Contents

Chapter

Introduction

The central theme in this book is to investigate and explore various properties of the class of *totally nonnegative matrices.*

At first it may appear that the notion of total positivity is artificial; however, this class of matrices arises in a variety of important applications. For example, totally positive (or nonnegative) matrices arise in statistics (see [GM96, Goo86, Hei94]), mathematical biology (see [GM96]), combinatorics (see [BFZ96, Bre95, Gar96, Peñ98]), dynamics (see [GK02]), approximation theory (see [GM96, Pri68, CP94b]), operator theory (see [Sob75]), and geometry (see [Stu88]). Historically, the theory of totally positive matrices originated from the pioneering work of Gantmacher and Krein ([GK60]) and was elegantly brought together in a revised version of their original monograph [GK02]). Also, under the influence of I. Schoenberg (see, for example, [Sch30]), Karlin published an influential treatise, *Total Positivity* ([Kar68]), which mostly concerns totally positive kernels but also deals with the discrete version, totally positive matrices. Since then there has been a considerable amount of work accomplished on total positivity, some of which is contained in the exceptional survey paper by T. Ando ([And87]). Two more recent survey papers have appeared ([Fal01] [Hog07, Chap. 21]) and both take the point of view of bidiagonal factorizations of totally nonnegative matrices.

Before we continue, as mentioned in the Preface, we issue here a word of warning. In the existing literature, the terminology used is not always consistent with what is presented above. Elsewhere in the literature the term totally positive (see below) corresponds to "strictly totally positive" (see, for example, the text [Kar68]) and the term totally nonnegative defined here sometimes corresponds to "totally positive" (as in some of the papers in [GM96]).

0.0 DEFINITIONS AND NOTATION

The set of all m-by-n matrices with entries from a field $\mathbb{F}$ will be denoted by $M_{m,n}(\mathbb{F})$. If $\mathbb{F} = \mathbb{R}$, the set of all real numbers, then we may shorten this notation to $M_{m,n}$. Our matrices will, for the most part, be real entried. In the case $m = n$, the matrices are square and $M_{n,n}(\mathbb{F})$ or $M_{m,n}$ will be shortened to $M_n(\mathbb{F})$ or M_n. For $A \in M_{m,n}(\mathbb{F})$, the notation $A = [a_{ij}]$ will indicate that the entries of A are $a_{ij} \in \mathbb{F}$, for $i = 1, 2, \ldots, m$ and $j = 1, 2, \ldots, n$. The transpose of a matrix A will be denoted by A^T.

For $A \in M_{m,n}(\mathbb{F})$, $\alpha \subseteq \{1, 2, \ldots, m\}$, and $\beta \subseteq \{1, 2, \ldots, n\}$, the submatrix of A lying in rows indexed by α and columns indexed by β will be denoted by $A[\alpha, \beta]$. Similarly, $A(\alpha, \beta)$ is the submatrix obtained from A by deleting the rows indexed by α and columns indexed by β. Throughout the book, we let α^c denote the *complement* of the index set α. So, in particular, $A(\alpha, \beta) = A[\alpha^c, \beta^c]$. If $A \in M_n(\mathbb{F})$ and $\alpha = \beta$, then the principal submatrix $A[\alpha, \alpha]$ is abbreviated to $A[\alpha]$, and the complementary principal submatrix is $A(\alpha)$. On the other hand, if $\alpha \neq \beta$, then the submatrix $A[\alpha, \beta]$ is referred to as a *nonprincipal submatrix of A*. In the special case when $\alpha = \{1, 2, \ldots, k\}$ with $1 \leq k \leq n$, we refer to the principal submatrix $A[\alpha]$ as a *leading principal submatrix*. In the same manner, for an n-vector $x \in \mathbb{F}^n$, $x[\alpha]$ denotes the entries of x in the positions indexed by α and $x(\alpha)$ denotes the complementary vector. If $x = [x_i] \in \mathbb{F}^n$, then we let $\mathrm{diag}(x_i)$ denote the n-by-n diagonal matrix with main diagonal entries x_i. For brevity, we often denote the sets $\{1, 2, \ldots, m\}$ and $\{1, 2, \ldots, n\}$ by M and N, respectively.

A *minor* in a given matrix A is by definition the determinant of a (square) submatrix of $A \in M_{m,n}(\mathbb{F})$. Here and throughout, $\det(\cdot)$ denotes the determinant of a square matrix. For example, if $|\alpha| = |\beta|$, the minor $\det A[\alpha, \beta]$ may also be denoted by $A_{\alpha, \beta}$, and the principal minor (if A is square) $\det A[\alpha]$ may be abbreviated to A_α.

We are interested here in those matrices $A \in M_{m,n}$ with *all* minors positive (nonnegative) and, occasionally, with all minors $A_{\alpha, \beta}$, with $|\alpha| = |\beta| \leq k$ positive (nonnegative). The former are called the *totally positive* (*totally nonnegative*) matrices and are denoted by TP (TN); we denote the latter as TP_k (TN_k). TP (TN) is used both as an abbreviation and as a name of a set. The reader should consult [HJ85, HJ91] or any other standard reference on matrix theory for standard notation and terminology not defined herein. In terms of compounds, TP_k, for example, simply means that the first through the kth compounds are entry-wise positive.

Several related classes of matrices are both of independent interest and useful in developing the theory of TP (TN) matrices. We give our notation for those classes here; we continue to use the subscript k to denote a requirement made on all minors of no more than k rows. No subscript indicates that the requirement is made on all minors. Each class is defined irrespective of the m and n in $M_{m,n}$, except that a subscript k makes sense only for $k \leq m, n$.

By InTN (InTN_k), we denote the square and *invertible* TN matrices (TN_k matrices), and by IrTN (IrTN_k), we mean the square and *irreducible* TN matrices (TN_k matrices). The intersection of these two classes is denoted by IITN (IITN_k); this class is also referred to as the *oscillatory* matrices [GK60], which are defined as those TN matrices with a TP integral power. The equivalence will be demonstrated in Chapter 2. We may also use the symbol OSC to denote oscillatory matrices.

A TN (TN_k) matrix may have a zero line (row or column), which often leads to an exception to convenient statements. So we denote TN (TN_k) matrices with no zero lines as TN′ (TN'_k).

A square, triangular matrix cannot be TP, but one that is TN *and* has all its minors positive unless they are identically 0 because only of the triangular structure is called *triangular TP*; these are denoted by ΔTP. Similarly, ΔTN denotes the triangular, TN matrices. An adjective "upper" or "lower" denotes the form of triangularity. Finally, the positive diagonal matrices ($\mathcal{D}_n$) are the intersections of upper ΔTP and lower ΔTP matrices, and thus are InTN.

For convenience of reference we present these classes in tabular form in Table 1. Note: Omission of subscript k means *all* minors.

TP_k :	all minors on no more than k rows are positive
TN_k :	all minors on no more than k rows are nonnegative
TN'_k :	TN_k with no all zero lines
InTN_k :	TN_k and invertible
IrTN_k :	TN_k and irreducible
IITN_k :	TN_k, irreducible and invertible
ΔTP:	triangular TP; all minors that may be positive are positive
ΔTN :	triangular TN
$\mathcal{D}_n$:	positive diagonal matrices in M_n

Table 1: Various Classes of Matrices of Interest

The ith standard basis vector is the n-vector whose only nonzero entry occurs in the ith component and that entry is a one, and is denoted by e_i; the (i,j)th standard basis matrix, the m-by-n matrix whose only nonzero entry is in the (i,j)th position and this entry is a one, will be denoted by E_{ij}. Observe that if $m = n$, then $E_{ij} = e_i e_j^T$. We also let e denote the n-vector consisting of all ones (the size of e will be determined from the context). Finally, we let $J_{m,n}$ ($J_{n,n} \equiv J_n$) and I_n denote the m-by-n matrix of all ones and the n-by-n identity matrix, respectively. The subscript is dropped when the sizes of these matrices is clear from the context.

0.1 JACOBI MATRICES AND OTHER EXAMPLES OF TN MATRICES

An n-by-n matrix $A = [a_{ij}]$ is called a *diagonal* matrix if $a_{ij} = 0$ whenever $i \neq j$. For example,

$$A = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 4 & 0 & 0 \\ 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 3 \end{bmatrix}$$

is a 4-by-4 diagonal matrix. We may also write A as $A = \text{diag}\,(1,4,3,3)$.

An n-by-n matrix $A = [a_{ij}]$ is referred to as a *tridiagonal* matrix if $a_{ij} = 0$ whenever $|i - j| > 1$. For example,

$$A = \begin{bmatrix} 1 & 4 & 0 & 0 \\ 3 & 4 & 1 & 0 \\ 0 & 0 & 3 & 4 \\ 0 & 0 & 1 & 3 \end{bmatrix}$$

is a 4-by-4 tridiagonal matrix. The entries of $A = [a_{ij}]$ that lie in positions $(i, i+1)$ $i = 1, \dots n - 1$ are referred to as the *superdiagonal*, and the entries of A in positions $(i, i-1)$ $i = 2, \dots, n$ are called *subdiagonal*.

Hence a matrix is tridiagonal if the only nonzero entries of A are contained on its sub-, main, and superdiagonal.

Tridiagonal matrices are perhaps one of the most studied classes of matrices. Much of the reason for this is that many algorithms in linear algebra require significantly less computational labor when they are applied to tridiagonal matrices. Some elementary examples include

(1) eigenvalues,

(2) solving linear systems,

(3) LU factorization,

(4) evaluating determinants.

For instance, the determinant of a tridiagonal matrix $A = [a_{ij}]$ can be evaluated by solving the recursive equation

$$\det A = a_{11} \det A[\{2, \dots, n\}] - a_{12}a_{21} \det A[\{3, \dots, n\}]. \tag{1}$$

Equation (1) is a simple consequence of the Laplace expansion of $\det A$ along row (or column) 1. Furthermore, both $A[\{2, \dots, n\}]$ and $A[\{3, \dots, n\}]$ are (smaller sized) tridiagonal matrices.

The recursive relation (1) also ties the eigenvalues of a (symmetric) tridiagonal matrix to certain orthogonal polynomials (Chebyshev), as the characteristic polynomials of A and its principal submatrices also satisfy a relation like (1). In addition, (1) arises in the Runge-Kutta methods for solving partial differential equations.

Suppose the tridiagonal matrix A has the form

$$A = \begin{bmatrix} a_1 & b_1 & 0 & \cdots & & 0 \\ c_1 & a_2 & \ddots & \ddots & & \vdots \\ 0 & \ddots & \ddots & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & & b_{n-1} \\ 0 & \cdots & \cdots & & c_{n-1} & a_n \end{bmatrix}. \tag{2}$$

Observe that a tridiagonal matrix A in the form (2) is called symmetric if $b_i = c_i$, $i = 1, \dots, n - 1$ and is called irreducible if both $b_i \neq 0$ and $c_i \neq 0$, for $i = 1, \dots, n - 1$.

An n-by-n matrix is called a P_0-matrix if all its principal minors are non-negative. We let P_0 denote the class of all P_0-matrices. Furthermore, an n-by-n matrix is called a P-matrix if all its principal minors are positive. We let P denote the class of all P-matrices.

Suppose A in (2) is an irreducible nonnegative ($a_i, b_i, c_i \geq 0$) tridiagonal matrix. If $a_i \geq b_i + c_{i-1}$ $i = 1, \ldots, n$, where $c_0 \equiv 0, b_n \equiv 0$, then A is a P_0-matrix. To verify this claim we need the following fact. If A is tridiagonal, then the principal submatrix $A[\{1, 2, 4, 6, 7, 8\}]$ can be written as $A[\{1, 2, 4, 6, 7\}] = A[\{1, 2\}] \oplus A[\{4\}] \oplus A[\{6, 7, 8\}]$, where $\oplus$ denotes direct sum. In particular,

$$\det A[\{1, 2, 4, 6, 7\}] = \det A[\{1, 2\}] \cdot \det A[\{4\}] \cdot \det A[\{6, 7, 8\}].$$

Hence to calculate any principal minor of a tridiagonal matrix, it is enough to compute principal minors based on associated contiguous index sets. Moreover, any principal submatrix of a tridiagonal matrix based on contiguous index sets is again a tridiagonal matrix. To verify that any tridiagonal of the form (2) that is nonnegative, irreducible and satisfies $a_i \geq b_i + c_{i-1}$ (row diagonal dominance) is a P_0-matrix, it is sufficient, by induction, to verify that $\det A \geq 0$.

Consider the case $n = 3$. Then by (1),

$$\begin{aligned}
\det A &= a_1 \det A[\{2, 3\}] - b_1 c_1 \det A[\{3\}] \\
&= a_1(a_2 a_3 - b_2 c_2) - b_1 c_1 a_3 \\
&= a_1 a_2 a_3 - a_1 b_2 c_2 - b_1 c_1 a_3 \\
&\geq a_1(c_1 + b_2)a_3 - a_1 b_2 c_2 - b_1 c_1 a_3 \\
&= a_1 b_2(a_3 - c_2) + a_3 c_1(a_1 - b_1) \\
&> 0.
\end{aligned}$$

In addition, since $\det A[\{3\}] > 0$ it follows from (1) that

$$\det A \leq a_1 \det A[\{2, 3\}]. \tag{3}$$

Consider the general case and let A' be the n-by-n matrix obtained from A by using an elementary row operation to eliminate c_1. Then, as $a_1 > b_1 > 0$,

$$\det A = a_1 \det \begin{bmatrix}
a_2 - \frac{b_1 c_1}{a_1} & b_2 & 0 & \cdots & 0 \\
c_2 & a_3 & b_3 & \cdots & 0 \\
\vdots & \ddots & \ddots & \ddots & \vdots \\
0 & \vdots & \ddots & \ddots & b_{n-1} \\
0 & \cdots & \cdots & c_{n-1} & a_n
\end{bmatrix}.$$

Observe that this matrix is an irreducible, nonnegative tridiagonal matrix that still satisfies the row dominance relation and hence by induction has a positive determinant. Thus $\det A > 0$.

Furthermore, we note that by (1) we have

$$\det A \leq a_1 \det A[\{2, \ldots, n\}].$$

So, if a nonnegative, irreducible tridiagonal matrix satisfies the simple dominance relation it is a P_0-matrix. In fact, even more is true.

Since tridiagonal matrices satisfy $a_{ij} = 0$ whenever $|i - j| > 1$ it is a simple consequence that

$$\det A[\{i_1, \ldots, i_k\}, \{j_1, \ldots, j_k\}] = 0$$

whenever there exists an $s \in \{1, \ldots, k\}$ such that $|i_s - j_s| > 1$.

Furthermore, if $A[\{i_1, \ldots, i_k\}, \{j_1, \ldots, j_k\}]$ is any submatrix of A that satisfies $|i_s - j_s| \leq 1, s = 1, \ldots, k$, then $\det A[\{i_1, \ldots, i_k\}, \{j_1, \ldots, j_k\}]$ is a product of principal minors and nonprincipal minors, where the principal minors are based on contiguous index sets and the nonprincipal minors are simply entries from the super- and/or subdiagonal.

For example, suppose A is a 10-by-10 tridiagonal matrix of the form in (2).

Then

$$\det A[\{1, 2, 3, 5, 6, 8, 10\}, \{2, 3, 4, 5, 6, 8, 9\}] = b_1 b_2 b_3 \det A[\{5, 6\}] a_8 c_9.$$

Consequently, any irreducible, nonnegative tridiagonal matrix that satisfies the condition $a_i \geq b_i + c_{i-1}$ has all its minors nonnegative.

We remark here that the irreducibility assumption was needed for convenience only. If the tridiagonal matrix is reducible, then we simply concentrate on the direct summands and apply the same arguments as above.

The upshot of the previous discussion is that if A is a nonnegative tridiagonal matrix that also satisfies $a_i \geq b_i + c_{i-1}$, then A is TN.

In the pivotal monograph [GK60], the chapter on oscillatory matrices began by introducing a class of tridiagonal matrices the authors called *Jacobi matrices*. An n-by-n matrix $\mathcal{J} = [a_{ij}]$ is a Jacobi matrix if $\mathcal{J}$ is a tridiagonal matrix of the form

$$\begin{bmatrix} a_1 & -b_1 & \cdots & & 0 \\ -c_1 & \ddots & \ddots & & \vdots \\ 0 & \ddots & \ddots & & -b_{n-1} \\ 0 & \cdots & & -c_{n-1} & a_n \end{bmatrix}.$$

Further, $\mathcal{J}$ is called a *normal Jacobi matrix* if, in addition, $b_i, c_i \geq 0$. Jacobi matrices were of interest to Gantmacher and Krein for at least two reasons. First, many of the fundamental properties of general oscillatory matrices can be obtained by an examination of analogous properties of Jacobi matrices. Second Jacobi matrices are among the basic types of matrices that arose from studying the oscillatory properties of an elastic segmental continuum (no supports between the endpoints) under small transverse oscillations.

One of the first things that we observe about Jacobi matrices is that they are not TN or even entrywise nonnegative. However, they are not that far off, as we shall see in the next result.

The following result on the eigenvalues of a tridiagonal matrix is well known, although we present a proof here for completeness.

Lemma 0.1.1 *Let T be an n-by-n tridiagonal matrix in the form*

$$T = \begin{bmatrix} a_1 & b_1 & 0 & 0 & 0 & \cdots & 0 \\ c_1 & a_2 & b_2 & 0 & 0 & \cdots & 0 \\ 0 & c_2 & a_3 & b_3 & 0 & \cdots & 0 \\ \vdots & & \ddots & \ddots & \ddots & & \vdots \\ 0 & \cdots & 0 & c_{n-3} & a_{n-2} & b_{n-2} & 0 \\ 0 & \cdots & 0 & 0 & c_{n-2} & a_{n-1} & b_{n-1} \\ 0 & \cdots & 0 & 0 & 0 & c_{n-1} & a_n \end{bmatrix}. \tag{4}$$

If $b_i c_i > 0$ for $i = 1, 2, \ldots, n-1$, then the eigenvalues of T are real and have algebraic multiplicity one (i.e., are simple). Moreover, T is similar (via a positive diagonal matrix) to a symmetric nonnegative tridiagonal matrix.

Proof. Let $D = [d_{ij}]$ be the n-by-n diagonal matrix where $d_{11} = 1$, and for $k > 1$, $d_{kk} = \sqrt{\frac{b_1 b_2 \cdots b_{k-1}}{c_1 c_2 \cdots c_{k-1}}}$. Then it is readily verified that

$$DTD^{-1} = \begin{bmatrix} a_1 & \sqrt{b_1 c_1} & 0 & \cdots & 0 \\ \sqrt{b_1 c_1} & a_2 & \sqrt{b_2 c_2} & \cdots & 0 \\ 0 & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \sqrt{b_{n-2} c_{n-2}} & a_{n-1} & \sqrt{b_{n-1} c_{n-1}} \\ 0 & \cdots & 0 & \sqrt{b_{n-1} c_{n-1}} & a_n \end{bmatrix}.$$

Since DTD^{-1} is symmetric and T and DTD^{-1} have the same eigenvalues, this implies that the eigenvalues of T are real. Suppose λ is an eigenvalue of T. Then $\lambda I - T$ is also of the form (4). If the first row and last column of $\lambda I - T$ are deleted, then the resulting matrix is an $(n-1)$-by-$(n-1)$ upper triangular matrix with no zero entries on its main diagonal, since $b_i c_i > 0$, for $i = 1, 2, \ldots, n-1$. Hence this submatrix has rank $n-1$. It follows that $\lambda I - T$ has rank at least $n-1$. However, $\lambda I - T$ has rank at most $n-1$ since λ is an eigenvalue of T. So by definition λ has geometric multiplicity one. This completes the proof, as DTD^{-1} is also a nonnegative matrix. $\square$

A diagonal matrix is called a *signature matrix* if all its diagonal entries are ± 1. Furthermore, two n-by-n matrices A and B are said to be *signature similar* if $A = SBS^{-1}$, where S is a signature matrix. It is not difficult to verify that a normal Jacobi matrix is signature similar to a nonnegative tridiagonal matrix. Furthermore, the eigenvalues of a normal Jacobi matrix are real and distinct.

Observe that if an irreducible normal Jacobi matrix satisfies the dominance condition $(a_i \geq |b_i| + |c_i|)$, then it is similar to a symmetric TN tridiagonal matrix. In particular (since a symmetric TN matrix is an example of a positive semidefinite matrix [HJ85]), the eigenvalues are nonnegative and distinct by Lemma 0.1.1.

Moreover, if the above normal Jacobi matrix is also invertible, then the eigenvalues are positive and distinct.

Since the eigenvalues are distinct, each eigenvalue has exactly one eigenvector (up to scalar multiple), and these (unique) eigenvectors satisfy many more interesting "sign" properties.

Let

$$\mathcal{J} = \begin{bmatrix} a_1 & b_1 & 0 & \cdots & & 0 \\ c_1 & a_2 & \ddots & \ddots & & \vdots \\ 0 & \ddots & \ddots & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & b_{n-1} \\ 0 & \cdots & \cdots & c_{n-1} & a_{n-1} \end{bmatrix}, \quad \text{with } b_i c_i > 0.$$

Given $\mathcal{J}$ above, define

$$D_0(\lambda) \equiv 1,$$
$$D_k(\lambda) = \det((\mathcal{J} - \lambda I)[1, \ldots, k]),$$
$$D_n(\lambda) = \text{char. poly. of } \mathcal{J}.$$

Recall that the functions D_k satisfy the recurrence relation

$$D_{k+1}(\lambda) = (a_{k+1} - \lambda) D_k(\lambda) - b_k c_k D_{k-1}(\lambda) \tag{5}$$

for $k = 1, 2, \ldots, n - 1$. Evidently,

$$D_0(\lambda) \equiv 1,$$
$$D_1(\lambda) = a_1 - \lambda,$$
$$D_2(\lambda) = (a_2 - \lambda) D_1(\lambda) - b_1 c_1 D_0(\lambda)$$
$$= (a_2 - \lambda)(a_1 - \lambda) - b_1 c_1,$$

and so on.

A key observation that needs to be made is the following:

"When $D_k(\lambda)$ vanishes ($1 < k < n$), the polynomials $D_{k-1}(\lambda)$ and $D_{k+1}(\lambda)$ differ from zero and have opposite sign."

Suppose that, for some k, $1 < k < n$, this property fails. That is, for fixed λ_0,

$$D_k(\lambda_0) = D_{k-1}(\lambda_0) = D_{k+1}(\lambda_0) = 0.$$

Go back to the recurrence relation (5), and notice that if the above inequalities hold, then

$$D_{k-2}(\lambda_0) = 0, \text{ and } D_{k+2}(\lambda_0) = 0.$$

Consequently, $D_k(\lambda_0) = 0$ for all $k = 2, 3, \ldots, n$. But then observe

$$D_3(\lambda_0) = (a_k - \lambda) D_2(\lambda_0) - b_2 c_2 D_1(\lambda_0),$$

which implies $D_1(\lambda_0) = 0$. Now we arrive at a contradiction since

$$D_2(\lambda_0) = (a_2 - \lambda) D_1(\lambda_0) - b_1 c_1 D_0(\lambda_0) = 0 = 0 - b_1 c_1,$$

and $b_1 c_1 > 0$.

Since we know that the roots are real and distinct, when λ passes through a root of $D_k(\lambda)$, the product $D_k D_{k-1}$ must reverse sign from $+$ to $-$. A consequence of this is that the roots of successive D_k's strictly interlace. Using this interlacing property we can establish the following simple but very important property of the functions $D_k(\lambda)$. Between each two adjacent roots of the polynomial $D_k(\lambda)$ lies exactly one root of the polynomial $D_{k-1}(\lambda)$ (that is, the roots *interlace*).

Observation: The sequence

$$D_{n-1}(\lambda), D_{n-2}(\lambda), \ldots, D_0(\lambda)$$

has $j - 1$ sign changes when $\lambda = \lambda_j$ where $\lambda_1 < \lambda_2 < \cdots < \lambda_n$ are the distinct eigenvalues of $\mathcal{J}$.

Further note that $D_k(\lambda) = (-\lambda)^k + \cdots$, and so it follows that $\lim_{\lambda \to -\infty} D_k(\lambda) = \infty$.

Finally, we come to the coordinates of the eigenvectors of $\mathcal{J}$. Assume for the moment that for each i, $b_i = 1$ and $c_i = 1$ (for simplicity's sake).

Suppose

$$x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

is an eigenvector of $\mathcal{J}$ corresponding to the eigenvalue λ. Then examining the vector equation $\mathcal{J}x = \lambda x$ yields the system of equations

$$
\begin{aligned}
(a_1 - \lambda)x_1 - x_2 &= 0, \\
-x_1 + (a_2 - \lambda)x_2 - x_3 &= 0, \\
-x_{n-2} + (a_{n-1} - \lambda)x_{n-1} - x_n &= 0, \\
-x_{n-1} + (a_n - \lambda)x_n &= 0.
\end{aligned}
$$

Since $D_n(\lambda) = 0$, it follows that $D_{n-1}(\lambda) \neq 0$; hence the first $n-1$ equations are linearly independent, and the nth equation is a linear combination of the previous $n - 1$.

Looking more closely at the first $n - 1$ equations we arrive at the relation

$$x_k = (a_{k-1} - \lambda)x_{k-1} - x_{k-2}.$$

Hence the coordinates of x satisfy the same relation as $D_{k-1}(\lambda)$.

Hence $x_k = C D_{k-1}(\lambda)$, $C \neq 0$ constant. Thus, since at $\lambda = \lambda_j$ the sequence $D_{n-1}, \ldots, D_0$ has exactly $j - 1$ sign changes, it follows that the eigenvector corresponding to λ_j has exactly $j - 1$ sign changes. The interlacing of sign changes in the eigenvectors follows as well.

We close with a remark that offers a nice connection between tridiagonal TN matrices and a certain type of additive closure (a property not enjoyed by general TN matrices). Suppose A is a TN matrix. Then a simple examination of the 2-by-2 minors of A involving exactly one main diagonal entry will be enough to verify the following statement. The matrix A is tridiagonal if and only if $A + D$ is TN for all positive diagonal matrices D. This result also brings out the fact that tridiagonal P_0 matrices are, in fact, TN.

0.1.1 Other Examples of TN Matrices

In the previous subsection we demonstrated that any entrywise nonnegative tridiagonal P_0-matrix was in fact TN. Other consequences include (1) an entry-wise nonnegative invertible P_0 tridiagonal matrix is both InTN and a P-matrix, and (2) an entrywise nonnegative invertible P_0 tridiagonal matrix is IITN and hence oscillatory.

From tridiagonal matrices, it is natural to consider their inverses as another example class of InTN matrices.

Example 0.1.2 (Inverses of tridiagonal matrices) Let T be an InTN tridiagonal matrix. Then $ST^{-1}S$ is another InTN matrix. For example, if

$$T = \begin{bmatrix} 2 & 1 & 0 \\ 1 & 2 & 1 \\ 0 & 2 & 1 \end{bmatrix}, \text{ then}$$

$$ST^{-1}S = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 2 \\ 1 & 2 & 3 \end{bmatrix}.$$

In fact, note that the matrix

$$[\min\{i, j\}] = \begin{bmatrix} 1 & 1 & 1 & \cdot & 1 \\ 1 & 2 & 2 & \cdot & 2 \\ 1 & 2 & 3 & \cdot & 3 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 2 & 3 & \cdots & n \end{bmatrix}$$

is the inverse of a tridiagonal matrix and is also InTN.

We also acknowledge that these matrices are referred to as "single-pair matrices" in [GK60, pp. 79–80], are very much related to "Green's matrices" in [Kar68, pp. 110–112], and are connected with "matrices of type D" found in [Mar70a].

Tridiagonal and inverse tridiagonal TN matrices have appeared in numerous places throughout mathematics. One instance is the case of the Cayley transform. Let $A \in M_n(\mathbb{C})$ such that $I + A$ is invertible. The *Cayley transform* of A, denoted by $\mathcal{C}(A)$, is defined to be

$$\mathcal{C}(A) = (I + A)^{-1}(I - A).$$

The Cayley transform was defined in 1846 by A. Cayley. He proved that if A is skew-Hermitian, then $\mathcal{C}(A)$ is unitary, and conversely, provided, of course, that $\mathcal{C}(A)$ exists. This feature is useful, for example, in solving matrix equations subject to the solution being unitary by transforming them into equations for skew-Hermitian matrices. One other important property of the Cayley transform is that it can be viewed as an extension to matrices of the conformal mapping

$$T(z) = \frac{1 - z}{1 + z}$$

from the complex plane into itself. In [FT02] the following result was proved connecting essentially nonnegative tridiagonal matrices and TN matrices. Recall that a matrix is called *essentially nonnegative* if all its off-diagonal entries are nonnegative.

Theorem 0.1.3 *Let $A \in M_n(\mathbb{R})$ be an irreducible matrix. Then A is an essentially nonnegative tridiagonal matrix with $|\lambda| < 1$ for all eigenvalues λ of A if and only if $I + A$ and $(I - A)^{-1}$ are TN matrices. In particular, $\mathcal{C}(-A) = (I - A)^{-1}(I + A)$ is a factorization into TN matrices.*

Proof. To verify necessity, observe that if $I + A$ is TN, then A is certainly essentially nonnegative. Also, since $(I - A)^{-1}$ is TN, it follows that $I - A$ is invertible and has the checkerboard sign pattern (i.e., the sign of its (i, j)th entry is $(-1)^{i+j}$). Hence $a_{i,i+2} = 0$ and $a_{i+2,i} = 0$ for all $i \in \{1, 2, \ldots, n-2\}$, and since A is irreducible and $I + A$ is TN, $a_{i,j} = 0$ for $|i - j| > 1$. That is, A is tridiagonal. The remaining conclusion now readily follows.

For the converse, if A is an essentially nonnegative irreducible tridiagonal matrix with $|\lambda| < 1$ for all eigenvalues λ of A, then $I + A$ is a nonnegative tridiagonal P-matrix and thus totally nonnegative (see previous section). Similarly, $I - A$ is a tridiagonal M-matrix since $|\lambda| < 1$ for all eigenvalues λ of A (recall that an n-by-n P-matrix is called an M-matrix if all off-diagonal entries are nonpositive), and hence $(I - A)^{-1}$ is TN (this follows from the remarks in the previous section). Since TN matrices are closed under multiplication, $\mathcal{C}(-A) = (I - A)^{-1}(I + A)$ is TN. $\square$

A natural question arising from Theorem 0.1.3 is whether in every factorization $\hat{F} = (I - A)^{-1}(I + A)$ of a totally nonnegative matrix $\hat{F}$ the factors $(I - A)^{-1}$ and $(I + A)$ are TN. We conclude with an example showing that neither of these factors need be TN.

Consider the TN matrix

$$\hat{F} = \begin{bmatrix} 1 & .9 & .8 \\ .9 & 1 & .9 \\ 0 & .9 & 1 \end{bmatrix}$$

and consider $A = -\mathcal{C}(\hat{F})$. Then $\hat{F} = \mathcal{C}(-A) = (I - A)^{-1}(I + A)$, where neither

$$(I - A)^{-1} = \begin{bmatrix} 1 & .45 & .4 \\ .45 & 1 & .45 \\ 0 & .45 & 1 \end{bmatrix} \quad \text{nor} \quad I + A = \begin{bmatrix} .8203 & .3994 & .2922 \\ .6657 & .5207 & .3994 \\ -.2996 & .6657 & .8203 \end{bmatrix}$$

is TN.

Example 0.1.4 (Vandermonde) Vandermonde matrices have been an example class of matrices that have garnered much interest for some time. In many respects it is not surprising that certain Vandermonde matrices are TP (see Chapter 1 for more discussion). For n real numbers $0 < x_1 < x_2 <$

$\ldots x_n$, the Vandermonde matrix $V(x_1, x_2, \cdots < x_n)$ is defined to be

$$V(x_1, x_2, \ldots, x_n) = \begin{bmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^{n-1} \\ 1 & x_2 & x_2^2 & \cdots & x_2^{n-1} \\ \vdots & \vdots & & \vdots & \vdots \\ 1 & x_n & x_n^2 & \cdots & x_n^{n-1} \end{bmatrix},$$

and is a TP matrix (see [GK60, p. 111]). Recall the classical determinantal formula,

$$\det V(x_1, x_2, \ldots, x_n) = \prod_{i>j}(x_i - x_j),$$

which is positive whenever $0 < x_1 < x_2 < \cdots < x_n$.

The reader is also encouraged to consult [GK60] for similar notions involving generalized Vandermonde matrices. Totally positive Vandermonde matrices were also treated in the survey paper on bidiagonal factorizations of TN matrices [Fal01], which we discuss further below.

It is really quite remarkable that the inequalities $0 < x_1 < x_2 < x_3$ are actually sufficient to guarantee that *all* minors of V above are positive. Indeed, any 2-by-2 submatrix of V has the form

$$A = \begin{bmatrix} x_i^{\alpha_1} & x_i^{\alpha_2} \\ x_j^{\alpha_1} & x_j^{\alpha_2} \end{bmatrix},$$

where $i, j \in \{1, 2, 3\}$, $i < j$, $\alpha_1, \alpha_2 \in \{0, 1, 2\}$, and $\alpha_1 < \alpha_2$. Since $0 < x_i < x_j$, it follows that $\det A > 0$. Hence, in this case, A is TP. For an arbitrary k-by-k submatrix of V, consider the following argument, which is similar to that given in [GK60] and in [Fal01]. Any k-by-k submatrix of $V(x_1, \ldots, x_n)$ is of the form $A = [x_{l_i}^{\alpha_j}]$, where $l_1, \ldots, l_k \in \{1, \ldots, n\}$, $l_1 < l_2 < \cdots < l_k$, $\alpha_1, \ldots, \alpha_k \in \{0, 1, \ldots, n-1\}$, and $\alpha_1 < \alpha_2 < \cdots < \alpha_k$. Let $f(x) = c_1 x^{\alpha_1} + c_2 x^{\alpha_2} + \cdots + c_k x^{\alpha_k}$ be a real polynomial. Descartes's rule of signs ensures that the number of positive real roots of $f(x)$ does not exceed the number of sign changes among the coefficients $c_1, \ldots, c_k$, so $f(x)$ has at most $k-1$ positive real roots. Therefore, the system of equations

$$f(x_{l_i}) = \sum_{j=1}^{k} c_j x_{l_i}^{\alpha_j} = 0, \ i = 1, 2, \ldots, k$$

has only the trivial solution $c_1 = c_2 = \cdots = c_k = 0$, so $\det A = \det[x_{l_i}^{\alpha_j}] \neq 0$.

To establish that $\det A > 0$, we use induction on the size of A. We know this to be true when the size of A is 2, so assume $\det A > 0$ whenever the size of A is at most $k-1$. Let $A = [x_{l_i}^{\alpha_j}]$ be k-by-k, in which $l_1, \ldots, l_k \in \{1, \ldots, n\}$, $l_1 < l_2 < \cdots < l_k$, $\alpha_1, \ldots, \alpha_k \in \{0, 1, \ldots, n-1\}$, and $\alpha_1 < \alpha_2 < \cdots < \alpha_k$. Then expanding $\det A$ along the last row gives

$$\det A = g(x_{l_k}) = a_k x_{l_k}^{\alpha_k} - a_{k-1} x_{l_k}^{\alpha_{k-1}} + \cdots + (-1)^{k-1} a_1 x_{l_k}^{\alpha_1}.$$

But $a_k = \det[x_{l_s}^{\alpha_t}]$, where $s, t \in \{1, 2, \ldots, k-1\}$, so the induction hypothesis ensures that $a_k > 0$. Thus $0 \neq g(x_{l_k}) \to \infty$ as $x_{l_k} \to \infty$, so $\det A = g(x_{l_k}) > 0$. The conclusion is that $V(x_1, \ldots, x_n)$ is TP whenever $0 < x_1 < \cdots < x_n$.

Example 0.1.5 (Cauchy matrix) An n-by-n matrix $C = [c_{ij}]$ is called a *Cauchy matrix* if the entries of C are given by

$$c_{ij} = \frac{1}{x_i + y_j},$$

where $x_1, x_2, \ldots, x_n$ and $y_1, y_2, \ldots, y_n$ are two sequences of real numbers (chosen accordingly so that c_{ij} is well defined). A Cauchy matrix is TP whenever $0 < x_1 < x_2 < \cdots < x_n$ and $0 < y_1 < y_2 < \cdots < y_n$ (see [GK02, pp. 77–78]).

We observe that the above claim readily follows from the well-known (Cauchy) identity

$$\det C = \frac{\prod_{i<k}(x_i - x_k) \prod_{i<k}(y_i - y_k)}{\prod_{i,k}(x_i + y_k)}.$$

As a particular instance, consider the matrix $A_a = [1/(a^{i-j} + 1)]$, where $a \neq 1$. Then $A_a = B_a D$, where $D = \text{diag}(a, a^2, \cdots a^n)$, and $B_a = [1/(a^i + a^j)]$. It is now easy to see that both B_a and A_a are TP for all $a > 0$ and $a \neq 1$.

Example 0.1.6 (Pascal matrix) Consider the n-by-n matrix $P_n = [p_{ij}]$ whose first row and column entries are all equal to 1, and, for $2 \leq i, j \leq n$, define $p_{ij} = p_{i-1,j} + p_{i,j-1}$ (Pascal's identity). Then a 4-by-4 symmetric Pascal matrix is given by

$$P_4 = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & 4 \\ 1 & 3 & 6 & 10 \\ 1 & 4 & 10 & 20 \end{bmatrix}.$$

The fact that P_n is TP will follow from the existence of a bidiagonal factorization, which is explained in much more detail in Chapter 2.

Example 0.1.7 (Routh-Hurwitz matrix) Let $f(x) = \sum_{i=0}^{n} a_i x^i$ be an nth degree polynomial in x. The n-by-n *Routh-Hurwitz matrix* is defined to be

$$\text{RH} = \begin{bmatrix} a_1 & a_3 & a_5 & a_7 & \cdots & 0 & 0 \\ a_0 & a_2 & a_4 & a_6 & \cdots & 0 & 0 \\ 0 & a_1 & a_3 & a_5 & \cdots & 0 & 0 \\ 0 & a_0 & a_2 & a_4 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \cdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & a_{n-1} & 0 \\ 0 & 0 & 0 & 0 & \cdots & a_{n-2} & a_n \end{bmatrix}.$$

A specific example of a Routh-Hurwitz matrix for an arbitrary polynomial of degree six, $f(x) = \sum_{i=0}^{6} a_i x^i$, is given by

$$\text{RH} = \begin{bmatrix} a_1 & a_3 & a_5 & 0 & 0 & 0 \\ a_0 & a_2 & a_4 & a_6 & 0 & 0 \\ 0 & a_1 & a_3 & a_5 & 0 & 0 \\ 0 & a_0 & a_2 & a_4 & a_6 & 0 \\ 0 & 0 & a_1 & a_3 & a_5 & 0 \\ 0 & 0 & a_0 & a_2 & a_4 & a_6 \end{bmatrix}.$$

A polynomial $f(x)$ is *stable* if all the zeros of $f(x)$ have negative real parts. It is proved in [Asn70] that $f(x)$ is stable if and only if the Routh-Hurwitz matrix formed from f is TN.

Example 0.1.8 (More examples)

The matrix $A = [e^{-\sigma(\alpha_i - \beta_j)^2}]$ is TP whenever $\sigma > 0$, $0 < \alpha_1 < \alpha_2 < \cdots < \alpha_n$, and $0 < \beta_1 < \beta_2 < \cdots < \beta_n$. It is worth noting that A may also be viewed as a positive diagonal scaling of a generalized Vandermonde matrix given by $V = [e^{2\sigma\alpha_i\beta_j}]$, which is also TP.

If the main diagonal entries of a TP matrix A are all equal to one, then it is not difficult to observe a "drop-off" effect in the entries of A as you move away from the main diagonal. For example, if $i < j$, then

$$a_{ij} \leq a_{i,i+1}a_{i+1,i+2}\cdots a_{j-1,j}.$$

A similar inequality holds for $i > j$.

Thus a natural question to ask is how much of a drop-off is necessary in a positive matrix to ensure that the matrix is TP. An investigation along these lines was carried in [CC98]. They actually proved the following interesting fact. If $A = [a_{ij}]$ is an n-by-n matrix with positive entries that also satisfy

$$a_{ij}a_{i+1,j+1} \geq c_0 a_{i,j+1}a_{i+1,j},$$

where $c_0 \approx 4.07959562349$, then A is TP. The number c_0 is actually the unique real root of $x^3 - 5x^2 + 4x - 1$.

More recently, this result was refined by [KV06] where they prove that, if $A = [a_{ij}]$ is an n-by-n matrix with positive entries and satisfies $a_{ij}a_{i+1,j+1} \geq ca_{i,j+1}a_{i+1,j}$, with $c \geq 4\cos^2(\frac{\pi}{n+1})$, then A is TP.

These conditions are particularly appealing for both Hankel and Toeplitz matrices. Recall that a *Hankel matrix* is an $(n+1)$-by-$(n+1)$ matrix of the form

$$\begin{bmatrix} a_0 & a_1 & \cdots & a_n \\ a_1 & a_2 & \cdots & a_{n+1} \\ \vdots & \vdots & \cdots & \vdots \\ a_n & a_{n+1} & \cdots & a_{2n} \end{bmatrix}.$$

So if the positive sequence $\{a_i\}$ satisfies $a_{k-1}a_{k+1} \geq ca_k^2$, then the corresponding Hankel matrix is TP. An $(n+1)$-by-$(n+1)$ *Toeplitz matrix* is of the form

$$\begin{bmatrix} a_0 & a_1 & a_2 & \cdots & a_n \\ a_{-1} & a_0 & a_1 & \cdots & a_{n-1} \\ a_{-2} & a_{-1} & a_0 & \cdots & a_{n-2} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ a_{-n} & a_{-(n-1)} & a_{-(n-2)} & \cdots & a_0 \end{bmatrix}.$$

Hence if the positive sequence $\{a_i\}$ satisfies $a_k^2 \geq ca_{k-1}a_{k+1}$, then the corresponding Toeplitz matrix is TP.

A *sequence* $a_0, a_1, \ldots$ *of real numbers is called totally positive* if the two-way infinite matrix given by

$$
\begin{bmatrix}
a_0 & 0 & 0 & \cdots \\
a_1 & a_0 & 0 & \cdots \\
a_2 & a_1 & a_0 & \cdots \\
\vdots & \vdots & \vdots &
\end{bmatrix}
$$

is TP. An infinite matrix is TP if all its minors are positive. Notice that the above matrix is a Toeplitz matrix. Studying the functions that generate totally positive sequences was a difficult and important step in the area of TP matrices; $f(x)$ generates the sequence $a_0, a_1, \ldots$ if $f(x) = a_0 + a_1 x + a_2 x^2 + \cdots$. In [ASW52], and see also [Edr52], it was shown that the above two-way infinite Toeplitz matrix is TP (i.e., the corresponding sequence is totally positive) if and only if the generating function $f(x)$ for the sequence $a_0, a_1, \ldots$ has the form

$$
f(x) = e^{\gamma x} \frac{\prod_{\nu=1}^{\infty}(1 + \alpha_v x)}{\prod_{\nu=1}^{\infty}(1 - \beta_v x)},
$$

where $\gamma, \alpha_v, \beta_v \geq 0$, and $\sum \alpha_v$, and $\sum \beta_v$ are convergent.

0.2 APPLICATIONS AND MOTIVATION

Positivity has roots in every aspect of pure, applied, and numerical mathematics. The subdiscipline, total positivity, also is entrenched in nearly all facets of mathematics. Evidence of this claim can be found in the insightful and comprehensive proceedings [GM96]. This collection of papers was inspired by Karlin's contributions to total positivity and its applications. This compilation contains 23 papers based on a five-day meeting held in September 1994. The papers are organized by area of application. In particular, the specialties listed are (in order of appearance)

(1) Spline functions

(2) Matrix theory

(3) Geometric modeling

(4) Probability and mathematical biology

(5) Approximation theory

(6) Complex analysis

(7) Statistics

(8) Real analysis

(9) Combinatorics

(10) Integral equations

The above list is by no means comprehensive, but it is certainly extensive and interesting (other related areas include differential equations, geometry, and function theory). A recent application includes new advances in accurate eigenvalue computation and connections with singular value decompositions on TN matrices (see, for example, [KD02, Koe05, Koe07]).

Historically speaking, total positivity came about in essentially three forms:

(1) Gantmacher/Krein—oscillations of vibrating systems

(2) Schoenberg—real zeros of a polynomials, spline function with applications to mathematical analysis, and approximation

(3) Karlin—integral equations, kernels, and statistics.

All these have led to the development of the theory of TP matrices by way of rigorous mathematical analysis.

We will elaborate on the related applications by choosing two specific examples to highlight.

(1) Statistics/probability

(2) CAGD/spline functions

On our way to discussing these specific applications, which in many ways represent a motivation for exploring TP matrices, we begin with a classic definition, which may have been the impetus for both Karlin and Schoenberg's interest in TP matrices.

A real function (or kernel) $k(x, y)$ in two variables, along with two linearly ordered sets X, Y is said to be *totally positive of order n* if for every

$$x_1 < x_2 < \cdots < x_n; \quad y_1 < y_2 < \cdots < y_n, \quad x_i \in X, \quad y_j \in Y,$$

we have

$$\det \begin{bmatrix} k(x_1, y_1) & \cdots & k(x_1, y_n) \\ \vdots & \ddots & \vdots \\ k(x_n, y_1) & \cdots & k(x_n, y_n) \end{bmatrix} > 0.$$

We use the term *TN for a function $k(x, y)$* if the inequalities above are not necessarily strict.

There are many examples of totally positive functions. For instance,

(1) $k(x, y) = e^{xy}$ (see [Kar68, p. 15]), $x, y \in (-\infty, \infty)$,

(2) $k(x, y) = x^y$, $x \in (0, \infty)$, $y \in (-\infty, \infty)$

are both examples (over their respective domains) of totally positive functions.

Upon closer examination of the function $k(x, y)$ in (2), we observe that for a fixed n and given $0 < x_1 < x_2 < \cdots < x_n$ and $y_1 < y_2 < \cdots < y_n$,

the n-by-n matrix $(k(x_i, y_j))_{i,j=1}$ is a generalized Vandermonde matrix. For instance, if $n = 4$, then the matrix in question is given by

$$
\begin{bmatrix}
x_1^{y_1} & x_1^{y_2} & x_1^{y_3} & x_1^{y_4} \\
x_2^{y_1} & x_2^{y_2} & x_2^{y_3} & x_2^{y_4} \\
x_3^{y_1} & x_3^{y_2} & x_3^{y_3} & x_3^{y_4} \\
x_4^{y_1} & x_4^{y_2} & x_4^{y_3} & x_4^{y_4}
\end{bmatrix}.
$$

An important instance is when a totally positive function of order n, say $k(x, y)$, can be written as $k(x, y) = f(x - y)$ $(x, y \in \mathbb{R})$. In this case, the function $f(u)$ is called a *Pólya frequency function of order n*, and is often abbreviated as PF_n (see [Kar68], for example). Further along these lines, if $k(x, y) = f(x - y)$ but $x, y \in \mathbb{Z}$, then $f(u)$ is said to be a *Pólya frequency sequence of order n*.

As an example, we mention a specific application of totally positive functions of order 2 (and PF_2) to statistics.

Observe that a function $h(u)$, $u \in \mathbb{R}$ is PF_2 if:

(1) $h(u) > 0$ for $-\infty < u < \infty$, and

(2) $\det \begin{pmatrix} h(x_1 - y_1) & h(x_1 - y_2) \\ h(x_2 - y_1) & h(x_2 - y_2) \end{pmatrix} > 0$
for all $-\infty < x_1 < x_2 < \infty$, $-\infty < y_1 < y_2 < \infty$.

In the book [BP75, p. 76], it is noted that conditions (1), (2) on $h(u)$ above are equivalent to either

(3) $\log h(u)$ is concave on $\mathbb{R}$, or

(4) for fixed $\triangle > 0$, $\dfrac{h(u + \triangle)}{h(u)}$ is decreasing in u for $a \leq u \leq b$ when

$$
a = \inf_{h(u)>0} y, \qquad b = \sup_{h(u)>0} u.
$$

Recall that for a continuous random variable u, with probability density function $f(u)$, its cumulative distribution function is defined to be $F(u) = \int_{-\infty}^{u} f(t)dt$. Then the reliability function, $\overline{F}$, is given by $\overline{F}(u) = 1 - F(u)$. A distribution function $F(u)$ is called an *increasing failure rate distribution* if the function

$$
\overline{F}(x|t) = \frac{\overline{F}(t + x)}{\overline{F}(t)}
$$

is decreasing in $t \in \mathbb{R}$ for each $x \geq 0$. It turns out that the condition $\overline{F}(x|t)$ decreasing is equivalent to the failure rate function

$$
r(t) = \frac{f(t)}{\overline{F}(t)}
$$

being an increasing function.

Hence we have that a distribution function $F(u)$ is an increasing failure rate distribution if and only if $\overline{F} = 1 - F$ is PF_2.

Return to totally positive functions. Suppose k is a function of $X \times Y$; then we will consider the derived "determinant function." Define the set (following [Kar68])

$$\triangle_p(X) = \{\overline{x} = (x_1, x_2, \ldots, x_p) | \ x_1 < \cdots < x_p, \ x_i \in X\}$$

(ordered p-tuples).

The determinant function

$$K_{[p]}(\overline{x}, \overline{y}) = \det \begin{bmatrix} k(x_1, y_1) & \cdots & k(x_1, y_n) \\ \vdots & \ddots & \vdots \\ k(x_n, y_1) & \cdots & k(x_n, y_n) \end{bmatrix}$$

defined on $\triangle_p(X) \times \triangle(Y)$ is called the *compound kernel of order p induced* by $k(x, y)$. If $k(x, y)$ is a totally positive function of order n, then $K_{[p]}(\overline{x}, \overline{y})$ is a nonnegative function on $\triangle_p(X) \times \triangle_p(Y)$ for each $p = 1, \ldots, n$.

As an example, consider the (indicator) function $k(x, y)$ defined

$$k(x, y) = \begin{cases} 1, & 1 \le x \le y \le b, \\ 0, & a \le y < x \le b. \end{cases}$$

Then k is a totally nonnegative function, and for arbitrary

$$1 \le x_1 < x_2 < \cdots < x_n \le b,$$
$$a \le y_1 < y_2 < \cdots < y_n \le b$$

we have

$$K_p[\overline{x}, \overline{y}] = \begin{cases} 1 & \text{if} & x_1 \le y_1 \le x_2 \le y_2 \le \cdots \le x_n \le y_n, \\ 0 & & \text{otherwise.} \end{cases}$$

The proof is straightforward and is omitted.

Another example class of totally positive functions is the exponential family of (statistical) distributions. Consider the function

$$k(x, y) = \beta(y)e^{xy}$$

with respect to sigma finite measure $d\mu(x)$. Examples of such distributions include the normal distribution with variance 1, exponential distributions, and the gamma distribution.

There is another connection between totally positive matrices and correlation matrices.

Following the paper [SS06], and the references therein, it is known by empirical analysis that correlation matrices of forward interest rates (which have applications in risk analysis) seem to exhibit spectral properties that are similar to TP matrices. In particular, it is shown that the correlation matrix R of yields can be approximately described by the correlation function

$$\rho_{ij} = \exp(-\beta|t_j - t_i|), \beta > 0.$$

If indices with maturities in the above model are identified, then $\rho_{ij} = \rho^{|i-j|}$, where $\rho = \exp(-\beta)$.

Then

$$
R = \begin{bmatrix}
1 & \rho & \rho^2 & \cdots & \rho^{n-1} \\
\rho & \ddots & \ddots & \ddots & \vdots \\
\rho^2 & \ddots & \ddots & \ddots & \rho^2 \\
\vdots & \ddots & \ddots & \ddots & \rho \\
\rho^{n-1} & \cdots & \rho^2 & \rho & 1
\end{bmatrix}
\quad \text{(special Toeplitz matrix)}.
$$

It is well known that R above is oscillatory. For example, appealing to an old result in [Edr52] and referring to [Kar68] (making use of Polyá frequency sequences) to demonstrate that R is TN, before establishing that R is OSC.

Without going into details here, we note that R is essentially (up to re-signing) the inverse of the IITN tridiagonal matrix

$$
T = \frac{1}{(1 - \rho^2)} \begin{bmatrix}
1 & \rho & 0 & \cdots & 0 \\
\rho & 1+\rho^2 & \rho & \ddots & 0 \\
0 & \rho & \ddots & \ddots & 0 \\
\vdots & \ddots & \ddots & 1+\rho^2 & \rho \\
0 & \cdots & 0 & \rho & 1
\end{bmatrix}.
$$

Observe that T satisfies the row dominance conditions mentioned in the previous section of this chapter. Consequently T is OSC, and hence R is OSC.

Further applications to risk analysis overlap with the theory of TP matrices.

In [Jew89], where choices between risky prospects are investigated, it is cited under what conditions it is true that

$$
E(u(Y + X)) = u(Y + C) \Rightarrow E(v(Y + X)) \leq E(v(Y + C)),
$$

where
X, Y—risky prospects,
u, v—increasing utilities with v more risk averse than u?

Clearly the answer will depend on what is assumed about the joint distribution of X and Y.

Jewitt proves (along with a slight generalization) that the above condition holds whenever Y is independent of X and has a log-concave density. Recall from the previous discussion that log-concavity is intimately related to the density being a totally positive function of order 2 ($k(x, y) = h(x - y)$ a totally positive function of order 2 is equivalent to being log-concave).

The theory of Pólya frequency functions (PF) is quite immeasurable and its history spans 80 years. Further, their connection to TP matrices is very strong and has produced numerous important results and associated applications.

A *Pólya frequency function* of order k (PF_k) is a function f in a one real variable u, $u \in \mathbb{R}$ for which $k(x, y) = f(x - y)$ is a totally nonnegative

function of order k for $x, y \in \mathbb{R}$. Schoenberg was one of the pioneers in the study of Pólya frequency functions.

A basic example of a PF function is

$$f(u) = e^{-\gamma u^2}, \quad \gamma > 0.$$

To verify, observe that

$$f(x - y) = e^{-\gamma(x-y)^2} = e^{-\gamma x^2} e^{-\gamma y^2} e^{2\gamma xy}.$$

Since e^{xy} is a totally nonnegative function, and the other factors are totally positive functions, their product above is a totally positive function.

If the argument of a PF_k function is restricted to the integers, we call the resulting sequence a *Pólya frequency sequence*.

For brevity, we discuss the case of "one-sided" Pólya frequency sequences: $k(m, n) = a_{n-m}$, generates a sequence $\{a_\ell\}$, with the extra stipulation that $a_\ell = 0$ for $\ell < 0$. A (one-sided) sequence is said to be a PF_∞ sequence if the corresponding Kernel written as an infinite matrix

$$A = \begin{bmatrix} a_0 & a_1 & \cdots & \cdots \\ 0 & a_0 & a_1 & \cdots \\ \vdots & 0 & a_0 & a_1 \\ \vdots & \ddots & \ddots & \ddots \end{bmatrix}$$

is TN, and similarly $\{a_\ell\}$ is PF_k if A is TN$_k$. Observe that if $a_0 > 0$ and A is TP$_2$, then the sequence $\{a_\ell\}$ is either finite or consists entirely of positive numbers.

To a given sequence $\{a_\ell\}$, we associate the usual generating function $f(z) = \sum_{\ell=0}^{\infty} a_\ell z^\ell$. Further, if $\{a_n\}$ is a PF_2 sequence, then $f(z)$ has a nonzero radius of convergence. Some examples of PF_∞ sequences are

(1) $f(z) = 1 + \alpha z, \quad \alpha > 0,$

(2) $f(z) = \dfrac{1}{1 - \beta z}, \quad \beta > 0.$

Observe that the (discrete) kernel corresponding to (2) has the form

$$A = \begin{pmatrix} 1 & \beta & \beta^2 & \cdots & \cdots \\ 0 & 1 & \beta & \beta^2 & \cdots \\ \vdots & 0 & 1 & \beta & \ddots \\ \vdots & \vdots & \ddots & \ddots & \ddots \\ 0 & 0 & \cdots & 0 & 1 \end{pmatrix}.$$

It can be easily verified that A is TN and that the radius of convergence of $f(z) = (1 - \beta z)^{-1}$ is given by $|z| < 1/\beta$.

There is an "algebra" associated with PF_r sequences. If $f(z)$ and $g(z)$ generate one-sided PF_r sequences, then so do

$$f(z) \cdot g(z) \quad \text{and} \quad \frac{1}{f(-z)}.$$

Hence combining examples (1) and (2) above, we arrive at the function

$$\frac{\prod_{i=1}^{n}(1 + \alpha_i z)}{\prod_{j=1}^{n}(1 - \beta_j z)},$$

with $\alpha_i > 0$ and $\beta_j > 0$, which generates a one-sided PF sequence. In fact, it can be shown that $e^{\gamma z}$, $\gamma > 0$ generates a PF sequence. Hence the function

$$f(z) = e^{\gamma z} \frac{\prod_{i=1}^{\infty}(1 + \alpha_i z)}{\prod_{j=1}^{\infty}(1 - \beta_j z)},$$

generates a one-sided PF sequence provided $\alpha_i \geq 0$, $\beta_j \geq 0$ and

$$\sum \alpha_i < \infty, \quad \sum \beta_j < \infty.$$

A crowning achievement along these lines is a representation of one-sided PF sequences in 1952 [Edr52]. A function $f(z) = \sum a_n z^n$ with $f(0) = 1$ generates a one-sided PF sequence if and only if it has the form

$$f(z) = e^{\gamma z} \frac{\prod_{i=0}^{\infty}(1 + \alpha_1 z)}{\prod_{j=0}^{\infty}(1 - \beta_j z)},$$

where $\gamma \geq 0$, $\alpha_i \geq 0$, $\beta_j \geq 0$, and $\sum(\alpha_i + \beta_i) < \infty$.

It is useful to note that Whitney's reduction theorem for totally positive matrices [Whi52] was a key component to proving one direction of the aforementioned result that appeared in [ASW52].

Edrei later considered doubly infinite sequences and their associated generating functions; see [Edr53a] and [Edr53b]. A doubly infinite sequence of real numbers $\{a_n\}_{-\infty}^{\infty}$ is said to be *totally positive* if for every k and sets of integers $s_1 < s_2 < \cdots < s_k$ and $t_1 < t_2 < \cdots < t_k$, the determinant of $C = [a_{s_i - t_j}]$ for $i, j = 1, 2, \ldots, k$ is nonnegative. We also note that some of Edrei's work on scalar Toeplitz matrices has been extended to block Toeplitz matrices (see, for example, [DMS86]), including a connection via matrix factorization.

We close this section on applications with a discussion on totally positive systems of functions. We incorporate two example applications within such systems of functions:

(1) spline functions,

(2) TP bases.

Given a system of real-valued functions $u = (u_0, u_1, \ldots, u_n)$ defined on $I \subseteq \mathbb{R}$, the *collocation matrix of u at* $t_0 < t_1, < \cdots < t_m$ in I is defined as the $(m+1)$-by-$(n+1)$ matrix

$$M = \begin{pmatrix} u_0(t_0) & u_1(t_0) & \cdots & u_n(t_0) \\ u_0(t_1) & u_1(t_1) & \cdots & u_n(t_1) \\ \vdots & \vdots & \ddots & \vdots \\ u_0(t_m) & u_1(t_m) & \cdots & u_n(t_m) \end{pmatrix}.$$

We say that a *system of functions u is totally positive* if all its collocation matrices are TP. The most basic system of functions that are totally positive is of course $1, x, x^2, \ldots, x^n$ when $I = (0, \infty)$. In this case the associated collocation matrices are just Vandermonde matrices.

We say that $(u_0, u_1, \ldots, u_n)$ is a normalized totally positive system if

$$\sum_{i=0}^{n} u_i(x) = 1 \text{ for } x \in I.$$

If, in addition, the system $(u_0, u_1, \ldots, u_n)$ is linearly independent, then we call $(u_0, u_1, \ldots, u_n)$ a *totally positive basis* for the space $\mathrm{Span}\{u_0, \ldots, u_n\}$. For example, $\{1, x, x^2, \ldots, x^n\}$ is a totally positive basis for $\mathbb{P}_n$, the space of polynomials of degree at most n.

Let $u = (u_0, \ldots, u_n)$ be a TP basis defined on I, and let $U = \mathrm{span}(u_0, \ldots, u_n)$. It follows that for any TP matrix A the basis v defined by

$$v^T = u^T A$$

is again a TP basis. The converse, which is an extremely important property of TP bases, is known to hold, namely, that all TP bases are generated in this manner. The key of course is choosing the appropriate starting basis. A TP basis u defined on I is called a *B-basis* if all TP bases v of U satisfy $v^T = A^T u$ for some invertible TN matrix A.

Some examples of B-bases include

(1) the Bernstein basis defined on $[a, b]$

$$b_i(t) = \frac{\binom{n}{i}}{(b-a)^n}(t-a)^i(b-t)^{n-i};$$

(2) the B-spline basis of the space of polynomial splines on a given interval with a prescribed sequence of knots.

TP bases and normalized TP bases have a natural geometric meaning which leads to representations with optimal shape-preserving properties.

Given a sequence of positive functions $(u_0, \ldots, u_n)$ on $[a, b]$ with $\sum u = 1$ and a sequence $(A_0, A_1, \ldots, A_n)$ of points in $\mathbb{R}^k$, we define the curve

$$\gamma(t) = \sum_{i=0}^{n} c_0 \cdot u_i(t), \quad t \in [a, b].$$

The functions $u_0, \ldots, u_n$ are typically called the *blending functions* and the points on $A_0, A_1, \ldots, A_n$ the *control points*. Let $P(A_0, \ldots, A_n)$ denote the polygonal arc with vertices $A_o, \ldots, A_n$. Often $P(A_0, \ldots, A_n)$ is called the *control polygon of the curve γ.*

If $u_0, \ldots, u_n$ is a normalized TP basis, then the curve γ preserves many of the shape properties of the control polygon. For example, the variation-diminution properties of TP matrices implies that the monotonicity and convexity of the control polygon are inherited by the curve γ. It is for these reasons that TP and B-bases are important in computer-aided geometric design.

As mentioned in example (2), the so-called B-spline's are deeply connected with certain aspects of total positivity. We describe briefly the concepts of spline functions and B-splines and their relationships to TP matrices.

Let $\triangle = \{x_i, i = 1, \ldots, r$ with $x_1 < x_2 < x_3 < \cdots < x_r\}$, $k > 0$, and let $m = (m_1, m_2, \ldots, m_r)$ be a sequence of integers with $1 \leq m_i \leq k+1$ for each i.

Suppose $[a, b]$ is a given interval with $x_i \in [a, b]$, and define $x_0 = a$, $x_{r+1} = b$. A piecewise polynomial function f of degree k on each of the intervals $[x_i, x_u)$ for $i \leq r-1$ and on $[x_r, x_{r+1}]$, and continuous of order $k - m_i$ (the function and the associated derivatives) at the knot x_i is called a *spline of degree k with knots x_i of multiplicity m_i.*

Spline functions have a rich and deep history and have applications in many areas of mathematics, including in approximation theory.

One important aspect to the theory for splines is determining an appropriate basis for the space spanned by the splines described above. B-splines turn out to be a natural such basis. To describe B-splines we need to introduce a finer partition based on the knots $x_1 < \cdots < x_r$ above and on $[a, b]$ (here we assume $[a, b]$ is a finite interval for simplicity). Define the new knots

$$y_1 \leq y_2 \leq \cdots < y_{k+1} < a,$$
$$y_{k+1} = y_{k+3} = \cdots = y_{k+m_1+1} = x_1,$$
$$y_{k+m_1+2} = \cdots = y_{k+m_1+m_2+1} = x_2,$$
$$\vdots$$
$$y_{k+n-m_r+2} = y_{k+n-m_r+3} = \cdots = y_{k+n+1} = x_r,$$
$$(n = \textstyle\sum m_i)$$
$$b \leq y_{k+n+2} \leq \cdots \leq y_{n+2k+2}.$$

Observe that $y_{i+k+1} > y_i$ for all i. Given this finer partition we define the functions

$$B_i(x) = (y_{i+k+1} - y_i)[y_i \cdots y_{i+k+1}](y - x)_t^k,$$

$$a \leq x \leq b, \qquad 1 \leq i \leq n+k+1,$$

when $u_t = \max\{0, u\}$ and $[y_i, \ldots, y_{i+k+1}]f(x, y)$ denotes the $(k+1)$st divided difference of f with respect to the variable y with arguments $y_i, \ldots, y_{i+k+1}$.

The functions $B_i(x)$ are called B-splines and form a basis for the space spanned (with compact support) all spline of degree k with knots x_i of multiplicity m_i.

Consider strictly increasing real numbers $y_1 < \cdots < y_{n+k+1}$, and let $B_i(x)$ be as defined above, with $[y_i, y_{i+k+1}]$ as the support of B_i.

For distinct real numbers $t_1 < t_2 < \cdots < t_n$, we let

$$M = \begin{pmatrix} B_1(t_1) & \cdots & B_n(t_1) \\ B_1(t_2) & \cdots & B_n(t_2) \\ \vdots & \vdots & \vdots \\ B_1(t_n) & \cdots & B_n(t_n) \end{pmatrix}$$

be the corresponding collocation matrix.

Schoenberg/Whitney [SW 53] proved that M is invertible if and only if $y_i < t_i < y_{i+k+1}$ for $i = 1, \ldots, n$.

In fact, much more is true about the matrix M when $y_2 < t_i < y_{i+k+1} \forall_i$. M is a TN matrix, as demonstrated by Karlin in [Kar68]. In [dB76] it was verified that all minors of M are nonnegative and a particular minor is positive (when $y_i < t_i < y_{i+k+1}$) if and only if all main diagonal entries of that minor are positive. Such matrices are also known as *almost totally positive matrices*.

0.3 ORGANIZATION AND PARTICULARITIES

This book is divided into 11 chapters including this introductory chapter on TN matrices. The next chapter, Chapter 1, carefully spells out a number of basic and fundamental properties of TN matrices along with a compilation of facts and results from core matrix theory that are useful for further development of this topic. For our purposes, as stated earlier, a matrix will be considered over the real field, unless otherwise stated. Further, it may be the case that some of the results will continue to hold over more general fields, although they will only be stated for matrices over the reals.

Chapter 2 introduces and methodically develops the important and useful topic of bidiagonal factorization. In addition to a detailed description regarding the existence of such factorizations for TN matrices, we also discuss a number of natural consequences and explain the rather important associated combinatorial objects known as planar diagrams.

The next four chapters, Chapters 3–6, highlight the fundamental topics: recognition of TN matrices (Chapter 3); sign variation of vectors (Chapter 4); spectral properties (Chapter 5); and determinantal inequalities (Chapter 6).

The remaining four chapters cover a wide range of topics associated with TN matrices that are of both recent and historical interest. Chapter 7 contains a detailed account of the distribution of rank deficient submatrices within a TN matrix, including a discussion of the important notion of row and column inclusion. Chapter 8 introduces the Hadamard (or entrywise)

product and offers a survey of the known results pertaining to Hadamard products of TN matrices, including a note on Hadamard powers. In Chapter 9, we explore the relatively modern idea of matrix completion problems for TN matrices. This chapter also includes a section on single entry perturbations, known as retractions, which turn out to be a useful tool for other problems on TN matrices. In Chapter 10 we conclude with a brief review of a number of subtopics connected with TN matrices, including powers and roots of TN matrices, TP/TN polynomial matrices, subdirect sums of TN matrices, and Perron complements of TN matrices.

For the reader we have included a brief introductory section with each chapter, along with a detailed account of the necessary terms and notation used within that chapter. We have aimed at developing, in detail, the theory of totally nonnegative matrices, and as such this book is sequential in nature. Hence some results rely on notions in earlier chapters, and are indicated accordingly by references or citations as necessary. On a few occasions some forward references have also been used (e.g., in section 2.6 a reference is made to a result to appear in Chapter 3). We have tried to keep such instances of forward references to a minimum for the benefit of the reader.

This work is largely self-contained with some minor exceptions (e.g., characterization of sign-regular transformations in Chapter 4). When a relevant proof has been omitted, a corresponding citation has been provided. In addition, some of the proofs have been extracted in some way from existing proofs in the literature, and in this case the corresponding citation accompanies the proof.

We have also included, at the end, a reasonably complete list of references on all facets of total nonnegativity. Finally, for the benefit of the reader, an index and glossary of symbols (nomenclature) have also been assembled.

Index