Acta Numerica (1992), –199(242

Theory of Algorithms for Unconstrained Optimization

Jorge Nocedal*

1. Introduction

A few months ago, while preparing a lecture to an audience that included engineers and numerical analysts, I asked myself the question: from the point of view of a user of nonlinear optimization routines, how interesting and practical is the body of theoretical analysis developed in this field? To make the question a bit more precise, I decided to select the best optimization methods known to date – those methods that deserve to be in a subroutine library – and for each method ask: what do we know about the behavior of this method, *as implemented in practice*? To make my task more tractable, I decided to consider only algorithms for unconstrained optimization.

I was surprised to find that remarkable progress has been made in the last 15 years in the theory of unconstrained optimization, to the point that it is reasonable to say that we have a good understanding of most of the techniques used in practice. It is reassuring to see a movement towards practicality: it is now routine to undertake the analysis under realistic assumptions, and to consider optimization algorithms as they are implemented in practice. The depth and variety of the theoretical results available to us today have made unconstrained optimization a mature field of numerical analysis.

Nevertheless there are still many unanswered questions, some of which are fundamental. Most of the analysis has focused on global convergence and rate of convergence results, and little is known about average behavior, worst case behavior and the effect of rounding errors. In addition, we do not have theoretical tools that will predict the efficiency of methods for large scale problems.

In this article I will attempt to review the most recent advances in the theory of unconstrained optimization, and will also describe some important open questions. Before doing so, I should point out that the value of the theory of optimization is not limited to its capacity for explaining the behavior of the most widely used techniques. The question

* Department of Electrical Engineering and Computer Science, Northwestern University, Evanston, IL 60208, USA

posed in the first paragraph: "what do we know about the behavior of the most popular algorithms?" is not the only important question. We should also ask how useful is the theory when designing new algorithms, i.e. how well can it differentiate between efficient and inefficient methods. Some interesting analysis will be discussed in this regard. We will see that the weaknesses of several classical algorithms that have fallen out of grace, such as the Fletcher-Reeves conjugate gradient method and the Davidon-Fletcher-Powell variable metric method, are fairly well understood. I will also describe several theoretical studies on optimization methods that have not yet enjoyed widespread popularity, but that may prove to be highly successful in the future.

I have used the terms "theoretical studies" and "convergence analysis", without stating precisely what I mean by them. In my view, convergence results fall into one of the four following categories.

1 Global convergence results. The questions in this case are: will the iterates converge from a remote starting point? Are all cluster points of the set of iterates solution points?

2 Local convergence results. Here the objective is to show that there is a neighborhood of a solution and a choice of the parameters of the method for which convergence to the solution can be guaranteed.

3 Asymptotic rate of convergence. This is the speed of the algorithm, as it converges to the solution (which is not necessarily related to its speed away from the solution).

4 Global efficiency or global rate of convergence. There are several measures; one of them estimates the function reduction at every iteration. Another approach is to study the worst case global behavior of the methods.

Most of the literature covers results in categories (1)-(3). Global efficiency results, category (4), can be very useful but are difficult to obtain. Therefore it is common to restrict these studies to convex problems (Nemirovsky and Yudin, 1983), or even to strictly convex quadratic objective functions (Powell, 1986). Global efficiency is an area that requires more attention and where important new results can be expected.

To be truly complete, the four categories of theoretical studies mentioned above should also take into account the effect of rounding errors, or noise in the function (Hamming 1971). However, we will not consider these aspects here, for this would require a much more extensive survey. The term global optimization is also used to refer to the problem of finding the global minimum of a function. We will not discuss that problem here, and reserve the term "global convergence" to denote the properties described in 1.

2. The Most Useful Algorithms for Unconstrained Optimization

Since my goal is to describe recent theoretical advances for practical methods of optimization, I will begin by listing my selection of the most useful optimization algorithms. I include references to particular codes in subroutine libraries instead of simply referring to mathematical algorithms. However the routines mentioned below are not necessarily the most efficient implementations available, and are given mainly as a reference. Most of the algorithms listed below are described in the books by (Dennis and Schnabel, 1983), (Fletcher, 1987) and (Gill, Murray and Wright, 1981).

- *The conjugate gradient method, or extensions of it.* Conjugate gradient methods are useful for solving very large problems and can be particularly effective on some types of multiprocessor machines. An efficient code implementing the Polak-Ribiére version of the conjugate gradient method, with restarts, is the routine VA14 of the Harwell subroutine library (Powell, 1977). A robust extension of the conjugate gradient method, requiring a few more vectors of storage, is implemented in the routine CONMIN (Shanno and Phua, 1980).
- *The BFGS variable metric method.* Good line search implementations of this popular variable metric method are given in the IMSL and NAG libraries. The BFGS method is fast and robust, and is currently being used to solve a myriad of optimization problems.
- *The partitioned quasi-Newton method for large scale optimization.* This method, developed by (Griewank and Toint, 1982c), is designed for partially-separable functions. These types of functions arise in numerous applications, and the partitioned quasi-Newton method takes good advantage of their structure. This method is implemented in the Harwell routine VE08, and will soon be superseded by a more general routine of the Lancelot package which is currently being developed by Conn, Gould and Toint.
- *The limited memory BFGS method for large scale optimization.* This method resembles the BFGS method but avoids the storage of matrices. It is particularly useful for large and unstructured problems. It is implemented in the Harwell routine VA15 (Liu and Nocedal, 1989).
- *Newton's method.* A good line search implementation is given in the NAG library, whereas the IMSL library provides a trust region implementation (Dennis and Schnabel, 1983), (Gay, 1983). A truncated Newton method for large problems, which requires only function and gradients, is given by (Nash, 1985).
- *The Nelder-Meade simplex method for problems with noisy functions.* An implementation of this method is given in the IMSL library.

In the following sections I will discuss recent theoretical studies on many of these methods. I will assume that the reader is familiar with the fundamental techniques of unconstrained optimization, which are described, for example in the books by (Dennis and Schnabel, 1983), (Fletcher, 1987) and (Gill, Murray and Wright, 1981). We will concentrate on line search methods because most of our knowledge on trust region methods for unconstrained optimization was obtained before 1982, and is described in the excellent survey papers by (Moré and Sorensen, 1984) and (Moré, 1983). However in section 8 we will briefly compare the convergence properties of line search and trust region methods.

3. The Basic Convergence Principles

One of the main attractions of the theory of unconstrained optimization is that a few general principles can be used to study most of the algorithms. In this section, which serves as a technical introduction to the paper, we describe some of these basic principles. The analysis that follows gives us a flavor of what theoretical studies on line search methods are, and will be frequently quoted in subsequent sections.

Our problem is to minimize a function of n variables,

$$\min f(x), \tag{3.1}$$

where f is smooth, and its gradient g is available. We consider iterations of the form

$$x_{k+1} = x_k + \alpha_k d_k, \tag{3.2}$$

where d_k is a search direction and α_k is a steplength obtained by means of a one-dimensional search. In conjugate gradient methods the search direction is of the form

$$d_k = -g_k + \beta_k d_{k-1}, \tag{3.3}$$

where the scalar β_k is chosen so that the method reduces to the linear conjugate gradient method when the function is quadratic and the line search is exact. Another broad class of methods defines the search direction by

$$d_k = -B_k^{-1} g_k \tag{3.4}$$

where B_k is a nonsingular symmetric matrix. Important special cases are given by:

$$\begin{aligned} B_k &= I \quad \text{(the steepest descent method)} \\ B_k &= \nabla^2 f(x_k) \quad \text{(Newton's method).} \end{aligned}$$

Variable metric methods are also of the form (3.4), but in this case B_k is not only a function of x_k, but depends also on B_{k-1} and x_{k-1}.

All these methods are implemented so that d_k is a descent direction, i.e. so that $d_k^T g_k < 0$, which guarantees that the function can be decreased by taking a small step along d_k. For the Newton-type methods (3.4) we can ensure that d_k is a descent direction by defining B_k to be positive definite. For conjugate gradient methods obtaining descent directions is not easy and requires a careful choice of the line search strategy. Throughout this section we will assume that the optimization method is of the form (3.2) where d_k is a descent direction.

The convergence properties of line search methods can be studied by measuring the goodness of the search direction and by considering the length of the step. The quality of the search direction can be studied by monitoring the angle between the steepest descent direction $-g_k$ and the search direction. Therefore we define

$$\cos\theta_k := -g_k^T d_k / \|g_k\| \, \|d_k\|. \tag{3.5}$$

The length of the step is determined by a line search iteration. A strategy that will play a central role in this paper consists in accepting a positive steplength α_k if it satisfies the two conditions:

$$f(x_k + \alpha_k d_k) \leq f(x_k) + \sigma_1 \alpha_k g_k^T d_k \tag{3.6}$$

$$g(x_k + \alpha_k d_k)^T d_k \geq \sigma_2 g_k^T d_k, \tag{3.7}$$

where $0 < \sigma_1 < \sigma_2 < 1$. The first inequality ensures that the function is reduced sufficiently, and the second prevents the steps from being too small. We will call these two relations the *Wolfe conditions*. It is easy to show that if d_k is a descent direction, if f is continuously differentiable and if f is bounded below along the ray $\{x_k + \alpha d_k | \, \alpha > 0\}$, then there always exist steplengths satisfying (3.6)-(3.7) (Wolfe, 1969, 1971). Algorithms that are guaranteed to find, in a finite number of iterations, a point satisfying the Wolfe conditions have been developed by Lemaréchal (1981), Fletcher (1987) and Moré and Thuente (1990).

This line search strategy allows us to establish the following useful result due to Zoutendijk. At first, the result appears to be obscure, but its power and simplicity will soon become evident. We will give a proof so that the reader can have a clear idea of how it depends on the properties of the function and line search. This result was essentially proved by Zoutendijk (1970) and Wolfe (1969 and 1971). The starting point of the algorithm is denoted by x_1.

Theorem 3.1 Suppose that f is bounded below in $\mathbf{R}^n$ and that f is continuously differentiable in a neighborhood $\mathcal{N}$ of the level set $\mathcal{L} := \{x : f(x) \leq f(x_1)\}$. Assume also that the gradient is Lipschitz continuous, i.e., there exists a constant $L > 0$ such that

$$\|g(x) - g(\tilde{x})\| \leq L\|x - \tilde{x}\|, \tag{3.8}$$

for all $x, \tilde{x} \in \mathcal{N}$. Consider any iteration of the form (3.2), where d_k is a descent direction and α_k satisfies the Wolfe conditions (3.6)-(3.7). Then

$$\sum_{k\geq 1} \cos^2\theta_k \; \|g_k\|^2 < \infty. \tag{3.9}$$

Proof. From (3.7) we have that

$$(g_{k+1} - g_k)^T d_k \geq (\sigma_2 - 1) {g_k}^T d_k.$$

On the other hand, the Lipschitz condition (3.8) gives

$$(g_{k+1} - g_k)^T d_k \leq \alpha_k L \|d_k\|^2.$$

Combining these two relations we obtain

$$\alpha_k \geq (\frac{\sigma_2 - 1}{L}) {g_k}^T d_k / \|d_k\|^2. \tag{3.10}$$

Using the first Wolfe condition (3.6) and (3.10), we have

$$f_{k+1} \leq f_k + \sigma_1 (\frac{\sigma_2 - 1}{L}) ({g_k}^T d_k)^2 / \|d_k\|^2.$$

We now use definition (3.5) to write this relation as

$$f_{k+1} \leq f_k + c \cos^2\theta_k \|g_k\|^2,$$

where $c = \sigma_1(\sigma_2 - 1)/L$. Summing this expression and recalling that f is bounded below we obtain

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

which concludes the proof. □

We shall call inequality (3.9) the *Zoutendijk condition*. Let us see how Zoutendijk's condition can be used to obtain global convergence results. Suppose that an iteration of the form (3.2) is such that

$$\cos\theta_k \geq \delta > 0, \tag{3.11}$$

for all k. Then we conclude directly from (3.9) that

$$\lim_{k\to\infty} \|g_k\| = 0. \tag{3.12}$$

In other words, if the search direction does not tend to be orthogonal to the gradient, then the sequence of gradients converges to zero. This implies, for example that the method of steepest descent, with a line search satisfying the Wolfe conditions, gives (3.12), since in this case we have $\cos\theta_k = 1$ for all k. Thus to make the steepest descent method "globally convergent" it is only necessary to perform an adequate line search.

For line search methods of the form (3.2), the limit (3.12) is the best type of global convergence result that can be obtained – we cannot guarantee that the method converges to minimizers, but only that it is attracted by stationary points.

Consider now the Newton-type method (3.2),(3.4), and assume that the condition number of the matrices B_k is uniformly bounded, i.e. that for all k

$$\|B_k\|\,\|B_k^{-1}\| \leq \Delta,$$

for some constant $\Delta > 0$. Then from (3.5) we have that

$$\cos\theta_k \geq 1/\Delta.$$

As before, we use Zoutendijk's condition (3.9) to obtain the global convergence result (3.12). We have therefore shown that Newton's method or the variable metric methods are globally convergent if the matrices B_k are positive definite (which is needed for the descent condition), if their condition number is bounded, and if the line search satisfies the Wolfe conditions. For a more thorough discussion see (Ortega and Rheinboldt, 1970).

For some algorithms, such as conjugate gradient methods, it is not possible to show the limit (3.12), but only a weaker result, namely

$$\liminf_{k\to\infty} \|g_k\| = 0. \tag{3.13}$$

We can also obtain this type of result from Zoutendijk's condition (3.9), but this time the method of proof is contradiction. Suppose that (3.13) does not hold, which means that the gradients remain bounded away from zero, i.e. there exists $\gamma > 0$ such that for all k

$$\|g_k\| \geq \gamma. \tag{3.14}$$

Then from (3.9) we conclude that

$$\cos\theta_k \to 0. \tag{3.15}$$

In other words, the algorithm can only fail, in the sense of (3.14), if the whole sequence $\{\cos\theta_k\}$ converges to 0. Therefore to establish (3.13) it suffices to show that a subsequence $\{\cos\theta_{k_j}\}$ is bounded away from zero.

For example, any line search method can be made globally convergent, in the sense of (3.13), by interleaving steepest descent steps. To be more precise, consider any method of the form (3.2) where d_k is a descent direction for all k, and where α_k is chosen to satisfy the Wolfe conditions. Suppose, in addition, that at every m steps, where m is some pre-selected integer, we define $d_k = -g_k$. Since for these steepest descent steps $\cos\theta_k = 1$, the previous discussion shows that the limit (3.13) is obtained.

It would seem that designing optimization algorithms with good convergence properties is easy, since all we need to ensure is that the search direction does not tend to become orthogonal to the gradient, or that steepest descent steps are interleaved regularly. Indeed, since the gradient g_k is always available, we can compute $\cos\theta_k$ at every iteration and apply the following angle test: if $\cos\theta_k$ is less than some pre-selected constant, then modify the search direction by turning it towards the steepest descent direction. Such angle tests have been proposed many times in the literature, and ensure global convergence, but are undesirable for the following reasons.

In addition to global convergence we would like the methods to converge rapidly. After all, if all we want to achieve is global convergence we should be satisfied with the steepest descent method. It is well-known, however, that steepest descent is very slow and that much faster algorithms can be designed. A classical result of Dennis and Moré states that the iteration (3.2) is superlinearly convergent if and only if

$$\alpha_k d_k = d_k^{\mathrm{N}} + o(\|d_k^{\mathrm{N}}\|), \tag{3.16}$$

where d_k^{N} is the Newton step (Dennis and Moré, 1974). Therefore to attain a fast rate of convergence it is necessary that we approximate the Newton direction asymptotically. An angle test may prevent us from doing so. For example, the BFGS variable metric method described in §5 can generate ill-conditioned approximations B_k of the Hessian. It is difficult, however, to determine if this is undesirable or if the matrices B_k are approximating well an ill-conditioned Hessian matrix. To decide this requires knowledge of the problem that we do not possess. We have learned that it is preferable not to interfere with the BFGS method and to let the matrices B_k evolve freely, because convergence is usually obtained and the rate is superlinear.

By far the most substantial argument against angle tests is this: the best implementations of the methods listed in §2 do not need them; it has been found that other types of safeguards are more effective. We will return to this.

(Dennis and Moré, 1977) prove a result that is of great practical value because it suggests how to estimate the initial trial value in the line search of a variable metric method. They show that for an iteration in which the the search directions approach the Newton direction, the steplength $\alpha_k = 1$ satisfies the Wolfe conditions for all large k, provided $\sigma_1 < 1/2$. Thus the unit trial steplength should always be used in variable metric methods.

Let us summarize what we have discussed so far. Zoutendijk's condition plays a central role when studying the global convergence properties of line search methods. Most of the global convergence analyses use it explicitly or follow similar approaches. The Dennis-Moré (3.16) condition is fundamental to the study of rates of convergence. It states that a method is superlinearly convergent if and only if the direction and the length of the step approximate those of Newton's method, asymptotically. Many variable metric methods are superlinearly convergent, and this is proved by simply verifying that (3.16) holds.

So far, we have only talked about one type of line search, namely the one satisfying the Wolfe conditions, and it would be misleading to suggest that this is the only useful strategy. Indeed many convergence results can also be proved for other line searches, as we will discuss in later sections. A popular strategy, called backtracking, consists of successively decreasing the steplength, starting from an initial guess, until a sufficient function reduction is obtained; see for example (Ortega and Rheinboldt, 1970). A backtracking line search is easy to implement and is well-suited for constrained problems.

Let us now discuss global efficiency analyses. One of the earliest results concerns the steepest descent method, with exact line searches, when applied to quadratic problems. This result is characteristic of global efficiency studies, which are established under very restrictive assumptions, and yet provide useful insight into the methods.

Suppose that f is the quadratic function

$$f(x) = \frac{1}{2}x^T Ax, \tag{3.17}$$

where A is symmetric and positive definite. Consider the steepest descent method with exact line searches

$$x_{k+1} = x_k - \alpha_k g_k, \tag{3.18}$$

where

$$\alpha_k = {g_k}^T g_k / {g_k}^T A g_k. \tag{3.19}$$

A simple computation (Luenberger, 1984) shows that

$$f_{k+1} = \left[1 - \frac{({g_k}^T g_k)^2}{({g_k}^T A g_k)({g_k}^T A^{-1} g_k)}\right] f_k. \tag{3.20}$$

This gives the function reduction at each iteration, and it is interesting that we have an equality. However this relation could not be used to estimate, a priori, how many iterations will be required to obtain a certain function reduction because it depends on gradient values which are unknown. Nevertheless, it is clear that the quotient in (3.20) can be bounded in terms of quantities involving only the matrix A. To do this, we use the Kantorovich inequality to obtain (Luenberger, 1984)

$$\frac{{g_k}^T g_k}{({g_k}^T A g_k)({g_k}^T A k^{-1} g_k)} \geq \frac{4\lambda_1\lambda_n}{(\lambda_1 + \lambda_n)^2},$$

where $\lambda_1 \leq \ldots \leq \lambda_n$ are the eigenvalues of A. By substituting this in (3.20) we obtain the simple relation

$$f_{k+1} \leq \left[\frac{\lambda_n - \lambda_1}{\lambda_n + \lambda_1}\right]^2 f_k. \tag{3.21}$$

This is the worst-case global behavior of the steepest descent method (3.18) - (3.19) on the quadratic problem (3.17), but it can be argued that this is also the average behavior (Akaike, 1959). Note that this global efficiency result also shows that asymptotic rate of convergence of the sequence $\{f(x)\}$ is linear, with a constant that depends on the condition number of A. Clearly, if λ_n/λ_1 is large, the term inside the square brackets in (3.21) is close to 1 and convergence will be slow.

Does this analysis help our understanding of the steepest descent method with inexact line searches on general nonlinear functions? The answer is definitely "yes". If at the solution point x_* the Hessian matrix is positive definite then, near x_*, f can be approximated well by a strictly convex quadratic, and the previous analysis is relevant - except that an inexact line search can make matters worse. However, if the line search always performs one quadratic interpolation, then the steplength will be exact asymptotically, and one can show that the rate of convergence is linear with constant $[\lambda_n - \lambda_1 \ / \ \lambda_n + \lambda_1]^2$, where $\lambda_1 \leq \ldots \leq \lambda_n$ are now the eigenvalues of the Hessian $\nabla^2 f(x_*)$.

This global efficiency result has been presented in some detail because it is illustrative of such studies in optimization methods: a simple model problem is chosen, and by direct computation, recurrence relations are established to determine the function reduction. Such relations are difficult to obtain for general nonlinear functions, but Nemirovsky and Yudin are able to derive several interesting results for convex functions. Their work is described in the book (Nemirovsky and Yudin, 1983) and in subsequent papers. We will now give a very brief description of their approach, to show its flavor.

Suppose that f is a strongly convex and continuously differentiable function. Suppose also that the gradient satisfies the Lipschitz condition (3.8) for all $x \in \mathbf{R}^n$. Let us denote a lower bound on the smallest eigenvalue of the Hessian $\nabla^2 f(x)$ by m. Nemirovsky and Yudin define the global estimate of the rate of convergence on an iterative method as a function $h(x_1 - x_*, m, L, k) :\rightarrow \mathbf{R}$ such that for any objective function f and for any $k \geq 1$ we have

$$f_k - f_* \leq c_1 h(x_1 - x_*, m, L, k),$$

where c_1 is a constant, k is the iteration number, L is the Lipschitz constant, and x_* is the solution point.

The faster the rate at which h converges to 0 as $k \rightarrow \infty$, the more efficient the method. Nemirovsky and Yudin (see also Nesterov, 1988) show that there is a lower bound on the rate of convergence of h.

Theorem 3.2 Consider an optimization method which, at every iteration k, evaluates the function f and gradient g at N_k auxiliary points whose convex hull has dimension less than or equal to l. Then for all k

$$h(x_1 - x_*, m, L, k) \geq c_2 \|x_1 - x_*\|^2 \min \left[([l+1]k)^{-2}, e^{-\sqrt{\frac{m}{L}} c_3 k (l-1)} \right], \tag{3.22}$$

where c_2 depends on m and L, and c_3 is a constant.

In this framework, a method is optimal if its efficiency mapping h is bounded above by the right hand side of (3.22), where c_2 and c_3 are allowed to be any constants. Nemirovsky and Yudin show that the well-known conjugate gradient and variable metric methods are not optimal, and (Nesterov, 1983) proposes a conjugate gradient method that achieves the optimal bound. In this theoretical framework optimization algorithms are ranked according to their worst case behavior. We will discuss this in more detail in later sections.

This concludes our outline of some basic principles used in the theoretical analysis of optimization methods. Two classical books giving an exhaustive treatment of this subject are (Ostrowsky, 1966) and (Ortega and Rheinboldt, 1970). Much of what is known about the theory of quasi-Newton methods is described in the survey paper by (Dennis and Moré, 1977) and in Dennis and Walker (1981). More recent survey papers include (Dennis and Schnabel, 1987), (Schnabel, 1989), (Toint, 1986a) and (Powell, 1985). In the following sections we focus on recent theoretical developments which are, to a great extent, not covered in these articles.

4. Conjugate Gradient Methods

The introduction of the conjugate gradient method by Fletcher-Reeves, in the 1960s, marks the beginning of the field of large scale nonlinear optimization. Here was a technique that could solve very large problems, since it requires storage of only a few vectors, and could do so much more rapidly than the steepest descent method. The definition of a large problem has changed drastically since then, but the conjugate gradient method has remained one of the most useful techniques for solving problems large enough to make matrix storage impractical. Numerous variants of the method of Fletcher and Reeves have been proposed over the last 20 years, and many theoretical studies have been devoted to them. Nevertheless, nonlinear conjugate gradient methods are perhaps the least understood methods of optimization.

The recent development of limited memory and discrete Newton methods have narrowed the class of problems for which conjugate gradient methods are recommended. Nevertheless, in my view, conjugate gradient methods are still the best choice for solving very large problems with relatively inexpensive objective functions (Liu and Nocedal, 1989). They can also be more suitable than limited memory methods on several types of multiprocessor computers (Nocedal, 1990).

The theory of conjugate gradient methods for nonlinear optimization is fascinating. Unlike the linear conjugate gradient method for the solution of systems of equations, which is known to be optimal (in some sense), some nonlinear conjugate gradient methods possess surprising, and sometimes bizarre properties. The theory developed so far offers fascinating glimpses into their behavior, but our knowledge remains fragmentary. I view the development of a comprehensive theory of conjugate gradient methods as one of the outstanding challenges in theoretical optimization, and I believe that it will come to fruition in the near future. This theory would not only be a significant mathematical accomplishment, but could result in the discovery of a superior conjugate gradient method.

The original conjugate gradient method proposed by (Fletcher and Reeves, 1964) is given by

$$d_k = -g_k + \beta_k^{\text{FR}} d_{k-1}, \tag{4.1}$$

$$x_{k+1} = x_k + \alpha_k d_k, \tag{4.2}$$

where α_k is a steplength parameter, and where

$$\beta_k^{\text{FR}} = \begin{cases} 0 & \text{for } k = 1 \\ \|g_k\|^2/\|g_{k-1}\|^2 & \text{for } k \geq 2. \end{cases} \tag{4.3}$$

When applied to strictly quadratic objective functions this method reduces to the linear conjugate gradient method provided α_k is the exact minimizer (Fletcher, 1987). Other choices of the parameter β_k in (4.1) also possess this property, and give rise to distinct algorithms for nonlinear problems. Many of these variants have been studied extensively, and the best choice of β_k is generally believed to be

$$\beta_k^{\text{PR}} = g_k^T(g_k - g_{k-1})/\|g_{k-1}\|^2, \tag{4.4}$$

and is due to Polak and Ribiére (1969).

The numerical performance of the Fletcher-Reeves method (4.3) is somewhat erratic: it is sometimes as efficient as the Polak-Ribiére method, but it is often much slower. It is safe to say that the Polak-Ribiére method is, in general, substantially more efficient than the Fletcher-Reeves method.

In many implementations of conjugate gradient methods, the iteration (4.1) is restarted every n steps by setting β_k equal to zero, i.e. taking a steepest descent step. This ensures global convergence, as was discussed in section 3. However many theoretical studies consider the iteration without restarts (Powell, 1977, 1984a), (Nemirovsky and Yudin, 1983), and there are good reasons for doing so. Since conjugate gradient methods are useful for large problems, it is relevant to consider their behavior as $n \to \infty$. When n is large (say 10,000) we expect to solve the problem in less than n iterations, so that a restart would not be performed. We can also argue that we would like to study the behavior of large sequences of unrestarted conjugate gradient iterations to discover patterns in their behavior. We will see that this approach has been very successful in explaining phenomena observed in practice. Therefore in this section we will only consider conjugate gradient methods without restarts.

The first practical global convergence result is due to Al-Baali (1985) and applies to the Fletcher-Reeves method. To establish this result it is necessary that the line search satisfy the *strong Wolfe conditions*

$$f(x_k + \alpha_k d_k) \leq f(x_k) + \sigma_1 \alpha_k g_k^T d_k \tag{4.5}$$

$$|g(x_k + \alpha_k d_k)^T d_k| \leq -\sigma_2 g_k^T d_k, \tag{4.6}$$

where $0 < \sigma_1 < \sigma_2 < \frac{1}{2}$. Note that if a stplength α_k satisfies the strong Wolfe conditions, then it satisfies the usual Wolfe conditions (3.6)-(3.7). Therefore Zoutendijk's result (3.9) will hold, provided we can show that the search directions of the Fletcher-Reeves method are descent directions. Al-Baali does this, obtaining the following global convergence result. Throughout this section we assume that the starting point is such that the level set $\mathcal{L} := \{x : f(x) \leq f(x_1)\}$ is bounded, that in some neighborhood $\mathcal{N}$ of $\mathcal{L}$, the objective function f is continuously differentiable, and that its gradient is Lipschitz continuous.

Theorem 4.1 Consider the Fletcher-Reeves method (4.1)-(4.2), where the steplength satisfies the strong Wolfe conditions (4.5)-(4.6). Then there is a constant $c > 0$ such that

$$g_k^T d_k \leq -c\,\|g_k\|^2, \tag{4.7}$$

for all $k \geq 1$, and

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

This result is interesting in many respects. The relation (4.7) is established by induction in a novel and elegant fashion. It shows that the strong Wolfe conditions are sufficient to ensure the descent property of the Fletcher-Reeves method. Prior to this result it was thought that an *ad hoc* and complicated line search would be required to guarantee descent. Relation (4.7) appears to play an important role in conjugate gradient methods, and we will encounter it again below. This theorem is also attractive because it applies to the algorithm as implemented in practice, and because the assumptions on the objective function are not restrictive.

Theorem 4.1 can be generalized to other iterations related to the Fletcher-Reeves method. Touati-Ahmed and Storey (1990) show that Theorem 4.1 holds for all methods of the form (4.1)-(4.2), which satisfy the strong Wolfe conditions, and with any β_k such that $0 \leq \beta_k \leq \beta_k^{\text{FR}}$. Gilbert and Nocedal (1990) extend this to any method with $|\beta_k| \leq \beta_k^{\text{FR}}$, and show that this result is tight in the following sense: there exists a smooth function f, a starting point x_1 and values of β_k satisfying

$$|\beta_k| \leq c\beta_k^{\text{FR}},$$

for some $c > 1$, such that the sequence of gradient norms $\{\|g_k\|\}$ generated by (4.1)-(4.2) is bounded away from zero.

This is our first encounter with a negative convergence result for conjugate gradient methods. It shows that the choice of the parameter β_k is crucial. An analysis of conjugate gradient methods with inexact line searches, shows that unless β_k is carefully chosen, the length of the search direction d_k can grow without bound causing the algorithm to fail. In the results mentioned so far, only the size of β_k with respect to β_k^{FR} plays an important role in ensuring global convergence. We will see later that a more subtle property of β_k determines the efficiency of the iteration.

Powell (1977) has given some arguments that explain, at least partially, the poor performance of the Fletcher-Reeves method in some problems: if a very small step is generated away from the solution, then due to the definition (4.3), it is likely, that subsequent steps will also be very short. We will not give the supporting facts for this argument, but only mention that the analysis is simple, and also shows that the Polak-Ribiére method would not slow down in these circumstances. This propensity for short steps, causes the Fletcher-Reeves algorithm to sometimes stall away from the solution, and this behavior can be observed in practice. For example, I have observed that when solving the minimal surface problem (Toint, 1983) with 961 variables, the Fletcher-Reeves method generates tiny steps for hundreds of iterations, and is only able to terminate this pattern after a restart is performed.

Powell (1977) and Nemirovsky and Yudin (1983) give global efficiency results that provide further evidence of the inefficiency of the Fletcher-Reeves method. The simplest analysis is that of Powell, who shows that if the Fletcher Reeves method, with exact line searches, enters a region in which the function is the two-dimensional quadratic

$$f(x) = \frac{1}{2}x^T x,$$

then the angle between the gradient g_k and the search direction d_k stays constant. Therefore, if this angle is close to 90° the method will converge very slowly. Indeed since this angle can be arbitrarily close to 90°, the Fletcher-Reeves method can be slower than the steepest descent method. Powell also shows that the Polak-Ribiére method behaves quite differently in these circumstances, for if a very small step is generated, the next search direction tends to the steepest descent direction, preventing a sequence of tiny steps from happening.

With all the arguments given in favor of the Polak-Ribiére method, we would expect to be able to prove, for it, a global convergence result similar to Theorem 4.1. That this is not possible follows from a remarkable result of Powell (1984a). He shows that the Polak-Ribiére method with exact line searches can cycle infinitely, without approaching a solution point. Since the steplength of Powell's example would probably be accepted by any practical line search algorithm, it appears unlikely that a satisfactory global convergence result will ever be found for the Polak-Ribiére method.

Powell establishes his negative result by an algebraic *tour de force*. He assumes that the line search always finds the first stationary point, and shows that there is a twice continuously differentiable function of three variables and a starting point such that the sequence of gradients generated by the Polak-Ribiére method stays bounded away from zero. Since Powell's example requires that some consecutive search directions become almost contrary, and since this can only be achieved (in the case of exact line searches) when $\beta_k < 0$, (Powell, 1986) suggests modifying the Polak-Ribiére method by setting

$$\beta_k = \max\{\beta_k^{\text{PR}}, 0\}. \tag{4.8}$$

Thus if a negative value of β_k^{PR} occurs, this strategy will restart the iteration along the steepest descent direction.

Gilbert and Nocedal (1990) show that this modification of the Polak-Ribiére method is globally convergent both for exact and inexact line searches. If negative values of β_k^{PR} occurred infinitely often, global convergence would follow, as discussed in section 3, because an infinite number of steepest descent steps would be taken. Thus Gilbert and Nocedal consider the case where $\beta_k^{\text{PR}} > 0$ for all sufficiently large k, and show that in this case $\liminf \|g_k\| = 0$, provided the line search has the following two properties: (i) it satisfies the strong Wolfe conditions, (ii) it satisfies (4.7) for some constant c. Gilbert and Nocedal discuss how to implement such a line search strategy for any conjugate gradient method with $\beta_k \geq 0$. We will now describe their analysis, which is quite different from that used by Al-Baali for the study of the Fletcher-Reeves method.

The use of inexact line searches in conjugate gradient methods requires careful consideration. In contrast with the Fletcher-Reeves method, the strong Wolfe conditions (4.5)-(4.6) no longer guarantee the descent property for the Polak-Ribiére or other conjugate gradient methods. It turns out, however, that if β_k is always non-negative it is possible to find a line search strategy that will provide the descent property. To see this note that from (4.1) we have

$$g_k{}^T d_k = -\|g_k\|^2 + \beta_k g_k{}^T d_{k-1}. \tag{4.9}$$

Therefore, to obtain descent for an inexact line search algorithm, one needs to ensure that the last term is not too large. Suppose that we perform a line search along the descent direction d_{k-1}, enforcing the Wolfe (or strong Wolfe) conditions, to obtain x_k. If $g_k{}^T d_{k-1} \leq 0$, the non-negativity of β_k implies that the sufficient descent condition (4.7) holds. On the other hand, if (4.7) is not satisfied then it must be the case that $g_k{}^T d_{k-1} > 0$, which means that a one-dimensional minimizer has been bracketed. It is then easy to apply a line search algorithm, such as that given by Lemaréchal (1981), Fletcher (1987) or Moré and Thuente (1990), to reduce $|g_k{}^T d_{k-1}|$ sufficiently and obtain (4.7). Note that the only condition imposed so far on β_k is that it be non-negative.

To obtain global convergence for other conjugate gradient methods we need to impose another condition on β_k, and interestingly enough, it is the property that makes the Polak-Ribiére method avoid the inefficiencies of the Fletcher-Reeves method. We say that a method has Property (*) if a small step, $\alpha_{k-1}d_{k-1}$ in a region away from the solution implies that β_k will be small. A precise definition is given in (Gilbert and Nocedal, 1990). It isolates an important property of the Polak-Ribiére method: the tendency to turn towards the steepest descent direction if a small step is generated away from the solution. The global convergence result of Gilbert and Nocedal is as follows.

Theorem 4.2 Consider any method of the form (4.1)-(4.2) with the following three properties: (i) $\beta_k \geq 0$ for all k; (ii) the line search satisfies the Wolfe conditions (3.6)-(3.7) and the sufficient descent condition (4.7); (iii) Property $(*)$ holds. Then $\liminf \|g_k\| = 0$.

This is one of the most general convergence results known to date. However it is not clear if the restriction $\beta_k \geq 0$ is essential, in some way, and should always be imposed in conjugate gradient methods, or if it only simplifies the analysis. It is also not known if the cycling of the Polak-Ribiére method predicted by Powell can occur in practice; to my knowledge it has never been observed. (Lukšan, 1991a) performed numerical tests with several conjugate gradient methods that restrict β_k^{PR} to be non-negative, as well as methods that are constrained by β_k^{FR}. The results are interesting, but inconclusive, and more research is needed.

How fast is the convergence of conjugate gradient methods? Let us first answer this question under the assumption that exact line searches are made. (Crowder and Wolfe, 1972) show that the rate of convergence is linear, and give an example that shows that the rate cannot be Q-superlinear. (Powell, 1976b) studies the case in which the conjugate gradient method enters a region where the objective function is quadratic, and shows that either finite termination occurs, or the rate of convergence is linear. (Cohen, 1972) and (Burmeister, 1973) show that, for general objective functions, the rate of convergence is n-step quadratic, i.e.

$$\|x_{k+n} - x_*\| = O(\|x_k - x_*\|^2),$$

and Ritter (1980) strengthens the result to

$$\|x_{k+n} - x_*\| = o(\|x_k - x_*\|^2).$$

(Powell, 1983) gives a slightly better result and performs numerical tests on small problems to measure the rate observed in practice. Faster rates of convergence can be established (Schuller, 1974), (Ritter, 1980), under the assumption that the search directions are uniformly linearly independent, but this does not often occur in practice. Several interesting results assuming asymptotically exact line searches are given by Baptist and Stoer (1977) and Stoer (1977). We will not discuss any of these rate of convergence results further because they are not recent and are described, for example, in (Powell, 1983).

(Nemirovsky and Yudin, 1983) devote some attention to the global efficiency of the

Fletcher-Reeves and Polak-Ribiére methods with exact line searches. For this purpose they define a measure of "laboriousness" and an "optimal bound" for it among a certain class of iterations. They show that on strongly convex problems, not only do the Fletcher-Reeves and Polak-Ribiére methods fail to attain the optimal bound, but they also construct examples in which both methods are slower than the steepest descent method. Subsequently (Nesterov, 1983) presents an algorithm that attains this optimal bound. It is related to PARTAN – the method of parallel tangents (Luenberger, 1984), and is unlikely to be effective in practice, but this has not been investigated, to the best of my knowledge. Some extensions of Nesterov's algorithm have been proposed by (Güler, 1989).

Let us now consider extensions of the conjugate gradient method. Motivated by the inefficiencies of the Fletcher-Reeves method, and guided by the desire to have a method that cannot converge to point where the gradient is non-zero, (Powell, 1977) proposed a conjugate gradient method which restarts automatically using a three-term recurrence iteration introduced by (Beale, 1972). This method has been implemented in the Harwell routine VE04 and outperforms the Fletcher-Reeves and Polak-Ribiére methods, but requires more storage. (Shanno and Phua, 1980) proposed a different extension of the conjugate gradient method that uses even more storage, and which resembles a variable metric iteration. It has been implemented in the highly successful and popular code CONMIN. This method, which is not simple to describe, also uses automatic restarts. The iteration is of the form

$$d_k = -H_k g_k,$$

where H_k is a positive definite and symmetric matrix. Since this ensures that the search directions are descent directions, the line search need only satisfy the usual Wolfe conditions (3.6)-(3.7). (Shanno, 1978a, 1978b) shows that this algorithm is globally convergent, with inexact line searches, on strongly convex problems. The convergence properties on non-convex problems are not known; in fact, CONMIN is related to the BFGS variable metric method, whose global convergence properties on non-convex problems are not yet understood, as we will discuss in the next section.

It is interesting to note that for all the conjugate gradient methods described in this section, and for their extensions, increased storage results in fewer function evaluations. The Fletcher-Reeves method requires 4 n-vectors of storage, Polak-Ribiére 5, VE04 6 and CONMIN 7. In terms of function evaluations, their ranking corresponds to the order in which they were just listed – with CONMIN at the top.

Are automatic restarts useful? This remains controversial. (Gill and Murray, 1979) speculate that the efficiency of VE04 and CONMIN is due to the fact that they make good use of the additional information they store, rather than to the effects of restarting. I agree with this assessment, and as we will see when we discuss limited memory methods, it is possible to design methods that are more effective than CONMIN and use no restarts. In my view, an undesirable feature of all the restarting criteria proposed so far is that they do not rule out the possibility of triggering a restart at every step, hence degrading the speed of convergence of the methods. Indeed, I have observed examples in which CONMIN restarts at every iteration and requires an excessive number of function evaluations.

I will end this section with a question that has intrigued me for some time: have we failed to discover the "right" implementation of the conjugate gradient method? Is there a simple iteration of the form (4.1)-(4.2) which performs significantly better than all the methods proposed so far, and which has all the desirable convergence properties? Given the huge number of articles proposing new variations of the conjugate gradient method, without much success, the answer would seem to be "no". However I have always felt that the answer is "yes" – but I could say no more.

5. Variable Metric Methods

We have seen that in order to obtain a superlinearly convergent method it is necessary to approximate the Newton step asymptotically – this is the principle of Dennis and Moré (3.16). How can we do this without actually evaluating the Hessian matrix at every iteration? The answer was discovered by (Davidon, 1959), and was subsequently developed and popularized by (Fletcher and Powell, 1963). It consists of starting with *any* approximation to the Hessian matrix, and at each iteration, update this matrix by incorporating the curvature of the problem measured along the step. If this update is done appropriately, one obtains some remarkably robust and efficient methods, called variable metric methods. They revolutionized nonlinear optimization by providing an alternative to Newton's method, which is too costly for many applications. There are many variable metric methods, but since 1970, the BFGS method has been generally considered to be the most effective. It is implemented in all major subroutine libraries and is currently being used to solve optimization problems arising in a wide spectrum of applications.

The theory of variable metric methods is beautiful. The more we study them, the more remarkable they seem. We now have a fairly good understanding of their properties. Much of this knowledge has been obtained recently, and we will discuss it in this section. We will see that the BFGS method has interesting self-correcting properties, which account for its robustness. We will also discuss some open questions that have resisted an answer for many years. Variable metric methods, aside from being highly effective in practice, are intricate mathematical objects, and one could spend a lifetime discovering new properties of theirs. Ironically, our many theoretical studies of variable metric methods have not resulted in the discovery of new methods, but have mainly served to explain phenomena observed in practice. However it is hard to predict the future of this area, which has given rise to many surprising developments.

The BFGS method is a line search method. At the k-th iteration, a symmetric and positive definite matrix B_k is given, and a search direction is computed by

$$d_k = -B_k^{-1} g_k. \tag{5.1}$$

The next iterate is given by

$$x_{k+1} = x_k + \alpha_k d_k, \tag{5.2}$$

where the stepsize α_k satisfies the Wolfe conditions (3.6)-(3.7). It has been found that it is best to implement BFGS with a very loose line search: typical values for parameters in (3.6)-(3.7) are $\sigma_1 = 10^{-4}$ and $\sigma_2 = 0.9$. The Hessian approximation is updated by

$$B_{k+1} = B_k - \frac{B_k s_k s_k^T B_k}{s_k^T B_k s_k} + \frac{y_k y_k^T}{y_k^T s_k}, \tag{5.3}$$

where, as before,

$$y_k = g_{k+1} - g_k, \qquad s_k = x_{k+1} - x_k. \tag{5.4}$$

Note that the two correction matrices on the right hand side of (5.3) have rank one. Therefore by the interlocking eigenvalue theorem (Wilkinson, 1965), the first rank-one correction matrix, which is subtracted, decreases the eigenvalues – we will say that it "shifts the eigenvalues to the left". On the other hand, the second rank-one matrix, which is added, shifts the eigenvalues to the right. There must be a balance between these eigenvalue shifts, for otherwise the Hessian approximation could either approach singularity or become arbitrarily large, causing a failure of the method.

A global convergence result for the BFGS method can be obtained by careful consideration of these eigenvalue shifts. This is done by Powell (1976a), who uses the trace and the determinant to measure the effect of the two rank-one corrections on B_k. He is able to show that if f is convex, then for any positive definite starting matrix B_1 and any starting point x_1, the BFGS method gives $\liminf \|g_k\| = 0$. If in addition the sequence $\{x_k\}$ converges to a solution point at which the Hessian matrix is positive definite, then the rate of convergence is superlinear.

This analysis has been extended by Byrd, Nocedal and Yuan (1987) to the restricted Broyden class of quasi-Newton methods in which (5.3) is replaced by

$$B_{k+1} = B_k - \frac{B_k s_k s_k^T B_k}{s_k^T B_k s_k} + \frac{y_k y_k^T}{y_k^T s_k} + \phi(s_k^T B_k s_k) v_k v_k^T, \tag{5.5}$$

where $\phi \in [0, 1]$, and

$$v_k = \left[\frac{y_k}{y_k^T s_k} - \frac{B_k s_k}{s_k^T B_k s_k} \right].$$

The choice $\phi = 0$ gives rise to the BFGS update, whereas $\phi = 1$ defines the DFP method – the first variable metric method proposed by Davidon, Fletcher and Powell (see e.g. (Fletcher, 1987)). Byrd, Nocedal and Yuan prove global and superlinear convergence on convex problems, for all methods in the restricted Broyden class, *except for DFP*. Their approach breaks down when $\phi = 1$, and leaves that case unresolved. Indeed the following question has remained unanswered since 1976, when Powell published his study on the BFGS method.

Open Question I.
Consider the DFP method with a line search satisfying the Wolfe conditions (3.6)-(3.7). Assume that f is strongly convex, which implies that there is a unique minimizer x_*. Do the iterates generated by the DFP method converge to x_*, for any starting point x_1 and any positive definite starting matrix B_1?

It is rather surprising that, even though the DFP method has been known for almost

30 years, we have little idea of what the answer to this basic question will turn out to be. DFP can be made to perform extremely poorly on convex problems, making a negative result plausible. On the other hand, the method has never been observed to fail; in fact even in the worst examples we can see the DFP method creeping towards a solution point. The most we can say is that the DFP method is globally convergent on convex functions if the line searches are exact (Powell, 1971, 1972), or that if it converges to a point, and line searches are exact, then the gradient at this point must be zero (Pu and Yu, 1988). It may also seem puzzling to the reader that global convergence has been established for $\phi = 0.999$, say, but not for $\phi = 1$. Wouldn't a continuity argument show that if the result holds for all $\phi < 1$ then it must also hold for $\phi = 1$? To answer this question, and to describe the self-correcting properties of the BFGS method, mentioned above, we will now discuss in some detail the convergence analyses of Powell, and Byrd, Nocedal and Yuan.

Let us begin by considering only the BFGS method, and let us assume that the function f is strongly convex, i.e. that there exist positive constants m and M such that

$$m\|z\|^2 \leq z^T G(x) z \leq M\|z\|^2 \tag{5.6}$$

for all $z, x \in \mathbf{R}^n$, where G denotes the Hessian matrix of f. Computing the trace of (5.3) we obtain

$$Tr(B_{k+1}) = Tr(B_k) - \frac{\|B_k s_k\|^2}{s_k^T B_k s_k} + \frac{\|y_k\|^2}{y_k^T s_k}. \tag{5.7}$$

It turns out that the middle term on the right hand side of this equation depends on $\cos\theta_k$, the angle between the steepest descent direction and the search direction, which was used extensively in section 3. To see this, we first note that

$$f_{k+1} - f_k = g_k^T s_k + \frac{1}{2} s_k^T G(\xi_k) s_k,$$

for some ξ_k between x_{k+1} and x_k. Thus, using the first Wolfe condition (3.6) we have

$$\sigma_1 g_k^T s_k \geq g_k^T s_k + \frac{1}{2} s_k^T G(\xi_k) s_k. \tag{5.8}$$

Next we use (5.6) and the definition (3.5) of $\cos\theta_k$ to obtain

$$(1-\sigma_1)\|g_k\|\|s_k\| \cos\theta_k \geq \frac{1}{2} m \|s_k\|^2,$$

which implies that

$$\|s_k\| \leq c_2 \|g_k\| \cos\theta_k, \tag{5.9}$$

where $c_2 = 2(1-\sigma_1)/m$. Since $B_k s_k = -\alpha_k g_k$, using (5.9) we obtain

$$\begin{aligned} \frac{\|B_k s_k\|^2}{s_k^T B_k s_k} &= \frac{\alpha_k^2 \|g_k\|^2}{\alpha_k \|s_k\|\|g_k\| \cos\theta_k} \\ &= \frac{\alpha_k \|g_k\|}{\|s_k\| \cos\theta_k} \\ &\geq \frac{\alpha_k}{c_2 \cos^2\theta_k}. \end{aligned} \tag{5.10}$$

We have thus shown that the term that tends to decrease the trace can be proportional to $\alpha_k/\cos^2\theta_k$. Let us now consider the last term in the trace equation (5.7). From the definition of y_k we have that

$$y_k = \overline{G} s_k, \tag{5.11}$$

where

$$\overline{G} = \int_0^1 G(x_k + \tau s_k) d\tau. \tag{5.12}$$

Let us define $z_k = \overline{G}^{\frac{1}{2}} s_k$, where $\overline{G}^{\frac{1}{2}}\overline{G}^{\frac{1}{2}} = \overline{G}$. Then from (5.11) and (5.6)

$$\begin{aligned} \frac{y_k^T y_k}{y_k^T s_k} &= \frac{s_k^T \overline{G}^2 s_k}{s_k^T \overline{G} s_k} \\ &= \frac{z_k^T \overline{G} z_k}{z_k^T z_k} \\ &\leq M. \end{aligned} \tag{5.13}$$

Therefore the term that tends to increase the trace is bounded above for all k, on convex problems. We obtain from (5.7) and (5.10)

$$Tr(B_{k+1}) \leq Tr(B_k) - \frac{\alpha_k}{c_2 \cos^2\theta_k} + M. \tag{5.14}$$

This relation allows insight into the behavior of the BFGS method. The discussion that follows is not rigorous, but all the statements made below can be established rigorously.

Suppose for the moment that the steplengths α_k are bounded below. If the algorithm produces iterations for which $\cos\theta_k$ is not very small, it will advance towards the solution, but some of the eigenvalues of $\{B_k\}$ could become large because the middle term on the right hand side of (5.14) could be significantly smaller than M. If, as a result of having an excessively large Hessian approximation B_k, steps with very small $\cos\theta_k$ are produced, little progress may be achieved, but a self correcting mechanism takes place: the middle term in (5.14) will be larger than M, thus decreasing the trace. This self-correction property is in fact very powerful. The smaller $\cos\theta_k$ is, the faster the reduction in the trace relation.

Suppose now that the steplengths α_k tend to zero. It is easy to see (Byrd, Nocedal and Yuan, 1987; p. 1179) that this is due to the existence of very small eigenvalues in B_k, which cannot be monitored by the means of the trace. Fortunately, it turns out that the BFGS update formula has a strong self-correcting property with respect to the determinant, which can be used to show that, in fact, α_k is bounded away from zero in mean. Indeed, the determinant of (5.3) is given by (Pearson (1969))

$$det(B_{k+1}) = det(B_k)\frac{y_k^T s_k}{s_k^T B_k s_k}. \tag{5.15}$$

Note that when $s_k^T B_k s_k$ is small relative to $y_k^T s_k = s_k^T \overline{G} s_k$, the determinant increases, reflecting the fact that the small curvature of our model is corrected, thus increasing some eigenvalues.

In conclusion, the trace relation shows that, for strongly convex problems, the eigenvalues of the matrices B_k cannot become too large, and the determinant relation shows that they cannot become too small. This can be used to show that the method is convergent, and by verifying the Dennis-Moré condition (3.16), one deduces that the rate of convergence is superlinear.

Let us now consider the restricted Broyden class (5.5) with $\phi \in [0,1)$. The analysis proceeds along similar lines. The trace relation is now (Byrd, Nocedal and Yuan, 1987)

$$Tr(B_{k+1}) \le Tr(B_k) + M + \frac{\phi \alpha_k}{c_1} - \frac{(1-\phi)\alpha_k}{c_2 \cos^2 \theta_k} + \frac{2\phi M \alpha_k}{m c_1 \cos \theta_k}, \tag{5.16}$$

where $c_1 = (1-\sigma_2)/M$. Note that the second and the third terms on the right hand side of (5.16) produce a shift to the right in the eigenvalues, in the sense that they increase the trace. The fourth term on the right hand side of (5.16) produces a shift to the left, which can be very strong when $\cos\theta_k$ is small. The last term can produce a shift in *either* direction. A crucial fact is that this last term, of uncertain sign, is inversely proportional to $\cos\theta_k$, whereas the negative fourth term is inversely proportional to $\cos^2\theta_k$. Therefore, when $\cos\theta_k$ is tiny, we still have a guaranteed decrease in the trace relation. This can be used to show that the Hessian approximation B_k cannot grow without bound.

The determinant relation, for any $\phi \in [0,1]$, can be shown to satisfy,

$$det(B_{k+1}) \ge det(B_k) \frac{y_k^T s_k}{s_k^T B_k s_k}, \tag{5.17}$$

which is essentially the same as for the BFGS update, and so we can reason as before to deduce that small eigenvalues are efficiently corrected. These arguments can be made rigorous, and can be used to establish global and superlinear convergence for any method in the restricted Broyden class using $\phi \in [0,1)$.

Why does this analysis not apply to the DFP method? It turns out that small eigenvalues do not cause problems, because (5.17) holds when $\phi = 1$, showing that the method possesses the self-correcting property with respect to the determinant mentioned above. Therefore if very small eigenvalues occur, the DFP method will be able to increase them quickly. Difficulties, however, can arise due to large eigenvalues. Note that the fourth term on the right hand side of (5.16), which plays a crucial role in preventing the trace from growing, is no longer present. The only term capable of decreasing the trace is the last term in (5.16). In addition to being of uncertain sign, this term is smaller in magnitude than the fourth term in (5.16), when $\cos\theta_k$ is small. Thus it is not certain that a shift to the left will occur, and even if it does we cannot expect it to be as strong as for other methods in the Broyden class. Therefore we can expect the DFP method to either develop excessively large Hessian approximations B_k, or at the very least, to have difficulties in reducing a large initial Hessian approximation. Numerical tests confirm these observations, which also seem to agree with a global efficiency study of Powell (1986), which we discuss later on.

We have assumed all along that the Wolfe conditions are always satisfied. Are the good properties of the BFGS method strongly dependent on them? This question is of practical importance, because for problems with inequality constraints it is often not possible to satisfy the second Wolfe condition (3.7). Fortunately it is proved by (Byrd and Nocedal, 1989) that the BFGS updating formula has excellent properties as long as it perceives positive curvature – regardless of how large the function reduction or the change in the gradient are. We now formally state one of these properties.

Theorem 5.1 Let $\{B_k\}$ be generated by the BFGS formula (5.3) where, B_1 is symmetric and positive definite, and where for all $k \geq 1$, y_k and s_k are any vectors that satisfy

$$\frac{y_k^T s_k}{s_k^T s_k} \geq m > 0 \tag{5.18}$$

$$\frac{\|y_k\|^2}{y_k^T s_k} \leq M. \tag{5.19}$$

Then for any $p \in (0,1)$ there exists a constant β_1, such that, for any $k > 1$, the relation

$$\cos\theta_j \geq \beta_1 \tag{5.20}$$

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

This result states that, even though we cannot be sure that all the $\cos\theta_k$ will be bounded below, we can be sure that this is the case for most of them. This is enough to obtain certain global convergence results. For example, Theorem 5.1 can be used to show that the BFGS method using a backtracking line search is globally convergent on convex problems. Various results of this type have also been obtained by (Werner, 1978 and 1989); see also (Warth and Werner, 1977).

The recent analysis on variable metric methods has not only produced new results, but as can be expected, has also provided simpler tools for performing the analysis. (Byrd and Nocedal, 1989) show that it is easier to work simultaneously with the trace and determinant relations. For this purpose they define, for any positive definite matrix B, the function

$$\psi(B) = \mathrm{tr}(B) - \ln(\det(B)), \tag{5.21}$$

where ln denotes the natural logarithm. It is easy to see that $\psi(B) > \ln[\mathrm{cond}(B)]$, so that global convergence can be established by analyzing the behavior of $\psi(B_k)$. Moreover the function ψ can also be used to establish superlinear convergence without having to explicitly verify the Dennis-Moré condition (3.16); this is explained in (Byrd and Nocedal, 1989).

5.1. Non-Convex Objective Functions

All the results for the BFGS method discussed so far depend on the assumption that the objective function f is convex. At present, few results are available for the case in which f is a more general nonlinear function. Even though the numerical experience of many years suggests that the BFGS method always converges to a solution point, this has not been proved.

Open Question II. Consider the BFGS method with a line search satisfying the Wolfe conditions (3.6)-(3.7). Assume that f is twice continuously differentiable and bounded below. Do the iterates satisfy $\liminf \|g_k\| = 0$, for any starting point x_1 and any positive definite starting matrix B_1?

This is one of the most fundamental questions in the theory of unconstrained optimization, for BFGS is perhaps the most commonly used method for solving nonlinear optimization problems. It is remarkable that the answer to this question has not yet been found. Nobody has been able to construct an example in which the BFGS method fails, and the most general result available to us, due to (Powell, 1976a), is as follows.

Theorem 5.2 Suppose that f is differentiable and bounded below. Consider the BFGS method with a line search satisfying the Wolfe conditions (3.6)-(3.7). Then the limit $\liminf \|g_k\| = 0$ is obtained for any starting point x_1 and any positive definite starting matrix B_1 if

$$\{\frac{y_k^T y_k}{y_k^T s_k}\} \tag{5.22}$$

is bounded above for all k.

We showed earlier (see (5.13)) that in the convex case (5.22) is always bounded, regardless of how the step s_k is chosen. However in the non-convex case, in which the Hessian matrix can be indefinite or singular, the quotient (5.22) can be arbitrarily large, and only the line search could control its size. It is not known if the Wolfe conditions ensure that (5.22) is bounded, and if not, it would be interesting to find a practical line search that guarantees this.

Now that the global behavior of variable metric methods on convex problems is reasonably well-understood, it is time that we made some progress in the case when f is a general nonlinear function. Unfortunately establishing any kind of practical results in this context appears to be extremely difficult.

The 1970s witnessed the development of a very complete *local* convergence theory for variable metric methods. The main results, due to (Broyden, Dennis and Moré, 1973) and (Dennis and Moré, 1974) have been used extensively for the analysis of both constrained and unconstrained methods, and are very well summarized in (Dennis and Moré, 1977) and (Dennis and Schnabel, 1983). A typical result is as follows. Suppose that x_* is a minimizer where the Hessian is positive definite. If x_1 is sufficiently close to x_* and B_1 is sufficiently close to $\nabla^2 f(x_*)$, then the iterates generated by the BFGS or DFP methods, with unit steplengths, converge to x_* superlinearly.

Another interesting result of Dennis and Moré makes no assumptions on the Hessian approximations, and states that if the iterates generated by BFGS or DFP satisfy

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

then the rate of convergence is superlinear. (Griewank and Toint, 1982b) extended this result to the restricted Broyden class. A stronger result for BFGS is implicit in the analysis of (Griewank, 1991) and (Byrd, Tapia and Zhang, 1990): if the iterates converge (in any way) then the convergence rate must be superlinear.

A more general local convergence theory for least change secant methods has been developed by (Dennis and Walker, 1981). This work is important because it unifies several local convergence analyses, and because it can be used to design methods for special applications. Recently, (Martínez, 1990) presented a theoretical framework that applies to some methods not covered by the theory of Dennis and Walker; see also (Martínez, 1991).

5.2. Global Efficiency of the BFGS and DFP Methods

(Nemirovsky and Yudin, 1983) note that to obtain efficiency measures of optimization methods on general objective functions appears to be an unproductive task, because only very pessimistic results can be established. Therefore they restrict their attention to convex problems, and make some interesting remarks on the properties of the DFP method. They do not resolve the question of whether DFP is optimal, in their sense, but note that DFP is not invariant under the scaling of f. They use this fact to show that, by badly scaling f, the DFP method can develop very large Hessian approximations and advance slowly. Their construction exploits the weakness of DFP with respect to large Hessian approximation mentioned above.

Powell (1986) is able to obtain much insight into the global behavior of BFGS and DFP by focusing on a narrower class of problems. He considers a strictly convex quadratic objective function of two variables, and studies the DFP and BFGS methods with steplengths of one. Since both methods are invariant under a linear change of variables, he assumes without loss of generality that $G(x_*) = I$, as this results when making the change of variables from x to $x_* + G(x_*)^{\frac{1}{2}}(x - x_*)$. Therefore Powell considers the objective function

$$f(u, v) = \frac{1}{2}(u^2 + v^2), \tag{5.23}$$

and analyzes the behavior of DFP and BFGS for different choices of the starting point x_1 and the starting matrix B_1. Due to the special form of the objective function, the secant equation $B_{k+1}s_k = y_k$, which is satisfied at each iteration by both DFP and BFGS, takes the form

$$B_{k+1}(x_{k+1} - x_k) = (x_{k+1} - x_k).$$

This shows that B_k always has one unit eigenvalue, and can assume that for all k,

$$B_k = \begin{pmatrix} 1 & 0 \\ 0 & \lambda_k \end{pmatrix}.$$

The DFP and BFGS iterations can be studied by measuring how fast λ_k converges to 1. Powell derives recurrence relations expressing λ_{k+2} in terms of λ_{k+1} and λ_k, and from them, estimates the total number of iterations required to obtain the solution to a given accuracy. These recurrence relations can also be used to estimate the function reduction at each step, and to predict how many iterations will be required before superlinear convergence takes place.

The results show vast differences of performance between the DFP and BFGS methods when the initial eigenvalue λ_1 is large. Powell shows that, in this case, the number of iterations required by the DFP method to obtain the solution with good accuracy can be of order λ_1. In contrast, the BFGS method requires only $\log_{10}\lambda_1$ iterations, in the worst case. The analysis shows that if λ_1 is large, and if the starting point is unfavorable, then the DFP method may decrease λ_k by at most one at every iteration.

When λ_1 is small, both methods are very efficient. The BFGS method requires only $\log_{10}(\log_{10})\lambda_1^{-1}$ iterations before superlinear convergence steps take place, whereas for the DFP method this occurs after only one or two iterations.

This analysis depends heavily on the assumption that unit steplengths are always taken. It is therefore relevant to ask if this is a reasonable assumption for problem (5.23). Powell shows that an algorithm using a backtracking line search, would accept the unit steplength in these circumstances. This would also be the case for other line search strategies that only demand a sufficient decrease in the function. However, a line search that requires the two Wolfe conditions may not accept the unit steplength in some iterations, if the initial eigenvalue λ_1 is large. Therefore Powell's analysis has some limitations, but the predictions of this analysis can be observed in some non-quadratic problems, as we now discuss.

Byrd, Nocedal and Yuan (1987) test methods in Broyden's class with a line search satisfying the Wolfe conditions. The objective function is strongly convex; it is the sum of a quadratic and a small quartic term. The problem has two variables and the starting matrix is chosen as a diagonal matrix with eigenvalues 1 and 10^4. The BFGS method obtained the solution to high accuracy in 15 iterations. It was able to decrease the trace of B_k from 10^4 to 3 in only 10 iterations. In contrast, the DFP method required 4041 iterations to obtain the solution (which is amazingly close to the estimate given by Powell). It took, for example, 3000 iterations for DFP to decrease the trace from 10^4 to 1100. These results agree closely with the theoretical predictions given above because the objective function is nearly quadratic – the quartic term is small.

What should we expect if we use $\phi = 0.999$ in this problem? Not surprisingly, we find that very many iterations are needed. However it is interesting that the number of iterations was 2223 – much less than for DFP. Thus a tiny change in ϕ, away from one, has a marked effect in performance.

5.3. Is BFGS the best variable metric method?

The search for a variable metric method that is more efficient than the BFGS method began in the 1970s and has not ceased. In fact a new burst of research has taken place in the last few years, and some of the new ideas may provide practical improvements in performance.

(Davidon, 1975) proposed a method in which B_{k+1} is chosen to be the member of the Broyden class that minimizes the condition number of $B_k^{-1}B_{k+1}$, subject to preserving positive definiteness. The resulting value of ϕ_k sometimes lies outside [0,1], and often coincides with the value of ϕ_k that defines the the Symmetric Rank-One method. We will discuss the Symmetric Rank-One method in section 6, and it suffices to say here that it possesses some important computational and theoretical properties. Unlike the Symmetric Rank-One method, however, Davidon's method is guaranteed to generate positive definite Hessian approximations B_k, and can be implemented without any safeguards. Nevertheless interest in the method died after numerical tests failed to show an improvement over the BFGS method, and since the theoretical study by (Schnabel, 1978) suggested that the advantages of using Davidon's approach were likely to be modest.

Recently several authors have taken a new look at the idea of deriving optimally conditioned updates, using different measures than the one proposed by Davidon. (Dennis and Wolkowicz, 1991) use the function

$$\omega(B) = \frac{\mathrm{tr}(B)}{n\det(B)},$$

to obtain a new class of updates. (Fletcher, 1991) notes that the optimal updates given by the ψ-function (5.21) are BFGS or DFP, depending on how the variational problem is posed. Other work in this area includes (Al-Baali, 1990), (Lukšan, 1991b), (Nazareth and Mifflin, 1991), (Yuan, 1991) and (Hu and Storey, 1991). A different approach, in which the secant equation is not imposed, has been investigated by (Yuan and Byrd, 1991). Even though these studies are interesting, it is too soon to know if any of these new methods can perform significantly better than the BFGS method.

The analysis of section 5.2, on the two-dimensional quadratic, suggests that the BFGS method is better at correcting small eigenvalues than large ones. Could we modify the method so as to strengthen its ability to correct large eigenvalues? Some authors feel that this can be done by using negative values for the parameter ϕ_k in Broyden's class. It is easy to explain the reason for this conjecture. Note that if $\phi_k < 0$, the fourth term in the right hand side of (5.16) remains negative and increases in magnitude, and the third term becomes negative. This suggests that, when $\phi_k < 0$, the algorithm is better able to correct large eigenvalues. Care should be taken because there is a negative value ϕ_k^c for which the update becomes singular (for values less than ϕ_k^c, the updated matrix becomes indefinite; see for example (Fletcher, 1987)). Zhang and Tewarson (1988) performed numerical tests with fixed negative values of ϕ_k, and their results show a moderate but consistent improvement over the BFGS method. They also prove that, for convex problems, global and linear convergence can be established for negative values of ϕ_k, provided that for all k,

$$(1-\nu)\phi_k^c \le \phi_k \le 0, \tag{5.24}$$

where ν is an arbitrary constant in (0,1). However (Byrd, Liu and Nocedal, 1990) show that this algorithm is not superlinearly convergent, in general. They show that designing a superlinearly convergent method which uses negative values of ϕ_k is possible, but is difficult to implement in practice.

One can also attempt to improve variable metric methods by introducing automatic scaling strategies that adjust the size of the matrix B_k. If properly done, this could alleviate, for example, the difficulties that DFP has with large eigenvalues. An idea proposed by Oren and Luenberger (1974) consists of multiplying B_k by a scaling factor ϑ_k before the update takes place. For example, for the BFGS method, the update would be of the form

$$B_{k+1} = \vartheta_k \left[B_k - \frac{B_k s_k s_k^T B_k}{s_k^T B_k s_k} \right] + \frac{y_k y_k^T}{y_k^T s_k}. \tag{5.25}$$

Several choices for ϑ_k have been proposed by Oren and Luenberger (1974), Oren (1982), and in the references cited in these papers. The choice

$$\vartheta_k = \frac{y_k^T s_k}{s_k^T B_k s_k} \tag{5.26}$$

is often recommended and has been tested in practice. The original motivation for self-scaling methods arises from the analysis of quadratic objective functions, and the main results also assume that exact line searches are performed. Disappointing numerical results were reported by several researches (see for example Shanno and Phua (1980)), and these results are explained by the analysis of (Nocedal and Yuan, 1991). They show that the method (5.25)-(5.26), using a line search that satisfies the Wolfe conditions, produces good search directions which allow superlinear convergence to take place if, in addition, the size of the step is correctly chosen. It turns out, however, that to estimate this stepsize, it is normally necessary to use an extra function evaluation, which makes the approach inefficient. Nocedal and Yuan give an example in which the stepsizes needed for superlinear convergence alternate between $\frac{1}{2}$ and 2, and note that this type of behavior can be observed in practice and is responsible for the relative inefficiency of the self-scaling method compared to the unscaled BFGS method.

For these reasons, the Oren-Luenberger scaling is now commonly applied only after the first iteration of a variable metric method. A quite different, and perhaps more promising strategy has been proposed by (Powell, 1987), and further developed by (Lalee and Nocedal, 1991) and Siegel (1991). Powell's idea is to work with the factorization

$$H_k = Z_k Z_k^T \tag{5.27}$$

of the inverse Hessian approximation H_k. This factorization has been used by (Goldfarb and Idnani, 1983) for quadratic programming and has the advantage that it can be used easily when inequality constraints are present. Powell shows that by introducing an orthogonal rotation that makes the first column of Z_k a multiple of s_k, the BFGS update of H_k can be obtained via a simple update to Z_k:

$$z_i^* = \begin{cases} s_k / \sqrt{s_k^T y_k} & i = 1 \\ z_i - (\frac{y_k^T z_i}{s_k^T y_k}) s_k & i = 2, \ldots, n, \end{cases}$$

where z_i and z_i^* are the i-th columns of Z_k and Z_{k+1} respectively. $Z_{k+1} Z_{k+1}^T$ gives H_{k+1}.

Note that the curvature information gathered during the most recent information is contained in the first column of Z_{k+1}, and that all other columns are obtained by a simple operation. Since in the BFGS update we wish to reduce the possibility of having an over-estimate of the Hessian, or equivalently an underestimate of the inverse Hessian, Powell proposes to increase all columns of Z_{k+1} so that their norms are at least equal to a parameter which depends on the norm of the first column.

(Lalee and Nocedal, 1991) extend Powell's idea to allow scaling down columns that are too large, as well as scaling up those that are too small. They give conditions on the scaling parameters in order for the algorithm to be globally and superlinearly convergent. (Siegel, 1991) proposes a slightly different scaling strategy. At every iteration, he only scales up the last l columns of the matrix Z_k, where l is a non-increasing integer. The parameter l does not change if the search direction d_k is in the span of the first $n - l$

columns of Z_k, or close to it. Otherwise, l is decreased by 1. These column scaling methods appear to work very well in practice, but there is not enough data yet to draw any firm conclusions.

6. The Symmetric Rank-One Method

One of the most interesting recent developments in unconstrained optimization has been the resurgence of the symmetric rank-one method (SR1). Several new theoretical and experimental studies have reversed the general perception of this method. Instead of being considered "fatally flawed", the SR1 method is now regarded by many researchers as a serious contender of the BFGS method for unconstrained problems, and as the most suitable quasi-Newton method for applications in which positive definite updates cannot be generated, such as constrained problems. The SR1 method remains controversial, and it is difficult to predict if the enthusiasm for this method is temporary, or if it will find a permanent place in optimization subroutine libraries.

The symmetric rank-one update is given by

$$B_{k+1} = B_k + \frac{(y_k - B_k s_k)(y_k - B_k s_k)^T}{s_k^T (y_k - B_k s_k)}. \tag{6.1}$$

It was first discovered by Davidon (1959) in his seminal paper on quasi-Newton methods, and re-discovered by several authors. The SR1 method can be derived by posing the following simple problem. Given a symmetric matrix B_k and the vectors s_k and y_k, find a new symmetric matrix B_{k+1} such that $B_{k+1} - B_k$ has rank one, and such that

$$B_{k+1} s_k = y_k.$$

It is easy to see that if $(y_k - B_k s_k)^T s_k \neq 0$, then the unique solution is (6.1), whereas if $y_k = B_k s_k$ then the solution is $B_{k+1} = B_k$. However if $(y_k - B_k s_k)^T s_k = 0$ and $y_k \neq B_k s_k$, there is no solution to the problem, and this case clouds what is otherwise a clean and simple argument. To prevent the method from failing, one can simply set $B_{k+1} = B_k$ when the denominator in (6.1) is close to zero, but this could prevent the method from converging rapidly.

It was noted early on that the SR1 method has some very interesting properties, provided it does not break down. For example Fiacco and McCormick (1968) show that the SR1 method without line searches finds the solution of a strongly convex quadratic function in at most $n + 1$ steps, if the search directions are linearly independent and if the denominator in (6.1) is always non-zero. In this case B_{n+1} equals the Hessian of the quadratic function. It is significant that this result does not require exact line searches, as is the case for the BFGS and DFP methods.

However, the fact that the denominator in (6.1) can vanish, introduces numerical instabilities and a possible breakdown of the method. Since this can happen even for quadratic functions, and since (6.1) does not always generate positive definite matrices, which complicates a line search implementation, the SR1 method fell out of favor. It was rarely used in practice, even though very good computational results had been obtained with safeguarded implementations (Dixon (1972)). The feeling in the early seventies was that the method has some intrinsic weaknesses, and that the BFGS method was clearly preferable.

The revival of the SR1 method began, interestingly enough, during the development of the partitioned quasi-Newton method of Griewank and Toint (1982c). As we will discuss in the next section, the curvature condition $s^T y > 0$ cannot always be expected to hold for all element functions, and therefore the BFGS method cannot always be applied. Therefore the implementation of the partitioned quasi-Newton method by Toint (Harwell routine VE08) uses the SR1 update when BFGS cannot be applied. This happens often; in particular after an SR1 update has been applied all subsequent updates are performed by means of SR1. The partitioned quasi-Newton method performs very well in practice, giving a first indication of the success of the SR1 method – but this work drew less attention than it deserved.

The SR1 method came to the limelight with a sequence of papers by Conn, Gould and Toint (1988a, 1988b, 1991). The first two papers deal with trust region methods for bound constrained problems, and report better results for SR1 than for BFGS. The authors speculate that the success of SR1 may be due to its superior ability to approximate the Hessian matrix at the solution. This is investigated in the third paper, in which the following result is established.

Theorem 6.1 Suppose that f is twice continuously differentiable, and that its Hessian is bounded and Lipschitz continuous. Let $\{x_k\}$ be the iterates generated by the SR1 method and suppose that $x_k \to x_*$ for some $x_* \in \mathbf{R}^n$. Suppose in addition that, for all k,

$$|s_k^T(y_k - B_k s_k)| \geq r\|s_k\| \, \|y_k - B_k s_k\|, \tag{6.2}$$

for some $r \in (0,1)$, and that the the steps s_k are uniformly linearly independent. Then

$$\lim_{k\to\infty} \|B_k - \nabla^2 f(x_*)\| = 0.$$

Condition (6.2) is often used in practice to ensure that the SR1 update is well behaved: if it is violated then the update is skipped. Conn, Gould and Toint (1991) report that the assumption of uniform linear independence of the search directions holds in most of their runs, and that the Hessian approximations generated by the SR1 method are often more accurate than those generated by BFGS or DFP.

Osborne and Sun (1988) propose a modification in which the Hessian approximation is scaled before the SR1 update is applied. They analyze this method and report good numerical results. In an interesting recent paper, Khalfan, Byrd and Schnabel (1991) make further contributions to the theory of the SR1 method, and present numerical results that, to some extent, conflict with those of Conn, Gould and Toint (1991). They consider both a line search and a trust region implementation and observe that, for the problems they tested, the Hessian approximations generated by the SR1 method are on the average only slightly more accurate than those produced by the BFGS method. They report that in about one third of their problems neither method produces close approximations to the Hessians at the solution.

These results suggest that the assumptions of Theorem 6.1 may not always be satisfied in practice . Therefore Khalfan, Byrd and Schnabel study whether the steps generated by the SR1 method are uniformly linearly independent and find that this is often not the case. They conclude that the efficiency of the SR1 method is unlikely to be due to the properties given in Theorem 6.1, and pursue an analysis that is not based on the linear independence assumption. They prove several results which we now describe.

The first result is related to the Dennis-Moré condition for superlinear convergence, and assumes that unit steplengths are taken. It states that if x_* is a minimizer such that $\nabla^2 f(x_*)$ is positive definite, and if

$$e_k \equiv \|x_k - x^*\|$$

and

$$\frac{\|(B_k - \nabla^2 f(x_*))s_k\|}{\|s_k\|},$$

are sufficiently small, then

$$\|x_k + s_k - x_*\| \leq c_1 \left[\frac{\|(B_k - \nabla^2 f(x_*))s_k\|}{\|s_k\|} e_k + c_2 e_k^2 \right]$$

where c_1 and c_2 are constants.

This bound suggests that some kind of quadratic rate is possible. To establish this, however, Khalfan, Byrd and Schnabel must assume that the matrices $\{B_k\}$ are positive definite and bounded. This appears, at first, to be a very unrealistic assumption, but the authors note that this is very often the case in their numerical tests. We now formally state this second result on the SR1 method.

Theorem 6.2 Suppose that the iterates generated by the SR1 method converge to x_* – a minimizer such that $\nabla^2 f(x_*)$ is positive definite. Assume that for all $k \geq 0$ the condition (6.2) is satisfied and that the matrices B_k are positive definite and uniformly bounded above in norm. Then the rate of convergence is $2n$-step q-quadratic, i.e.

$$\limsup_{k \to \infty} \frac{e_{k+2n}}{e_k^2} \leq \infty.$$

These new results are, of course, not as strong as the global convergence results described for the BFGS method, but one should keep in mind that the renewed interest in the SR1 method is very recent. Therefore substantial advances in this area can be expected.

7. Methods for Large Problems

Every function f with a sparse Hessian is partially separable, i.e. it can be written in the form

$$f(x) = \sum_{i=1}^{ne} f_i(x), \tag{7.1}$$

where each of the ne element functions f_i depends only on a few variables. This statement is proved by (Griewank and Toint, 1981a), and provides the foundation for their partitioned quasi-Newton method for large-scale optimization. The idea behind this method is to exploit the partially separable structure (7.1) and update an approximation B_k^i to the Hessian of each element function f_i. These matrices, which are often very accurate, can be assembled to define an approximation B_k to the Hessian of f. There is one complication: even if $\nabla^2 f(x_*)$ is positive definite, some of the element functions may be concave, so that the BFGS method cannot always be used. In this case Griewank and Toint use the SR1 update formula, and implement safeguards that skip the update if it is suspect.

The search direction of the partitioned quasi-Newton method, as implemented by (Toint, 1983), is determined by solving the system

$$\left(\sum_{i=1}^{ne} B_k^i \right) d_k = -g_k \tag{7.2}$$

inside a trust region, using a truncated conjugate gradient iteration. If a direction of negative curvature is detected, the conjugate gradient iteration is terminated, and d_k is set to this direction of negative curvature. After this, a line search is performed along d_k. This method is described and analyzed by (Griewank and Toint, 1982b, 1982c, 1984); the implementation just outlined corresponds to the Harwell routine VE08.

The partitioned quasi-Newton method performs very well in practice, and represents one of the major algorithmic advances in nonlinear optimization. We should note that many practical problems are directly formulated in the form (7.1), and that many other problems can be recast in that form. Thus the partitioned quasi-Newton method is of wide applicability.

To establish global convergence results, similar to those for the BFGS method on convex problems, it is necessary to assume that all the element functions f_i are convex. Under this assumption (Griewank, 1991) shows that the partitioned quasi-Newton method is globally convergent, even if the system (7.2) is solved inexactly. Griewank also relaxes the smoothness conditions on the gradients of the element functions f_i, and establishes rate of convergence results under the assumption that these gradients are only Lipschitzian, rather than differentiable. Griewank's analysis completely describes the behavior of the partitioned quasi-Newton method in the convex case, and strengthens earlier work by (Toint, 1986b).

A very different approach for solving large problems ignores the structure of the problem, and uses the information of the last few iterations to define a variable metric approximation of the Hessian. This, so-called limited memory BFGS method, has proved to be very useful for solving certain large unstructured problems, and is in fact competitive with the partitioned quasi-Newton method on partially separable problems in which the number of variables entering into the element functions f_i exceeds 5 or 6 (Liu and Nocedal, 1989).

The limited memory BFGS method is very similar to the standard BFGS method – the only difference is in the matrix update. Instead of storing the matrices H_k that approximate the inverse Hessian, one stores a certain number, say m, of pairs $\{s_i, y_i\}$ that define them implicitly. The product $H_k g_k$, which defines the search direction, is obtained by performing a sequence of inner products involving g_k and the m most recent vector pairs $\{s_i, y_i\}$. This is done efficiently by means of a recursive formula (Nocedal, 1980). After computing the new iterate, we delete the oldest pair from the set $\{s_i, y_i\}$ and replace it by the newest one. Thus the algorithm always keeps the m most recent pairs $\{s_i, y_i\}$ to define the iteration matrix. It has been observed that scaling can be highly beneficial for large problems and several strategies for doing this have been studied by (Gilbert and Lemaréchal, 1989).

The limited memory BFGS method is suitable for large scale problems because it has been observed in practice that small values of m (say $m \in [3, 7]$) give satisfactory results. It is not understood why this method is as fast as the standard BFGS method on many problems. Another interesting open question is how to design a strategy for selecting the most useful corrections pairs – not simply the most recent ones – to improve the performance of the method.

Since the Dennis-Moré condition (3.16) cannot possibly hold for the limited memory BFGS method, its rate of convergence must be linear. (Liu and Nocedal, 1989) prove that the limited memory BFGS method is globally and linearly convergent on convex problems for any starting point, and for several useful scaling strategies. It is interesting to note that, as implemented by Liu and Nocedal, the method does not possess quadratic termination. A different limited memory method, that combines cycles of BFGS and conjugate gradient directions has been developed by (Buckley and LeNir, 1983).

Newton's method is, of course, the best method for solving many types of problems. Both line search and trust region implementations have been developed for the large-scale case; see (Steihaug, 1983), (Nash, 1985), (O'Leary, 1982) and (Toint, 1986a). The convergence properties of implementations of Newton's method in which the linear system

$$\nabla^2 f(x_k) d_k = -g_k \tag{7.3}$$

is solved inaccurately were first considered by (Dembo, Eisenstat and Steihaug, 1982) and by (Bank and Rose, 1981). Several interesting recent papers generalizing this work, and focusing on specific methods for solving the linear system (7.3), include (Brown and Saad, 1989 and 1990), (El Hallabi and Tapia, 1989), (Martínez, 1990) and (Eisenstat and Walker, 1991). Non-monote Newton methods, i.e. methods in which function values are allowed to increase at some iterations, have been analyzed by (Grippo, Lampariello and Lucidi, 1990a, 1990b); the numerical results appear to be very satisfactory. Non-monotone methods may prove to be very useful for solving highly nonlinear problems.

8. Remarks on Other Methods

I have concentrated on recent theoretical studies on methods for solving general unconstrained minimization problems. Due to space limitations I have not discussed the solution of systems of nonlinear equations or nonlinear least squares. The Nelder-Meade method is known to fail, so that establishing a global convergence result for it is not possible. Recently there has been research on modifications of the Nelder-Meade method to improve its performance, and it is possible to establish global convergence for some of them. For a description of this work see (Torczon, 1991).

As mentioned earlier, I have not reviewed trust region methods because most of their theoretical studies (for unconstrained problems) are not recent and are reviewed by (Moré and Sorensen, 1984). Nevertheless, I would like to briefly contrast their properties with those of line search methods.

Trust region methods do not require the Hessian approximations B_k to be positive definite. In fact, very little is required to establish global convergence: it is only necessary to assume that the norm of the matrices $\|B_k\|$ does not increase at a rate that is faster than linear (Powell, 1984b). In contrast, for line search methods one needs to ensure that the condition number of the Hessian approximations $\|B_k\|$ does not grow too rapidly. This requires control on both the largest and smallest eigenvalues of B_k, making the analysis more complex than for trust region methods. It is also possible to show that for trust region methods the sequence of iterates always has an accumulation point at which the gradient is zero and the Hessian is positive semi-definite. This is better than the result $\liminf \|g_k\| = 0$ which is the most that can be proved for line search methods.

Thus the theory of trust region methods has several advantages over that of line search methods, but both approaches seem to perform equally well in practice. Line search methods are more commonly used because they have been known for many years and because they can be simpler to implement. At present, line search and trust region methods coexist, and it is difficult to predict if one of these two approaches will become dominant. This will depend on the theoretical and algorithmic advances that the future has in store.

9. Acknowledgements

I am very grateful to Marucha Lalee, Christy Hartung and Peihuang Lu for their help in the preparation of this article. I like to acknowledge support from the Department of Energy, grant DE-FG02-87ER25047-A001, and the National Science Foundation, grant CCR-9101359.

10. *

REFERENCES

H. Akaike (1959), "On a successive transformation of probability distribution and its application to the analysis of the optimum gradient method, *Ann. Inst. Statist. Math.* **11**, 1-17.

M. Al-Baali (1985), "Descent property and global convergence of the Fletcher-Reeves method with inexact line search", *IMA Journal of Numerical Analysis* **5**, 121–124.

M. Al-Baali (1990), "Variational Quasi-Newton Methods for Unconstrained Optimization", Technical Report, Department of Mathematics, University of Damascus (Syria).

R.E. Bank and D.J. Rose, "Global approximate Newton methods, *Numerische Mathematik* **37**, 279–295.

P. Baptist and J. Stoer (1977), "On the relation between quadratic termination and convergence properties of minimization algorithms, Part II, Applications", *Numerische Mathematik* **28**, 367–392.

E.M.L. Beale (1972), "A derivation of conjugate gradients", in *Numereical Methods for Nonlinear Optimization* (F.A. Lootsma, ed.), Academic Press.

P.N. Brown and Y. Saad (1989), "Globally convergent techniques in nonlinear Newton-Krylov algorithms", Technical Report UCRL-102434, Lawrence Livermore National Laboratory.

P.N. Brown and Y. Saad (1990), "Hybrid Krylov methods for nonlinear systems of equations", *SIAM J. Sci. Stat. Comput.* **11**, 450–481.

G.G. Broyden, J.E. Dennis and J.J. Moré (1973), "On the local and superlinear convergence of quasi-Newton methods", *J. Inst. Math. Appl.* **12**, 223-246.

A. Buckley and A. LeNir (1983), "QN-like variable storage conjugate gradients",*Mathematical Programming* **27**, 155–175.

W. Burmeister (1973), "Die Konvergenzordnung des Fletcher-Powell Algorithmus", *Z. Angew. Math. Mech.* **53**, 693–699.

R. Byrd, D. Liu and J. Nocedal (1990), "On the Behavior of Broyden's Class of Quasi-Newton Methods", Tech. Report NAM 01, Northwestern University, EECS Dept.

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

R.H. Byrd, J. Nocedal and Y. Yuan (1987), "Global convergence of a class of quasi-Newton methods on convex problems", *SIAM J. Numer. Anal.* **24**, 1171–1190.

R.H. Byrd, R.A. Tapia and Y. Zhang (1990), "An SQP Augmented Lagrangian Bfgs Algorithm for Constrained Optimization", Tech. Report, University of Colorado (Boulder).

A. Cohen (1972), "Rate of convergence of several conjugate gradient algorithms", *SIAM J. Numer. Anal.* **9**, 248–259.

A.R. Conn, N.I.M. Gould, and PH.L. Toint (1988a), "Global convergence of a class of trust region algorithms for optmization with simple bounds", *SIAM J. Numer. Anal.* **25**, 433–460.

A. R. Conn, N.I.M. Gould, and PH. L. Toint (1988b), "Testing a class of methods for solving minimization problems with simple bounds on the variables", *Mathematics of Computation* **50**, 399–430.

A. R. Conn, N.I.M. Gould, and PH. L. Toint (1991), "Convergence of quasi-Newton matrices generated by the symmetric rank one update", *Math. Prog.* **2**, 177–195.

H.P. Crowder and P. Wolfe (1972), "Linear convergence of the conjugate gradient method", *IBM Journal of Research and Development* **16**, 431–433.

W.C. Davidon (1959), "Variable metric methods for minimization", Argonne National Lab Report (Argonne, IL).

W.C. Davidon (1975), "Optimally conditioned optimization algorithms without line searches", *Math. Prog.* **9**, 1-30.

R.S. Dembo, S.C. Eisenstat, and T. Steihaug (1982), "Inexact Newton methods", *SIAM J. Numer. Anal.* **19**, 400–408.

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

J.E. Dennis, Jr. and J.J. Moré (1977), "Quasi-Newton methods, motivation and theory", *SIAM Rev.* **19**, 46–89.

J.E. Dennis, Jr. and R.B. Schnabel (1983), *Numerical Methods for Unconstrained Optimization and Nonlinear Equations* , Prentice-Hall, Inc., (Englewood Cliffs, NJ).

J.E. Dennis, Jr. and R.B. Schnabel (1987), "A view of unconstrained optimization", Technical Report CU-CS-376-87, University of Colorado (Boulder), to appear in: *Handbooks in Operations Research and Management Science* **1**, Optimization (G.L. Nemhauser, A.H.G. Rinnooy Kan, and M.J. Todd, eds.), North-Holland (Amsterdam).

J.E. Dennis, Jr. and H.F. Walker (1981), "Convergence theorems for least-change secant update methods," *SIAM J. Numer. Anal.* **18**, 949–987, **19**, 443.

J.E. Dennis, Jr. and H. Wolkowicz (1991), "Sizing and least change secant methods", to appear.

L.C.W. Dixon (1972), "The choice of step length, a crucial factor in the performance of variable metric algorithms", in *Numerical Methods for Nonlinear Optimization* (F.A. Lootsma, ed.), Academic Press (London).

S. Eisenstat and H. Walker (1990), "Globally convergent inexact Newton methods", Yale Technical Report.

M. El Hallabi and R.A. Tapia (1989), "A global convergence theory for arbitrary norm trust-region methods for nonlinear equations", Tech. Rep. TR87-25, Rice University, Department of Mathematical Sciences.

A.V. Fiacco and G.P. McCormick (1968), *Nonlinear Programmming* , John Wiley & Sons (New York).

R. Fletcher (1987), *Practical Methods of Optimization* **1**, Unconstrained Optimization, John Wiley & Sons (New York).

R. Fletcher and M.J.D. Powell (1963), "A rapidly convergent descent method for minimization", *Comput. J.* **6**, 163–168.

D. Gay (1983), "Subroutines for unconstrained minimization using a model/ trust-region approach", *ACM Trans. Math. Soft.* **9,4**, 503–524.

J.C. Gilbert and C. Lemaréchal (1989), "Some numerical experiments with variable storage quasi-Newton algorithms", *Mathematical programming* **45**, 407–436.

J. C. Gilbert and J. Nocedal (1990), "Global convergence properties of conjugate gradient methods for optimization", Rapport de Recherche, INRIA (Paris).

P.E. Gill and W. Murray (1979), "Conjugate-gradient methods for large-scale nonlinear optimization", Technical report SOL 79-15, Dept. of Operations Research, Stanford University.

P. E. Gill, W. Murray and M. H. Wright (1981), *Practical Optimization*, Academic Press (London).

D. Goldfarb, and A. Idnani (1983), "A numerically stable dual method for solving strictly convex quadric programs", *Math. Prog.* **27**, 1–33.

A. Griewank (1991), "The global convergence of partitioned BFGS on problems with convex decompositions and Lipschitzian gradients", *Math. Prog.* **50**, 141–175.

A. Griewank and Ph.L. Toint (1982a), "On the unconstrained optimization of partially separable objective functions", in *Nonlinear Optimization 1981* (M.J.D. Powell, ed.), Academic Press (London), 301–312.

A. Griewank and Ph.L. Toint (1982b), "Local convergence analysis of partitioned quasi-Newton updates", *Numerische Mathematik* **39**, 429–448.

A. Griewank and Ph.L. Toint (1982c), "Partitioned variable metric updates for large structured optimization problems", *Numer. Math.* **39**, 119–137.

A. Griewank and Ph.L. Toint (1984), "Numerical experiments with partially separable optimization problems", in *Numerical Analysis: Proceedings Dundee 1983* , Lecture Notes in Mathematics 1066 (D.F. Griffiths, ed.), Springer Verlag, (Berlin), 203–220.

L. Grippo, F. Lampariello and S. Lucidi (1990a), " A quasi-discrete Newton algorithm with a nonmonote stabilization technique", *J. Optim. Theory Appl.* **64**, 485–500.

L. Grippo, F. Lampariello and S. Lucidi (1990b), " A class of nonmonotone stabilization methods in unconstrained optimization", Technical report R-290, Consiglio Nazionale delle Ricerche.

O. Güler (1989), "Optimal algorithms for smooth convex programming", Working Paper Series No. 89-17, Department of Management Sciences, The University of Iowa.

R.W. Hamming (1971), *Introduction to Applied Numerical Analysis* , McGraw-Hill.

Y.F. Hu and C. Storey (1991), "On optimally and near-optimally conditioned quasi-Newton updates", Technical Report A141, Department of Mathematical Sciences, Loughborough University of Technology (Leicestershire).

H. Khalfan, R.H. Byrd, and R.B. Schnabel (1990), "A theoretical and experimental study of the symmetric rank one update", Technical Report CU-CS-489-90, University of Colorado (Boulder).

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

C. Lemaréchal (1981), "A view of line searches", in *Optimization and Optimal Control* , (Auslander, Oettli and Stoer, eds.), Lecture Notes in Control and Information Science 30, Springer Verlag, 59–78.

D.C. Liu and J. Nocedal (1989), "On the limited memory BFGS method for large scale optimization", *Mathematical Programming* **45**, 503–528.

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

L. Lukšan (1991a), "Computational experience with improved conjugate gradient methods for unconstrained minimization", Technical Report No. 488, Institute of Computer & Information Sciences, Czechoslovak Academy of Sciences (Prague).

L. Lukšan (1991b), "On variationally derived scaling and preconvex variable metric updates", Technical Report No. 496, Institute of Computer & Information Sciences, Czechoslovak Academy of Sciences (Prague).

J.M. Martínez (1990), "Local convergence theory of inexact Newton methods based on structured least change updates", *Math. Comp.* **55**, 143–168.

J.M. Martínez (1991), "On the Relation between two local convergence theories of least change update methods", Tech. Rept. IMECC-UNICAMP.

J. Moré (1983), "Recent developments in algorithms and software for trust region methods", in *Mathematical Programming, The State of the Art* (A. Bachem, M. Grotschel, G. Korte, eds), Springer-Verlag, 256–287.

J. Moré and D.C. Sorensen (1984), "Newton's method", in *Studies in Numerical Analysis* (G.H. Golub, ed), The Mathematical Association of America, 29–82.

J. Moré, and D.J. Thuente (1990) "On line search algorithms with guaranteed sufficient decrease", Mathematics and Computer Science Division Preprint MCS-P153-0590, Argonne National Laboratory (Argonne, IL).

S.G. Nash (1985), "Preconditioning of truncated-Newton methods", *SIAM Journal on Scientific and Statistical Computing* **6**, 599–616.

J.L. Nazareth and R.B. Mifflin (1991), "The least prior deviation quasi-Newton update", Technical Report, Department of Pure and Applied Mathematics, Washington State University.

A.S. Nemirovsky and D. B. Yudin (1983), *Problem Complexity and Method Efficiency* , Wiley.

Y.E. Nesterov (1983), "A method of solving a convex programming problem with convergence rate $0(1/k^2)$", *Soviet Mathematics Doklady* **27**, 372-376.

Y.E. Nesterov (1988), "On an approach to the construction of optimal methods of minimization of smooth convex functions", *Ekonomika i Matem.* **24**, 509–517.

J. Nocedal (1980), "Updating quasi-Newton matrices with limited storage", *Mathematics of Computation* **35**, 773–782.

J. Nocedal (1990), "The performance of several algorithms for large scal unconstrained optimization", in *Large-Scale Numerical Optimization* (T.F. Coleman and Y. Li, eds), SIAM (Philadelphia), 138–151.

J. Nocedal and Ya-xiang Yuan (1991), "Analysis of a self-scaling quasi-Newton method", Technical Report NAM-02, Department of Electrical Engineering and Computer Science, Northwestern University (Evanston, IL).

D.P. O'Leary (1982), "A discrete Newton algorithm for minimizing a function of many variables", *Math. Prog.* **23**, 20–33.

S.S. Oren (1982), "Perspectives on self-scaling variable metric algorithms", *J. Opt. Theory and Appl.* **37**, 137–147.

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

J.M. Ortega and W.C.Rheinboldt (1970), *Iterative Solution of Nonlinear Equations in Several Variables*, Academic Press.

M.R. Osborne and L.P. Sun (1988), "A new approach to the symmetric rank-one updating algorithm", Rept. NMO/01, Department of Statistics, IAS, Australian National University.

A. Ostrowski (1966), *Solution of Equations and Systems of Equations* , second edition, Academic Press (New York).

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

E. Polak and G. Ribière (1969), "Note sur la convergence de methodes de directions conjugées", *Rev. Française Informat Recherche Operationelle, 3e Année* **16**, 35–43.

M.J.D. Powell (1971), "On the convergence of the variable metric algorithm", *J. Inst. Math. Appl.* **7**, 21–36.

M.J.D. Powell (1972), "Some properties of the variable metric method", in *Numerical Methods for Non-linear Optimization* (F.A. Lootsma, ed.), Academic Press (London).

M.J.D. Powell (1976a), "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.), SIAM Publications (Philadelphia).

M.J.D. Powell (1976b), "Some convergence properties of the conjugate gradient method", *Mathematical Programming* **11**, 42–49.

M.J.D. Powell (1977), "Restart procedures of the conjugate gradient method", *Mathematical Programming* **2**, 241–254J.

M.J.D. Powell (1983), "On the rate of convergence of variable metric algorithms for unconstrained optimization", Report DAMTP 1983/NA7, Department of Applied Mathematics and Theoretical Physics, University of Cambridge (Cambridge).

M.J.D. Powell (1984a), "Nonconvex minimization calculations and the conjugate gradient method", in *Lecture Notes in Mathematics* **1066**, Springer-Berlag (Berlin), 122–141.

M.J.D. Powell (1984b), " On the global convergence of trust region algorithms for unconstrained minimization", *Mathematical Programming* **29**, 297-303.

M.J.D. Powell (1985), "Convergence properties of algorithms for nonlinear optimization", Report DAMTP 1985/NA1, Department of Applied Mathematics and Theoretical Physics, University of Cambridge (Cambridge).

M.J.D. Powell (1986), "How bad are the BFGS and DFP methods when the objective function is quadratic?", *Math. Prog.* **34**, 34–47.

M.J.D. Powell (1987), "Update conjugate directions by the BFGS formula", *Math. Prog.* **38**, 29–46.

D.G. Pu and W.C. Yu (1988), "On the convergence property of the DFP algorithm", *Journal of Qüfu Normol University 14/3* , 63–69.

K. Ritter (1980), "On the rate of superlinear convergence of a class of variable metric methods", *Numer. Math.* **35**, 293–313.

R.B. Schnabel (1978), "Optimal conditioning in the convex class of rank two updates", *Math. Prog.* **15**, 247–260.

R.B. Schnabel (1989), "Sequential and parallel methods for unconstrained optimization", in *Mathematical Programming, Recent Developments and Applications* (M. Iri and K. Tanabe, eds.), Kluwer Academic Publishers, 227–261.

G. Schuller (1974), "On the order of convergence of certain quasi-Newton methods", *Numer. Math.* **23**, 181-192.

D.F. Shanno (1978a), "On the convergence of a new conjugate gradient algorithm", *SIAM Journal on Numerical Analysis* **15**, 1247–1257.

D.F. Shanno (1978b), "Conjugate gradient methods with inexact searches", *Mathematics of Operations Research* **3**, 244–256.

D.F. Shanno and K.H. Phua (1980), "Remark on algorithm 500: minimization of unconstrained multivariate functions", *ACM Transactions on Mathematical Software* **6**, 618–622.

D. Siegel (1991), "Modifying the BFGS update by a new column scaling technique", Technical Report DAMTP 1991/NA5, Department of Applied Mathematics and Theoretical Physics, University of Cambridge.

T. Steihaug (1983), "The conjugate gradient method and trust regions in large scale optimization", *SIAM J. Num. Anal.* **20**, 626–637.

J. Stoer (1977), "On the relation between quadratic termination and convergence properties of minimization algorithms", *Numer. Math.* **28**, 343–366.

Ph.L. Toint (1983), "VE08AD, a routine for partially separable optimization with bounded variables", Harwell Subroutine Library, A.E.R.E. (UK).

Ph.L. Toint (1986), "Global convergence of the partitioned BFGS algorithm for convex partially separable optimization", *Math. Programming* **36**, No. 3, 290–306.

Ph.L. Toint (1986a), "A view of nonlinear optimization in a large number of variables", Technical Report nr 86/16, Facultés Universitaires de Namur.

Ph.L. Toint (1986b), "Global convergence of the partioned BFGS algorithm for convex partially separable optimization", *Math. Prog.* **36**, 290–306.

V. Torczon (1991), "On the convergence of the multidimensional search algorithm", *SIAM J. Optimization* **1**,1, 123–145.

D. Touati-Ahmed and C. Storey (1990), "Efficient hybrid conjugate gradient techniques", *Journal of Optimization Theory and Applications* **64**, 379–397.

W. Warth and J. Werner (1977), "Effiziente Schrittweitenfunktionen bei unrestringierten Optimierungsaufgaben", *Computing* **19, 1**, 59–72.

J. Werner (1978), "Uber die globale konvergenz von Variable-Metric Verfahren mit nichtexakter Schrittweitenbestimmung", *Numerische Mathematik* **31** 321–334.

J. Werner (1989), "Global convergence of quasi-Newton methods with practical line searches", NAM-Bericht Nr. 67, Institut für Numerische und Angewandte Mathematik der Universität Göttingen.

J.H. Wilkinson (1965), *The Algebraic Eigenvalue Problem* , Oxford University Press (London).

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

P. Wolfe (1971), "Convergence conditions for ascent methods. II: Some corrections", *SIAM Rev.* **13**, 185–188.

Y. Yuan (1991), "A modified BFGS algorithm for unconstrained optimization", *IMA J. Numerical Analysis* **11**.

Y. Yuan and R. Byrd (1991), "Non-quasi-Newton updates for unconstrained optimization", Technical Report, Department of Computer Science, University of Colorado (Boulder).

Y. Zhang and R.P. Tewarson (1988), "Quasi-Newton algorithms with updates from the pre-convex part of Broyden's family", *IMA J. of Numer. Anal.* **8**, 487–509.

G. Zoutendijk (1970), "Nonlinear Programming, Computational Methods", in *Integer and Nonlinear Programming* (J. Abadie, ed.), North-Holland (Amsterdam), 37–86.