MATHEMATICS OF COMPUTATION, VOLUME 24, NUMBER 110, APRIL, 1970

The Convergence of Single-Rank Quasi-Newton Methods

By C. G. Broyden

Abstract. Analyses of the convergence properties of general quasi-Newton methods are presented, particular attention being paid to how the approximate solutions and the iteration matrices approach their final values. It is further shown that when Broyden's algorithm is applied to linear systems, the error norms are majorised by a superlinearly convergent sequence of an unusual kind.

1. Introduction. During recent years certain new methods [2], [3], [4], [7], [8], [9], [12], have been advanced for solving simultaneous nonlinear equations, and for determining the unconstrained optimum of a function whose first partial derivatives are available as explicit expressions. We write the equations in the form

$$f(x) = 0, \tag{1.1}$$

where f, x and 0 are nth order vectors, and denote $f(x_i)$ by f_i, where x_i is an approximation to the solution of Eq. (1.1). Since the vector function $f(x)$ may be a vector of first partial derivatives of some scalar function, the optimisation problem may in principle be solved in this way.

The methods referred to involve a step vector p_i which is normally computed from the equation

$$p_i = -H_i f_i, \tag{1.2}$$

where H_i is some as yet unspecified nth order matrix. The step vector is then scaled by a factor t_i and the next approximation to the solution, x_{i+1}, is given by

$$x_{i+1} = x_i + p_i t_i. \tag{1.3}$$

The factor t_i is chosen to satisfy certain requirements discussed in Section 3 (below) and the product $p_i t_i$ is often referred to as a step.

Since this product occurs frequently in the analysis, we denote it subsequently by s_i. A further simplification of notation is achieved by observing that much of the discussion, apart from Section 5, is concerned only with changes that occur during a single iteration. We therefore omit the subscript i and replace $i + 1$ by the subscript unity.

We consider now the choice of the matrix H. If this is taken to be the inverse Jacobian evaluated at x, and t is set equal to unity, Eqs. (1.1)-(1.3) define Newton's method. This is probably the best available method if the Jacobian may be evaluated explicitly and a good initial estimate of the solution is known. If, however, the former condition is not satisfied, it is only feasible that H should approximate in some way

Received October 7, 1968, revised September 26, 1969.
AMS Subject Classifications. Primary 6510, 6550.
Key Words and Phrases. Nonlinear algebraic systems, quasi-Newton methods, single-rank methods.

the local inverse Jacobian. In the class of methods under discussion, H_1 has been chosen to satisfy the equation

$$H_1 y = s, \tag{1.4}$$

where

$$y = f_1 - f. \tag{1.5}$$

The motivation underlying this requirement is discussed more fully in [3] and [4].

In all the methods referred to above, the matrix H_1 has been computed from H by adding to the latter a correction matrix of either rank one or rank two. We consider here only the former type of correction, which in its most general form may be written

$$H_1 = H - (Hy - s)q^T/q^T y, \tag{1.6a}$$

where q is an arbitrary nth order vector, subject only to the restriction that

$$q^T y \neq 0. \tag{1.6b}$$

We consider, moreover, applications of algorithms of this type only to the solution of the linear system

$$f(x) \equiv Ax - b, \tag{1.7}$$

where A is an nth order nonsingular matrix and b an nth order vector. Although for most of the algorithms discussed the final form of H is known, less is known of the way in which the successive values of H approach this final form and the successive values of x approach the solution. In particular, we are interested in those algorithms which ensure that some norm of the error e_i, where

$$e_i = x_i - A^{-1}b, \tag{1.8}$$

decreases monotonically, since algorithms of this type have performed extremely well in practice. Finally, we prove a convergence theorem for a modification of Broyden's algorithm [3], when applied to linear systems, and investigate briefly how the behaviour of the modified algorithm differs from that of the original.

2. General Single-Rank Methods. In order to analyse the convergence properties of methods of this type when applied to the linear system (1.7), it is convenient to define two distinct but related matrices, each of which gives some indication of the discrepancy between A^{-1} and the current value of H. We define these matrices, E and R, by

$$E = A^{-1}B - I \tag{2.1a}$$

and

$$R = HA - I, \tag{2.1b}$$

where B is defined to be H^{-1}. Clearly, E and R are null if and only if H is the inverse of A, and they are related by the equation

$$I + E = (I + R)^{-1}. \tag{2.2}$$

Now, it may readily be verified that if B is as defined above, B_1 may be expressed in terms of B by an equation analogous to Eq. (1.6a), namely,

$$B_1 = B - (Bs - y)q^TB/q^TBs. \tag{2.3}$$

Since, for the linear system (1.7), Eqs. (1.3) and (1.5) give

$$y = As, \tag{2.4}$$

some simple manipulation involving Eqs. (1.6a), (2.1) and (2.3) then yields

$$E_1 = E(I - sq^TB/q^TBs) \tag{2.5a}$$

and

$$R_1 = R(I - sq^TA/q^TAs). \tag{2.5b}$$

We note that E_1 and R_1 are independent of the step-length since s appears in both the numerator and denominator of the above equations, that both are singular since

$$E_1s = 0 \tag{2.6a}$$

and

$$R_1s = 0, \tag{2.6b}$$

and that the postmultiplying factors in Eqs. (2.5) are the same, except that the true Jacobian in Eq. (2.5b) is replaced by the current approximation in Eq. (2.5a).

Equations (2.5) enable us to analyse the behaviour of the sequences of matrices $\{E_i\}$ and $\{R_i\}$ for particular choices of the arbitrary vector q. We shall be particularly interested in the distinction between methods in which the rank of E is reduced as the iteration proceeds and those in which some norm of E is reduced.

A further question that must be asked about these methods concerns their stability. It was shown in [4] that it could be disastrous if either H or B should become singular, and this suggests that a desirable feature of these methods is that the condition number of H should be kept as small as possible. Let then $\|\cdot\|$ denote the spectral norm of a matrix, and define the condition number $k(\cdot)$ as the product of the spectral norm of a matrix and that of its inverse. We shall be interested in obtaining bounds for $k(H)$ for the methods under discussion. Our initial theorem, which represents a first step in this direction, obtains bounds for $k(H)$ in terms of $\|E\|$ and $\|R\|$.

THEOREM 1. *Let A, H and H_1 be nonsingular, where H_1 and H are related by Eq.* (1.6a), *and let E and R be defined by Eqs.* (2.5). *If $\phi(E)$, $\phi(R)$ and ψ are defined by*

$$\phi(E) = \max(1, \|E\| - 1), \tag{2.7a}$$

$$\phi(R) = \max(1, \|R\| - 1), \tag{2.7b}$$

and

$$\psi = (1 + \|E\|)(1 + \|R\|), \tag{2.8}$$

then

$$k(A)/\psi \leqq k(H) \leqq \psi k(A) \tag{2.9}$$

and

$$k(H_1) \geqq \phi(E_1)\phi(R_1)/k(A). \tag{2.10}$$

Proof. Rearranging Eqs. (2.1) gives

$$A = H^{-1}(I + R), \qquad A^{-1} = (I + E)H,$$

and taking norms and combining yields the first inequality (2.9). A similar rearrangement of Eqs. (2.1) gives

$$H^{-1} = A(I + E), \qquad H = (I + R)A^{-1},$$

which on taking norms yields the second inequality. A third arrangement of the same equations, with the current values replaced by the updated ones, is

$$I + E_1 = A^{-1}H_1^{-1}, \qquad I + R_1 = H_1A,$$

so that

$$\|I + E_1\| \leqq \|A^{-1}\|\|H_1^{-1}\| \tag{2.11a}$$

$$\|I + R_1\| \leqq \|A\|\|H_1\|. \tag{2.11b}$$

We note though that E_1 is singular, satisfying Eq. (2.6a), so that since

$$\|I + E_1\| \geqq \frac{\|(I + E_1)s\|}{\|s\|},$$

$$\|I + E_1\| \geqq 1. \tag{2.12a}$$

Similarly

$$\|I + R_1\| \geqq 1. \tag{2.12b}$$

The triangle inequality, however, yields

$$\|I + E_1\| \geqq \|E_1\| - 1$$

so that, from inequality (2.12a),

$$\|I + E_1\| \geqq \phi(E_1).$$

Since a similar result holds for R_1, it follows from the inequalities (2.11) that

$$k(A)k(H_1) \geqq \phi(E_1)\phi(R_1),$$

completing the proof.

COROLLARY 1. *$k(H)$ is bounded above if either*

$$\|E\| \leqq e < 1 \tag{a}$$

or

$$\|R\| \leqq r < 1. \tag{b}$$

Proof. (a) Equation (2.2) may be rearranged to give

$$R = -(I + E)^{-1}E, \tag{2.13}$$

so that

$$\|R\| \leqq e/(1 - e).$$

Thus $\psi \leqq (1 + e)/(1 - e)$, and the result follows from Theorem 1. The proof of (b) is similar.

COROLLARY 2. *$k(H_1)$ may be made arbitrarily large by making either $\|E_1\|$ or $\|R_1\|$ sufficiently large.*

Proof. Since $\phi(R_1)$ is bounded below by unity, and $\phi(E_1)$ may be made arbitrarily large, the result follows from inequality (2.10). A similar result holds if R_1 replaces E_1.

Corollary 2 establishes that the boundedness of $\|E_1\|$ and $\|R_1\|$ is necessary as well as sufficient for $k(H_1)$ to be bounded. We note that we *cannot* prove a result similar to Corollary 2, with H_1 replaced by an arbitrary H, since an essential part of the proof is that $\phi(R_1)$ is bounded below, and this follows from the singularity of R_1. This is guaranteed, since H_1 is the result of a single-rank update (see Eq. (2.6b)), but will not in general hold for an arbitrary H. We do note, though, from the theorem, that $k(H)$ is bounded if both $\|E\|$ and $\|R\|$ are bounded, and that this sufficient condition is true for all H.

In order to proceed with the discussion we next prove a simple lemma.

LEMMA 1. *If x, y are two vectors of order n such that $x^Ty = 1$, I is the unit matrix of order n and*

$$\sigma = \|I - xy^T\|,$$

then

$$\sigma = \|x\|\|y\|.$$

Proof. It is well known (see e.g. [11]) that σ^2 is the largest eigenvalue of M, where $M = (I - xy^T)^T(I - xy^T)$. Now M has $n - 2$ eigenvalues equal to unity whose corresponding eigenvectors are any $n - 2$ vectors orthogonal both to one another and to x and y. Since $x^Ty = 1$, one of the two remaining eigenvalues is equal to zero, and its corresponding eigenvector is x. This leaves one eigenvalue, which we shall call λ. Since the sum of the eigenvalues of a matrix is equal to its trace,

$$\lambda + 0 + (n - 2) = \text{tr}(M).$$

However, since $x^Ty = 1$,

$$\text{tr}(M) = n - 2 + \|x\|^2\|y\|^2.$$

Hence $\lambda = \|x\|^2\|y\|^2$, and since by Cauchy's inequality $\|x\|^2\|y\|^2 \geqq (x^Ty)^2$, it follows that $\lambda \geqq 1$. Thus, λ is the largest eigenvalue of M, and it is therefore equal to σ^2. So, $\sigma = \|x\|\|y\|$ and the lemma is proved.

COROLLARY. *The spectral norm of $I - xy^T$ is equal to unity if and only if x is proportional to y.*

Proof. The proof follows immediately from the use of Cauchy's inequality in proving Lemma 1.

We now apply this lemma to Eqs. (2.5) and obtain

$$\|E_1\| \leqq \rho\|E\| \tag{2.15a}$$

$$\|R_1\| \leqq \theta\|R\|, \tag{2.15b}$$

where ρ and θ are defined by

$$\rho = \|s\|\|q^TB\|/|q^TBs| \tag{2.16a}$$

and

$$\theta = \|s\|\|q^TA\|/|q^TAs|. \tag{2.16b}$$

It follows immediately from Cauchy's inequality that $\rho \geqq 1$ and $\theta \geqq 1$ so that we cannot infer from Eqs. (2.15) that any reduction of the spectral norm of the error matrices occurs (although we shall see later that one algorithm reduces the Euclidean norm of E). Since we obtain the best bound for $k(H_1)$ by minimising $\|E_1\|$ and $\|R_1\|$, we are interested in algorithms for which $\rho = 1$ and $\theta = 1$. From the proof of Cauchy's inequality, it follows that $\rho = 1$ if and only if $q^T B = \lambda s^T$, where λ is any nonzero scaling factor, and $\theta = 1$ if and only if $q^T A = \mu s^T$, where μ is also an arbitrary nonzero scaling factor. Since $p = st$, we may obtain the two algorithms satisfying these conditions by choosing $q^T = p^T H$ ($\rho = 1$) and $q^T = p^T A^{-1}$ ($\theta = 1$). We observe that the first of these give Broyden's algorithm and the second gives an algorithm that is not computationally realisable, since the inverse Jacobian is unknown.

We have so far concentrated on the norms of the error matrices, and we turn now to their ranks. Clearly, if the rank can be reduced to zero, we have achieved the inverse Jacobian and are then able to obtain the exact solution of the linear equations. We first prove that, in general, E and R have the same rank.

LEMMA 2. *If $I + E$ and $I + R$ are both nonsingular, then E and R have the same rank.*

Proof. Let x be a vector satisfying the equation $Ex = 0$, but otherwise arbitrary. Then, from Eq. (2.13), $Rx = 0$ so that the rank of R cannot exceed that of E. We can similarly show, since Eq. (2.2) is symmetric in E and R, that the rank of E cannot exceed that of R, proving the lemma.

Our next lemma establishes sufficient conditions to ensure the reduction of the ranks of E and R to zero after at most n steps. We revert, temporarily, to our original use of subscripts.

LEMMA 3. *Let s_i, $i = 0, 1, \ldots, r-1$, be r consecutive steps, where $r \leqq n$, and let E_0 be arbitrary. If the matrices E_{i+1}, $i = 0, 1, \ldots, r-1$, are given by Eq.* (2.5a) *and ρ_i and θ_i, $i = 0, 1, \ldots, r-1$ are bounded, then sufficient conditions for E_r to have rank not exceeding $n - r$, are that*

$$q_i^T B_i s_j = 0, \qquad 0 \leqq j < i \leqq r-1, \tag{2.17a}$$

or

$$q_i^T y_j = 0, \qquad 0 \leqq j < i \leqq r-1. \tag{2.17b}$$

Proof. E_r may, from Eq. (2.5a), be written

$$E_r = E_0 M_r, \tag{2.18a}$$

where

$$M_r = (I - s_0 v_0^T)(I - s_1 v_1^T) \cdots (I - s_{r-1} v_{r-1}^T) \tag{2.18b}$$

and where $v_i^T = q_i^T B_i / q_i^T B_i s_i$. It follows from the definition of v_i^T and from Eq. (2.17a) that

$$v_i^T s_i = 1 \tag{2.19a}$$

and

$$v_i^T s_j = 0, \qquad 0 \leqq j < i \leqq r-1. \tag{2.19b}$$

Now Eqs. (2.19) imply that the vectors $s_j, 0 \leqq j \leqq r-1$, are linearly independent, and it follows immediately from Eqs. (2.18b) and (2.19) that

$$M_r s_j = 0, \qquad 0 \leqq j \leqq r - 1.$$

Hence, from Eq. (2.18a),

$$E_r s_j = 0, \qquad 0 \leqq j \leqq r - 1, \tag{2.20}$$

and the rank of E_r cannot exceed $n - r$, proving the first part of the lemma. To prove the second part, we note that a similar argument, with Eq. (2.5b) replacing Eq. (2.5a), establishes that if

$$q_i^T A s_j = 0, \qquad 0 \leqq j < r \leqq r - 1, \tag{2.21}$$

then the rank of R_r cannot exceed $n - r$. The lemma then follows from Lemma 2, and the equation $y_j = As_j$.

COROLLARY. *If, in Eqs.* (2.17), $r = n$, *then* E_n *is null.*

We note that, in order to prove Lemma 3, it is necessary for the vectors s_j, $0 \leqq j \leqq r - 1$ to be linearly independent. We also note though that this is the only property that we require of them; in particular, we do not require that they are computed using Eq. (1.2) and $s_i = p_i t_i$.

This completes the general discussion of the properties of the error matrices. Before we apply these results to particular algorithms, we discuss the choice of the step-length parameter t.

3. The Choice of t. This parameter is normally chosen when solving a general set of nonlinear equations in one of three ways:

(1) If f is the vector of first partial derivatives of a scalar function F, it is chosen to minimise F along p, resulting in the relationship

$$f_1^T p = 0. \tag{3.1}$$

(2) It is chosen to minimise or reduce $\| f \|$.

(3) It is put equal to unity.

The reasons for these choices are as follows. Choice (1), which is only appropriate to minimisation problems, has two main justifications. The first, which is valid for any F, is that it gives the "best" improvement at each step, it being implicitly assumed that this is in itself desirable. The second justification concerns the case when F is quadratic, so that $f(x)$ is given by

$$f(x) \equiv Ax - b, \tag{3.2}$$

and where A is now assumed to be symmetric and positive definite. For the algorithms developed by Fletcher and Powell [8], Broyden [4], and Pearson [10], and where Eq. (3.2) applies, it is necessary to choose t to satisfy Eq. (3.1), in order to obtain n-step convergence (it being again assumed that this is in itself desirable).

Choice (2) is made in order to obtain a monotonically decreasing sequence of norms of f, in order to control the convergence of the algorithm when applied to strongly nonlinear systems. Unfortunately, this control may seriously inhibit convergence for at least one algorithm (see [5] and Section 5 below). It was found for this algorithm that choice (3) was far superior provided that good initial approximations both to the solution and to the Jacobian could be found. We note also, that setting $t = 1$ gives the closest resemblance to Newton's method. Because of this, and because choosing t to be unity has been proposed for at least two other algorithms, [2], [7],

we examine more closely the effects of this choice. We assume that p is computed using Eq. (1.2), and that f is given by Eq. (3.2), where A is now only assumed to be nonsingular.

It follows from Eqs. (1.2) and (1.3) that

$$x_1 = x - Hf. \tag{3.3}$$

Define now the vector e by

$$e = x - A^{-1}b, \tag{3.4}$$

where A and b relate to Eq. (3.2). Then, from Eqs. (3.2)–(3.4),

$$e_1 = (I - HA)e,$$

which becomes, from Eqs. (2.1b) and (2.2),

$$e_1 = -Re \tag{3.5a}$$

or

$$e_1 = (I + E)^{-1}Ee. \tag{3.5b}$$

Since s is given by $s = x_1 - x$, it follows from Eq. (3.4) that $s = e_1 - e$ so that, from Eq. (3.5b),

$$Es = -e_1 \tag{3.6a}$$

and from Eq. (3.5a),

$$Rs = (I + R)e_1. \tag{3.6b}$$

We now obtain lower bounds for $\|E_1\|$ and $\|R_1\|$ in terms of ρ and θ. Eqs. (2.5) may be written

$$E_1 = E - Esq^TB/q^TBs, \tag{3.7a}$$

$$R_1 = R - Rsq^TA/q^TAs, \tag{3.7b}$$

so that, since $\|xy^T\| = \|x\|\|y\|$,

$$\|E_1\| \geqq \|Es\|\|q^TB\|/|q^TBs| - \|E\|, \tag{3.8a}$$

$$\|R_1\| \geqq \|Rs\|\|q^TA\|/|q^TAs| - \|R\|. \tag{3.8b}$$

Hence, from Eqs. (2.16) and (3.6),

$$\|E_1\| \geqq \rho\|e_1\|/\|s\| - \|E\| \tag{3.9a}$$

$$\|R_1\| \geqq \theta\|(I + R)e_1\|/\|s\| - \|R\|. \tag{3.9b}$$

These results, together with Theorem 1, show that, if e_1 is not null, then it is not only sufficient but also necessary for both ρ and θ to be bounded, in order that $k(H_1)$ should be bounded. The stability of a particular algorithm is thus critically dependent upon the magnitude of these parameters, and since they may be readily expressed in terms of known quantities, they afford a simple means of analysing the stability properties of any given algorithm.

We now examine the dependence of the vector error e_1 upon E and R. Writing ε for $\|e\|$ and σ for $\|E\|$ and taking norms of Eq. (3.5) gives

(3.10a) $$\varepsilon_1 \leqq \|R\|\varepsilon$$

and, if $\sigma < 1$,

(3.10b) $$\varepsilon_1 \leqq \sigma\varepsilon/(1 - \sigma).$$

The bound on $\varepsilon_1/\varepsilon$ thus increases monotonically with $\|R\|$, and with σ for $0 \leqq \sigma < 1$.

It follows from Eqs. (3.10) that, if the norms of E and R increase substantially, it is possible that the norms of subsequent vector errors may increase, and although this may not matter if the equations are linear, we believe that it could be serious in the nonlinear case since it may aggravate the effects of any nonlinearities present. It is, of course, always possible to prevent a large increase in the vector error norm by enforced residual norm reduction (choice (2) for t), but this can itself, as has been stated, lead to slow convergence.

A further difficulty caused by permitting the norms of the matrix errors to increase is that, particularly in n-step methods, it is possible for large but cancelling corrections to be added to the iteration matrix with the result that after n steps this latter may be the difference of two or more much "larger" matrices. This cancellation may result in excessive rounding errors, as has been noted by Bard [1].

For these reasons we believe that, if t is taken to be unity, it is desirable that the norms of the matrix errors should be kept as small as possible, even if this means sacrificing n-step convergence for linear systems. Perhaps the best way of achieving this ideal would be, from Eq. (3.10a), to choose an algorithm for which $\|R_1\| \leqq \|R\|$ but unfortunately such an algorithm, as was seen in the previous section, is not computationally realisable. Possibly the closest practical approximation is not to permit σ to increase and this, as has been seen, is in fact possible.

4. Particular Methods. We consider now particular methods, where x is given by

(4.1) $$x_1 = x - Hf,$$

that is, where p is given by Eq. (1.2), and t is taken to be unity. The results, regarding the matrix error norm σ derived in the previous section, are thus applicable. We note, however, that the restriction on the form s_i imposed by Eq. (4.1) is not necessary to establish rank-reduction.

A. *The Secant Method.* In this method, [2], [12], q_i is specifically chosen to satisfy

(4.2) $$q_i^T y_j = 0, \qquad 0 \leqq j < i \leqq n - 1.$$

The algorithm is thus, from Lemma 3, rank-reducing when applied to linear equations.

We note that the directions of the vectors q_j, with the exception of $j = n - 1$, are not uniquely determined, and it is thus possible for either ρ_i or θ_i to become infinite. If, however, q_i is chosen to be the projection of y_i orthogonal to y_j, $j = 0, 1, \ldots, i - 1$, one of these possibilities does not arise. For, if we define

(4.3) $$Y_r = [y_0, y_1, \ldots, y_{r-1}], \qquad r = 1, 2, \ldots, n - 1,$$

(4.4a) $$P_0 = I,$$

(4.4b) $$P_r = I - Y_r(Y_r^T Y_r)^{-1} Y_r^T, \qquad r = 1, 2, \ldots, n - 1,$$

then P_r is symmetric,

$$P_r^2 = P_r \tag{4.5}$$

and q_r is given by

$$q_r = P_r y_r. \tag{4.6}$$

It follows from Eqs. (2.4) and (2.16b) that θ_r is infinite only if $q_r^T y_r = 0$ and $\|q_r\| > 0$. But since P_r is both symmetric and idempotent, $q_r^T y_r = 0$ only if $y_r^T P_r^T P_r y_r = 0$, i.e., only if $P_r y_r = 0$. But, this, from Eq. (4.6), implies that $\|q_r\| = 0$, so that θ_r is either finite or not defined. The latter case occurs, from Eq. (4.4b) if and only if

$$y_r = \sum_{j=0}^{r-1} y_j \alpha_j, \tag{4.7}$$

where the α_j's are constants whose precise values are irrelevant to this analysis. Premultiplication of Eq. (4.7) by A^{-1} then gives, since $y_j = As_j$,

$$s_r = \sum_{j=0}^{r-1} s_j \alpha_j. \tag{4.8}$$

Now, by hypothesis $q_i^T y_j = 0$, $0 \leqq j < i \leqq r-1$ so that, from Eqs. (2.20) and (4.8), $E_r s_r = 0$, and hence, from Eq. (3.6a), $e_{r+1} = 0$. It follows that q_r is null and H_{r+1} is not defined only if e_{r+1} is null and no further iterations are necessary.

Although this analysis shows that the ultimate catastrophe of a division by zero cannot occur, it is still possible for H to become arbitrarily badly conditioned in the course of the iteration. For, from Eq. (2.16b) it follows that if q is given by Eq. (4.6) then, since $y = As$,

$$\theta = \|A^{-1}y\| \, \|q^T A\| / y^T P^2 y$$

or

$$\theta = \|A^{-1}y\| \, \|q^T A\| / \|q\|^2,$$

and standard norm inequalities then yield

$$\|y\| / (k(A)\|q\|) \leqq \theta \leqq k(A)\|y\| / \|q\|. \tag{4.9}$$

Since after the first iteration P is singular, we infer from Eqs. (4.6) and (4.9) that $\|y\|/\|q\|$, and hence θ, may become arbitrarily large, so that, if e_1 is not null, $k(H_1)$ may also become arbitrarily large.

For nonlinear equations, i usually exceeds $n-1$, and q_i is then chosen to satisfy

$$q_i^T y_j = 0, \qquad i - n + 1 \leqq j \leqq i - 1. \tag{4.10}$$

This choice of q is also apparently not entirely satisfactory, since again neither ρ nor θ are bounded. Barnes [2] suggests carrying out the matrix update only if ρ, as defined by Eq. (2.16a), is less than 10^4. This would imply in the linear case that $\sigma_1 \leqq 10^4 \sigma$, and it follows from Theorem 1, Corollary 1 that $k(H_1)$ is bounded if $\sigma < 10^{-4}$. It follows, moreover, from Eq. (2.3) that this restriction upon ρ implies the existence, but not the nonsingularity, of the sequence of approximate Jacobians $\{B_i\}$, but it also follows that massive increases in σ are possible and this may make the method unstable.

B. *The Symmetric Method.* This method has been outlined by Broyden [4] and by Murtagh and Sargent [9]. The particular form where t is unity has been discussed

by Davidon [7]. It is only used if the Jacobian is known to be symmetric, e.g., if f is a vector of first partial derivatives, but despite its proposed application to the minimisation problem, it differs fundamentally from the Davidon-Fletcher-Powell and Pearson [10] algorithms in that it is not necessary to satisfy Eq. (3.1), in order to obtain n-step convergence when f is linear. Indeed, as we shall see, if this equation is satisfied by the symmetric algorithm with $t = 1$, then immediate breakdown occurs.

The vectors q_i for the symmetric algorithm are chosen so that H_i is symmetric for all i, and this gives the unique value of q to be

$$q = Hy - s. \tag{4.11}$$

But $y = As$, so that q may be written, from Eq. (2.1b),

$$q = Rs. \tag{4.12}$$

We first establish that the algorithm is rank-reducing. Eq. (2.1b) gives, since $B = H^{-1}$,

$$BR = A - B. \tag{4.13}$$

But since both A and B are symmetric, it follows that $BR = R^TB$, so that from Eq. (4.12),

$$q^TB = s^T(A - B). \tag{4.14}$$

This becomes, from Eq. (2.1a),

$$q_r^TB_r = -s_r^TAE_r, \tag{4.15}$$

(reverting temporarily to the original use of subscripts). We now proceed by induction. Assume that

$$q_i^TB_is_j = 0, \qquad 0 \leqq j < i \leqq r - 1. \tag{4.16}$$

Then, from Eq. (2.20), $E_rs_j = 0$, $0 \leqq j \leqq r - 1$ so that, from Eq. (4.15)

$$q_r^TB_rs_j = 0, \qquad 0 \leqq j \leqq r - 1, \tag{4.17}$$

and Eq. (4.16) now holds with r replacing $r - 1$. The induction is initiated by proving that $q_1^TB_1s_0 = 0$, and this follows immediately from Eqs. (2.6a) and (4.15). That the algorithm is rank-reducing now follows immediately from Lemma 3. We note that this result does not require that $s = -Hf$, but only that the s_j's are linearly independent.

The stability properties of the symmetric method with $t = 1$ are obtained from Eqs. (1.2), (1.5) and (4.11) which yield, since $s = p$,

$$q = Hf_1. \tag{4.18}$$

The values of ρ and θ are then given, from Eqs. (2.1b), (2.16) and the symmetry of H, by

$$\rho = \|s\|\,\|f_1\|/|s^Tf_1| \tag{4.19a}$$

and

$$\theta = \|s\|\,\|f_1^T(I + R)\|/|f_1^T(I + R)s|. \tag{4.19b}$$

It is thus possible in the symmetric method, as in the secant method, for either ρ or θ

to become arbitrarily large. Indeed, if Eq. (3.1) is satisfied, then, from Eq. (4.19a), ρ becomes infinite and immediate breakdown occurs. This result is somewhat bizarre, since it is often considered desirable that Eq. (3.1) should be satisfied when minimising functions, and it is only for function minimisation that the symmetric algorithm has been seriously suggested.

When using the symmetric algorithms for general function minimisation, Davidon [7] suggests that the matrix update be carried out only if certain conditions are satisfied. After considerable simplification, these conditions are seen to be

$$\gamma^{-1} \leqq (1 - \alpha/\alpha) \quad \text{if } \gamma > 0, \tag{4.20a}$$

$$|\gamma^{-1}| \leqq (\beta - 1/\beta) \quad \text{if } \gamma < 0, \tag{4.20b}$$

where

$$\gamma = s^T f_1 / f_1^T H f_1 \tag{4.21}$$

and the arbitrary constants α and β satisfy

$$0 < \alpha < 1 < \beta. \tag{4.22}$$

Now Eqs. (2.1b) and (3.9b) give $Rs = HAe_1$, and, since $f_1 = Ae_1$, it follows from Eqs. (4.19) and (4.21) that

$$\rho = \|s\| \, \|f_1\| \phi / |\gamma| \tag{4.23a}$$

and

$$\theta = \|s\| \, \|f_1^T (I + R)\| \phi / |1 + \gamma|, \tag{4.23b}$$

where

$$\phi^{-1} = |f_1^T H f_1|. \tag{4.23c}$$

Davidon suggested that reasonable values of α and β might be $\alpha = 10^{-3}$ and $\beta = 10$, and with these values we obtain bounds for ρ and θ to be, from Eqs. (4.23),

$$\rho \leqq 999.0 \|s\| \, \|f_1\| \phi \tag{4.24a}$$

$$\theta \leqq 0.999 \|s\| \, \|f_1^T (I + R)\| \phi \qquad \gamma > 0, \tag{4.24b}$$

$$\rho \leqq 0.9 \|s\| \, \|f_1\| \phi \tag{4.24c}$$

$$\theta \leqq 9.0 \|s\| \, \|f_1^T (I + R)\| \phi \qquad \gamma < 0. \tag{4.24d}$$

It follows that, even if ϕ is bounded above, substantial increases in $\|E\|$ are possible for $\gamma > 0$. Since in general it is not possible to guarantee that H is positive definite for the symmetric method, it follows from Eq. (4.23c) that no such bound on ϕ exists, so that if e_1 is not null an indefinite worsening of $k(H_1)$ is possible.

C. *Broyden's Method.* In this method, [3], [5], q is given by

$$q^T = p^T H, \tag{4.25}$$

and it follows immediately from Lemma 3 that the algorithm is rank-reducing if $p_i^T p_j = 0$, $i > j$. This enables a good approximation to the local inverse Jacobian to be obtained by taking steps along the coordinate axes, although there then exists the possibility of instability if the initial matrix differs too much from the final approxi-

mation. If, however, p_i is computed by Eq. (1.2), the algorithm is not rank-reducing. The values of ρ and θ are, from Eqs. (2.1b), (2.16), and (4.25),

$$\rho = 1 \tag{4.26a}$$

and

$$\theta = \|s\| \, \|s^T(I + R)\|/|s^T(I + R)s|, \tag{4.26b}$$

and it follows immediately that the algorithm is completely stable if $\|R\| < 1$ or, from Theorem 1, Corollary 1, $\|E\| < 1$, since $\|E_1\| \leqq \|E\|$.

Of the three methods discussed, we see that if $p = -Hf$ and $t = 1$, then the first two methods are rank-reducing and the third is not. Thus Broyden's algorithm will not solve linear systems in a finite number of steps and this has been held to be a disadvantage of the method. On the other hand, the basic forms of the other algorithms have also been found wanting [2], [7], and the effect of the proposed modifications is to weaken their rank-reducing character. A disturbing feature of these algorithms is the fact that these modifications may be necessary, no matter how small $\|E\|$ may become. This means that, in their basic forms, these algorithms can convert a good approximation to the inverse Jacobian into a poor one, with all the implications discussed in Section 3, and this behaviour is in marked contrast to Broyden's algorithm which becomes more stable as $\|E\|$ tends to zero. Whether or not their rank-reducing character is more effective in solving nonlinear systems than the norm-reducing and asymptotically stable character of Broyden's method, is at the moment open to conjecture. We look to extensive numerical experiment and more sophisticated theory to provide the ultimate resolution of this question.

5. Further Properties of Broyden's Method. In the method as proposed by Broyden [3], the step length parameter t was chosen to satisfy the second condition of Section 3 (above). We first prove a convergence theorem for the modification of this method given by choosing $t = 1$, and then discuss less formally in Section 6 the effect of choosing t to reduce $\|f\|$. Before proving the theorem though, we prove the following lemma.

LEMMA 4. *If E is an $n \times n$ matrix and x, y are two vectors of order n, then*

$$\|E + xy^T\|_E^2 = \|E\|_E^2 + 2y^TE^Tx + \|x\|^2\|y\|^2,$$

where the subscript E denotes the Euclidean matrix norm.

Proof.

$$\begin{aligned}
\|E + xy^T\|_E^2 &= \operatorname{tr}[(E + xy^T)^T(E + xy^T)] \\
&= \operatorname{tr}(E^TE + E^Txy^T + yx^TE + yx^Txy^T) \\
&= \operatorname{tr}(E^TE) + \operatorname{tr}(E^Txy^T) + \operatorname{tr}(yx^TE) + \operatorname{tr}(yx^Txy^T) \\
&= \|E\|_E^2 + 2y^TE^Tx + y^Tyx^Tx \\
&= \|E\|_E^2 + 2y^TE^Tx + \|y\|^2\|x\|^2.
\end{aligned}$$

THEOREM 2. *If Broyden's algorithm is applied to the linear system* (3.2), $p_i = -H_if_i$, *t_i is unity and $\sigma_0 < 1$, then the Euclidean norm ε_r of the rth error vector satisfies the inequality*

$$\varepsilon_r \leqq (k/r^{1/2})^r \varepsilon_0, \tag{5.1}$$

where k is a constant depending upon E_0.

Proof. If the value of q_i given by Eq. (4.25) is substituted in Eq. (2.5a), we obtain

$$E_{i+1} = E_i - E_i p_i p_i^T / p_i^T p_i, \tag{5.2}$$

so that, from Lemma 4,

$$\phi_{i+1}^2 = \phi_i^2 - \|E_i p_i\|^2 / \|p_i\|^2, \tag{5.3}$$

where ϕ_i denotes the Euclidean norm of E_i. Now, Eqs. (1.3) and (1.8) yield, since $t_i = 1$,

$$p_i = e_{i+1} - e_i,$$

and Eq. (3.5b) may be rearranged to give

$$e_{i+1} - e_i = -(I + E_i)^{-1} e_i,$$

so that, if $\sigma_i < 1$,

$$\|p_i\| \leqq \varepsilon_i / (1 - \sigma_i). \tag{5.4}$$

Since $s_i = p_i$ Eq. (3.6a) may be written

$$E_i p_i = -e_{i+1} \tag{5.5}$$

and combining Eqs. (5.4) and (5.5) with (5.3), finally yields the inequality

$$\phi_{i+1}^2 \leqq \phi_i^2 - \varepsilon_{i+1}^2 (1 - \sigma_i)^2 / \varepsilon_i^2. \tag{5.6}$$

We now consider the vector error norms. Since $\sigma_i \leqq \phi_i$ (see e.g. [11]), it follows from inequality (3.10b) that

$$\varepsilon_{i+1}^2 \leqq \phi_i^2 \varepsilon_i^2 / (1 - \sigma_i)^2, \tag{5.7}$$

and we convert this inequality into an equation by introducing a parameter θ_i such that

$$\varepsilon_{i+1}^2 = \theta_i \phi_i^2 \varepsilon_i^2 / (1 - \sigma_i)^2. \tag{5.8}$$

It follows immediately that

$$0 \leqq \theta_i \leqq 1. \tag{5.9}$$

Combining Eqs. (5.6) and (5.8) gives

$$\phi_{i+1}^2 \leqq (1 - \theta_i) \phi_i^2,$$

so that

$$\phi_r^2 \leqq \phi_0^2 \prod_{j=1}^{r} (1 - \theta_{r-j}), \qquad r \geqq 1. \tag{5.10}$$

Now the values of σ_i decrease monotonically, so that $\sigma_i \leqq \sigma_0$, $i \geqq 1$ and hence, from Eq. (5.8),

$$\varepsilon_{i+1}^2 \leqq \theta_i \phi_i^2 \varepsilon_i^2 / (1 - \sigma_0)^2. \tag{5.11}$$

We now show that

$$\varepsilon_r^2 \leqq k^{2r}\varepsilon_0^2 \prod_{j=1}^{r} P_j(\theta_{r-j}), \tag{5.12}$$

where

$$k = \phi_0/(1 - \sigma_0)$$

and

$$P_j(\theta) \equiv \theta(1 - \theta)^{j-1}. \tag{5.13}$$

The proof is by induction. Assume that Eq. (5.12) holds for r. Then from Eqs. (5.10)–(5.12) it follows that

$$\varepsilon_{r+1}^2 \leqq k^{2(r+1)}\varepsilon_0^2\theta_r \prod_{j=1}^{r} [(1 - \theta_{r-j})P_j(\theta_{r-j})].$$

This equation may be written, from Eq. (5.13),

$$\varepsilon_{r+1}^2 \leqq k^{2(r+1)}\varepsilon_0^2 P_1(\theta_r) \prod_{j=1}^{r} P_{j+1}(\theta_{r-j})$$

and, on putting $i = j + 1$, it becomes

$$\varepsilon_{r+1}^2 \leqq k^{2(r+1)}\varepsilon_0^2 \prod_{i=2}^{r+1} P_i(\theta_{r+1-i}),$$

being thus Eq. (5.12), with r replaced by $r + 1$. It suffices now to show that Eq. (5.12) is true for $r = 1$, and this is done merely by setting $i = 0$ in Eq. (5.11).

It now remains to obtain an upper bound for the product of the polynomials appearing in Eq. (5.12). We do this by assuming that the θ_{r-j} are independent and finding the maximum value of each factor in the product, subject only to the condition, given by Eq. (5.9), that $0 \leqq \theta \leqq 1$. On examining Eq. (5.13), we see that $P_j(\theta)$ has a simple zero at $\theta = 0$ and a zero of order $j - 1$ at $\theta = 1$. Differentiation gives a maximum at $\theta = 1/j$, $j \geqq 2$, for which the value of $P_j(\theta)$ is $(j - 1)^{j-1}/j^j$. The maximum value of $P_1(\theta)$, $0 \leqq \theta \leqq 1$, is clearly unity. Substituting these maximum values in Eq. (5.12) gives

$$\varepsilon_r^2 \leqq k^{2r}\varepsilon_0^2/r^r$$

and, taking square roots, yields the theorem.

We note the following features of the algorithm that emerge from the above proof. From Eq. (3.10b), we see that strictly monotonic convergence, i.e., $\varepsilon_{i+1} < \varepsilon_i$ for all $i \geqq 0$, is guaranteed provided that $\sigma_0 < \frac{1}{2}$. The algorithm, however, converges if $\sigma_0 < 1$. Although the vector error may become arbitrarily small, there is no guarantee that H_i will tend to A^{-1}, since it is possible that an exact solution may be obtained, no matter how large σ_i might be. This is reinforced by Eq. (5.6), which shows that a minimal improvement in the vector error is associated with a maximal improvement in the matrix error, and conversely.

If, on the other hand, H does approach A^{-1}, so that σ and $\|R\|$ are small, we may infer that the correction added to H is small in norm, from the fact that $\sigma_1 \leqq \sigma$. A

precise bound may be obtained simply as follows. If we write $H_1 = H + C$, it follows from Eqs. (1.6a), (4.25) and $s = p$ that

$$C = (Hy - s)s^T H/s^T Hy. \tag{5.14}$$

This becomes, from Eqs. (2.1b) and (2.4)

$$C = Rss^T H/s^T(I + R)s,$$

so that, if $\|R\| < 1$,

$$\|C\|/\|H\| \leqq \|R\|/(1 - \|R\|). \tag{5.15}$$

This result shows that, as H approaches A^{-1}, the corrections to H become relatively small. The massive cancellations described by Bard [1] thus cannot occur, and we also infer that errors in the correction term due to rounding have no appreciable effect. We would thus expect this particular matrix update to be extremely stable as the solution is approached, and this expectation has been abundantly realised in practice.

Although no tests have been carried out with the algorithm on linear systems, the convergence observed during the final stages of solving a nonlinear problem, when the Jacobian may be regarded as being substantially constant, has certainly been in accordance with Eq. (5.1). Further discussion of the convergence of the algorithm, with results of numerical experiments, may be found in [5].

6. Some Effects of Enforced Residual Norm Reduction. The results we obtain in this section are valid for all algorithms for which the step direction p is given by $p = -Hf$. They are thus applicable to all three algorithms discussed in Section 4 (above).

It follows immediately from Eq. (3.10b) that, if $t = 1$, a sufficient condition for $\varepsilon_1 < \varepsilon$ is that $\sigma < \frac{1}{2}$. To obtain a similar result for $\|f\|$, we note from Eqs. (3.2), (3.4) and (3.5b) that

$$f_1 = AE(I + E)^{-1}A^{-1}f, \tag{6.1}$$

so that, if $\sigma < 1$,

$$\|f_1\| \leqq k(A)\sigma\|f\|/(1 - \sigma). \tag{6.2}$$

It follows that $\|f_1\| < \|f\|$, with $t = 1$, if

$$\sigma < 1/(1 + k(A)). \tag{6.3}$$

This is clearly, since $k(A) \geqq 1$, a greater restriction than $\sigma < \frac{1}{2}$, to which it reduces in the optimum case when $k(A) = 1$.

Despite the fact that Eq. (6.3) is not a necessary condition, it is possible to construct examples for which

$$1/(1 + k(A)) < \sigma < \tfrac{1}{2}, \tag{6.4}$$

and for which the step with $t = 1$ reduces $\|e\|$ but increases $\|f\|$. Thus, in these cases, enforced residual norm reduction has a positive effect, despite the fact that without its use the vector error norm is reduced. To analyse this effect, we consider the function $f(t)$, where

$$f(t) = Ax - b, \tag{6.5a}$$

$$x = x_i - H_i f_i t \tag{6.5b}$$

and A is nonsingular (to avoid confusion, we return to the use of subscripts to denote particular values of the variables). The function $\|f(t)\|^2$ is thus a strictly convex quadratic function of t, and, in consequence, possesses a unique minimum. Now Eqs. (6.5) give

$$f(t) \equiv (I - tAH_i)f_i, \tag{6.6}$$

so that

$$\frac{d\|f(t)\|^2}{dt} = 2(t f_i^T H_i^T A^T A H_i f_i - f_i^T A H_i f_i). \tag{6.7}$$

Thus, from Eq. (2.1b)

$$[d\|f(t)\|^2/dt]_{t=0} = -2f_i^T(I + AR_iA^{-1})f_i, \tag{6.8}$$

and if

$$\|AR_iA^{-1}\| < 1, \tag{6.9}$$

it follows that $[d\|f(t)\|^2/dt]_{t=0}$ will be negative. Hence, it is possible to reduce or minimise $\|f\|$ by choosing a positive value of t if

$$\|R_i\| < 1/k(A), \tag{6.10}$$

since this implies inequality (6.9).

To analyse the behaviour of the vector errors, we note that $e(t) = A^{-1}f(t)$ so that, from Eqs. (2.1b) and (6.6),

$$e(t) = [I - t(I + R_i)]e_i. \tag{6.11}$$

A similar calculation to that for $f(t)$ then yields

$$[d\|e(t)\|^2/dt]_{t=0} = -2e_i^T(I + R_i)e_i, \tag{6.12}$$

so that a sufficient condition for a positive value of t to reduce or minimise $\|e(t)\|$ is that

$$\|R_i\| < 1. \tag{6.13}$$

Now, if Eq. (6.10) is satisfied, then so is Eq. (6.13), and enforced residual norm reduction does in fact reduce the error norms, although perhaps by not as much as would be achieved by letting $t = 1$. It is, however, possible for Eq. (6.13) to be satisfied but not Eq. (6.10). Now (6.10) is a sufficient condition, but examples may be constructed, where

$$1/k(A) < \|R_i\| < 1, \tag{6.14}$$

and where $[d\|f(t)\|^2/dt]_{t=0}$ is positive.

Thus, in order to reduce or minimise $\|f\|$ in this case, a negative value of t must be chosen, but since $\|R_i\| < 1$, this implies, since $\|e(t)\|$ is strictly convex, that $\|e\|$ must *increase*. Under these circumstances, one would expect enforced residual norm reduction to inhibit convergence, and experimental evidence reported in [5] shows that this does indeed occur when using Broyden's algorithm. Neither must it be assumed

that this phenomenon can only occur if H_i is a very poor approximation to A^{-1}, for this condition is specifically excluded by the requirement that $\|R_i\| < 1$. We are thus faced with the dilemma that a technique used to prevent divergence when a long way from the solution may inhibit convergence when close to it, and there is no obvious way of detecting the transition from the one state to the other. It is for this reason that continuation methods, which go back to Davidenko [6] and earlier, may well assume a more prominent position in the array of techniques for solving nonlinear simultaneous equations, since these methods involve solving a sequence of problems, where a good initial estimate of the solution is available for each problem in the sequence.

The author is grateful to Mr. B. L. Meek of Queen Elizabeth College, London, and to the referee, for their helpful comments and criticism.

University of Essex
Computing Centre
Wivenhoe Park, Colchester, Essex
England

1. YONATHAN BARD, "On a numerical instability of Davidon-like methods," *Math. Comp.*, v. 22, 1968, pp. 665–666. MR **38** #858.
2. J. G. P. BARNES, "An algorithm for solving non-linear equations based on the secant method," *Comput. J.*, v. 8, 1965, pp. 66–72. MR **31** #5330.
3. C. G. BROYDEN, "A class of methods for solving nonlinear simultaneous equations," *Math. Comp.*, v. 19, 1965, pp. 577–593. MR **33** #6825.
4. C. G. BROYDEN, "Quasi-Newton methods and their application to function minimisation," *Math. Comp.*, v. 21, 1967, pp. 368–381. MR **36** #7317.
5. C. G. BROYDEN, "A new method of solving nonlinear simultaneous equations," *Comput. J.*, v. 12, 1969, pp. 95–100.
6. D. F. DAVIDENKO, "On a new method of numerical solution of systems of nonlinear equations," *Dokl. Akad. Nauk SSSR*, v. 88, 1953, pp. 601–602. (Russian) MR **14**, 906.
7. W. C. DAVIDON, "Variance algorithm for minimization," *Comput. J.*, v. 10, 1967/68, pp. 406–410. MR **36** #4790.
8. R. FLETCHER & M. J. D. POWELL, "A rapidly convergent descent method for minimization," *Comput. J.*, v. 6, 1963/64, pp. 163–168. MR **27** #2096.
9. B. A. MURTAGH & R. W. H. SARGENT, *A Constrained Minimization Method with Quadratic Convergence*, Chapter 14, "Optimization" (Edited by R. Fletcher), Academic Press, London & New York, 1969.
10. JOHN D. PEARSON, "On variable metric methods of minimization," *Comput. J.*, v. 12, 1969, pp. 171–178.
11. J. H. WILKINSON, *The Algebraic Eigenvalue Problem*, Clarendon Press, Oxford, 1965. MR **32** #1894.
12. PHILIP WOLFE, "The secant method for solving nonlinear equations," *Comm. Assoc. Comput. Mach.*, v. 2, 1959, pp. 12–13.