J. Inst. Maths Applics (1970) **6**, 76–90

The Convergence of a Class of Double-rank Minimization Algorithms
1. General Considerations

C. G. BROYDEN
Computing Centre, University of Essex,
Wivenhoe Park, Colchester, Essex

[Received 7 March 1969 and in revised form 19 May 1969]

This paper presents a more detailed analysis of a class of minimization algorithms, which includes as a special case the DFP (Davidon–Fletcher–Powell) method, than has previously appeared. Only quadratic functions are considered but particular attention is paid to the magnitude of successive errors and their dependence upon the initial matrix. On the basis of this a possible explanation of some of the observed characteristics of the class is tentatively suggested.

1. Introduction

PROBABLY the best-known algorithm for determining the unconstrained minimum of a function of many variables, where explicit expressions are available for the first partial derivatives, is that of Davidon (1959) as modified by Fletcher & Powell (1963). This algorithm has many virtues. It is simple and does not require at any stage the solution of linear equations. It minimizes a quadratic function exactly in a finite number of steps and this property makes convergence of this algorithm rapid, when applied to more general functions, in the neighbourhood of the solution. It is, at least in theory, stable since the iteration matrix $\mathbf{H}_i$, which transforms the ith gradient into the ith step direction, may be shown to be positive definite.

In practice the algorithm has been generally successful, but it has exhibited some puzzling behaviour. Broyden (1967) noted that $\mathbf{H}_i$ does not always remain positive definite, and attributed this to rounding errors. Pearson (1968) found that for some problems the solution was obtained more efficiently if $\mathbf{H}_i$ was reset to a positive definite matrix, often the unit matrix, at intervals during the computation. Bard (1968) noted that $\mathbf{H}_i$ could become singular, attributed this to rounding error and suggested the use of suitably chosen scaling factors as a remedy.

In this paper we analyse the more general algorithm given by Broyden (1967), of which the DFP algorithm is a special case, and determine how for quadratic functions the choice of an arbitrary parameter affects convergence. We investigate how the successive errors depend, again for quadratic functions, upon the initial choice of iteration matrix paying particular attention to the cases where this is either the unit matrix or a good approximation to the inverse Hessian. We finally give a tentative explanation of some of the observed experimental behaviour in the case where the function to be minimized is not quadratic.

2. Basic Theory

Define a quadratic function $F(\mathbf{x})$ by

$$F(\mathbf{x}) \equiv \tfrac{1}{2}\mathbf{x}^T\mathbf{A}\mathbf{x}-\mathbf{b}^T\mathbf{x}+c, \tag{2.1}$$

where $\mathbf{A}$ is an nth order positive definite matrix and $\mathbf{b}$ is a vector of order n. If we denote grad F by $\mathbf{f}(\mathbf{x})$ then

$$\mathbf{f}(\mathbf{x}) \equiv \mathbf{A}\mathbf{x}-\mathbf{b}, \tag{2.2}$$

and the unique minimum of $F(\mathbf{x})$ occurs at $\mathbf{s}$, where

$$\mathbf{s} = \mathbf{A}^{-1}\mathbf{b}. \tag{2.3}$$

Let $\mathbf{x}_i$ be an approximation to $\mathbf{s}$ and define the error $\mathbf{e}_i$ at $\mathbf{x}_i$ by

$$\mathbf{e}_i = \mathbf{x}_i-\mathbf{s}. \tag{2.4}$$

If we denote $\mathbf{f}(\mathbf{x}_i)$ by $\mathbf{f}_i$ then, from equations (2.2)–(2.4), it follows that

$$\mathbf{f}_i = \mathbf{A}\mathbf{e}_i. \tag{2.5}$$

Let now $\mathbf{x}_{i+1}$ be a new approximation to $\mathbf{s}$ where

$$\mathbf{x}_{i+1} = \mathbf{x}_i+\mathbf{p}_i t_i, \tag{2.6}$$

$\mathbf{p}_i$ is an arbitrary vector, t_i is that value of t which minimizes $F(\mathbf{x}(t))$ and where

$$\mathbf{x}(t) = \mathbf{x}_i+\mathbf{p}_i t. \tag{2.7}$$

The minimization of $F(\mathbf{x}(t))$ implies, as is well known (see e.g. Fletcher & Powell, 1963), that the gradient of $F(\mathbf{x})$ at $\mathbf{x}_{i+1}$ is orthogonal to $\mathbf{p}_i$ so that

$$\mathbf{p}_i^T\mathbf{f}_{i+1} = 0. \tag{2.8}$$

Now from equations (2.4) and (2.6) it follows that

$$\mathbf{e}_{i+1}-\mathbf{e}_i = \mathbf{p}_i t_i, \tag{2.9}$$

and pre-multiplying this equation by $\mathbf{p}_i^T\mathbf{A}$, together with equations (2.5) and (2.8), yields

$$-\mathbf{p}_i^T\mathbf{A}\mathbf{e}_i = \mathbf{p}_i^T\mathbf{A}\mathbf{p}_i t_i, \tag{2.10}$$

so that the unique value of t_i which minimizes $F(\mathbf{x}(t))$ is given by

$$t_i = \frac{-\mathbf{p}_i^T\mathbf{A}\mathbf{e}_i}{\mathbf{p}_i^T\mathbf{A}\mathbf{p}_i}. \tag{2.11}$$

Substituting this value in equation (2.9) and rearranging gives the basic error equation

$$\mathbf{e}_{i+1} = (\mathbf{I}-\mathbf{p}_i\mathbf{p}_i^T\mathbf{A}/\mathbf{p}_i^T\mathbf{A}\mathbf{p}_i)\mathbf{e}_i, \tag{2.12}$$

and we now derive a sufficient condition for n-step convergence of processes to which the above equation is applicable.

LEMMA 1. *If*

$$\mathbf{p}_i^T\mathbf{A}\mathbf{p}_j = 0, \qquad 1 \leqslant j < i \leqslant n, \tag{2.13}$$

then $\mathbf{e}_{n+1}$ *is null for arbitrary* $\mathbf{e}_1$.

Proof. If $\mathbf{I}$ is the unit matrix of order n and $\mathbf{x}_i, \mathbf{y}_i$, $i = 1, 2, \ldots, n$, are nth order vectors then necessary and sufficient conditions for the matrix $(\mathbf{I}-\mathbf{x}_1\mathbf{y}_1)(\mathbf{I}-\mathbf{x}_2\mathbf{y}_2^T)\ldots(\mathbf{I}-\mathbf{x}_n\mathbf{y}_n^T)$ to be null are that

$$\mathbf{y}_i^T\mathbf{x}_i = 1, \qquad 1 \leqslant i \leqslant n, \tag{2.14a}$$

and

$$\mathbf{y}_i^T\mathbf{x}_j = 0, \qquad 1 \leqslant j < i \leqslant n. \tag{2.14b}$$

(For proof see e.g. Broyden, 1967.)
The Lemma follows from this result and equation (2.12).

3. The Generalized Method

In this method (Broyden, 1967) the vector $\mathbf{p}_i$ is given by

$$\mathbf{p}_i = -\mathbf{H}_i\mathbf{f}_i, \tag{3.1}$$

where $\mathbf{H}_i$ is positive definite. $\mathbf{H}_1$ is chosen to be an arbitrary positive definite matrix (often the unit matrix) and $\mathbf{H}_{i+1}$ is given by

$$\mathbf{H}_{i+1} = \mathbf{H}_i - \mathbf{H}_i\mathbf{y}_i\mathbf{w}_i^T + \mathbf{p}_i t_i \mathbf{q}_i^T, \qquad i = 1, 2, \ldots, \tag{3.2a}$$

where

$$\mathbf{y}_i = \mathbf{f}_{i+1} - \mathbf{f}_i, \tag{3.2b}$$

$$\mathbf{q}_i^T = \alpha_i \mathbf{p}_i^T - \beta_i \mathbf{y}_i^T \mathbf{H}_i, \tag{3.2c}$$

$$\mathbf{w}_i^T = \gamma_i \mathbf{y}_i^T \mathbf{H}_i + \beta_i t_i \mathbf{p}_i^T, \tag{3.2d}$$

$$\alpha_i = (1 + \beta_i \mathbf{y}_i^T \mathbf{H}_i \mathbf{y}_i)/\mathbf{p}_i^T \mathbf{y}_i, \tag{3.2e}$$

$$\gamma_i = (1 - \beta_i t_i \mathbf{p}_i^T \mathbf{y}_i)/\mathbf{y}_i^T \mathbf{H}_i \mathbf{y}_i. \tag{3.2f}$$

The parameter β_i is arbitrary and setting it equal to zero gives the DFP method (Fletcher & Powell, 1963). It was shown by Broyden (1967) that the matrices $\mathbf{H}_i$ constructed in this way are always positive definite if $\beta_i \geqslant 0$.

In order to analyse the convergence of this algorithm we define three more quantities. Let $\mathbf{B}$ be the positive definite matrix that satisfies the equation

$$\mathbf{B}^2 = \mathbf{A}, \tag{3.3}$$

and define $\mathbf{z}_i$, $\mathbf{K}_i$ by

$$\mathbf{z}_i = \mathbf{B}\mathbf{e}_i, \tag{3.4}$$

and

$$\mathbf{K}_i = \mathbf{B}\mathbf{H}_i\mathbf{B}. \tag{3.5}$$

These definitions coupled with equations (2.12) and (3.2) then yield, after some tedious and uninstructive algebra involving relationships established in Section 2 (above),

$$\mathbf{z}_{i+1} = \mathbf{z}_i - \mathbf{K}_i\mathbf{z}_i t_i, \tag{3.6}$$

$$\mathbf{K}_{i+1} = \mathbf{K}_i - \mathbf{K}_i^2\mathbf{z}_i t_i^2(\gamma_i \mathbf{z}_i^T \mathbf{K}_i^2 + \beta_i \mathbf{z}_i^T \mathbf{K}_i) + \mathbf{K}_i\mathbf{z}_i t_i^2((\alpha_i/t_i)\mathbf{z}_i^T\mathbf{K}_i - \beta_i \mathbf{z}_i^T\mathbf{K}_i^2), \tag{3.7}$$

and

$$t_i = \mathbf{z}_i^T\mathbf{K}_i\mathbf{z}_i / \mathbf{z}_i^T\mathbf{K}_i^2\mathbf{z}_i. \tag{3.8}$$

It follows immediately from the last three equations and equations (3.2e, f) that

$$\mathbf{z}_{i+1}^T\mathbf{K}_i\mathbf{z}_i = 0, \tag{3.9}$$

and

$$\mathbf{K}_{i+1}\mathbf{K}_i\mathbf{z}_i = \mathbf{K}_i\mathbf{z}_i. \tag{3.10}$$

The last two equations are the key equations to this part of the analysis, the proof of n-step convergence of the process relying heavily upon them.

THEOREM 1. *The generalized algorithm exhibits n-step convergence.*

Proof. The proof is inductive.

Assume that

$$\mathbf{z}_i^T\mathbf{K}_j\mathbf{z}_j = 0, \qquad 1 \leqslant j \leqslant i-1, \tag{3.11a}$$

and

$$\mathbf{K}_i\mathbf{K}_j\mathbf{z}_j = \mathbf{K}_j\mathbf{z}_j, \qquad 1 \leqslant j \leqslant i-1, \tag{3.11b}$$

where i is some arbitrary integer in the range $1 < i < n$. Then

$$\mathbf{z}_i^T\mathbf{K}_i\mathbf{K}_j\mathbf{z}_j = 0, \tag{3.12a}$$

and

$$\mathbf{z}_i^T\mathbf{K}_i^2\mathbf{K}_j\mathbf{z}_j = 0. \tag{3.12b}$$

It follows immediately from equations (3.6) and (3.7) that

$$\mathbf{z}_{i+1}^T\mathbf{K}_j\mathbf{z}_j = 0, \qquad 1 \leqslant j \leqslant i-1, \tag{3.13a}$$

and

$$\mathbf{K}_{i+1}\mathbf{K}_j\mathbf{z}_j = \mathbf{K}_j\mathbf{z}_j, \qquad 1 \leqslant j \leqslant i-1, \tag{3.13b}$$

and combining these with equations (3.9) and (3.10) gives equations (3.11a, b) with the i replaced by $i+1$. The induction is completed by noting that equations (3.11a, b) are satisfied for $i = 2$, $j = 1$ by putting i equal to unity in equations (3.9) and (3.10). It follows that after n steps equation (3.12a) is satisfied for $1 \leqslant j < i \leqslant n$ so that, from equations (3.3)–(3.5),

$$\mathbf{e}_i^T\mathbf{A}\mathbf{H}_i\mathbf{A}\mathbf{H}_j\mathbf{A}\mathbf{e}_j = 0, \qquad 1 \leqslant j < i \leqslant n.$$

This becomes, from equations (2.5) and (3.1),

$$\mathbf{p}_i^T\mathbf{A}\mathbf{p}_j = 0, \qquad 1 \leqslant j < i \leqslant n, \tag{3.14}$$

and n-step convergence then follows from Lemma 1.

We note that, from equation (3.14), the step vectors $\mathbf{p}_i$ are conjugate with respect to $\mathbf{A}$ so that the method is one of "conjugate directions" (Hestenes & Stiefel, 1952). Moreover, from equations (3.3)–(3.5) and equations (3.11a),

$$\mathbf{e}_i^T\mathbf{A}\mathbf{H}_j\mathbf{A}\mathbf{e}_j = 0, \qquad 1 \leqslant j < i \leqslant n,$$

which becomes, from equations (2.5) and (3.1),

$$\mathbf{f}_i^T\mathbf{p}_j = 0, \qquad 1 \leqslant j < i \leqslant n. \tag{3.15}$$

This establishes the well-known result that the gradient is orthogonal to all the previous steps, so that the algorithm progressively minimizes $F(\mathbf{x})$ in the subspaces defined by the vectors $\mathbf{p}_i$.

4. Preliminary Analysis

It was demonstrated in the previous section that the generalized algorithm converges for quadratic functions in at most n steps. Circumstances exist, however, when convergence occurs in fewer than n steps, an extreme case being that where the initial iteration matrix is $\mathbf{A}^{-1}$ and only one step is necessary. It thus seems likely that under the appropriate conditions convergence may occur in any number of steps between 1 and n, and it also seems likely that the precise number will depend upon the initial conditions. We shall show, in fact, that the number of steps required is equal to the number of linearly independent vectors in the sequence $\mathbf{z}_1, \mathbf{K}_1\mathbf{z}_1, \mathbf{K}_1^2\mathbf{z}_1, \ldots$, and we shall denote this number by m.

We shall also be concerned with how the sequence of errors $\{\mathbf{e}_i\}$ is governed by the values of $\mathbf{K}_1$ and $\mathbf{z}_1$, and in order to analyse this it is necessary to obtain expressions

for $\mathbf{e}_i$ in terms of $\mathbf{K}_1$ and $\mathbf{z}_1$. It is found that a convenient way of achieving this is by expressing $\mathbf{z}_1$ in terms of m orthogonal components, one of which is annihilated at every step of the process.

To investigate then the effect of the choice of $\mathbf{K}_1$ and $\mathbf{z}_1$ on the subsequent behaviour of the algorithm we begin by defining a class of polynomial central to this part of the theory.

Definition 1

Define $\pi_{r,j}$ to be the class of matrix polynomials of degree r in $\mathbf{K}_j$ whose lowest power of $\mathbf{K}_j$ is at least the first.

THEOREM 2. *If $\mathbf{K}_i$ and $\mathbf{z}_i$ are as previously defined, $\mathbf{K}_i$ is positive definite and $\beta_i \geqslant 0$ then*

$$\mathbf{K}_{i+1}^m \mathbf{z}_{i+1} = \mathbf{P}_{m+1,i}\mathbf{z}_i, \tag{4.1}$$

where m is any positive integer, and where

$$\mathbf{P}_{m+1,i} \in \pi_{m+1,i}.$$

Proof. This is by induction. Assume the validity of equation (4.1). Then, from equation (3.7),

$$\mathbf{K}_{i+1}^{m+1}\mathbf{z}_{i+1} = \mathbf{K}_i\mathbf{P}_{m+1,i}\mathbf{z}_i - \mathbf{K}_i^2\mathbf{z}_i\lambda_i + \mathbf{K}_i\mathbf{z}_i\mu_i, \tag{4.2}$$

where λ_i and μ_i are scalars. Now $\mathbf{K}_i\mathbf{P}_{m+1,i} \in \pi_{m+2,i}$ so that equation (4.2) may be written

$$\mathbf{K}_{i+1}^{m+1}\mathbf{z}_{i+1} = \mathbf{P}_{m+2,i}\mathbf{z}_i. \tag{4.3}$$

Note that, for $m \geqslant 1$, $\mathbf{P}_{m+2,i} \in \pi_{m+2,i}$ since the terms involving $\mathbf{K}_i$ and $\mathbf{K}_i^2$ in equation (4.2) do not affect the degree of the polynomial. Thus if equation (4.1) is true for m it is also true for $m+1$. To initiate the induction we note that equations (3.6) and (3.7) yield

$$\mathbf{K}_{i+1}\mathbf{z}_{i+1} = \mathbf{K}_i\mathbf{z}_i\phi_i - \mathbf{K}_i^2\mathbf{z}_i\psi_i, \tag{4.4}$$

where ϕ_i and ψ_i are scalars. To show that $\psi_i \neq 0$, and hence that the term in equation (4.4) involving $\mathbf{K}_i^2$ does not vanish, premultiply equation (4.4) by $\mathbf{z}_{i+1}^T$ giving, from equation (3.9),

$$\mathbf{z}_{i+1}^T\mathbf{K}_{i+1}\mathbf{z}_{i+1} = \mathbf{z}_{i+1}^T\mathbf{K}_i^2\mathbf{z}_i\psi_i. \tag{4.5}$$

Now it follows from equation (3.5) that $\mathbf{K}_i$ is positive definite if and only if $\mathbf{H}_i$ is positive definite. Now $\mathbf{K}_i$ is positive definite by hypothesis so that $\mathbf{H}_i$ is positive definite. Moreover $\beta_i \geqslant 0$ so that (Broyden, 1967) $\mathbf{H}_{i+1}$ and hence $\mathbf{K}_{i+1}$ are both positive definite. It now follows from equation (4.5) that $\psi_i \neq 0$, so that equation (4.4) may be written

$$\mathbf{K}_{i+1}\mathbf{z}_{i+1} = \mathbf{P}_{2,i}\mathbf{z}_i, \tag{4.6}$$

where $\mathbf{P}_{2,i} \in \pi_{2,i}$. This establishes equations (4.1) with $m = 1$ completing both the induction and the proof.

Corollary. If $\mathbf{Q}_{m,i+1} \in \pi_{m,i+1}$ then

$$\mathbf{Q}_{m,i+1}\mathbf{z}_{i+1} = \mathbf{Q}_{m+1,i}\mathbf{z}_i, \tag{4.7}$$

where

$$\mathbf{Q}_{m+1,i} \in \pi_{m+1,i}.$$

Proof. Apply Theorem 2 to each term in $\mathbf{Q}_{m,i+1}$ in turn.

THEOREM 3. *If $\beta_i \geqslant 0$, $j = 1, 2, \ldots, i-1$, and $\mathbf{K}_1$ is positive definite then*

$$\mathbf{K}_i\mathbf{z}_i = \mathbf{R}_{i,1}\mathbf{z}_1, \tag{4.8}$$

where

$$\mathbf{R}_{i,1} \in \pi_{i,1}.$$

Proof. Replacing i by $i-1$ in equation (4.6) gives

$$\mathbf{K}_i\mathbf{z}_i = \mathbf{P}_{2,i-1}\mathbf{z}_{i-1}, \tag{4.9}$$

and successive use of the corollary of Theorem 2 yields the theorem.

The effect of Theorem 3 is to establish a relationship between the vectors $\mathbf{K}_i\mathbf{z}_i$, $i = 1, 2, \ldots m$, and the vectors $\mathbf{K}_1^i\mathbf{z}_1$, $i = 1, 2, \ldots, m$. To simplify the subsequent notation we omit the subscript unity, so that

$$\mathbf{K} = \mathbf{K}_1 \tag{4.10a}$$

and

$$\mathbf{z} = \mathbf{z}_1. \tag{4.10b}$$

We assume, moreover, for the remainder of this section that $\mathbf{H}_1$ and hence $\mathbf{K}$ are positive definite.

Now equation (4.8) may be written

$$\mathbf{K}_i\mathbf{z}_i = [\mathbf{Kz}, \mathbf{K}^2\mathbf{z}, \ldots, \mathbf{K}^i\mathbf{z}]\mathbf{w}_i, \tag{4.11}$$

where the elements of $\mathbf{w}_i$ are the coefficients of $\mathbf{R}_{i,1}$. Since $\mathbf{R}_{i,1}$ is of degree i the last element of $\mathbf{w}_i$ is not zero so that $\mathbf{w}_i$ is not null. *It follows that it is impossible for $\mathbf{K}_i\mathbf{z}_i$, and hence $\mathbf{z}_i$, to be null if the vectors $\mathbf{Kz}, \mathbf{K}^2\mathbf{z}, \ldots, \mathbf{K}^i\mathbf{z}$ are linearly independent* and we shall subsequently prove the converse of this statement. To proceed with the analysis we assume that only the first m, where $1 \leqslant m \leqslant n$, terms of the sequence $\{\mathbf{K}^i\mathbf{z}\}$, $i = 1, 2, \ldots$, are linearly independent and we define the matrix $\mathbf{M}$ by

$$\mathbf{M} = [\mathbf{Kz}, \mathbf{K}^2\mathbf{z}, \ldots, \mathbf{K}^m\mathbf{z}]. \tag{4.12}$$

Clearly $\mathbf{M}$ is an $n \times m$ matrix of rank m. Equation (4.11) may now be written

$$\mathbf{K}_i\mathbf{z}_i = \mathbf{M}\mathbf{v}_i, \qquad 1 \leqslant i \leqslant m, \tag{4.13}$$

where $\mathbf{v}_i$ is a vector whose first i elements are the same as those of $\mathbf{w}_i$ and whose remaining elements are zero. Equation (4.13) may itself be written

$$[\mathbf{K}_1\mathbf{z}_1, \mathbf{K}_2\mathbf{z}_2, \ldots, \mathbf{K}_m\mathbf{z}_m] = \mathbf{MV}, \tag{4.14}$$

where $\mathbf{V}$ is an $m \times m$ upper triangular matrix whose ith column is $\mathbf{v}_i$.

Now the vectors $\mathbf{K}_i\mathbf{z}_i$, $i = 1, 2, \ldots, m$ form, from equation (3.12a), an orthogonal set and since $\mathbf{V}$ is upper triangular, equation (4.14) shows that a basis for the first r of those vectors, $1 \leqslant r \leqslant m$, is provided by the first r terms of the sequence $\{\mathbf{K}^i\mathbf{z}\}$. Since the vectors $\mathbf{K}_i\mathbf{z}_i$ are orthogonal it seems natural to normalize them and to investigate their possible use as a basis for the $\mathbf{z}_i$'s. Denote the normalized form of $\mathbf{K}_i\mathbf{z}_i$ by $\mathbf{q}_i$. Then

$$[\mathbf{K}_1\mathbf{z}_1, \mathbf{K}_2\mathbf{z}_2, \ldots, \mathbf{K}_m\mathbf{z}_m]\mathbf{R} = \mathbf{Q}, \tag{4.15}$$

where

$$\mathbf{Q} = [\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_m], \tag{4.16}$$

and $\mathbf{R}$ is diagonal and chosen so that

$$\mathbf{Q}^T\mathbf{Q} = \mathbf{I}. \tag{4.17}$$

It follows from equations (4.14) and (4.15) that

$$\mathbf{Q} = \mathbf{MU}, \tag{4.18}$$

where

$$\mathbf{U} = \mathbf{VR}. \tag{4.19}$$

The choice of $\mathbf{R}$ is unique apart from the signs of its non-zero (diagonal) elements, and we choose these arbitrarily to make the diagonal elements of $\mathbf{U}$ positive. Note that from equation (4.18) they cannot be zero since $\mathbf{U}$ is upper triangular and both $\mathbf{M}$ and $\mathbf{Q}$ have rank m.

We show now that subject to the sign convention adopted $\mathbf{Q}$ is determined uniquely by $\mathbf{K}$ and $\mathbf{z}$. Equations (4.17) and (4.18) yield

$$\mathbf{U}^T\mathbf{M}^T\mathbf{MU} = \mathbf{I}, \tag{4.20}$$

so that

$$\mathbf{UU}^T = (\mathbf{M}^T\mathbf{M})^{-1}, \tag{4.21}$$

and the result follows from the uniqueness of the Choleski decomposition given the sign of the square roots.

We have thus implicitly constructed from the sequence $\{\mathbf{K}^i\mathbf{z}\}$ a unique set of orthonormal vectors $\mathbf{q}_i$. We show now that the vectors $\mathbf{z}_i$, $i = 1, 2,..., m$ may be expressed extremely simply in terms of these vectors.

THEOREM 4. *If* $\mathbf{Q}$ *is as defined above there exists a vector* $\mathbf{c}$ *such that*

$$\mathbf{z} = \mathbf{Qc}. \tag{4.22}$$

Proof. Since by hypothesis $\mathbf{K}^{m+1}\mathbf{z}$ is a linear combination of the vectors $\mathbf{K}^i\mathbf{z}$, $i = 2,..., m$, there exist numbers α_i, $i = 1, 2,..., m+1$ such that

$$\sum_{i=1}^{m+1} \mathbf{K}^i\mathbf{z}\alpha_i = 0, \tag{4.23}$$

which becomes, on pre-multiplying by $\mathbf{K}^{-1}$,

$$\sum_{i=1}^{m+1} \mathbf{K}^{i-1}\mathbf{z}\alpha_i = 0. \tag{4.24}$$

Hence $\alpha_1 \neq 0$, since if it were zero it would imply the linear dependence of the vectors $\mathbf{Kz}_1$, $\mathbf{K}^2\mathbf{z}_1,..., \mathbf{K}^m\mathbf{z}$. It follows again from equation (4.24) that $\mathbf{z}$ may thus be expressed in terms of these vectors, so that we may write

$$\mathbf{z} = \mathbf{Ma} \tag{4.25}$$

where $\mathbf{a} = [a_i]$ and $a_i = -\alpha_{i+1}/\alpha_1$. Now $\mathbf{U}$ is non-singular so that we may define $\mathbf{c}$ by

$$\mathbf{c} = \mathbf{U}^{-1}\mathbf{a}. \tag{4.26}$$

The theorem follows from equations (4.18), (4.25) and (4.26).

The relationship between $\mathbf{z}$ and $\mathbf{z}_i$, $i = 2, 3,..., m$, is provided by the following theorem.

THEOREM 5. *Define* $\mathbf{z}$,$\mathbf{Q}$,$\mathbf{c}$ *as above. Then*

$$\mathbf{z}_i = \mathbf{Qc}_i, \qquad i = 1, 2,..., m, \tag{4.27}$$

where $\mathbf{c}_i$ *is equal to* $\mathbf{c}$ *with its first* $i-1$ *elements set equal to zero.*

Proof. This is by induction. Assume the validity of equation (4.27) for some i, $1 \leqslant i \leqslant m$. Then, if $\mathbf{c} = [\gamma_j]$,

$$\mathbf{z}_i = \sum_{j=1}^{m} \mathbf{q}_j\gamma_j. \tag{4.28}$$

But, from equations (3.6) and (4.15) and (4.16),

$$\mathbf{z}_{i+1} = \mathbf{z}_i - \mathbf{q}_i t_i \rho_i, \tag{4.29}$$

where $|\rho_i| = \|\mathbf{K}_i\mathbf{z}_i\|$, so that

$$\mathbf{z}_{i+1} = \mathbf{q}_i(\gamma_i - t_i\rho_i) + \sum_{j=i+1}^{m} \mathbf{q}_j\gamma_j. \tag{4.30}$$

But, from equations (3.9), (4.15) and (4.16),

$$\mathbf{q}_i^T\mathbf{z}_{i+1} = 0, \tag{4.31}$$

so that, since the vectors $\mathbf{q}_i$ are orthonormal,

$$\gamma_i = t_i\rho_i. \tag{4.32}$$

This establishes that if equation (4.8) is true for i it is true for $i+1$, and its truth for $i = 1$ follows from equation (4.10b) and Theorem 4. The theorem follows.

This theorem and the preceding one show that the vectors $\mathbf{z}_i$, $i = 1, 2,..., m$, and hence the error vectors $\mathbf{e}_i$ and the vectors of independent variables $\mathbf{x}_i$, are independent of the values of the parameter β_i that appears in equations (3.2c, d, e, f). They depend entirely upon the initial values $\mathbf{K}$ and $\mathbf{z}$, since these values determine $\mathbf{M}$ which in turn gives (subject to the arbitrary choice of signs) unique values of $\mathbf{Q}$ and $\mathbf{c}$. A further point is that the total number of iterations m depends only upon $\mathbf{K}$ and $\mathbf{z}$. Since β_i is arbitrary it follows that if $m < n$ the final value of $\mathbf{H}_i$ is to a certain extent arbitrary and in this case may be but a poor approximation to $\mathbf{A}^{-1}$.

5. Analysis of the Vector Errors

In the previous section we established a general relationship between the vectors $\mathbf{z}_i$ and the initial matrix $\mathbf{K}$. In this section we derive specific properties of the error vectors $\mathbf{e}_i$ when $\mathbf{K}$ satisfies certain conditions. In particular we consider the cases where $\mathbf{K} = \mathbf{A}$, and where it approximates to the unit matrix. These cases occur when the initial matrix is respectively the unit matrix and a good approximation to $\mathbf{A}^{-1}$. We shall, moreover, in the former case, show that the method becomes the method of conjugate gradients.

THEOREM 6. *Let* $\mathbf{K}$ *and* $\mathbf{Q}$ *be as previously defined. Then*

$$\mathbf{KQ} = \mathbf{QT}, \tag{5.1}$$

where $\mathbf{T} = [t_{ij}]$ *is symmetric, positive definite, tridiagonal, non-negative and irreducible.*

Proof. Equations (5.1) and the orthogonality of the columns of $\mathbf{Q}$ give

$$\mathbf{T} = \mathbf{Q}^T\mathbf{KQ}, \tag{5.2}$$

and the properties of symmetry and positive definiteness follow immediately. To prove the remainder of the theorem we note that, by hypothesis, $\mathbf{K}^{m+1}\mathbf{z}$ is a linear combination of the vectors $\mathbf{K}^i\mathbf{z}$, $i = 1, 2,..., m$, so that

$$\mathbf{K}^{m+1}\mathbf{z} = \mathbf{Mh},$$

where $\mathbf{h}$ is some mth order vector. Thus

$$\mathbf{KM} = \mathbf{MH}, \tag{5.3}$$

where

$$\mathbf{H} = [\mathbf{s}_2, \mathbf{s}_2,..., \mathbf{s}_m, \mathbf{h}], \tag{5.4}$$

and $\mathbf{s}_i$ is the ith column of the mth order unit matrix. Clearly $\mathbf{H}$ is upper Hessenberg and it is moreover non-singular, for since $\mathbf{K}$ is non-singular the product $\mathbf{KM}$ must have rank m, and hence, from equation (5.3), $\mathbf{H}$ has rank m. Now $\mathbf{T}$ becomes, from equations (4.18), (5.1) and (5.3),

$$\mathbf{T} = \mathbf{U}^T\mathbf{M}^T\mathbf{M}\mathbf{H}\mathbf{U}, \tag{5.5}$$

and this becomes, from equation (4.20),

$$\mathbf{T} = \mathbf{U}^{-1}\mathbf{H}\mathbf{U}. \tag{5.6}$$

Since the product of an upper triangular and an upper Hessenberg matrix is itself upper Hessenberg it follows from equation (5.6) that $\mathbf{T}$ is upper Hessenberg. It is thus, since it is symmetric, tridiagonal. To show that it is non-negative we note that since it is positive definite its diagonal elements must be positive. To obtain its off-diagonal elements we define the matrix $\mathbf{C} = [c_{ij}]$ by

$$\mathbf{C} = \mathbf{H}\mathbf{U}, \tag{5.7}$$

so that, from equation (5.6),

$$\mathbf{C} = \mathbf{U}\mathbf{T}. \tag{5.8}$$

Equating the expressions for $c_{i+1,i}$, $i = 1, 2, \ldots, n-1$, from equations (5.7) and (5.8) yields, from equation (5.4),

$$u_{i,i} = u_{i+1,i+1}t_{i+1,i}, \qquad i = 1, 2, \ldots, n-1. \tag{5.9}$$

Now $u_{i,i} > 0$, $i = 1, 2, \ldots, n$ so equation (5.9) establishes the non-negativity and irreducibility of $\mathbf{T}$, completing the proof.

Corollary

$$\mathbf{Q}^T\mathbf{K}^{-1}\mathbf{Q} = \mathbf{T}^{-1}. \tag{5.10}$$

Proof. Pre-multiplication by $\mathbf{K}^{-1}$ and post-multiplication by $\mathbf{H}^{-1}$ of equation (5.3) gives

$$\mathbf{K}^{-1}\mathbf{M} = \mathbf{M}\mathbf{H}^{-1}, \tag{5.11}$$

so that, from equations (4.18) and (4.20),

$$\mathbf{Q}^T\mathbf{K}^{-1}\mathbf{Q} = \mathbf{U}^{-1}\mathbf{H}^{-1}\mathbf{U}. \tag{5.12}$$

The corollary follows from equation (5.6).

We prove now two lemmas.

LEMMA 2. *The vector $\mathbf{c} = [\gamma_j]$ is the first column of $\mathbf{T}^{-1}$ multiplied by a scaling factor whose modulus is $\|\mathbf{K}\mathbf{z}\|$.*

Proof. By definition $\mathbf{K}\mathbf{z} = \mathbf{q}_1\rho_1$ where $|\rho_1| = \|\mathbf{K}\mathbf{z}\|$. But from Theorem 4, $\mathbf{K}\mathbf{z} = \mathbf{K}\mathbf{Q}\mathbf{c}$ so that, from the above and Theorem 6, $\mathbf{Q}\mathbf{T}\mathbf{c} = \mathbf{q}_1\rho_1$. Hence, pre-multiplying by $\mathbf{Q}^T$,

$$\mathbf{T}\mathbf{c} = \mathbf{s}_1\rho_1, \tag{5.13}$$

where $\mathbf{s}_1$ is the first column of the mth order unit matrix, proving the lemma.

LEMMA 3. *The last $m-i$ elements of $\mathbf{T}\mathbf{c}_i$ are zero.*

Proof. From Theorem 5, $\mathbf{c}_i$ is obtained from $\mathbf{c}$ by altering its first $i-1$ elements. Since $\mathbf{T}$ is tridiagonal this implies that only the first i elements of $\mathbf{T}\mathbf{c}_i$ differ from those of $\mathbf{T}\mathbf{c}$. The remaining $m-i$ must then be zero from Lemma 2, completing the proof.

We are now in a position to prove our first convergence theorem about the generalized algorithm.

THEOREM 7. The residual vectors $\mathbf{f}_i$ obtained by the generalized algorithm are conjugate with respect to the initial iteration matrix.

Proof. It follows from equations (2.5), (3.3)–(3.5) and (4.10a) that

$$\mathbf{f}_i^T\mathbf{H}_1\mathbf{f}_j = \mathbf{z}_i^T\mathbf{K}\mathbf{z}_j,$$

which becomes, from Theorems 5 and 6,

$$\mathbf{f}_i^T\mathbf{H}_1\mathbf{f}_j = \mathbf{c}_i^T\mathbf{T}\mathbf{c}_j.$$

Assume that $i > j$. The first $i-1$ elements of $\mathbf{c}_i$ are zero from Theorem 5 and the last $m-j$ elements of $\mathbf{T}\mathbf{c}_j$ are zero from Lemma 3. Since $i > j$, $m-j > i$ so that each term in the inner product of $\mathbf{c}_i$ and $\mathbf{T}\mathbf{c}_j$ has at least one zero factor, and thus $\mathbf{f}_i^T\mathbf{H}_1\mathbf{f}_j = 0$. If $j > i$ the same result follows from symmetry, proving the theorem.

Corollary 1. If $\mathbf{H}_1 = t\mathbf{I}$, where t is a positive scalar, the method becomes identical to the method of conjugate gradients.

Proof. It was established in Section 3 that the method is one of conjugate directions, and it was shown by Hestenes & Stiefel (1952) that any conjugate direction method having orthogonal residuals is essentially a conjugate gradient method.

Corollary 2. If $\mathbf{H}_1 = t\mathbf{I}$ then $\|\mathbf{e}_{i+1}\| < \|\mathbf{e}_i\|$, $i = 1, 2, \ldots, m$.

Proof. This follows from the first corollary and the known properties of the conjugate gradient method.

We have thus established that if the initial iteration matrix for the generalized algorithm is the unit matrix the norms of the errors decrease strictly monotonically. The second case occurs where $\mathbf{H}_1$ is a close approximation to $\mathbf{A}^{-1}$. In these circumstances intuition suggests that convergence will be strict monotonic, and it is natural to ask how nearly $\mathbf{H}_1$ must approach $\mathbf{A}^{-1}$ for this to be true. Although we do not, in fact, prove the strict monotonicity of convergence we do show that the norms of successive errors are bounded by a strictly monotonically decreasing sequence, and that the more closely $\mathbf{H}_1$ approximates $\mathbf{A}^{-1}$ the more rapidly does this sequence converge.

Before proving the theorem, however, we prove the following Lemma.

LEMMA 4. *The spectral norm of a matrix cannot be less than the Euclidean norm of one of its columns.*

Proof. By definition

$$\|\mathbf{A}\| = \max_{\|\mathbf{x}\|=1} \|\mathbf{A}\mathbf{x}\|, \tag{5.14}$$

so that. if $\mathbf{A}$ has n columns and if $\mathbf{s}_i$ is the ith column of the nth order unit matrix,

$$\|\mathbf{A}\mathbf{s}_i\| \leqslant \|\mathbf{A}\|.$$

The lemma follows.

We now obtain error bounds on $\|\mathbf{e}_i\|$, $i = 2, 3, \ldots, m$, in the case where $\mathbf{H}_1$ is a good approximation to $\mathbf{A}^{-1}$.

THEOREM 8. *If $\mathbf{K}$ is as previously defined, and*

$$\mathbf{K} = \mathbf{I} + \mathbf{E}^*, \tag{5.15}$$

where $\|\mathbf{E}^*\| = \sigma < 1$, *then*

$$\|\mathbf{e}_i\| \leqslant k\frac{(1+\sigma)}{(1-\sigma)}\sigma^{i-1}\|\mathbf{e}_1\|, \tag{5.16}$$

where k *is the square root of the condition number of* $\mathbf{A}$, *i.e. where*

$$k^2 = \|\mathbf{A}\|\,\|\mathbf{A}^{-1}\|. \tag{5.17}$$

Proof. Equations (5.2) and (5.15) yield

$$\mathbf{T} = \mathbf{I}+\mathbf{E}, \tag{5.18}$$

where $\mathbf{E} = \mathbf{Q}^T\mathbf{E}^*\mathbf{Q}$ and since $\mathbf{Q}^T\mathbf{Q} = \mathbf{I}$, $\|\mathbf{E}\| \leqslant \sigma$. Since $\mathbf{E}$ is convergent it follows from equation (5.18) that

$$\mathbf{T}^{-1} = \sum_{j=0}^{\infty}(-\mathbf{E})^j. \tag{5.19}$$

Define the vector $\mathbf{t}_i^*$ to be the first column of $\mathbf{T}^{-1}$ with the first $i-1$ elements set equal to zero. Now $\mathbf{T}$ is, by Theorem 6, tridiagonal so that, from equation (5.18), $\mathbf{E}$ is tridiagonal. It follows from this and equation (5.19) that only the terms $\mathbf{E}^j$, $j \geqslant i-1$, can make a non-zero contribution to $\mathbf{t}_i^*$. Thus $\mathbf{t}_i^*$ is equal to the first column of the matrix $\sum_{j=i-1}^{\infty}(-\mathbf{E})^j$ with the first $i-1$ elements set equal to zero. Clearly $\|\mathbf{t}_i^*\|$ is less than or equal to the norm of the first column of $\sum_{j=i-1}^{\infty}(-\mathbf{E})^j$ and this, from Lemma 4, is itself less than or equal to $\|\sum_{j=i-1}^{\infty}(-\mathbf{E})^j\|$. Since

$$\|\sum_{j=i-1}^{\infty}(-\mathbf{E})^j\| \leqslant \frac{\sigma^{i-1}}{(1-\sigma)}, \tag{5.20}$$

it follows that

$$\|\mathbf{t}_i^*\| \leqslant \frac{\sigma^{i-1}}{(1-\sigma)}. \tag{5.21}$$

Now from the definition of $\mathbf{t}_i^*$, Lemma 2 and the relationship between $\mathbf{c}$ and $\mathbf{c}_i$ established by Theorem 5 it follows that

$$\mathbf{c}_i = \mathbf{t}_i^*\rho_1,$$

where $|\rho_1| = \|\mathbf{Kz}\|$. Hence, from equations (5.15) and (5.21),

$$\|\mathbf{c}_i\| \leqslant \frac{(1+\sigma)}{(1-\sigma)}\sigma^{i-1}\|\mathbf{z}\|. \tag{5.22}$$

Now equations (3.3) and (3.4) give

$$\|\mathbf{e}_i\|^2 = \mathbf{z}_i^T\mathbf{A}^{-1}\mathbf{z}_i,$$

which becomes, from Theorem 5,

$$\|\mathbf{e}_i\|^2 = \mathbf{c}_i^T\mathbf{Q}^T\mathbf{A}^{-1}\mathbf{Q}\mathbf{c}_i. \tag{5.23}$$

Thus

$$\|\mathbf{e}_i\| \leqslant \|\mathbf{A}^{-1}\|^{\frac{1}{2}}\frac{(1+\sigma)}{(1-\sigma)}\sigma^{i-1}\|\mathbf{z}\|,$$

and since $\mathbf{z} = \mathbf{Be}_1$ and $\|\mathbf{B}\| = \|\mathbf{A}\|^{\frac{1}{2}}$ the theorem follows.

Corollary. If $\mathbf{A} = \mathbf{I}+\mathbf{E}$, $\|\mathbf{E}\| = \sigma$ and $\sigma < 1$ the successive errors generated by the method of conjugate gradients satisfy

$$\|\mathbf{e}_i\| \leqslant \left(\frac{1+\sigma}{1-\sigma}\right)^{\frac{1}{2}}\sigma^{i-1}\|\mathbf{e}_1\|.$$

Proof. If $\mathbf{A} = \mathbf{I}+\mathbf{E}$, then

$$\|\mathbf{A}\|\,\|\mathbf{A}^{-1}\| \leqslant (1-\sigma)/(1-\sigma),$$

and substituting this result in equation (5.16) proves the corollary.

Theorem 8 provides bounds on the successive error norms $\|\mathbf{e}_i\|$ but does not ensure that convergence is strict monotonic. It does, however, show that if $\mathbf{H}_1$ approximates in norm sufficiently closely to $\mathbf{A}^{-1}$ then convergence is rapid. On the other hand the second corollary of Theorem 7 guarantees strict monotonic convergence if $\mathbf{H}_1$ is the scaled unit matrix despite the fact that, since the scaling is arbitrary, the norm of $\mathbf{H}_1\mathbf{A}^{-1}$ may be arbitrarily large. For other initial matrices the behaviour depends upon the precise relationship between $\mathbf{H}_1$ and $\mathbf{A}$, although it is possible in all cases to obtain a bound on $\|\mathbf{e}_{i+1}\|$ in terms of $\|\mathbf{e}_i\|$. This, from equation (3.4) and the fact that $\|\mathbf{z}_{i+1}\| < \|\mathbf{z}_i\|$, is seen to be

$$\|\mathbf{e}_{i+1}\| < (\|\mathbf{A}\|\,\|\mathbf{A}^{-1}\|)^{\frac{1}{2}}\|\mathbf{e}_i\|, \tag{5.24}$$

so that, if the problem is badly conditioned, substantial increases in $\|\mathbf{e}_i\|$ are possible.

6. Further Analysis of the Sequence $\{\mathbf{K}_i\}$

Equation (3.7) shows that $\mathbf{K}_{i+1}$ depends on both $\mathbf{K}_i$ and β_i, and since β_i is arbitrary $\mathbf{K}_{i+1}$ is also, to a certain extent, arbitrary. But $\mathbf{K}_i$ is itself arbitrary, depending upon the choice of β_{i-1}, and it might therefore be thought that $\mathbf{K}_{i+1}$ would also depend, through $\mathbf{K}_i$, upon β_{i-1}. We now prove that this is not so, and examine more closely the double-rank matrix updating procedure and the dependence of the matrix sequence $\{\mathbf{K}_i\}$ upon the parameters β_i.

THEOREM 9. *The matrix* $\mathbf{K}_i$ *depends only upon the initial matrix* $\mathbf{K}$ *and vector* $\mathbf{z}$ *apart from a single arbitrary additive term of rank one.*

Proof. From equations (3.7), (4.4) and (4.15) it follows that $\mathbf{K}_{j+1}$ may be expressed as

$$\mathbf{K}_{j+1} = \mathbf{K}_j+[\mathbf{q}_j \quad \mathbf{q}_{j+1}]\begin{bmatrix}\omega_j & \xi_j\\ \xi_j & \eta_j\end{bmatrix}\begin{bmatrix}\mathbf{q}_j^T\\ \mathbf{q}_{j+1}^T\end{bmatrix}, \tag{6.1}$$

where ω_j, ξ_j, and η_j are constants whose values we determine subsequently. Putting $j = 1,2,\ldots,i-1$, in equation (6.1) then gives

$$\mathbf{K}_i = \mathbf{K}_i^*+\mathbf{q}_i\eta_{i-1}\mathbf{q}_i^T, \tag{6.2a}$$

where

$$\mathbf{K}_i^* = \mathbf{K}+\mathbf{q}_1\omega_1\mathbf{q}_1^T+\sum_{j=2}^{i-1}\mathbf{q}_j(\omega_j+\eta_{j-1})\mathbf{q}_j^T+\sum_{j=1}^{i-1}\xi_j(\mathbf{q}_j\mathbf{q}_{j+1}^T+\mathbf{q}_{j+1}\mathbf{q}_j^T). \tag{6.2b}$$

Now equations (3.11b) and (4.15) imply that

$$\mathbf{K}_i\mathbf{q}_j = \mathbf{q}_j, \qquad 1 \leqslant j \leqslant i-1, \tag{6.3}$$

so that the orthonormality of the vectors $\mathbf{q}_j$ in conjunction with equations (6.2) and (6.3) gives

$$\omega_1 = 1-\mathbf{q}_1^T\mathbf{K}\mathbf{q}_1, \tag{6.4a}$$

$$\omega_j+\eta_{j-1} = 1-\mathbf{q}_j^T\mathbf{K}\mathbf{q}_j, \qquad 2 \leqslant j \leqslant i-1, \tag{6.4b}$$

and

$$\xi_j = -\mathbf{q}_j^T\mathbf{K}\mathbf{q}_{j+1}, \qquad 1 \leqslant j \leqslant i-1. \tag{6.4c}$$

Thus since the vectors $\mathbf{q}_j$ depend only upon $\mathbf{K}$ and $\mathbf{z}$ (see Section 4 above), equations (6.2b) and (6.4) imply that $\mathbf{K}_i^*$ is determined solely by $\mathbf{K}$ and $\mathbf{z}$, proving the theorem.

It follows from the preceding theorem and the fact that $\mathbf{K}_{i+1}$ depends upon β_i that η_i must also depend upon β_i, and we now derive the precise form of this dependence. If, in order to simplify notation, we denote $\mathbf{z}_i^T\mathbf{K}_i^j\mathbf{z}_i$ by θ_j, $j = 1, 2, 3$, and 4, equation (3.8) may be written

$$\theta_2 t_i = \theta_1, \tag{6.5}$$

and equations (3.6) and (3.7) then give

$$\mathbf{K}_{i+1}\mathbf{z}_{i+1} = \mathbf{K}_i\mathbf{z}_i[1-\beta_i t_i^2(\theta_2-\theta_3 t_i)]-\mathbf{K}_i^2\mathbf{z}_i[t_i+\gamma_i t_i^2(\theta_2-\theta_3 t_i)]. \tag{6.6}$$

Now equations (3.1)–(3.5) yield, after some tedious manipulation,

$$\gamma_i t_i^2 = (1-\beta_i t_i^2\theta_2)/\theta_3, \tag{6.7}$$

and substituting this value for $\gamma_i t_i^2$ in equation (6.6) gives

$$\mathbf{K}_{i+1}\mathbf{z}_{i+1} = (\mathbf{K}_i\mathbf{z}_i-\mathbf{K}_i^2\mathbf{z}_i\theta_2/\theta_3)\rho_i, \tag{6.8a}$$

where

$$\rho_i = 1-\beta_i t_i^2(\theta_2-\theta_3 t_i). \tag{6.8b}$$

It follows immediately that

$$\mathbf{z}_{i+1}^T\mathbf{K}_{i+1}^2\mathbf{z}_{i+1} = (\theta_4\theta_2^2/\theta_3^2-\theta_2)\rho_i^2. \tag{6.9}$$

We consider now equations (3.7) and (6.1). The only term in equation (6.1) that can give rise to a term of the form $\mathbf{K}_i^2\mathbf{z}_i\mathbf{z}_i^T\mathbf{K}_i^2$ in equation (3.7) is the term $\mathbf{q}_{i+1}\eta_i\mathbf{q}_{i+1}^T$, and since the magnitude of the terms arising from both equations must be identical we obtain, from equations (6.8) and (6.9),

$$-\gamma_i t_i^2 = (\theta_2/\theta_3)^2\eta_i/(\theta_4\theta_2^2/\theta_3^2-\theta_2),$$

which simplifies, from equation (6.7), to

$$\eta_i = k(\beta_i t_i^2\theta_2-1), \tag{6.10a}$$

where

$$k = \left(\frac{\theta_4}{\theta_3}-\frac{\theta_3}{\theta_2}\right). \tag{6.10b}$$

Hence η_i varies linearly with β_i and since $\theta_2 > 0$ and $k \geqslant 0$ (by Cauchy's inequality) it follows that in general η_i increases with β_i.

Theorem 9 indicates that despite the fact that $\mathbf{K}_i$ depends upon $i-1$ arbitrary parameters β_j there is only one arbitrary term in its composition. To see how this can occur we observe that a term of the form $\mathbf{q}_j\mathbf{q}_j^T$ not only occurs when $\mathbf{K}_{j-1}$ is transformed to $\mathbf{K}_j$ but also when $\mathbf{K}_j$ is transformed to $\mathbf{K}_{j+1}$. We note moreover that only such terms are involved in more than one transformation, and that, from

Theorem 9, the amount of the component $\mathbf{q}_j\mathbf{q}_j^T$ that occurs during the *first* transformation is arbitrary. This implies that the amount of this component involved in the second transformation depends upon and in fact compensates for the arbitrary amount involved in the first transformation, a conclusion which is stated explicitly by equation (6.4b). Thus, when minimizing quadratic functions, the effect of an injudicious choice of β_i is entirely eliminated in the next iteration but it cannot reasonably be hoped that this desirable attribute will be retained when the function to be minimized is non-quadratic.

7. Conclusions

Since the preceding analysis refers only to quadratic functions any general inferences drawn must necessarily be speculative. One would expect, however, similar behaviour to that described above for functions which approximate in some sense to quadratic functions, and in general this similarity has been observed experimentally. In particular the result that the sequence of approximate solutions is independent of the choice of the arbitrary parameter β_i seems to hold fairly well for certain non-quadratic functions. Greenstadt (1967) reports that a method he denotes by "variation I", where β_i is chosen so that $\alpha_i = 0$, seems to be competitive with the DFP method with no clear-cut implications as to which is the better. Previous unpublished work of the present author, where various arbitrary choices of β_i were made, also supports this view.

We now consider the evidence of Bard (1968) and Pearson (1968). Bard reported that the iteration matrix $\mathbf{H}_i$ could become singular and Pearson noted that the convergence of the DFP algorithm could sometimes be substantially improved by periodically resetting $\mathbf{H}_i$ to some positive definite matrix. This improvement would be explained by the hypothesis that the poor performance of the DFP algorithm without resetting was due to the matrix $\mathbf{H}_i$ becoming singular or nearly so, giving rise to the situation described by Broyden (1967). The best matrix to which to reset $\mathbf{H}_i$ is probably the inverse Hessian evaluated at that point, but this would require a substantial amount of additional computation. Since, for quadratic functions, the choice of the unit matrix as initial iteration matrix guarantees that the error norms decrease strictly monotonically it would seem that this is probably the best substitute. Now although Bard has suggested that all quasi-Newton algorithms are prone to the iteration matrix becoming singular we wish to advance the hypothesis that some algorithms are more so than others, and that in particular the DFP algorithm would be expected to exhibit this behaviour. The basis for this assertion is given by equation (6.10). Since for the DFP method $\beta_i = 0$ for all i it follows from the fact that k is in general positive that η_i is in general negative for all i, so that $\mathbf{K}_i$ is obtained from $\mathbf{K}_i^*$ by subtracting an unknown positive multiple of $\mathbf{q}_i\mathbf{q}_i^T$. A simple argument based upon the use of Rayleigh quotients then shows that in general the smallest eigenvalue of $\mathbf{K}_i$ is less than that of $\mathbf{K}_i^*$, leading in general to the conclusion that $\mathbf{K}_i$ is more ill-conditioned than $\mathbf{K}_i^*$. It is possible that if the function to be minimized is strongly non-quadratic then not only might $\mathbf{K}_i$ be very nearly singular (although still positive definite), but the missing component might not be restored in the next iteration. In this case it would be unlikely that a rapid restoration of the deficient component would take place unless the

iteration matrix was reset, and we suggest that something of this nature occurred during the course of the computation described by Pearson. Since η_i is essentially arbitrary, and ill-conditioned matrices are best avoided in practical computation, it would appear reasonable to choose η_i having regard to the condition number of $\mathbf{K}_{i+1}$, and although choosing η_i to minimize this condition number would require excessive computation elementary considerations of this nature preclude the possibility of a negative value of η_i. We therefore suggest that a reasonable value of this parameter is in fact zero and note that this gives rise to a new algorithm. The properties of this new algorithm, both theoretical and experimental, are the principal concern of the second part of this paper.

The author is indebted to Dr John Greenstadt for sending him an advance copy of his paper on variable metric methods, and to the referee for his comments and advice.

References

BARD, Y. 1968 *Math. Comp.* **22,** 665–666.
BROYDEN, C. G. 1967 *Math. Comp.* **21,** 368–381.
DAVIDON, W. C. 1959 *Variable metric method for minimization.* A.E.C. Research and Development Report. ANL–5990 (Rev. TID–4500, 14th Edition).
FLETCHER, R. & POWELL, M. J. D. 1963/64 *Comput. J.* **6,** 163–168.
GREENSTADT, JOHN 1967 *Variations of Variable Metric Methods.* IBM New York Sci. Centre, Rep. No. 320–2901
HESTENES, M. R. & STIEFEL, E. 1952 *J. Res. Natn. Bur. Stand.* **49,** 409.
PEARSON, JOHN D. 1968 *On Variable Metric Methods of Minimization.* Research Analysis Corporation Advanced Research Department, Technical Paper RAC–TP–302.

J. Inst. Maths Applics (1970) **6,** 91–95

A Note on the Envelope Construction for Group Velocity in Dispersive, Anisotropic Wave Systems

J. A. SHERCLIFF
University of Warwick, Coventry

[Received 11 February 1969]

1. Introduction

WE ARE CONCERNED here with systems of two-dimensionally propagating waves which are dispersive with respect both to frequency and the orientation of the wave-normal, i.e. the dispersion relation for harmonic waves is of the form

$$f(k,\omega,\psi) = 0 \tag{1}$$

or

$$g(\omega,k_x,k_y) = 0, \tag{2}$$

where k = magnitude of the wave number vector $\mathbf{k}$ (with Cartesian components k_x and k_y), ω = frequency and ψ = inclination of $\mathbf{k}$ to the x-axis, as shown in Fig. 1.

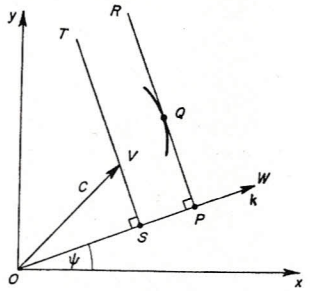

FIG. 1.

It is well-known that lines PR drawn normal to the wave number vector $\mathbf{k}$ at a distance $OP = \omega/k$ (the phase velocity) from the origin, envelope a curve which is, to a suitable scale, a constant-phase line, e.g. a crest or trough, for waves excited continuously at the origin at constant frequency ω. An equivalent statement is that the constant phase curve is the polar reciprocal of the locus of W, the extremity of the wave number vector $\mathbf{k}$, with respect to a circle of radius $\omega^{\frac{1}{2}}$ (Lighthill, 1965).

Also, if the point of tangency is Q, then OQ is in the direction of the group velocity $\mathbf{C}$, but OQ equals the group velocity in magnitude only when the waves are not dispersive with respect to frequency, i.e. when the dispersion relation is of the form $\omega/k = f(\psi)$. Examples of this type are well-known for instance in magneto-acoustics, where the envelope construction is commonly exploited for finding group velocity.