MATHEMATICS OF COMPUTATION, VOLUME 25, NUMBER 114, APRIL, 1971

The Convergence of an Algorithm for Solving Sparse Nonlinear Systems

By C. G. Broyden

Abstract. A new algorithm for solving systems of nonlinear equations where the Jacobian is known to be sparse is shown to converge locally if a sufficiently good initial estimate of the solution is available and if the Jacobian satisfies a Lipschitz condition. The results of numerical experiments are quoted in which systems of up to 600 equations have been solved by the method.

1. Introduction. As more efficient techniques [7], [8] are developed for handling large matrices of random sparseness, it is inevitable that attention will eventually turn to the problem of solving large sparse nonlinear systems. We regard a nonlinear algebraic system as sparse if a large proportion of the elements of its Jacobian are zero over the domain in which we are interested. Indeed, problems of this type have already been solved [4], [9], using Newton's method, but since, for certain problems, Newton's method is inferior to methods of quasi-Newton type [1], [2], it is natural to attempt to modify these latter methods to enable them to cope with sparse systems.

A drawback becomes immediately evident when considering possible modifications to Broyden's method [1]. One of the advantages possessed by the latter is that an approximation to the inverse Jacobian is used, thereby making it unnecessary to solve a set of linear equations at every iteration. This device cannot readily be retained when solving sparse systems since the inverse of a sparse matrix is in general not sparse, and the consequent benefits would then be destroyed. A major advantage of Broyden's method thus disappears at one stroke. Neither is it possible to use the equivalent Broyden update on the Jacobian since this consists of adding a single-rank matrix to the old approximate Jacobian to form the new one. In general, this single-rank correction is not sparse so that again the advantages conferred by sparseness would be lost. It is possible, however, to modify this single-rank correction so that it is not only sparse but also enables the new approximate Jacobian to satisfy the quasi-Newton equation, and this has been done by Schubert [14]. The properties of this modified algorithm, both theoretical and experimental, form the subject of this paper.

In Section 2, a device for specifying the positions of the nonzero elements of the Jacobian is established and a formal description of the new algorithm is given. Section 3 is devoted to providing local convergence proofs under the assumption that the Jacobian satisfies a Lipschitz condition of order one while in Section 4 the results of certain numerical experiments are discussed. An overall assessment of the algorithm appears in the final section.

Received March 30, 1970, revised September 14, 1970.
AMS 1969 *subject classifications.* Primary 6510, 6550.
Key words and phrases. Sparse, algebraic, nonlinear, systems.

2. The New Algorithm. A prerequisite for the discussion of algorithms specific to sparse systems is a device for indicating precisely which elements of the Jacobian matrix are known to be always zero. Of the many available, we choose one that uses a set of diagonal matrices S_j, $j = 1, 2, \cdots, n$, where the kth diagonal element of S_j is zero if it is known that $\partial f_i/\partial x_k = 0$, and unity otherwise. It follows that if $J(x)$ denotes the Jacobian matrix [6] evaluated at x, and if u_i is the jth column of the unit matrix of order n, then

$$u_i^T J(x) = u_i^T J(x) S_j, \qquad j = 1, 2, \cdots, n. \tag{2.1}$$

Since our approximation B to $J(x)$ is similarly sparse, it follows that

$$u_i^T B = u_i^T B S_j, \qquad j = 1, 2, \cdots, n. \tag{2.2}$$

We are now in a position to describe the algorithm, ignoring any initialisation procedure and denoting new values by the subscript unity.

1. Compute p, where

$$p = -B^{-1} f. \tag{2.3}$$

2. Compute x_1, where

$$x_1 = x + pt \tag{2.4}$$

and t is some scalar, not equal to zero, whose choice we discuss below.

3. Compute f_1 and test for convergence. If this has not occurred,

4. Compute y, where

$$y = f_1 - f. \tag{2.5}$$

5. Compute B_1, where

$$B_1 = B - \sum_{i=1}^{n} u_i u_i^T (Bp - yt^{-1}) \frac{p_i^T}{p_i^T p_i} \tag{2.6}$$

and

$$p_i = S_i p. \tag{2.7}$$

6. Repeat from step 1.

The scalar t is either set equal to unity or chosen so that $||f_1|| < ||f||$. For a fuller discussion of these strategies, see [1] and [3]. In the subsequent sections of the paper, we restrict ourselves to the case where $t = 1$, and it then follows from Eqs. (2.3), (2.5) and (2.6) that B_1 is given by

$$B_1 = B + \sum_{i=1}^{n} u_i u_i^T f_1 \frac{p_i^T}{p_i^T p_i}. \tag{2.8}$$

We note from Eqs. (2.4)–(2.6) that since $p_i^T p = p_i^T p_i$, the quasi-Newton equation

$$B_1(x_1 - x) = f_1 - f \tag{2.9}$$

is satisfied by B_1. We further note from Eq. (2.6) that

$$u_i^T B_1 = u_i^T \left[B - (Bp - yt^{-1}) \frac{p_i^T}{p_i^T p_i} \right]$$

so that since $p_i^T = p_i^T S_i$, if Eq. (2.2) is satisfied it follows that

$$u_i^T B_1 = u_i^T B_1 S_i, \tag{2.10}$$

so that the sparseness characteristics of B are preserved. The correction matrix added to B to give B_1 is not, however, a rank-one matrix, its rank being determined by the sparseness characteristics of the Jacobian and the presence of odd zeroes in the vectors y and p. A further consequence of Eqs. (2.2) and (2.7) is that Eq. (2.6) may be written

$$B_1 = B - \sum_{i=1}^{n} u_i u_i^T (Bp_i - yt^{-1}) \frac{p_i^T}{p_i^T p_i}. \tag{2.11}$$

Now, it follows from Eq. (2.3) that p is null only if f is null, and this, by definition, occurs only at the solution. It is, however, possible that p_i might be null for some j when p is not null and since this could cause a division failure when evaluating B_1, we now consider how this contingency may be overcome.

It is shown in the next section that subject to certain conditions

$$u_i^T y = u_i^T J(\bar{x}_i) p_i t,$$

where $J(\bar{x}_i)$ is the Jacobian of $f(x)$ evaluated at some point, depending upon j, lying on the line joining x and x_1. Thus, Eq. (2.11) may be written

$$B_1 = B - \sum_{i=1}^{n} u_i u_i^T [B - J(\bar{x}_i)] \frac{p_i p_i^T}{p_i^T p_i},$$

so that the correction applied to B does not depend directly upon $||p_i||$. In the event that p_i is null, the jth term in the sum could well be ignored. This is equivalent to replacing the matrix $p_i p_i^T / p_i^T p_i$ by the null matrix and we note that this does not invalidate the convergence proofs appearing in the next section. Perhaps the best practical procedure is to ignore the jth term in the update if $||p_i|| < \delta ||p||$, where δ is some suitable small quantity. This device was not found to be necessary when carrying out the tests described in Section 4 (below).

We give now a less formal description of the matrix update that illustrates its relationship to the update given by Broyden [1]. Thus:

1. Compute the correction given by Broyden [1, Method 1]. In general, the correction, if added to B, would give a nonsparse matrix.

2. Set those elements of the correction matrix corresponding to the zero elements of B themselves to zero. Note that if this correction is added to B as it stands, the new matrix will not, in general, satisfy the quasi-Newton equation (2.9) though it will be sparse.

3. Scale each row of the sparse correction independently so that when the scaled sparse correction is added to B the resulting B_1 does satisfy the quasi-Newton equation.

It is readily seen that this is equivalent to the correction formalised in Eq. (2.6), and to the method as proposed by Schubert [14]. It has the interesting property that if $f_i(x)$ is a function of x_i alone for all j, so that the Jacobian becomes diagonal, the algorithm reduces to the secant algorithm applied in parallel to all the f_i's. If, on the other hand, every S_i is the unit matrix, the method becomes Broyden's original method. Thus, in the cases of maximum and minimum sparseness the algorithm reverts to known methods, which may themselves in turn be regarded as special cases of the new method.

3. A Local Convergence Proof. This section is devoted to obtaining local convergence proofs for the algorithm for two different choices of the initial approximation to the Jacobian. Note that $||\cdot||$ denotes the Euclidean norm in the case of both vector and matrix. We distinguish, where necessary, the spectral matrix norm by the subscript s.

THEOREM 1. *Let s be a solution of the equations in question so that*

$$f(s) = 0. \tag{3.1}$$

Assume that there exists a neighbourhood N of s such that

(1) *∇f_i satisfies a Lipschitz condition on N of order one with constant L_i,*

(2) *$x_0 \in N$, where x_0 is the initial approximation to the solution,*

$$||e_0|| \leqq \mu_1/\alpha L, \tag{3}$$

where $e_0 = x_0 - s$, $\mu_1 \simeq 0.244$, $\alpha = ||A^{-1}||_s$, A is the Jacobian of $f(x)$ at s, and

$$L^2 = \sum_{i=1}^{n} L_i^2, \tag{3.2}$$

$$||E_0||^2 \leqq \left(\frac{\theta_1 - \frac{1}{2}\alpha L||e_0||}{\alpha(1 + \theta_1)}\right)^2 - \frac{L^2||e_0||^2}{1 - \theta_1^2} \tag{4}$$

where $E_0 = B_0 - A$, B_0 is the initial approximation to the Jacobian and $\theta_1 \simeq 0.671$.

Then, if x_r is the rth approximation to the solution generated by the algorithm with $t = 1$, and $e_r = x_r - s$,

$$||e_r|| \leqq \theta_1^r||e_0||. \tag{3.3}$$

Proof. We first consider the change in the approximation B to the Jacobian that occurs as a result of the matrix update, and, as before, we use the subscript unity to denote new values. If we define the matrix error E by $E = B - A$, we have, from Eq. (2.11),

$$E_1 = E - \sum_{i=1}^{n} u_i u_i^T[(A + E)p_i - y]\frac{p_i^T}{p_i^T p_i},$$

so that

$$u_i^T E_1 = u_i^T E\left(I - \frac{p_i p_i^T}{p_i^T p_i}\right) - u_i^T(Ap_i - y)\frac{p_i^T}{p_i^T p_i}, \qquad j = 1, 2, \cdots, n.$$

Now, if $x \in N$ and $x_1 \in N$, it follows from a result quoted by Kantorovich and Akilov [13, p. 660], and Eqs. (2.1), (2.5) and (2.7) that

$$u_i^T y = u_i^T J(x + p\bar{t}_i)p_i,$$

where $0 < \bar{t}_i < 1$. Thus, since $A = J(s)$,

$$u_i^T E_1 = u_i^T E\left(I - \frac{p_i p_i^T}{p_i^T p_i}\right) - u_i^T[J(s) - J(x + p\bar{t}_i)]\frac{p_i p_i^T}{p_i^T p_i}, \tag{3.4}$$

so that, since $p_i p_i^T/p_i^T p_i$ is both symmetric and idempotent,

$$||u_i^T E_1||^2 \leqq ||u_i^T E||^2 + ||u_i^T[J(s) - J(x + p\bar{t}_i)]||^2.$$

It follows from the first condition of the theorem and the definition of e that

$$||u_i^T E_1||^2 \leqq ||u_i^T E||^2 + L_i^2||e + p\bar{t}_i||^2; \tag{3.5}$$

and if $||e_1|| < ||e||$, so that $||e + p|| < ||e||$, it further follows from the fact that $0 < \bar{t}_i < 1$ that $||e + p\bar{t}_i|| < ||e||$. Substituting this result in inequality (3.5) and summing over j, then gives, from Eq. (3.2),

$$||E_1||^2 \leqq ||E||^2 + L^2||e||^2. \tag{3.6}$$

We consider now the relationship between the successive vector errors e and e_1. Condition (1) of the theorem, together with Eq. (3.2), implies that, for $x, y \in N$, $||J(x) - J(y)|| \leqq L||x - y||$ so that, since the spectral norm cannot exceed the Euclidean norm,

$$||J(x) - J(y)||_s \leqq L||x - y||. \tag{3.7}$$

If we now define r by

$$r = f(x) - Ae, \tag{3.8}$$

it follows from Eq. (3.7) and a lemma quoted by e.g. Dennis [12] that

$$||r|| \leqq \tfrac{1}{2}L||e||^2. \tag{3.9}$$

Now $x_1 = x - B^{-1}f(x)$ so that, from Eq. (3.8) and the definition of e,

$$e_1 = e - B^{-1}(Ae + r). \tag{3.10}$$

Now, since $E = B - A$,

$$B^{-1} = (I + A^{-1}E)^{-1}A^{-1} \quad \text{and} \quad I - B^{-1}A = (I + A^{-1}E)^{-1}A^{-1}E$$

so that Eq. (3.10) becomes

$$e_1 = (I + A^{-1}E)A^{-1}(Ee - r).$$

Thus, if $\alpha||E|| < 1$, taking norms and applying Eq. (3.9) yields

$$\frac{||e_1||}{||e||} \leqq \frac{\alpha(||E|| + \frac{1}{2}L||e||)}{1 - \alpha||E||}. \tag{3.11}$$

The proof now proceeds inductively, and we use the subscripts i and r to denote ith and rth approximations, respectively.

Assume that $||e_i|| \leqq \theta||e_{i-1}||$, $i = 1, 2, \cdots, r$, where $0 < \theta < 1$. Then, from Eq. (3.6),

$$\begin{aligned} ||E_r||^2 &\leqq ||E_0||^2 + L^2||e_0||^2(1 + \theta^2 + \cdots + \theta^{2r-2}) \\ &< ||E_0||^2 + \frac{L^2||e_0||^2}{1 - \theta^2}. \end{aligned} \tag{3.12}$$

Now, a sufficient condition that $||e_{r+1}|| \leqq \theta||e_r||$ is, from Eq. (3.11), $||E_r|| \leqq (\theta - \frac{1}{2}\alpha L||e_r||)/\alpha(1 + \theta)$. Since we require that $||e_r|| < ||e_0||$, $r > 0$, this condition is satisfied if

$$||E_r|| \leqq \frac{\theta - \frac{1}{2}\alpha L||e_0||}{\alpha(1 + \theta)} \tag{3.13}$$

and it thus follows from Eq. (3.12) that $||e_{r+1}|| \leqq \theta||e_r||$ if

$$||E_0||^2 + \frac{L^2||e_0||^2}{1 - \theta^2} \leqq \left(\frac{\theta - \frac{1}{2}\alpha L||e_0||}{\alpha(1 + \theta)}\right)^2,$$

or

$$\alpha^2||E_0||^2 \leqq \left(\frac{\theta - \frac{1}{2}\mu}{1 + \theta}\right)^2 - \frac{\mu^2}{1 - \theta^2}, \tag{3.14a}$$

where

$$\mu = \alpha L||e_0|| . \tag{3.14b}$$

It is clearly only possible to satisfy Eq. (3.14a) if

$$\left(\frac{\theta - \frac{1}{2}\mu}{1 + \theta}\right)^2 - \frac{\mu^2}{1 - \theta^2} \geqq 0$$

or

$$\mu \leqq \frac{2\theta(1 - \theta^2)^{1/2}}{(1 - \theta^2)^{1/2} + 2(1 + \theta)}, \tag{3.15}$$

and since we wish to obtain the most generous bound for $||e_0||$, we find that value of θ, $0 < \theta < 1$, that maximises μ. Denoting this value by θ_1 and the corresponding value of μ by μ_1, we find on differentiating Eq. (3.15) that $\theta_1 \simeq 0.671$, $\mu_1 \simeq 0.244$.

Thus, provided that $||e_0||$ and $||E_0||$ satisfy the third and fourth conditions of the theorem, and $||e_i|| \leqq \theta_1||e_{i-1}||$, $i = 1, 2, \cdots, r$, then $||e_{r+1}|| \leqq \theta_1||e_r||$, so that $||e_{r+m}|| \leqq \theta_1^m||e_r||$, $m = 1, 2, \cdots$. It is trivial to demonstrate that if the conditions of the theorem are satisfied $||E_0||$ satisfies Eq. (3.13) with $r = 0$, so that $||e_1|| \leqq \theta_1||e_0||$. Thus, $||e_r|| \leqq \theta_1^r||e_0||$, completing the proof.

THEOREM 2. *If*

(1) $x_0 \in N$,

(2) $B_0 = J(x_0)$,

(3) $||e_0|| \leqq \mu_2/\alpha L$, *where* $\mu_2 \simeq 0.205$, *then* $||e_r|| \leqq \theta_2^r||e_0||$ *where* $\theta_2 \simeq 0.724$.

Proof. From the argument used to establish Eq. (3.6), it follows that

$$||J(s + e_0) - J(s)|| \leqq L||e_0||$$

so that, from the second condition of the theorem and from the definitions of A, E_0 and e_0,

$$||E_0|| \leqq L||e_0|| .$$

It then follows from Eq. (3.14a) that if $||e_i|| \leqq \theta||e_{i-1}||$, $i = 1, 2, \cdots, r$, $||e_{r+1}|| \leqq \theta||e_r||$ if

$$\mu^2\left(\frac{2 - \theta^2}{1 - \theta^2}\right) \leqq \left(\frac{\theta - \frac{1}{2}\mu}{1 + \theta}\right)^2 \quad \text{or} \quad \mu \leqq \frac{2\theta(1 - \theta^2)^{1/2}}{(1 - \theta^2)^{1/2} + 2(1 + \theta)(2 - \theta^2)^{1/2}}.$$

The value of θ, which we denote by θ_2, that maximises the above expression is $\theta_2 = 0.724$ and the corresponding value of μ is $\mu_2 = 0.205$. The proof is completed by induction in a manner similar to that of Theorem 1.

We note that the third condition of Theorem 1 shows that the bound on $||e_0||$ is inversely proportional to both L and α, emphasising the obvious point that the

better conditioned A is and the less nonlinear the equations are, the greater the tolerable initial error becomes. The fourth condition also corroborates an intuitively obvious result, that the better the initial estimate x_0, the greater the tolerance that may be extended towards the initial Jacobian approximation B_0.

In the case where B_0 is equal to $J(x_0)$, the bound on $||e_0||$ is reduced slightly since an increase in $||e_0||$ cannot be compensated by a reduction in $||E_0||$. We note that, as in Theorem 1, the bound on $||e_0||$ is inversely proportional to both $||A^{-1}||$ and L.

Another interesting feature emerges from Eq. (3.4). The first term on the right-hand side of this equation is due to using an approximation to A instead of A itself, and the second term is due to the presence of nonlinearities in the equations to be solved. Despite the fact that the sources of these terms are quite independent, the terms themselves are not, being in fact orthogonal. This feature contributes markedly to obtaining the convergence proof since it means that these error terms cannot reinforce each other. In the nonsparse algorithm, this characteristic is still present, as may be seen by putting p_j, $j = 1, 2, \cdots, n$, equal to p.

We finally note that these convergence proofs do not entirely do justice to the alogrithm since they ignore the reduction of the matrix error norm that may occur due to the first term on the right-hand side of Eq. (3.4). We conjecture that in most cases convergence is superlinear (a conjecture supported by the experimental evidence to date), although we have been unable to prove this in the general nonlinear case. We have, however, been able to prove superlinear convergence for the parent algorithm when applied to linear systems [3].

4. Experimental Results. We describe here briefly the results of some numerical experiments carried out in ALGOL using an ICL 1909. The performance of the new algorithm is compared with that of Newton's method where the Jacobian was computed using forward differences with an increment to each variable of 0.001. Since a reasonably good initial estimate was available for all the problems attempted, it was sufficient that the step-length scaling factor t was set equal to unity, and indeed only problems of this type need be considered since more difficult problems may be broken down using a Davidenko path technique [2], [5]. The first iteration was the same in both cases since a forward-difference initial approximation to the Jacobian was used in the new algorithm.

The comparison between methods is based upon the number of evaluations of individual elements of the residual vector f, the inherent assumption being that each element requires as much work to compute as any other. It is not satisfactory to compare the numbers of evaluations of the vector function f since when obtaining the jth row of the Jacobian by forward differences it is necessary to compute the jth element of f only one more time than the number of nonzero elements of that row. For every problem the iteration was terminated when $||f||$ was exceeded by some arbitrary prescribed tolerance.

The test problems used were chosen since they gave rise to Jacobians of band form. The reason for this choice was the immediate availability of a band-matrix linear equations routine, and the nonavailability of a similar routine to handle randomly sparse matrices. Since the theory of Sections 2 and 3 (above) applies equally to randomly sparse or banded matrices it was thought that this, coupled with experimental results on the band matrices (problems 7–9 show no dependence

TABLE 1. *Type* 1. $\epsilon = 10^{-6}$

Problem	n	k_1	I_N	I_B	N_N	N_B
1	5	0.1	3	5	59	43
2	5	0.5	3	4	59	38
3	10	0.5	3	5	124	88
4	20	0.5	4	5	332	178
5	600	0.5	4	5	10,192	5,398
6	600	2.0	4	7	10,192	6,598

of performance upon band structure), would provide sufficient justification for publication of the new algorithm.

The problems are of two basic types, thus:

Type 1. $f_i = (3 - k_1 x_i)x_i + 1 - x_{i-1} - 2x_{i+1}$.

Type 2. $f_i = (k_1 + k_2 x_i^2)x_i + 1 - k_3 \sum_{j=i-r_1;\, j\neq i}^{i+r_2} (x_j + x_j^2)$.

In each case x_i, $i < 1$ or $i > n$, is regarded as zero. The parameters k_1, k_2 and k_3 enable the amount of nonlinearity to be varied while r_1 and r_2 permit the bandwidth to be altered in equations of Type 2. Note that r_1 and r_2 are independent so that the band is not necessarily symmetric about the principal diagonal.

The results of the tests are summarized in three tables. The initial estimate of the solution in each case was $x_i = -1$, $i = 1, 2, \cdots, n$, and the particular combination of parameters for each problem is given in the tables. The symbols I_N, I_B, N_N and N_B denote the numbers of iterations and residual element evaluations for Newton's method and the new method respectively, and ϵ refers to the final tolerance on $||f||$.

For problems of Type 1, (Table 1) the number of iterations required by the new method is somewhat higher than that required by Newton's method but even for problem 6 the number of element evaluations is substantially lower. It is seen that both methods coped adequately with a system of 600 equations. Problems 2, 3, and 4 were among those attempted by Schubert [14] whose results are in agreement with those quoted in Table 1 of the present paper.

Table 2 shows the effect, more precisely, the absence of effect, when the band structure of Type 2 problems is altered keeping the bandwidth constant, and Table 3 shows how the algorithms compare as the amount of nonlinearity is varied. It is clear from this table that for the types of equation tested the ratio N_B/N_N increases as the nonlinearities become more pronounced but even in the worst case (problem 23)

TABLE 2. *Type* 2. $n = 100$, $k_1 = k_2 = k_3 = 1.0$, $\epsilon = 10^{-6}$

Problem	r_1	r_2	I_N	I_B	N_N	N_B
7	3	3	4	8	3252	1588
8	2	4	4	8	3248	1587
9	5	1	4	8	3236	1584

TABLE 3. *Type* 2. $n = 50, r_1 = r_2 = 5, \epsilon = 10^{-6}$

Problem	k_1	k_2	k_3	I_N	I_B	N_N	N_B
10	1	1	1	4	8	2330	970
11	2	1	1	5	10	2900	1070
12	1	2	1	5	11	2900	1120
13	3	2	1	5	11	2900	1120
14	2	3	1	5	15	2900	1320
15	3	3	1	5	16	2900	1370
16	2	2	1	5	11	2900	1120
17	1	2	2	4	7	2330	920
18	2	2	2	4	9	2330	1020
19	2	3	2	4	11	2330	1120
20	2	4	1	5	20	2900	1570
21	2	5	1	5	23	2900	1720
22	3	4	1	5	19	2900	1520
23	3	5	1	5	24	2900	1770

this ratio does not exceed 2/3. The total number of iterations, however, is up by a factor of 5. Clearly, whether or not the new method is competitive depends upon the cost of computing f relative to solving the band equations, and this will of course vary with different problems.

A final comparison may be made, through problems 1–4, with the Broyden's original method [1]. This solved these problems to comparable accuracy in 5, 5, 7 and 8 iterations respectively compared with the new method's 5, 4, 5 and 5. It thus appears that for some sparse equations the new method is better, in terms of iterations, than the old one and simple theoretical considerations support this conclusion.

5. Discussion and Conclusions. The results quoted in the previous section show that for mildly nonlinear problems the new method offers a useful if modest improvement upon Newton's method, but that this improvement tends to vanish as the nonlinearities become more pronounced. It also appears, since systems of up to 600 equations were tested, that size alone provides no difficulties for the new method. The convergence proofs of Section 3, guaranteeing convergence provided certain conditions are satisfied, bestow a certain amount of respectability upon the algorithm and confidence in its use. The practical superiority of the new method over Newton's method though depends heavily upon the relative amounts of computation required in the evaluation of f and the solution of the band system. If f is extremely laborious to compute, and such functions may occur in practice, then the new method will be superior.

We consider now the computational aspects of the method. If $t = 1$ the matrix update is given by Eq. (2.8) and it is necessary only to store the vector p in addition to those matrices and vectors whose storage is required by Newton's method, namely x, f and B (the last in packed form with some marker device for indicating the position of the nonzero elements). It is true that Eq. (2.8) requires the availability of the n

vectors p_i, but these may be obtained directly from p using the marker device. Even if $t \neq 1$, only one further vector need be stored so that from the standpoint of computer storage the method appears to be extremely attractive until one remembers that most routines for solving sparse linear systems destroy the original matrix. Since this is required for updating, it is necessary to retain a copy so that the "packed" matrix storage is doubled. This might be a serious disadvantage of the method but one which could perhaps be overcome, for certain problems, by using a device due to Bennett [11], whereby the factors L and U of B are stored and updated.

One concludes therefore that the new method cannot be regarded as being the automatic choice in every case where the Jacobian is difficult to evaluate. For large problems or small computers, storage requirements may be paramount, and even if the storage is available, the software being used may make it not readily accessible. Thus, the potential user of the new algorithm must decide for or against its use only after a comprehensive survey of the computing facilities available to him.

The author is extremely grateful to Mr. K. Fielding, now of the University of Surrey, for undertaking the numerical experiments and to the referee for suggesting improvements to the convergence proof.

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

1. C. G. BROYDEN, "A class of methods for solving nonlinear simultaneous equations," *Math. Comp.*, v. 19, 1965, pp. 577–593. MR **33** #6825.

2. C. G. BROYDEN, "A new method of solving nonlinear simultaneous equations," *Comput. J.*, v. 12, 1969/70, pp. 94–99. MR **39** #6509.

3. C. G. BROYDEN, "The convergence of single-rank quasi-Newton methods," *Math. Comp.*, v. 24, 1970, pp. 365–382.

4. A. CHANG, *Applications of Sparse Matrix Methods in Electric Power Systems Analysis*, Proc. Sympos. on Sparse Matrices and Their Applications (IBM Watson Research Center, 1968), RA 1 #11707, Watson Research Center, Yorktown Heights, New York, 1969.

5. D. F. DAVIDENKO, "An application of the method of variation of parameters to the construction of iterative formulas of increased accuracy for numerical solutions of nonlinear equations," *Dokl. Akad. Nauk SSSR*, v. 162, 1962, pp. 499–502 = *Soviet Math. Dokl.*, v. 3, 1962, pp. 702–706. MR **31** #2838.

6. A. A. GOLDSTEIN, *Constructive Real Analysis*, Harper & Row, New York, 1967. MR **36** #705.

7. F. G. GUSTAVSON, W. LINIGER & R. WILLOUGHBY, "Symbolic generation of an optimal Crout algorithm for sparse systems of linear equations," *J. Assoc. Comput. Mach.*, v. 17, 1970, pp. 87–109.

8. R. P. TEWARSON, *The Gaussian Elimination and Sparse Systems*, Proc. Sympos. on Sparse Matrices and Their Applications (IBM Watson Research Center, 1968), Watson Research Center, Yorktown Heights, New York, 1968.

9. W. F. TINNEY & C. E. HART, "Power flow solutions by Newton's method," *IEEE Trans.*, v. PAS-86, 1967, pp. 1449–1460.

10. J. H. WILKINSON, *The Algebraic Eigenvalue Problem*, Clarendon Press, Oxford, 1965. MR **32** #1894.

11. J. M. BENNETT, "Triangular factors of modified matrices," *Numer. Math.*, v. 7, 1965, pp. 217–221. MR **31** #1766.

12. J. E. DENNIS, *On Convergence of Newton-Like Methods*, Numerical Methods for Non-linear Algebraic Equations edited by P. Rabinowitz, Gordon & Breach, London, 1970.

13. L. V. KANTOROVIČ & G. P. AKILOV, *Functional Analysis in Normed Spaces*, Fitzmatgiz, Moscow, 1959; English transl., Internat. Series of Monographs in Pure and Appl. Math., vol. 46, Macmillan, New York, 1964. MR **22** #9837; MR **35** #4699.

14. L. K. SCHUBERT, "Modification of a quasi-Newton method for nonlinear equations with a sparse Jacobian," *Math. Comp.*, v. 25, 1970, pp. 27–30.