J. Inst. Maths Applics (1971) **7**, 21–36

On the Convergence of the Variable Metric Algorithm

M. J. D. POWELL

Theoretical Physics Division, U.K.A.E.A. Research Group,
Atomic Energy Research Establishment, Harwell

[Received 10 November 1969]

The variable metric algorithm is a frequently used method for calculating the least value of a function of several variables. However it has been *proved* only that the method is successful if the objective function is quadratic, although in practice it treats many types of objective functions successfully. This paper extends the theory, for it proves that successful convergence is obtained provided that the objective function has a strictly positive definite second derivative matrix for all values of its variables. Moreover it is shown that the rate of convergence is super-linear.

1. The Variable Metric Algorithm

THE VARIABLE METRIC ALGORITHM was invented and published by W. C. Davidon (1959), and now it is one of the most frequently used and most successful techniques for calculating the least value of a differentiable function of several real variables, $F(x_1, x_2, \ldots x_n)$ say. Therefore, during the last ten years, some attempts have been made to prove the convergence of the method, but this is not easy because the algorithm is intended to be applied to general differentiable functions. This paper proves convergence in the case that $F(x_1, x_2, \ldots x_n)$ has continuous second derivatives, and satisfies a strict convexity condition. Previously convergence had been proved only in the very special case when the objective function is exactly quadratic (Davidon, 1959). Of course we would prefer to have theorems that apply to non-convex functions, but these would be complicated by the possibility of local minima of the objective function. Moreover I believe that in this case it can happen that the variable metric algorithm fails to find any local minimum of $F(x_1, x_2, \ldots x_n)$, although I have not yet discovered any specific examples.

We use a very basic definition of the variable metric algorithm. The original description (Davidon, 1959) included many empirical devices, so a simpler description was published by Fletcher & Powell (1963). For our theorems the definition is even more basic, because we suppose that all calculations can be worked out exactly.

The algorithm is iterative, and it requires the evaluation of certain values of $F(x_1, x_2, \ldots x_n)$ and its first derivative vector. To begin the kth iteration we require a vector of variables, $\mathbf{x}^{(k)}$, the gradient vector at this point, $\mathbf{g}^{(k)}$, and a positive definite symmetric $n \times n$ matrix $H^{(k)}$. Before the first iteration the vector $\mathbf{x}^{(1)}$ has to be set by the user of the algorithm, and the matrix $H^{(1)}$ is set to some positive definite matrix, for example the unit matrix.

The kth iteration begins by testing the gradient vector $\mathbf{g}^{(k)}$. If it is zero the algorithm finishes, and otherwise the search direction

$$\mathbf{d}^{(k)} = -H^{(k)}\mathbf{g}^{(k)} \tag{1}$$

is calculated; note that it cannot be the zero vector. Next the function of one variable

$$\phi(\lambda) \equiv F(\mathbf{x}^{(k)}+\lambda\mathbf{d}^{(k)}) \tag{2}$$

is considered, and the value of λ, $\lambda^{(k)}$ say, that gives the least value of $\phi(\lambda)$ is found. Sometimes $\lambda^{(k)}$ does not exist, because $\phi'(\lambda) < 0$ for all $\lambda \geqslant 0$, but we shall see that our conditions on $F(\mathbf{x})$ imply that each value of $\lambda^{(k)}$ is well-defined. The vector of variables that is to be used by the next iteration is set to the value

$$\begin{aligned}\mathbf{x}^{(k+1)} &= \mathbf{x}^{(k)}+\lambda^{(k)}\mathbf{d}^{(k)}\\ &= \mathbf{x}^{(k)}+\boldsymbol{\delta}^{(k)}\end{aligned} \tag{3}$$

say. Finally the kth iteration calculates $H^{(k+1)}$ by applying the formula

$$H^{(k+1)} = H^{(k)}-\frac{H^{(k)}\boldsymbol{\gamma}^{(k)}\boldsymbol{\gamma}^{(k)T}H^{(k)}}{(\boldsymbol{\gamma}^{(k)T}H^{(k)}\boldsymbol{\gamma}^{(k)})}+\frac{\boldsymbol{\delta}^{(k)}\boldsymbol{\delta}^{(k)T}}{(\boldsymbol{\delta}^{(k)T}\boldsymbol{\gamma}^{(k)})}, \tag{4}$$

where $\boldsymbol{\gamma}^{(k)}$ is the difference

$$\boldsymbol{\gamma}^{(k)} = \mathbf{g}^{(k+1)}-\mathbf{g}^{(k)}, \tag{5}$$

and where the superscript "T" denotes the transpose of a column vector.

It is straightforward to show that formula (4) never requires a division by zero. First we note that the definition of $\lambda^{(k)}$ provides the equations

$$\mathbf{g}^{(k+1)T}\mathbf{d}^{(k)} = \mathbf{g}^{(k+1)T}\boldsymbol{\delta}^{(k)} = 0. \tag{6}$$

Therefore the first denominator of formula (4) is the expression

$$\begin{aligned}\boldsymbol{\gamma}^{(k)T}H^{(k)}\boldsymbol{\gamma}^{(k)} &= \mathbf{g}^{(k+1)T}H^{(k)}\mathbf{g}^{(k+1)}+\mathbf{g}^{(k)T}H^{(k)}\mathbf{g}^{(k)}+2\mathbf{g}^{(k+1)T}\mathbf{d}^{(k)}\\ &= \mathbf{g}^{(k+1)T}H^{(k)}\mathbf{g}^{(k+1)}+\mathbf{g}^{(k)T}H^{(k)}\mathbf{g}^{(k)}\\ &> 0,\end{aligned} \tag{7}$$

because of the positive definiteness of $H^{(k)}$. Secondly we note that the inequality

$$\begin{aligned}\phi'(0) &= \mathbf{g}^{(k)T}\mathbf{d}^{(k)}\\ &= -\mathbf{g}^{(k)T}H^{(k)}\mathbf{g}^{(k)}\\ &< 0\end{aligned} \tag{8}$$

implies that $\lambda^{(k)} \neq 0$. Therefore, using equation (6) again, we find that the second denominator is the expression

$$\begin{aligned}\boldsymbol{\delta}^{(k)T}\boldsymbol{\gamma}^{(k)} &= -\boldsymbol{\delta}^{(k)T}\mathbf{g}^{(k)}\\ &= \lambda^{(k)}\mathbf{g}^{(k)T}H^{(k)}\mathbf{g}^{(k)}\\ &\neq 0.\end{aligned} \tag{9}$$

Therefore the calculation of $H^{(k+1)}$ from formula (4) is straightforward. Fletcher & Powell (1963) have proved that, because $\boldsymbol{\delta}^{(k)} \neq \mathbf{0}$, this formula has the property that if $H^{(k)}$ is positive definite, then $H^{(k+1)}$ is also positive definite.

In the case that $F(\mathbf{x})$ is a positive definite quadratic function, it has been proved that $\mathbf{g}^{(k)} = 0$ (Davidon, 1959) for some $k \leqslant n+1$, and of course the condition $\mathbf{g}^{(k)} = \mathbf{0}$ implies that $\mathbf{x}^{(k)}$ is the position of the minimum of $F(\mathbf{x})$. However, when $F(\mathbf{x})$ is not quadratic, the condition $\mathbf{g}^{(k)} = \mathbf{0}$ is usually not obtained, in which case the variable metric algorithm generates an infinite sequence of points $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \mathbf{x}^{(3)}, \ldots$ Many numerical experiments show that this sequence converges very quickly to the

position of the minimum of $F(\mathbf{x})$, and now we make this statement firm in a quite general case when the second derivative matrix of $F(\mathbf{x})$ is not constant.

Goldfarb (1969) and McCormick & Pearson (1969) have given conditions under which our main convergence theorems would be valid, but in fact the last condition of the second theorem of McCormick & Pearson is not generally satisfied by the variable metric algorithm. Goldfarb shows that convergence can be proved if the eigenvalues of the matrices $H^{(k)}$, $k = 1, 2 \ldots$ (see equation (4)) are bounded away from zero and are bounded above, but an interesting feature of our paper is that it proves convergence of the variable metric algorithm *before* proving that the matrices $H^{(k)}$ satisfy the conditions stated by Goldfarb. The present ordering of the theorems is not a deliberate method of differing from published work, but it is due to the fact that the given theorems are difficult to establish, and so I have not yet found any other way to prove them.

2. The conditions on $F(\mathbf{x})$

This paper proves the convergence of the variable metric algorithm in the case that $F(\mathbf{x})$ has continuous second derivatives, that have the property that there exists a positive constant ε such that, for all $\mathbf{x}$, the eigenvalues of the second derivative matrix of $F(\mathbf{x})$ at $\mathbf{x}$ are not less than ε. In this section we deduce a number of fundamental lemmas from these conditions that are needed to prove the convergence theorems.

LEMMA 1. *For each vector of variables* $\mathbf{y}$, *the set of vectors* $\mathbf{x}$ *that satisfy the condition* $F(\mathbf{x}) \leqslant F(\mathbf{y})$ *is closed, convex and bounded.*

Proof of Lemma 1. The closure of the set is a consequence of the continuity of the function $F(\mathbf{x})$, and the convexity of the set follows from the fact that $F(\mathbf{x})$ is a convex function. Therefore we just have to prove that the set is bounded.

We let $\mathbf{d}$ be any direction through $\mathbf{y}$ that is normalized, $\|\mathbf{d}\| = 1$. Here, and throughout this paper, the vector norm is Euclidean. We consider the function of one variable

$$\psi(\lambda) \equiv F(\mathbf{y}+\lambda\mathbf{d}). \tag{10}$$

From the conditions on $F(\mathbf{x})$ and the normalization of $\mathbf{d}$, we deduce the inequality $\psi''(\lambda) \geqslant \varepsilon$, from which it follows that the function

$$r(\lambda) \equiv \psi(\lambda)-\psi(0)-\lambda\mathbf{d}^T\mathbf{g}(\mathbf{y})-\tfrac{1}{2}\lambda^2\varepsilon \tag{11}$$

is convex, where $\mathbf{g}(\mathbf{y})$ is the gradient of $F(\mathbf{x})$ at $\mathbf{y}$. Now $r(\lambda)$ is chosen to satisfy $r(0) = r'(0) = 0$, so we deduce the inequality

$$\psi(\lambda) \geqslant \psi(0)+\lambda\mathbf{d}^T\mathbf{g}(\mathbf{y})+\tfrac{1}{2}\lambda^2\varepsilon. \tag{12}$$

But the right-hand side of this inequality exceeds $\psi(0)$ if

$$|\lambda| > 2\|\mathbf{g}(\mathbf{y})\|/\varepsilon \geqslant 2|\mathbf{d}^T\mathbf{g}(\mathbf{y})|/\varepsilon. \tag{13}$$

Therefore, because the direction of $\mathbf{d}$ is arbitrary, $F(\mathbf{x})$ exceeds $F(\mathbf{y})$ if $\|\mathbf{x}-\mathbf{y}\| > 2\|\mathbf{g}(\mathbf{y})\|/\varepsilon$. Therefore the set of points $\mathbf{x}$ satisfying the condition $F(\mathbf{x}) \leqslant F(\mathbf{y})$ is bounded. Lemma 1 is proved.

An important corollary of the lemma and the fact that the function $F(\mathbf{x})$ is continuous is that the least value of $F(\mathbf{x})$ is attained at some finite point, $\boldsymbol{\xi}$ say. Moreover the least value of $F(\mathbf{x})$ is attained at only one point, because in the last paragraph we remarked that $F(\mathbf{x})$ exceeds $F(\boldsymbol{\xi})$ if $\|\mathbf{x}-\boldsymbol{\xi}\| > 2\|\mathbf{g}(\boldsymbol{\xi})\|/\varepsilon$, and the right-hand side of this inequality is zero because $\mathbf{g}(\boldsymbol{\xi}) = \mathbf{0}$.

Note that the lemma also implies that the parameter $\lambda^{(k)}$, used in equation (3), is finite, for $\mathbf{x}^{(k+1)}$ is calculated so that the inequality $F(\mathbf{x}^{(k+1)}) < F(\mathbf{x}^{(k)})$ is satisfied.

LEMMA 2. *If* $\mathbf{y}$ *and* $\mathbf{z}$ *are any two members of the set of vectors* $\mathbf{x}$ *satisfying the condition* $F(\mathbf{x}) \leqslant F(\mathbf{x}^{(1)})$, *then the ratios* $\|\boldsymbol{\delta}\|/\|\boldsymbol{\gamma}\|$, $\|\boldsymbol{\gamma}\|/\|\boldsymbol{\delta}\|$, $(\boldsymbol{\delta}^T\boldsymbol{\gamma})/\|\boldsymbol{\delta}\|^2$, $\|\boldsymbol{\delta}\|^2/(\boldsymbol{\delta}^T\boldsymbol{\gamma})$, $(\boldsymbol{\delta}^T\boldsymbol{\gamma})/\|\boldsymbol{\gamma}\|^2$ and $\|\boldsymbol{\gamma}\|^2/(\boldsymbol{\delta}^T\boldsymbol{\gamma})$ *are all bounded, where* $\boldsymbol{\delta}$ *and* $\boldsymbol{\gamma}$ *are the vectors*

$$\left.\begin{aligned}\boldsymbol{\delta} &= \mathbf{z}-\mathbf{y}\\ \boldsymbol{\gamma} &= \mathbf{g}(\mathbf{z})-\mathbf{g}(\mathbf{y})\end{aligned}\right\}, \tag{14}$$

$\mathbf{g}(\mathbf{x})$ *being the gradient of* $F(\mathbf{x})$ *at* $\mathbf{x}$.

Proof of Lemma 2. Because of the Schwarz inequality, $(\boldsymbol{\delta}^T\boldsymbol{\gamma}) \leqslant \|\boldsymbol{\delta}\|\,\|\boldsymbol{\gamma}\|$, we see that it is sufficient to prove only that the ratios $\|\boldsymbol{\gamma}\|/\|\boldsymbol{\delta}\|$ and $\|\boldsymbol{\delta}\|^2/(\boldsymbol{\delta}^T\boldsymbol{\gamma})$ are finite.

To bound the ratio $\|\boldsymbol{\gamma}\|/\|\boldsymbol{\delta}\|$, we let $G(\mathbf{x})$ be the second derivative matrix of $F(\mathbf{x})$ at $\mathbf{x}$, and note that direct differentiation gives the equation

$$\frac{d}{d\theta}[\mathbf{g}(\mathbf{y}+\theta\boldsymbol{\delta})] = G(\mathbf{y}+\theta\boldsymbol{\delta})\boldsymbol{\delta}, \tag{15}$$

whence we deduce the identity

$$\boldsymbol{\gamma} = \int_0^1 G(\mathbf{y}+\theta\boldsymbol{\delta})\boldsymbol{\delta}\, d\theta. \tag{16}$$

Therefore a corollary of the triangle inequality for norms gives the bound

$$\begin{aligned}\|\boldsymbol{\gamma}\| &\leqslant \int_0^1 \|G(\mathbf{y}+\theta\boldsymbol{\delta})\boldsymbol{\delta}\|\, d\theta\\ &\leqslant \|\boldsymbol{\delta}\| \int_0^1 \|G(\mathbf{y}+\theta\boldsymbol{\delta})\|\, d\theta\\ &\leqslant \|\boldsymbol{\delta}\| \max_{0\leqslant\theta\leqslant 1} \|G(\mathbf{y}+\theta\boldsymbol{\delta})\|,\end{aligned} \tag{17}$$

where $\|G\|$ denotes the Euclidean norm of the matrix G. Now Lemma 1 states that the set of points $\mathbf{x}$ satisfying the condition $F(\mathbf{x}) \leqslant F(\mathbf{x}^{(1)})$ is compact, and we have imposed the condition that the elements of $G(\mathbf{x})$ are continuous functions of $\mathbf{x}$. Therefore the number

$$E(\mathbf{x}^{(1)}) = \max_{x} \|G(\mathbf{x})\|,\ F(\mathbf{x}) \leqslant F(\mathbf{x}^{(1)}), \tag{18}$$

is finite. Therefore, because of the convexity of the set $\{\mathbf{x};\ F(\mathbf{x}) \leqslant F(\mathbf{x}^{(1)})\}$, we deduce from inequalities (17) and (18) the bound

$$\|\boldsymbol{\gamma}\| \leqslant E(\mathbf{x}^{(1)})\|\boldsymbol{\delta}\|. \tag{19}$$

To establish the other bound, we integrate the identity

$$\frac{d}{d\theta}[\boldsymbol{\delta}^T\mathbf{g}(\mathbf{y}+\theta\boldsymbol{\delta})] = \boldsymbol{\delta}^T G(\mathbf{y}+\theta\boldsymbol{\delta})\boldsymbol{\delta}, \tag{20}$$

and obtain the equation

$$\int_0^1 \boldsymbol{\delta}^T G(\mathbf{y}+\theta\boldsymbol{\delta})\boldsymbol{\delta}\, d\theta = \boldsymbol{\delta}^T\boldsymbol{\gamma}. \tag{21}$$

Now the conditions on the second derivatives of $F(\mathbf{x})$ imply that this integrand is not less than $\varepsilon\|\boldsymbol{\delta}\|^2$, so we deduce the inequality,

$$\|\boldsymbol{\delta}\|^2 \leqslant (\boldsymbol{\delta}^T\boldsymbol{\gamma})/\varepsilon. \tag{22}$$

Lemma 2 is proved.

LEMMA 3. *For all vectors* $\mathbf{x}$, *the inequality*

$$\|\mathbf{g}(\mathbf{x})\|^2 \geqslant \varepsilon\{F(\mathbf{x})-F(\xi)\} \tag{23}$$

holds, where $F(\xi)$ *is the least value of* $F(\mathbf{x})$.

Proof of Lemma 3. Because the function $\psi(\theta) = F(\mathbf{x}+\theta\{\xi-\mathbf{x}\})$ is convex, the inequality

$$F(\mathbf{x}+\theta\{\xi-\mathbf{x}\}) \geqslant F(\mathbf{x})+\theta(\xi-\mathbf{x})^T\mathbf{g}(\mathbf{x}) \tag{24}$$

is satisfied. In particular the value $\theta = 1$ gives the condition

$$\begin{aligned} F(\mathbf{x})-F(\xi) &\leqslant -(\xi-\mathbf{x})^T\mathbf{g}(\mathbf{x}) \\ &\leqslant \|\mathbf{g}(\mathbf{x})\|\,\|\xi-\mathbf{x}\|. \end{aligned} \tag{25}$$

Now expression (22) and the Schwarz inequality provide the bound

$$\|\xi-\mathbf{x}\| \leqslant \|\mathbf{g}(\xi)-\mathbf{g}(\mathbf{x})\|/\varepsilon, \tag{26}$$

and $\mathbf{g}(\xi) = \mathbf{0}$. Therefore Lemma 3 is a consequence of inequality (25).

LEMMA 4. *The sequence of points generated by the variable metric algorithm has the property that the sums* $\Sigma\|\delta^{(k)}\|^2$ *and* $\Sigma\|\gamma^{(k)}\|^2$ *are convergent, where* $\delta^{(k)}$ *and* $\gamma^{(k)}$ *are defined by equations* (3) *and* (5).

Proof of Lemma 4. We let $\psi(\theta) = F(\mathbf{x}^{(k+1)}-\theta\delta^{(k)})$, and note that the conditions on the derivatives of $F(\mathbf{x})$ imply the inequality $\psi''(\theta) \geqslant \varepsilon\|\delta^{(k)}\|^2$. Therefore, because equation (6) shows that $\psi'(0) = 0$, we deduce the relation

$$\psi(\theta) \geqslant \psi(0)+\tfrac{1}{2}\varepsilon\|\delta^{(k)}\|^2\theta^2. \tag{27}$$

Substituting $\theta = 1$ gives the result

$$F(\mathbf{x}^{(k)})-F(\mathbf{x}^{(k+1)}) \geqslant \tfrac{1}{2}\varepsilon\|\delta^{(k)}\|^2, \tag{28}$$

so by summing over k we obtain the inequality

$$\sum_{k=1}^{\infty}\|\delta^{(k)}\|^2 \leqslant 2\{F(\mathbf{x}^{(1)})-F(\xi)\}/\varepsilon, \tag{29}$$

where $F(\xi)$ is the least value of $F(\mathbf{x})$. Therefore $\Sigma\|\delta^{(k)}\|^2$ is convergent, and, because of Lemma 2, $\Sigma\|\gamma^{(k)}\|^2$ is also convergent. The lemma is proved.

3. The Convergence Theorems

THEOREM 1. *If the objective function* $F(\mathbf{x})$ *has continuous second derivatives, and if there exists a positive constant* ε *that is not greater than the least eigenvalue of any second derivative matrix of* $F(\mathbf{x})$, *then the sequence of points* $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \mathbf{x}^{(3)}, \ldots$, *generated by the variable metric algorithm, converges to* ξ, *which is the position of the least value of* $F(\mathbf{x})$.

Proof of Theorem 1. The proof of the theorem is quite long, because we may not assume that the matrices $H^{(k)}$ (see equation (4)) are uniformly bounded, and we may not assume that their least eigenvalues are bounded away from zero. The method of our proof is to define $\Gamma^{(k)}$ to be the matrix $[H^{(k)}]^{-1}$, to work out an expression for the trace of $\Gamma^{(k)}$, and to show that this expression implies a contradiction unless the sequence of gradients $\mathbf{g}^{(k)}$, $k = 1, 2, 3, \ldots$, tends to zero.

The expression that we require for the trace of $\Gamma^{(k)}$ is derived by a considerable amount of algebra, so it is convenient to simplify our notation. From this point on

we may omit the iteration number superscripts, and we will distinguish between (k) and $(k+1)$ by using the superscript "*" in the second case. For example in place of equation (4) we may write

$$H^* = H - \frac{H\gamma\gamma^T H}{(\gamma^T H\gamma)} + \frac{\delta\delta^T}{(\delta^T\gamma)}. \tag{30}$$

First we note that equation (30) implies that the inverse of the matrix H^* is the matrix

$$\Gamma^* = \left(I - \frac{\gamma\delta^T}{(\delta^T\gamma)}\right)\Gamma\left(I - \frac{\delta\gamma^T}{(\delta^T\gamma)}\right) + \frac{\gamma\gamma^T}{(\delta^T\gamma)}. \tag{31}$$

We do not provide the details that justify this statement, because it is easy to verify that the product $H^*\Gamma^*$ is equal to the unit matrix.

Next we express the trace of Γ^* in terms of the trace of Γ, and obtain from expression (31) the identity

$$\text{Tr}\,(\Gamma^*) = \text{Tr}\,(\Gamma) - 2\frac{(\delta^T\Gamma\gamma)}{(\delta^T\gamma)} + \frac{(\delta^T\delta\Gamma)(\gamma^T\gamma)}{(\delta^T\gamma)^2} + \frac{(\gamma^T\gamma)}{(\delta^T\gamma)}. \tag{32}$$

The middle two terms of expression (32) simplify in the following way, which uses equations (1), (3), (5), and (6)

$$\begin{aligned} -2\frac{(\delta^T\Gamma\gamma)}{(\delta^T\gamma)} + \frac{(\delta^T\Gamma\delta)(\gamma^T\gamma)}{(\delta^T\gamma)^2} &= \lambda^{(k)}\left[\frac{2(\mathbf{g}^T\gamma)}{(\delta^T\gamma)} + \frac{(-\mathbf{g}^T\delta)(\gamma^T\gamma)}{(\delta^T\gamma)^2}\right] \\ &= \lambda^{(k)}\frac{2(\mathbf{g}^T\gamma) + (\gamma^T\gamma)}{(\delta^T\gamma)} \\ &= \frac{\|\mathbf{g}^*\|^2 - \|\mathbf{g}\|^2}{(\mathbf{g}^T H\mathbf{g})}. \end{aligned} \tag{33}$$

Now the following identity is valid

$$\frac{1}{(\mathbf{g}^{*T}H^*\mathbf{g}^*)} = \frac{1}{(\mathbf{g}^{*T}H\mathbf{g}^*)} + \frac{1}{(\mathbf{g}^T H\mathbf{g})}. \tag{34}$$

To prove it we use equations (4), (5) and (6) to obtain the relation

$$\begin{aligned} \mathbf{g}^{*T}H^*\mathbf{g}^* &= \mathbf{g}^{*T}\left[H - \frac{H\gamma\gamma^T H}{(\gamma^T H\gamma)}\right]\mathbf{g}^* \\ &= \mathbf{g}^T\left[H - \frac{H\gamma\gamma^T H}{(\gamma^T H\gamma)}\right]\mathbf{g}. \end{aligned} \tag{35}$$

Then we replace γ by $(\mathbf{g}^* - \mathbf{g})$, and use equations (1), (6) and (7) to deduce the identity

$$\begin{aligned} \mathbf{g}^{*T}H^*\mathbf{g}^* &= \mathbf{g}^T\left[H - \frac{H\mathbf{g}\mathbf{g}^T H}{(\mathbf{g}^T H\mathbf{g})}\right]\mathbf{g} \\ &= \frac{(\mathbf{g}^T H\mathbf{g})(\mathbf{g}^{*T}H\mathbf{g}^*)}{(\mathbf{g}^T H\mathbf{g}) + (\mathbf{g}^{*T}H\mathbf{g}^*)}. \end{aligned} \tag{36}$$

Equation (34) is obtained by inverting the two sides of equation (36). Therefore equation (32) is equivalent to the relation

$$\text{Tr}\,(\Gamma^*) = \text{Tr}\,(\Gamma) + \frac{\|\mathbf{g}^*\|^2}{(\mathbf{g}^{*T}H^*\mathbf{g}^*)} - \frac{\|\mathbf{g}\|^2}{(\mathbf{g}^T H\mathbf{g})} - \frac{\|\mathbf{g}^*\|^2}{(\mathbf{g}^{*T}H\mathbf{g}^*)} + \frac{(\gamma^T\gamma)}{(\delta^T\gamma)}. \tag{37}$$

We now use equation (37) to express $\mathrm{Tr}\,(\Gamma^{(j+1)})$ in terms of $\mathrm{Tr}\,(\Gamma^{(j)})$ for $j = 1, 2, \ldots k$. Thus we deduce the equation

$$\mathrm{Tr}\,(\Gamma^{(k+1)}) = \mathrm{Tr}\,(\Gamma^{(1)}) + \frac{\|\mathbf{g}^{(k+1)}\|^2}{(\mathbf{g}^{(k+1)T}H^{(k+1)}\mathbf{g}^{(k+1)})} - \frac{\|\mathbf{g}^{(1)}\|^2}{(\mathbf{g}^{(1)T}H^{(1)}\mathbf{g}^{(1)})} - \\ - \sum_{j=1}^{k} \frac{\|\mathbf{g}^{(j+1)}\|^2}{(\mathbf{g}^{(j+1)T}H^{(j)}\mathbf{g}^{(j+1)})} + \sum_{j=1}^{k} \frac{\|\boldsymbol{\gamma}^{(j)}\|^2}{(\boldsymbol{\delta}^{(j)T}\boldsymbol{\gamma}^{(j)})}. \quad (38)$$

Therefore because of Lemma 2 there exists a number B, depending on $\mathbf{x}^{(1)}$ but not on k, such that the inequality

$$\mathrm{Tr}\,(\Gamma^{(k+1)}) \leqslant \frac{\|\mathbf{g}^{(k+1)}\|^2}{(\mathbf{g}^{(k+1)T}H^{(k+1)}\mathbf{g}^{(k+1)})} - \sum_{j=1}^{k} \frac{\|\mathbf{g}^{(j+1)}\|^2}{(\mathbf{g}^{(j+1)T}H^{(j)}\mathbf{g}^{(j+1)})} + Bk \quad (39)$$

holds.

The greater part of the remainder of the proof shows that, if the theorem is not true, then the sum of the last two terms of the inequality (39) is negative, because most numbers in the sequence $(\mathbf{g}^{(j+1)T}H^{(j)}\mathbf{g}^{(j+1)})$, $j = 1, 2, 3 \ldots$, tend to zero.

To show that most of the numbers $(\mathbf{g}^{(j+1)T}H^{(j)}\mathbf{g}^{(j+1)})$ are very small, we consider the trace of the matrix $H^{(k+1)}$. From equation (4) this trace is the expression

$$\mathrm{Tr}\,(H^{(k+1)}) = \mathrm{Tr}\,(H^{(1)}) - \sum_{j=1}^{k} \frac{\|H^{(j)}\boldsymbol{\gamma}^{(j)}\|^2}{(\boldsymbol{\gamma}^{(j)T}H^{(j)}\boldsymbol{\gamma}^{(j)})} + \sum_{j=1}^{k} \frac{\|\boldsymbol{\delta}^{(j)}\|^2}{(\boldsymbol{\delta}^{(j)T}\boldsymbol{\gamma}^{(j)})}. \quad (40)$$

Because the matrix $H^{(k+1)}$ is positive definite, the right-hand side of this equation is positive. Therefore, because of Lemma 2, there exists a number M, independent of k, such that the inequality

$$\sum_{j=1}^{k} \frac{\|H^{(j)}\boldsymbol{\gamma}^{(j)}\|^2}{(\boldsymbol{\gamma}^{(j)T}H^{(j)}\boldsymbol{\gamma}^{(j)})} < Mk \quad (41)$$

holds. Now the Schwarz inequality provides the relation

$$(\boldsymbol{\gamma}^{(j)T}H^{(j)}\boldsymbol{\gamma}^{(j)})^2 \leqslant \|H^{(j)}\boldsymbol{\gamma}^{(j)}\|^2\|\boldsymbol{\gamma}^{(j)}\|^2, \quad (42)$$

and equation (7) provides the inequality $(\boldsymbol{\gamma}^{(j)T}H^{(j)}\boldsymbol{\gamma}^{(j)}) > (\mathbf{g}^{(j+1)T}H^{(j)}\mathbf{g}^{(j+1)})$. Therefore from expression (41) we deduce the relation

$$\sum_{j=1}^{k} \frac{(\mathbf{g}^{(j+1)T}H^{(j)}\mathbf{g}^{(j+1)})}{\|\boldsymbol{\gamma}^{(j)}\|^2} < Mk. \quad (43)$$

Therefore at least two thirds of the terms that are summed must satisfy the inequality

$$(\mathbf{g}^{(j+1)T}H^{(j)}\mathbf{g}^{(j+1)}) < 3M\|\boldsymbol{\gamma}^{(j)}\|^2, \quad (44)$$

for otherwise the left-hand side of expression (43) would exceed the right-hand side. Now Lemma 4 states that $\Sigma\|\boldsymbol{\gamma}^{(j)}\|^2$ is finite, so at least two thirds of the terms $(\mathbf{g}^{(j+1)T}H^{(j)}\mathbf{g}^{(j+1)})$, $j = 1, 2, 3 \ldots$, form a sequence that converges to zero.

We now consider the numerators of the terms that are summed in expression (39). If Theorem 1 is false, then the limit of the monotonically decreasing sequence $F(\mathbf{x}^{(k)})$, $k = 1, 2, 3 \ldots$, exceeds $F(\xi)$. Therefore Lemma 3 shows that the numbers $\|\mathbf{g}^{(j+1)}\|^2$, $j = 1, 2, 3 \ldots$, are bounded away from zero; say the inequality

$$\|\mathbf{g}^{(j+1)}\|^2 \geqslant c > 0 \quad (45)$$

holds. We let K be an integer such that the right-hand side of expression (44) is less than $c/3B$ for all $j \geqslant K$.

Now if we substitute $k = 3K$ in expression (39), we find that at least K terms of the summation exceed the value $3B$, so the inequality

$$\mathrm{Tr}\,(\Gamma^{(k+1)}) < \frac{\|\mathbf{g}^{(k+1)}\|^2}{(\mathbf{g}^{(k+1)T}H^{(k+1)}\mathbf{g}^{(k+1)})} \tag{46}$$

is satisfied for at least one value of k. This inequality was deduced from the assumption that Theorem 1 is false, so we complete the proof of the theorem by showing that the inequality is a contradiction.

The trace of a symmetric matrix is the sum of its eigenvalues, so the trace of a positive definite matrix is an upper bound on the greatest eigenvalue. Therefore the inverse of the trace is a lower bound on the least eigenvalue of the inverse of the matrix. In particular inequality (46) implies that all the eigenvalues of $H^{(k+1)}$ exceed the amount $(\mathbf{g}^{(k+1)T}H^{(k+1)}\mathbf{g}^{(k+1)})/\|\mathbf{g}^{(k+1)}\|^2$. But this amount is a Rayleigh quotient, so it is an upper bound on the least eigenvalue of $H^{(k+1)}$, which is a contradiction. Theorem 1 is proved.

We now prove that the rate of convergence of $F(\mathbf{x}^{(k)})$, $k = 1, 2, 3, \ldots$, to $F(\xi)$ is at least linear.

THEOREM 2. *If the objective function $F(\mathbf{x})$ has continuous second derivatives, and if there exists a positive constant ε that is not greater than the least eigenvalue of any second derivative matrix of $F(\mathbf{x})$, then there exists a number $\mu < 1$, dependent on $\mathbf{x}^{(1)}$ but not on k, such that the sequence of function values $F(\mathbf{x}^{(k)})$, $k = 1, 2, 3 \ldots$, satisfies the inequality*

$$F(\mathbf{x}^{(k)}) - F(\xi) \leqslant \mu^k[F(\mathbf{x}^{(1)}) - F(\xi)], \tag{47}$$

where ξ is the position of the least value of $F(\mathbf{x})$.

Proof of Theorem 2. Already we have remarked that the inequality (46) is a contradiction, so from expression (39) we deduce the relation

$$\sum_{j=1}^{k} \frac{\|\mathbf{g}^{(j+1)}\|^2}{(\mathbf{g}^{(j+1)T}H^{(j)}\mathbf{g}^{(j+1)})} \leqslant Bk. \tag{48}$$

Therefore, following an argument like the one which was used to deduce inequality (44), we note that for at least two thirds of the integers $j = 1, 2, \ldots k$ the inequality

$$\|\mathbf{g}^{(j+1)}\|^2 \leqslant 3B(\mathbf{g}^{(j+1)T}H^{(j)}\mathbf{g}^{(j+1)}) \tag{49}$$

is satisfied. Therefore the inequalities (44) and (49) are both satisfied for at least one third of the integers, in which case the condition

$$\|\mathbf{g}^{(j+1)}\|^2 < 9BM\|\boldsymbol{\gamma}^{(j)}\|^2 \tag{50}$$

is obtained, whence from Lemma 3 the inequality

$$\varepsilon[F(\mathbf{x}^{(j+1)}) - F(\xi)] < 9BM\|\boldsymbol{\gamma}^{(j)}\|^2 \tag{51}$$

is satisfied.

To treat the right-hand side of inequality (51) we note that Lemma 2 shows that $\|\boldsymbol{\gamma}^{(j)}\|^2$ is not more than a certain constant multiple of $\|\boldsymbol{\delta}^{(j)}\|^2$. Moreover the relation

(28) provides a convenient bound on $\|\delta^{(j)}\|^2$, so we deduce that there exists a constant h such that the inequality

$$F(\mathbf{x}^{(j+1)})-F(\xi) < h[F(\mathbf{x}^{(j)})-F(\mathbf{x}^{(j+1)})] \tag{52}$$

holds. Therefore we find that for one third of the values of j the inequality

$$F(\mathbf{x}^{(j+1)})-F(\xi) < \left(\frac{h}{h+1}\right)[F(\mathbf{x}^{(j)})-F(\xi)] \tag{53}$$

is obtained. For the remaining values of j the algorithm ensures that $[F(\mathbf{x}^{(j+1)})-F(\xi)] < [F(\mathbf{x}^{(j)})-F(\xi)]$, so Theorem 2 is proved, the value of μ in the inequality (47) being the cube root of $h/(h+1)$.

The following lemma is an important consequence of Theorem 2.

Lemma 5. The sum $\sum_{k=1}^{\infty} \|\delta^{(k)}\|$ *is convergent.*

Proof of Lemma 5. Equation (28) provides the inequality

$$\|\delta^{(k)}\|^2 \leqslant 2[F(\mathbf{x}^{(k)})-F(\mathbf{x}^{(k+1)})]/\varepsilon, \tag{54}$$

and the definition of ξ implies that $[F(\mathbf{x}^{(k)})-F(\mathbf{x}^{(k+1)})] \leqslant [F(\mathbf{x}^{(k)})-F(\xi)]$. Therefore from Theorem 2 we deduce the result

$$\|\delta^{(k)}\| \leqslant (\sqrt{\mu})^k\{2[F(\mathbf{x}^{(1)})-F(\xi)]/\varepsilon\}^{\frac{1}{2}}, \tag{55}$$

which shows that $\Sigma\|\delta^{(k)}\|$ is convergent. The lemma is proved.

The proofs of Theorems 3 and 4 make frequent use of this lemma. In particular they require the numbers $\Delta^{(k)}$, $k = 1, 2, 3, \ldots$ which are defined by the equation

$$\Delta^{(k)} = \sum_{j=k}^{\infty} \|\delta^{(j)}\|. \tag{56}$$

Note that $\Delta^{(k)}$ is an upper bound on the distance $\|\mathbf{x}^{(k)}-\xi\|$.

Our last two theorems require one extra condition on the function $F(\mathbf{x})$. It is that there exists a constant L such that, for all $\mathbf{x}$ satisfying the condition $F(\mathbf{x}) \leqslant F(\mathbf{x}^{(1)})$, the inequality

$$\left|\frac{\partial^2 F(\mathbf{x})}{\partial x_i \partial x_j}-\frac{\partial^2 F(\xi)}{\partial x_i \partial x_j}\right| \leqslant L\|\mathbf{x}-\xi\|, \; i, j = 1, 2, \ldots, n, \tag{57}$$

is obtained. Lemma 1 and the definition of a derivative imply that this condition is always satisfied if $F(\mathbf{x})$ is three times differentiable at ξ.

The main purpose of inequality (57) is that it implies the following lemma.

LEMMA 6. *For* $k = 1, 2, 3, \ldots$, *the inequality*

$$\|\gamma^{(k)}-G(\xi)\delta^{(k)}\| \leqslant nL\Delta^{(k)}\|\delta^{(k)}\| \tag{58}$$

is satisfied, where $\delta^{(k)}$, $\gamma^{(k)}$ *and* $\Delta^{(k)}$ *are defined by equations* (3), (5) *and* (56), *and where* $G(\xi)$ *is the second derivative matrix of* $F(\mathbf{x})$ *at* ξ.

Proof of Lemma 6. We apply the mean value theorem to equation (16) and deduce that, for $i = 1, 2, \ldots n$, there exists a number θ_i from the interval [0, 1] such that

$$\gamma_i^{(k)} = [G(\mathbf{x}^{(k)}+\theta_i\delta^{(k)})\delta^{(k)}]_i, \tag{59}$$

where the subscript "i" denotes the ith component of an n-component vector. Therefore we have the identity

$$[\gamma^{(k)}-G(\xi)\delta^{(k)}]_i = [\{G(\mathbf{x}^{(k)}+\theta_i\delta^{(k)})-G(\xi)\}\delta^{(k)}]_i. \tag{60}$$

Now inequality (57) and the definition of $\Delta^{(k)}$ imply that the absolute value of every component of the matrix $\{G(\mathbf{x}^{(k)}+\theta_i\delta^{(k)})-G(\xi)\}$ is not greater than $L\Delta^{(k)}$, so equation (60) provides the inequality

$$|[\gamma^{(k)}-G(\xi)\delta^{(k)}]_i| \leqslant n^{\frac{1}{2}}L\Delta^{(k)}\|\delta^{(k)}\|, \tag{61}$$

from which we deduce the truth of Lemma 6.

We state one other result as a lemma because it is required by Theorems 3 and 4.

LEMMA 7. *If P is any non-singular $n\times n$ matrix, and if we make two parallel applications of the variable metric algorithm, the first to the function $F(\mathbf{x})$, starting at the point $\mathbf{x}^{(1)}$, with the positive definite matrix $H^{(1)}$, and the second to the function $\bar{F}(\mathbf{x}) = F(P^{-1}\mathbf{x})$, starting at the point $\bar{\mathbf{x}}^{(1)} = P\mathbf{x}^{(1)}$, with the positive definite matrix $PH^{(1)}P^T$, then the following equations are satisfied for all values of k*

$$\bar{F}(\bar{\mathbf{x}}^{(k)}) = F(\mathbf{x}^{(k)}) \tag{62}$$

$$\bar{\mathbf{x}}^{(k)} = P\mathbf{x}^{(k)} \tag{63}$$

$$\bar{\mathbf{g}}^{(k)} = (P^{-1})^T\mathbf{g}^{(k)} \tag{64}$$

and

$$\bar{H}^{(k)} = PH^{(k)}P^T, \tag{65}$$

where the bars distinguish the second application of the variable metric algorithm.

Proof of Lemma 7. Equation (63) and the definition of $\bar{F}(\mathbf{x})$ would imply that equations (62) and (64) are valid, and equations (63) and (65) hold for $k = 1$. Therefore we just have to check that if equations (63) and (65) hold for any value of k, then they also hold for $(k+1)$. In fact it is straightforward to verify that this property is obtained from the definition of the variable metric algorithm, so the lemma is proved.

We now prove Theorem 3, which states that the elements of the matrices $\Gamma^{(k)}$ and $H^{(k)}$ do not grow without limit as k tends to infinity, provided that the third derivative of $F(\mathbf{x})$ exists at the minimum point ξ.

THEOREM 3. *If the function $F(\mathbf{x})$ satisfies the conditions of Theorems* 1 *and* 2, *and if also there exists a constant L such that, for all $\mathbf{x}$ satisfying the condition $F(\mathbf{x}) \leqslant F(\mathbf{x}^{(1)})$, the inequality*

$$\left|\frac{\partial^2 F(\mathbf{x})}{\partial x_i \partial x_j} - \frac{\partial^2 F(\xi)}{\partial x_i \partial x_j}\right| \leqslant L\|\mathbf{x}-\xi\|, \quad i, j = 1, 2, \ldots, n, \tag{66}$$

is obtained, then the matrices $H^{(k)}$, $k = 1, 2, 3, \ldots$, generated by the variable metric algorithm, are uniformly bounded. Moreover the eigenvalues of the matrices $H^{(k)}$ are bounded away from zero.

Proof of Theorem 3. We let $G(\xi)$ be the second derivative matrix of $F(\mathbf{x})$ at the minimum ξ, and we consider the effect of applying the variable metric algorithm to the function $\bar{F}(\mathbf{x}) = F([G(\xi)]^{-\frac{1}{2}}\mathbf{x})$, starting at the point $\bar{\mathbf{x}}^{(1)} = [G(\xi)]^{\frac{1}{2}}\mathbf{x}^{(1)}$ with the positive definite matrix $\bar{H}^{(1)} = [G(\xi)]^{\frac{1}{2}}H^{(1)}[G(\xi)]^{\frac{1}{2}}$. Lemma 7 states that instead of the matrices $H^{(k)}$, $k = 1, 2, 3, \ldots$, we would calculate the matrices $\bar{H}^{(k)} = [G(\xi)]^{\frac{1}{2}}H^{(k)}[G(\xi)]^{\frac{1}{2}}$. Therefore, we note that the matrices $H^{(k)}$ have the properties claimed in the statement of the theorem if and only if these properties are obtained by the matrices $\bar{H}^{(k)}$. Moreover the types of condition that are imposed on the derivatives of $F(\mathbf{x})$ are inherited by the derivatives of $\bar{F}(\mathbf{x})$. Therefore Theorem 3 is valid for the function $F(\mathbf{x})$ if and only if it is valid for the function $\bar{F}(\mathbf{x})$. But the definition of

$F(\mathbf{x})$ is such that its second derivative matrix at its minimum point is equal to the unit matrix. Therefore there is no loss of generality in proving Theorem 3 subject to the condition $G(\xi) = I$, which we do because this property provides substantial simplifications to the algebra of the proof. In particular we deduce from Lemma 6 the inequality

$$\|\gamma^{(k)}-\delta^{(k)}\| \leqslant nL\Delta^{(k)}\|\delta^{(k)}\|. \tag{67}$$

The method of proof that is used to show that the matrices $H^{(k)}$ are bounded is a direct one. From equation (4) we deduce that the ratio $\|H^{(k+1)}\|/\max[1, \|H^{(k)}\|]$ is bounded by a number, $m^{(k)}$ say, that is only a small amount greater than one, and that tends to one as k tends to infinity. Finally we show that the product $\Pi m^{(k)}$ is finite.

To bound $\|H^{(k+1)}\|/\max[1, \|H^{(k)}\|]$ we define the matrix

$$Q^{(k)} = I-\frac{\delta^{(k)}\delta^{(k)T}}{(\delta^{(k)T}\delta^{(k)})}+\frac{\gamma^{(k)}\gamma^{(k)T}}{(\gamma^{(k)T}\delta^{(k)})}, \tag{68}$$

and we note that $Q^{(k)}$ is positive definite, and that the equation $Q^{(k)}\delta^{(k)} = \gamma^{(k)}$ is satisfied. This definition of $Q^{(k)}$ is useful because it gives the inequality

$$\|[Q^{(k)}]^{\frac{1}{2}}H^{(k+1)}[Q^{(k)}]^{\frac{1}{2}}\| \leqslant \max\{1, \|[Q^{(k)}]^{\frac{1}{2}}H^{(k)}[Q^{(k)}]^{\frac{1}{2}}\|\}. \tag{69}$$

We now prove this assertion, using the fact that the Euclidean norm of a symmetric matrix is the largest of the absolute values of its eigenvalues.

We define the vector

$$\mathbf{z}^{(k)} = [Q^{(k)}]^{\frac{1}{2}}\delta^{(k)} = [Q^{(k)}]^{-\frac{1}{2}}\gamma^{(k)}, \tag{70}$$

and we express $\delta^{(k)}$ and $\gamma^{(k)}$ in terms of $\mathbf{z}^{(k)}$ in equation (4). Thus we obtain the identity

$$\begin{aligned}[Q^{(k)}]^{\frac{1}{2}}H^{(k+1)}[Q^{(k)}]^{\frac{1}{2}} &= J^{(k)}-\frac{J^{(k)}\mathbf{z}^{(k)}\mathbf{z}^{(k)T}J^{(k)}}{(\mathbf{z}^{(k)T}J^{(k)}\mathbf{z}^{(k)})}+\frac{\mathbf{z}^{(k)}\mathbf{z}^{(k)T}}{(\mathbf{z}^{(k)T}\mathbf{z}^{(k)})}\\ &= R^{(k)}+\frac{\mathbf{z}^{(k)}\mathbf{z}^{(k)T}}{(\mathbf{z}^{(k)T}\mathbf{z}^{(k)})}\end{aligned} \tag{71}$$

say, where $J^{(k)} = [Q^{(k)}]^{\frac{1}{2}}H^{(k)}[Q^{(k)}]^{\frac{1}{2}}$. We have to show that the eigenvalues of the matrix (71) are bounded by $\max[1, \|J^{(k)}\|]$. The matrix $R^{(k)}$ is obtained by subtracting a symmetric, rank one matrix from $J^{(k)}$, so $\|R^{(k)}\|$ can exceed $\|J^{(k)}\|$ only if the least eigenvalue of $R^{(k)}$ is so negative that its modulus exceeds $\|J^{(k)}\|$. But $R^{(k)}$ has an eigenvalue of zero, which must be the least eigenvalue because $J^{(k)}$ is positive definite. Therefore $\|R^{(k)}\| \leqslant \|J^{(k)}\|$. The addition of the term $\mathbf{z}^{(k)}\mathbf{z}^{(k)T}/(\mathbf{z}^{(k)T}\mathbf{z}^{(k)})$ to $R^{(k)}$ changes only one eigenvalue. It increases the eigenvalue whose eigenvector is $\mathbf{z}^{(k)}$ from zero to one. Therefore the inequality (69) holds.

From expression (69) we deduce the inequality

$$\begin{aligned}\|H^{(k+1)}\| &\leqslant \|[Q^{(k)}]^{-\frac{1}{2}}\|^2 \max\{1, \|[Q^{(k)}]^{\frac{1}{2}}\|^2\|H^{(k)}\|\}\\ &= \|[Q^{(k)}]^{-1}\| \max\{1, \|Q^{(k)}\|\,\|H^{(k)}\|\},\end{aligned} \tag{72}$$

because the matrix $Q^{(k)}$ is symmetric. We define the number

$$m^{(k)} = \max\{1, \|[Q^{(k)}]^{-1}\|\} \,.\, \max\{1, \|Q^{(k)}\|\}, \tag{73}$$

and note that inequality (72) provides the bound

$$\begin{aligned}\|H^{(k+1)}\| &\leqslant m^{(k)} \max\{1, \|H^{(k)}\|\}\\ &\leqslant \prod_{j=1}^{k} m^{(j)} \max\{1, \|H^{(1)}\|\}.\end{aligned} \tag{74}$$

Therefore we now calculate bounds on $\|[Q^{(k)}]^{-1}\|$ and $\|Q^{(k)}\|$, in order to bound $m^{(k)}$.

The definition (68) gives the bound

$$\|Q^{(k)}-I\| = \left\|\frac{\gamma^{(k)}\gamma^{(k)T}}{(\gamma^{(k)T}\delta^{(k)})}-\frac{\delta^{(k)}\delta^{(k)T}}{(\delta^{(k)T}\delta^{(k)})}\right\| \leqslant$$

$$\leqslant \left\|\frac{(\gamma^{(k)}-\delta^{(k)})\gamma^{(k)T}}{(\gamma^{(k)T}\delta^{(k)})}\right\|+\left\|\frac{\delta^{(k)}(\gamma^{(k)}-\delta^{(k)})^T}{(\gamma^{(k)T}\delta^{(k)})}\right\|+\left\|\frac{\delta^{(k)}\delta^{(k)T}}{(\gamma^{(k)T}\delta^{(k)})}-\frac{\delta^{(k)}\delta^{(k)T}}{(\delta^{(k)T}\delta^{(k)})}\right\|. \tag{75}$$

Because the Euclidean norm of the rank one matrix $\mathbf{u}\mathbf{v}^T$ is equal to $\|\mathbf{u}\|\,\|\mathbf{v}\|$, the first term on the right-hand side of this inequality is equal to $\|\gamma^{(k)}-\delta^{(k)}\|\,\|\gamma^{(k)}\|/(\gamma^{(k)T}\delta^{(k)})$, the second term is equal to $\|\delta^{(k)}\|\,\|\gamma^{(k)}-\delta^{(k)}\|/(\gamma^{(k)T}\delta^{(k)})$, and the third term is equal to

$$\|\delta^{(k)}\|^2\left|\frac{1}{(\gamma^{(k)T}\delta^{(k)})}-\frac{1}{(\delta^{(k)T}\delta^{(k)})}\right| \leqslant \frac{\|\gamma^{(k)}-\delta^{(k)}\|\,\|\delta^{(k)}\|}{(\gamma^{(k)T}\delta^{(k)})}. \tag{76}$$

Therefore Lemma 2 implies that there is a constant D, independent of the iteration number, such that the inequality

$$\|Q^{(k)}-I\| \leqslant D\|\gamma^{(k)}-\delta^{(k)}\|/\|\delta^{(k)}\| \tag{77}$$

holds, so from expression (67) we deduce the condition

$$\|Q^{(k)}-I\| \leqslant nDL\Delta^{(k)}. \tag{78}$$

Now equation (56) implies that $\Delta^{(k)}$ tends to zero as k tends to infinity, so as well as obtaining the bound

$$\|Q^{(k)}\| \leqslant 1+nDL\Delta^{(k)} \tag{79}$$

from inequality (78), we also note that there exists a constant η such that the condition

$$\|[Q^{(k)}]^{-1}\| \leqslant 1+\eta\Delta^{(k)}. \tag{80}$$

holds.

We substitute these bounds in the definition (73), and obtain the result

$$m^{(k)} \leqslant (1+\eta\Delta^{(k)})(1+nDL\Delta^{(k)}). \tag{81}$$

Therefore, because of inequality (74), to prove the first part of the theorem it is sufficient to show that the product $\Pi(1+\eta\Delta^{(k)})(1+nDL\Delta^{(k)})$ is bounded. It is a well known theorem of analysis that this product is finite if the sum $\Sigma\Delta^{(k)}$ is absolutely convergent.

From equations (55) and (56) we deduce the inequality

$$\Delta^{(k)} \leqslant (\sqrt{\mu})^k\{2[F(\mathbf{x}^{(1)})-F(\xi)]/\varepsilon\}^{\frac{1}{2}}/(1-\sqrt{\mu}). \tag{82}$$

Therefore $\Sigma\Delta^{(k)}$ is convergent, and the first part of Theorem 3 is proved.

We also have to prove that the eigenvalues of the matrices $H^{(k)}$ are bounded away from zero. It is sufficient to show that the matrices $\Gamma^{(k)} \equiv [H^{(k)-1}]$, $k = 1, 2, 3, \ldots$, are uniformly bounded. We apply a direct method like the one we have just used, and we find that we can replace H by Γ in inequality (74).

To obtain this result we use equation (70) to express $\delta^{(k)}$ and $\gamma^{(k)}$ in terms of $\mathbf{z}^{(k)}$ in equation (31). Thus we deduce the identity

$$[Q^{(k)}]^{-\frac{1}{2}}\Gamma^{(k+1)}[Q^{(k)}]^{-\frac{1}{2}} = \left(I-\frac{\mathbf{z}^{(k)}\mathbf{z}^{(k)T}}{(\mathbf{z}^{(k)T}\mathbf{z}^{(k)})}\right)\{[Q^{(k)}]^{-\frac{1}{2}}\Gamma^{(k)}[Q^{(k)}]^{-\frac{1}{2}}\}$$

$$\left(I-\frac{\mathbf{z}^{(k)}\mathbf{z}^{(k)T}}{(\mathbf{z}^{(k)T}\mathbf{z}^{(k)})}\right)+\frac{\mathbf{z}^{(k)}\mathbf{z}^{(k)T}}{(\mathbf{z}^{(k)T}\mathbf{z}^{(k)})} = S^{(k)}+\frac{\mathbf{z}^{(k)}\mathbf{z}^{(k)T}}{(\mathbf{z}^{(k)T}\mathbf{z}^{(k)})} \tag{83}$$

say. Now because the matrix $(I-\mathbf{z}\mathbf{z}^T/\mathbf{z}^T\mathbf{z})$ is a symmetric projection operator, we have the inequality $\|S^{(k)}\| \leqslant \|[Q^{(k)}]^{-\frac{1}{2}}\Gamma^{(k)}[Q^{(k)}]^{-\frac{1}{2}}\|$. Moreover the addition of the term $\mathbf{z}^{(k)}\mathbf{z}^{(k)T}/(\mathbf{z}^{(k)T}\mathbf{z}^{(k)})$ to $S^{(k)}$ changes only one eigenvalue: it increases the eigenvalue whose eigenvector is $\mathbf{z}^{(k)}$ from zero to one. Therefore we deduce a relation that is similar to inequality (69), namely the condition

$$\|[Q^{(k)}]^{-\frac{1}{2}}\Gamma^{(k+1)}[Q^{(k)}]^{-\frac{1}{2}}\| \leqslant \max\{1, \|[Q^{(k)}]^{-\frac{1}{2}}\Gamma^{(k)}[Q^{(k)}]^{-\frac{1}{2}}\|\}, \tag{84}$$

from which we obtain the inequality

$$\begin{aligned}\|\Gamma^{(k+1)}\| &\leqslant \|Q^{(k)}\| \max\{1, \|[Q^{(k)}]^{-1}\| \, \|\Gamma^{(k)}\|\} \\ &\leqslant m^{(k)} \max\{1, \|\Gamma^{(k)}\|\} \\ &\leqslant \prod_{j=1}^{k} m^{(j)} \max\{1, \|\Gamma^{(1)}\|\}. \end{aligned} \tag{85}$$

We have already proved that the product $\Pi m^{(j)}$ is bounded, so now the proof of Theorem 3 is complete.

Although Theorem 3 has some intrinsic interest, its main value is that it enables us to prove that the variable metric algorithm converges superlinearly.

THEOREM 4. *If the function $F(\mathbf{x})$ satisfies the conditions of Theorem 3, then the ratio $\|\mathbf{x}^{(k+1)}-\xi\|/\|\mathbf{x}^{(k)}-\xi\|$ tends to zero as k tends to infinity, where the vectors $\mathbf{x}^{(k)}$, $k = 1, 2, 3, \ldots$, are the points generated by the variable metric algorithm, and where $F(\xi)$ is the least value of $F(\mathbf{x})$.*

Proof of Theorem 4. First we apply the argument that was used at the beginning of Theorem 3, and from the definition $\bar{F}(\mathbf{x}) = F([G(\xi)]^{-\frac{1}{2}}\mathbf{x})$ and equation (63) we deduce the identity

$$\bar{\mathbf{x}}^{(k)}-\bar{\xi} = [G(\xi)]^{-\frac{1}{2}}\{\mathbf{x}^{(k)}-\xi\}, \tag{86}$$

where $\bar{\xi}$ is the position of the least value of $\bar{F}(\mathbf{x})$. Thus we find that again there is no loss of generality in supposing that $G(\xi)$ is the unit matrix, so we do assume that $G(\xi) = I$, in order to simplify the algebra of the proof.

The basis of the proof of the theorem is a relation between $\|\Gamma^{(k+1)}-I\|$ and $\|\Gamma^{(k)}-I\|$. Because $G(\xi) = I$, and because there are similarities between $G(\mathbf{x}^{(k)})$ and $\Gamma^{(k)}$, we expect a study of the sequence $\|\Gamma^{(k)}-I\|$, k = 1, 2, 3, ..., to yield some interesting results. I have some examples that show that this sequence need not converge to zero, but the sequence does imply that the ratio $\|\mathbf{g}^{(k+1)}\|/\|\mathbf{g}^{(k)}\|$ tends to zero. Thus we prove the theorem.

To study the sequence of numbers $\|\Gamma^{(k)}-I\|$, we deduce from equation (31) the relation

$$(\Gamma^*-I) = \left(I-\frac{\gamma\gamma^T}{(\gamma^T\gamma)}\right)(\Gamma-I)\left(I-\frac{\gamma\gamma^T}{(\gamma^T\gamma)}\right)+T, \tag{87}$$

where T is the matrix

$$T = \left\{\frac{\gamma\gamma^T}{(\delta^T\gamma)}-\frac{\gamma\gamma^T}{(\gamma^T\gamma)}\right\}+\left\{\frac{\Gamma\gamma\gamma^T}{(\gamma^T\gamma)}-\frac{\Gamma\delta\gamma^T}{(\delta^T\gamma)}\right\}+ \left\{\frac{\gamma\gamma^T\Gamma}{(\gamma^T\gamma)}-\frac{\gamma\delta^T\Gamma}{(\delta^T\gamma)}\right\}+\left\{\frac{\gamma\gamma^T(\delta^T\Gamma\delta)}{(\delta^T\gamma)^2}-\frac{\gamma\gamma^T(\gamma^T\Gamma\gamma)}{(\gamma^T\gamma)^2}\right\}. \tag{88}$$

We have grouped the terms of T into pairs so that, using Lemma 2, inequality (67)

and the boundedness of Γ, it can be shown that $\|T\|$ is bounded by a multiple of $\Delta^{(k)}$, following a method like the one we applied to inequality (75). However to prove Theorem 4 it is necessary to use Frobenius norms of matrices instead of Euclidean norms, in which case $\|T\|$ remains bounded by a multiple of $\Delta^{(k)}$, because the Frobenius norm of an $n\times n$ matrix is at most $\sqrt{n}$ times the Euclidean norm. We indicate Frobenius norms by the subscript "F", and we let the bound on $\|T^{(k)}\|_F$ be $W\Delta^{(k)}$. Therefore equation (87) gives the inequality

$$\|\Gamma^{(k+1)}-I\|_F \leqslant \left\|\left(I-\frac{\gamma^{(k)}\gamma^{(k)T}}{(\gamma^{(k)T}\gamma^{(k)})}\right)(\Gamma^{(k)}-I)\left(I-\frac{\gamma^{(k)}\gamma^{(k)T}}{(\gamma^{(k)T}\gamma^{(k)})}\right)\right\|_F + W\Delta^{(k)} = \beta^{(k)}+W\Delta^{(k)} \tag{89}$$

say.

To relate $\beta^{(k)}$ to $\|\Gamma^{(k)}-I\|_F$, we recall that the square of the Frobenius norm of the matrix C, say, is equal to $\Sigma\|Ce_i\|^2$, where $\{e_1, e_2, \ldots, e_n\}$ is any set of orthonormal vectors. In particular by letting $\varepsilon_1 = \gamma^{(k)}/\|\gamma^{(k)}\|$, we deduce the inequality

$$\|\Gamma^{(k)}-I\|_F^2-\{\beta^{(k)}\}^2 \geqslant \|(\Gamma^{(k)}-I)\gamma^{(k)}\|^2/\|\gamma^{(k)}\|^2. \tag{90}$$

Further, by dividing by the factor $\{\|\Gamma^{(k)}-I\|_F+\beta^{(k)}\}$ we obtain the condition

$$\|\Gamma^{(k)}-I\|_F-\beta^{(k)} \geqslant \|(\Gamma^{(k)}-I)\gamma^{(k)}\|^2/2V\|\gamma^{(k)}\|^2, \tag{91}$$

where V is a bound on $\|\Gamma^{(k)}-I\|_F$ that is independent of k. Note that condition (91) depends on the property $\beta^{(k)} \leqslant V$. Therefore from inequality (89) we find the relation

$$\|\Gamma^{(k+1)}-I\|_F \leqslant \|\Gamma^{(1)}-I\|_F-\sum_{j=1}^{k}\frac{\|(\Gamma^{(j)}-I)\gamma^{(j)}\|^2}{2V\|\gamma^{(j)}\|^2}+W\sum_{j=1}^{k}\Delta^{(j)}. \tag{92}$$

Now we know from inequality (82) that $\Sigma\Delta^{(j)}$ is convergent, so we deduce that the sum $\Sigma\|(\Gamma^{(j)}-I)\gamma^{(j)}\|^2/\|\gamma^{(j)}\|^2$ is also convergent. This remark is interesting and useful because, if it happens that the sequence of matrices $\Gamma^{(j)}$ fails to converge to the unit matrix, then the sum shows that there are restrictions on the *directions* of the vectors $\gamma^{(j)}$. We let σ equal the value of the sum.

We now use Lemma 2, Theorem 3 and inequalities (67) and (82) to deduce that the different sum $\Sigma\|\Gamma^{(j)}\delta^{(j)}-\gamma^{(j)}\|^2/\|\gamma^{(j)}\|^2$ is also absolutely convergent. Indeed we have the bound

$$\sum_{j=1}^{\infty}\frac{\|\Gamma^{(j)}\delta^{(j)}-\gamma^{(j)}\|^2}{\|\gamma^{(j)}\|^2} \leqslant \sum_{j=1}^{\infty}\frac{2\{\|(\Gamma^{(j)}-I)\gamma^{(j)}\|^2+\|\Gamma^{(j)}\delta^{(j)}-\Gamma^{(j)}\gamma^{(j)}\|^2\}}{\|\gamma^{(j)}\|^2} \leqslant$$

$$2\sigma+2\{\max_j\|\Gamma^{(j)}\|^2\}\left\{\max_j\frac{\|\delta^{(j)}\|^2}{\|\gamma^{(j)}\|^2}\right\}(nL)^2\sum_{j=1}^{\infty}[\Delta^{(j)}]^2 < \infty. \tag{93}$$

To bound the numerator of this sum, we use the Schwarz inequality to obtain the expression

$$\|\Gamma^{(j)}\delta^{(j)}-\gamma^{(j)}\| \geqslant [\{-H^{(j)}g^{(j+1)}\}^T\{\Gamma^{(j)}\delta^{(j)}-\gamma^{(j)}\}]/\|H^{(j)}g^{(j+1)}\|$$
$$= (g^{(j+1)T}H^{(j)}g^{(j+1)})/\|H^{(j)}g^{(j+1)}\|, \tag{94}$$

the last line being a consequence of equations (1) and (6). Therefore, because of Theorem 3, we deduce that the sum $\Sigma\|g^{(j+1)}\|^3/\|\gamma^{(j)}\|^2$ is convergent. Further, the elementary inequality

$$\|g^{(j+1)}-g^{(j)}\|^2 \leqslant 2\{\|g^{(j)}\|^2+\|g^{(j+1)}\|^2\} \tag{95}$$

shows that the sum $\Sigma\|\mathbf{g}^{(J+1)}\|^2/\{\|\mathbf{g}^{(J)}\|^2+\|\mathbf{g}^{(J+1)}\|^2\}$ is also convergent, so we can now make the useful deduction

$$\lim_{k\to\infty}\|\mathbf{g}^{(k+1)}\|/\|\mathbf{g}^{(k)}\| = 0. \tag{96}$$

To complete the proof of the theorem, we note that Lemma 2 shows that the expression

$$\frac{\|\mathbf{x}^{(k+1)}-\xi\|}{\|\mathbf{g}^{(k+1)}-\mathbf{g}(\xi)\|}\Big/\frac{\|\mathbf{x}^{(k)}-\xi\|}{\|\mathbf{g}^{(k)}-\mathbf{g}(\xi)\|} = \frac{\|\mathbf{x}^{(k+1)}-\xi\|}{\|\mathbf{x}^{(k)}-\xi\|}\Big/\frac{\|\mathbf{g}^{(k+1)}\|}{\|\mathbf{g}^{(k)}\|} \tag{97}$$

is bounded. Therefore Theorem 4 is an immediate consequence of statement (96).

4. Discussion

The theorems show that one can expect the variable metric algorithm to be entirely successful whenever the objective function has a positive definite second derivative matrix for all vectors of variables $\mathbf{x}$. However this condition is usually not obtained, so we ought to try to understand the success of the algorithm when the objective function is not convex. The purpose of this section is to make a few remarks on the non-convex case.

In particular we point out that the theorems presented in this paper are sometimes relevant to non-convex functions, because the conditions on $F(\mathbf{x})$ have to be obtained only for values of $\mathbf{x}$ that satisfy the inequality $F(\mathbf{x}) \leqslant F(\mathbf{x}^{(1)})$. Moreover the structure of the variable metric algorithm is such that any calculated vector $\mathbf{x}^{(k)}$ can be regarded as a starting point for the later iterations. Therefore if the algorithm is applied to a non-convex function, and if it happens that a point $\mathbf{x}^{(k)}$ is calculated, such that our derivative conditions are met for all $\mathbf{x}$ satisfying the condition $F(\mathbf{x}) \leqslant F(\mathbf{x}^{(k)})$, then we expect convergence to the minimum at a super-linear rate. Moreover if the sequence of points $\mathbf{x}^{(k)}$ converges to a local minimum of $F(\mathbf{x})$, that is not the global minimum, then it may also be possible to apply the theorems to infer super-linear convergence, by isolating the domain of $\mathbf{x}$ to a neighbourhood of the local minimum.

However we are not able to draw any conclusions about the behaviour of the algorithm when the estimates $\mathbf{x}^{(k)}$ are in a region where the second derivative matrices of $F(\mathbf{x})$ are not positive definite. My experience is that usually the linear searches for a minimum, made on each iteration, tend to avoid such regions, for often it is possible to verify by calculation that a few iterations of the variable metric algorithm lead to a satisfactory convex region, in which case we now understand why the algorithm is so successful. But occasional numerical examples show very slow progress. Therefore in the non-convex case it would be worthwhile to have a characterization of the possible limit points of the sequence $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \mathbf{x}^{(3)}, \ldots$ I believe that there is no need for the gradient of $F(\mathbf{x})$ to be zero at these limit points, in which case substantial revisions should be made to the algorithm, because often it is available to inexperienced computer users as a general purpose method for calculating the least value of an arbitrary differentiable function.

Dr P. Wolfe encouraged me to prove these theorems, and many of the ideas that are presented are a direct result of our discussions. Dr R. Fletcher also contributed to

some of these meetings, and he showed me the relevance of eigenvalues to the proof of Theorem 3.

References

DAVIDON, W. C. 1959 *Variable metric method for minimization.* A.E.C. Research and Development Report, ANL-5990 (Rev.).

FLETCHER, R. & POWELL, M. J. D. 1963 *Comput. J.* **6**, 163–168.

GOLDFARB, D. 1969 In R. Fletcher (Ed.), *Optimization*, pp. 273–281. London: Academic Press.

MCCORMICK, G. P. & PEARSON, J. D. 1969 In R. Fletcher (Ed.), *Optimization*, pp. 307–325. London: Academic Press.