Variable Metric Methods for Constrained Optimization

M. J. D. Powell
D.A.M.T.P., University of Cambridge, Silver Street, Cambridge, CB3 9EW, England

Abstract. Variable metric methods solve nonlinearly constrained optimization problems, using calculated first derivatives and a single positive definite matrix, which holds second derivative information that is obtained automatically. The theory of these methods is shown by analysing the global and local convergence properties of a basic algorithm, and we find that superlinear convergence requires less second derivative information than in the unconstrained case. Moreover, in order to avoid the difficulties of inconsistent linear approximations to constraints, careful consideration is given to the calculation of search directions by unconstrained minimization subproblems. The Maratos effect and relations to reduced gradient algorithms are studied briefly.

1. Introduction

The methods to be considered are intended to calculate the least value of a function $F(\underset{\sim}{x})$, subject to equality and inequality constraints

$$\left.\begin{array}{ll} c_i(\underset{\sim}{x})=0, & i=1,2,\ldots,m' \\ c_i(\underset{\sim}{x})\geqslant 0, & i=m'+1,\ldots,m \end{array}\right\} \tag{1.1}$$

on the vector of variables $\underset{\sim}{x}\in\mathbb{R}^n$, where the functions $F(\underset{\sim}{x})$ and $\{c_i(\underset{\sim}{x});\ i=1,2,\ldots,m\}$ are real and differentiable. Either or both of the integers m' and $(m-m')$ may be zero. Each method generates a sequence of points $\{\underset{\sim}{x}_k;\ k=1,2,3,\ldots\}$ in $\mathbb{R}^n$, where $\underset{\sim}{x}_1$ is provided by the user, and where the sequence should converge to a solution, $\underset{\sim}{x}^*$ say, of the given nonlinear programming problem. It is assumed that the functions and their first derivatives can be calculated for any $\underset{\sim}{x}$.

In the unconstrained case, $m=m'=0$, the calculation of $\underset{\sim}{x}_{k+1}$ from $\underset{\sim}{x}_k$ by a variable metric method depends on the quadratic approximation

$$\bar{F}_k(\underset{\sim}{x})=F(\underset{\sim}{x}_k)+(\underset{\sim}{x}-\underset{\sim}{x}_k)^T\underset{\sim}{\nabla}F(\underset{\sim}{x}_k)+\tfrac{1}{2}(\underset{\sim}{x}-\underset{\sim}{x}_k)^T B_k(\underset{\sim}{x}-\underset{\sim}{x}_k) \tag{1.2}$$

to $F(\underset{\sim}{x})$, where B_k is an $n\times n$ positive definite symmetric matrix that is chosen automatically. Often $\underset{\sim}{x}_{k+1}$ is the vector $[\underset{\sim}{x}_k-B_k^{-1}\underset{\sim}{\nabla}F(\underset{\sim}{x}_k)]$ in order to minimize $\bar{F}_k(\underset{\sim}{x})$, provided that this value is acceptable to a "line search" or "trust region" technique, that is present to force convergence from poor starting approximations (see Fletcher [11] or Gill, Murray and Wright [16], for instance). The matrix B_k serves to provide information about second derivatives of $F(\underset{\sim}{x})$ that was obtained on previous iterations, but no second derivatives are calculated expli-

A. Bachem et al. (eds.), *Mathematical Programming The State of the Art*
© Springer-Verlag Berlin Heidelberg 1983

citly. The use of a positive definite matrix for this purpose in the unconstrained case characterizes a variable metric method.

When constraints are present, we define an algorithm to be a variable metric method if second derivative information from previous iterations, used in the calculation of $\underset{\sim}{x}_{k+1}$ from $\underset{\sim}{x}_k$, is also held in a symmetric matrix B_k, which now may be positive definite or positive semi-definite, and whose dimensions are at most $n \times n$. Allowing the rank of B_k to be less than n is necessary if one is to enjoy the savings that can be obtained from "reduced second derivative matrices" (see Fletcher [12], for instance). It is important to note that only one matrix is used for curvature information from both the objective and constraint functions. Often it is suitable to regard B_k as an approximation to the Hessian of the Lagrangian function of the given calculation at the solution $\underset{\sim}{x}^*$.

The following basic method is used by some variable metric algorithms to calculate $\underset{\sim}{x}_{k+1}$ from $\underset{\sim}{x}_k$ in the usual case when the dimensions of B_k are $n \times n$. The search direction $\underset{\sim}{d}_k$ is defined to be the vector $\underset{\sim}{d}$ in $\mathbb{R}^n$ that minimizes the function

$$\bar{F}_k(\underset{\sim}{x}_k+\underset{\sim}{d}) = F(\underset{\sim}{x}_k) + \underset{\sim}{d}^T \underset{\sim}{\nabla} F(\underset{\sim}{x}_k) + \tfrac{1}{2}\underset{\sim}{d}^T B_k \underset{\sim}{d} \tag{1.3}$$

subject to the linear constraints

$$\left.\begin{array}{ll} c_i(\underset{\sim}{x}_k) + \underset{\sim}{d}^T \underset{\sim}{\nabla} c_i(\underset{\sim}{x}_k) = 0, & i=1,2,\ldots,m' \\ c_i(\underset{\sim}{x}_k) + \underset{\sim}{d}^T \underset{\sim}{\nabla} c_i(\underset{\sim}{x}_k) \geqslant 0, & i=m'+1,\ldots,m \end{array}\right\}. \tag{1.4}$$

Therefore the calculation of $\underset{\sim}{d}_k$ is a convex quadratic programming problem. Then $\underset{\sim}{x}_{k+1}$ is given the value

$$\underset{\sim}{x}_{k+1} = \underset{\sim}{x}_k + \alpha_k \underset{\sim}{d}_k, \tag{1.5}$$

where α_k is a positive step length that is chosen by a line search procedure to give a reduction

$$W(\underset{\sim}{x}_{k+1}, \underset{\sim}{\mu}) < W(\underset{\sim}{x}_k, \underset{\sim}{\mu}) \tag{1.6}$$

in the line search objective function

$$W(\underset{\sim}{x}, \underset{\sim}{\mu}) = F(\underset{\sim}{x}) + \sum_{i=1}^{m'} \mu_i |c_i(\underset{\sim}{x})| + \sum_{i=m'+1}^{m} \mu_i \max[0, -c_i(\underset{\sim}{x})]. \tag{1.7}$$

Here $\underset{\sim}{\mu} \in \mathbb{R}^m$ is a vector of positive parameters that are held constant, except perhaps for some automatic adjustments on the early iterations to provide suitable values. We let α_k be the first number in a monotonically decreasing sequence $\{\alpha_{k1}, \alpha_{k2}, \alpha_{k3}, \ldots\}$ that is allowed by a condition that is a little stronger than inequality (1.6), where $\alpha_{k1}=1$, and where the ratios $\{\alpha_{kj+1}/\alpha_{kj};\ j=1,2,3,\ldots\}$ are all in a closed subinterval of (0, 1), for example [0.1, 0.5]. In practice one might use the freedom in each ratio to minimize the predicted value of $W(\underset{\sim}{x}_k + \alpha_{kj+1}\underset{\sim}{d}_k, \underset{\sim}{\mu})$.

It is proved in Section 2 that, under certain conditions, this basic method generates a sequence $\{\underset{\sim}{x}_k; k=1,2,3,\ldots\}$ whose limit points are all Kuhn-Tucker points of the given calculation. These conditions are less severe than those that occur in Han's [17] original analysis of global convergence, but even weaker conditions have been found by Chamberlain [5]. The restrictions on B_k are very mild indeed. Because they allow $B_k = I$ on each iteration, variable metric methods can be related to projected gradient algorithms.

It is well known, however, that if $B_k = I$ on each iteration, and if there are no constraints on the variables, then a version of the steepest descent method is obtained, that is often useless because it converges too slowly. Therefore the more successful variable metric algorithms for unconstrained optimization generate the matrices $\{B_k; k = 1, 2, 3, \ldots\}$ in such a way that the ratio

$$\|[B_k - \nabla^2 F(\underline{x}^*)]\underline{d}_k\| / \|\underline{d}_k\| \tag{1.8}$$

tends to zero as $k \to \infty$, if the sequence $\{\underline{x}_k; k = 1, 2, 3, \ldots\}$ converges to $\underline{x}^*$. In this case, if $\nabla^2 F(\underline{x}^*)$ is positive definite, then Q-superlinear convergence,

$$\lim_{k \to \infty} \|\underline{x}_k + \underline{d}_k - \underline{x}^*\| / \|\underline{x}_k - \underline{x}^*\| = 0, \tag{1.9}$$

is obtained (Dennis and Moré [9]). Section 3 gives analogous conditions on B_k for superlinear convergence in constrained calculations, much of the theory being taken from Powell [22]. Of course we find that, if the solution $\underline{x}^*$ is at the intersection of n constraint boundaries whose normals are linearly independent, and if the strict complementarity condition holds, then the final rate of convergence is independent of B_k, which suggests correctly that constrained calculations often converge more rapidly than unconstrained ones.

Unfortunately, even though the limit (1.9) may be obtained, condition (1.6) may prevent the choice $\underline{x}_{k+1} = \underline{x}_k + \underline{d}_k$ for every value of k. This phenomenon, which is known as the "Maratos effect", does not occur often in practice. However, Section 4 gives an example to show that the effect can cause a variable metric algorithm to be highly inefficient, and it describes briefly some remedies that have been proposed recently.

A much more common difficulty is that the linearizations (1.4) of the given nonlinear constraints (1.1) may introduce inconsistencies. For example, let $n = 1$ and let the constraints $x \leqslant 1$ and $x^2 \geqslant 0$ be linearized at $x_k = 3$. We obtain the inequalities $3 + d \leqslant 1$ and $9 + 6d \geqslant 0$, which have no solution. Therefore it is sometimes important to choose a search direction that does not satisfy the conditions (1.4). Suitable methods are suggested by Bartholomew-Biggs [2], Fletcher [13] and Powell [21]. Each of the recommended search directions can be defined as the vector $\underline{d}$ that solves an unconstrained minimization problem, whose objective function is $\bar{F}_k(\underline{x}_k + \underline{d})$ plus penalty terms that depend on the left hand sides of the linear constraints (1.4). This work is the subject of Section 5. Because it allows one to introduce further constraints on the search direction, it makes possible the use of trust regions in variable metric algorithms for constrained optimization.

In the final section some topics are mentioned briefly that are important to variable metric methods, but that cannot be given more consideration because of time and space restrictions. Particular attention is given to the relations between reduced gradient and variable metric algorithms.

2. Global Convergence

In this section we consider the convergence of the sequence $\{\underset{\sim}{x}_k;\, k=1,2,3,\ldots\}$ when each iteration calculates $\underset{\sim}{x}_{k+1}$ from $\underset{\sim}{x}_k$ by the basic method that is described in Section 1. Therefore we assume that the constraints (1.4) are consistent, and that each quadratic programming problem that determines a search direction has a bounded solution.

We make use of the Kuhn-Tucker conditions that hold at the solution of the quadratic programming problem that determines $\underset{\sim}{d}_k$. These conditions state the existence of Lagrange parameters $\{\lambda_i^{(k)};\, i=1,2,\ldots,m\}$ that satisfy the bounds

$$\lambda_i^{(k)} \geqslant 0, \qquad i=m'+1,\ldots,m, \tag{2.1}$$

and the equations

$$\underset{\sim}{\nabla} F(\underset{\sim}{x}_k)+B_k \underset{\sim}{d}_k=\sum_{i=1}^{m} \lambda_i^{(k)} \underset{\sim}{\nabla} c_i(\underset{\sim}{x}_k), \tag{2.2}$$

$$\lambda_i^{(k)}\{c_i(\underset{\sim}{x}_k)+\underset{\sim}{d}_k^T \underset{\sim}{\nabla} c_i(\underset{\sim}{x}_k)\}=0, \qquad i=1,2,\ldots,m. \tag{2.3}$$

Further, the conditions (1.4) are obtained when $\underset{\sim}{d}=\underset{\sim}{d}_k$. It follows that, if $\underset{\sim}{d}_k$ is zero, then $\underset{\sim}{x}_k$ is a Kuhn-Tucker point of the main calculation. Because one usually terminates a variable metric algorithm in this case, we assume from now on that none of the calculated vectors of variables are Kuhn-Tucker points of the given nonlinear programming problem.

The parameters $\{\mu_i;\, i=1,2,\ldots,m\}$ of the line search objective function (1.7) have to be such that the reduction (1.6) can be obtained for a positive value of the step length α_k. In order to identify suitable conditions on $\underset{\sim}{\mu}$, we note that the Lagrangian function

$$\bar{L}_k(\underset{\sim}{x}_k+\underset{\sim}{d})=\bar{F}_k(\underset{\sim}{x}_k+\underset{\sim}{d})-\sum_{i=1}^{m} \lambda_i^{(k)}\{c_i(\underset{\sim}{x}_k)+\underset{\sim}{d}^T \underset{\sim}{\nabla} c_i(\underset{\sim}{x}_k)\} \tag{2.4}$$

is a quadratic function of $\underset{\sim}{d}$, that is convex because its second derivative matrix is B_k, and that takes its least value when $\underset{\sim}{d}=\underset{\sim}{d}_k$ because of equation (2.2). Thus we deduce the relation

$$\begin{aligned}\bar{L}_k(\underset{\sim}{x}_k) &= \bar{L}_k(\underset{\sim}{x}_k+\underset{\sim}{d}_k)+\tfrac{1}{2}\underset{\sim}{d}_k^T B_k \underset{\sim}{d}_k \\ &\geqslant \bar{L}_k(\underset{\sim}{x}_k+\underset{\sim}{d}_k).\end{aligned} \tag{2.5}$$

Moreover the function

$$\begin{aligned}\bar{W}_k(\underset{\sim}{x}_k+\underset{\sim}{d}) &= \bar{F}_k(\underset{\sim}{x}_k+\underset{\sim}{d})+\sum_{i=1}^{m'} \mu_i|c_i(\underset{\sim}{x}_k)+\underset{\sim}{d}^T \underset{\sim}{\nabla} c_i(\underset{\sim}{x}_k)| \\ &\quad+\sum_{i=m'+1}^{m} \mu_i \max[0,\, -c_i(\underset{\sim}{x}_k)-\underset{\sim}{d}^T \underset{\sim}{\nabla} c_i(\underset{\sim}{x}_k)]\end{aligned} \tag{2.6}$$

is an approximation to $W(\underset{\sim}{x}_k+\underset{\sim}{d},\underset{\sim}{\mu})$ that satisfies the condition

$$\bar{W}_k(\underset{\sim}{x}_k+\underset{\sim}{d})=W(\underset{\sim}{x}_k+\underset{\sim}{d},\underset{\sim}{\mu})+o(\|\underset{\sim}{d}\|), \tag{2.7}$$

provided that the objective and constraint functions have continuous first derivatives. Therefore we compare $\bar{W}_k(\underset{\sim}{x}_k+\underset{\sim}{d})$ with $\bar{L}_k(\underset{\sim}{x}_k+\underset{\sim}{d})$.

The difference between these functions is bounded below by the expression

$$\bar{W}_k(\underset{\sim}{x}_k+\underset{\sim}{d})-\bar{L}_k(\underset{\sim}{x}_k+\underset{\sim}{d})\geqslant\sum_{i=1}^{m'}\{\mu_i-|\lambda_i^{(k)}|\}|c_i(\underset{\sim}{x}_k)+\underset{\sim}{d}^T\underset{\sim}{\nabla}c_i(\underset{\sim}{x}_k)|$$
$$+\sum_{i=m'+1}^{m}\{\mu_i-\lambda_i^{(k)}\}\max[0,\ -c_i(\underset{\sim}{x}_k)-\underset{\sim}{d}^T\underset{\sim}{\nabla}c_i(\underset{\sim}{x}_k)],\tag{2.8}$$

which depends on inequality (2.1). We require the components of $\underset{\sim}{\mu}$ to satisfy the bounds

$$\mu_i>|\lambda_i^{(k)}|,\qquad i=1,2,\ldots,m,\tag{2.9}$$

in order that the relation

$$\bar{W}_k(\underset{\sim}{x}_k+\underset{\sim}{d})>\bar{L}_k(\underset{\sim}{x}_k+\underset{\sim}{d})\tag{2.10}$$

is obtained if at least one of the conditions (1.4) does not hold.

In this case, if $\underset{\sim}{x}_k$ is infeasible with respect to the given nonlinear constraints (1.1), we have the bound

$$\begin{aligned}\bar{W}_k(\underset{\sim}{x}_k)&>\bar{L}_k(\underset{\sim}{x}_k)\\&\geqslant\bar{L}_k(\underset{\sim}{x}_k+\underset{\sim}{d}_k)=\bar{W}_k(\underset{\sim}{x}_k+\underset{\sim}{d}_k),\end{aligned}\tag{2.11}$$

while, if $\underset{\sim}{x}_k$ is a feasible point, we have the condition

$$\begin{aligned}\bar{W}_k(\underset{\sim}{x}_k)&=\bar{L}_k(\underset{\sim}{x}_k)\\&>\bar{L}_k(\underset{\sim}{x}_k+\underset{\sim}{d}_k)=\bar{W}_k(\underset{\sim}{x}_k+\underset{\sim}{d}_k),\end{aligned}\tag{2.12}$$

where the strict inequality depends on the observation that the term $\underset{\sim}{d}_k^T B_k \underset{\sim}{d}_k$ of expression (2.5) is positive, because otherwise $\underset{\sim}{d}_k=0$ would be a solution of the quadratic programming calculation. Hence the number

$$r_k=\bar{W}_k(\underset{\sim}{x}_k)-\bar{W}_k(\underset{\sim}{x}_k+\underset{\sim}{d}_k)\tag{2.13}$$

is positive. This remark is important because the convexity of the function (2.6) and equation (2.7) give the relation

$$\begin{aligned}W(\underset{\sim}{x}_k,\underset{\sim}{\mu})-W(\underset{\sim}{x}_k+\alpha\underset{\sim}{d}_k,\underset{\sim}{\mu})&=\bar{W}_k(\underset{\sim}{x}_k)-\bar{W}_k(\underset{\sim}{x}_k+\alpha\underset{\sim}{d}_k)+o(\alpha)\\&\geqslant\alpha[\bar{W}_k(\underset{\sim}{x}_k)-\bar{W}_k(\underset{\sim}{x}_k+\underset{\sim}{d}_k)]+o(\alpha)\\&=\alpha r_k+o(\alpha),\qquad 0\leqslant\alpha\leqslant1.\end{aligned}\tag{2.14}$$

Therefore the required reduction (1.6) can be achieved by choosing α_k to be sufficiently small and positive.

Further, we may replace the condition (1.6) on the step length by the inequality

$$W(\underset{\sim}{x}_{k+1},\underset{\sim}{\mu})\leqslant W(\underset{\sim}{x}_k,\underset{\sim}{\mu})-\sigma\alpha_k r_k,\tag{2.15}$$

where σ is any constant from the open interval (0, 1), for example $\sigma=0.1$ is usually suitable in practice. This is the stronger condition that is mentioned in Section 1, and it is important to the following convergence theorem.

Theorem 1. If the sequence $\{x_k; k=1,2,3,\ldots\}$ is calculated in the way that has been described, if the points of this sequence and the points $\{x_k+d_k; k=1,2,3,\ldots\}$ remain in a closed, bounded and convex region of $\mathbb{R}^n$ in which the objective and constraint functions have continuous first derivatives, if the matrices $\{B_k; k=1,2,3,\ldots\}$ are uniformly bounded, and if the components of μ satisfy the condition

$$\mu_i \geqslant |\lambda_i^{(k)}| + \rho \tag{2.16}$$

for all i and k, where ρ is a positive constant, then all limit points of the sequence $\{x_k; k=1,2,3,\ldots\}$ are Kuhn-Tucker points of the given nonlinear programming problem.

Proof. Let η be a small positive constant, and consider the iterations on which the number r_k, defined by equation (2.13), exceeds η. Because continuity of first derivatives in a compact domain is equivalent to uniform continuity, and because the vector d_k in expression (2.14) is bounded, it follows that there is a positive constant $\beta(\eta)$ such that inequality (2.15) holds for any α_k in the interval $[0, \beta(\eta)]$. Thus, remembering the way in which the step length is chosen (see Section 1), we deduce that $[W(x_k, \mu) - W(x_{k+1}, \mu)]$ is bounded away from zero if $r_k > \eta$. However, the reductions in the line search objective function tend to zero, because $\{W(x_k, \mu); k=1,2,3,\ldots\}$ is a monotonically decreasing sequence that is bounded below. Therefore r_k also tends to zero, which is the condition

$$\lim_{k\to\infty} [\bar{W}_k(x_k) - \bar{W}_k(x_k+d_k)] = 0. \tag{2.17}$$

Because at least one of the expressions (2.11) and (2.12) is satisfied for each k, equation (2.17) gives the limits

$$\lim_{k\to\infty} [\bar{W}_k(x_k) - \bar{L}_k(x_k)] = 0 \tag{2.18}$$

and

$$\lim_{k\to\infty} [\bar{L}_k(x_k) - \bar{L}_k(x_k+d_k)] = 0. \tag{2.19}$$

By letting $d=0$ in inequality (2.8), and by using condition (2.16), we deduce from expression (2.18) that every limit point of the sequence $\{x_k; k=1,2,3,\ldots\}$ satisfies the nonlinear constraints (1.1).

Our method of proof allows us to simplify the notation by assuming without loss of generality that $\{x_k; k=1,2,3,\ldots\}$ has only one limit point, x^* say. We define I^* to be the set of indices of the inequality constraints that are satisfied as equations at x^*, and for each k we let δ_k be the distance

$$\delta_k = \min_{y\in C} \|\nabla F(x_k) - y\|_2, \tag{2.20}$$

where $C \subset \mathbb{R}^n$ is the convex cone of points that can be expressed in the form

$$y = \sum_{i=1}^{m'} v_i \nabla c_i(x^*) + \sum_{i\in I^*} v_i \nabla c_i(x^*), \tag{2.21}$$

the coefficients $\{v_i; i=1,2,\dots,m'\}$ and $\{v_i; i\in I^*\}$ being real numbers that are unconstrained and non-negative respectively. It remains to prove that $\delta_k\to 0$, so it is sufficient to establish the condition

$$\lim_{k\to\infty} \|\nabla F(\underline{x}_k)-\sum_{i=1}^{m'}\lambda_i^{(k)}\nabla c_i(\underline{x}^*)-\sum_{i\in I^*}\lambda_i^{(k)}\nabla c_i(\underline{x}^*)\|=0, \tag{2.22}$$

where the Lagrange parameters $\{\lambda_i^{(k)}; i=1,2,\dots,m\}$ have been defined already.

Of course we make use of the fact that equation (2.2) gives the expression

$$\lim_{k\to\infty} \|\nabla F(\underline{x}_k)-\sum_{i=1}^{m}\lambda_i^{(k)}\nabla c_i(\underline{x}_k)+B_k\underline{d}_k\|=0. \tag{2.23}$$

Because the theorem states that the Lagrange multipliers are bounded, this limit is preserved if $\nabla c_i(\underline{x}_k)$ is replaced by $\nabla c_i(\underline{x}^*)$ for all i. Moreover, because equations (2.5) and (2.19) imply that $B_k^{1/2}\underline{d}_k\to 0$, we deduce from the boundedness of the matrices $\{B_k; k=1,2,3,\dots\}$ that expression (2.23) remains valid if the term $B_k\underline{d}_k$ is deleted. It follows that equation (2.22) is true if $\lambda_i^{(k)}$ tends to zero as $k\to\infty$, where i is the index of any inequality constraint that is not in I^*.

Let i be such an index, and let k be so large that $c_i(\underline{x}_k)$ is positive. By giving further attention to the *derivation* of inequality (2.8), we deduce the condition

$$\bar{W}_k(\underline{x}_k)-\bar{L}_k(\underline{x}_k)\geqslant\lambda_i^{(k)}c_i(\underline{x}_k). \tag{2.24}$$

Since $\lambda_i^{(k)}\geqslant 0$, and since $c_i(\underline{x}_k)\to c_i(\underline{x}^*)>0$, it follows from expressions (2.18) and (2.24) that $\lambda_i^{(k)}$ tends to zero as $k\to\infty$, which completes the proof of the theorem. □

This theorem compares favourably with other global convergence results that have been published for variable metric algorithms, because it allows the matrices $\{B_k; k=1,2,3,\dots\}$ to be positive semi-definite, and because the gradient vectors of the active constraints at $\underline{x}^*$ do not have to be linearly independent. We recall from Section 1, however, that the algorithm that is analysed has the disadvantage that useful search directions may exist when there is no solution to the quadratic programming problem that normally defines $\underline{d}_k$. We return to this question in Section 5.

We consider next the choice of the parameters $\{\mu_i; i=1,2,\dots,m\}$ of the line search objective function. Large constant values are not recommended, because variable metric methods for constrained optimization are faster than reduced gradient algorithms only if the calculated points $\{\underline{x}_k; k=1,2,3,\dots\}$ are allowed to move away from the boundaries of curved active constraints. Therefore an ideal algorithm would adjust $\underline{\mu}$ automatically, but the techniques that have been proposed already are rather crude. Chamberlain [4] shows that a method that I suggested can lead to cycling instead of convergence, so now I [23] prefer to set the components of $\underline{\mu}$ to very small positive values initially, which are increased if necessary during the calculation so that a positive step length can give the reduction (1.6) in the line search objective function. It is easy to obtain the property that, if $\underset{\sim}{\mu}$ remains bounded, then the number of iterations that change its value is finite. Thus convergence theorems that depend on constant $\underset{\sim}{\mu}$ are valid. It seems however that, given any technique for choosing $\underset{\sim}{\mu}$, it is possible to find pathological examples to show that the technique is inefficient, but serious difficulties are unusual in practice.

3. Superlinear Convergence

We assume in this section that a variable metric algorithm for constrained optimization calculates each search direction $\underset{\sim}{d}_k$ by the basic method of Section 1, and that the sequence $\{\underset{\sim}{x}_k; k=1,2,3,\ldots\}$ converges to a Kuhn-Tucker point $\underset{\sim}{x}^*$ of the given nonlinear programming problem. We study the important question of choosing the matrices $\{B_k; k=1,2,3,\ldots\}$ so that, if the step length of the line search is set to one for all sufficiently large k, which is the condition

$$\underset{\sim}{x}_{k+1}=\underset{\sim}{x}_k+\underset{\sim}{d}_k, \qquad k \geqslant k_0, \tag{3.1}$$

where k_0 is a constant, then the rate of convergence of the sequence $\{\underset{\sim}{x}_k; k=1,2,3,\ldots\}$ is superlinear. Our theory depends on several conditions that are usual in this kind of analysis.

Because superlinear convergence is obtained only if $\|\underset{\sim}{d}_k\|$ tends to zero as $k \rightarrow \infty$, we assume without loss of generality that any inequality constraints of the given nonlinear programming problem are satisfied as equations at $\underset{\sim}{x}^*$. We also make the usual assumption that any inequality constraints can be treated as equations for sufficiently large k, but it does lose generality unless the constraint gradients are linearly independent at $\underset{\sim}{x}^*$ and the strict complementarity condition holds. Therefore in this section $m'=m$ and the constraints on the variables are the equations

$$c_i(\underset{\sim}{x})=0, \qquad i=1,2,\ldots,m. \tag{3.2}$$

We require three more conditions. They are that all functions are twice continuously differentiable, that the first derivatives of the constraints $\{\underset{\sim}{\nabla} c_i(\underset{\sim}{x}^*); i=1,2,\ldots,m\}$ are linearly independent, and that second order sufficiency holds. To state the details of this last condition, we define the Lagrange parameters $\{\lambda_i^*; i=1,2,\ldots,m\}$ by the equation

$$\underset{\sim}{\nabla} F(\underset{\sim}{x}^*)=\sum_{i=1}^{m} \lambda_i^* \underset{\sim}{\nabla} c_i(\underset{\sim}{x}^*), \tag{3.3}$$

and we let G^* be the second derivative matrix

$$G^*=\nabla^2 F(\underset{\sim}{x}^*)-\sum_{i=1}^{m} \lambda_i^* \nabla^2 c_i(\underset{\sim}{x}^*). \tag{3.4}$$

Second order sufficiency states that, if $\underset{\sim}{d}$ is any non-zero vector that is orthogonal to the gradients $\{\underset{\sim}{\nabla} c_i(\underset{\sim}{x}^*); i=1,2,\ldots,m\}$, then $\underset{\sim}{d}^T G^* \underset{\sim}{d}$ is positive. Two consequences of these conditions are that the matrix

$$J(\underset{\sim}{x})=\left(\begin{array}{c|c} G^* & -N(\underset{\sim}{x}) \\ \hline -N(\underset{\sim}{x})^T & 0 \end{array}\right) \tag{3.5}$$

is nonsingular at $\underset{\sim}{x}=\underset{\sim}{x}^*$, where $N(\underset{\sim}{x})$ is the $n \times m$ matrix that has the columns $\{\underset{\sim}{\nabla} c_i(\underset{\sim}{x});\ i=1,2,\ldots,m\}$, and that there exists a positive number τ such that the matrix

$$G^*+\tau \sum_{i=1}^{m} \underset{\sim}{\nabla} c_i(\underset{\sim}{x}^*) \underset{\sim}{\nabla} c_i(\underset{\sim}{x}^*)^T \tag{3.6}$$

is positive definite (see Fletcher [12], for instance).

We will find it useful to take the point of view that the equations

$$\left.\begin{array}{l} \underset{\sim}{\nabla} F(\underset{\sim}{x})-\sum_{i=1}^{m} \lambda_i \underset{\sim}{\nabla} c_i(\underset{\sim}{x})=0 \\ -c_i(\underset{\sim}{x})=0, \quad i=1,2,\ldots,m \end{array}\right\} \tag{3.7}$$

are a square system in $(m+n)$ unknowns, namely the components of $\underset{\sim}{x}$ and $\underset{\sim}{\lambda}$. The point $(\underset{\sim}{x}^*, \underset{\sim}{\lambda}^*)$ is a solution of the system, and here the Jacobian is the nonsingular matrix $J(\underset{\sim}{x}^*)$. Therefore, if Newton's method for solving nonlinear equations is applied inductively to the system, and if the calculated points in $\mathbb{R}^{m+n}$ converge to $(\underset{\sim}{x}^*, \underset{\sim}{\lambda}^*)$, then the rate of convergence is superlinear. Further, this statement remains true if the true Jacobian matrix is replaced by $J(\underset{\sim}{x})$ for each $(\underset{\sim}{x}, \underset{\sim}{\lambda})$. Thus we deduce the limit

$$\lim_{k\to\infty} \frac{\|\underset{\sim}{x}_k+\underset{\sim}{\delta}_k-\underset{\sim}{x}^*\|+\|\underset{\sim}{\lambda}_k+\underset{\sim}{\eta}_k-\underset{\sim}{\lambda}^*\|}{\|\underset{\sim}{x}_k-\underset{\sim}{x}^*\|+\|\underset{\sim}{\lambda}_k-\underset{\sim}{\lambda}^*\|}=0, \tag{3.8}$$

provided that $\underset{\sim}{\lambda}_k$ is sufficiently close to $\underset{\sim}{\lambda}^*$, where $\underset{\sim}{\delta}_k \in \mathbb{R}^n$ and $\underset{\sim}{\eta}_k \in \mathbb{R}^m$ are defined by the system

$$J(\underset{\sim}{x}_k)\left(\begin{array}{c} \underset{\sim}{\delta}_k \\ \hline \underset{\sim}{\eta}_k \end{array}\right)=\left(\begin{array}{c} -\underset{\sim}{\nabla} F(\underset{\sim}{x}_k)+N(\underset{\sim}{x}_k)\underset{\sim}{\lambda}_k \\ \hline \underset{\sim}{c}(\underset{\sim}{x}_k) \end{array}\right). \tag{3.9}$$

However, $\underset{\sim}{\delta}_k$ is independent of $\underset{\sim}{\lambda}_k$ because expression (3.9) is equivalent to the equations

$$J(\underset{\sim}{x}_k)\left(\begin{array}{c} \underset{\sim}{\delta}_k \\ \hline \underset{\sim}{\eta}_k+\underset{\sim}{\lambda}_k \end{array}\right)=\left(\begin{array}{c} -\underset{\sim}{\nabla} F(\underset{\sim}{x}_k) \\ \hline \underset{\sim}{c}(\underset{\sim}{x}_k) \end{array}\right). \tag{3.10}$$

Therefore we may let $\underset{\sim}{\lambda}_k=\underset{\sim}{\lambda}^*$ in expression (3.8), in order to obtain the condition

$$\lim_{k\to\infty} \|\underset{\sim}{x}_k+\underset{\sim}{\delta}_k-\underset{\sim}{x}^*\| / \|\underset{\sim}{x}_k-\underset{\sim}{x}^*\|=0. \tag{3.11}$$

This limit is used in the second half of the proof of the following theorem, which extends to the constrained case the Q-superlinear convergence result of Dennis and Moré [9]. A similar theorem is proved by Boggs, Tolle and Wang [3], but their conditions on $\{B_k;\ k=1,2,3,\ldots\}$ and $\{\underset{\sim}{x}_k;\ k=1,2,3,\ldots\}$ are stronger than ours.

Theorem 2. Let the conditions of the first three paragraphs of this section be satisfied, and for each k let σ_k be the number

$$\sigma_k = \min_{\underset{\sim}{\phi} \in \mathbb{R}^m} \|(B_k - G^*)\underset{\sim}{d}_k - \sum_{i=1}^{m} \phi_i \underset{\sim}{\nabla} c_i(\underset{\sim}{x}_k)\|_2, \tag{3.12}$$

where G^* is the matrix (3.4). Then the Q-superlinear convergence condition

$$\lim_{k \to \infty} \|\underset{\sim}{x}_k + \underset{\sim}{d}_k - \underset{\sim}{x}^*\| / \|\underset{\sim}{x}_k - \underset{\sim}{x}^*\| = 0 \tag{3.13}$$

is equivalent to the limit

$$\lim_{k \to \infty} \sigma_k / \|\underset{\sim}{d}_k\| = 0. \tag{3.14}$$

Proof. First we assume that condition (3.13) holds. Thus, because the Lagrangian function

$$L(\underset{\sim}{x}) = F(\underset{\sim}{x}) - \sum_{i=1}^{m} \lambda_i^* c_i(\underset{\sim}{x}) \tag{3.15}$$

is stationary at $\underset{\sim}{x}^*$, we deduce the relation

$$\lim_{k \to \infty} \|\underset{\sim}{\nabla} L(\underset{\sim}{x}_k + \underset{\sim}{d}_k)\| / \|\underset{\sim}{d}_k\| = 0. \tag{3.16}$$

Equation (3.4), the continuity of second derivatives, equation (2.2) and the definition (3.12) imply the inequality

$$\begin{aligned}
\|\underset{\sim}{\nabla} L(\underset{\sim}{x}_k + \underset{\sim}{d}_k)\| &= \|\underset{\sim}{\nabla} L(\underset{\sim}{x}_k) + G^* \underset{\sim}{d}_k\| + o(\|\underset{\sim}{d}_k\|) \\
&= \|\underset{\sim}{\nabla} F(\underset{\sim}{x}_k) - \sum_{i=1}^{m} \lambda_i^* \underset{\sim}{\nabla} c_i(\underset{\sim}{x}_k) + G^* \underset{\sim}{d}_k\| + o(\|\underset{\sim}{d}_k\|) \\
&= \|(G^* - B_k)\underset{\sim}{d}_k + \sum_{i=1}^{m} (\lambda_i^{(k)} - \lambda_i^*) \underset{\sim}{\nabla} c_i(\underset{\sim}{x}_k)\| + o(\|\underset{\sim}{d}_k\|) \\
&\geqslant \sigma_k + o(\|\underset{\sim}{d}_k\|).
\end{aligned} \tag{3.17}$$

Therefore the limit (3.14) is a consequence of expression (3.16).

To prove the converse result, we compare $\underset{\sim}{d}_k$ with the vector $\underset{\sim}{\delta}_k$ that is defined by equation (3.10). Therefore we write equation (2.2) and the linear constraints on $\underset{\sim}{d}_k$ in the form

$$\left(\begin{array}{c|c} B_k & -N(\underset{\sim}{x}_k) \\ \hline -N(\underset{\sim}{x}_k)^T & 0 \end{array}\right) \left(\begin{array}{c} \underset{\sim}{d}_k \\ \hline \underset{\sim}{\lambda}^{(k)} \end{array}\right) = \left(\begin{array}{c} -\underset{\sim}{\nabla} F(\underset{\sim}{x}_k) \\ \hline \underset{\sim}{c}(\underset{\sim}{x}_k) \end{array}\right), \tag{3.18}$$

which we subtract from expression (3.10) to obtain the identity

$$J(\underset{\sim}{x}_k) \left(\begin{array}{c} \underset{\sim}{\delta}_k - \underset{\sim}{d}_k \\ \hline \underset{\sim}{\eta}_k + \underset{\sim}{\lambda}_k - \underset{\sim}{\lambda}^{(k)} \end{array}\right) = \left(\begin{array}{c} (B_k - G^*)\underset{\sim}{d}_k \\ \hline 0 \end{array}\right). \tag{3.19}$$

Further, the argument that gives equation (3.10) from (3.9) also provides the identity

$$J(\underset{\sim}{x}_k) \left(\begin{array}{c} \underset{\sim}{\delta}_k - \underset{\sim}{d}_k \\ \hline \underset{\sim}{\eta}_k + \underset{\sim}{\lambda}_k - \underset{\sim}{\lambda}^{(k)} + \underset{\sim}{\phi}_k \end{array}\right) = \left(\begin{array}{c} (B_k - G^*)\underset{\sim}{d}_k - N(\underset{\sim}{x}_k)\underset{\sim}{\phi}_k \\ \hline 0 \end{array}\right), \tag{3.20}$$

where we let ϕ_k be the value of ϕ that minimizes expression (3.12). Since $J(\underset{\sim}{x}_k)$ tends to a nonsingular matrix, it follows that the bound

$$\|\underset{\sim}{\delta}_k - \underset{\sim}{d}_k\| = O(\sigma_k) \tag{3.21}$$

is obtained as $k \to \infty$.

Therefore, if condition (3.14) holds, we have $\|\underset{\sim}{\delta}_k - \underset{\sim}{d}_k\| = o(\|\underset{\sim}{d}_k\|)$. Because equation (3.11) can be expressed in the form $\|\underset{\sim}{\delta}_k - (\underset{\sim}{x}^* - \underset{\sim}{x}_k)\| = o(\|\underset{\sim}{x}_k - \underset{\sim}{x}^*\|)$, it follows that the limit

$$\|\underset{\sim}{d}_k - (\underset{\sim}{x}^* - \underset{\sim}{x}_k)\| = o(\|\underset{\sim}{x}_k - \underset{\sim}{x}^*\|) \tag{3.22}$$

is obtained. This limit is the same as the required condition (3.13), which completes the proof of the theorem. □

A corollary of this theorem is that, if the conditions of this section hold, then there exist positive definite matrices $\{B_k;\ k = 1, 2, 3, \dots\}$ that give superlinear convergence, even though G^* may have some negative eigenvalues. To prove this statement we recall that the matrix (3.6) is positive definite. Therefore there exists an integer k_1 such that we may make the choice

$$B_k = G^* + \tau \sum_{i=1}^{m} \underset{\sim}{\nabla} c_i(\underset{\sim}{x}_k) \underset{\sim}{\nabla} c_i(\underset{\sim}{x}_k)^T, \qquad k \geq k_1. \tag{3.23}$$

In this case the numbers $\{\sigma_k;\ k \geq k_1\}$ are all zero, so superlinear convergence is a consequence of Theorem 2.

We note that the superlinear convergence condition (3.14) reduces to expression (1.8) in the unconstrained case, and that superlinear convergence is independent of B_k when $m = n$ and the constraint gradients are linearly independent.

We now take the point of view that $\{\underset{\sim}{\nabla} c_i(\underset{\sim}{x}_k);\ i = 1, 2, \dots m\}$ are known, and that we wish to obtain superlinear convergence by choosing B_k in a way that is independent of $\underset{\sim}{d}_k$. Condition (3.14) holds if B_k has the form

$$B_k = G^* + \sum_{i=1}^{m} \sum_{j=1}^{m} \underset{\sim}{\nabla} c_i(\underset{\sim}{x}_k) S_{ij}^{(k)} \underset{\sim}{\nabla} c_j(\underset{\sim}{x}_k)^T + o(\|\underset{\sim}{d}_k\|), \tag{3.24}$$

where $S^{(k)}$ is a symmetric $m \times m$ matrix. Hence, ignoring the $o(\|\underset{\sim}{d}_k\|)$ term, there are $\frac{1}{2}[n(n+1) - m(m+1)]$ degrees of freedom in B_k to be determined. However, the following theorem gives a superlinear convergence result that requires the elements of B_k to satisfy only $\frac{1}{2}(n-m)(n-m+1)$ conditions. Thus the reduction in the number of conditions is substantial when n is large and m is close to n. We note in Section 6 that it is related to reduced gradient methods.

This theorem is a little stronger than Theorem 1 of Powell [22], because it does not require $\underset{\sim}{d}^T B_k \underset{\sim}{d}$ to be bounded below by a constant positive multiple of $\|\underset{\sim}{d}\|^2$, when $\underset{\sim}{d}$ is any vector that is orthogonal to the gradients $\{\underset{\sim}{\nabla} c_i(\underset{\sim}{x}^*);\ i = 1, 2, \dots, m\}$.

Theorem 3. Let the conditions of the first three paragraphs of this section be satisfied, let the matrices $\{B_k;\ k = 1, 2, 3, \dots\}$ be bounded, and for each k let P_k be the symmetric projection matrix such that, for any $\underset{\sim}{y} \in \mathbb{R}^n$, $P_k \underset{\sim}{y}$ is the vector of least Euclidean length of the form

$$P_k \underset{\sim}{y} = \underset{\sim}{y} - \sum_{i=1}^{m} \phi_i \underset{\sim}{\nabla} c_i (\underset{\sim}{x}_k). \tag{3.25}$$

If the condition

$$\lim_{k \to \infty} \|P_k (B_k - G^*) P_k \underset{\sim}{d}_k\| / \|\underset{\sim}{d}_k\| = 0 \tag{3.26}$$

holds, then the variable metric algorithm gives the two-step superlinear rate of convergence

$$\lim_{k \to \infty} \|\underset{\sim}{x}_{k+1} - \underset{\sim}{x}^*\| / \|\underset{\sim}{x}_{k-1} - \underset{\sim}{x}^*\| = 0. \tag{3.27}$$

Proof. First we show that $\|\underset{\sim}{d}_k\|$ is not much larger than $\|\underset{\sim}{x}_k - \underset{\sim}{x}^*\|$. Because the right hand side of equation (3.12) is just $\|P_k (B_k - G^*) \underset{\sim}{d}_k\|$, and because P_k is the operator

$$P_k = I - N(\underset{\sim}{x}_k)[N(\underset{\sim}{x}_k)^T N(\underset{\sim}{x}_k)]^{-1} N(\underset{\sim}{x}_k)^T, \tag{3.28}$$

the limit (3.26) implies the bound

$$\begin{aligned} \sigma_k &= \|P_k (B_k - G^*) N(\underset{\sim}{x}_k)[N(\underset{\sim}{x}_k)^T N(\underset{\sim}{x}_k)]^{-1} N(\underset{\sim}{x}_k)^T \underset{\sim}{d}_k\| + o(\|\underset{\sim}{d}_k\|) \\ &\leqslant M_1 \|\underset{\sim}{x}_k - \underset{\sim}{x}^*\| + o(\|\underset{\sim}{d}_k\|), \end{aligned} \tag{3.29}$$

where M_1 is a constant, and where the last line depends on the constraint

$$\underset{\sim}{c}(\underset{\sim}{x}_k) + N(\underset{\sim}{x}_k)^T \underset{\sim}{d}_k = 0 \tag{3.30}$$

in the quadratic programming problem that determines $\underset{\sim}{d}_k$. Since equation (3.21) is still valid, expression (3.29) gives the relation

$$\|\underset{\sim}{d}_k\| - \|\underset{\sim}{\delta}_k\| \leqslant M_2 \|\underset{\sim}{x}_k - \underset{\sim}{x}^*\| + o(\|\underset{\sim}{d}_k\|) \tag{3.31}$$

for some constant M_2. It follows from the limit (3.11) that $\|\underset{\sim}{d}_k\|$ is bounded by the inequality

$$\|\underset{\sim}{d}_k\| \leqslant M_3 \|\underset{\sim}{x}_k - \underset{\sim}{x}^*\|, \tag{3.32}$$

where M_3 is another positive constant.

As in Theorem 2, our proof of superlinear convergence depends on the closeness of $\underset{\sim}{d}_k$ to $\underset{\sim}{\delta}_k$. In order to use equation (3.21) again, we note that the first part of expression (3.29) and the constraint (3.30) give the limit

$$\begin{aligned} \sigma_k &= O(\|\underset{\sim}{c}(\underset{\sim}{x}_k)\|) + o(\|\underset{\sim}{d}_k\|) \\ &= O(\|\underset{\sim}{d}_{k-1}\|^2) + o(\|\underset{\sim}{d}_k\|), \end{aligned} \tag{3.33}$$

where the last line depends on the remark that $\underset{\sim}{c}(\underset{\sim}{x}_k)$ is the error of the approximation

$$\underset{\sim}{c}(\underset{\sim}{x}_{k-1} + \underset{\sim}{d}_{k-1}) \approx \underset{\sim}{c}(\underset{\sim}{x}_{k-1}) + N(\underset{\sim}{x}_{k-1})^T \underset{\sim}{d}_{k-1}. \tag{3.34}$$

It follows from expressions (3.21), (3.32) and (3.33) that the limit

$$\|\underset{\sim}{\delta}_k - \underset{\sim}{d}_k\| = o(\|\underset{\sim}{x}_{k-1} - \underset{\sim}{x}^*\| + \|\underset{\sim}{x}_k - \underset{\sim}{x}^*\|) \tag{3.35}$$

is obtained. Therefore the triangle inequality and equation (3.11) imply the relation

$$\|\underset{\sim}{x}_k+\underset{\sim}{d}_k-\underset{\sim}{x}^*\| \leqslant \|\underset{\sim}{x}_k+\underset{\sim}{\delta}_k-\underset{\sim}{x}^*\|+\|\underset{\sim}{\delta}_k-\underset{\sim}{d}_k\| \\ = o(\|\underset{\sim}{x}_{k-1}-\underset{\sim}{x}^*\|+\|\underset{\sim}{x}_k-\underset{\sim}{x}^*\|). \tag{3.36}$$

The required result (3.27) now follows from the fact that $\|\underset{\sim}{x}_k-\underset{\sim}{x}^*\|$ is bounded above by the sum

$$\|\underset{\sim}{x}_{k-1}-\underset{\sim}{x}^*\|+\|\underset{\sim}{d}_{k-1}\| \leqslant (1+M_3)\|\underset{\sim}{x}_{k-1}-\underset{\sim}{x}^*\|, \tag{3.37}$$

so the proof of Theorem 3 is complete. □

It is interesting to compare Theorems 2 and 3 in the case when the constraints (3.2) depend on only the first m components of $\underset{\sim}{x}$. We consider the partitioned matrices

$$B_k=\left(\begin{array}{c|c} B_{11}^{(k)} & B_{12}^{(k)} \\ \hline B_{21}^{(k)} & B_{22}^{(k)} \end{array}\right) \tag{3.38}$$

and

$$G^*=\left(\begin{array}{c|c} G_{11}^* & G_{12}^* \\ \hline G_{21}^* & G_{22}^* \end{array}\right), \tag{3.39}$$

where the dimensions of $B_{11}^{(k)}$ and G_{11}^* are $m\times m$. The superlinear convergence result of Theorem 2 is independent of the submatrix $B_{11}^{(k)}$, while Theorem 3 is independent of the submatrices $B_{11}^{(k)}$, $B_{12}^{(k)}$, and $B_{21}^{(k)}$, provided that B_k is uniformly bounded. Thus superlinear convergence can be obtained when the rank of B_k is only $(n-m)$.

A useful technique for generating the matrices $\{B_k;\ k=1,2,3,\ldots\}$ so that two-step superlinear convergence is obtained is described in [21] and analysed in [22]. It applies the well-known BFGS formula from unconstrained optimization, but the change in gradient of the objective function $[\underset{\sim}{\nabla} F(\underset{\sim}{x}_{k+1})-\underset{\sim}{\nabla} F(\underset{\sim}{x}_k)]$ is replaced by the change in gradient of an estimate of the Lagrangian function. Specifically the difference

$$\underset{\sim}{\nabla} F(\underset{\sim}{x}_{k+1})-\underset{\sim}{\nabla} F(\underset{\sim}{x}_k)-\sum_{i=1}^{m}\lambda_i^{(k)}\{\underset{\sim}{\nabla} c_i(\underset{\sim}{x}_{k+1})-\underset{\sim}{\nabla} c_i(\underset{\sim}{x}_k)\} \tag{3.40}$$

is used, where the multipliers $\{\lambda_i^{(k)};\ i=1,2,\ldots,m\}$ are still the Lagrange parameters of the quadratic programming problem that determines $\underset{\sim}{d}_k$, expect that a further modification is sometimes made to preserve positive definiteness. Thus superlinear convergence is achieved without the explicit calculation of any second derivatives.

4. The Maratos Effect

We consider the following calculation:

$$\left.\begin{array}{ll} \text{minimize} & F(\underset{\sim}{x})=-x_1+10(x_1^2+x_2^2-1), \\ \text{subject to} & x_1^2+x_2^2-1=0, \\ \text{starting at} & (x_1,x_2)=(0.8,\,0.6). \end{array}\right\} \tag{4.1}$$

Powell [23] reports that, if it is solved by a variable metric method of the type that is analysed in Section 2, then after 35 iterations the estimate of the solution (1,0) is (0.9887, 0.1613) and convergence occurs at a very slow linear rate. However, if one gives up line searches and uses the formula

$$\underset{\sim}{x}_{k+1} = \underset{\sim}{x}_k + \underset{\sim}{d}_k \tag{4.2}$$

instead, then six decimals accuracy are obtained in only five iterations. The reason is that the iteration (4.2), which gives superlinear convergence, is not allowed by the line search condition (1.6). This unfortunate phenomenon is called the "Maratos effect", because it is observed and considered in his Ph.D. dissertation [18].

It is easy to analyse the effect in the example (4.1) when $\underset{\sim}{x}_k$ is feasible and $B_k = G^* = I$. If the components of $\underset{\sim}{x}_k$ are $(\cos\theta, \sin\theta)$, then the components of $\underset{\sim}{x}_k + \underset{\sim}{d}_k$ are $(\cos\theta + \sin^2\theta,\ \sin\theta[1-\cos\theta])$. Thus the iteration (4.2) would converge quadratically, but the line search objective function (1.7) has the value

$$W(\underset{\sim}{x}_k + \underset{\sim}{d}_k, \mu) = -\cos\theta + (9+\mu)\sin^2\theta, \tag{4.3}$$

which exceeds $W(\underset{\sim}{x}_k, \mu) = -\cos\theta$ for every $\mu \geqslant 0$. Therefore the condition (1.6) demands a step length that prevents a superlinear rate of convergence.

The effect would not be serious if it occurred on only a few iterations of a constrained optimization calculation, but the example (4.1) shows that it can persist. However, if the starting point of the example is moved away from the constraint, then usually the line search objective function allows step lengths of one. Because I disagree with the writers who suggest that the effect is so rare that it can be ignored, the remainder of this section mentions some useful remedies.

The remedy that is proposed by Mayne and Polak [19] can be derived from the following remarks. Suppose that $\underset{\sim}{x}_k$ is feasible and is close to the Kuhn-Tucker point $\underset{\sim}{x}^*$, and that B_k gives the superlinear convergence condition

$$\|\underset{\sim}{x}_k + \underset{\sim}{d}_k - \underset{\sim}{x}^*\| = o(\|\underset{\sim}{d}_k\|), \tag{4.4}$$

where $\underset{\sim}{d}_k$ is calculated by the basic method of Section 1. As before, let $\{\lambda_i^*;\ i = 1, 2, \ldots, m\}$ be the Lagrange parameters at $\underset{\sim}{x}^*$, and let $L(\underset{\sim}{x})$ be the Lagrangian function (3.15). Then, assuming that $\{\mu_i \geqslant |\lambda_i^*|;\ i = 1, 2, \ldots, m\}$ we have $W(\underset{\sim}{x}_k, \underset{\sim}{\mu}) \geqslant L(\underset{\sim}{x}_k)$, and, assuming the second order sufficiency condition, we have the bounds

$$\left.\begin{aligned} L(\underset{\sim}{x}_k) &\geqslant L(\underset{\sim}{x}^*) + \eta\|\underset{\sim}{d}_k\|^2 \\ L(\underset{\sim}{x}_k + \underset{\sim}{d}_k) &= L(\underset{\sim}{x}^*) + o(\|\underset{\sim}{d}_k\|^2) \end{aligned}\right\}, \tag{4.5}$$

where η is a positive constant. It follows that the inequality

$$W(\underset{\sim}{x}_k + \underset{\sim}{d}_k, \underset{\sim}{\mu}) \leqslant W(\underset{\sim}{x}_k, \underset{\sim}{\mu}) - \eta\|\underset{\sim}{d}_k\|^2 + o(\|\underset{\sim}{d}_k\|^2) + V(\underset{\sim}{x}_k + \underset{\sim}{d}_k) \tag{4.6}$$

holds, where $V(\underset{\sim}{x})$ is the function

$$V(\underset{\sim}{x}) = \sum_{i=1}^{m} \lambda_i^* c_i(\underset{\sim}{x}) + \sum_{i=1}^{m'} \mu_i |c_i(\underset{\sim}{x})| + \sum_{i=m'+1}^{m} \mu_i \max[0, -c_i(\underset{\sim}{x})]. \tag{4.7}$$

Thus the Maratos effect does not occur if $V(\underset{\sim}{x}_k+\underset{\sim}{d}_k)$ is $o(\|\underset{\sim}{d}_k\|^2)$, but usually it is $O(\|\underset{\sim}{d}_k\|^2)$, because the linear conditions (1.4) imply that any violations of the given nonlinear constraints (1.1) at $(\underset{\sim}{x}_k+\underset{\sim}{d}_k)$ are of this magnitude.

Therefore the technique of Mayne and Polak [19] gives $\underset{\sim}{x}_{k+1}$ the form

$$\underset{\sim}{x}_{k+1}=\underset{\sim}{x}_k+\underset{\sim}{d}_k+\tilde{\underset{\sim}{d}}_k, \tag{4.8}$$

where $\tilde{\underset{\sim}{d}}_k$ is a small correction to $\underset{\sim}{d}_k$ such that $V(\underset{\sim}{x}_k+\underset{\sim}{d}_k+\tilde{\underset{\sim}{d}}_k)$ is normally $O(\|\underset{\sim}{d}_k\|^3)$. Specifically, they define I_k to be the set of constraint indices

$$I_k=\{1,2,\ldots,m'\}\cup\{i:\lambda_i^{(k)}>0\}, \tag{4.9}$$

and they let $\tilde{\underset{\sim}{d}}_k$ be the vector of least Euclidean length that satisfies the equations

$$c_i(\underset{\sim}{x}_k+\underset{\sim}{d}_k)+\tilde{\underset{\sim}{d}}_k^T\underset{\sim}{\nabla}c_i(\underset{\sim}{x}_k)=0, \quad i\in I_k. \tag{4.10}$$

Thus, in non-degenerate cases, $\|\tilde{\underset{\sim}{d}}_k\|$ is of magnitude $\|\underset{\sim}{d}_k\|^2$, and, provided that the indices of the constraints that make positive contributions to $V(\underset{\sim}{x}_k+\underset{\sim}{d}_k+\tilde{\underset{\sim}{d}}_k)$ are all in I_k, we have $V(\underset{\sim}{x}_k+\underset{\sim}{d}_k+\tilde{\underset{\sim}{d}}_k)=O(\|\underset{\sim}{d}_k\|^3)$ as required. However, if $\underset{\sim}{x}_k$ is not sufficiently close to $\underset{\sim}{x}^*$, the line search objective function may not allow the value (4.8). In this case the step length α_k is calculated as described in Section 1, but $\underset{\sim}{x}_{k+1}$ has the form

$$\underset{\sim}{x}_{k+1}=\underset{\sim}{x}_k+\alpha_k\underset{\sim}{d}_k+\alpha_k^2\tilde{\underset{\sim}{d}}_k, \tag{4.11}$$

in order that condition (2.15) can still be satisfied by choosing sufficiently small positive step lengths.

When there are no inactive inequality constraints in the quadratic programming calculation that determines $\underset{\sim}{d}_k$, the following modification to the method of the previous paragraph is sometimes useful. We replace the conditions (1.4) by the equations

$$\underset{\sim}{d}^T\underset{\sim}{\nabla}c_i(\underset{\sim}{x}_k)=0, \quad i=1,2,\ldots,m, \tag{4.12}$$

before calculating $\underset{\sim}{d}_k$, so the function values $\{c_i(\underset{\sim}{x}_k);\ i=1,2,\ldots,m\}$ are not required. Then $\tilde{\underset{\sim}{d}}_k$ is determined as before to satisfy expression (4.10). Coleman and Conn [7] study the asymptotic convergence properties of this technique when $\underset{\sim}{x}_{k+1}$ is given the value (4.8) on every iteration, and they show that the Maratos effect does not occur.

A different remedy is proposed by Chamberlain, Lemaréchal, Pedersen and Powell [6]. It is based on the observation that, if $\underset{\sim}{x}_{k+1}$ has the value (4.2) on every iteration, and if the superlinear convergence of Section 3 is obtained, then, for sufficiently large k, the line search objective function satisfies the inequality

$$W(\underset{\sim}{x}_{k+2},\underset{\sim}{\mu})<W(\underset{\sim}{x}_k,\underset{\sim}{\mu}). \tag{4.13}$$

Thus the disadvantages of the Maratos effect can be avoided by allowing an iteration to increase the line search objective function if, at the beginning of the iteration, the value of $W(\underset{\sim}{x},\underset{\sim}{\mu})$ is the least that has been calculated. The "watchdog technique" of Chamberlain et al. [6] applies this idea. If a new least value of $W(\underset{\sim}{x},\underset{\sim}{\mu})$ does not occur during a prescribed number of iterations, then the vector of variables is reset to the value that gave the least value of $W(\underset{\sim}{x}, \underset{\sim}{\mu})$, and then one insists on a line search that will make a further reduction in $\tilde{W}(\underset{\sim}{x}, \underset{\sim}{\mu})$. In this way one usually obtains the property that a limit point of the calculated sequence $\{\underset{\sim}{x}_k; k=1, 2, 3, \ldots\}$ is a Kuhn-Tucker point of the given nonlinear programming problem. If an iteration increases $W(\underset{\sim}{x}, \underset{\sim}{\mu})$, then a reduction is required in the estimate of the Lagrangian function

$$L_k(\underset{\sim}{x}) = F(\underset{\sim}{x}) - \sum_{i=1}^{m} \lambda_i^{(k)} c_i(\underset{\sim}{x}), \tag{4.14}$$

in order to try to prevent excessive changes to the variables. It is proved that, even with this refinement, the Maratos effect does not prevent a superlinear rate of convergence. Moreover, the technique can avoid some of the inefficiencies, discussed in the last paragraph of Section 2, that may be caused by a poor choice of $\underset{\sim}{\mu}$.

5. Unconstrained Calculations of the Search Direction

Instead of using the basic quadratic programming method of Section 1, the search direction of an iteration of a variable metric algorithm may be defined by an unconstrained minimization calculation. Usually the unconstrained problem depends on parameters in such a way that, if the parameters are sufficiently large, and if the constraints (1.4) are consistent, then $\underset{\sim}{d}_k$ is obtained, where in this section we reserve the notation $\underset{\sim}{d}_k$ for the solution of the quadratic programming problem of Section 1. However, two advantages of the unconstrained approach are that it can provide useful directions when the quadratic programming problem has no solution, and that it allows the use of "trust regions". These ideas are explained well by Fletcher [12], [13] and [14], and we consider them briefly.

The usual unconstrained approach is to let the search direction be the value of $\underset{\sim}{d}$ that minimizes the function

$$\begin{aligned}\bar{W}_k(\underset{\sim}{x}_k + \underset{\sim}{d}) = \bar{F}_k(\underset{\sim}{x}_k + \underset{\sim}{d}) &+ \sum_{i=1}^{m'} \mu_i |c_i(\underset{\sim}{x}_k) + \underset{\sim}{d}^T \underset{\sim}{\nabla} c_i(\underset{\sim}{x}_k)| \\ &+ \sum_{i=m'+1}^{m} \mu_i \max[0, -c_i(\underset{\sim}{x}_k) - \underset{\sim}{d}^T \underset{\sim}{\nabla} c_i(\underset{\sim}{x}_k)],\end{aligned} \tag{5.1}$$

where we are deliberately using the same notation as equation (2.6), and where the parameters $\{\mu_i;\ i=1, 2, \ldots, m\}$ are non-negative. This calculation has a unique solution, $\bar{\underset{\sim}{d}}_k$ say, whenever the function (5.1) is strictly convex. We note that $\bar{\underset{\sim}{d}}_k = \underset{\sim}{d}_k$ if and only if the terms under the summation signs of expression (5.1) are zero when $\underset{\sim}{d} = \bar{\underset{\sim}{d}}_k$. This condition is necessary because $\underset{\sim}{d}_k$ satisfies the constraints (1.4), and it is also sufficient because otherwise the replacement of $\bar{\underset{\sim}{d}}_k$ by $\underset{\sim}{d}_k$ would reduce the function (5.1).

It is plausible that, if the constraints (1.4) are consistent, then large values of $\{\mu_i; i=1, 2, \ldots, m\}$ force the terms under the summation signs of expression

(5.1) to be zero at $\underset{\sim}{d}=\bar{\underset{\sim}{d}}_k$. In fact $\bar{\underset{\sim}{d}}_k=\underset{\sim}{d}_k$ if the bounds

$$\mu_i \geqslant |\lambda_i^{(k)}|, \quad i=1,2,\ldots,m, \tag{5.2}$$

are obtained, where $\{\lambda_i^{(k)};\ i=1,2,\ldots,m\}$ are still the Lagrange parameters at the solution of the quadratic programming problem that determines $\underset{\sim}{d}_k$. To prove this assertion we note that the right hand side of expression (2.8) is non-negative, and we recall that the function $\{\bar{L}_k(\underset{\sim}{x}_k+\underset{\sim}{d});\ \underset{\sim}{d}\in\mathbb{R}^n\}$ is least when $\underset{\sim}{d}=\underset{\sim}{d}_k$. Thus we deduce the condition

$$\bar{W}_k(\underset{\sim}{x}_k+\underset{\sim}{d}) \geqslant \bar{L}_k(\underset{\sim}{x}_k+\underset{\sim}{d}_k), \quad \underset{\sim}{d}\in\mathbb{R}^n. \tag{5.3}$$

Since $\bar{W}_k(\underset{\sim}{x}_k+\underset{\sim}{d}_k)=\bar{F}_k(\underset{\sim}{x}_k+\underset{\sim}{d}_k)=\bar{L}_k(\underset{\sim}{x}_k+\underset{\sim}{d}_k)$, it follows that $\underset{\sim}{d}=\underset{\sim}{d}_k$ minimizes $\bar{W}_k(\underset{\sim}{x}_k+\underset{\sim}{d})$, which completes the proof.

The search direction $\bar{\underset{\sim}{d}}_k$ has the advantage over $\underset{\sim}{d}_k$ that, if it is non-zero, it automatically satisfies the descent condition

$$W(\underset{\sim}{x}_k+\alpha\bar{\underset{\sim}{d}}_k,\underset{\sim}{\mu}) < W(\underset{\sim}{x}_k,\underset{\sim}{\mu}) \tag{5.4}$$

when α is small and positive, which is due to the bound (2.7) and the convexity of $\bar{W}_k(\underset{\sim}{x}_k+\underset{\sim}{d})$. Therefore, one may be able to keep the parameters $\{\mu_i;\ i=1,2,\ldots,m\}$ of the line search objective function smaller than the values that occur in the theory of Section 2, which sometimes reduces the number of iterations. However, small values of $\underset{\sim}{\mu}$ are not always suitable because, if $\underset{\sim}{x}^*$ is a Kuhn-Tucker point at which the gradients of the active constraints are linearly independent, and if $\underset{\sim}{\lambda}^*$ is the vector of Lagrange multipliers at $\underset{\sim}{x}^*$, then the search directions $\{\bar{\underset{\sim}{d}}_k;\ k=1,2,3,\ldots\}$ provide convergence to $\underset{\sim}{x}^*$ only if $\mu_i\geqslant|\lambda_i^*|$ for $i=1,2,\ldots,m$. Clearly $\underset{\sim}{\mu}$ should be increased if $\bar{\underset{\sim}{d}}_k=0$ and $\underset{\sim}{x}_k$ is not feasible, unless it is believed that the given constraints (1.1) are inconsistent. Also it is usually worthwhile to increase $\underset{\sim}{\mu}$ if the last two terms of expression (5.1) make a non-zero contribution to $\bar{W}_k(\underset{\sim}{x}_k+\bar{\underset{\sim}{d}}_k)$ on several consecutive iterations.

Another technique for generating search directions when the constraints (1.4) cannot be satisfied is suggested by Powell [21]. In order to describe it we let V_k be the set of indices of the inequality constraints that are violated at $\underset{\sim}{x}_k$, and we let the set S_k contain the indices of the remaining inequality constraints. The largest value of ξ in [0, 1], $\tilde{\xi}$ say, is found such that the conditions

$$\left.\begin{array}{ll} \xi c_i(\underset{\sim}{x}_k)+\underset{\sim}{d}^T\underset{\sim}{\nabla} c_i(\underset{\sim}{x}_k)=0, & i=1,2,\ldots,m' \\ \xi c_i(\underset{\sim}{x}_k)+\underset{\sim}{d}^T\underset{\sim}{\nabla} c_i(\underset{\sim}{x}_k)\geqslant 0, & i\in V_k \\ c_i(\underset{\sim}{x}_k)+\underset{\sim}{d}^T\underset{\sim}{\nabla} c_i(\underset{\sim}{x}_k)\geqslant 0, & i\in S_k \end{array}\right\} \tag{5.5}$$

hold for some $\underset{\sim}{d}\in\mathbb{R}^n$. If $\tilde{\xi}=0$, which is always allowed by $\underset{\sim}{d}=0$, or if $\tilde{\xi}$ is less than a small prescribed tolerance, then the calculation finishes because it is assumed that the constraints (1.1) have no feasible point. Otherwise the search direction is defined to be the value of $\underset{\sim}{d}$ that minimizes the quadratic function (1.3) subject to the conditions (5.5), where ξ is a constant that is chosen from the interval $(0,\tilde{\xi}]$. I used to set $\xi=\tilde{\xi}$ on all iterations, but now I prefer the value $\xi=0.9\tilde{\xi}$ when $\tilde{\xi}<1$ [24], in order that the conditions (5.5) allow some freedom in $\underset{\sim}{d}$ to reduce the objective function (1.3).

If the first line of expression (5.5) is replaced by the inequalities

$$|c_i(\underset{\sim}{x}_k) + \underset{\sim}{d}^T \underset{\sim}{\nabla} c_i(\underset{\sim}{x}_k| \leqslant (1-\xi)|c_i(\underset{\sim}{x}_k)|, \quad i=1,2,\ldots,m', \tag{5.6}$$

which is a suitable change because it allows more freedom in $\underset{\sim}{d}$ without increasing the predicted constraint violations $\{|c_i(\underset{\sim}{x}_k) + \underset{\sim}{d}^T \underset{\sim}{\nabla} c_i(\underset{\sim}{x}_k)|; i=1,2,\ldots,m'\}$, then the method that has just been described can be expressed as an unconstrained minimization calculation. The objective function of this calculation depends on the ratios

$$\left.\begin{array}{ll} r_i^{(k)}(\underset{\sim}{d}) = |c_i(\underset{\sim}{x}_k) + \underset{\sim}{d}^T \underset{\sim}{\nabla} c_i(\underset{\sim}{x}_k)| / |c_i(\underset{\sim}{x}_k)|, & i=1,2,\ldots,m' \\ r_i^{(k)}(\underset{\sim}{d}) = \dfrac{\max[0, -c_i(\underset{\sim}{x}_k) - \underset{\sim}{d}^T \underset{\sim}{\nabla} c_i(\underset{\sim}{x}_k)]}{\max[0, -c_i(\underset{\sim}{x}_k)]}, & i=m'+1,\ldots,m \end{array}\right\}, \tag{5.7}$$

where $r_i^{(k)}(\underset{\sim}{d})$ is zero if its numerator is zero, but in all other cases a zero denominator makes $r_i^{(k)}(\underset{\sim}{d})$ unbounded. We note that, if $\xi \in [0, 1]$, then $\underset{\sim}{d}$ satisfies condition (5.6) and the last two lines of expression (5.5) if the function

$$R_k(\underset{\sim}{d}) = \max_{1 \leqslant i \leqslant m} r_i^{(k)}(\underset{\sim}{d}) \tag{5.8}$$

is at most $(1-\xi)$. Conversely, if $R_k(\underset{\sim}{d}) \leqslant 1$, then the constraints on $\underset{\sim}{d}$ allow $\xi = 1 - R_k(\underset{\sim}{d})$. Hence $(1-\tilde{\xi})$ is the least possible value of $R_k(\underset{\sim}{d})$, where $\tilde{\xi}$ is still the greatest ξ in [0, 1] such that the conditions on $\underset{\sim}{d}$ hold for some $\underset{\sim}{d} \in \mathbb{R}^n$. Therefore, if $\tilde{\underset{\sim}{d}}_k$ is the vector $\underset{\sim}{d}$ that minimizes the convex objective function

$$\tilde{W}_k(\underset{\sim}{x}_k + \underset{\sim}{d}) = \bar{F}_k(\underset{\sim}{x}_k + \underset{\sim}{d}) + \tilde{\zeta} R_k(\underset{\sim}{d}), \tag{5.9}$$

where $\tilde{\zeta}$ is a large positive constant, then $R_k(\tilde{\underset{\sim}{d}}_k)$ is close to or equal to $(1-\tilde{\xi})$, and $\tilde{\underset{\sim}{d}}_k$ minimizes $\bar{F}_k(\underset{\sim}{x}_k + \underset{\sim}{d})$ subject to the constraint $R_k(\underset{\sim}{d}) \leqslant R_k(\tilde{\underset{\sim}{d}}_k)$. Thus, except for the change to the first line of expression (5.5), $\tilde{\underset{\sim}{d}}_k$ is the search direction that would be given by the method of the previous paragraph if ξ were equal to $1 - R_k(\tilde{\underset{\sim}{d}}_k)$.

However, there are advantages in preferring $\bar{W}(\underset{\sim}{x}_k + \underset{\sim}{d})$ to $\tilde{W}(\underset{\sim}{x}_k + \underset{\sim}{d})$ for calculating the search directions of a variable metric algorithm. One is that, unless $\underset{\sim}{x}_k$ is a stationary point of the line search objective function, then it is possible to reduce $W(\underset{\sim}{x}, \underset{\sim}{\mu})$ by moving from $\underset{\sim}{x}_k$ along $\underset{\sim}{d}_k$. Further, non-zero terms in the sums of expression (5.1) indicate when consideration should be given to increasing the components of $\underset{\sim}{\mu}$. On the other hand, if one takes the point of view that a search direction is not acceptable if it makes a constraint violation larger when the step length is close to zero, then, for large ζ, expression (5.9) provides an acceptable direction if any exist, but expression (5.1) may fail to do so.

Because the line search objective function $W(\underset{\sim}{x}, \underset{\sim}{\mu})$ is an L_1 penalty function for the given nonlinear programming problem, and because expression (5.9) is closely related to L_∞ penalty functions, we ask whether least squares penalty functions make a useful contribution to the calculation of search directions for variable metric algorithms. We consider this question when all the constraints are equations $(m' = m)$, so we let $\hat{\underset{\sim}{d}}_k(\hat{\zeta})$ be the value of $\underset{\sim}{d}$ that minimizes the convex quadratic function

$$\hat{W}_k(\underset{\sim}{x}_k + \underset{\sim}{d}) = \bar{F}_k(\underset{\sim}{x}_k + \underset{\sim}{d}) + \hat{\zeta} \sum_{i=1}^{m} \{c_i(\underset{\sim}{x}_k) + \underset{\sim}{d}^T \underset{\sim}{\nabla} c_i(\underset{\sim}{x}_k)\}^2, \tag{5.10}$$

where $\hat{\zeta}$ is a positive parameter. If the basic method of Section 1 has a solution $\underset{\sim}{d}_k$, then $\hat{\underset{\sim}{d}}_k(\hat{\zeta})$ tends to $\underset{\sim}{d}_k$ if $\hat{\zeta}$ is made very large, but in general, unlike the other two unconstrained minimization procedures that have been mentioned for calculating search directions, $\hat{\underset{\sim}{d}}_k(\hat{\zeta})$ is not equal to $\underset{\sim}{d}_k$ for finite parameter values. However, $\hat{\underset{\sim}{d}}_k(\hat{\zeta})$ is not a new choice of search direction, because it occurs in the very successful REQP algorithm of Bartholomew-Biggs [1], [2].

In order to prove this statement we recall that equations (11) and (12) of [2] and equations (6) and (7) of [1] state that the REQP search direction is the vector $\underset{\sim}{d}$ that minimizes the function

$$\bar{F}_k(\underset{\sim}{x}_k+\underset{\sim}{d})=F(\underset{\sim}{x}_k)+\underset{\sim}{d}^T \underset{\sim}{\nabla} F(\underset{\sim}{x}_k)+\tfrac{1}{2}\underset{\sim}{d}^T B_k \underset{\sim}{d}, \tag{5.11}$$

subject to the constraints

$$c_i(\underset{\sim}{x}_k)+\underset{\sim}{d}^T \underset{\sim}{\nabla} c_i(\underset{\sim}{x}_k)=-\tfrac{1}{2} r u_i, \quad i=1,2,\ldots,m, \tag{5.12}$$

where $r>0$, and where $\underset{\sim}{u}\in\mathbb{R}^m$ is defined by the nonsingular system of equations

$$[\tfrac{1}{2} r I+N_k^T B_k^{-1} N_k]\underset{\sim}{u}=N_k^T B_k^{-1} \underset{\sim}{\nabla} F(\underset{\sim}{x}_k)-\underset{\sim}{c}(\underset{\sim}{x}_k). \tag{5.13}$$

Here N_k is the matrix $N(\underset{\sim}{x}_k)$ of Section 3, whose columns are the gradients $\{\underset{\sim}{\nabla} c_i(\underset{\sim}{x}_k);\ i=1,2,\ldots,m\}$. We show that the REQP search direction is the value of $\underset{\sim}{d}$ that gives the least value of the function (5.10) when $\hat{\zeta}=1/r$.

It follows from expressions (5.10), (5.11) and (5.12) that $\hat{\underset{\sim}{d}}_k(\hat{\zeta})=\hat{\underset{\sim}{d}}_k$, say, is the REQP search direction if and only if it satisfies the conditions

$$c_i(\underset{\sim}{x}_k)+\hat{\underset{\sim}{d}}_k^T \underset{\sim}{\nabla} c_i(\underset{\sim}{x}_k)=-(1/2\hat{\zeta})u_i, \quad i=1,2,\ldots,m, \tag{5.14}$$

which in vector form are the equation

$$\underset{\sim}{c}(\underset{\sim}{x}_k)+N_k^T \hat{\underset{\sim}{d}}_k=-(1/2\hat{\zeta})\underset{\sim}{u}. \tag{5.15}$$

Therefore, since $\underset{\sim}{u}$ is defined by the system (5.13), we have only to verify that, if the left hand side of expression (5.15) is multiplied by the matrix $[-I-2\hat{\zeta}N_k^T B_k^{-1} N_k]$, we obtain the vector $[N_k^T B_k^{-1} \underset{\sim}{\nabla} F(\underset{\sim}{x}_k)-\underset{\sim}{c}(\underset{\sim}{x}_k)]$. Because the gradient of the function (5.10) is zero at $\underset{\sim}{d}=\hat{\underset{\sim}{d}}_k$, we have the identity

$$\underset{\sim}{\nabla} F(\underset{\sim}{x}_k)+B_k \hat{\underset{\sim}{d}}_k+2\hat{\zeta}N_k[\underset{\sim}{c}(\underset{\sim}{x}_k)+N_k^T \hat{\underset{\sim}{d}}_k]=0, \tag{5.16}$$

which implies the relation

$$-2\hat{\zeta}N_k^T B_k^{-1} N_k[\underset{\sim}{c}(\underset{\sim}{x}_k)+N_k^T \hat{\underset{\sim}{d}}_k]=N_k^T B_k^{-1} \underset{\sim}{\nabla} F(\underset{\sim}{x}_k)+N_k^T \hat{\underset{\sim}{d}}_k. \tag{5.17}$$

Thus we deduce the equation

$$-[I+2\hat{\zeta}N_k^T B_k^{-1} N_k][\underset{\sim}{c}(\underset{\sim}{x}_k)+N_k^T \hat{\underset{\sim}{d}}_k]=-\underset{\sim}{c}(\underset{\sim}{x}_k)+N_k^T B_k^{-1} \underset{\sim}{\nabla} F(\underset{\sim}{x}_k), \tag{5.18}$$

which is the required result.

The minimization of the function (5.10) gives a definition of the REQP search direction that is easier to understand than the usual definition that depends on expressions (5.11), (5.12) and (5.13). In particular, it is clear that the search direction is well defined if the constraint gradients $\{\underset{\sim}{\nabla} c_i(\underset{\sim}{x}_k);\ i=1,2,\ldots,m\}$ are linearly dependent, which, as Bartholomew-Biggs [1] points out, is "not so obvious" when his definition is employed. The main advantage of calculating the search direction from the function (5.10), instead of from expression (5.1) or (5.9), is that one only has to solve a single system of linear equations. There is the disadvantage, however, that the REQP search direction is not zero when $\underset{\sim}{x}_k$ is a Kuhn-Tucker point and $\hat{\zeta}$ is finite, unless $\underset{\sim}{\nabla} F(\underset{\sim}{x}_k)$ happens to be zero.

As well as providing search directions when linear approximations to constraints are inconsistent, the unconstrained minimization techniques of this section allow bounds on search directions to be imposed. For example, it is straightforward to minimize any of the functions (5.1), (5.9) and (5.10) subject to the conditions

$$|d_i| \leqslant h_k, \qquad i=1,2,\ldots,n, \tag{5.19}$$

on the components of $\underset{\sim}{d}$, where h_k is a positive parameter. Algorithms that include such bounds are called "trust region methods", and they have several advantages. For example, the use of trust regions can improve greatly the global convergence properties of a variable metric method for constrained optimization. For detailed information, including the description of an algorithm that chooses the parameters $\{h_k;\ k=1,2,3,\ldots\}$ automatically, the paper [13] by Fletcher is recommended.

6. Further Considerations

There are many important questions on variable metric algorithms that have not been considered so far. Some of them are mentioned briefly in this section, but the lack of attention that they receive here does not imply that they are less important than the subjects of the earlier sections.

At present the number of variables that can occur in practice is restricted mainly by the work of calculating each search direction $\underset{\sim}{d}_k$. The usual defence to this disadvantage is to point out that in many important applications the calculation of functions and gradients is so laborious that the time to solve quadratic programming subproblems is insignificant, but faster ways of choosing search directions should be investigated. For small values of n, Schittkowski's [25] extensive comparison of computer programs for constrained optimization shows that, of the methods considered, the REQP method [2] of Bartholomew-Biggs requires least computing time, while Powell's algorithm [21] uses the smallest number of function and gradient evaluations. Thus variable metric algorithms perform very well, but the present state of development is such that the gains in computer time over other methods for constrained optimization are usual only when the number of variables is small.

The reported [25] speed of the REQP algorithm is due to the use of an "active set" method for inequality constraints. An active set method provides and revises automatically a list of inequality constraints that are to be treated as equations in the calculation of each search direction, the remaining inequalities being ignored. Thus the quadratic programming subproblem of Section 1 to determine $\underset{\sim}{d}_k$ is reduced to the solution of a system of linear equations, which usually saves much computer time on each iteration. However, as indicated in the previous paragraph, the use of an active set method may increase the total number of iterations. An interesting discussion of this subject is given by Murray and Wright [20].

Active set methods have several other uses, because they are needed whenever one wishes to apply to inequality constraints a technique that is intended for the case when all constraints are equations. In particular they occur in reduced gradient algorithms for constrained optimization. If there are m independent active constraints, these algorithms allow $(n-m)$ of the components of $\underset{\sim}{x}$ to be unconstrained, and the remaining components are calculated to satisfy the active constraints.

Theorem 3 is closely related to reduced gradient algorithms when its conditions hold, which include the assumption that all constraints are equations $(m'=m)$. In order to explain this remark, we compare the "reduced second derivative matrix" of the reduced gradient algorithm at $\underset{\sim}{x}=\underset{\sim}{x}^*$ with the matrix $P_k G^* P_k$ of Theorem 3 in the limit as $k\to\infty$. The reduced second derivative matrix, $\hat{G}^*$ say, has dimensions $(n-m)\times(n-m)$, and it depends on the choice of the $(n-m)$ unconstrained variables, which we suppose are the first $(n-m)$ components of $\underset{\sim}{x}$. For any sufficiently small $\hat{\underset{\sim}{h}}\in\mathbb{R}^{n-m}$, we let $\underset{\sim}{h}\in\mathbb{R}^n$ be the vector whose first $(n-m)$ components are $\{\hat{h}_i;\, i=1,2,\ldots,n-m\}$, and whose last m components are such that the constraints are satisfied at $(\underset{\sim}{x}^*+\underset{\sim}{h})$. Then $\hat{G}^*$ is defined by the relation

$$F(\underset{\sim}{x}^*+\underset{\sim}{h})=F(\underset{\sim}{x}^*)+\tfrac{1}{2}\hat{\underset{\sim}{h}}^T\hat{G}^*\hat{\underset{\sim}{h}}+o(\|\hat{\underset{\sim}{h}}\|^2). \tag{6.1}$$

Further, because $F(\underset{\sim}{x})$ is equal to the Lagrangian function (3.15) when $\underset{\sim}{x}$ is feasible, we may also define $\hat{G}^*$ by the equation

$$L(\underset{\sim}{x}^*+\underset{\sim}{h})=L(\underset{\sim}{x}^*)+\tfrac{1}{2}\hat{\underset{\sim}{h}}^T\hat{G}^*\hat{\underset{\sim}{h}}+o(\|\hat{\underset{\sim}{h}}\|^2). \tag{6.2}$$

Since the reduced gradient algorithm seeks the unconstrained minimum of the function $\Phi(\hat{\underset{\sim}{h}})\equiv F(\underset{\sim}{x}^*+\underset{\sim}{h})$, it can achieve superlinear convergence from calculated first derivatives when a suitable approximation to $\hat{G}^*$ is available.

In order to relate $\hat{G}^*$ to G^*, we compare expression (6.2) with the equation

$$L(\underset{\sim}{x}^*+\underset{\sim}{h})=L(\underset{\sim}{x}^*)+\tfrac{1}{2}\underset{\sim}{h}^T G^*\underset{\sim}{h}+o(\|\underset{\sim}{h}\|^2), \tag{6.3}$$

and, for $i=1,2,\ldots,n-m$, we let $\underset{\sim}{h}_i$ be the vector

$$\underset{\sim}{h}_i=\lim_{\varepsilon\to 0}\underset{\sim}{h}_i(\varepsilon)/\varepsilon, \tag{6.4}$$

where $\underset{\sim}{h}_i(\varepsilon)$ is the value of $\underset{\sim}{h}$ that is defined by the method of the previous paragraph when $\hat{\underset{\sim}{h}}$ is ε times the i-th coordinate vector in $\mathbb{R}^{n-m}$. It follows that the elements of $\hat{G}^*$ are given by the equation

$$\hat{G}^*_{ij}=\underset{\sim}{h}_i^T G^*\underset{\sim}{h}_j. \tag{6.5}$$

Now, since $\underset{\sim}{x}^*$ and $[\underset{\sim}{x}^*+\underset{\sim}{h}_i(\varepsilon)]$ are both feasible, the limit (6.4) implies that $\underset{\sim}{h}_i$ is orthogonal to the constraint gradients $\{\underset{\sim}{\nabla} c_i(\underset{\sim}{x}^*);\, i=1,2,\ldots,m\}$. In other words each $\underset{\sim}{h}_i$ is in the column space of P^*, where P^* is the limit of the symmetric projection matrices $\{P_k;\ k=1,2,3,\ldots\}$ that occur in Theorem 3. Therefore equation (6.5) shows that $\hat{G}^*$ can be derived from the matrix $P^*G^*P^*$. Conversely, because the vectors $\{\underline{h}_i;\ i=1,2,\ldots,n-m\}$ span the column space of P^*, $P^*G^*P^*$ can be obtained from $\hat{G}^*$.

These remarks suggest that reduced gradient and variable metric algorithms require equivalent second derivative information to achieve superlinear convergence, but the main difference between these classes of methods is that reduced gradient procedures include an inner iteration in order to satisfy nonlinear constraints sufficiently accurately. Because of the usefulness of reduced gradient procedures for solving large structured problems, it might be very valuable to merge these two approaches to constrained optimization calculations. The papers of Coleman and Conn [7], [8] give further attention to this subject.

Another question that deserves more attention is the choice of the multipliers $\{\lambda_i^{(k)};\ i=1,2,\ldots,m\}$ in the expression (3.40) that is used to calculate B_{k+1} from B_k. As Chamberlain [4] shows, the method that is suggested in Section 3 may make the sequence of matrices $\{B_k;\ k=1,2,3,\ldots\}$ unbounded. Several estimates of Lagrange parameters are dicussed by Gill and Murray [15]. Moreover, Dixon [10] and Schittkowski [26] have investigated recently the possibility of replacing the line search objective function (1.7) by a function that has continuous first derivatives.

For information on programming considerations, when implementing a variable metric method for constrained optimization, the book by Gill, Murray and Wright [16] is recommended. It gives careful attention to suitable procedures for the matrix calculations that occur. A Fortran listing of a variable metric algorithm is available from the author [24].

It has been shown in this paper that the ideas and theory of variable metric methods are an important part of the subject of mathematical programming, and that there is no clear dividing line between these methods and several other classes of algorithms for the solution of nonlinear optimization problems.

References

[1] M. C. Bartholomew-Biggs, "Recursive quadratic programming methods for nonlinear constraints", in *Nonlinear Optimization 1981,* ed. M. J. D. Powell (Academic Press, London, 1982) pp. 213–221.

[2] M. C. Biggs, "Constrained minimization using recursive quadratic programming: some alternative subproblem formulations", in *Towards Global Optimization,* eds. L. C. W. Dixon and G. P. Szegö (North Holland Publishing Co., Amsterdam, 1975) pp. 341–349.

[3] P. T. Boggs, J. W. Tolle and P. Wang, "On the local convergence of quasi-Newton methods for constrained optimization", *SIAM Journal on Control and Optimization,* Vol. 20 (1982) pp. 161–171.

[4] R. M. Chamberlain, "Some examples of cycling in variable metric algorithms for

constrained optimization", *Mathematical Programming,* Vol. 16 (1979) pp. 378–383.

[5] R. M. Chamberlain, *The Theory and Application of Variable Metric Methods to Constrained Optimization Problems,* Ph. D. Dissertation (University of Cambridge, 1980).

[6] R. M. Chamberlain, C. Lemaréchal, H. C. Pedersen and M. J. D. Powell, "The watchdog technique for forcing convergence in algorithms for constrained optimization", *Mathematical Programming Study,* Vol. 16 (1982) pp. 1–17.

[7] T. F. Coleman and A. R. Conn, "Nonlinear programming via an exact penalty function: asymptotic analysis", Report CS-80-30 (University of Waterloo, 1980).

[8] T. F. Coleman and A. R. Conn, "Nonlinear programming via an exact penalty function: global analysis", Report CS-80-31 (University of Waterloo, 1980).

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

[10] L. C. W. Dixon, "On the convergence properties of variable metric recursive quadratic programming methods", Report NOC-110 (The Hatfield Polytechnic, 1980).

[11] R. Fletcher, *Practical Methods of Optimization, Vol. 1, Unconstrained Optimization* (John Wiley & Sons, Chichester, 1980).

[12] R. Fletcher, *Practical Methods of Optimization, Vol. 2, Constrained Optimization* (John Wiley & Sons, Chichester, 1981).

[13] R. Fletcher, "Numerical experiments with an exact L_1 penalty function method", in *Nonlinear Programming 4,* eds. O. L. Mangasarian, R. R. Meyer and S. M. Robinson (Academic Press, New York, 1981) pp. 99–129.

[14] R. Fletcher, "Methods for nonlinear constraints", in *Nonlinear Optimization 1981,* ed. M. J. D. Powell (Academic Press, London, 1982) pp. 185–211.

[15] P. E. Gill and W. Murray, "The computation of Lagrange multiplier estimates for constrained minimization", *Mathematical Programming,* Vol. 17 (1979) pp. 32–60.

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

[17] S. P. Han, "A globally convergent method for nonlinear programming", *Journal of Optimization Theory and Applications,* Vol. 22 (1977) pp. 297–309.

[18] N. Maratos, *Exact Penalty Function Algorithms for Finite Dimensional and Control Optimization Problems,* Ph. D. Dissertation (Imperial College, University of London, 1978).

[19] D. Q. Mayne and E. Polak, "A superlinearly convergent algorithm for constrained optimization problems", *Mathematical Programming Study,* Vol. 16 (1982) pp. 45–61.

[20] W. Murray and M. H. Wright, "Computation of the search direction in constrained optimization algorithms", *Mathematical Programming Study,* Vol. 16 (1982) pp. 62–83.

[21] M. J. D. Powell, "A fast algorithm for nonlinearly constrained optimization algorithms", in *Numerical Analysis, Dundee, 1977, Lecture Notes in Mathematics 630,* ed. G. A. Watson (Springer-Verlag, Berlin, 1978) pp. 144–157.

[22] M. J. D. Powell, "The convergence of variable metric methods for nonlinearly constrained optimization calculations", in *Nonlinear Programming 3,* eds. O. L. Mangasarian, R. R. Meyer and S. M. Robinson (Academic Press, New York, 1978) pp. 27–63.

[23] M. J. D. Powell, "Extensions to subroutine VFO2AD", in *System Modeling and Optimization, Lecture Notes in Control and Information Sciences 38,* eds. R. F. Drenick and F. Kozin (Springer-Verlag, New York, 1982) pp. 529–538.

[24] M. J. D. Powell, "VMCWD: a Fortran subroutine for constrained optimization", Report DAMTP-1982/NA4 (University of Cambridge, 1982).

[25] K. Schittkowski, *Nonlinear Programming Codes, Lecture Notes in Economics and Mathematical Systems 183* (Springer-Verlag, Berlin, 1980).

[26] K. Schittkowski, "On the convergence of a sequential quadratic programming method with an augmented Lagrangian line search function", Report SOL-82-4 (Stanford University, 1982).