To be Submitted for Publication

LBL-7236
Preprint c. 2

UPDATE METHODS FOR HIGHLY STRUCTURED SYSTEMS OF NONLINEAR EQUATIONS

J. H. Avila and P. Concus

RECEIVED
LAWRENCE
BERKELEY LABORATORY
MAR 8 1978
LIBRARY AND
DOCUMENTS SECTION

November 1977

Prepared for the U. S. Department of Energy
under Contract W-7405-ENG-48;
Also issued as Institute for
Advanced Computation Report 5537

TWO-WEEK LOAN COPY

This is a Library Circulating Copy
which may be borrowed for two weeks.
For a personal retention copy, call
Tech. Info. Division, Ext. 5716

LAWRENCE BERKELEY LABORATORY
UNIVERSITY OF CALIFORNIA
BERKELEY

LBL-7236
c. 2

DISCLAIMER

This document was prepared as an account of work sponsored by the United States Government. While this document is believed to contain correct information, neither the United States Government nor any agency thereof, nor the Regents of the University of California, nor any of their employees, makes any warranty, express or implied, or assumes any legal responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by its trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof, or the Regents of the University of California. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof or the Regents of the University of California.

LBL-7236

UPDATE METHODS FOR HIGHLY STRUCTURED SYSTEMS OF NONLINEAR EQUATIONS

J. H. Avila[†] and P. Concus[‡]

November 1977

ABSTRACT

Often it is desirable to incorporate known structure of the Jacobian matrix into numerical algorithms for solving systems of nonlinear equations by Broyden update methods. Here we develop a technique for accomplishing this for a certain highly structured system of equations arising as the discrete form of an eigenvalue problem for a nonlinear elliptic partial differential equation of nuclear reactor analysis. We demonstrate how the technique, which is formulated in terms of a pseudo-inverse, can be extended to incorporate other, more general, structures. As an example, we obtain known minimal update formulas for preserving sparsity and symmetry as special cases of the technique. Numerical results are given for the nonlinear eigenvalue problem.

[†]Institute for Advanced Computation, P.O. Box 5071, Sunnyvale, California 94086.

[‡]Lawrence Berkeley Laboratory, University of California, Berkeley, California 94720.

1. INTRODUCTION

In the numerical solution of sparse systems of nonlinear equations by quasi-Newton methods it is usually desirable, but often difficult, to impose the known structure of the Jacobian matrix on its approximates. For example, in its simplest form, the update method introduced by Broyden [1] will preserve neither sparsity nor symmetry. A technique for preserving sparsity for the update method has been introduced by Shubert [2], and several ways to preserve symmetry are discussed by Dennis and Moré [3]. In this paper we develop a technique for preserving even more complex structures.

We begin by examining a system with a special structure arising from a nonlinear eigenvalue problem. By utilizing the minimal update aspect of most quasi-Newton methods and employing properties of the pseudo-inverse, we show how to impose the desired structure on the approximate Jacobians. Then we show how our technique can be used to impose structure in more general nonlinear systems. In its general setting the technique includes, as special cases, the minimal update methods of [2] and [3].

2. A NONLINEAR EIGENVALUE PROBLEM

The motivation for our study arose in connection with the following nonlinear eigenvalue problem arising in the study of neutron diffusion in a nuclear reactor (see [4], chapter 4).

Let (x,z) be a point in the rectangle R $(0 \leq x \leq L,\ 0 \leq z \leq H)$; find $\phi(x,z)$ and the smallest real scalar λ such that

$$\partial^2\phi/\partial x^2 + \partial^2\phi/\partial y^2 - \frac{1}{\beta}\phi + \lambda\frac{\alpha}{\beta}\phi = 0 \ . \tag{1}$$

Here α and β are functions of x, y, and ϕ; $\alpha, \beta > 0$. There are appropriate boundary conditions on the boundary of R that need not concern us here and a normalization condition that we shall give below in discrete form. For the problems under consideration the smallest eigenvalue λ is simple, real, and positive.

By placing a grid on R $(x_i=(i-1)\Delta x,\ i=1, \ldots, I;\ z_k=(k-1)\Delta z,\ k=1, \ldots, K)$ and taking the standard five-point difference approximation in (1), we can write the finite difference formulation as

$$A\phi - \frac{1}{\beta}\phi + \lambda\frac{\alpha}{\beta}\phi = 0 \ , \tag{2}$$

which we denote by

$$A\phi + E(\lambda,\phi) = 0 \ , \tag{3}$$

where A is a matrix of order $n = I \times K$ and ϕ is a vector belonging to R^n. We impose the normalization condition that

$$s^T\phi - 1 = 0 \tag{4}$$

for some appropriate vector s belonging to R^n.

If z represents the vertical direction and x the horizontal direction, then for a class of reactors and an appropriate choice of Δx, α and β depend only on values of ϕ below any given point and not on values laterally. To be precise, at (x_i, z_k) one has $\alpha_{i,k} = \alpha_{i,k}(\phi_{i,k}, \phi_{i,k-1}, \ldots, \phi_{i,1})$ and a similar expression for $\beta_{i,k}$. Furthermore, the dependence is on the sum of the values, i.e., $\alpha_{i,k} = \alpha_{i,k}\left(\sum_{j=1}^{k} \phi_{i,j}\right)$ and similarly for $\beta_{i,k}$.

We wish to find ϕ and the smallest value of the scalar λ such that (3) and (4) are satisfied.

One way to solve this nonlinear eigenvalue problem is to use Newton's method [5] on the system of equations (3,4). The Jacobian matrix J of (3,4) is

$$J = \begin{bmatrix} A + E_\phi & E_\lambda \\ s^T & 0 \end{bmatrix}. \tag{5}$$

If we number the nodes from 1 to n by running through the i,k indices in increasing order, cycling through k most rapidly, we find that E_ϕ has the block diagonal form

$$E_\phi = \begin{bmatrix} D_1 & & & 0 \\ & D_2 & & \\ & & \ddots & \\ 0 & & & D_I \end{bmatrix}, \tag{6}$$

with each D_i a matrix of order K. Moreover, each D_i has the form

$$(7) \qquad D_i = \begin{bmatrix} b_1^{(i)} & & & & & \\ a_2^{(i)} & b_2^{(i)} & & & 0 & \\ a_3^{(i)} & a_3^{(i)} & b_3^{(i)} & & & \\ \cdot & & & \cdot & & \\ \cdot & & & & \cdot & \\ \cdot & & & & & \cdot \\ a_K^{(i)} & a_K^{(i)} & \cdot & \cdot & \cdot & a_K^{(i)} \quad b_K^{(i)} \end{bmatrix} .$$

The fact that the elements in each row to the left of the diagonal are identical follows from the dependence of α and β on the sum of the elements of ϕ corresponding to grid points below any given grid point. The block diagonal structure of E_ϕ follows from the lateral ϕ-independence.

For the problem under consideration the functions α and β, their derivative with respect to ϕ, and hence the Jacobian (5) are very expensive to compute. For this reason the quasi-Newton methods seem a natural alternative for solving (3,4). The difficulty in using such methods lies in trying to impose the structure (6,7) on E_ϕ.

For a set of n nonlinear equations in n variables

$$F(x) = 0,$$

where x belongs to R^n, the traditional Broyden's method in its simplest form is given by [1]

$$(8) \qquad B_j(x_{j+1} - x_j) = -F(x_j) \quad , \quad j=0, 1, \ldots$$

where x_0 is an initial estimate of the solution, B_0 is a given approximation to the Jacobian $F'(x_0)$, and

$$B_{j+1} = B_j + \frac{(y_j - B_j w_j) w_j^T}{w_j^T w_j} , \tag{9}$$

with

$$y_j = F(x_{j+1}) - F(x_j) , \qquad w_j = x_{j+1} - x_j . \tag{10}$$

For our problem, we proceed with the above method as follows. We pick an estimate for E_ϕ, which we denote by B_0. In particular B_0 has the required structure. We then pick initial guesses λ_0 and ϕ_0 such that $s^T\phi_0=1$. For each j=0, 1, 2, ... we solve

$$\begin{cases} (A + B_j)\Delta\phi_j + E_\lambda^j \Delta\lambda_j = A\phi_j + E(\lambda_j,\phi_j) \equiv r_j \\ s^T\Delta\phi_j = 0 \end{cases} \tag{11}$$

which solution we denote by

$$\Delta\lambda_j = \frac{s^T(A + B_j)^{-1} r_j}{s^T(A + B_j)^{-1} E_\lambda^j} \tag{12}$$

and

$$\Delta\phi_j = (A + B_j)^{-1} r_j - (A + B_j)^{-1} E^j \Delta\lambda_j . \tag{13}$$

Then

$$\begin{cases} \phi_{j+1} = \phi_j - \Delta\phi_j \\ \lambda_{j+1} = \lambda_j - \Delta\lambda_j \end{cases} . \tag{14}$$

We defer discussing the updating of B_j until the next section. One condition, however, that B_{j+1} must satisfy is the secant condition

$$(15) \qquad (A + B_{j+1})\Delta\phi_j + E_\lambda^{j+1}\Delta\lambda_j = r_j - r_{j+1} \quad .$$

Let $B_{j+1} = B_j + \Delta B_j$, then (15) implies

$$(16) \qquad \Delta B_j \Delta\phi_j = r_j - r_{j+1} - E_\lambda^{j+1}\Delta\lambda_j - (A + B_j)\Delta\phi_j \equiv v_j \quad .$$

Clearly, if B_j and ΔB_j possess the desired structure, then so will B_{j+1}. We shall determine how to find ΔB_j possessing the structure defined above, such that (16) is satisfied.

3. STRUCTURED UPDATES

We begin by placing our updating technique in a more general setting. In this way, as we shall see in Section 4, the technique can be extended to incorporate other kinds of structure in the Jacobian.

First, we establish the following result.

Lemma. Let the $n \times m$ matrix U be of the form

$$U = \begin{bmatrix} u_1^T & & & 0 \\ & u_2^T & & \\ & & \ddots & \\ 0 & & & u_n^T \end{bmatrix}, \tag{17}$$

where u_ℓ is a vector in R^{k_ℓ} and $\sum_{\ell=1}^{n} k_\ell = m$.

(i) If $u_\ell \neq 0$ for $1 \leq \ell \leq n$, then the pseudo-inverse $U^\dagger$ of U is given by the $m \times n$ matrix

$$U^\dagger = \begin{bmatrix} \frac{u_1}{u_1^T u_1} & & & 0 \\ & \frac{u_2}{u_2^T u_2} & & \\ & & \ddots & \\ 0 & & & \frac{u_n}{u_n^T u_n} \end{bmatrix}. \tag{18}$$

(ii) <u>If</u> $u_\ell \equiv 0$ <u>for some values of</u> ℓ, <u>then</u> $U^\dagger$ <u>is given by</u> (18) <u>with the columns corresponding to those values of</u> ℓ <u>replaced by zeros</u>.

<u>Proof</u>. The result can be obtained from the observation, which can be verified directly, that $U^\dagger$ satisfies the four requisite properties of the (unique) Moore-Penrose pseudo-inverse -- namely, $U^\dagger U$ and $UU^\dagger$ are symmetric, $UU^\dagger U = U$, and $U^\dagger UU^\dagger = U^\dagger$.

A somewhat more constructive proof can be given, which will point more clearly toward later extensions. When $u_\ell \neq 0$ for $1 \leq \ell \leq n$, the $n \times n$ diagonal matrix

$$UU^T = \begin{bmatrix} u_1^T u_1 & & & \bigcirc \\ & u_2^T u_2 & & \\ & & \ddots & \\ \bigcirc & & & u_n^T u_n \end{bmatrix}$$

is of full rank, and we can form

$$U^* = U^T (UU^T)^{-1} . \tag{19}$$

One can verify easily that U^* satisfies the requisite properties of $U^\dagger$ listed in the preceding paragraph, hence U^* is equal to $U^\dagger$. More generally, when UU^T may not be of full rank, we form

$$U^* = U^T (UU^T)^\dagger ,$$

where $(UU^T)^\dagger$, the pseudo-inverse of UU^T, is the diagonal matrix

whose elements are

$$[(UU^T)^+]_\ell = \begin{cases} (u_\ell^T u_\ell)^{-1} & , \text{ if } u_\ell \not\equiv 0 \\ 0 & , \text{ if } u_\ell \equiv 0 \end{cases} .$$

This U^* is $U^\dagger$, in general.

We now return to the question of how to preserve structure within update methods. The general Broyden update formula (9) is usually derived in one of two ways. The first is to require that the secant condition

$$B_{j+1} w_j = y_j \tag{20}$$

be satisfied (with w_j and y_j given by (10)) together with the condition that

$$B_{j+1} z = B_j z \tag{21}$$

for all z orthogonal to w_j. The second is to require that (20) be satisfied together with the minimal update condition

$$\| B_{j+1} - B_j \|_F \text{ is minimized.} \tag{22}$$

Here $\| \cdot \|_F$ denotes the Frobenius norm.

By using (20) and (22), the updating technique that preserves the structure that we seek can be obtained directly.

In establishing the desired result, we first simplify (16) somewhat. In particular, from the structure of ΔB_j described by (6) and (7), it is clear that we can consider the problem for just one diagonal block. That is, we wish to find a $K \times K$ matrix C such that

(23) $$Cu = v$$

and such that

(24) $$\|C\|_F = \text{minimum.}$$

The matrix C corresponds to one block of ΔB_j, u replaces one segment of the vector $\Delta\phi_j$, and v replaces the corresponding segment of v_j in (16). In addition C is required to have the structure given by

(25) $$C = \begin{bmatrix} b_1 & & & & & \\ a_2 & b_2 & & & 0 & \\ a_3 & a_3 & b_3 & & & \\ \cdot & & & \cdot & & \\ \cdot & & & & \cdot & \\ \cdot & & & & & \cdot \\ a_K & a_K & a_K & \cdots & a_K & b_K \end{bmatrix} .$$

Let the elements of u and v be denoted by

(26) $$u = (\mu_1, \mu_2, \ldots, \mu_K)^T \quad , \quad v = (\nu_1, \nu_2, \ldots, \nu_K)^T$$

and define S_ℓ to be

$$S_1 = 0$$

$$S_\ell = (\ell - 1)^{-1/2} \sum_{i=1}^{\ell-1} \mu_i \quad , \quad 2 \le \ell \le K \quad .$$

Then the update technique is contained in the following result.

<u>Theorem</u>. <u>The</u> $K \times K$ <u>matrix</u> C <u>of the form</u> (25) <u>satisfying</u> (23) <u>and</u> (24) <u>is given by</u>

(27a) $$\begin{cases} b_\ell = \mu_\ell \nu_\ell / (S_\ell^2 + \mu_\ell^2) \quad , \quad 1 \le \ell \le K \\ a_\ell = (\ell - 1)^{-1/2} \nu_\ell S_\ell / (S_\ell^2 + \mu_\ell^2) \quad , \quad 2 \le \ell \le K \end{cases}$$

if $S_\ell^2 + \mu_\ell^2 \neq 0$. *If* $\mu_\ell = S_\ell = 0$ *for some values of* ℓ, *then*

$$b_\ell = 0 \quad , \quad a_\ell = 0 \tag{27b}$$

for those values; when (23) *and* (25) *are not consistent,* (23) *will be satisfied in the least-squares sense.*

Proof. Using (25) and (26), we obtain that (23) is equivalent to the set of equations

$$\begin{bmatrix} \mu_1 & & & & & \\ & \mu_2 & \mu_1 & & & \\ & & \mu_3 & \sum_{i=1}^{2} \mu_i & & \\ & & & \ddots & \ddots & \\ & & & & \ddots & \\ & & & & \mu_K & \sum_{i=1}^{K-1} \mu_i \end{bmatrix} \begin{bmatrix} \frac{b_1}{b_2} \\ \frac{a_2}{b_3} \\ \frac{a_3}{\cdot} \\ \cdot \\ \cdot \\ \frac{\cdot}{b_K} \\ a_K \end{bmatrix} = \begin{bmatrix} \nu_1 \\ \nu_2 \\ \nu_3 \\ \cdot \\ \cdot \\ \cdot \\ \nu_K \end{bmatrix} . \tag{28}$$

We denote (28) as

$$Uc = v \ . \tag{29}$$

Let D be the diagonal matrix

$$D = \mathrm{diag}(1, \sqrt{1}, 1, \sqrt{2}, 1, \sqrt{3}, \ldots, 1, \sqrt{K-1}, 1) \ .$$

Then if we scale c with D we find that $\|Dc\|_2$, the Euclidean norm of Dc, satisfies

$$\|Dc\|_2 = \|C\|_F \ .$$

With this scaling (29) becomes

$$UD^{-1}\tilde{c} = v \ , \tag{30}$$

where

$$\tilde{c} = Dc \ .$$

The properties of the pseudo-inverse are such that the least squares solution of (30)

$$\tilde{c} = (UD^{-1})^{\dagger} v \ , \tag{31}$$

minimizes $\|\tilde{c}\|_2 = \|Dc\|_2$; (31) is equivalent to

$$c = D^{-1}(UD^{-1})^{\dagger} v \ .$$

The matrix UD^{-1} is of the form given in the Lemma, and (27a,b) follow. If the structure (25) is not consistent with the values of u and v in the secant condition (23)--that is, if $\nu_\ell \neq 0$ and $u_\ell = S_\ell = 0$ for some ℓ--then the properties of the pseudo-inverse are such that the solution for C will minimize $\|v-Uc\|_2 = \|v-Cu\|_2$.

For our application we have been interested primarily in the case for which (25) represents the actual form of the Jacobian, so that (23) is satisfied; but such a restriction is not necessary for the method.

It is interesting to note that if the structure of C is unspecified, then the set of equations (23) can be rewritten as an equivalent set similar to (28), except that each diagonal block is simply u^T. By using the Lemma and setting $C = B_{j+1} - B_j$ one then obtains the standard Broyden update formula (9).

In the next section, we examine how other structural forms of the Jacobian can be preserved within the context of the update methods. In this connection we give here a more general result analogous to Theorem 1, which also follows directly from the minimal least-squares property of the pseudo-inverse. As before, we consider condition (23) but permit C to be any real $n \times m$ matrix, with u and v given vectors in R^m and R^n, respectively. A structure $S(C)$ is imposed on C, which when combined with (23) yields the equation

$$Uc = w \ , \tag{32}$$

where c is a vector whose components are the unknown elements of C. We assume that a diagonal scaling matrix D exists so that $\|Dc\|_2 = \|C\|_F$, and we write the minimal least squares solution to the correspondingly scaled form of (32) as

$$c = D^{-1}(UD^{-1})^{\dagger}w \ . \tag{33}$$

There follows immediately

<u>Theorem 2</u>. <u>The matrix</u> C <u>having structure</u> $S(C)$ <u>and satisfying</u> (23) <u>and</u> (24), <u>with</u> C, u, v <u>as above</u>, <u>is given by</u> (33). <u>If</u> $S(C)$ <u>and</u> (23) <u>are not consistent</u>, <u>or if otherwise</u> (32) <u>has no solution</u>, <u>then</u> (32) <u>will be satisfied in the least squares sense</u>.

For many update methods, such as those considered in this and the following sections, $m = n$, $S(C)$ restricts certain elements of C to be equal to other elements of C or to zero, $w = v$, and $Uc = Cu$.

4. OTHER STRUCTURES

Often an important requirement of an approximate Jacobian of a nonlinear system is that known elements be incorporated into it. For example, in the solution of

$$f_i(x) = 0 \quad , \quad i=1, 2, \ldots, n$$

assume that

$$\frac{\partial f_i}{\partial x_k} \text{ is given for } k \in I_i$$

with I_i a subset of the integers 1, 2, ..., n. One wishes to require that the approximation B_j to the Jacobian satisfy (8), (22), and the condition that

$$(B_j)_{i,k} \text{ is specified for } k \in I_i, \text{ for each } i.$$

Schubert [2] has solved this problem using an approach different from ours. We show how the same result can be obtained with the technique presented in section 3. Let C be $B_{j+1} - B_j$. Then in our context, at each iterative step we wish to determine C subject to the conditions (23), (24) and

$$(C)_{i,k} = 0 \quad \text{for } k \in I_i \ .$$

Let $\bar{u}_i$ and $\bar{c}_i$ denote the $(n-r_i)$-dimensional vectors formed from u and the i-th row C_i^T of C in (23) by deleting the r_i elements of u (or columns of C) corresponding to the index set I_i.

Then (23) is equivalent to

$$\begin{bmatrix} \bar{u}_1^T & & & \bigcirc \\ & \bar{u}_2^T & & \\ & & \ddots & \\ \bigcirc & & & \bar{u}_n^T \end{bmatrix} \begin{bmatrix} \bar{c}_1 \\ \bar{c}_2 \\ \vdots \\ \bar{c}_n \end{bmatrix} = v \quad .$$

Using the Lemma, one can determine the pseudo-inverse of the matrix on the left and find that

$$\bar{c}_i = \frac{\nu_i \bar{u}_i}{\bar{u}_i^T \bar{u}_i}$$

if $\bar{u}_i \not\equiv 0$ (or $\bar{c}_i \equiv 0$ if $u_i \equiv 0$), where ν_i is the ith component of the vector v. Here, no scaling is necessary.

In principle, symmetry can be treated in a similar way. However, the structure of the matrix U does not lend itself to an application of the Lemma for obtaining the pseudo-inverse in this case. Even so we can show that an explicit update formula is obtainable by the procedure described above. Consider (23) and (24) with the condition that C be symmetric. If n is the order of the matrix C, then the problem of determining C becomes the problem of determining the least squares solution of

$$\left[\begin{array}{cccc|cccc|c|cc|c}
\mu_1 & \mu_2 & \cdots & \mu_n & 0 & 0 & \cdots & 0 & & & & \\
 & \mu_1 & & 0 & \mu_2 & \mu_3 & \cdots & \mu_n & \cdot & & 0 & 0 \\
 & & \ddots & & & \mu_2 & & 0 & & & & \\
0 & & & \mu_1 & 0 & & \ddots & \mu_2 & & \mu_{n-1} & \mu_n & \\
 & & & & & & & & & 0 & \mu_{n-1} & \mu_n
\end{array}\right]
\left[\begin{array}{c}
c_{1,1} \\ c_{1,2} \\ \vdots \\ c_{1,n} \\ \hline c_{2,2} \\ c_{2,3} \\ \vdots \\ c_{2,n} \\ \hline \vdots \\ \hline c_{n-1,n-1} \\ c_{n-1,n} \\ \hline c_{n,n}
\end{array}\right]
=
\begin{bmatrix} \nu_1 \\ \nu_2 \\ \vdots \\ \nu_n \end{bmatrix},$$

(34)

with the notation of (26).

Write (34) as

$$Uc = v \ . \tag{35}$$

We scale c so that the Euclidean norm of the resulting vector has the same value as the Frobenius norm of C by choosing the scaling matrix

$$D = \mathrm{diag}(1, \underbrace{\sqrt{2}, \sqrt{2}, \ldots, \sqrt{2}}_{n-1}, 1, \underbrace{\sqrt{2}, \ldots, \sqrt{2}}_{n-2}, \ldots, 1, \sqrt{2}, 1).$$

Equation (35) is equivalent to

$$UD^{-1}\tilde{c} = v \ , \tag{36}$$

with

$$\tilde{c} = Dc.$$

Note that $(UD^{-1})(UD^{-1})^T$ is of full rank (for the non-trivial case $u \neq 0$) and that $(UD^{-1})^{\dagger} = (UD^{-1})^T[(UD^{-1})(UD^{-1})^T]^{-1}$. Thus

$$c = D^{-2}U^T(UD^{-2}U^T)^{-1}v \tag{37}$$

is the desired least-squares solution of (36).

It is straightforward to establish that

$$UD^{-2}U^T = \frac{1}{2}(u^Tu I + uu^T) ,$$

where I is the $n \times n$ identity matrix. Using the Sherman-Morrison formula, we obtain

$$(UD^{-2}U^T)^{-1} = \frac{2}{d^2}I - \frac{1}{d^4}uu^T ,$$

where

$$d^2 = u^Tu .$$

Finally, using (37) and the definition of C, we obtain

$$C = \frac{vu^T + uv^T}{d^2} - \frac{v^Tuuu^T}{d^4} .$$

This result is the Powell symmetric Broyden update (see [3]).

5. NUMERICAL RESULTS

Some numerical experiments were performed for the nonlinear eigenvalue problem of Section 2 by the method described there, using the updating technique of Section 3 and Theorem 1. For these experiments it was more convenient to update the approximations to $d(1/\beta)/d\phi$ and $d(\alpha/\beta)/d\phi$ separately than to update a single approximation to $dE/d\phi$. The results for a typical example are given below.

The values of I and K are 8 and 16, respectively. Initially ϕ is set to a constant satisfying the normalization condition (4). Next, α and β are computed from ϕ, and λ is set to an initial value λ_0. As an initial approximation to the Jacobian we take the "linear part" of E_ϕ in (5), obtained by neglecting the ϕ dependence in α and β.

In the table below are tabulated the values of λ and the maximum absolute difference in ϕ obtained at each iteration for $\lambda_0 = 0.90$. The converged eigenvector ϕ has all its components positive, as would be predicted on physical grounds.

j	$\|\phi_j - \phi_{j-1}\|_\infty$	λ_j
0		0.90
1	1.35(+0)	0.9407976
2	5.16(-1)	0.9517847
3	7.95(-2)	0.9501302
4	9.69(-3)	0.9501903
5	8.91(-5)	0.9501893
6	1.94(-6)	0.9501893

By comparison, Newton's method, with $d(1/\beta)/d\phi$ and $d(\alpha/\beta)/d\phi$ obtained by differencing, required four iterations for λ to be accurate

to seven digits and for $\|\phi_j - \phi_{j-1}\|_\infty < 3.0(-4)$; substantially more computer time per iteration was required than for the update method because of the numerous additional evaluations of α and β.

ACKNOWLEDGMENTS

We wish to thank Alan Cline and John E. Dennis, Jr. for the interest they took in this paper during its presentation at the 1977 SIAM Fall Meeting and for the resulting improvements that arose from their comments. We thank also Richard R. Underwood for helpful comments made earlier. This work was performed in part under the auspices of the General Electric Company and of the U.S. Department of Energy.

REFERENCES

[1] C.G. Broyden, A Class of Methods for solving nonlinear simultaneous equations, Math. Comp., 19 (1965), pp. 577-593.

[2] L.K. Schubert, Modification of a quasi-Newton method for nonlinear equations with a sparse Jacobian, Math. Comp., 24 (1970), pp. 27-30.

[3] J.E. Dennis and J.J. Moré, Quasi-Newton methods, motivation and theory, SIAM Review, 19 (1977), pp. 46-89.

[4] S. Glasstone and A. Sesonske, Nuclear Reactor Engineering, Van Nostrand, 1967.

[5] P.M. Anselone and L.B. Rall, The solution of characteristic value-vector problems by Newton's method, Numerische Mathematik, 11 (1968), pp. 38-45.

This report was done with support from the Department of Energy. Any conclusions or opinions expressed in this report represent solely those of the author(s) and not necessarily those of The Regents of the University of California, the Lawrence Berkeley Laboratory or the Department of Energy.

TECHNICAL INFORMATION DEPARTMENT
LAWRENCE BERKELEY LABORATORY
UNIVERSITY OF CALIFORNIA
BERKELEY, CALIFORNIA 94720