Groups of Banded Matrices with Banded Inverses

Gilbert Strang

Department of Mathematics, MIT

Abstract

A product $A = F_1 \ldots F_N$ of invertible block-diagonal matrices will be banded with a banded inverse. We establish this factorization with the number N controlled by the bandwidths w and not by the matrix size n. When A is an orthogonal matrix, or a permutation, or banded plus finite rank, the factors F_i have $w = 1$ and generate that corresponding group. In the case of infinite matrices, conjectures remain open.

1 Introduction

Banded matrices with banded inverses are unusual, but these exceptional matrices do exist. Block diagonal matrices are the first examples. Products of block diagonal matrices give many more examples (including useful ones). The main theorem in an earlier paper [10] is that all examples are produced this way, from multiplying invertible block diagonal matrices of bandwidth 1.

When A and B are banded with banded inverses, A^{-1} and AB also have those properties. This group of matrices is all of $\mathrm{GL}(n)$ in the finite case (every matrix has bandwidth less than n). For singly infinite matrices this appears to be a new group. In both cases the key point of the theorem is that the number N of block diagonal factors is controlled by the bandwidth w and *not by the matrix size* n.

Theorem 1 The factors in $A = F_1 F_2 \ldots F_N$ can be chosen block diagonal, with 2 by 2 and 1 by 1 blocks. Then each generator F and F^{-1} has bandwidth ≤ 1.

Here w is the larger of the bandwidths of A and A^{-1}. So $A_{ij} = 0$ and $\left(A^{-1}\right)_{ij} = 0$ for $|i - j| > w$. The number of factors F could be as large as $C w^2$ (just to carry out ordinary elimination) but N does not increase with n.

Important banded matrices with banded inverses arise in constructing orthogonal polynomials on the unit circle [5]. They also yield filter banks with perfect reconstruction, the key to wavelets. Those are block Toeplitz matrices in the wavelet case, and "CMV matrices" in other cases. Our earlier paper [10] applied an observation of Ilya Spitkovsky to separate these matrices into the minimum number $N = w$ of factors F (each with bandwidth 1). We believe that these special (and short) factorizations can lead to useful algorithms.

For other A, the main step of the proof is to reach $A = BC$, two factors with diagonal blocks of size $2w$. One of those factors begins with a block of size w, and it is this "offset" between the block positions in B and C that makes the proof work.

The form of this "offset product" is important even for $w = 1$:

$$BC = \begin{bmatrix} 1 & & & & & & \\ & 2 & 3 & & & & \\ & 4 & 5 & & & & \\ & & & 6 & 7 & & \\ & & & 8 & 9 & & \\ & & & & & 10 & \end{bmatrix} \begin{bmatrix} 1 & 2 & & & & & \\ 3 & 4 & & & & & \\ & & 5 & 6 & & & \\ & & 7 & 8 & & & \\ & & & & 9 & 10 & \\ & & & & 11 & 12 & \end{bmatrix}. \tag{1}$$

The second row of BC has 2 times $[3 \;\; 4]$ followed by 3 times $[5 \;\; 6]$. The third row of BC has 4 times $[3 \;\; 4]$ followed by 5 times $[5 \;\; 6]$. A pair of *singular matrices* lies side by side in the product BC :

$$\begin{bmatrix} 1 & 2 & & & \\ 6 & 8 & 15 & 18 & \\ 12 & 16 & 25 & 30 & \\ & & 42 & 48 & \cdot \;\; \cdot \\ & & 56 & 64 & \cdot \;\; \cdot \\ & & & & \cdot \;\; \cdot \end{bmatrix} = \begin{bmatrix} 1 & [1 \;\; 2] & & & \\ \begin{bmatrix} 2 \\ 4 \end{bmatrix} & [3 \;\; 4] & \begin{bmatrix} 3 \\ 5 \end{bmatrix} & [5 \;\; 6] & \\ & & \begin{bmatrix} 6 \\ 8 \end{bmatrix} & [7 \;\; 8] & \cdot \;\; \cdot \\ & & & & \cdot \;\; \cdot \\ & & & & \cdot \;\; \cdot \end{bmatrix}$$

When a column of B multiplies a row of C, the nonzeros sit in a 2 by 2 matrix *of rank* 1. These 2 by 2 matrices don't overlap in BC. So when those column-row products are added (a legal way to compute BC), the result is a "*CMV matrix*." This matrix BC has *two diagonals of singular* 2 *by* 2 *blocks*.

The inverse matrix $(BC)^{-1} = C^{-1}B^{-1}$ has a similar multiplication in the opposite order. The pattern of nonzeros is just the transpose of the pattern above. This product is another CMV matrix (singular 2 by 2 blocks side by side).

This paper will describe a corresponding factorization for other (smaller or larger) groups of matrices. Here are four of those groups, not an exhaustive list :

1. **Banded orthogonal matrices**. The inverse is the transpose (so automatically banded). The factors F will now be orthogonal matrices with $w = 1$: block diagonal with small orthogonal blocks.

2. **Banded permutation matrices**. In this case $w = \max|i - p(i)|$ measures the maximum movement from i in $(1,\dots,n)$ to $p(i)$ in $(p(1),\dots,p(n))$. Each factor F is then a permutation with $w = 1$; it exchanges disjoint pairs of neighbors.

   We conjectured that fewer than $2w$ factors F would be sufficient. Greta Panova has found a beautiful proof [6]. Other constructions [1, 9] also yield $N \le 2w - 1$.

3. **Banded plus finite rank**. For infinite matrices A, banded with banded inverse, we enlarge to a group B by including also $A + Z$. The perturbation Z allows any matrix of finite rank such that $A + Z$ is invertible. Then its inverse will be $A^{-1} + Y$, also perturbed with rank $(Y) \le \text{rank}(Z)$, and we have a group.

   In this case, we include the new factors $F = I + (\text{rank } 1)$ along with the block diagonal F's. These factors generate the enlarged group.

4. **Cyclically banded matrices**. Cyclic means that "n is adjacent to 1." The distance from diagonal i to diagonal j is the smaller of $|i-j|$ and $n-|i-j|$. The cyclic bandwidth of A is the largest distance for which $A_{ij} \neq 0$. Then w_c is the larger of the cyclic bandwidths of A and A^{-1}.

   The natural conjecture is $A = F_1 \dots F_N$ with factors that have cyclic bandwidth 1 (so that F_{1n} and F_{n1} may be nonzero). The number N should be controlled by w_c. No proof of this conjecture is to be found in the present paper.

2 Banded orthogonal matrices

We need to recall how to obtain $A = BC$ with block diagonal factors B and C. We construct *orthogonal* B and C when A is orthogonal. Then the final step will separate B and C into orthogonal factors with $w = 1$.

The key is to partition A into blocks H and K of size $2w$. *Each of those blocks has rank exactly* w. This comes from a theorem of Asplund about ranks of submatrices of A, when the inverse matrix has bandwidth w. The display shows a matrix for which A and A^{-1} have bandwidth at most $w = 2$.

$$
\begin{array}{cc|cccc|cccc}
 & & \mathrm{X} & \mathrm{x} & \mathrm{x} & & & & & \\
 & & \mathrm{x} & \mathrm{X} & \mathrm{x} & \mathrm{x} & & & & \\
H_1 & & \mathrm{x} & \mathrm{x} & \mathrm{X} & \mathrm{x} & \mathrm{x} & & & K_1 \\
 & & & \mathrm{x} & \mathrm{x} & \mathrm{X} & \mathrm{x} & \mathrm{x} & & \\
\hline
 & & & & \mathrm{x} & \mathrm{x} & \mathrm{X} & \mathrm{x} & \mathrm{x} & \\
 & & & & & \mathrm{x} & \mathrm{x} & \mathrm{X} & \mathrm{x} & \mathrm{x} \\
H_2 & & & & & & \mathrm{x} & \mathrm{x} & \mathrm{X} & \mathrm{x} \\
 & & & & & & & \mathrm{x} & \mathrm{x} & \mathrm{X}
\end{array}
$$

Asplund's theorem applies to submatrices like the K's that are above subdiagonal w, and like the H's that are below superdiagonal w. "Those submatrices have rank $\leq w$[2, 11]." In our case the ranks are exactly w because each set of $2w$ rows of A must have full rank $2w$ (since A is invertible).

The two columns of H_1 are orthonormal. Choose a 4 by 4 orthogonal matrix Q_1, such that we get two columns correct in the identity matrix :

$$
Q_1 H_1 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \\ 0 & 0 \end{bmatrix} = R.
$$

Since the four rows of $Q_1[H_1 \quad K_1] = [R \quad Q_1 K_1]$ are orthonormal, the first two rows of $Q_1 K_1$ must be zero. Then there is a 4 by 4 orthogonal matrix ${C_1}^{\mathrm{T}}$ acting on the *columns*

of $Q_1 K_1$ that produces R^{T} in the next two rows:

$$Q_1 K_1 {C_1}^{\mathrm{T}} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 0^{\mathrm{T}} \\ R^{\mathrm{T}} \end{bmatrix} \quad \begin{matrix} \text{in rows} & 1-4 \\ \text{columns} & 3-6 \end{matrix}$$

At this stage the first $2w = 4$ rows are set. $Q_1 A\, {C_1}^{\mathrm{T}}$ agrees with the first four rows of I. If $B_1 = Q_1^{-1}$ is placed into the first block of B, and C_1 in rows/columns 5 to 8 of C, then four rows of $A = BC$ are now correct.

Moving to the next four rows, our goal is to change those into four more rows of the identity matrix. This will put the 8 by 10 submatrix of H's and K's in its final form:

$$\begin{bmatrix} Q_1 & \\ & Q_2 \end{bmatrix} \begin{bmatrix} H_1 & K_1 & & \\ & & H_2 & K_2 \end{bmatrix} \begin{bmatrix} I_2 & & & \\ & {C_1}^{\mathrm{T}} & & \\ & & & {C_2}^{\mathrm{T}} \end{bmatrix} = \left[\begin{array}{cc|cc} R & & 0^{\mathrm{T}} & \\ & & R^{\mathrm{T}} & \\ \hline & 0 & R & 0^{T} \\ & & & R^{\mathrm{T}} \end{array}\right]$$

Columns 3 and 4 end with the zero matrix as indicated, because those columns are orthonormal and they already have 1's from R^{T}.

Rows 5 to 8 are now in exactly the same situation that we originally met for rows 1 to 4. There is an orthogonal matrix Q_2 that produces R in columns 5 and 6, as shown. The rest of rows 5 and 6 must contain 0^{T}, as shown. Then an orthogonal ${C_2}^{\mathrm{T}}$ produces R^{T} in rows 7 and 8. We now have eight rows of $Q A\, C^{\mathrm{T}} = I$. The construction continues to rows 9 to 12, and onward:

$$QAC^{\mathrm{T}} = \begin{bmatrix} Q_1 & & & \\ & Q_2 & & \\ & & \ddots & \end{bmatrix} A \begin{bmatrix} I_2 & & & \\ & {C_1}^{\mathrm{T}} & & \\ & & {C_2}^{\mathrm{T}} & \\ & & & \ddots \end{bmatrix} = I.$$

Q and C are block diagonal with orthogonal 4 by 4 blocks, except for the 2 by 2 block I_2. We have shown that every banded orthogonal matrix A with $w = 2$ can be factored into $A = Q^{\mathrm{T}} C = BC$. The reasoning is the same for any w.

Theorem 2 Every banded orthogonal matrix A, finite or singly infinite, has orthogonal block diagonal factors $A = Q^{-1} C = BC$. The blocks B_i and C_i have size $2w$ except that $C_0 = I$ has size w (to produce the offset).

This completes the main step in the factorization—we have B and C with orthogonal blocks B_i and C_i. The final step is to reach 2 by 2 blocks. A straightforward construction succeeds for each B_i and C_i. At the end we factor all B_i at once and all C_i at once.

Lemma Any orthogonal matrix Q of size $2w$ can be reduced to the identity matrix by a sequence of multiplications, $G_M \ldots G_1\, Q = I$. Each factor G differs from I only in a 2 by

2 orthogonal block (a plane rotation):

$$G = \begin{bmatrix} I_m & & & \\ & c & -s & \\ & s & c & \\ & & & I_p \end{bmatrix}. \tag{2}$$

Proof Start at the bottom of column 1. Choose $c = \cos\theta$ and $s = \sin\theta$ to produce zero from $s\,Q_{n-1,1} + c\,Q_{n,1}$ in that corner of $G_1 Q$. (Take $s = 1$ and $c = 0$ in case $Q_{n-1,1} = 0$.) Move up column 1, producing a new zero with each factor G. At the top of column 1, choose signs to produce 1 as the diagonal entry in this unit vector.

Continue from the bottom of column 2. The new factors G that produce zeros below the diagonal will not affect the zeros in column 1. The $(1,2)$ entry in column 2 is already zero because the columns remain orthogonal at every step. At the end, we have the columns of I from $M = O\left(w^2\right)$ orthogonal matrices G_i.

In the non-orthogonal case, elimination in this order (climbing up each column) yields $N = O\left(w^2\right)$ factors F in Theorem 1. Each zero is produced by a "Gauss matrix" that has a single nonzero next to its diagonal part I. The $O\left(w^3\right)$ estimate in [10] was over-generous.

This lemma applies to each orthogonal block B_i (of size $2w$) in the matrix B. The key point is that we can *simultaneously* reduce all those blocks to I. The matrices $G_1, \ldots, G_M$ for all the different blocks of B go along the diagonals of $F_1, \ldots, F_M$. Then $F_M \ldots F_1 B = I$.

Similarly the lemma applies to the blocks C_i of C. Then we have $A = BC = \left(F_1^{-1} \ldots F_M^{-1}\right)\left(F_{M+1}^{-1} \ldots F_{2M}^{-1}\right)$. This completes the orthogonal case.

3 Wavelet Matrices

The matrices that lead to wavelets are banded with banded inverses. Furthermore they are block Toeplitz (away from the first and last rows) with 2 by 2 blocks. In this case the factors F will also be block Toeplitz (away from those boundary rows). This means that each F_i will have a 2 by 2 block repeating down the diagonal.

In the analysis of wavelets it is often convenient to work with *doubly infinite* matrices A_∞ (purely block Toeplitz, with no boundary rows). Suppose A_∞ and its inverse have bandwidth $w = N$, coming from N blocks centered on each pair of rows. We showed in [10] how to find N block diagonal factors in $A_\infty = F_1 \ldots F_N$: First find $N-1$ factors G_i, each with two singular 2 by 2 blocks per row ($w = 2$). Then split each G_i into block diagonal factors $F_{i1} F_{i2}$. Those are offset as in equation (1) above, and we choose the factors so that F_{i2} is *not* offset compared to $F_{i+1,1}$. Then $F_{i2} F_{i+1,1}$ is a block diagonal F, and the banded matrix A_∞ has N factors:

$$\begin{aligned} A_\infty &= (F_{11}F_{12})(F_{21}F_{22})\cdots(F_{N-1,1}F_{N-1,2}) \\ &= F_{11}\ (F_{12}F_{21})(F_{22}F_{31})\cdots(F_{N-2,2}F_{N-1,1})\ F_{N-1,2}. \end{aligned}$$

The 4-coefficient Daubechies wavelet matrix has $N = 2$. It is the single matrix G_1. It was factored in [5] into block diagonal F_{11} times F_{12}:

$$\begin{bmatrix} \bullet & & & \\ & 1+\sqrt{3} & -1+\sqrt{3} & \\ & 1-\sqrt{3} & 1+\sqrt{3} & \\ & & & \bullet \end{bmatrix} \begin{bmatrix} \sqrt{3} & -1 & & \\ 1 & \sqrt{3} & & \\ & & \sqrt{3} & -1 \\ & & 1 & \sqrt{3} \end{bmatrix} = \begin{bmatrix} \bullet & \bullet & & \\ 1+\sqrt{3} & 3+\sqrt{3} & 3-\sqrt{3} & 1-\sqrt{3} \\ 1-\sqrt{3} & -3+\sqrt{3} & 3+\sqrt{3} & -1-\sqrt{3} \\ & & \bullet & \bullet \end{bmatrix}$$

Again, column 2 times row 2 gives one singular block and column 3 times row 3 gives the other. Dividing by the lengths $\sqrt{8}$ and $\sqrt{4}$ of their rows, those factors become orthogonal matrices. Their product shows the $4+4$ Daubechies coefficients. The $6+6$ coefficient matrix for the next orthogonal Daubechies wavelet was factored by Raghavan [8].

These factors are plane rotations through $\pi/12$ and $\pi/6$. TeKolste also observed [12] that the Daubechies matrices are offset products of plane rotations. He recognized that the N rotation angles add to $\pi/4$. This connects to the fact that the coefficients in the second row of A_∞ add to zero (a highpass filter). There should be additional conditions on the angles to determine the multiplicity of this zero in the highpass frequency response (the polynomial whose coefficients come from that second row of A_∞).

A task for the future is to construct useful matrices A_∞ and A_n starting from well-chosen factors. Wavelets need not be orthogonal (the most popular choices are not). And they need not be restricted to block Toeplitz matrices. It seems feasible to construct time-varying wavelets by starting with factors F in which the blocks are not constantly repeated down the diagonal. The matrix still has a banded inverse.

4 Banded Permutation Matrices

The bandwidth of a permutation is the maximum distance $|i - p(i)|$ that any entry is moved. Thus $w = 1$ for a "parallel exchange of neighbors" like 2, 1, 4, 3. The matrix F for this permutation is block diagonal with a 2 by 2 block for each exchange.

It is straightforward to reach the identity by a sequence of these F's. At each step we move from left to right, exchanging pairs of neighbors that are in the wrong order:

$$4\,5\,6\,1\,2\,3 \to 4\,5\,\mathbf{1\,6}\,2\,3 \to 4\,\mathbf{1\,5\,2\,6}\,3 \to \mathbf{1\,4\,2\,5\,3\,6} \to 1\,\mathbf{2\,4\,3\,5}\,6 \to 1\,2\,\mathbf{3\,4}\,5\,6.$$

The original permutation has $w = 3$ (in fact all entries have $|i - p(i)| = 3$). Five steps were required to reach the identity. So the banded permutation matrix is the product of $N = 5 = 2w - 1$ matrices F.

We conjectured in [10] that $N \le 2w - 1$ in all cases. A beautiful proof is given by Greta Panova [6], using the "wiring diagram" to decide the sequence of exchanges in advance. A second proof [1] by Albert, Li, Strang, and Yu confirms that the algorithm illustrated above also has $N \le 2w - 1$. A third proof is proposed by Samson and Ezerman [9].

5 Finite Rank Perturbations

We now consider the larger set of invertible matrices $M = A + Z$, when A and A^{-1} have bandwidth $\leq w$ and Z has rank $\leq r$. The inverse matrix M^{-1} has the same properties, from the Woodbury-Morrison formula (which has many discoverers). Write Z as $-UV$ where U has independent columns and V has independent rows. Then the formula yields

$$M^{-1} = (A - UV)^{-1} = A^{-1} + Y = A^{-1} + A^{-1}U(I - VA^{-1}U)^{-1}VA^{-1}. \tag{3}$$

The rank of Y is not greater than the ranks (both $\leq r$) of U and V.

The product M_1M_2 will be the sum of A_1A_2 with bandwidth $\leq 2w$, and $A_1Z_2 + A_2Z_1 + Z_1Z_2$ with rank $\leq 3r$. So we have a group B of invertible matrices $M = A + Z$, in which A and A^{-1} belong to our original group and rank(Z) is finite. As before, B is all of GL(n) for finite size n. It is apparently a new group for $n = \infty$.

We want to describe a set of generators for this group. They will include the same block diagonal factors F with bandwidth 1, together with new factors of the form $I + (\text{rank } 1)$. We show how to express $A^{-1}M = I + A^{-1}Z$ using at most r of these new factors.

Theorem 3 If $L = A^{-1}Z$ has rank r, there are vectors $u_1, v_1, \ldots, u_r, v_r$ so that

$$I + L = \left(I + u_r v_r^{\mathrm{T}}\right) \ldots \left(I + u_1 v_1^{\mathrm{T}}\right). \tag{4}$$

Proof We will choose vectors such that $v_i^{\mathrm{T}} u_j = 0$ for $i > j$. Then if the columns of V and U are $v_1, \ldots, v_r$ and $u_1, \ldots, u_r$, the product $V^{\mathrm{T}}U$ is upper triangular.

Under this condition, the expression (4) reduces to $I + u_r v_r^{\mathrm{T}} + \cdots + u_1 v_1^{\mathrm{T}} = I + UV^{\mathrm{T}}$. So the goal is to achieve $UV^{\mathrm{T}} = L$ with $V^{\mathrm{T}}U$ upper triangular.

The usual elimination steps reduce L to a matrix W with only r nonzero rows. Thus $EL = W$ or $L = E^{-1}W = \left(E^{-1}\right)_r W_r$, where we keep only those first r rows of W and the first r columns of E^{-1}. In the opposite order, $W_r \left(E^{-1}\right)_r$ may not be triangular, but every matrix is similar to an upper triangular matrix :

$$\text{For some } r \text{ by } r \text{ matrix } S, \quad S\, W_r \left(E^{-1}\right)_r S^{-1} \text{ is upper triangular.}$$

Now take $V^{\mathrm{T}} = SW_r$ and $U = \left(E^{-1}\right)_r S^{-1}$. Then $V^{\mathrm{T}}U$ is triangular and $UV^{\mathrm{T}} = L$.

6 Cyclically Banded Matrices

A cyclically banded n by n matrix A (with bandwidth w) has three nonzero parts. There can be w nonzero diagonals in the upper right corner and lower left corner, in addition to the $2w + 1$ diagonals in the main band. Call the new parts in the corners b and c :

$$A = \begin{bmatrix} & & & b \\ & & 0 & \\ & a & & \\ 0 & & & \\ c & & & \end{bmatrix}$$

$A_{ij} = 0$ if $|i - j| > w$
and also $n - |i - j| > w$.

Thus "cyclic bandwidth $w = 1$" now allows $A_{1n} \neq 0$ and $A_{n1} \neq 0$.

There is a corresponding infinite periodic matrix that has bandwidth w in the normal way. "Periodic" means that $A_{i+n,j+n} = A_{ij}$ for all integers i and j. When we identify 0 with n and every i with $n+i$, the corner parts b and c move to fill the gaps at the start and end of the main band. Here is an example with $b = 4$ and $c = 3$ and all 1's in the tridiagonal a :

$$\text{Periodic } A_\infty = \left[\begin{array}{cc|ccc|cc} 1 & 1 & 3 & & & & \\ \hline & 4 & 1 & 1 & & & \\ & & 1 & 1 & 1 & & \\ & & & 1 & 1 & 3 & \\ \hline & & & & 4 & 1 & 1 \end{array}\right] \text{ with cyclic } w = 1$$

Notice that the periodic matrix A_∞ is block Toeplitz. The banded blocks a repeat down the main diagonal, and blocks b and c (all 3 by 3) go down the adjacent diagonals. Multiplication AB of cyclically banded matrices corresponds to $A_\infty B_\infty$ for periodic matrices. Inversion A^{-1} corresponds to $(A_\infty)^{-1}$. The cyclic bandwidth for A is the normal bandwidth for A_∞.

The reader will make the same conjecture as the author :

Conjecture If A and A^{-1} have cyclic bandwidth $\leqslant w$, they are products of $N = O(w^2)$ factors for which F and F^{-1} have cyclic bandwidth $\leqslant 1$. The number N does not depend on n.

For cyclically banded permutation matrices this conjecture is proved by Greta Panova. Her wiring diagrams move to a cylinder, for periodicity. The number of factors still satisfies $N \leqslant 2w - 1$ (for permutations). The parallel transpositions in a factor F can include an exchange of 1 with n. Two new factors are allowed that also have cyclic bandwidth 1 — these are the cyclic shifts. They permute $1,\dots,n$ to $n,1,\dots,n-1$ or to $2,\dots,n,1$. The inverse of one cyclic shift is the other.

Allow us to close the paper with preliminary comments on a proof of the conjecture. The non-cyclic proof began with Asplund's condition on a matrix A whose inverse has bandwidth $\leqslant w$. *All submatrices of A above subdiagonal w and below superdiagonal w have* rank $\leqslant w$. We expect this condition to extend to periodic infinite matrices A_∞, but one change is important : The cyclic shift S (and its periodic form S_∞) is inverted by its transpose. The finite matrix obeys Asplund's rule :

$$S = \begin{bmatrix} 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \quad \begin{array}{l} \text{has rank 1 above subdiagonal 1 :} \\ \quad \text{then } S^{-1} \text{ has upper bandwidth 1} \\ \text{has rank 3 below superdiagonal 3 :} \\ \quad \text{then } S^{-1} \text{ has lower bandwidth 3.} \end{array}$$

But S_∞ needs a revised rule. It is lower triangular (rank zero above the diagonal) but its inverse (the transpose shift) is upper triangular.

The second obstacle, perhaps greater, lies in the factorization of A_∞ into block diagonal periodic matrices $B_\infty C_\infty$. As in Section 2 above, their diagonal blocks may have size $2w$

(independent of n). The further factorization into periodic matrices F with 2 by 2 blocks would be straightforward, but how to reach B_∞ and C_∞? The elimination steps we used earlier are now looking for a place to start.

In our original proof of the (non-cyclic) factorization into $A = F_1 \dots F_N$, a decomposition of "Bruhat type" was the first step — in order to reach triangular factors:

$$A = (\text{triangular})\ (\text{permutation})\ \ (\text{triangular}) = LPU. \tag{5}$$

With P factored into F's, this leaves the triangular L and U to be factored. L, P, U are all banded with banded inverses; their bandwidths come from A and A^{-1}. Note that Bruhat places P *between* the triangular factors, where numerical linear algebra places it first. (The standard Bruhat triangular factors are both upper triangular.)

In the periodic case, P_∞ will be an *affine permutation*. The banded periodic matrix A_∞ is naturally associated with a rational matrix function $a(z)$. The blocks in A_∞ are the coefficients in $a(z)$. For our block tridiagonal matrix, this function will be $bz^{-1} + a + cz$.

Then the triangular $L_\infty P_\infty U_\infty$ factorization of A_∞ is associated with a factorization of this n by n matrix function, as in

$$a(z) = l(z)\, p(z)\, u(z) \ \text{ with } \ p(z) = \operatorname{diag}(z^{k_1}, \dots, z^{k_n}). \tag{6}$$

The integers $k_1, \dots, k_n$ are the (left) partial indices of $a(z)$. They determine the periodic permutation matrix P_∞; the 1 in row i lies in column $i + k_i n$. The factors $u(z)$ and $l(z)$ are analytic for $|z| < 1$ and $|z| > 1$. Their matrix coefficients appear as the blocks in the triangular factors U_∞ and L_∞.

Matrix factorization theory began with the fundamental paper of Plemelj [7]. A remarkable bibliography and a particularly clear exposition and proof of (6) are in the survey paper [4] by Gohberg, Kaashoek, and Spitkovsky. They include a small example that we convert from $a(z) = l(z)\, p(z)\, u(z)$ to $A_\infty = L_\infty\, P_\infty\, U_\infty$ with $U_\infty = I$. Rows 1 to 4 show the pattern in these infinite block Toeplitz matrices:

$$a(z) = \begin{bmatrix} z & 0 \\ 1 & z^{-1} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ z^{-1} & 1 \end{bmatrix} \begin{bmatrix} z & \\ & z^{-1} \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = l\ p\ u$$

$$A_\infty = \begin{bmatrix} 0 & 0 & 0 & 0 & 1 & 0 & & \\ 0 & 1 & 1 & 0 & 0 & 0 & & \\ & & 0 & 0 & 0 & 0 & 1 & 0 \\ & & 0 & 1 & 1 & 0 & 0 & 0 \end{bmatrix} = L_\infty\, P_\infty =$$

$$= \begin{bmatrix} 0 & 0 & 1 & 0 & 0 & 0 & & \\ 1 & 0 & 0 & 1 & 0 & 0 & & \\ & & 0 & 0 & 1 & 0 & 0 & 0 \\ & & 1 & 0 & 0 & 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} 0 & 0 & 0 & 0 & 1 & 0 & & \\ 0 & 1 & 0 & 0 & 0 & 0 & & \\ & & 0 & 0 & 0 & 0 & 1 & 0 \\ & & 0 & 1 & 0 & 0 & 0 & 0 \end{bmatrix}.$$

Important: The inverse of A_∞ is also banded in this example because the determinant of $a(z)$ is a monomial. The inverse of this L_∞ reverses signs below the diagonal, and the inverse of a permutation P_∞ is always the transpose. With indices $k_1 = 1$ and $k_2 = -1$ and $n = 2$, rows 1 and 2 of P_∞ have ones in columns $1 + 2 = 3$ and $2 - 2 = 0$.

Two key questions stand out :

1. Block diagonal factorization of the periodic matrices U_∞ and L_∞.
2. Nonperiodic infinite matrices A, banded with banded inverses: Do they factor into $F_1 \dots F_N$ (block diagonal and shifts)?

References

1. C. Albert, C.-K. Li, G. Strang and G. Yu, Permutations as products of parallel transpositions, submitted to SIAM J. Discrete Math (2010).

2. E. Asplund, Inverses of matrices $\{a_{ij}\}$ which satisfy $a_{ij} = 0$ for $j > i + p$, Math. Scand. 7 (1959) 57-60.

3. I. Daubechies, Orthonormal bases of compactly supported wavelets, Comm. Pure Appl. Math. 41 (1988) 909-996.

4. I. Gohberg, M. Kaashoek, and I. Spitkovsky. An overview of matrix factorization theory and operator applications, Operator Th. Adv. Appl., Birkhäuser (2003) 1-102.

5. V. Olshevsky, P. Zhlobich, and G. Strang, Green's matrices, Lin. Alg. Appl. 432 (2010) 218-241.

6. G. Panova, Factorization of banded permutations, arXiv.org/abs/1007.1760 (2010).

7. J. Plemelj, Riemannsche Funktionenscharen mit gegebener Monodromiegruppe, Monat. Math. Phys. 19 (1908) 211-245.

8. V.S.G. Raghavan, Banded Matrices with Banded Inverses, M.Sc. Thesis, MIT (2010).

9. M.D. Samson and M.F. Ezerman, Factoring permutation matrices into a product of tridiagonal matrices, arXiv.org/abs/1007.3467 (2010).

10. G. Strang, Fast transforms : Banded matrices with banded inverses, Proc. Natl. Acad. Sci. 107 (2010) 12413-12416.

11. G. Strang and T. Nguyen, The interplay of ranks of submatrices, SIAM Review 46 (2004) 637-646.

12. K. TeKolste, private communication (2010).

Department of Mathematics, MIT
Cambridge MA 02139
gs@math.mit.edu

AMS Subject Classification 15A23
Keywords: Banded matrix, banded inverse, group generators, factorization, permutations.