states that any method of the form (1.2)–(1.5) satisfies

$$\sum_{k=1}^{\infty} \cos^2\theta_k \|g_k\|^2 < \infty.$$

The contradiction shows that $\liminf \|g_k\| = 0$. □

Thus we have proved that the self-scaling BFGS method, with inexact line searches, is globally convergent on general convex functions.

3. Linear convergence

We make the following assumptions on the objective function.

Assumptions 3.1. (1) The objective function f is twice continuously differentiable. (2) The level set $D = \{x \in \mathbb{R}^n : f(x) \leq f(x_1)\}$ is convex, and there exist positive constants m and M such that

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

for all $z \in \mathbb{R}^n$ and all $x \in D$. Note that this implies that f has a unique minimizer x_* in D. (3) The Hessian matrix G is Lipschitz continuous at x_*.

Under these assumptions, Theorem 2.1 implies that $x_k \to x_*$. We now investigate the rate of convergence of the iteration.

Since (1.6) and (2.6) are invariant under the transformations

$$\tilde{s}_k = G_*^{1/2} s_k, \qquad \tilde{y}_k = G_*^{-1/2} y_k, \qquad \tilde{B}_k = G_*^{-1/2} \bar{B}_k G_*^{-1/2},$$

we can assume, without loss of generality, that $G_* = I$. Let us define

$$\varepsilon_k = x_k - x_*.$$

By the Mean Value Theorem,

$$\begin{aligned} y_k &= \int_0^1 G_k(x_k + \tau s_k)\, \mathrm{d}\tau\, s_k \\ &= \int_0^1 [G_k(x_k + \tau s_k) - G_*]\, \mathrm{d}\tau s_k + \int_0^1 G_*\, \mathrm{d}\tau\, s_k. \end{aligned} \tag{3.2}$$

Thus, since G is Lipschitz continuous at x_*, we have

$$\|y_k\| = \|s_k\|(1 + 0(\sigma_k)), \tag{3.3}$$

where

$$\sigma_k = \max\{\|\varepsilon_k\|, \|\varepsilon_{k+1}\|\}.$$

Also from (3.2),

$$y_k^{\mathrm{T}} s_k = \|s_k\|^2(1 + \mathrm{O}(\sigma_k)), \tag{3.4}$$

and using this in (2.9), we obtain

$$M_k / m_k = 1 + \mathrm{O}(\sigma_k). \tag{3.5}$$

Substituting this in (2.11)

$$\mathrm{tr}(\bar{B}_{k+1}) = \mathrm{tr}(\bar{B}_k) + q_k[1 + \mathrm{O}(\sigma_k) - 1/\cos^2\theta_k] \tag{3.6}$$

$$= \mathrm{tr}(\bar{B}_k) + q_k \mathrm{O}(\sigma_k) + q_k[1 - 1/\cos^2\theta_k]. \tag{3.7}$$

Since the last term is non-positive, and since

$$q_k \leqslant \mathrm{tr}(\bar{B}_k), \tag{3.8}$$

we obtain

$$\mathrm{tr}(\bar{B}_{k+1}) \leqslant \mathrm{tr}(\bar{B}_k)(1 + \mathrm{O}(\sigma_k)). \tag{3.9}$$

We can now establish a linear convergence result.

Theorem 3.1. *Let x_1 be a starting point for which f satisfies Assumptions* 3.1. *Then for any positive definite starting matrix B_1, the self-scaling BFGS method* (1.2)–(1.8) *generates a sequence $\{x_k\}$ that satisfies*

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

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

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

for all k.

Proof. Since we know that the iterates converge to x_*, we have that $\sigma_k \to 0$. Thus using (3.9) we have that for large k,

$$[\operatorname{tr}(\bar{B}_{k+1})]^n \leq \operatorname{tr}(\bar{B}_k)]^n (1 + a_1 \sigma_k), \tag{3.10}$$

for some constant a_1. On the other hand, the line search conditions (1.4)–(1.5) and the assumptions on f give (see for example Byrd, Nocedal and Yuan, 1987, p. 1175)

$$f_{k+1} - f_* \leq (1 - c_1 \cos^2 \theta_k)(f_k - f_*) \tag{3.11}$$

for some constant $0 < c_1 \leq 1$. Combining (3.10) and (3.11), asuming that $\cos^2 \theta_k \geq \frac{1}{2}$ and that k is large enough, we obtain

$$\begin{aligned}[\operatorname{tr}(\bar{B}_{k+1})]^{2n}(f_{k+1} - f_*) &\leq [\operatorname{tr}(\bar{B}_k)]^{2n}(f_k - f_*)(1 - \tfrac{1}{2}c_1)(1 + a_1\sigma_k)^2 \\ &\leq [\operatorname{tr}(\bar{B}_k)]^{2n}(f_k - f_*),\end{aligned} \tag{3.12}$$

since $\sigma_k \to 0$. We note that the relation (3.12) also holds when $\cos^2 \theta_k < \frac{1}{2}$, since in that case (3.6) implies that, for large k, $\operatorname{tr}(\bar{B}_{k+1}) < \operatorname{tr}(\bar{B}_k)$, and since $f_{k+1} \leq f_k$, by (3.11). Therefore, we see from (3.12) that the sequence $\{[\operatorname{tr}(\bar{B}_k)]^{2n}(f_k - f_*)\}$ is bounded, i.e. there exists a constant $\hat{m}$ such that

$$[\operatorname{tr}(\bar{B}_k)]^{2n}(f_k - f_*) \leq \hat{m}, \tag{3.13}$$

for all k. Moreover by Taylor's theorem

$$f_k - f_* \geq \tfrac{1}{2} m \|\varepsilon_k\|^2, \tag{3.14}$$

where m is the lower bound in (3.1). Using this in (3.13) we obtain

$$[\operatorname{tr}(\bar{B}_k)]^{2n} \|\varepsilon_k\|^2 \leq \bar{m} \tag{3.15}$$

where $\bar{m} = 2\hat{m}/m$. Similarly, from (3.13) and the condition $f_{k+1} \leq f_k$ we have that

$$[\operatorname{tr}(\bar{B}_k)]^{2n}(f_{k+1} - f_*) \leq \hat{m},$$

and thus applying (3.14) for $k+1$ we obtain

$$[\operatorname{tr}(\bar{B}_k)]^{2n} \|\varepsilon_{k+1}\|^2 \leq \bar{m}. \tag{3.16}$$

Using (3.15) and (3.16) in (3.10) we have

$$\begin{aligned}[\operatorname{tr}(\bar{B}_{k+1})]^n &\leq [\operatorname{tr}(\bar{B}_k)]^n + a_1[\operatorname{tr}(\bar{B}_k)]^n \sigma_k \\ &\leq [\operatorname{tr}(\bar{B}_k)]^n + \check{m}, \end{aligned} \tag{3.17}$$

where $\check{m} = a_1 \bar{m}^{1/2}$. Thus we have shown that $[\operatorname{tr}(\bar{B}_{k+1})]^n$ grows at most linearly; we shall make use of this fact later on.

We will now consider the behavior of the trace relation (3.6) when $\cos^2\theta_k$ is small. Thus let us suppose that

$$\cos^2\theta_k < \tfrac{1}{2}. \tag{3.18}$$

Then $1 + 2\cos^2\theta_k < 2$, which in turn implies that

$$\frac{1 + 2\cos^2\theta_k}{2\cos^2\theta_k} < \frac{1}{\cos^2\theta_k}$$

or

$$-\frac{1}{\cos^2\theta_k} < -1 - \frac{1}{2\cos^2\theta_k}. \tag{3.19}$$

Substituting this in (3.6) we have

$$\begin{aligned}\operatorname{tr}(\bar{B}_{k+1}) &\leq \operatorname{tr}(\bar{B}_k) + q_k\left[1 + O(\sigma_k) - 1 - \frac{1}{2\cos^2\theta_k}\right] \\ &\leq \operatorname{tr}(\bar{B}_k) - \tfrac{1}{2} q_k \left[\frac{1}{\cos^2\theta_k} - 1\right], \end{aligned} \tag{3.20}$$

since $O(\sigma_k) \leq \frac{1}{2}$ for large k.

Let $\mu_1^{(k)} \leq \cdots \leq \mu_n^{(k)}$ be the eigenvalues of $\bar{B}_k$. Then from (2.8),

$$q_k \geq \mu_1^{(k)} = \det(\bar{B}_k)/\mu_2^{(k)} \cdots \mu_n^{(k)}. \tag{3.21}$$

By the geometric/arithmetic mean inequality

$$\left[\frac{\operatorname{tr}(\bar{B}_k)}{n-1}\right]^{n-1} > \left[\frac{\mu_2^{(k)} + \cdots + \mu_n^{(k)}}{n-1}\right]^{n-1} \geq \mu_2^{(k)} \cdots \mu_n^{(k)}.$$

Using this in (3.21), and recalling that $\det(\bar{B}_k) = \det(B_1)$ we have

$$\begin{aligned} q_k &\geq (n-1)^{n-1} \det(B_1)/[\operatorname{tr}(\bar{B}_k)]^{n-1} \\ &= c_2/[\operatorname{tr}(\bar{B}_k)]^{n-1}, \end{aligned} \tag{3.22}$$

where $c_2 = (n-1)^{n-1}\det(B_1)$. Substituting this bound in (3.20) and defining $\tilde{c} = \frac{1}{2}c_2$, we have

$$\operatorname{tr}(\bar{B}_{k+1}) \leq \operatorname{tr}(\bar{B}_k) - \tilde{c}\left[\frac{1}{\cos^2\theta_k} - 1\right] \Big/ [\operatorname{tr}(\bar{B}_k)]^{n-1}. \tag{3.23}$$

Moreover by (3.18) and (3.6) we have $\mathrm{tr}(\bar{B}_{k+1}) \leq \mathrm{tr}(\bar{B}_k)$, for large k, so that by(3.23)

$$[\mathrm{tr}(\bar{B}_{k+1})]^n \leq [\mathrm{tr}(\bar{B}_{k+1})][\mathrm{tr}(\bar{B}_k)]^{n-1}$$

$$\leq [\mathrm{tr}(\bar{B}_k)]^n - \tilde{c}\left[\frac{1}{\cos^2\theta_k} - 1\right]. \tag{3.24}$$

This relation states that, when (3.18) holds, $[\mathrm{tr}(\bar{B}_k)]^n$ decreases — and the reduction is proportional to $1/\cos^2\theta_k - 1$.

We now apply (3.17) (which holds for all k) when $\cos^2\theta_k \geq \frac{1}{2}$, and apply (3.24) when $\cos^2\theta_k < \frac{1}{2}$. We obtain

$$0 < [\mathrm{tr}(\bar{B}_{k+1})]^n \leq [\mathrm{tr}(B_1)]^n - \sum_{\substack{i=1 \\ \cos^2\theta_i < 1/2}}^{k} \tilde{c}\left[\frac{1}{\cos^2\theta_i} - 1\right] + \sum_{\substack{i=1 \\ \cos^2\theta_i \geq 1/2}}^{k} \check{m},$$

and hence

$$\sum_{\substack{i=1 \\ \cos^2\theta_i < 1/2}}^{k} \left[\frac{1}{\cos^2\theta_i} - 1\right] \leq c_3 k$$

or

$$\sum_{\substack{i=1 \\ \cos^2\theta_i < 1/2}}^{k} \frac{1}{\cos^2\theta_i} \leq c_4 k, \tag{3.25}$$

for some constants c_3 and c_4. From this last relation we obtain

$$\sum_{i=1}^{k} \frac{1}{\cos^2\theta_i} \leq c_4 k + \sum_{\substack{i=1 \\ \cos^2\theta_i \geq 1/2}}^{k} \frac{1}{\cos^2\theta_i} \leq c_5 k,$$

where $c_5 = c_4 + 2$. Applying the geometric/arithmetic mean inequality we have

$$\prod_{i=1}^{k} \frac{1}{\cos^2\theta_i} \leq c_5^k. \tag{3.26}$$

We now conclude the proof as in Powell (1976). Let

$$I_k = \{i\colon 1 \leq i \leq k, \cos^2\theta_i < 1/c_5^2\}$$

$$J_k = \{i\colon 1 \leq i \leq k, \cos^2\theta_i \geq 1/c_5^2\}.$$

From (3.26)

$$c_5^k \geq \prod_{i \in I_k} \frac{1}{\cos^2\theta_i} \prod_{i \in J_k} \frac{1}{\cos^2\theta_i} \geq \prod_{i \in I_k} \frac{1}{\cos^2\theta_i} \geq c_5^{2|I_k|}.$$

This implies that $|I_k| \leq \frac{1}{2}k$, and therefore $|J_k| \geq \frac{1}{2}k$. Using the latter in (3.11) we obtain

$$f_{k+1} - f_* \leq c_6^k (f_1 - f_*),$$

where $c_6=[1-c_1/c_5^2]^{1/2}<1$. Therefore f_k converges to f_* R-linearly. It is then easy to show (see for example Byrd and Nocedal, 1989, p. 733) that

$$\sum_{k=1}^{\infty} \|\varepsilon_k\|<\infty, \tag{3.27}$$

and that x_k converges to x_* R-linearly. □

4. Can superlinear convergence be obtained?

We will now show that the search direction generated by the self-scaling BFGS method approaches the Newton direction, asymptotically. From (3.9),

$$\operatorname{tr}(\bar{B}_{k+1}) \leqslant \operatorname{tr}(B_1) \prod_{i=1}^{k} [1+\mathrm{O}(\sigma_i)].$$

Since (3.27) implies that $\sum_{i=1}^{\infty} \sigma_i<\infty$, we apply the basic result (2.20) on infinite products to obtain

$$\prod_{i=1}^{\infty} [1+\mathrm{O}(\sigma_i)]<\infty.$$

Therefore $\{\operatorname{tr}(\bar{B}_k)\}$ is bounded, and since $\det(\bar{B}_k)$ is constant, we conclude from (2.8) that

$$c_7 \leqslant q_k \leqslant c_8,$$

for some positive constants c_7 and c_8. Using this in (3.7) and recalling that $\operatorname{tr}(\bar{B}_{k+1})>0$ we have

$$\sum_{i=1}^{\infty} q_i\left[\frac{1}{\cos^2\theta_i}-1\right]<\operatorname{tr}(B_1)+c_8 \sum_{i=1}^{\infty} \mathrm{O}(\sigma_i)<\infty.$$

We conclude that

$$\cos^2\theta_k \to 1. \tag{4.1}$$

Since we have assumed that $G_*=I$, this means that the search direction approaches the Newton direction asymptotically, and superlinear convergence would be obtained if the steplength α_k is chosen appropriately. Byrd, Liu and Nocedal (1991, Lemma 3.2), show that if (4.1) and

$$\frac{s_k^{\mathrm{T}} B_k s_k}{s_k^{\mathrm{T}} y_k} \to 1 \tag{4.2}$$

hold, then setting $\alpha_k=1$ for all large k gives a Q-superlinear rate of convergence (this is just a restatement of the well-known Dennis and Moré (1974) condition). The BFGS method satisfies (4.1) and (4.2), so that asymptotically only one function evaluation per iteration is needed if the trial value $\alpha_k=1$ is always used in the line search. However the self-scaling BFGS method does not satisfy (4.2), and to obtain superlinear convergence it may require non-unit steplengths. This is illustrated by an example given below in which the desired steplengths satisfy $\alpha_{2k} \to \frac{1}{2}$ and $\alpha_{2k+1} \to 2$. The question, therefore, is whether it is possible to predict the correct steplength and avoid the function evaluations required by a line search iteration.

For general steplengths, Lemma 3.2 of Byrd, Liu and Nocedal (1990) implies that superlinear convergence is obtained only if (4.1) and

$$\frac{\alpha_k s_k^{\mathrm{T}} y_k}{s_k^{\mathrm{T}} B_k s_k} \to 1 \tag{4.3}$$

hold. Suppose that an exact line search is performed at x_k to obtain x_{k+1}, i.e. $x_{k+1} = x_k - \alpha_k^* d_k$ where α_k^* is the steplength to a one-dimensional minimizer of f along d_k. Then $g_{k+1}^{\mathrm{T}} s_k = 0$, and using (1.2), (1.3) and (1.7) we obtain

$$\frac{s_k^{\mathrm{T}} B_k s_k}{s_k^{\mathrm{T}} y_k} = \frac{s_k^{\mathrm{T}} B_k s_k}{-s_k^{\mathrm{T}} g_k} = \alpha_k^*.$$

It is clear from this relation that α_k^* satisfies (4.3), so that the self-scaling method with exact line searches is superlinearly convergent. Unfortunately exact line searches are impractical, and we need to look for other strategies.

It turns out that by making two function evaluations per search direction we can compute a value of α_k satisfying (4.3). We can use any initial guess, say $\alpha_k^{(1)} = 1$, evaluate f and g at that point and make a quadratic interpolation to obtain α_k. It is easy to see that (4.3) holds for this choice of α_k. This is the best strategy we know for obtaining superlinear convergence — and it is expensive.

We will now present a simple example that shows that non-unit steplengths may be needed by the self-scaling BFGS method to obtin superlinear convergence. This shows that the condition (4.2) is not satisfied.

Consider the two-dimensional quadratic problem

$$\min f(x) = \tfrac{1}{2}(u^2 + v^2), \tag{4.4}$$

with $x = (u, v)$, which has the unique solution

$$x_* = (0, 0)^{\mathrm{T}}. \tag{4.5}$$

We define the functions

$$\tau(\lambda, \beta) = \lambda(1 + \beta^2)^2, \tag{4.6}$$

$$\rho(\lambda, \beta) = (1 + \lambda\beta^2)^2 \tag{4.7}$$

and

$$\xi(\lambda, \beta) = \frac{(\beta^2 + 1)^2 \rho(\lambda, \beta)\tau(\lambda, \beta)}{[\rho(\lambda, \beta) + \beta^2 \tau(\lambda, \beta)]^2}. \tag{4.8}$$

In what follows we will assume that $\lambda > 1$. We first note that

$$\xi(\lambda, 1) = \frac{4\rho(\lambda, 1)\tau(\lambda, 1)}{[\rho(\lambda, 1)+\tau(\lambda, 1)]^2},$$

so that using the elementary inequality $(a+b)^2 \geqslant 4ab$ we conclude that

$$\xi(\lambda, 1) \leqslant 1. \tag{4.9}$$

We also note that

$$\lim_{\beta \to +\infty} \xi(\lambda, \beta) = \lambda. \tag{4.10}$$

For any $\lambda \geqslant 1$ and $\beta \geqslant 1$, it follows from the definitions (4.6)–(4.7) that

$$\tau(\lambda, \beta) \leqslant \rho(\lambda, \beta) \leqslant \lambda\tau(\lambda, \beta). \tag{4.11}$$

Using (4.8) and (4.11) we have

$$\begin{aligned} 1 - \frac{\xi(\lambda, \beta)}{\lambda} &= \frac{\rho(\lambda\rho - \tau) + 2\beta^2(\lambda - 1)\rho\tau + \beta^4(\lambda\tau - \rho)\tau}{\lambda(\rho + \beta^2\tau)^2} \\ &\geqslant \frac{2\beta^2(\lambda - 1)\rho\tau}{\lambda(\rho + \beta^2\tau)^2} \\ &\geqslant \frac{2\beta^2(\lambda - 1)}{\lambda^2(1+\beta^2)^2} \geqslant 0, \end{aligned} \tag{4.12}$$

where $\rho = \rho(\lambda, \beta)$ and $\tau = \tau(\lambda, \tau)$.

Let us now define the sequence

$$\lambda_k = 2 + (\tfrac{1}{2})^k, \quad k = 1, 2, \ldots. \tag{4.13}$$

Since $\xi(\lambda_k, \beta)$ is a continuous function of β, since $\{\lambda_k\}$ is a decreasing sequence, and since $\lambda_k > 1$, we see from (4.9) and (4.10) that, for each k, there exists $\beta_k > 1$ such that

$$\xi(\lambda_k, \beta_k) = \lambda_{k+1}. \tag{4.14}$$

As $\lambda_{k+1}/\lambda_k \to 1$, we have from (4.14) that $\xi(\lambda_k, \beta_k)/\lambda_k \to 1$, so that by (4.12),

$$\lim_{k\to\infty} \frac{2\beta_k^2(\lambda_k - 1)}{\lambda_k^2(1+\beta_k^2)^2} = 0.$$

This implies that

$$\lim_{k\to\infty} \beta_k = +\infty. \tag{4.15}$$

The sequences $\{\lambda_k\}$ and $\{\beta_k\}$ will help us exhibit the desired behavior of the self-scaling method. Suppose that at iteration $2k$ the current iterate is of the form

$$x_{2k} = \varepsilon_k \begin{bmatrix} 1 \\ \dfrac{1}{\lambda_k \beta_k} \end{bmatrix}, \tag{4.16}$$

and the Hessian approximation is

$$B_{2k} = \begin{bmatrix} 1 & 0 \\ 0 & \lambda_k \end{bmatrix}, \tag{4.17}$$

where we assume that

$$\varepsilon_k \beta_k = o(1). \tag{4.18}$$

From (4.16), (4.18), (4.15) and (4.5) we have that

$$\|x_{2k} - x_*\| = |\lambda_k \varepsilon_k \beta_k|(1 + o(1)). \tag{4.19}$$

The search direction is given by

$$d_{2k} = -\varepsilon_k \begin{bmatrix} 1 \\ \beta_k \end{bmatrix}. \tag{4.20}$$

Due to the special form of the objective function we have that

$$y_k = s_k, \tag{4.21}$$

for all k, so that the scaling factor (1.8) is

$$\phi_{2k} = \frac{d_{2k}^{\mathrm{T}} d_{2k}}{d_{2k}^{\mathrm{T}} B_{2k} d_{2k}} = (1 + \beta_k^2)/(1 + \lambda_k \beta_k^2). \tag{4.22}$$

We now compute the next Hessian approximation. Using (1.6), (4.21), (4.17), (4.20), and (4.22), we have

$$\begin{aligned} B_{2k+1} &= \phi_{2k}\left[B_{2k} - \frac{B_{2k} d_{2k} d_{2k}^{\mathrm{T}} B_{2k}}{d_{2k}^{\mathrm{T}} B_{2k} d_{2k}} \right] + \frac{d_{2k} d_{2k}^{\mathrm{T}}}{d_{2k}^{\mathrm{T}} d_{2k}} \\ &= \phi_{2k} \frac{\lambda_k}{1 + \lambda_k \beta_k^2} \begin{bmatrix} \beta_k^2 & -\beta_k \\ -\beta_k & 1 \end{bmatrix} + \frac{1}{1 + \beta_k^2} \begin{bmatrix} 1 & \beta_k \\ \beta_k & \beta_k^2 \end{bmatrix} \\ &= \frac{1}{\rho_k (1 + \beta_k^2)} \begin{bmatrix} \beta_k^2 \tau_k + \rho_k & \beta_k[\rho_k - \tau_k] \\ \beta_k[\rho_k - \tau_k] & \beta_k^2 \rho_k + \tau_k \end{bmatrix}, \end{aligned} \tag{4.23}$$

where $\tau_k = \tau(\lambda_k, \beta_k)$ and $\rho_k = \rho(\lambda_k, \beta_k)$. We choose the step-size to be

$$\alpha_{2k} = \frac{\lambda_k(\rho_k + \beta_k^2 \tau_k) + \tau_k - \rho_k}{\tau_k (1 + \beta_k^2)}, \tag{4.24}$$

because it gives a superlinear step for large k. Indeed, the next iterate is

$$x_{2k+1} = x_{2k} + \alpha_{2k} d_{2k} = \frac{(1 - \lambda_k)\varepsilon_k}{\tau_k (1 + \beta_k^2)} \begin{bmatrix} \rho_k + \beta_k^2 \tau_k \\ \beta_k(\rho_k - \tau_k) \end{bmatrix}, \tag{4.25}$$

and a short computation shows that

$$\|x_{2k+1}-x^*\| = (\lambda_k - 1)|\varepsilon_k|(1+o(1)) = o(\|x_{2k}-x^*\|). \tag{4.26}$$

We now compute a new iteration of the self-scaling method. Due to the special form of f we also have that $g_k = x_k$, for all k, so that by (4.25) and (4.23), the new search direction satisfies

$$\frac{1}{\rho_k(1+\beta_k^2)}\begin{bmatrix} \beta_k^2\tau_k+\rho_k & \beta_k[\rho_k-\tau_k] \\ \beta_k[\rho_k-\tau_k] & \beta_k^2\rho_k+\tau_k \end{bmatrix} d_{2k+1} = \frac{(\lambda_k-1)\varepsilon_k}{\tau_k(1+\beta_k^2)}\begin{bmatrix} \rho_k+\beta_k^2\tau_k \\ \beta_k(\rho_k-\tau_k) \end{bmatrix}.$$

By observation we conclude that d_{2k+1} is a multiple of $(1,0)^{\mathrm{T}}$; more precisely

$$d_{2k+1} = \frac{(\lambda_k-1)\rho_k\varepsilon_k}{\tau_k}\begin{bmatrix} 1 \\ 0 \end{bmatrix}. \tag{4.27}$$

Thus it follows that

$$\phi_{2k+1} = \frac{d_{2k+1}^{\mathrm{T}} d_{2k+1}}{d_{2k+1}^{\mathrm{T}} B_{2k+1} d_{2k+1}} = \rho_k(1+\beta_k^2)/(\rho_k + \beta_k^2\tau_k), \tag{4.28}$$

and consequently

$$\begin{aligned} B_{2k+2} &= \phi_{2k+1}\left[B_{2k+1} - \frac{B_{2k+1}d_{2k+1}d_{2k+1}^{\mathrm{T}}B_{2k+1}}{d_{2k+1}^{\mathrm{T}}B_{2k+1}d_{2k+1}}\right] + \frac{d_{2k+1}d_{2k+1}^{\mathrm{T}}}{d_{2k+1}^{\mathrm{T}}d_{2k+1}} \\ &= \phi_{2k+1}\frac{1}{\rho_k(1+\beta_k^2)}\begin{bmatrix} 0 & 0 \\ 0 & \dfrac{(\beta_k^2+1)^2\rho_k\tau_k}{\rho_k+\beta_k^2\tau_k} \end{bmatrix} + \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \\ &= \begin{bmatrix} 1 & 0 \\ 0 & \lambda_{k+1} \end{bmatrix}, \end{aligned} \tag{4.29}$$

where the last step follows from (4.14). Note that these formulae are independent of the stepsize α_{2k+1}. We now define

$$\alpha_{2k+1} = \frac{\lambda_{k+1}\beta_{k+1}(\rho_k+\beta_k^2\tau_k) - \beta_k(\rho_k-\tau_k)}{\lambda_{k+1}\beta_{k+1}\rho_k(1+\beta_k^2)}, \tag{4.30}$$

and a direct calculation using (4.25) and (4.27) shows that

$$x_{2k+2} = x_{2k+1} + \alpha_{2k+1}d_{2k+1} = \varepsilon_{k+1}\begin{bmatrix} 1 \\ \lambda_{k+1}\beta_{k+1} \end{bmatrix}, \tag{4.31}$$

where

$$\varepsilon_{k+1} = \frac{(1-\lambda_k)\beta_k(\rho_k-\tau_k)\varepsilon_k}{\tau_k(1+\beta_k^2)\lambda_{k+1}\beta_{k+1}}. \tag{4.32}$$

Using (4.32), (4.11), the boundedness of λ_k, (4.15) and (4.26) we obtain

$$\begin{aligned}\|x_{2k+2}-x^*\| &= \lambda_{k+1}\beta_{k+1}|\varepsilon_{k+1}|(1+o(1)) \\ &= (\lambda_k - 1)|\varepsilon_k(\rho_k-\tau_k)|\tau_k^{-1}(\beta_k+\beta_k^{-1})^{-1}(1+o(1)) \\ &= O(|\varepsilon_k/\beta_k|) \\ &= o(|\varepsilon_k|) \\ &= o(\|x_{2k+1}-x^*\|). \end{aligned} \tag{4.33}$$

Thus we obtain once more a superlinear step.

We deduce, by induction, that the sequence $\{x_k\}$ generated by the self-scaling BFGS method can satisfy the cyclical pattern (4.16)–(4.33) for all k. Now, taking limits in (4.24) and (4.30) we obtain

$$\alpha_{2k} \to 2 \quad \text{and } \alpha_{2k+1} \to \tfrac{1}{2}.$$

Our example therefore shows that the self-scaling BFGS method can converge superlinearly with $\alpha_{2k} \to 2$ and $\alpha_{2k+1} \to \frac{1}{2}$.

In practice, however, it may not be easy to guess the correct steplength, and a practical algorithm that attempts to use only one function evaluation per search direction may exhibit a linear–superlinear–superlinear type of convergence, as we now discuss. Let $\{\bar{x}_k\}$ be a sequence generated as follows. The iterate $\bar{x}_{3k}$ is given by x_{2k}, as defined in the derivation above. Remembering that a stepsize near 2 would give a superlinear step, we see that a stepsize of one would be accepted by the line search conditions (the function decreases along that direction, and a shorter step will reduce the function), so that

$$\|\bar{x}_{3k+1}\| / \|\bar{x}_{3k}\| \to \tfrac{1}{2}.$$

Then the new Hessian approximation at $\bar{x}_{3k+1}$ will be given by (4.23), which for large k is very close to Diag[0.5, 1]. Hence at the next iteration a unit step will give a superlinear step. We can assume that $\bar{x}_{3k+2}$ takes us to x_{2k+1} in the above derivation. Note that this last step would not change the Hessian approximation much, and hence we can assume that the latter is still given by (4.23). Now at the $3k+2$th iteration, as the superlinear step requires a stepsize close to 0.5, the unit trial steplength will give a point having about the same function value as the current iterate. Then the line search technique would adjust the stepsize correctly to a number close to $\frac{1}{2}$, which we could assume to be (4.30), and consequently $\bar{x}_{3k+3}$ is x_{2k+2}.

Therefore, the sequence $\{\bar{x}_k\}$ converges in a linear–superlinear–superlinear pattern. This is exhibited in the following numerical test.

5. Numerical results

We present a numerical example using problem (4.4). The initial point is $(10^{15}, 10^{20})$, the initial matrix is Diag[1, 2], and the stopping condition is

$$\|g_k\| \leq 10^{-20}. \tag{5.1}$$

The runs were performed on a Sparcstaion 1, in double precision. We ran the self-scaling BFGS method both with an exact line search as well as with an inexact line search that satisfies the Wolfe conditions (1.4)–(1.5) with $\delta_1 = 0.01$ and $\delta_2 = 0.9$. The results are given in Table 1 and Table 2 respectively.

In the exact line search case, we observe that α_k converges to $\frac{1}{2}$ and 2 alternatively. In the inexact line search case, we find that the iterates converge in a linear-superliner-superlinear pattern. We also note that in the second of the two consecutive superlinear steps, two function values were calculated. These results confirm the predictions of our analysis.

For purposes of comparison we also present in Table 3 the results of the (unscaled) BFGS method with an inexact line search. We see that the rate of convergence is superlinear and that the unit stepsize is used at every iteration.

We have also performed numerical tests with the self-scaling BFGS method using the Moré, Garbow and Hillstrom (1971) collection of test functions. We tried both well scaled and badly scaled starting matrices, and observed that, in terms of function evaluations, the standard BFGS method is superior to the self-scaling method most of the time. Moreover, we observed that the self-scaling method required almost twice as many function evaluations as the BFGS method on several test problems. In these cases, the behavior of the self-scaling method was similar to that described in Sections 4 and 5.

Table 1

Self-scaling BFGS method with exact line search

x_k	B_k		d_k	α_k
10^{15}	1.0	0.0	-10^{15}	
10^{20}	0.0	2.0	-0.5×10^{20}	2.0
-10^{15}	0.5	10^{-5}	2×10^{15}	
2×10^{10}	10^{-5}	1.0	-4×10^{10}	0.5
0.875	1.0	2.0×10^{-5}	-0.475	
3.994×10^{4}	2.0×10^{-5}	2.0	-1.997×10^{4}	2.0
-7.620×10^{-2}	0.5	1.191×10^{-5}	0.152	
1.815×10^{-6}	1.191×10^{-5}	1.0	-3.629×10^{-6}	0.5
1.804×10^{-16}	1.0	2.382×10^{-5}	-8.606×10^{-17}	
7.923×10^{-12}	2.382×10^{-5}	2.0	7.923×10^{-12}	2.0
8.287×10^{-18}	0.5	1.086×10^{-5}	-1.657×10^{-17}	
-1.800×10^{-22}	1.086×10^{-5}	1.0	3.600×10^{-22}	0.5
1.540×10^{-32}				
7.433×10^{-28}				

Table 2

Self-scaling BFGS method with inexact line search

x_k	B_k		d_k	α_k
10^{15}	1.0	0.0	-10^{15}	1.0
10^{20}	0.0	2.0	-0.5×10^{20}	
0.0	0.5	10^{-5}	10^{15}	1.0
0.5×10^{20}	10^{-5}	1.0	-0.5×10^{20}	
10^{15}	0.5	-10^{-5}	-2×10^{15}	1.0(failed)
-2×10^{10}	-10^{-5}	1.0	14488	0.5
-0.125	1.0	7.244×10^{-12}	5.256×10^{-2}	1.0
-2×10^{10}	7.244×10^{-12}	2.0	10^{10}	
-7.244×10^{-2}	0.5	2.628×10^{-12}	9.232×10^{-2}	1.0
-10^{10}	2.628×10^{-12}	1.0	10^{10}	
1.988×10^{-2}	0.5	4.616×10^{-12}	-3.976×10^{-2}	1.0(failed)
3.815×10^{-6}	4.616×10^{-12}	1.0	-3.815×10^{-6}	0.5
-1.830×10^{-10}	1.0	-9.594×10^{-5}	9.150×10^{-11}	1.0
1.907×10^{-6}	-9.594×10^{-5}	2.0	-9.537×10^{-7}	
-9.150×10^{-11}	0.5	-4.797×10^{-5}	9.150×10^{-11}	1.0
9.537×10^{-7}	-4.797×10^{-5}	1.0	-9.537×10^{-7}	
-8.677×10^{-19}	0.5	-4.797×10^{-5}	1.735×10^{-18}	1.0(failed)
0.0	-4.797×10^{-5}	1.0	8.235×10^{-23}	0.5
-1.997×10^{-27}				
4.163×10^{-23}				

Table 3

BFGS method with inexact line search

x_k	B_k		d_k	α_k
10^{15}	1.0	0.0	-10^{15}	1.0
10^{20}	0.0	2.0	-0.5×10^{20}	
0.0	1.0	-4×10^{-15}	-2×10^{6}	1.0
0.5×10^{20}	-4×10^{-15}	1.0	-0.5×10^{20}	
-2×10^{6}	1.0	-8×10^{-25}	2×10^{6}	1.0
-1.638×10^{4}	-8×10^{-25}	1.0	$1{\cdot}638\times10^{4}$	
-4×10^{-5}	1.0	-10^{-13}	4×10^{-5}	1.0
-7.276×10^{-12}	-10^{-13}	1.0	7.276×10^{-12}	
-3.523×10^{-19}	1.0	2.41×10^{-19}	3.523×10^{-19}	1.0
4.338×10^{-8}	2.41×10^{-19}	1.0	-4.338×10^{-18}	
0				
5.745×10^{-30}				

6. Final remarks

Several researchers (see for example Shanno and Phua, 1978) report disappointing numerical results with self-scaling quasi-Newton methods. Their results are consistent with our analysis. Indeed, even though our analysis is asymptotic, the main lesson we derive from it is that for the self-scaling method it is not easy to guess the correct steplength, which results in additional function evaluations. Therefore it would be desirable to design scaling strategies that do not suffer from the inefficiencies just mentioned. Some authors prefer to apply the scaling parameter at selected iterations — usually only during the first few iterations. This can be very useful for some problems. A rather different scaling technique, in which the columns of a factorization of the Hessian approximation are scaled, has been proposed by Powell (1987). This strategy has been improved by Siegel (1991) and Lalee and Nocedal (1991), who establish global and superlinear convergence results. The numerical resuts with these techniques appear to be very satisfactory.

Acknowledgements

We would like to thank Marucha Lalee and the two referees for carefully reading this article and making useful suggestions.

References

T.M. Apostol, *Mathematical Analysis* (Addison-Wesley, Reading, MA, 1957).

R.H. Byrd, D.C. Liu and J. Nocedal, "On the behavior of Broyden's class of quasi-newton methods," Report No. NAM 01, Department of Electrical Engineering and Computer Science, Northwestern University (Evanston, IL, 1990).

R.H. Byrd and J. Nocedal, "A tool for the analysis of quasi-Newton methods with application to unconstrained minimization," *SIAM Journal on Numerical Analysis* 26 (1989) 727-739.

R.H. Byrd, J. Nocedal and Y. Yuan, "Global convergence of a class of variable metric algorithms," *SIAM Journal on Numerical Analysis* 24 (1987) 1171-1190.

J.E. Dennis and J.J. Moré, "A characterization of superlinear convergence and its application to quasi-Newton methods," Mathematics of Computation 28 (1974) 549-560.

J.E. Dennis and H. Wolkowicz, "Sizing and least change secant methods," Technical Report, Department of Mathematical Sciences, Rice University (Houston, TX, 1991).

M. Lalee and J. Nocedal, "Automatic column scaling strategies for quasi-Newton methods," Report No. NAM 04, Department of Electrical Engineering and Computer Science, Northwestern University (Evanston, IL, 1991).

D.G. Luenberger, *Linear and Nonlinear Programming* (Addison-Wesley, Reading, MA, 1984, 2nd ed.).

J.J. Moré, B.S. Garbow and K.E. Hillstrom, "Testing unconstrained optimization software," *ACM Transactions on Mathematical Software* 7 (1981) 17-41.

S.S. Oren, "Perspectives on self-scaling variable metric algorithms," *Journal of Optimization Theory and Applications* 37 (1982) 137-147.

S.S. Oren and D.G. Luenberger, "Self-scaling variable metric (SSVM) algorithms, part I: Criteria and sufficient conditions for scaling a class of algorithms," *Management Science* 20 (1974) 845-862.

J.D. Pearson, "Variable metric methods for minimization," *Computer Journal* 12 (1969) 171-178.

M.J.D. Powell, "Some global convergence properties of a variable metric algorithm for minimization without exact line searches," in: R.W. Cottle and C.E. Lemke, eds., *Nonlinear Programming, SIAM-AMS Proceedings, Vol. IX* (SIAM, Philadelphia, PA, 1976) pp. 53-72.

M.J.D. Powell, "Update conjugate directions by the BFGS formula," *Mathematical Programming* 38 (1987) 29-46.

D.F. Shanno and K.H. Phua, "Matrix conditioning and nonlinear optimization," *Mathematical Programming* 14 (1978) 149-160.

D. Siegel, "Modifying the BFGS update by a new column scaling technique," Technical Report, Department of Applied Mathematics and Theoretical Physics, Cambridge University (Cambridge, UK, 1991).

P. Wolfe, "Convergence conditions for ascent methods," *SIAM Review* 11 (1969) 226-235.

P. Wolfe, "Convergence conditions for ascent methods II: some corrections," *SIAM Review* 13 (1971) 185-188.

G. Zoutendijk, "Nonlinear programming, computational methods," in: J. Abadie, ed., *Integer and Nonlinear Programming* (North-Holland, Amsterdam, 1970) pp. 37-86.