Banded Matrices with Banded Inverses and $A = LPU$

Gilbert Strang

Abstract. If A is a banded matrix with a banded inverse, then $A = BC = F_1 \ldots F_N$ is a product of block-diagonal matrices. We review this factorization, in which the F_i are tridiagonal and N is independent of the matrix size. For a permutation with bandwidth w, each F_i exchanges disjoint pairs of neighbors and $N < 2w$.

This paper begins the extension to infinite matrices. For doubly infinite permutations, the factors F now include the left and right shift. For banded infinite matrices, we discuss the triangular factorization $A = LPU$ (completed in a later paper on *The Algebra of Elimination*). Four directions for elimination give four factorizations LPU and UPL and $U_1 \pi U_2$ (Bruhat) and $L_1 \pi L_2$ with different L, U, P and π.

1. Introduction

This paper is about two factorizations of invertible matrices. One is the familiar $A = LU$, is the lower times upper triangular, which is a compact description of the elimination algorithm. A permutation matrix P may be needed to exchange rows. The question is whether P comes before L or after! Numerical analysts put P first, to order the rows so that all upper left principal submatrices become nonsingular (which allows LU). Algebraists write $A = LPU$, and in this form P is *unique*.

Most mathematicians think only of one or the other, and a small purpose of this paper is to present both. We also connect elimination starting at the $(n, 1)$ entry to the Bruhat factorization $A = U_1 \pi U_2$. In this form the most likely permutation π (between two upper triangular factors) is the reverse identity. In fact the four natural starting points $(1, 1)$, $(n, 1)$, (n, n), $(1, n)$ lead to four factorizations with different L, U, P, π:

$$A = LPU\,, \quad A = U\pi U\,, \quad A = UPL\,, \quad A = L\pi L.$$

The P's and π's are unique when A is invertible, and in each case elimination can choose row or column operations to produce these factorizations.

Our larger purpose is to discuss banded matrices that have banded inverses: $A_{ij} = 0$ and also $(A^{-1})_{ij} = 0$ for $|i - j| > w$. (The unique permutation P will share the same bandwidth w.) The purpose of our factorization $A = F_1 \ldots F_N$ is to make this bandedness evident: *The matrices F_i are block diagonal.* Then

AMS Subject Classification. PROC10959.

Key words and phrases. Banded matrix, Banded inverse, Bruhat permutation, Factorization, Group generators, Shifting index.

their inverses singly are also block diagonal, and the products $A = F_1 \dots F_N$ and $A^{-1} = F_N^{-1} \dots F_1^{-1}$ are both banded.

We established this factorization in *The Algebra of Elimination* [16] using Asplund's test for a banded inverse: All submatrices of A above subdiagonal w or below superdiagonal w have rank $\leq w$. The key point of the theorem is that *the number $N \leq Cw^2$ of factors F_i is controlled by w and not by the matrix size n.* This opens the possibility of infinite matrices (singly or doubly infinite).

We will not achieve here the complete factorizations of infinite matrices, but we do describe progress (as well as difficulties) for $A = LPU$. In one important case—*banded permutations* of $\mathbf{Z}$, represented by doubly infinite matrices—we introduce an idea that may be fruitful. The factors $F_1, \dots, F_N$ for these matrices include disjoint transpositions T of neighbors and also bi-infinite shifts S and S^{T}. All have $w = 1$:

$$T = \begin{bmatrix} \bullet & & \\ & \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} & \\ & & \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \end{bmatrix} \quad \text{and} \quad S = \begin{bmatrix} \bullet & \bullet & & & \\ 0 & 0 & 1 & & \\ & 0 & 0 & 1 & \\ & & & \bullet & \bullet \end{bmatrix}$$

Greta Panova used a neat variation [11] of the "wiring diagram" for P, to show that the number of factors is $N \leq 2w - 1$. This conjecture from [16] was for finite matrices. The extension to banded permutations of $\mathbf{Z}$ allows also s shift factors. We call $s(P)$ the *shifting index* of P (negative for $S^{\mathrm{T}} = S^{-1}$). (Important and recently discovered references are [9, 14, 15], please see below.) This shifting index has the property that

$$s(P_1 P_2) = s(P_1) + s(P_2). \tag{1.1}$$

It should also have a useful meaning for A, when $A = LPU$. For the periodic block Toeplitz case with block size B, $s(P)$ is the sum of k_i in the classical factorization of a matrix polynomial into $a(z) = \ell(z)p(z)u(z)$ with $p(z) = \operatorname{diag}(z^{k_1}, \dots, z^{k_B})$.

The original paper [16] outlined algorithms to produce the block diagonal factors F_i in particular cases with banded inverses:

1. Wavelet matrices are block Toeplitz (periodic) and doubly infinite (i, j in $\mathbf{Z}$). A typical pair of rows contains 2 by 2 blocks M_0 to M_{N-1}. The action of this A is governed by the matrix polynomial $M(z) = \sum M_j z^j$. The inverse is banded exactly when $\det M(z)$ has only one term cz^{N-1}. In the nondegenerate case, the number N counts the factors F_i and also equals the bandwidth w (after centering):

$$A = \begin{bmatrix} \cdots & & \\ M_0 \ \dots \ M_{N-1} & & \\ & M_0 \ \dots \ M_{N-1} & \\ & \cdots & \end{bmatrix} \quad \text{has factors} \quad F_i = \begin{bmatrix} \bullet & & & \\ & B_i & & \\ & & B_i & \\ & & & \bullet \end{bmatrix}.$$

Important point: The 2 by 2 blocks B_{i+1} in F_{i+1} are shifted by one row and column relative to B_i in F_i. Otherwise the product of F's would only be block diagonal.

2. CMV matrices. The matrices studied in [3, 10] have two blocks on each pair of rows of A. Those 2 by 2 blocks are *singular*, created by multiplying a typical

$F_1F_2 = A$ (notice again the shift in position of the blocks):

$$\begin{bmatrix} 1 & & & & \\ & 2 & 3 & & \\ & 4 & 5 & & \\ & & & 6 & 7 \\ & & & 8 & 9 \end{bmatrix} \begin{bmatrix} a & b & & & \\ c & d & & & \\ & & e & f & \\ & & g & h & \\ & & & & i \end{bmatrix} = \left[\begin{array}{cc|ccc} a & b & & & \\ 2c & 2d & 3e & 3f & \\ 4c & 4d & 5e & 5f & \\ \cline{3-5} & & 6g & 6h & 7i \\ & & 8g & 8h & 9i \end{array}\right] \tag{1.2}$$

A has bandwidth $w = 2$. Also $A^{-1} = F_2^{-1}F_1^{-1}$ has $w = 2$. Note how column 2 of F_1 times row 2 of F_2 produces the singular block with $2c$, $2d$, $4c$, $4d$ (and two other blocks are also singular). Necessarily A passes Asplund's test: Those three singular blocks assure that every admissible submatrix has rank ≤ 2. One submatrix is indicated, above the second subdiagonal of A.

In applications to orthogonal polynomials on the circle $|z| = 1$, CMV matrices are no longer block Toeplitz. The 2 by 2 blocks are all different as shown. We may think of them as "*time-varying*" wavelet matrices. Extending this analogy, we allow them to have $N > 2$ blocks centered along each pair of rows (then $w = N$). The factorization of A is recursive. Always F_{i+1} has its diagonal blocks shifted with respect to F_i, as in the multiplication above.

3. Orthogonal matrices. The original CMV matrices and the Daubechies wavelet matrices were banded and also *orthogonal*: $A^{\mathrm{T}}A = I$. It is natural to look for orthogonal factors F_i with $w = 1$. This can be achieved for all banded orthogonal matrices. Our example [10] is a CMV matrix and Daubechies matrix of particular interest: F_1 and F_2 are Toeplitz (periodic) and their blocks are rotations:

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

$$= \begin{bmatrix} \bullet & \bullet & & \\ 1+\sqrt{3} & 3+\sqrt{3} & 3-\sqrt{3} & 1-\sqrt{3} \\ 1-\sqrt{3} & -3+\sqrt{3} & 3+\sqrt{3} & -1-\sqrt{3} \\ & & \bullet & \bullet \end{bmatrix} \tag{1.3}$$

To normalize those columns to unit length, divide by $2\sqrt{2}$, 2, and $4\sqrt{2}$. The rotation angles are $\pi/12$ and $\pi/6$, adding to $\pi/4$.

4. Permutation matrices. The ordering $(3, 4, 1, 2)$ is associated with a 4 by 4 permutation matrix. The bandwidths of P and $P^{\mathrm{T}} = P^{-1}$ are $w = 2$. This is the greatest distance $w = \max|\,i - p(i)|$ that any entry must move:

$$(3, 4, 1, 2) \text{ corresponds to } P = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}.$$

Three "greedy steps" [1] will exchange disjoint neighbors to reach the order $(1, 2, 3, 4)$:

$$(3, 4, 1, 2) \rightarrow (3, \mathbf{1}, \mathbf{4}, 2) \rightarrow (\mathbf{1}, \mathbf{3}, \mathbf{2}, \mathbf{4}) \rightarrow (1, \mathbf{2}, \mathbf{3}, 4).$$

The product of the corresponding block diagonal matrices $F_1F_2F_3$ is P:

$$
(1.4) \qquad P = \begin{bmatrix} 1 & & & \\ & 0 & 1 & \\ & 1 & 0 & \\ & & & 1 \end{bmatrix} \begin{bmatrix} 0 & 1 & & \\ 1 & 0 & & \\ & & 0 & 1 \\ & & 1 & 0 \end{bmatrix} \begin{bmatrix} 1 & & & \\ & 0 & 1 & \\ & 1 & 0 & \\ & & & 1 \end{bmatrix}.
$$

In this example N reaches its maximum value $2w - 1 = 3$ for permutations of bandwidth w.

2. The Factorization $A = F_1 \ldots F_N$

A and A^{-1} are banded n by n matrices. We will display the steps of their factorization into block diagonal matrices. The factors are reached in two steps:

(1) Factor A into BC with diagonal blocks of sizes $w, 2w, 2w, \ldots$ for B and $2w, 2w, \ldots$ for C. As in equations (2), (3), (4), this shift between the two sets of blocks means that $A = BC$ need not be block diagonal.

(2) Break B and C separately into factors F with blocks of size 2 (or 1) along the diagonal. This is achieved in [17] by ordinary elimination, and is not repeated here. In principle we may need $O(w^2)$ steps, moving upward in successive columns $1, \ldots, 2w$ of each block in B and C.

We do want to explain the key idea behind Step 1, to reach $A = BC$.

Suppose A has bandwidth $w = 2$. If A^{-1} also has $w = 2$, Asplund's theorem [2, 18] imposes a rank condition on certain submatrices of A, above subdiagonal w or below superdiagonal w. All these submatrices must have rank $\leq w$.

To apply this condition we take the rows and the columns of A in groups of size $2w = 4$. The main diagonal is indicated by capital letters X, and Asplund's condition rank ≤ 2 applies to all H_i and K_i. All those ranks are **exactly 2** because each set of four rows has rank 4.

$$
\begin{array}{c|cc|cccc|cccc|c}
 & \mathrm{X} & \mathrm{x} & \mathrm{x} & & & & & & & & \\
H_1 & \mathrm{x} & \mathrm{X} & \mathrm{x} & \mathrm{x} & & & & & & & K_1 \\
 & \mathrm{x} & \mathrm{x} & \mathrm{X} & \mathrm{x} & \mathrm{x} & & & & & & \\
 & & \mathrm{x} & \mathrm{x} & \mathrm{X} & \mathrm{x} & \mathrm{x} & & & & & \\
\hline
 & & & \mathrm{x} & \mathrm{x} & \mathrm{X} & \mathrm{x} & \mathrm{x} & & & & \\
H_2 & & & & \mathrm{x} & \mathrm{x} & \mathrm{X} & \mathrm{x} & \mathrm{x} & & & K_2 \\
 & & & & & \mathrm{x} & \mathrm{x} & \mathrm{X} & \mathrm{x} & \mathrm{x} & & \\
 & & & & & & \mathrm{x} & \mathrm{x} & \mathrm{X} & \mathrm{x} & \mathrm{x} &
\end{array}
$$

Our plan is to diagonalize these submatrices $H_1, K_1, H_2, K_2, \ldots$ by row operations on the H's and column operations on the K's. The row steps can be done in parallel on $H_1, H_2, \ldots$ to give the blocks in B. The column steps give the blocks in C, and we fold into C the diagonal matrix reached at the end of the elimination.

Elimination on H_1: With rank 2, row operations will replace every x by zero. Rows 3 and 4 of the new K_1 must now be independent (since they have only zeros in H_1).

Elimination on K_1: With rank 2, upward row operations and then column operations will replace every x by zero. Columns 5 and 6 in the new H_2 must now be independent (since those columns start with zeros in the current K_1).

Elimination on H_2: Leftward column operations and then row operations will replace every x by zero. Eventually the whole A is reduced to a diagonal matrix.

3. Four Triangular Factorizations

The basic factorization is $A = LU$. The first factor has 1's on the diagonal, the second factor has nonzeros $d_1, \ldots, d_n$. Multiplying the k by k upper left submatrices gives $A_k = L_k U_k$, so a necessary condition for $A = LU$ is that every A_k is nonsingular. Executing the steps of elimination shows that this condition is also sufficient. After $k-1$ columns are zero below the main diagonal, the (k, k) entry will be $\det A_k / \det A_{k-1}$. Below this nonzero pivot d_k, row operations will achieve zeros in column k. Then continue to $k+1$. Inverting all those row operations by L leaves $A = LU$.

Note that column operations will give exactly the same result. At step k, subtract multiples of *column* k from later columns to clear out *row* k above the diagonal. After n steps we have a lower triangular L_c with the same pivots d_k on its diagonal. Recover A by inverting those column operations (add instead of subtract). This uses upper triangular matrices multiplying on the right, so $A = L_c U_c$. Moving the pivot matrix $D = \text{diag}\,(d_1, \ldots, d_n)$ from L_c to U_c must reproduce $A = L_c D^{-1} D U_c = LU$ as found by row operations, because those factors are unique.

If any submatrices A_k are singular, $A = LU$ is impossible. A permutation matrix P is needed. In numerical linear algebra (where additional row exchanges bring larger entries into the pivot positions) it is usual to imagine all exchanges done first. Then the reordered matrix PA factors into LU. In algebra (where the size of the pivot is not important) we keep the rows in place. Elimination is still executed by a lower triangular matrix. But the outcome may not be upper triangular, until we reorder the rows by factoring out P:

$$\begin{bmatrix} 0 & 2 & 5 \\ 1 & 0 & 4 \\ 0 & 0 & 3 \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 5 \\ 0 & 2 & 4 \\ 0 & 0 & 3 \end{bmatrix} = PU.$$

The elimination steps (which produced those zeros below the entries 1 and 2) are inverted by L. Then the original A is LPU.

To see that P is unique, consider any upper left submatrix a of A:

$$\begin{bmatrix} a & * \\ * & * \end{bmatrix} = \begin{bmatrix} \ell & 0 \\ * & * \end{bmatrix} \begin{bmatrix} p & * \\ * & * \end{bmatrix} \begin{bmatrix} u & * \\ 0 & * \end{bmatrix} \quad \text{gives } a = \ell p u. \tag{3.1}$$

If a has s rows and t columns, then ℓ is s by s and u is t by t—both with nonzero diagonals and both invertible! Therefore the s by t submatrix p has the same rank as a. Since the ranks of all upper left submatrices p are determined by A, the whole permutation P is uniquely determined in $A = LPU$ [5, 6, 8]. This simple step is all-important.

The 1's in P indicate pivots in A. This occurs in the i, j position when the rank of the i by j upper left submatrix a_{ij} *jumps above* the rank of $a_{i-1,\,j}$ and $a_{i,\,j-1}$. (By convention a_{i0} and a_{0j} have rank zero.) Again this criterion treats rows and columns equally. Elimination by row or by column operations leads to the same P.

A 2 by 2 example shows that the triangular L and U are no longer unique:

$$\begin{bmatrix} 0 & 1 \\ 1 & a \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ b & 1 \end{bmatrix} \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 1 & c \\ 0 & 1 \end{bmatrix} \text{ provided } a = b + c.$$

If elimination is by row operations, we will choose $b = a$ to clear out column 2 and reach $c = 0$. If elimination is by column operations, we will choose $c = a$ to clear out row 2 and reach $b = 0$. These particular choices of L and U are "reduced on the right" and "reduced on the left." The matrices PUP^{-1} and $P^{-1}LP$ are upper and lower triangular respectively (both are identity matrices in this example).

For each of these *reduced* factorizations—the normal choices when constructed by elimination—all three factors L,P,U are unique up to the diagonal pivot matrix D. If we include D as a fourth factor in $A = LPDU$, with diagonal 1's in L and U, then all reduced factors are unique.

Those paragraphs summarized the known algebra of elimination "with P in the middle." We want to add one trivial observation. It is prompted by the Bruhat factorization $A = U_1 \pi U_2$ with two upper triangular factors. The permutation π will now be the *reverse identity* matrix J for a generic matrix A. Notice that $U_1\pi$ is *not* lower triangular, and Bruhat in this generic case is not the same as $A = LU$. Our observation is that *four inequivalent factorizations of A come from four different starting points for elimination.*

Those starting points are the four corner entries of A. We indicate the shapes of the triangular factors in the four generic cases, when the eliminations proceed without meeting zeros in the natural pivot positions (if a zero does appear, the permutation will change from I or J):

(down and right) $A = LU \longleftarrow$ a_{11} a_{1n} $\longrightarrow A = L_1JL_2$ (down and left)

(up and right) $A = U_1JU_2 \longleftarrow$ a_{n1} a_{nn} $\longrightarrow A = UL$ (up and left)

Thus Bruhat comes from eliminating *starting at* a_{n1}. When column 1 is reduced to zero above this pivot, the next pivot position is $(n-1, 2)$. With no row exchanges, upward elimination will reach a "southeast" matrix. This becomes upper triangular when its rows are reordered by J:

$$A \longrightarrow \begin{bmatrix} & & & * \\ & & * & * \\ & * & * & * \\ * & * & * & * \end{bmatrix} = \begin{bmatrix} & & & 1 \\ & & 1 & \\ & 1 & & \\ 1 & & & \end{bmatrix} \begin{bmatrix} * & * & * & * \\ & * & * & * \\ & & * & * \\ & & & * \end{bmatrix} = JU_2.$$

The steps of upward elimination are inverted by an upper triangular matrix U_1 that brings back $A = U_1JU_2$.

If a pivot entry is zero, the reverse identity J changes to a different permutation π. These permutations come in a natural partial order (the Bruhat order) based on the number of transpositions of neighbors. $J = (n, \ldots, 1)$ comes first with the maximum number of transpositions.

4. Infinite Matrices

A is singly infinite if the indices i, j are natural numbers (i, j in $\mathbf{N}$), and doubly infinite if all integers are allowed (i, j in $\mathbf{Z}$). We mention three difficulties with the factorization of infinite matrices.

I. For singly infinite matrices, elimination starts with a_{11}. Even if there are no row exchanges, the factors L and U can be unbounded. The pivots can approach 0 and ∞. Consider a block diagonal matrix A with 2 by 2 blocks B_n :

$$B_n = \begin{bmatrix} \varepsilon_n & 1 \\ 1 & 0 \end{bmatrix} \qquad B_n^{-1} = \begin{bmatrix} 0 & 1 \\ 1 & -\varepsilon_n \end{bmatrix} \qquad \varepsilon_n \to 0.$$

B_n and B_n^{-1} stay bounded but the blocks in L and U will grow as $n \to \infty$:

$$B_n = L_n U_n = \begin{bmatrix} 1 & 0 \\ \varepsilon_n^{-1} & 1 \end{bmatrix} \begin{bmatrix} \varepsilon_n & 1 \\ 0 & -\varepsilon_n^{-1} \end{bmatrix}.$$

Thus L and U are unbounded.

II. For doubly infinite matrices, elimination has no natural starting point. Instead of a recursive algorithm, we need to describe the decisions at step k in terms of the original matrix A. Here is a reasonable formulation of that step:

For each k in $\mathbf{Z}$, remove all columns of A after column k to create a submatrix $A(k)$ ending at column k.

Define $I(k)$ as the set of all numbers i in $\mathbf{Z}$ such that row i of $A(k)$ is not a linear combination of previous rows of $A(k)$. The set $I(k)$ has these properties:

(1) $I(k)$ contains no numbers greater than $k + w$. By the bandedness of A, the rows beyond row $k + w$ are zero in $A(k)$.

(2) $I(k)$ contains every number $i \le k - w$. All the nonzeros in row i of A are also in row i of $A(k)$, by bandedness. Since A is invertible, that row i cannot be a combination of previous rows.

(3) $I(k)$ contains $I(k-1)$. If i is in $I(k-1)$, then row i of $A(k-1)$ is not a combination of previous rows of $A(k-1)$; so row i of $A(k)$ is not a combination of previous rows of $A(k)$.

(4) If multiples of previous rows of $A(k)$ are subtracted from later rows to form a matrix $B(k)$, the sets $I(k)$ are the same for $A(k)$ and $B(k)$.

LEMMA 4.1. *$I(k)$ contains exactly one row number that is not in $I(k-1)$. Call that new number $i(k)$. Every i in $\mathbf{Z}$ is $i(k)$ for one column number k.*

Reasoning : For each i that is not in $I(k-1)$, row i of $A(k-1)$ is a combination of previous rows of $A(k-1)$. Subtract from each of those rows i of $A(k)$ that combination of previous rows of $A(k)$. Then these rows i of the new matrix $B(k)$ (formed from $A(k)$) are all zero except possibly in its last column k.

> We must show that *exactly one* of these rows of $B(k)$ ends in a nonzero. Then its row number i (not in $I(k-1)$) is in $I(k)$. The permutation matrix P will have a one in that row $i(k)$, column k.

Suppose $I(k)$ contains *two* row numbers $i_1 < i_2$ that are not in $I(k-1)$. Then rows i_1 and i_2 of $B(k)$ have only zero entries before column k. Therefore row i_2 is a multiple of row i_1. Thus i_2 cannot belong to $I(k)$.

Suppose $I(k)$ contains *no* new row numbers, and equals $I(k-1)$. If i is not in $I(k-1)$, elimination can produce zeros in row i up to and including column k. Then the idea is to use column operations to produce zeros in all the remaining entries of column k. That is now a column of zeros, which contradicts the invertibility of the original matrix A.

Those column (and row) operations will produce zeros by using the pivots already located in positions $(i(j), j)$ for $j < k$. Yinghui Wang and I have discussed this sequence of steps. For finite matrices no additional hypothesis will be needed. But infinite matrices follow their own rules, and it is too early to give sufficient conditions for $A = LPU$ to be achieved.

III. The reader might enjoy a striking example of this third difficulty with infinite matrices. *These matrices have $AB = I$ but $Bx = 0$. Thus $(AB)x$ is different from $A(Bx)$*:

$$(4.1) \qquad A = \begin{bmatrix} 1 & 1 & 1 & 1 & \bullet \\ 0 & 1 & 1 & 1 & \bullet \\ 0 & 0 & 1 & 1 & \bullet \\ 0 & 0 & 0 & 1 & \bullet \\ \bullet & \bullet & \bullet & \bullet & \bullet \end{bmatrix} \quad B = \begin{bmatrix} 1 & -1 & 0 & 0 & \bullet \\ 0 & 1 & -1 & 0 & \bullet \\ 0 & 0 & 1 & -1 & \bullet \\ 0 & 0 & 0 & 1 & \bullet \\ \bullet & \bullet & \bullet & \bullet & \bullet \end{bmatrix} \quad x = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ \bullet \end{bmatrix}$$

The associative law $(AB)x = A(Bx)$ has failed! The sums and differences in A and B correspond to integrals and derivatives (and also $BA = I$). Rien Kaashoek and Richard Dudley showed us similar examples, and Alan Edelman pointed out the disturbing analogy with the fundamental theorem of calculus.

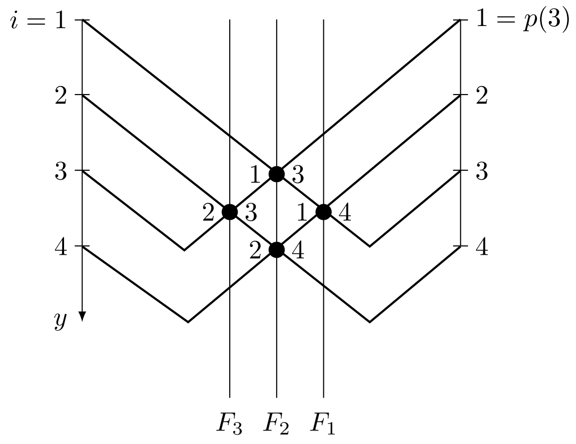
The usual proof of the associative law is a rearrangement of a double sum. For infinite series, that rearrangement is permitted when there is absolute convergence. (Changing every -1 in B to $+1$ will produce divergence in ABx.) More generally, $(AB)x = A(Bx)$ for bounded operators on a Banach space. Our problem is to stay within this framework when L and U can be unbounded.

Notice the relevance of associativity in our attempted proof above. Row operations reduced $A(k)$ to $B(k)$, and then row and column operations reduced column k to zero. Does this safely contradict the invertibility of A? (Sections of infinite matrices are analyzed by Lindner in [9]—a beautiful theory is developing.)

5. Banded Permutations and the Shifting Number

Factoring banded permutations is a combinatorial problem and Greta Panova showed how a "hooked wiring diagram" yields $P = F_1 \ldots F_N$ with $N < 2w$ factors. In the finite case [11], each factor F executes disjoint exchanges of neighbors. The intersections of wires indicate which neighbors to exchange. A second proof of $N < 2w$ is given in [1].

The diagram has wires from $1, 2, 3, 4$ to $3, 4, 1, 2$. This P has $w = 2$, each F has $w = 1$, and $N = 3$ factors are required. They were displayed in equation (4). The distance from left to right is $2w$, and all hooked lines have slope $-1/+1$.



In this example F_3 yields $1, 3, 2, 4$ by one transposition. The two exchanges in F_2 yield $3, 1, 4, 2$. Then F_1 produces $3, 4, 1, 2$. Three lines cannot meet at the same point, because two would be going in the same direction.

We must prove that intersections of hooked lines occur on at most $2w - 1$ verticals. *Suppose that $i < j$ but $p(i) > p(j)$.* The line through the left point $x = 0, y = i$ is $y = i + x$ (slope $+1$ because y increases downward). The line through the right point $x = 2w, y = p(j)$ is $y = p(j) - x + 2w$. Those lines meet (between their hooks) at $x = w + \frac{1}{2}(p(j) - i)$. We need to show that there are only $2w-1$ possible values for the integer $p(j)-i$. Then there will be only $2w-1$ possible values for x, and those $2w - 1$ vertical lines will include all the intersections.

Bandedness gives $p(j) - j \geq -w$. Adding $j - i > 0$ (which becomes $j - i \geq 1$ for integers) leaves $p(j) - i \geq 1 - w$. This is the desired bound on one side.

In the opposite direction $i - p(i) \geq -w$. Adding $p(i) - p(j) \geq 1$ leaves $i - p(j) \geq 1 - w$. So the only possibilities for $i - p(j)$ are the $2w - 1$ numbers $1 - w, \ldots, w - 1$.

The intersecting lines reveal the order for the transpositions F_i of neighbors, whose product is P. See [1, 13] for a greedy sequence of transpositions F_i. This factorization with $N < 2w$ extends to banded singly infinite permutations.

Turn now to *doubly infinite permutations* (of $\mathbf{Z}$). A new possibility appears, because the left shift matrix S is also a permutation with bandwidth $w = 1$ (so S and the right shift $S^T = S^{-1}$ become admissible factors F_i of P):

$$S(\ldots, x_0, x_1, \ldots) = (\ldots, x_1, x_2, \ldots) \text{ has } S_{ij} = 1 \text{ on the superdiagonal } j = i + 1.$$

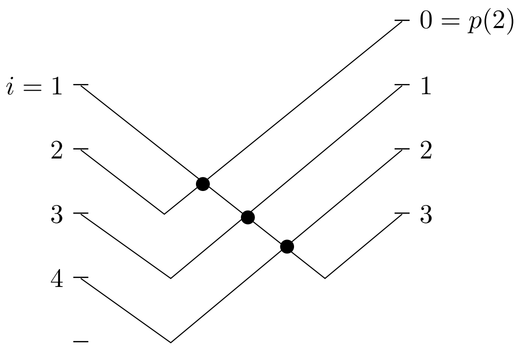
Theorem 5.1. *A banded permutation P of $\mathbf{Z}$ factors into $P = S^s F_1 \ldots F_N$ with $N < 2w$ and $|s| \leq w$. The shifting index $s(P)$ (positive or negative) has the property that*

$$s(P_1 P_2) = s(P_1) + s(P_2). \tag{5.1}$$

Proof. The pure shifts $P = S^w$ and $P = S^{-w}$ are extreme cases. For the factorization in general, we first untangle the hooked wires by a sequence of transpositions as before. After untangling, the diagram will show a shift by s. Our example is a permutation P that has period $B = 4$ and bandwidth $w = 2$:

$$p(4n+1) = 4n+3, \quad p(4n+2) = 4n, \quad p(4n+3) = 4n+1, \quad p(4n+4) = 4n+2$$

We draw one period of the wiring diagram for P: □



Three consecutive transpositions will untangle the wires. But our example still has a shift: $(1, 2, 3, 4) \to (0, 1, 2, 3)$ after the untangling. Therefore $P = S\,F_1\,F_2\,F_3$. Four rows of the matrix for P will show one period with bandwidth $w = 2$:

$$P \text{ includes } \begin{bmatrix} 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} = \begin{bmatrix} M_0 & M_1 \end{bmatrix}. \tag{5.2}$$

This permutation has shifting index $s = 1$. Every permutation factors in the same way into $P = S^s F_1 \ldots F_N$, untangling followed by possible shifts left or right. All factors are permutations of $\mathbf{Z}$ with bandwidth $w = 1$.

The rule for $s(P_1\,P_2)$ comes from this factorization. Each product FS^s is the same as $S^s f$, where f is constructed by moving all the 2 by 2 (and 1 by 1) blocks s places along the diagonal of F. Shifts in P_2 can then combine with shifts in P_1:

$$P_1\,P_2 = (S^{s_1} F_1 \ldots F_N)(S^{s_2} F_{N+1} \ldots F_M) = S^{s_1+s_2} f_1 \cdots f_N\, F_{N+1} \ldots F_M. \tag{5.3}$$

Thus the index for $P_1\,P_2$ is $s_1 + s_2 = s(P_1) + s(P_2)$.

The shifting index for a doubly infinite invertible matrix imitates the *Fredholm index* for a singly infinite matrix. That index is defined when the kernels of A and A^* are finite-dimensional:

$$\text{index}\,(A) = \dim\,(\text{kernel of } A) - \dim\,(\text{kernel of } A^*) \tag{5.4}$$

The index of $A_1 A_2$ is the sum of the separate indices. Thus index $(A) =$ index (P) if $A = LPU$ with invertible L and U. Similarly $s(A) = s(P)$ in the invertible doubly infinite case.

There is a nice connection between the Fredholm index and the shifting index. If we stay with permutations, we can sketch a simple proof of this connection:

THEOREM 5.2. *The shifting index of a banded doubly infinite permutation equals the Fredholm index of every singly infinite submatrix P_n (containing all entries P_{ij} with $i \geq n$ and $j \geq n$).*

PROOF. For permutations, the Fredholm index of P_n is just the number of zero columns minus the number of zero rows (both finite for banded P). Now remove a row and column (*vectors* r *and* c) to form P_{n+1}. If $r = c = (0, 0, \ldots)$ or if $r = c = (1, 0, 0, \ldots)$ this index is unchanged. Suppose $r = (0, 0, \ldots)$ but c contains a 1 from some row $i > n$ of P_n. Then the zero row r was removed but a new zero row i has been created in P_{n+1}. The index is again unchanged (and similarly if c is zero and r is nonzero). When both c and r have 1's, their removal creates a zero row and a zero column. So the index of P_n is independent of n.

For the doubly infinite shift S^s, all singly infinite sections $(S^s)_n$ have Fredholm index s. For $s > 0$, all sections start with s zero columns and have no zero rows. For $s < 0$, they start with $-s$ zero rows and have no zero columns. To complete the proof for any banded permutations, we express its factorization in the form $P = f_1 \ldots f_N \, S^s$ and show that the Fredholm index of every P_n stays at s (the shifting index of P).

The proof can use induction. When f_k exchanges rows n and $n+1$ of the permutation $Q = f_{k+1} \ldots f_N \, S^s$, it will also exchange those rows of the singly infinite section Q_n. The Fredholm index of Q_n is unchanged. From the first step in this proof we conclude that all exchanges of neighbors, from each factor f_k, leave the index of every section unchanged. So all those indices stay at $s = s(P)$.

After formulating this theorem on the two indices, we learned from Marko Lindner that it holds for a much wider class of doubly infinite matrices [see 9, 14, 15]. The original proof [14] is very much deeper, using K-theory. Our shifting index s is the "plus-index" in that literature, recent and growing and impressive. □

Note To compute $s(P)$ from our definition requires the factorization $P = S^s F_1 \ldots F_N$. A more intrinsic definition (if true) comes from the average shift from i to $p(i)$:

$$\text{Shifting index } s(P) = \lim_{T\to\infty} \left(\frac{1}{2T+1} \sum_{-T}^{T} (i - p(i)) \right). \tag{5.5}$$

This paper ends with a summary of the periodic (block Toeplitz) case, for which all information about A is contained in the matrix polynomial $M(z)$. The triangular factorization of $M(z)$ is a long-studied and beautiful problem. The discussion of this periodic case could extend to matrices that are not banded, but we don't go there.

6. Periodic Matrices (Block Toeplitz)

A singly or doubly infinite matrix has period B if

$$A(i+B, j+B) = A(i,j) \quad \text{for } i,j \text{ in } \mathbf{N} \text{ or } i,j \text{ in } \mathbf{Z}.$$

Rows 1 to B contain a sequence $\ldots, M_{-1}, M_0, M_1, \ldots$ of B by B blocks. In the doubly infinite case, those matrices are repeated up and down the "block diagonals" of A:

$$A = \begin{bmatrix} \bullet & \bullet & \bullet & \bullet & & & & \\ \bullet & M_0 & M_1 & M_2 & \dots & & & \\ \bullet & M_{-1} & M_0 & M_1 & M_2 & \dots & & \\ \bullet & M_{-2} & M_{-1} & M_0 & M_1 & M_2 & \dots & \\ & & & & & & & \dots \end{bmatrix}.$$

A singly infinite periodic matrix starts with M_0 in the first block as shown. The blocks above and to the left are not present.

Periodic matrices are "Toeplitz" or "stationary" or "linear time-invariant" by blocks. The natural approach to their analysis is through the B by B matrix function

$$M(z) = \sum M_j\, z^j \quad \text{(the symbol or the frequency response of } A\text{)}.$$

The matrix multiplication $y = Ax$ becomes a block multiplication $Y(z) = M(z)X(z)$ when we separate the components of x and y into blocks x_i and y_k of length B:

$$Y(z) = \sum y_k\, z^k = \left(\sum M_j\, z^j\right)\left(\sum x_i\, z^i\right) = M(z)X(z). \tag{6.1}$$

This convolution rule is the essential piece of algebra at the foundation of digital signal processing. The map $x \to X(z)$ is the Discrete Time Fourier Transform (in blocks). In this doubly infinite case, we may multiply $y = Ax$ and transform to get $Y(z)$, or we may transform first and multiply $M(z)X(z)$. Thus $FA = MF$. The singly infinite case has $i \geq 0$ in $\sum x_i\, z^i$, and we project $y = Ax$ to have $k \geq 0$ in $\sum y_k\, z^k$.

The two cases are different, but the symbol $M(z)$ governs both:

Doubly infinite

(1) A is banded if $M(z)$ has finitely many terms (a polynomial in z and z^{-1}).

(2) A is also invertible if $M(z)$ is invertible for every $|z| = 1$.

(3) A^{-1} is represented by $(M(z))^{-1}$ which involves a division by $\det M$. So A^{-1} is also banded if $\det M(z)$ is a monomial $c\, z^m$, $c \neq 0$.

(4) A is a permutation P if $M(z) = D(z)p$ is a diagonal matrix $\operatorname{diag}(z^{k_1}, \dots, z^{k_B})$ times a B by B permutation matrix p. Then $p_{ij} = 1$ corresponds to $P_{iJ} = 1$ when $J = j + k_i B$ (equal indices mod B).

(5) $A = LPU$ if $M(z) = L(z)P(z)U(z)$.

(6) The *shifting index* of P (and A) is the sum of partial indices $s = \sum k_i$.

A key point for us is that the factorization into $L(z)P(z)U(z)$ has been achieved. This theorem has a long and distinguished history beginning with Plemelj [12]. (G.D. Birkhoff's factorization corresponds to PLU.) A short direct proof, and much more, is in the valuable overview [7]. Notice that an independent proof of $A = LPU$ by elimination on infinite matrices would provide a new approach to the classical problem of factoring $M(z)$ into $L(z)P(z)U(z)$.

There are important changes in **1–6** for singly infinite matrices A, L, P, U. Those are still periodic (block Toeplitz). But LU is not periodic; its $1,1$ block is L_0U_0 but the $2,2$ block includes $L_{-1}U_1$. *The correct order for these block triangular* *matrices is* UL. This is the Wiener-Hopf factorization that solves singly infinite periodic systems. $A = UL$ is not achieved by elimination (which would have to start at a nonexistent lower right corner of A), but it follows from $M(z) = U(z)L(z)$.

We indicate the changes in (1)–(6) for the singly infinite case. Notice especially that the shifting index s becomes the Fredholm index in (6). But index zero is not the same as invertibility. So those properties are considered separately.

(1) U is banded when $U(z)$ is a matrix polynomial in z.

(2) U is invertible if $U(z)$ is invertible for $|z| \leq 1$. If U is bidiagonal, with the numbers u_0 and u_1 on diagonals 0 and 1, we need $|u_0| > |u_1|$.

(3) U^{-1} is represented by $(U(z))^{-1}$. U^{-1} is banded if det $U(z)$ is a nonzero constant.

(4) A singly infinite periodic permutation (invertible!) is block diagonal.

(5) $A = UL$ if $M(z) = U(z)L(z)$. This is Wiener-Hopf with $P(z)$ included in $L(z)$.

(6) If $P(z) = D(z)p$ with $D(z) = \text{diag}(z^{k_1}, \ldots, z^{k_B})$ times a permutation p, then the *Fredholm index* of the matrix P (and of $A = LPU$) is $\sum k_i$.

The example in section 5 (with period $B = 4$) illustrates the Fredholm index in the singly infinite case (6):

$$P = \begin{bmatrix} M_0 & M_1 & 0 & \bullet \\ 0 & M_0 & M_1 & \bullet \\ 0 & 0 & \bullet & \bullet \end{bmatrix} \quad M_0 = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix} \quad M_1 = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

In this case $\det(M_0 + M_1 z) = z$. The Fredholm index of P is 1. The kernel of P is spanned by $(1, 0, 0, \ldots)$. The diagonal matrix $D(z)$ is $\text{diag}(1, 1, z, 1)$. The multiplicative property of $\det(P_1(z)\, P_2(z))$ confirms that $\text{index}(P_1\, P_2) = \text{index}(P_1) + \text{index}(P_2)$. Those indices are the exponents of z in the determinants of $P_1(z)$ and $P_2(z)$.

Acknowledgements

Many friends have discussed sections of this paper with me. It is a pleasure to thank each of them for generous help.

Factoring CMV matrices (Vadim Olshevsky and Pavel Zhlobich)

Factoring matrix polynomials (Ilya Spitkovsky, Rien Kaashoek, and Patrick Dewilde)

Factoring banded permutations (Greta Panova, Alex Postnikov, and Chi-Kwong Li)

Bruhat factorizations (David Vogan and George Lusztig)

Shifting index and Fredholm index (Marko Lindner)

Factoring infinite matrices (Yinghui Wang)

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. M.J . Cantero, L. Moral and L. Velazquez, *Five-diagonal matrices and zeros of orthogonal polynomials on the unit circle*, Lin. Alg. Appl. 362 (2003) 29-56.
4. I. Daubechies, *Orthonormal bases of compactly supported wavelets*, Comm. Pure Appl. Math. 41 (1988) 909-996.
5. L. Elsner, *On some algebraic problems in connection with general eigenvalue algorithms*, Lin. Alg. Appl. 26 (1979) 123-138.
6. I. Gohberg and S. Goldberg, *Finite dimensional Wiener-Hopf equations and factorizations of matrices*, Lin. Alg. Appl. 48 (1982) 219-236.
7. I. Gohberg, M. Kaashoek, and I. Spitkovsky. *An overview of matrix factorization theory and operator applications*, Operator Th. Adv. Appl., Birkhäuser (2003) 1-102.
8. L. Yu. Kolotilina and A. Yu. Yeremin, *Bruhat decomposition and solution of sparse linear algebra systems*, Soviet J. Numer. Anal. Math. Modelling 2 (1987) 421-436.
9. M. Lindner, *Infinite Matrices and Their Finite Sections*, Birkhäuser 2006.
10. V. Olshevsky, P. Zhlobich, and G. Strang, *Green's matrices*, Lin. Alg. Appl. 432 (2010) 218-241.
11. G. Panova, *Factorization of banded permutations*, arXiv.org/abs/1007.1760 (2010).
12. J. Plemelj, *Riemannsche Funktionenscharen mit gegebener Monodromiegruppe*, Monat. Math. Phys. 19 (1908) 211-245.
13. M.D. Samson and M.F. Ezerman, *Factoring permutation matrices into a product of tridiagonal matrices*, arXiv.org/abs/1007.3467 (2010).
14. V.S. Rabinovich, S. Roch, and J. Roe, *Fredholm indices of band-dominated operators*, Integral Eqns. Oper. Th. 49 (2004) 221-238.
15. V.S. Rabinovich, S. Roch, and B. Silbermann, *Fredholm theory and finite section method for band-dominated operators*, Integral Eqns. Oper. Th. 30 (1998) 452-495.
16. G. Strang, *Fast transforms : Banded matrices with banded inverses*, Proc. Natl. Acad. Sci. 107 (2010) 12413-12416.
17. G. Strang, *Groups of banded matrices with banded inverses*, Proceedings of the AMS.
18. G. Strang and T. Nguyen, *The interplay of ranks of submatrices*, SIAM Review 46 (2004) 637-646.

Dept. of Mathematics MIT Cambridge MA 02139 USA
Current address: Dept. of Mathematics MIT Cambridge MA 02139 USA
E-mail address: gs@math.mit.edu

Proceedings of ICCM2010 (International Congress of Chinese Mathematicians, Beijing, Dec 2010)