MATHEMATICS OF COMPUTATION, VOLUME 31, NUMBER 140
OCTOBER 1977, PAGES 954–961

On Sparse and Symmetric Matrix Updating Subject to a Linear Equation

By Ph. L. Toint*

Abstract. A procedure for symmetric matrix updating subject to a linear equation and retaining any sparsity present in the original matrix is derived. The main feature of this procedure is the reduction of the problem to the solution of an n dimensional sparse system of linear equations. The matrix of this system is shown to be symmetric and positive definite. The method depends on the Frobenius matrix norm. Comments are made on the difficulties of extending the technique so that it uses more general norms, the main points being shown by a numerical example.

1. Introduction. Square matrix updating has become a very active field of research in linear algebra in the last few years, and its techniques are especially useful in algorithms for solving nonlinear systems of equations (see Broyden [1]) and in quasi-Newton methods for unconstrained optimization (see Davidon [2], Fletcher and Powell [3], Powell [7], Huang [6], for example). One common feature of these updating procedures is that the updated matrix satisfies a linear equation which, in the optimization field for example, has been called "quasi-Newton equation" or "DFP condition". Unfortunately, when the updated matrix is symmetric, these methods usually revise all the elements of the matrix; and therefore, the size of the problem that can be treated is often limited by the amount of computer storage that is available.

Different techniques have appeared for solving linear algebra problems of large dimension when their structure is sparse. For example, very good algorithms are now available to solve large and sparse systems of linear equations (see Reid [8]); and recently, Schubert presented in [9] a modification of Broyden's [1] method for solving nonlinear systems of equations which takes the sparsity of the problem into account. This method is of real interest but has the drawback that the resulting matrix is not symmetric, even when starting with a symmetric one. Therefore, its use is restricted to problems where the symmetry of the updated matrix is not important.

Most of the standard matrix updating techniques can be obtained by calculating the smallest correction matrix in an appropriate norm that causes the new matrix to satisfy some linear constraints; and this problem approach has some advantages in both theory and practice (see [4]). However, except for Schubert's method which

Received March 2, 1977.
AMS (MOS) subject classifications (1970). Primary 65F30; Secondary 15A24.
Key words and phrases. Matrix updating, quasi-Newton methods, unconstrained optimization.

*This work was done during a visit of the author in the Department of Applied Mathematics and Theoretical Physics, University of Cambridge, Cambridge (GB).

seems to be very successful, the linear constraints do not include sparsity conditions. We would like to keep the usual norms and find updating formulas for symmetric matrices that preserve known sparsity conditions. It is straightforward to pose this problem in a way that requires the solution of a large system of linear equations. However, we show that when the matrix norm is the Frobenius norm, we have only to solve a system that has as many variables as the dimension of the matrix to be updated, the matrix of this system being symmetric and positive definite and retaining the sparsity that is present in the original problem. Therefore, our results provide the possibility of solving very large nonlinear optimization calculations when the second derivative matrix has a known sparsity structure and is to be approximated by a symmetric matrix.

Section 2 of this paper presents a more detailed formulation of the problem and a practical updating algorithm. Section 3 is concerned with the formal derivation of the procedure while Section 4 discusses properties of the involved linear system. Additional remarks are given in Section 5.

2. Problem Formulation and Updating Procedure. Assume that A is an $n \times n$ sparse symmetric matrix of real numbers. Assume, moreover, that the sparsity conditions do not apply to the diagonal elements of A and that they are consistent with the symmetry of A. This paper is concerned with the problem of finding a matrix

$$A^* = A + E, \tag{1}$$

which is also symmetric ($A^{*T} = A^*$), which satisfies the condition

$$A^*x = y \tag{2}$$

for two given nonzero vectors x and y of R^n, and where the known sparsity structure that is obtained in A is preserved in A^*. We let the sparsity conditions be

$$A_{ij} = A^*_{ij} = 0 \qquad (i, j) \in I, \tag{3}$$

where I is a set of pairs of integers. We assume that the diagonal is not constrained by any sparsity conditions. We let J be the set of pairs of integers not belonging to I. Thus, $(i, i) \in J$ for all i. This assumption is made because this is the usual case, and it simplifies greatly the answer to the question whether a suitable A^* can be found. Since the conditions on A^* are generally not sufficient to determine it uniquely, we fix the remaining degrees of freedom by asking that the matrix A^* will be as close as possible to A with respect to the Frobenius norm, i.e. we minimize the expression

$$\|A - A^*\|_F \triangleq \left\{ \sum_{i=1}^{n} \sum_{j=1}^{n} (A_{ij} - A^*_{ij})^2 \right\}^{\frac{1}{2}}. \tag{4}$$

Therefore, to find the correction defined in Eq. (1) is to solve the problem

$$\tfrac{1}{2}\|E\|_F^2 \quad \text{is minimum} \tag{5}$$

subject to the linear constraints

$$Ex = y - Ax, \tag{6}$$

$$E_{ij} = 0, \quad (i, j) \in I, \tag{7}$$

$$E = E^T. \tag{8}$$

We now describe the recommended updating procedure. Its steps are justified in Section 3.

Define first, for $i = 1, \ldots, n$, the vectors $x(i)$ by the following formula

$$x(i)_j = \begin{cases} x_j, & (i, j) \in J, \\ 0, & (i, j) \in I. \end{cases} \tag{9}$$

Suppose for the moment that none of the vectors $x(i)$ are identically zero. Next build the matrix Q in the following way:

$$Q_{ij} = x(i)_j x(j)_i + \|x(i)\|^2 \delta_{ij} \quad \text{for } i = 1, \ldots, n; j = 1, \ldots, n, \tag{10}$$

where δ_{ij} is the Kronecker delta. We see that Q satisfies the sparsity conditions. It is also symmetric, and it is proved in Section 4 that it is positive definite. We calculate the vector λ by solving the linear system

$$Q\lambda = y - Ax. \tag{11}$$

The required correction may now be obtained from the simple formula

$$E_{ij} = \begin{cases} 0, & (i, j) \in I, \\ \lambda_i x_j + \lambda_j x_i, & (i, j) \in J, \end{cases} \tag{12}$$

and then A^* is defined by Eq. (1).

In the case where some of the vectors $x(i)$ are zero, we reduce the size of the problem. Specifically, if K is the set of values of i such that $x(i) = 0$, we set the ith row and column of E to zero, $i \in K$; and we use the formula (12) only for the values of i and j that are not in K. The corresponding values of λ are found from the linear equations that are obtained by deleting from the system (11) the ith row and column of Q, $i \in K$, and the ith component of the right-hand side, $i \in K$.

3. Justification of the Procedure. This section is devoted to the formal derivation of Eqs. (10), (11) and (12). It is convenient to let r be the vector

$$r = y - Ax. \tag{13}$$

By using (7) and (9), condition (6) may now be written as

$$\sum_{j=1}^{n} E_{ij} x(i)_j = r_i \quad \text{for } i = 1, \ldots, n. \tag{14}$$

We take symmetry into account by letting E have the form

$$E = \tfrac{1}{2}(B + B^T), \tag{15}$$

where B is now a matrix which does not need to be symmetric anymore. The whole

problem is now restated in terms of B as follows: find a matrix B such that

$$\frac{1}{8}\|B + B^T\|_F^2 \quad \text{is minimum} \tag{16}$$

subject to the conditions

$$\sum_{j=1}^{n} (B_{ij} + B_{ji})x(i)_j = 2r_i \quad \text{for } i = 1, \ldots, n. \tag{17}$$

Observe that condition (7) may be dropped because of the fact that (16), i.e. (5), will force to zero all the elements which do not appear explicitly in the constraints.

As in Greenstadt [5], let us write the Lagrangian function of the optimization problem (16)–(17):

$$\begin{aligned}\Phi(B, \lambda) = &\frac{1}{8}\sum_{i=1}^{n}\sum_{j=1}^{n} (B_{ij}^2 + B_{ji}^2 + 2B_{ij}B_{ji}) \\ &- \sum_{i=1}^{n} \lambda_i \left[\sum_{j=1}^{n} (B_{ij} + B_{ji})x(i)_j - 2r_i\right].\end{aligned} \tag{18}$$

Differentiation with respect to B_{ij} shows that we must satisfy the equation

$$\frac{\partial\Phi(B, \lambda)}{\partial B_{ij}} = \tfrac{1}{2}(B_{ij} + B_{ji}) - \lambda_i x(i)_j - \lambda_j x(j)_i = 0 \tag{19}$$

for $i = 1, \ldots, n$ and $j = 1, \ldots, n$.

Observe now that we may use (15) to rewrite (19) as

$$E_{ij} = \lambda_i x(i)_j + \lambda_j x(j)_i \quad \text{for } i, j = 1, \ldots, n, \tag{20}$$

and we may forget about the B matrix and use (14) in place of (17). This B matrix was just an artifact to differentiate the Lagrangian function of the problem in the space of symmetric matrices. Introducing now (20) in (14) yields

$$\sum_{j=1}^{n} [\lambda_i x(i)_j + \lambda_j x(j)_i]x(i)_j = r_i \quad \text{for } i = 1, \ldots, n, \tag{21}$$

which is

$$\lambda_i \sum_{j=1}^{n} [x(i)_j]^2 + \sum_{j=1}^{n} \lambda_j x(j)_i x(i)_j = r_i \quad \text{for } i = 1, \ldots, n. \tag{22}$$

We have found a linear system of equations in λ of the form

$$Q\lambda = r, \tag{23}$$

where the (i, j)th element of the matrix Q is defined by

$$Q_{ij} = x(i)_j x(j)_i + \sum_{k=1}^{n} [x(i)_k]^2 \delta_{ij}, \tag{24}$$

which is equivalent to the system (10)–(11). Since it is proved in the next section that Q is positive definite, the vector λ is well defined. It follows from Eq. (20) that E is given by

$$E_{ij} = (Q^{-1}r)_i x(i)_j + (Q^{-1}r)_j x(j)_i \tag{25}$$

for $i = 1, \ldots, n$ and $j = 1, \ldots, n$. This is exactly (12). We note that the symmetry and sparsity conditions are satisfied.

4. Properties of the Linear System $Q\lambda = r$. In order to solve the system (23), we need to be assured that it is nonsingular. The following theorem, by stating positive definiteness, ensures that there is a solution.

THEOREM 1. *If none of the vectors $x(i)$ $(i = 1, \ldots, n)$ are zero, then the matrix Q is positive definite, i.e.*

$$\forall z \in R^n;\ z \neq 0, \qquad z^T Q z > 0. \tag{26}$$

In order to prove (26), choose an arbitrary $z \in R^n$ which is not the zero vector. Then, by (24),

$$\begin{aligned} z^T Q z &= \sum_{i=1}^{n}\sum_{j=1}^{n} z_i Q_{ij} z_j = \sum_{i=1}^{n}\sum_{j=1}^{n} z_i x(i)_j x(j)_i z_j + \sum_{i=1}^{n}\sum_{k=1}^{n} [x(i)_k]^2 z_i^2 \\ &= \sum_{(i,j)\in J} [z_i x_i x_j z_j + z_i^2 x_j^2] = \frac{1}{2}\sum_{(i,j)\in J} [x_i z_j + x_j z_i]^2 \\ &= 2\sum_{i=1}^{n} z_i^2 x_i^2 + \frac{1}{2}\sum_{(i,j)\in J;\, i\neq j} (z_i x_j + z_j x_i)^2 \geqslant 0. \end{aligned} \tag{27}$$

Suppose $z^T Q z = 0$. Since z is not the zero vector, there exists a component of z, z_k say, that is nonzero, and the conditions

$$z_k x_k = 0, \tag{28}$$

$$z_k x_j + z_j x_k = 0, \qquad (i, j) \in J,\ j \neq k, \tag{29}$$

must be satisfied. It follows that x_k is zero and hence that $x_j = 0$, $(k, j) \in J$. But these conditions are equivalent to the statement that $x(k)$ is zero, which is a contradiction. Therefore the theorem is true.

COROLLARY. *The system* (23) *is singular if and only if at least one of the vectors $x(i)$ $(i = 1, \ldots, n)$ is zero.*

Proof. The "only if" statement has been proved already. To prove the reverse statement we suppose that $x(k) = 0$, and we let z be the kth coordinate vector in expression (27). One finds that $z^T Q z$ is zero, which completes the proof of the corollary.

We now justify the procedure that is given at the end of Section 2 for the case where some $x(i)$ are zero. Observe first that, if $x(k)$, say, is zero, then the left-hand side of (14) is zero when $i = k$. If r_k happens to be nonzero, then the constraints of the problem are incompatible. This may occur because of incorrect sparsity

requirements or because of rounding errors. Errors of this type cannot be corrected by the present calculation. Observe also that, because $x_k = 0$, the remaining components of r, namely r_i $(i \neq k)$, are independent of the kth column of E. Hence it is not helpful to admit nonzero elements into the kth row and column of E. We, therefore, satisfy condition (5) by setting this row and column to zero. This procedure may be repeated for each k such that $x(k)$ is zero, and the resulting reduced problem is nonsingular by Theorem 1.

Observe finally that, by (24), the matrix Q is also symmetric and has the same sparsity as the matrix A. Hence, algorithms for solving sparse symmetric and positive definite systems of linear equations may be used to solve (23). These algorithms are well developed and efficient (see [8], for example). The positive definiteness of Q allows the pivots of the procedure for solving the equations to be chosen from the diagonal.

5. Additional Remarks. The most useful feature of the proposed procedure is that the main part of the work is only to solve a linear system of n equations in n unknowns, with a positive definite matrix and all the sparsity that is present in the original problem. Therefore, very large systems may be treated.

It is also interesting to observe that the correction (12), sets the nonzero elements of E to those of a rank two matrix. It is likely that corrections of this type will provide several useful methods. Obviously, it would be even more valuable, as in [5] and [4], to take up the freedom in E by minimizing the expression

$$\|E\|_W^2 = \sum_{i=1}^{n} \sum_{j=1}^{n} (W^{1/2} E W^{1/2})_{ij}^2 \tag{30}$$

with W any symmetric positive definite matrix instead of the Frobenius norm (5). However, the fact that the elements B_{ij}, $(i, j) \in I$, are zero at the solution of the quadratic programming problem given by (16) and (17) is a consequence of the fact that the Frobenius norm is used. The linear constraints on E are the same as before, namely

$$\sum_{j=1}^{n} E_{ij} x(i)_j = r_i \quad \text{for } i = 1, \ldots, n, \tag{31}$$

$$E = E^T, \tag{32}$$

and

$$E_{ij} = 0 \quad \text{for } (i, j) \in I, \tag{33}$$

but now we have to introduce Lagrange multipliers for the sparsity conditions (33). By the same procedure as above, one can obtain, in correspondence with (20),

$$\begin{aligned} [WEW]_{ij} &= \lambda_i x(i)_j + \lambda_j x(j)_i, \qquad (i, j) \in J, \\ [WEW]_{ij} &= \gamma_{ij} + \gamma_{ji}, \qquad\qquad (i, j) \in I, \end{aligned} \tag{34}$$

where γ_{ij} are the Lagrange parameters corresponding to (33). Therefore, the substitution that gave Eq. (21) is no longer very useful. Observe that (34) shows that the

elements $[WEW]_{ij}$, $(i, j) \in J$, are the elements (i, j) of a rank two matrix, but that this property is not in general obtained for E, except in the case when there are no sparsity conditions. A numerical example will illustrate this point. Let $n = 5$ and define

$$W = \frac{1}{6}\begin{bmatrix} 4 & 0 & 0 & 0 & -2 \\ 0 & 3 & 0 & 0 & 0 \\ 0 & 0 & 3 & 0 & 0 \\ 0 & 0 & 0 & 3 & 0 \\ -2 & 0 & 0 & 0 & 4 \end{bmatrix}, \quad x = \begin{bmatrix} 1 \\ 0 \\ 1 \\ 0 \\ 1 \end{bmatrix} \quad \text{and} \quad r = \begin{bmatrix} 2 \\ 7 \\ 0 \\ 14 \\ 2 \end{bmatrix}.$$

Ask also that $E_{14} = E_{15} = E_{25} = 0$. The correction E must then be a quindiagonal matrix. In this case we find that the solution of the quadratic problem (31)–(33), is the correction:

$$E = \frac{1}{5}\begin{bmatrix} 6 & 15 & 4 & 0 & 0 \\ 15 & 0 & 20 & 0 & 0 \\ 4 & 20 & -8 & 40 & 4 \\ 0 & 0 & 40 & 0 & 30 \\ 0 & 0 & 4 & 30 & 6 \end{bmatrix},$$

which is not a rank two matrix since the upper left 3×3 determinant is equal to 1800/125. However, in agreement with (34), it may easily be verified that the quindiagonal part of

$$WEW = \frac{1}{30}\begin{bmatrix} 20 & 30 & 4 & -30 & -16 \\ 30 & 0 & 30 & 0 & -15 \\ 4 & 30 & -12 & 60 & 4 \\ -30 & 0 & 60 & 0 & 60 \\ -16 & -15 & 4 & 60 & 20 \end{bmatrix}$$

is a rank two matrix.

However, one easy generalization of the Frobenius norm is to use norms of the form

$$\|E\| = \left\{ \sum_{i=1}^{n} \sum_{j=1}^{n} t_{ij} E_{ij}^2 \right\}^{1/2}, \tag{35}$$

where the t_{ij} are arbitrary positive weighting factors. This form is obtained if the matrix W in expression (30) is diagonal. The formulae (10) and (12) become the equations

$$Q_{ij} = \frac{x(i)_j x(j)_i}{t_{ij}} + \sum_{k=1}^{n} \frac{[x(i)_k]^2}{t_{ik}} \delta_{ij}, \tag{36}$$

$$E_{ij} = \frac{1}{t_{ij}} [\lambda_i x(i)_j + \lambda_j x(j)_i] \tag{37}$$

for $i, j = 1, \ldots, n$, where λ is still the solution of the system $Q\lambda = r$. We note that the sparsity and positive definiteness of Q are preserved. One use of the weights t_{ij} is to allow differences in the scale of the variables.

Finally, if some of the diagonal elements of A and A^* are forced to be zero, then Theorem 1 is no longer true. However, because Eq. (27) still holds, the matrix Q is positive definite or positive semidefinite. It is now more difficult to recognize the semidefinite case at an early state. Fortunately, in the main application of this work, namely the estimation of second derivative matrices in unconstrained minimization calculations, we expect the diagonal elements of A to be nonzero. The applications to quasi-Newton methods in unconstrained optimization are fairly obvious. The proposed procedure provides a generalization of the method of Powell [7] to the sparse case. It is hoped that numerical experience will show a good behavior of this new algorithm on large sparse problems (problems occurring from the numerical solution of PDEs for example). Thus, the use of this sparse symmetric updating may provide the means of solving practical problems of very large dimensionality.

6. Conclusion. A symmetric and sparsity conserving matrix updating subject to a linear equation is proposed. The main feature of it is the reduction of the problem to the solution of an n dimensional sparse system of linear equations. Properties of the method are discussed and seem very encouraging. Interesting applications to quasi-Newton optimization methods may be expected in the near future.

Acknowledgement. The author wishes to express his gratitude to Professor M. J. D. Powell for having suggested the topic and for considerable help. The author also thanks the Royal Society and the European Science Exchange Program for their financial support.

Department of Mathematics
Facultés Universitaires Notre-Dame de la Paix
Namur, Belgium

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. W. C. DAVIDON, *Variable Metric Method for Minimization,* Report #ANL-5990 (Rev.), A.N.L. Research and Development Report, 1959.

3. 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.

4. D. GOLDFARB, "A family of variable-metric methods derived by variational means," *Math. Comp.*, v. 24, 1970, pp. 23–26. MR **41** #2896.

5. J. GREENSTADT, "Variations on variable-metric methods," *Math. Comp.*, v. 24, 1970, pp. 1–22. MR **41** #2895.

6. H. Y. HUANG, "Unified approach to quadratically convergent algorithms for function minimization," *J. Optimization Theory Appl.*, v. 5, 1970, pp. 405–423. MR **44** #6134.

7. M. J. D. POWELL, "A new algorithm for unconstrained optimization," *Nonlinear Programming* (J. B. ROSEN, O. L. MANGASARIAN & K. RITTER, Editors), Academic Press, New York, 1970, pp. 31–65. MR **42** #7043.

8. J. K. REID, *Two Fortran Subroutines for Direct Solution of Linear Equations Whose Matrix is Sparse, Symmetric and Positive Definite,* Report AERE-R. 7119, Harwell, 1972.

9. L. K. SCHUBERT, "Modification of a quasi-Newton method for nonlinear equations with a sparse Jacobian," *Math. Comp.*, v. 24, 1970, pp. 27–30. MR **41** #2923.