JOURNAL OF COMPUTATIONAL PHYSICS 84, 360–376 (1989)

Transformations of Matrices into Banded Form

LAURETTE S. TUCKERMAN

Center for Nonlinear Dynamics and Department of Physics, University of Texas, Austin, Texas 78712

Received March 11, 1988; revised September 13, 1988

A class of full or triangular matrices R is described for which there exist banded matrices B such that the product BR is also banded. The banded matrices yield recursion relations for solving systems of linear equations. Examples of such matrices (arising from second derivative operators acting on orthogonal function expansions) are used to illustrate the main theorem and its application. Practical considerations in efficient implementation are discussed. © 1989 Academic Press, Inc.

1. INTRODUCTION

The numerical solution of most scientific problems requires, at some stage, the solution of a system of linear equations. Typically, these equations arise from the discretization of differential operators in one or more dimension. In some cases, the operators are explicitly designed to have sparse or banded representations (e.g., one-dimensional finite difference derivative operators), and the algorithm for solution is straightforward. However, there are a number of matrix representations which are full (or nearly full) but which have special properties allowing rapid solution of their linear equation systems. These include Toeplitz and Vandermonde matrices [1], and fast Fourier transforms.

In this paper, we describe a different class of matrices whose operation count for solving the associated linear system can also be reduced. For each matrix R in this class, there exists a banded matrix B such that BR is also banded, with the same bandwidth as B. The advantages of constructing matrices B and BR are clear. Suppose that

$$Rf = g,$$

where R is an upper triangular N by N matrix (see Fig. 1). Calculating g from f (multiplication by R) or vice versa (backsolving with R) both require $O(N^2)$ steps. Storage of R requires $O(N^2)$ words of memory. Suppose we can write instead,

$$BRf = Bg,$$

where B and BR are upper triangular of bandwidth $J+1$, as in Fig. 1. Then to calculate g from f, we act with BR on f and then backsolve to get g. Both computations require only $O(JN)$ steps. To compute f from g, we reverse the roles of B and

0021-9991/89 $3.00

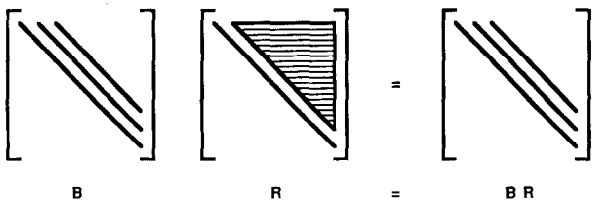

FIG. 1. The structure of the matrices B, R, and BR for an upper triangular matrix R in the case $J=2$.

BR, acting with B on g, then backsolving with BR. The matrices B and BR can be stored in $O(JN)$ words.

Another way of saying this is that there exists a recursion relation between $J+1$ terms of f and $J+1$ terms of g. To multiply f by the upper triangular matrix R, one would usually use all values of $f(m)$, $m \geq n$, in order to calculate the value $g(n)$. With a recursion relation, one would use instead a small number (here $J+1$) of values of f, together with a small number (here J) of previously calculated values of g (J initial conditions must also be specified to begin the recursion). An analogous statement can be made for solving, i.e., calculating f from g.

A rigorous statement and proof of our result will be given in Section 2; later, we will modify it to treat matrices of other kinds, such as lower triangular, infinite-dimensional, and full matrices, and also matrices with additional arbitrarily specified rows. Essentially, the requirement on the matrix R is that its off-diagonal nonzero elements be of the form

$$R(m, n) = S(m)\, T(n)$$

where S and T are N-vectors, or more generally,

$$R(m, n) = \sum_{j=1}^{J} S(m, j)\, T(j, n),$$

where S and T are N by J, and J by N matrices, respectively. Examples of such matrices abound. Here we present three examples arising from second derivative operators acting on orthogonal function expansions. For each operator, we also state the associated recursion relation, which we will derive in detail in Section 3.

EXAMPLE 1. The Laplacian in spherical coordinates contains the operator

$$\frac{1}{\sin\theta}\frac{d}{d\theta}\sin\theta\frac{d}{d\theta} - \frac{1}{\sin^2\theta}$$

which, when acting on the sine series $\sum_n f(n)\sin(n\theta)$, is represented by the matrix:

$$R(m, n) = \begin{cases} -2m & m < n \text{ and } m+n \text{ even} \\ -m(m+1) & m = n \\ 0 & \text{otherwise.} \end{cases} \tag{1.1}$$

Here $J=1$, $S(m,1)=-2m$, and $T(1,n)=1$. The recursion relation equivalent to $Rf=g$ is [2]:

$$(m+1)[f(m+2)-f(m)]=\frac{g(m)}{m}-\frac{g(m+2)}{m+2}.$$

EXAMPLE 2. The second derivative operator d^2/dx^2 acting on a series of Chebyshev polynomials $\sum_n f(n)\,T_n(x)$ is described by the matrix

$$R(m,n)=\begin{cases}(1/c_m)n(n^2-m^2) & m<n \text{ and } m+n \text{ even}\\ 0 & \text{otherwise,}\end{cases} \tag{1.2}$$

where $c_m=2$ for $m=0$, and $c_m=1$ otherwise. Here $J=2$. For $j=1$ we have $S(m,1)=1/c_m$ and $T(1,n)=n^3$. For $j=2$ we have $S(m,2)=m^2/c_m$ and $T(2,n)=-n$. The oft-used recursion relation [3, 4] equivalent to $Rf=g$ is

$$4m(m^2-1)\,f(m)=(m+1)\,c_{m-2}g(m-2)-2mg(m)+(m-1)\,g(m+2).$$

EXAMPLE 3. The operator

$$r\frac{d}{dr}r\frac{d}{dr}=r^2\frac{d^2}{dr^2}+r\frac{d}{dr}$$

arising in cylindrical coordinates acting on series of Chebyshev polynomials has the representation:

$$R(m,n)=\begin{cases}(1/c_m)n(n^2-m^2) & m<n \text{ and } m+n \text{ even}\\ m^2 & m=n\\ 0 & \text{otherwise.}\end{cases} \tag{1.3}$$

J, S and T are the same as for Example 2 and the recursion relation is:

$$\begin{aligned}&c_{m-2}(m+1)(m-2)^2\,f(m-2)-2m(2-m^2)\,f(m)+(m-1)(m+2)^2\,f(m-2)\\&\quad=(m+1)\,c_{m-2}g(m-2)-2mg(m)+(m-1)\,g(m+2).\end{aligned}$$

A recursion relation similar to those of Examples 2 and 3 has also been used to treat a spherical shell with a logarithmically transformed radial coordinate [5]. Additional examples can be found in [4, 6]. In fact, Examples 1–3 are all part of a general class in which the elements of R are sums of monomials:

EXAMPLE 4.

$$\begin{aligned}S(m,j)&=m^{\alpha_j}\\ T(j,n)&=n^{\beta_j},\end{aligned} \tag{1.4}$$

where the α_j's are all distinct.

Although the use of recursion relations for matrices of this form is widespread (e.g., [4, 6, 7]), their existence is often discovered on a case-by-case basis. This

leads to some misconceptions: for instance, it is sometimes stated (e.g., [5, 8]) that the only linear combinations of derivatives of Chebyshev series that engender recursion relations are those with constant coefficients, which is contradicted by Example 3. More general arguments, when cited to construct recursion relations, tend to rely on properties of orthogonal functions or other analytic properties specific to the application. The purely algebraic criterion presented here will illuminate recursion relations that are already known and provide new ones.

2. The Theorem and Proof

Our main result is:

THEOREM. *Let R be an upper triangular matrix of the form*

$$R(m,n)=\begin{cases}\sum_{j=1}^{J} S(m,j)T(j,n) & 1\leqslant m<n\leqslant N\\ R(m,n) & 1\leqslant m=n\leqslant N\\ 0 & \text{otherwise}\end{cases} \tag{2.1}$$

with $1\leqslant J<N$. Define the J by J matrices S_k and the vectors s_k of length J by

$$S_k(i,j)\equiv S(k+i,j) \tag{2.2a}$$

$$s_k(j)\equiv S(k,j) \tag{2.2b}$$

for $1\leqslant i,j\leqslant J$ and $k<N-J$ and suppose that s_k is in the column space of S_k^{T}. Then there exists an invertible banded matrix B, depending only on S, with J nonzero super-diagonals, such that BR is also banded with J nonzero super-diagonals.

The proof of the theorem relies on the reduction from N to J of the number of conditions necessary to transform a matrix to banded form.

Proof. If B is a banded matrix with J nonzero super-diagonals, so that $B(k,m)=0$ unless $k\leqslant m\leqslant k+J$, then BR is written

$$BR(k,n)=\sum_{i=0}^{\min(J,n-k)} B(k,k+i)R(k+i,n). \tag{2.3}$$

We need to show that $BR(k,n)=0$ for

$$1\leqslant n<k \tag{2.4a}$$

and for

$$k+J<n\leqslant N. \tag{2.4b}$$

Clearly for $n<k$, the sum in (2.3) is zero.

Suppose now that $k+J<N$. If $k+J<n\leqslant N$, then (2.3) becomes

$$BR(k,n)=\sum_{i=0}^{J} B(k,k+i)\sum_{j=1}^{J} S(k+i,j)T(j,n). \tag{2.5}$$

The sum in (2.5) will be zero if the $J+1$ values $B(k,k+i)$ (for k fixed, $0\leqslant i\leqslant J$) can be chosen so as to satisfy the J homogeneous equations:

$$\sum_{i=0}^{J} B(k,k+i)S(k+i,j)=0 \qquad \text{for} \quad 1\leqslant j\leqslant J. \tag{2.6}$$

Since B is upper triangular, its eigenvalues lie along the diagonal. Thus B will be invertible if $B(k,k)\neq 0$. Let us choose $B(k,k)=1$ and define the vector b_k consisting of the J remaining unknown super-diagonals by

$$b_k(i)\equiv B(k,k+i). \tag{2.7}$$

Recalling the definition of the vectors s_k and matrices S_k, (2.6) can be rewritten as

$$S_k^{\mathrm{T}} b_k=-s_k. \tag{2.8}$$

The conditions of the theorem on s_k and S_k guarantee the existence of a solution b_k to (2.8).

If $k+J\geqslant N$, then the range for (2.4b), $k+J<n\leqslant N$, is empty. Thus there are no conditions to impose and the $N-k+1$ values of $B(k,k+i)$ (for $0\leqslant i\leqslant N-k\leqslant J$), can be chosen arbitrarily. ∎

For completeness, we write the form of the nonzero entries of BR. For $k\leqslant n\leqslant k+J$, we have

$$\begin{aligned} BR(k,n)&=\sum_{i=0}^{n-k} B(k,k+i)R(m,n)\\ &=B(k,n)R(n,n)+\sum_{i=0}^{n-k-1} B(k,k+i)\sum_{j=1}^{J} S(k+i,j)T(j,n) \end{aligned}$$

which, using (2.6), can also be expressed as

$$BR(k,n)=B(k,n)R(n,n)-\sum_{i=n-k}^{J} B(k,k+i)\sum_{j=1}^{J} S(k+i,j)T(j,n). \tag{2.9}$$

3. Detailed Examples

The proof of the theorem is constructive; it provides an algorithm for constructing the matrix B via Eq. (2.8). We shall apply this algorithm to derive recursion relations associated with each of the examples of Section 1. However, the matrices

of Examples (1–3) are not exactly of the form specified in the theorem, because they preserve parity: $R(m, n) = 0$ unless $m + n$ is even. Such a matrix can be separated into two matrices, each involving only the odd or even indices, respectively. Effectively, the theorem is applied separately to each of the two $N/2$ by $N/2$ decoupled matrices; this, however, requires renaming the new matrices and indices. A simpler and equivalent procedure for treating parity-preserving matrices is to substitute

$$k + 2i \text{ for } k + i \quad \text{and} \quad k + 2J \text{ for } k + J$$
$$(n-k)/2 \text{ for } n-k \quad \text{and} \quad (N-k)/2 \text{ for } N-k,$$

wherever they appear in the theorem and proof. For example, definitions (2.2a) and (2.7) are changed to read

$$S_k(i, j) \equiv S(k + 2i, j)$$
$$b_k(i) = B(k, k + 2i),$$

so that the elements of b_k and of $S_k(., j)$ are all of the same parity as k, while definition (2.2b) of s_k is unchanged.

EXAMPLE 1. We recall that

$$R(m, n) = \begin{cases} -2m & m < n \text{ and } m + n \text{ even} \\ -m(m+1) & m = n \\ 0 & \text{otherwise,} \end{cases}$$

so that $J = 1$, $S(m, 1) = -2m$, and $T(1, n) = 1$.

According to the proof of the theorem, we find B by solving Eq. (2.8):

$$S_k^{\mathrm{T}} b_k = -s_k.$$

Since here $J = 1$, all of these quantities are scalars. In particular,

$$S_k \equiv S_k(1, 1) = S(k + 2, 1) = -2(k + 2) \quad s_k \equiv s_k(1) = S(k, 1) = -2k$$

Equation (2.8) then becomes

$$-2(k + 2)\, b_k = 2k,$$

whose solution is

$$b_k \equiv b_k(1) = -\frac{k}{k+2}.$$

Thus $B(k, k) = 1$, $B(k, k + 2) = b_k = -k/(k + 2)$, and all other elements of B are zero.

We then calculate the product matrix BR:

$$BR(k,n) = B(k,k)R(k,n) + B(k,k+2)R(k+2,n)$$

$$= R(k,n) - \frac{k}{k+2} R(k+2,n)$$

$$= \begin{cases} 0 & \text{for } n < k \\ -k(k+1) & \text{for } n = k \\ -2k - \dfrac{k}{k+2}(-(k+2)(k+3)) & \\ \quad = -2k + k(k+3) = k(k+1) & \text{for } n = k+2 \\ -2k - \dfrac{k}{k+2}(-2(k+2)) = 0 & \text{for } n > k+2 \\ 0 & \text{for } n+k \text{ odd.} \end{cases}$$

We see that BR indeed has the same banded structure as B.

The recursion relation equivalent to $Rf = g$ is derived by writing out the components of the "banded equation" $BRf = Bg$:

$$\sum_m (BR)(k,m) f(m) = \sum_m B(k,m) g(m)$$

$$BR(k,k) f(k) + BR(k,k+2) f(k+2) = B(k,k) g(k) + B(k,k+2) g(k+2)$$

$$-k(k+1) f(k) + k(k+1) f(k+2) = g(k) - \frac{k}{k+2} g(k+2),$$

which can be simplified to:

$$(k+1)[f(k+2) - f(k)] = \frac{g(k)}{k} - \frac{g(k+2)}{(k+2)}.$$

EXAMPLE 2. We recall that

$$R(m,n) = \begin{cases} (1/c_m) n(n^2 - m^2) & m < n \text{ and } m+n \text{ even} \\ 0 & \text{otherwise} \end{cases}$$

(where $c_m = 2$ for $m = 0$ and $c_m = 1$ otherwise), so that $J = 2$ and

$$j = 1: \qquad S(m,1) = \frac{1}{c_m}, \qquad T(1,n) = n^3$$

$$j = 2: \qquad S(m,2) = \frac{m^2}{c_m}, \qquad T(2,n) = -n.$$

This example is slightly more complicated because here $J=2$; so that in Eq. (2.8), S_k^{T} is a 2 by 2 matrix, while the right-hand side s_k and the unknown b_k are vectors of length 2. In particular,

$$S_k(i,1)=S(k+2i,1)=\frac{1}{c_{k+2i}}=1$$

$$S_k(i,2)=S(k+2i,2)=\frac{(k+2i)^2}{c_{k+2i}}=(k+2i)^2$$

$$s_k(1)=S(k,1)=\frac{1}{c_k}$$

$$s_k(2)=S(k,2)=\frac{k^2}{c_k}.$$

Equation (2.8) becomes

$$\begin{pmatrix}1 & 1\\ (k+2)^2 & (k+4)^2\end{pmatrix}\begin{pmatrix}b_k(1)\\ b_k(2)\end{pmatrix}=-\begin{pmatrix}\dfrac{1}{c_k}\\ \dfrac{k^2}{c_k}\end{pmatrix}.$$

This equation is easily solved by inverting S_k^{T}:

$$\begin{pmatrix}b_k(1)\\ b_k(2)\end{pmatrix}=\frac{-1}{(k+4)^2-(k+2)^2}\begin{pmatrix}(k+4)^2 & -1\\ -(k+2)^2 & 1\end{pmatrix}\begin{pmatrix}\dfrac{1}{c_k}\\ \dfrac{k^2}{c_k}\end{pmatrix}$$

$$=\frac{-1}{c_k[(k+4)^2-(k+2)^2]}\begin{pmatrix}(k+4)^2-k^2\\ -(k+2)^2+k^2\end{pmatrix}$$

$$=\frac{-1}{c_k(4k+12)}\begin{pmatrix}8k+16\\ -4k-4\end{pmatrix}=\frac{1}{c_k(k+3)}\begin{pmatrix}-2(k+2)\\ k+1\end{pmatrix}.$$

Thus,

$$\begin{aligned}B(k,k)&=1\\ B(k,k+2)&=\frac{-2(k+2)}{c_k(k+3)}\\ B(k,k+4)&=\frac{k+1}{c_k(k+3)}.\end{aligned}\tag{3.1}$$

We now calculate the product matrix BR:

$$BR(k,n)=B(k,k)R(k,n)+B(k,k+2)R(k+2,n)+B(k,k+4)R(k+4,n).$$

For $n \leqslant k$,

$$BR(k, n) = 0.$$

For $n = k + 2$,

$$BR(k, k+2) = B(k, k) R(k, k+2) = \frac{1}{c_k} (k+2)((k+2)^2 - k^2)$$

$$= \frac{4}{c_k} (k+2)(k+1).$$

For $n = k + 4$,

$$BR(k, k+4) = B(k, k) R(k, k+4) + B(k, k+2) R(k+2, k+4)$$

$$= \frac{1}{c_k} (k+4)((k+4)^2 - k^2)$$

$$+ \frac{-2(k+2)}{c_k(k+3)} \frac{1}{c_{k+2}} (k+4)((k+4)^2 - (k+2)^2)$$

$$= \frac{1}{c_k} (k+4)(8k+16) + \frac{-2(k+2)}{c_k(k+3)} (k+4)(4k+12)$$

$$= \frac{8}{c_k} [(k+4)(k+2) - (k+2)(k+4)] = 0.$$

Finally, for $n > k + 4$, the construction of the theorem guarantees the vanishing of $BR(k, n)$:

$$BR(k, n) = B(k, k) R(k, n) + B(k, k+2) R(k+2, n)$$

$$+ B(k, k+4) R(k+4, n)$$

$$= \frac{1}{c_k(k+3)} [(k+3) n(n^2 - k^2) - 2(k+2) n(n^2 - (k+2)^2)$$

$$+ (k+1) n(n^2 - (k+4)^2)]$$

$$= \frac{n}{c_k(k+3)} [((k+3) - 2(k+2) + (k+1))(n^2 - k^2)$$

$$- 2(k+2)(-4k-4) + (k+1)(-8k-16)]$$

$$= \frac{n}{c_k(k+3)} [8(k+2)(k+1) - 8(k+1)(k+2)] = 0.$$

In this example, BR is sparser than B. The diagonal element $BR(k, k)$ vanishes because $BR(k, k) = R(k, k) = 0$ for this matrix. The second super-diagonal

$BR(k, k+4)$ vanishes for a more subtle reason. For parity-preserving matrices, Eq. (2.9) becomes

$$BR(k, n) = B(k, n) R(n, n) - \sum_{i=(n-k)/2}^{J} B(k, k+2i) \sum_{j=1}^{J} S(k+2i, j) T(j, n).$$

We substitute $n = k+4$ to obtain

$$BR(k, k+4) = -B(k, k+4) \sum_{j=1}^{J} S(k+4, j) T(j, k+4).$$

For this particular matrix, it turns out that $\sum_{j=1}^{J} S(n, j) T(j, n) = 0$ for any n, because

$$S(n, 1) T(1, n) = n^3/c_n = -S(n, 2) T(2, n).$$

Thus, $BR(k, k+4) = 0$ as well.

The recursion relation equivalent to $Rf = g$ is:

$$\begin{aligned} BR(k, k+2) f(k+2) &= B(k, k)\, g(k) + B(k, k+2)\, g(k+2) \\ &\quad + B(k, k+4)\, g(k+4) \\ \frac{4}{c_k}(k+1)(k+2) f(k+2) &= g(k) + \frac{-2(k+2)}{c_k(k+3)} g(k+2) \\ &\quad + \frac{k+1}{c_k(k+3)} g(k+4). \end{aligned}$$

By substituting $m = k+2$,

$$\begin{aligned} \frac{4}{c_{m-2}}(m-1)\, m f(m) &= g(m-2) + \frac{-2m}{c_{m-2}(m+1)} g(m) \\ &\quad + \frac{m-1}{c_{m-2}(m+1)} g(m+2) \end{aligned}$$

and, dividing through by the coefficient of $f(m)$, we recover the more familiar form [3, 4]:

$$\begin{aligned} f(m) &= \frac{c_{m-2}}{4m(m-1)} g(m-2) - \frac{1}{2(m-1)(m+1)} g(m) \\ &\quad + \frac{1}{4m(m+1)} g(m+2) \end{aligned}$$

or

$$4m(m^2-1) f(m) = c_{m-2}(m+1)\, g(m-2) - 2m\ g(m) + (m-1)\, g(m+2). \qquad (3.2)$$

EXAMPLE 3. For this operator,

$$R(m,n)=\begin{cases}(1/c_m)n(n^2-m^2) & m<n \text{ and } m+n \text{ even}\\ m^2 & m=n\\ 0 & \text{otherwise.}\end{cases}$$

The off-diagonal elements of this matrix are identical to those of Example 2. The matrix B is again defined by (3.1) since the construction of B depends only on S. The product matrix BR will, however, be different:

$$BR(k,k)=k^2$$

$$BR(k,k+2)=\frac{2(k+2)}{c_k(k+3)}(k^2+4k+2)$$

$$BR(k,k+4)=\frac{k+1}{c_k(k+3)}(k+4)^2.$$

The recursion relation equivalent to $Rf=g$ is

$$\begin{aligned}k^2f(k)&+\frac{2(k+2)}{c_k(k+3)}(k^2+4k+2)\,f(k+2)\\&+\frac{k+1}{c_k(k+3)}(k+4)^2\,f(k+4)\\&=g(k)+\frac{-2(k+2)}{c_k(k+3)}\,g(k+2)+\frac{k+1}{c_k(k+3)}\,g(k+4).\end{aligned}$$

Again substituting $m=k+2$, and multiplying both sides by $c_{m-2}(m+1)$, we obtain

$$\begin{aligned}&c_{m-2}(m+1)(m-2)^2\,f(m-2)+2m(m^2-2)\,f(m)+(m-1)(m+2)^2\,f(m+2)\\&\quad=(m+1)\,c_{m-2}\,g(m-2)-2m\;\;g(m)+(m-1)\;g(m+2).\end{aligned}$$

The right-hand side (Bg) of this recursion relation is the same as that of (3.2).

4. DISCUSSION AND EXTENSIONS

In this section we will discuss the theorem proved above and modify it to treat matrices of other kinds, such as lower triangular, infinite-dimensional, and full matrices, and also matrices with additional arbitrarily specified rows. We will conclude the article by considering some points related to practical implementation, such as the Sherman–Morrison–Woodbury formula, LU decomposition, and permutation.

The matrix R need not be invertible for the theorem to hold, though the invertibility of BR depends on that of R. The bandwith of BR may be less than $J+1$ (as

we have seen in Example 2): in special cases, BR may be the identity, so that $B = R^{-1}$. A complete characterization of matrices whose inverses are banded is given in [9].

The sum of two matrices of form (2.1) (with J-values J_1 and J_2) is also of the same form. In general, the S and T matrices of the sum will have $J = J_1 + J_2$. This expanded S matrix may not satisfy the hypothesis of the theorem, but it is often possible to regroup terms, forming a smaller S with $J < J_1 + J_2$ to which the theorem is applicable.

We briefly discuss the hypothesis of the theorem. Note that s_k is in the column space of S_k^{T} if and only if s_k is orthogonal to every element (if any) of the null space of S_k; this provides an equivalent condition. A different condition on S, which is stronger than the hypothesis of the theorem but more easily understood, is to simply required that the matrices S_k be invertible for $k < N - J$; Examples 1–4 actually satisfy this stronger condition. For Example 4, this follows from the linear independence of the monomials m^{α_j} *over any interval of* m: the theorem thus applies for any value of J, and any set of distinct exponents α_j.

We now present some variations of the theorem. Many suggest themselves, the most obvious being:

COROLLARY. *Let* R *be a* lower *triangular matrix of the form*:

$$R(m, n) = \begin{cases} \sum_{j=1}^{J} S(m, j)\, T(j, n) & 1 \leqslant n < m \leqslant N \\ R(m, n) & 1 \leqslant n = m \leqslant N \\ 0 & \text{otherwise.} \end{cases} \tag{4.1}$$

Define the vectors s_k as before, but the matrices S_k by

$$S_k(i, j) \equiv S(k - i, j)$$

and suppose again that s_k is in the column space of S_k^{T}. Then there exists an invertible banded matrix B, depending only on S, with J nonzero sub-diagonals, such that BR is also banded with J nonzero sub-diagonals.

The proof of the theorem requires little modification. For $k > J + 1$, we must solve

$$B(k, k)\, S(k, j) + \sum_{i=1}^{J} S(k - i, j)\, B(k, k - i) = 0 \qquad \text{for} \quad 1 \leqslant j \leqslant J.$$

Again choosing $B(k, k) = 1$, and defining $b_k(i) \equiv B(k, k - i)$, this reduces to Eq. (2.8). For $k \leqslant J + 1$, the values of $B(k, k - i)$, $0 \leqslant i \leqslant k$, can be chosen arbitrarily.

The result can easily be extended to infinite dimensional arrays. The arrays can be either singly or doubly infinite, e.g., $[1, \infty]$ or $[-\infty, \infty]$. The only change is the absence of the arbitrary rows $N - J \leqslant k$ (for the upper triangular case), when there is an infinite upper bound, and $k \leqslant J + 1$ (for the lower triangular case), when there is an infinite lower bound. These rows $B(k, \cdot)$ are then specified, like the other

rows, as a nontrivial $(J+1)$-dimensional solution of a J-dimensional homogeneous system of equations.

Other modifications of the matrix R yield matrices which are banded on all but a few rows. Let us try to follow the proof of the theorem for a matrix R which is neither upper nor lower triangular:

$$R(m,n)=\begin{cases}\sum_{j=1}^{J} S(m,j)T(j,n) & m\neq n\\ R(m,n) & m=n.\end{cases} \tag{4.2}$$

Since R is neither upper nor lower triangular, we will let B be a general banded matrix with J_1 nonzero sub-diagonals and J_2 nonzero super-diagonals, where $J_1+J_2=J$. Then,

$$BR(k,n)=\sum_{m=\max(k-J_1,1)}^{\min(k+J_2,N)} B(k,m)R(m,n). \tag{4.3}$$

We would like $BR(k,n)$ to vanish for

$$1\leqslant n<k-J_1 \tag{4.4a}$$

and for

$$k+J_2<n\leqslant N. \tag{4.4b}$$

Consider first k such that $1\leqslant k-J_1$ and $k+J_2\leqslant N$. Then if n is in either range (4.4a) or range (4.4b), it is outside the limits of the sum in (4.3). Equation (4.3) becomes

$$BR(k,n)=\sum_{m=k-J_1}^{k+J_2} B(k,m)\sum_{j=1}^{J} S(m,j)T(j,n)$$

which will be zero if

$$\sum_{m=k-J_1}^{k+J_2} B(k,m)S(m,j)=0 \qquad \text{for} \quad 1\leqslant j\leqslant J. \tag{4.5}$$

To solve (4.5) for $B(k,\cdot)$, the vectors s_k must again be orthogonal to all null vectors of certain J by J matrices. (Care must be taken to ensure that B is invertible).

Now consider k such that $N<k+J_2$. We get, instead,

$$BR(k,n)=\sum_{m=k-J_1}^{N} B(k,m)R(m,n). \tag{4.6}$$

The range (4.4b) is empty, but the range (4.4a) is not. The number of unknown elements $B(k,\cdot)$ is insufficient to make (4.6) vanish, and therefore $BR(k,n)\neq 0$ for

$$1\leqslant n<k-J_1 \qquad \text{and} \qquad N<k+J_2. \tag{4.7a}$$

For $k - J_1 < 1$, by a similar argument, $BR(k, n) \neq 0$ for

$$k - J_1 < 1 \qquad \text{and} \qquad k + J_2 < n \leqslant N. \tag{4.7b}$$

Thus there are $J_1 + J_2 = J$ rows for which the matrix BR cannot be made banded. If we choose B to be upper triangular, i.e, $J_1 = 0$, as in Fig. 2a, then range (4.7b) does not exist, and BR is banded except for the J bottommost rows. If B is made lower triangular, range (4.7a) is empty and the nonzero rows are all located at the top, as in Fig. 2b.

However, the offending ranges disappear when R is an infinite-dimensional array. If R has an infinite upper bound, range (4.7a) does not exist, so an upper triangular B makes BR banded for all rows. Similarly, if R has an infinite *lower* bound and B is taken to be *lower* triangular, then BR has bandwith $J + 1$ everywhere.

The remaining nonzero entries of BR (i.e., for $1 \leqslant k - J_1 \leqslant n \leqslant k + J_2 \leqslant N$) are

$$\begin{aligned} BR(k, n) &= B(k, n) R(n, n) + \sum_{m = k - J_1}^{k + J_2} B(k, m) \sum_{j = 1}^{J} S(m, j) T(j, n) \\ &\quad - B(k, n) \sum_{j = 1}^{J} S(n, j) T(j, n) \\ &= B(k, n) \left(R(n, n) - \sum_{j = 1}^{J} S(n, j) T(j, n) \right). \end{aligned}$$

Another method is also available for inverting matrices (or solving linear systems) when R is of form (4.2). The Sherman–Morrison–Woodbury formula [10, 11, 1] provides a simple and powerful procedure for computing the inverse of a matrix from the inverse of a closely related matrix. Note that R is the sum of the diagonal matrix D defined by

$$D(n, n) \equiv R(n, n) - \sum_{j = 1}^{J} S(n, j) T(j, n)$$

and of the rank-J matrix ST:

$$(ST)(m, n) = \sum_{j = 1}^{J} S(m, j) T(j, n).$$

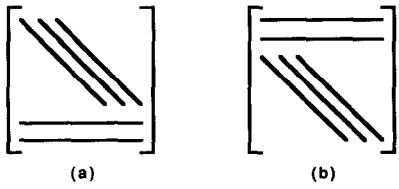

FIG. 2. The structure of the matrix BR for a full matrix R of form (4.2). (a) B has been chosen to be upper triangular ($J_1 = 0$ and $J_2 = J = 2$); (b) B is lower triangular ($J_2 = 0$ and $J_1 = J = 2$).

Then the inverse of R can be obtained from the inverses of D and of the J by J matrix,

$$C \equiv -(I + TD^{-1}S)$$

by the Sherman–Morrison–Woodbury formula:

$$R^{-1} = (D + ST)^{-1} = D^{-1} + D^{-1}SC^{-1}TD^{-1}.$$

Unlike the full matrix, the upper and lower triangular matrices (2.1) and (4.1) cannot be expressed as sums of diagonal and rank-J matrices.

The last matrix we treat is upper triangular except for the last M rows which are full:

$$R(m,n) = \begin{cases} \sum_{j=1}^{J} S(m,j)T(j,n) & m < n \text{ and } m \leqslant N-M \\ R(m,m) & m = n \\ R(m,n) & N-M < m \\ 0 & \text{otherwise,} \end{cases} \tag{4.8}$$

as shown in Fig. 3. Such matrices arise when boundary conditions are imposed on differential equations. Multiplying R by an upper triangular matrix B of bandwidth $J+1$ and adapting the proof of the theorem, we find that the matrix BR has bandwidth $J+1$ for rows $1 < k \leqslant N-M-J$. It has at most $J+M$ nonzero entries in rows $N-M-J < k \leqslant N-M$, and N entries in the last rows $N-M < k \leqslant N$, as in Fig. 3. If only one row is changed ($M=1$), then the bandwidth of BR remains $J+1$ except for the last unavoidably full row.

The consequences of adding $M \ll N$ full rows to an otherwise banded matrix BR need not be catastrophic. To take advantage of its sparse structure, we act with $(BR)^{-1}$ by backsolving if BR is triangular, and by LU or UL decomposition if it is not. In an otherwise banded matrix, sparseness is preserved by:

(1) LU decomposition when there are full rows at the bottom and/or columns on the right;

(2) UL decomposition when there are full rows at the top and/or columns on the left.

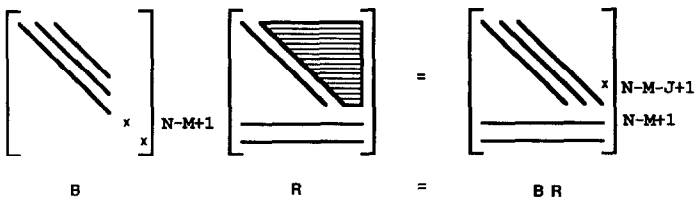

FIG. 3. The structure of the matrices B, R, and BR for a matrix R of form (4.8). R is upper triangular except for the M bottommost rows which are full. Here $M = J = 2$. BR has bandwidth $J+1$ in rows 1 through $N-M-J$, bandwidth $J+M$ in rows $N-M-J+1$ through $N-M$, and bandwidth N in rows $N-M+1$ through N.

In order to invert BR while taking advantage of its sparseness, it is necessary for BR to be diagonally dominant. Otherwise pivoting (which destroys sparseness) is necessary for numerical stability. It is often possible to make BR diagonally dominant by permuting its rows (partial pivoting), which may perturb its triangularity and may slightly increase the number of non-banded rows. Consider the second derivative matrix of Example 3, operating on Chebyshev polynomials of *either* odd *or* even parity, whose last row has been replaced by a full row corresponding to boundary conditions. (The standard case [4, 6] is the sum of the matrix of Example 2 and a multiple of the identity.) The banded matrix BR resulting from direct application of the theorem is upper triangular with an extra full row at the bottom, as shown in Fig. 4a, and is not diagonally dominant. However, a cyclic permutation of rows leads to a diagonally dominant matrix which is tridiagonal except for one full row at the top, as in Fig. 4b.

If full rows occur at both the top and bottom of the matrix, neither LU nor UL decomposition preserves sparseness. The Sherman–Morrison–Woodbury formula again suggests itself, since a matrix which is banded except on J arbitrarily distributed rows can be written as the sum of a banded matrix and of a rank-J matrix. However, the inverse of even a perfectly banded matrix is generally full and so should not be directly computed. The Sherman–Morrison–Woodbury formula can instead be used, in conjunction with the LU (or UL) decomposition of the perfectly banded matrix, to formulate a Greens function-like procedure [12, 1] for acting with $(BR)^{-1}$ on a vector.

Finally, a matrix R of any of the forms discussed above can be used as an economical preconditioner for the iterative solution of more general linear systems.

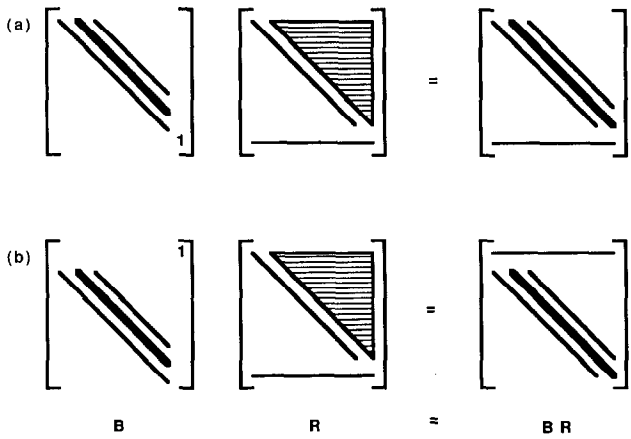

FIG. 4. The structure of the matrices B, R, and BR for an upper triangular matrix R with one full row at the bottom. The thick lines in B and BR indicate the maximum element of each row. In (b) the rows of B (and BR) have been permuted to yield diagonally dominant matrices.

ACKNOWLEDGMENTS

We thank Dwight Barkley, Richard Friesner, and Eric Kostelich for their suggestions. The author holds an NSF Mathematical Sciences Postdoctoral Research Fellowship, and began this work while at the Centre d'Etudes Nucleaires de Saclay (France). This work was partially supported by the Office of Naval Research Nonlinear Dynamics Program.

REFERENCES

1. W. PRESS, B. FLANNERY, S. TEUKOLSKY, AND W. VETTERLING, *Numerical Recipes: The Art of Scientific Computing* (Cambridge Univ. Press, Cambridge, 1986), pp. 43, 66.
2. P. S. MARCUS AND L. S. TUCKERMAN, *J. Fluid Mech.* **185**, 1 (1987).
3. D. ELLIOT, *Proc. Cambridge Philos. Soc.* **57**, 823 (1961).
4. D. GOTTLIEB AND S. A. ORSZAG, *Numerical Analysis of Spectral Methods* (SIAM, Philadelphia, 1977), pp. 119, 160.
5. S. C. R. DENNIS AND L. QUARTAPELLE, *J. Comput. Phys.* **61**, 218 (1985).
6. C. CANUTO, M. Y. HUSSAINI, A. QUATERONI, AND T. A. ZANG, *Spectral Methods in Fluid Dynamics* (Springer-Verlag, New York, 1988), pp. 127, 229.
7. C. W. CLENSHAW, *Proc. Cambridge Philos. Soc.* **53**, 134 (1957).
8. A. T. PATERA AND S. A. ORSZAG, *J. Fluid Mech.* **112**, 467 (1981).
9. W. W. BARRETT AND P. J. FEINSILVER, *Linear Algebra Appl.* **41**, 111 (1981).
10. J. SHERMAN AND W. J. MORRISON, *Ann. Math. Statist.* **20**, 261 (1949).
11. M. WOODBURY, Memorandum Report 42, Statistical Research Group, Princeton, 1950 (unpublished).
12. L. S. TUCKERMAN, *J. Comput. Phys.* **80**, 403 (1989).