國立交通大學理學院應用數學系

博士論文

某些矩陣的高-吳數

Gau-Wu numbers of certain matrices

研究生: 李信儀
Student: Hsin-Yi Lee
指導教授: 吳培元 教授
Advisor: Pei Yuan Wu

中華民國一百零三年六月
June, 2014

某些矩陣的高-吳數
Gau-Wu numbers of certain matrices

研 究 生：李信儀 Student: Hsin-Yi Lee
指導教授：吳培元 教授 Advisor: Dr. Pei Yuan Wu

國 立 交 通 大 學
應 用 數 學 系
博 士 論 文

A Thesis
Submitted to Department of Applied Mathematics
College of Science
National Chiao Tung University
in partial Fulfillment of the Requirements
for the Degree of
Doctor of Philosophy
in

Applied Mathematics

June 2014

Hsinchu, Taiwan, Republic of China

中華民國一百零三年六月
June, 2014

某些矩陣的高-吳數

研究生: 李信儀　　　　　　　　　　　　　　　　　　　　指導教授: 吳培元 教授

國立交通大學
應用數學系 博士班

摘要

對於一個 n 乘 n 的矩陣 A，令 $k(A)$ 表示數值域邊界上的點 $\langle Ax_j, x_j\rangle$ 所對應的正交單位向量 x_j 的最大個數。我們稱這個數 $k(A)$ 為高-吳數。若 A 為正規或二次矩陣，則其高-吳數 $k(A)$ 可以明確地被計算出來。而對於一個矩陣 A 形如 $B \oplus C$，我們證明了高-吳數為 2 時，其充分且必要條件為其中一個矩陣，稱之為 B，的數值域，完全落在另外一個矩陣 C 的數值域的內部且 $k(C)$ 為 2。對於一個不可約的矩陣 A，我們可以確切地決定何時其高-吳數等於 n。這些結果以及已知的 4 乘 4 矩陣的數值域的圖形，可用以決定任何一個 4 乘 4 可約矩陣的高-吳數。

此外，設 A 為一個 n 乘 n (n 大於或等於 2) 的非負矩陣，其形如下

$$\begin{bmatrix} 0 & A_1 & & 0 \\ & 0 & \ddots & \\ & & \ddots & A_{m-1} \\ 0 & & & 0 \end{bmatrix},$$

此處 m 大於或等於 3 並且對角線上所出現的零均為零方陣。若 A 的實部為不可約的矩陣，則其高-吳數 $k(A)$ 的上界為 m-1。再者，我們也得到這種矩陣的高-吳數達到其上界的充分且必要條件。除此之外，我們也研究了另外一類型的非負矩陣，稱之為雙隨機矩陣。我們證明了任何一個 3 乘 3 的雙隨機矩陣的高吳數必定為 3。另外，我們也決定了 4 乘 4 的雙隨機矩陣的數值域及其高-吳數。最後我們也考慮一般的 n 乘 n (n 大於或等於 5) 雙隨機矩陣，藉由其可能的數值域的圖形得到其高-吳數的下界。

Gau-Wu numbers of certain matrices

Student: Hsin-Yi Lee Advisor: Dr. Pei Yuan Wu

Department of Mathematics
National Chiao Tung University

ABSTRACT

For any n-by-n matrix A, let $k(A)$ stand for the maximal number of orthonormal vectors x_j such that the scalar products $\langle Ax_j, x_j\rangle$ lie in the boundary of the numerical range $W(A)$. This number $k(A)$ is called the Gau-Wu number of the matrix A. If A is a normal or a quadratic matrix, then the exact value of $k(A)$ can be computed. For a matrix A of the form $B \oplus C$, we show that $k(A) = 2$ if and only if the numerical range of one summand, say, B, is contained in the interior of the numerical range of the other summand C and $k(C) = 2$. For an irreducible matrix A, we can determine exactly when the value of $k(A)$ equals the size of A. These are then applied to determine $k(A)$ for a reducible matrix A of size 4 in terms of the shape of $W(A)$.

Moreover, if A is an n-by-n ($n \geq 2$) nonnegative matrix of the form

$$\begin{bmatrix} 0 & A_1 & & 0 \\ & 0 & \ddots & \\ & & \ddots & A_{m-1} \\ 0 & & & 0 \end{bmatrix},$$

where $m \geq 3$ and the diagonal zeros are zero square matrices, with irreducible real part, then $k(A)$ has an upper bound $m-1$. In addition, we also obtain necessary and sufficient conditions for $k(A) = m-1$ for such a matrix A. The other class of nonnegative matrices we study is the doubly stochastic ones. We prove that the value of $k(A)$ is equal to 3 for any 3-by-3 doubly stochastic matrix A. Next, for any 4-by-4 doubly stochastic matrix, we also determine its numerical range. This result can be applied to find the value of $k(A)$ for any doubly stochastic matrix A of size 4 in terms of the shape of $W(A)$. Furthermore, the lower bound of $k(A)$ is also found for a general n-by-n ($n \geq 5$) doubly stochastic matrix A via the possible shapes of $W(A)$.

誌　謝

本論文得以完成，首先要特別感謝我的指導教授　吳培元老師在這五年多來，無論是學業上或研究上都不斷地給予各種幫助。除了學術上的教導，我更從他的身上學習到做人處事的道理以及態度。非常感謝吳老師所教授的一切。另外要感謝口試委員黃毅青教授、簡茂丁教授、高華隆教授與王國仲教授的指點，使得本論文更臻完善。

在博士班求學的生涯中，感謝許多交大應用數學系教授的指導與助理們的幫助，如王國仲教授對於數值域中所探討的高-吳數所給予的寶貴建議與意見，讓我能看到自己的盲點，進而解決某一類矩陣的問題。其他還有林文偉講座教授、李明佳教授、王夏聲教授、林琦焜教授、白啓光教授、葉立明教授、許義容教授、張書銘教授等等，在課業與其他方面給我的協助。另外要感謝中央大學高華隆教授，當我在處理數值域中的問題時，他所給予的協助與鼓勵，讓我對這領域的看法有了全新的體驗與認識。

沒有師門的學長們的幫忙，很難度過這漫長的博士生涯。感謝逢甲大學張其棟教授教導我學會繪圖軟體的使用。更感謝中山大學蔡明誠博士在數不清多少個夜晚同我視訊。一方面檢視我的證明，另一方面給予不斷地鼓勵。使我體悟了這句名言『問松林，松林幾經冬？山川何如昔，風雲與古同。』另外要感謝系上的學長、同學與學弟們，包括高雄師範大學吳恭儉教授、呂明杰博士、黃韋強博士、黃皜文博士、李忠逵博士、陳德軒、龔柏任、黃俊銘、陳哲楷等等。還有其他透過網路認識的朋友們（族繁不及備載），非常感謝這一路上有他們的陪伴，使得我的研究生涯多采多姿。

最後要特別感謝家人，包括我的雙親、愛妻文鳳、姊妹們以及愛女采潔。感謝他們始終不斷地給我支持以及無私的付出與關懷，讓我能夠專心地在研究工作上持續邁進。尤其是愛女還年幼，雙親與妻子對女兒地細心照料，讓我不必為生活瑣碎以及孩子的事情而操心，進而能順利地取得博士學位，謹以此論文獻給你們。

Contents

1 Introduction

Let A be an n-by-n complex matrix. Its *numerical range* $W(A)$ is, by definition, the set $\{\langle Ax, x\rangle\colon\ x \in \mathbb{C}^n,\ \|x\| = 1\}$, where $\langle\cdot,\cdot\rangle$ and $\|\cdot\|$ denote the standard inner product and its associated norm in $\mathbb{C}^n$, respectively. One of the most important properties of the numerical range is its convexity. In fact, the study of the numerical range originates from the discovery of this property by Toeplitz [17] and Hausdorff [7]: the former proved that the boundary of the numerical range is always a convex curve, but left open the possibility that it may have interior holes while the latter, using a different approach, showed that this cannot happen. An interesting account on the history of this theorem can be found in [6].

For a matrix A, let A^* denote its adjoint, Re A its real part $(A + A^*)/2$ and Im A its imaginary part $(A - A^*)/2i$. The set of eigenvalues of A is denoted by $\sigma(A)$. For any subset $\triangle$ of $\mathbb{C}$, $\triangle^\wedge$ denotes its convex hull, that is, $\triangle^\wedge$ is the smallest convex set containing $\triangle$. We list below several important properties of the numerical range.

(1) $W(U^*AU) = W(A)$ for any unitary matrix U.

(2) $W(A)$ is a compact subset of $\mathbb{C}$.

(3) $W(aA + bI) = aW(A) + b$ for any scalars a and b.

(4) $W(\text{Re } A) = \text{Re } W(A)$ and $W(\text{Im } A) = \text{Im } W(A)$.

(5) If $A = \begin{bmatrix} B & * \\ * & * \end{bmatrix}$, then $W(B) \subseteq W(A)$.

(6) $\sigma(A) \subseteq W(A)$.

(7) If A is normal, then $W(A)$ is equal to $\sigma(A)^\wedge$.

(8) $W(\sum\limits_n \oplus A_n) = (\cup_n W(A_n))^\wedge$.

For other properties of the numerical range, the reader may consult [8, Chapter 1].

In Chapter 2, we consider the maximum number $k = k(A)$ for which there exist orthonormal vectors $x_1, ..., x_k \in \mathbb{C}^n$ with $\langle Ax_j, x_j\rangle$ in the boundary $\partial W(A)$ of $W(A)$

for all j. Note that $k(A)$ is also the maximum size of a compression of A with all its diagonal entries in $\partial W(A)$. Recall that a k-by-k matrix B is a *compression* of A if $B = V^*AV$ for some n-by-k matrix V with $V^*V = I_k$. Here I_k denotes the k-by-k identity matrix. In particular, if n equals k, then A and B are said to be *unitarily similar*, which we denote by $A \cong B$. The number $k(A)$ was introduced in [5] and [19] and is called the Gau-Wu number by [2]. It relates properties of the numerical range to the compressions of A. In particular, it was shown in [5, Lemma 4.1 and Theorem 4.4] that $2 \leq k(A) \leq n$ for any n-by-n ($n \geq 2$) matrix A, and $k(A) = \lceil n/2 \rceil$ for any S_n-matrix A ($n \geq 3$). Recall that an n-by-n matrix A is of *class* S_n if it is a *contraction*, that is, $\| A \| \equiv \max_{\|x\|=1} \|Ax\| \leq 1$, its eigenvalues are all in the open unit disc $\mathbb{D} \equiv \{z \in \mathbb{C} : |z| < 1\}$, and the rank of $I_n - A^*A$ equals one. In [19, Theorem 3.1], it was proven that, for an n-by-n ($n \geq 2$) weighted shift matrix A with weights $w_1, ..., w_n$, $k(A) = n$ if and only if either $|w_1| = \cdots = |w_n|$ or n is even and $|w_1| = |w_3| = \cdots = |w_{n-1}|$ and $|w_2| = |w_4| = \cdots = |w_n|$. Recall that an n-by-n ($n \geq 2$) matrix of the form

$$\begin{bmatrix} 0 & w_1 & & \\ & 0 & \ddots & \\ & & \ddots & w_{n-1} \\ w_n & & & 0 \end{bmatrix}$$

is called a *weighted shift matrix* with *weights* $w_1, ..., w_n$. Moreover, in [2] $k(A)$ is computed for two classes of n-by-n matrices as follows. An n-by-n matrix A is *almost normal* if it has $n-1$ orthogonal eigenvectors. Note that every almost normal matrix is unitarily similar to $A_n \oplus A_a$, where A_n is normal while A_a is almost normal and *unitarily irreducible* (cf. [14]). Recall that a matrix A is unitarily reducible if and only if A is unitarily similar to $A_1 \oplus A_2$ for some lower-dimensional matrices A_1 and A_2; otherwise, A is unitarily irreducible. In [2, Theorem 3], it was proven that, for any almost normal matrix A, $k(A) = l_1 + l_2$, where l_1 is the number of eigenvalues of

A_n located on $\partial W(A)$, counting their multiplicities, and

$$l_2 = \begin{cases} 0 & \text{if } W(A_a) \text{ lies in the interior of } W(A_n), \\ 2 & \text{if there exist distinct parallel supporting lines of } W(A) \\ & \text{passing through points of } W(A_a), \text{ or} \\ 1 & \text{otherwise.} \end{cases}$$

Furthermore, [2, Theorem 5] shows that if A is an n-by-n ($n \geq 3$) *tridiagonal Toeplitz matrix* of the form

$$\begin{bmatrix} a & c & 0 & \dots & 0 \\ b & a & c & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & b & a & c \\ 0 & \dots & 0 & b & a \end{bmatrix},$$

then

$$k(A) = \begin{cases} n & \text{if } |b| = |c|, \\ \lceil n/2 \rceil & \text{otherwise.} \end{cases}$$

We will show that if A is a normal or a quadratic matrix, then the exact value of $k(A)$ can be computed. Recall that a *quadratic matrix* A is one which satisfies $A^2 + z_1 A + z_2 I = 0$ for some scalars z_1 and z_2. For a matrix A of the form $B \oplus C$, we show that $k(A) = 2$ if and only if the numerical range of one summand, say, B is contained in the interior of the numerical range of the other summand C and $k(C) = 2$. For an irreducible matrix A, we can determine exactly when the value of $k(A)$ equals the size of A. These are then applied to determine $k(A)$ for a reducible matrix A of size 4 in terms of the shape of $W(A)$. These results also appeared in [10].

In Chapter 3, we continue to study $k(A)$ for two classes of n-by-n nonnegative matrices A. Recall that an n-by-n matrix $A = [a_{ij}]_{i,j=1}^n$ is a *nonnegative matrix*, denoted by $A \succeq 0$, if $a_{ij} \geq 0$ for all i and j. Recall also that a square matrix P is a *permutation matrix* if there is exactly one 1 on every row and every column and all other entries are 0. Note that any permutation matrix P is unitary with $P^* = P^T = P^{-1}$. Two square matrices A and B of the same size are *permutationally similar* if there is a permutation matrix P such that $P^TAP = B$, which is denoted by $A \cong_p B$. A matrix A is *permutationally reducible* if it is permutationally similar to a matrix of the form $\begin{bmatrix} B & C \\ 0 & D \end{bmatrix}$, where B and D are square matrices; otherwise, A is *permutationally irreducible*. This should not be confused with the notion of unitarily reducible (resp., irreducible) matrix. For nonnegative matrices, reducibility (resp., irreducibility) in general refers to the permutational one. Note that the reducibility (or irreducibility, for that matter) of nonnegative matrices is preserved under the permutational similarity, and the irreducibility of a nonnegative matrix A passes to that of $\operatorname{Re} A$. The converse of the latter is false as witness $A = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$. If A is an n-by-n ($n \geq 2$) nonnegative matrix of the form

$$\begin{bmatrix} 0 & A_1 & & 0 \\ & 0 & \ddots & \\ & & \ddots & A_{m-1} \\ 0 & & & 0 \end{bmatrix},$$

where $m \geq 3$ and the diagonal zeros are zero square matrices, with irreducible real part, then $k(A)$ has an upper bound $m-1$. In addition, we also obtain necessary and sufficient conditions for $k(A) = m-1$ for such a matrix A. The other class of nonnegative matrices we study is the *doubly stochastic* ones. Recall that an n-by-n nonnegative matrices A is doubly stochastic if its row sums and column sums are all equal to one. It is proven that the value of $k(A)$ can be determined for any doubly stochastic matrix A of size 3 or 4 in terms of the shape of $W(A)$. Note that the shapes of $W(A)$ can be determined completely by the tests given in [1, Theorems 1 and 3]. Moreover, the lower bound of $k(A)$, in general, is also found for an n-by-n ($n \geq 5$) doubly stochastic matrix via possible shapes of $W(A)$.

2 Gau-Wu numbers of direct sums of matrices

2.1 Introduction

In Section 2.2 below, we first determine the value of $k(A)$ for a normal matrix A (Proposition 2.2.1). Then we consider the direct sum $A = B \oplus C$, where the numerical ranges $W(B)$ and $W(C)$ are assumed to be disjoint. In this case, we show that the value of $k(A)$ is equal to the sum of $k_1(B)$ and $k_1(C)$ (Theorem 2.2.2), where $k_1(B)$ and $k_1(C)$ are defined as follows. We define $k_1(B)$ to be the maximum number k for which there are orthonormal vectors $x_1, \ldots, x_k$ in $\mathbb{C}^n$ such that $\langle Bx_i, x_i \rangle$ is in $\partial W(A) \cap \partial W(B)$ for all $i = 1, \ldots, k$, and similarly for $k_1(C)$. Based on the proof of Theorem 2.2.2, we obtain the same formula for $k(A)$ under a slightly weaker condition on B and C (Theorem 2.2.6). In Section 2.3, we give some applications of Theorem 2.2.6. The first one (Proposition 2.3.1) shows that the equality $k(A) = k_1(B) + k_1(C)$ holds for a matrix A of the form $B \oplus C$ with normal C. In particular, we are able to determine the value of $k(A)$ for any 4-by-4 reducible matrix A (Corollary 2.3.4 and Propositions 2.3.7 – 2.3.9). Moreover, the number $k(A \oplus (A + aI_n))$ can be determined for any n-by-n matrix A and nonzero complex number a (Proposition 2.3.10). At the end of Section 2.3, we propose several open questions on $k(B \oplus C)$ and give a partial answer for one of them (Proposition 2.3.11). That is, the equality $k(\oplus_{j=1}^{m} A) = m \cdot k(A)$ holds if the dimension of $H_\xi(A)$ equals one for each $\xi \in \partial W(A)$, where the subspace $H_\xi(A)$ is defined in the first paragraph of Section 2.2. By using this, we can determine the value of $k(A)$ for a quadratic matrix A (Corollary 2.3.12).

Note that all of the results in Sections 2.2 and 2.3 have also appeared in [10].

We end this section by fixing some notation. A finite square matrix A is called *positive definite*, denoted by $A > 0$, if A is Hermitian and $\langle Ax, x\rangle > 0$ for all $x \neq 0$. I_n is the n-by-n identity matrix. The n-by-n diagonal matrix with diagonals $\xi_1, ..., \xi_n$ is denoted by $\operatorname{diag}(\xi_1, ..., \xi_n)$. The *cardinal number* of a set S is $\#(S)$. The notation δ_{ij} is the *Kronecker delta*, that is, δ_{ij} has the value 1 if $i = j$, and the value 0 if otherwise. The *span* of a nonempty subset S of a vector space V, denoted by $\operatorname{span}(S)$, is the subspace consisting of all linear combinations of the vectors in S.

2.2 Direct sum

We start by reviewing a few basic facts concerning the boundary points of a numerical range. For an n-by-n matrix A, a point ξ in $\partial W(A)$ and a supporting line L of $W(A)$ which passes through ξ, there is a θ in $[0, 2\pi)$ such that the ray from the origin which forms angle θ from the positive x-axis is perpendicular to L. In this case, $\operatorname{Re}(e^{-i\theta}\xi)$ is the maximum eigenvalue of $\operatorname{Re}(e^{-i\theta}A)$ with the corresponding eigenspace $E_{\xi,L}(A) \equiv \ker \operatorname{Re}(e^{-i\theta}(A - \xi I_n))$. Let $K_\xi(A)$ denote the set $\{x \in \mathbb{C}^n : \langle Ax, x\rangle = \xi\|x\|^2\}$ and $H_\xi(A)$ the subspace spanned by $K_\xi(A)$. If the matrix A is clear from the context, we will abbreviate these to $E_{\xi,L}$, K_ξ and H_ξ, respectively. For other related properties, we refer the reader to [4, Theorem 1] and [19, Proposition 2.2]. The next proposition on the value of $k(A)$ for a normal matrix A is an easy consequence of [19, Lemma 2.9]. It can be regarded as a motivation for our study of this topic.

Proposition 2.2.1. *If A is an n-by-n normal matrix with p eigenvalues (counting multiplicity) in $\partial W(A)$, then $k(A) = p$.*

Proof. We may assume, after a unitary similarity, that A is a matrix of the form $B \oplus C$, where $B = \operatorname{diag}(\lambda_1, \ldots, \lambda_p)$ and $C = \operatorname{diag}(\lambda_{p+1}, \ldots, \lambda_n)$ with $\lambda_1, \ldots, \lambda_p \in \partial W(A)$ and $\lambda_{p+1}, \ldots, \lambda_n \in \operatorname{int} W(B)$. It follows from [19, Lemma 2.9] that $k(A) =$

$k(B \oplus C) = k(B) = p.$ ■

One of our main results of this section is the following theorem for $k(A)$ when A is a matrix of the form $B \oplus C$ with disjoint $W(B)$ and $W(C)$. Recall that the value of $k_1(B)$ is the maximum number k for which there are orthonormal vectors $x_1, \ldots, x_k$ in $\mathbb{C}^n$ such that $\langle Bx_i, x_i \rangle$ is in $\partial W(A) \cap \partial W(B)$ for all $i = 1, \ldots, k$. If the subset $\partial W(A) \cap \partial W(B)$ is empty, then we define $k_1(B) = 0$. The following theorem provides a formula for determining the value of $k(A)$ by $k_1(B)$ and $k_1(C)$.

Theorem 2.2.2. *Let $A = B \oplus C$, where B and C are n-by-n and m-by-m matrices, respectively. If the numerical ranges $W(B)$ and $W(C)$ are disjoint, then $k(A) = k_1(B) + k_1(C) \leq k(B) + k(C)$. In this case, $k(A) = k(B) + k(C)$ if and only if $k_1(B) = k(B)$ and $k_1(C) = k(C)$. In particular, $k(A) = m + n$ if and only if $k_1(B) = k(B) = n$ and $k_1(C) = k(C) = m$.*

This will be proven after the following lemma which is the case when C equals a 1-by-1 matrix $[c]$.

Recall that z is an *extreme point* of the convex subset Δ of $\mathbb{C}$ if z belongs to Δ and cannot be expressed as a convex combination of two other (distinct) points of Δ; otherwise, z is a *nonextreme point*. Recall also that a point z is a *corner* of a convex set Δ of the complex plane if z is in the closure of Δ and Δ has two supporting lines passing through z. If A is a finite matrix, $\xi = \langle Ax, x \rangle$ and $\|x\| = 1$, then x is called a unit vector corresponding to the point ξ in $W(A)$.

Lemma 2.2.3. *If $A = B \oplus [c]$ is an n-by-n matrix, where B is of size $n-1$ and c is a scalar, then $k(A) = k_1(B) + k_1([c])$.*

Proof. By Proposition 2.2.1, we may assume that the interior of the numerical range $W(B)$ is nonempty. If c is in the interior of $W(B)$, then $k(A) = k(B)$ by [19, Lemma 2.9]. Obviously, $k(B) = k_1(B)$ and $k_1([c]) = 0$ in this case. Hence it remains to consider the case when c is outside the interior of $W(B)$. That is, we will prove that $k(A) = k_1(B) + 1$ for $c \notin \operatorname{int} W(B)$. By the definition of $k(A)$, there are points $\xi_j = \langle Az_j, z_j\rangle$ in $\partial W(A)$, $j = 1, 2, \ldots, k(A)$, with $\langle z_i, z_j\rangle = \delta_{ij}$ for $i, j = 1, ..., k(A)$. Clearly, the inequality $k(A) \geq k_1(B) + 1$ holds. Assume that $k(A) \geq k_1(B) + 2$. Let $z_j = x_j \oplus y_j$ for each j. We claim that every x_j is a nonzero vector. Indeed, if $x_{j_0} = 0$ for some j_0, then $y_{j_0} \neq 0$ and $\langle z_j, z_{j_0}\rangle = \langle y_j, y_{j_0}\rangle = 0$ for all $j \neq j_0$. This implies that $y_j = 0$ for all $j \neq j_0$ and thus $k_1(B)$ is at least $k_1(B) + 1$, which is absurd. Hence the claim has been proven. From $\xi_j = \langle Az_j, z_j\rangle = \|x_j\|^2 b_j + \|y_j\|^2 c \in \partial W(A)$, where $b_j = \langle B(x_j/\|x_j\|), x_j/\|x_j\|\rangle$, it follows that ξ_j is in the (possibly degenerate) line segment $[c, b_j]$, and b_j is in the boundary of $W(B)$ for each j. We note that there are at least two nonzero y_j's; this is because if otherwise, then we obtain the inequality $k_1(B) \geq k_1(B)+1$, which is a contradiction. Hence we may assume that $y_1, ..., y_h \neq 0$, where $h \geq 2$, and that this h is the maximal such number.

If c is not in $W(B)$, then there are exactly two points p and q in the boundary of $W(B)$ such that the two line segments $[c, p]$ and $[c, q]$ are in the boundary of $W(A)$ and the relative interior of these two line segments are disjoint from the boundary of $W(B)$ by the fact that $W(A)$ is the convex hull of the union of $W(B)$ and the singleton $\{c\}$. Hence there are three cases to consider: the intersection of the boundary of $W(B)$ and the supporting line at p (resp., q) containing $[c, p]$ (resp., $[c, q]$) is (1) $\{p\}$ (resp., $\{q\}$), (2) a line segment $[p, p']$ (resp., $\{q\}$) or $\{p\}$ (resp., a line segment $[q, q']$), or (3) a line segment $[p, p']$ (resp., a line segment $[q, q']$) (cf. Figure 2.2.4). We need only prove case (2) since other cases can be done similarly.

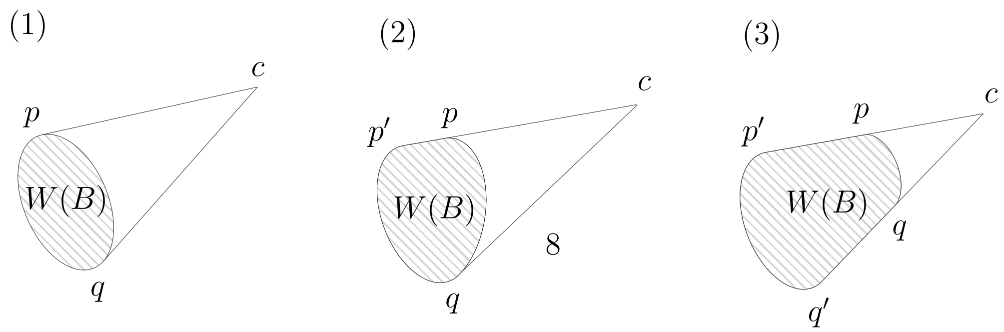

Figure 2.2.4

Define three (disjoint) subsets consisting of the corresponding unit vectors, and their cardinal numbers, respectively, in the following:

$$\begin{aligned} R &\equiv \{z_j : \xi_j \in [c, p')\} \text{ with } r \equiv \#(R), \\ S &\equiv \{z_j : \xi_j \in (c, q)\} \text{ with } s \equiv \#(S), \quad \text{and} \\ T &\equiv \{z_j : \xi_j \in \partial W(A) \backslash ([c, p') \cup (c, q))\} \text{ with } t \equiv \#(T). \end{aligned}$$

So, $k(A) = r + s + t$. Obviously, every $z_j \in T$ is of the form $x_j \oplus 0$. Moreover, we partition R into two disjoint subsets $R_1 \equiv \{z_j : y_j \neq 0\}$ and $R_2 \equiv \{z_j : y_j = 0\}$. We call their cardinal numbers r_1 and r_2, respectively. Without loss of generality, we may assume that $R_1 = \{z_1, ..., z_{r_1}\}$, $R_2 = \{z_{r_1+1}, ..., z_{r_1+r_2}\}$, $S = \{z_{r+1}, ..., z_{r+s}\}$, and $T = \{z_{r+s+1}, ..., z_{r+s+t}\}$, where $r_1 + r_2 = r$. This shows that $r_1 + s = h \geq 2$.

First assume that $s = 0$. Then $r_1 \geq 2$. For the clarity of the proof, the following method is called $(*)$. Since every y_j, $j = 1, \ldots, r_1$, is nonzero, we define the vectors $z'_j = (x_j/y_j) \oplus 1$ for these j's so that the vectors in $M \equiv \left\{ \left(z'_1 - z'_j\right) / \left\|z'_1 - z'_j\right\| \right\}_{j=2}^{r_1} = \left\{ \left(((x_1/y_1) - (x_j/y_j)) \oplus 0\right) / \left\|z'_1 - z'_j\right\| \right\}_{j=2}^{r_1}$ are linearly independent and are perpendicular to vectors in $T \cup R_2$. This together with [4, Theorem 1] shows that $\operatorname{span}(M) \subseteq \cup_{\eta \in [c,p']} K_\eta(A)$ and thus every unit vector in $\operatorname{span}(M)$ is a unit vector corresponding to some $\eta \in \partial W(B)$. Choosing an orthonormal basis $\{v_j \oplus 0\}_{j=2}^{r_1}$ for the subspace $\operatorname{span}(M)$, we deduce from the orthonormality of the vectors in $T \cup R_2 \cup \{v_j \oplus 0\}_{j=2}^{r_1}$ that

$$k_1(B) \geq t + r_2 + (r_1 - 1) = r + s + t - 1 = k(A) - 1 \geq k_1(B) + 1,$$

which is impossible. Hence we must have $s \geq 1$.

If $s = 1$, then $r_1 \geq 1$. A similar argument as above yields that

$$k_1(B) \geq \begin{cases} t + r_2 + 1 & \text{if } r_1 = 1, \text{ and} \\ t + r_2 + (r_1 - 1) + 1 & \text{if } r_1 \geq 2 \end{cases}$$

by considering the orthonormal subsets $T \cup R_2 \cup \{(x_{r+1}/\left\|x_{r+1}\right\|) \oplus 0\}$ and $T \cup R_2 \cup \{v_j \oplus 0\}_{j=2}^{r_1} \cup \{(x_{r+1}/\left\|x_{r+1}\right\|) \oplus 0\}$, where $\{v_j \oplus 0\}_{j=2}^{r_1}$ is an orthonormal subset of $\operatorname{span}(R_1)$, via applying $(*)$ on R_1. The above inequalities imply that

$$k_1(B) \geq \begin{cases} r + s + t - 1 \geq k(A) - 1 \geq k_1(B) + 1 & \text{if } r_1 = 1, \text{ and} \\ r + s + t - 1 \geq k(A) - 1 \geq k_1(B) + 1 & \text{if } r_1 \geq 2. \end{cases}$$

This is a contradiction. Hence $s \geq 2$.

If $r_1 = 0$, then applying $(*)$ on S, we reach a contradiction since

$$k_1(B) \geq t + r_2 + (s-1) = r + s + t - 1 = k(A) - 1 \geq k_1(B) + 1.$$

If $r_1 = 1$, then we obviously have the linear independence of the subset $N \equiv \left\{\left(z_1' - z_j'\right) / \left\|z_1' - z_j'\right\|\right\}_{j=r+2}^{r+s} = \left\{\left(\left((x_1/y_1) - (x_j/y_j)\right) \oplus 0\right) / \left\|z_1' - z_j'\right\|\right\}_{j=r+2}^{r+s}$ by applying $(*)$ on S again. Let $\{v_j \oplus 0\}_{j=r+2}^{r+s}$ be an orthonormal basis for the subspace $\operatorname{span}(N)$. Hence

$$k_1(B) \geq t + r_2 + (s-1) + 1 = r + s + t - 1 = k(A) - 1 \geq k_1(B) + 1$$

by the orthonormality of the vectors in $T \cup R_2 \cup \{v_j \oplus 0\}_{j=r+2}^{r+s} \cup \{(x_1/\left\|x_1\right\|) \oplus 0\}$. This is again a contradiction. If $r_1 \geq 2$, then applying Method I on S and R_1, we have the linear independence of the subsets $P \equiv \left\{\left(z_1' - z_j'\right) / \left\|z_1' - z_j'\right\|\right\}_{j=r+2}^{r+s} = \left\{\left(\left((x_1/y_1) - (x_j/y_j)\right) \oplus 0\right) / \left\|z_1' - z_j'\right\|\right\}_{j=r+2}^{r+s}$ and $Q \equiv \left\{\left(z_1' - z_j'\right) / \left\|z_1' - z_j'\right\|\right\}_{j=2}^{r_1} = \left\{\left(\left((x_1/y_1) - (x_j/y_j)\right) \oplus 0\right) / \left\|z_1' - z_j'\right\|\right\}_{j=2}^{r_1}$, respectively. Let $\{v_j \oplus 0\}_{j=r+2}^{r+s}$ be an orthonormal basis for $\operatorname{span}(P)$. Then $\operatorname{span}(P) \oplus \operatorname{span}(x \oplus y) = \operatorname{span}(S)$ for some unit vector $x \oplus y$ orthogonal to $\operatorname{span}(P)$. Clearly, x is a nonzero vector; this is because if otherwise, then $0 \oplus y (\in \text{span}\,(S))$ is orthogonal to $z_1 = x_1 \oplus y_1 (\in R_1)$, which contradicts the fact that y and y_1 are nonzero scalars. Let $\{v_j \oplus 0\}_{j=2}^{r_1}$ be an orthonormal basis for the subspace span (Q). Then we conclude that the subset $T \cup R_2 \cup \{v_j \oplus 0\}_{j=2}^{r_1} \cup \{v_j \oplus 0\}_{j=r+2}^{r+s} \cup \{(x/\|x\|) \oplus 0\}$ is orthonormal so that

$$k_1(B) \geq t + r_2 + (r_1 - 1) + (s - 1) + 1 = r + s + t - 1 = k(A) - 1 \geq k_1(B) + 1,$$

which is a contradiction. This completes the proof of case (2).

In case (1), we define three subsets consisting of the corresponding unit vectors, and their cardinal numbers, respectively, as follows:

$$\begin{aligned} R &\equiv \{z_j : \xi_j \in [c, p)\} \text{ with } r \equiv \#(R), \\ S &\equiv \{z_j : \xi_j \in (c, q)\} \text{ with } s \equiv \#(S), \quad \text{and} \\ T &\equiv \{z_j : \xi_j \in \partial W(A) \backslash ([c,p) \cup (c,q))\} \text{ with } t \equiv \#(T). \end{aligned}$$

As for case (3), we have

$$\begin{aligned} R &\equiv \{z_j : \xi_j \in [c, p')\} \text{ with } r \equiv \#(R), \\ S &\equiv \{z_j : \xi_j \in (c, q')\} \text{ with } s \equiv \#(S), \quad \text{and} \\ T &\equiv \{z_j : \xi_j \in \partial W(A) \backslash ([c,p') \cup (c,q'))\} \text{ with } t \equiv \#(T). \end{aligned}$$

As before, we partition R (resp., S) into two disjoint subsets $R_1 \equiv \{z_j : y_j \neq 0\}$ and $R_2 \equiv \{z_j : y_j = 0\}$ (resp., $S_1 \equiv \{z_j : y_j \neq 0\}$ and $S_2 \equiv \{z_j : y_j = 0\}$). Based on the arguments for case (2), we get a series of contradictions for each individual case. In a similar fashion, we remark that if $A = B \oplus cI_m$, where $c \notin W(B)$, then $k(A) = k_1(B) + k_1(cI_m) = k_1(B) + m$. This remark will be used in the remaining part of the proof.

To complete the proof, we let c be in the boundary of $W(B)$. Assume that $\partial W(B)$ contains no line segment. We infer that $c = b_j = \xi_j$ for $j = 1, ..., h$ since these ξ_j's are in the (possibly degenerate) line segment $[c, b_j]$ contained in the boundary of $W(B)$. Define a new vector $z_j' = (x_j/y_j) \oplus 1$ for each $j = 1, ..., h$. Then the subset $S \equiv \left\{ \left(z_1' - z_j'\right) / \left\| z_1' - z_j' \right\| \right\}_{j=2}^{h} = \left\{ \left(\left((x_1/y_1) - (x_j/y_j)\right) \oplus 0\right) / \left\| z_1' - z_j' \right\| \right\}_{j=2}^{h}$ is linearly independent. Since c is an extreme point of $W(A)$, we have $H_c(A) = K_c(A)$ by [4, Theorem 1] and $\operatorname{span}(S)$ is a subspace of $H_c(A)$. Let $\{v_j \oplus 0\}_{j=2}^{h}$ be an orthonormal basis for $\operatorname{span}(S)$. Then $c = \langle A(v_j \oplus 0), v_j \oplus 0\rangle = \langle Bv_j, v_j\rangle$ is in $\partial W(B)$ for $j = 2, \ldots, h$. Hence

$$k(B) \geq (h-1) + (k(A) - h) = k(A) - 1 \geq k(B) + 1.$$

This is a contradiction. So, we may assume that $\partial W(B)$ contains a line segment l such that c belongs to l. If c is not an extreme point of l, then we infer that $c = b_j = \xi_j$ or $\xi_j \in (c, b_j)$ for $j = 1, ..., h$ since x_j and y_j are nonzero vectors for these j's. Hence $z_j \in H_c(A)$ for $j = 1, ..., h$ by [4, Theorem 1]. Similar arguments show that $H_c(A)$ has an orthonormal subset $\{w_j \oplus 0\}_{j=2}^{h}$. Since $H_c(A) = \cup_{\eta \in l} K_\eta(A)$ by [4, Theorem 1], this implies that $w_j \oplus 0 \in K_{\eta_j}(A)$, where $\eta_j \in l$, for $j = 2, ..., h$. From $\eta_j = \langle A(w_j \oplus 0), w_j \oplus 0\rangle = \langle Bw_j, w_j\rangle \in l \subseteq \partial W(B)$, where $j = 2, ..., h$, we reach a contradiction since

$$k(B) \geq (h-1) + (k(A) - h) = k(A) - 1 \geq k(B) + 1.$$

For the remaining part of the proof, let c be an extreme point of l, where l is a line segment on the boundary of $W(B)$. We consider two cases: either (a) there is only one line segment in $\partial W(B)$ passing through c, or (b) there are exactly two line segments in $\partial W(B)$ passing through c. In case (a), since x_j and y_j are nonzero vectors for $j = 1, ..., h$, we infer that $c = b_j = \xi_j$ or $\xi_j \in (c, b_j)$ for these j's. This implies that $z_j \in H_\eta(A)$ by [4, Theorem 1], where η is not an extreme point of l. So, the same arguments as above lead us to a contradiction. For case (b), since c is a corner of $W(B)$, c is a reducing eigenvalue of B by [3, Theorem 1]. Thus B is unitarily similar to a matrix of the form $B' \oplus cI_{n'}$, where c is not an eigenvalue of B', and the size of B' and n' are both less than n. Obviously, $c \notin W(B')$. We apply the preceding remark as for the case of $c \notin W(B)$ to see that $k(A) = k(B' \oplus cI_{n'+1}) = k_1(B') + n' + 1$, and $k(B) = k(B' \oplus cI_{n'}) = k_1(B') + n'$. In addition, $k(B) = k_1(B)$ in this case. Hence we obtain that $k(A) = k_1(B) + 1$, which contradicts our assumption that $k(A) \geq k_1(B) + 2$. With this, we conclude the proof of the asserted equality. ∎

We remark that the part of the proof of Lemma 2.2.3 on $c \notin W(B)$ involves the following three cases (1), (2), and (3) depending on whether $\partial W(B)$ contains a line segment or otherwise. In case (1), we have $R = \{z_j : y_j \neq 0\}$ and $S = \{z_j : y_j \neq 0\}$, in (2) $R = R_1 \cup R_2$, where $R_1 = \{z_j : y_j \neq 0\}$ and $R_2 = \{z_j : y_j = 0\}$, and $S = \{z_j : y_j \neq 0\}$, and in (3) $R = R_1 \cup R_2$, where $R_1 = \{z_j : y_j \neq 0\}$ and $R_2 = \{z_j : y_j = 0\}$, and $S = S_1 \cup S_2$, where $S_1 = \{z_j : y_j \neq 0\}$ and $S_2 = \{z_j : y_j = 0\}$. Note that the key point is to handle R and S in (1), R_1 and S in (2), and R_1 and S_1 in (3), that is, all nonzero y_j's of the three cases. We find that the proofs of the three cases are almost the same. This observation can facilitate the proof of Theorem 2.2.2 as follows. If $\partial W(B)$ contains a line segment such that this line segment is a portion of $\partial W(A)$ and stretches to a point of $\partial W(C)$, then we take the same method as the proof of Lemma 2.2.3 on $c \notin W(B)$ to partition the corresponding R into $R_1 = \{z_j : y_j \neq 0\}$ and $R_2 = \{z_j : y_j = 0\}$. As mentioned above, we need only handle R_1. On the other hand, if $\partial W(B)$ contains no such line segments, then we need only handle the corresponding $R = \{z_j : y_j \neq 0\}$. From this, there is no difference between the proofs of the two cases. Hence we may assume, in the proof of Theorem 2.2.2, that $\partial W(B)$ and $\partial W(C)$ contain no line segments.

Before giving a proof of Theorem 2.2.2, we note several things. First of all, by Lemma 2.2.3, we may assume that both of the numerical ranges $W(B)$ and $W(C)$ are not singletons. Secondly, we may further assume that $\partial W(B)$ and $\partial W(C)$ contain no line segment by the above remark. Thirdly, since $W(A)$ is the convex hull of the union of $W(B)$ and $W(C)$, there are two line segments, called $[a, p]$ and $[b, q]$, in $\partial W(A)$, where $a, b \in \partial W(B)$ and $p, q \in \partial W(C)$. Fourthly, it is easy to check that $a \neq b$ and

$p \neq q$. Indeed, if $a = b$, then a is a corner. By [3, Theorem 1], we obtain that a is a reducing eigenvalue of A, and hence a is a reducing eigenvalue of B. This shows that $W(B)$ must contain a line segment, which contradicts our previous assumption. Similarly, we also have $p \neq q$. Combining the above, we have the following Figure 2.2.5 as the numerical range $W(A)$.

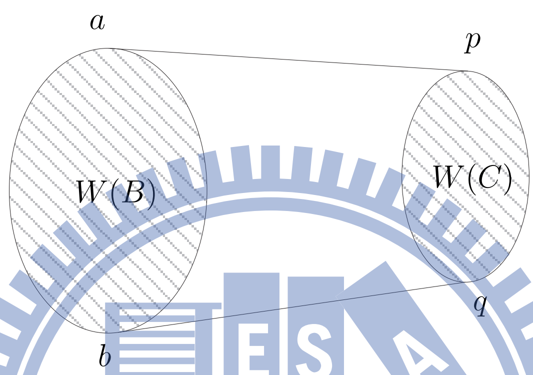

Figure 2.2.5

As before, by the definition of $k(A)$, there exist $\xi_j = \langle Az_j, z_j \rangle \in \partial W(A)$, $j = 1, 2, \ldots, k(A)$, where $z_j = x_j \oplus y_j$, and $\langle z_i, z_j \rangle = \delta_{ij}$ for $i, j = 1, ..., k(A)$. We define four (disjoint) subsets consisting of the corresponding unit vectors, and their cardinal numbers, respectively, as follows:

$$\begin{aligned} R &\equiv \{z_j : \xi_j \in (a, p)\} \text{ with } r \equiv \#(R), \\ S &\equiv \{z_j : \xi_j \in (b, q)\} \text{ with } s \equiv \#(S), \\ T_B &\equiv \{z_j : \xi_j \in \partial W(A) \cap \partial W(B)\} \text{ with } t_1 \equiv \#(T_B), \quad \text{and} \\ T_C &\equiv \{z_j : \xi_j \in \partial W(A) \cap \partial W(C)\} \text{ with } t_2 \equiv \#(T_C). \end{aligned}$$

Since the intersection of $W(B)$ and $W(C)$ is empty, and $\partial W(B)$ and $\partial W(C)$ contain no line segment, we may assume that

$$\begin{aligned} R &= \{z_j = x_j \oplus y_j : x_j \neq 0 \text{ and } y_j \neq 0\}_{j=1}^{r}, \\ S &= \{z_j = x_j \oplus y_j : x_j \neq 0 \text{ and } y_j \neq 0\}_{j=r+1}^{r+s}, \\ T_B &= \{z_j = x_j \oplus 0 : x_j \neq 0\}_{j=r+s+1}^{r+s+t_1}, \quad \text{and} \\ T_C &= \{z_j = 0 \oplus y_j : y_j \neq 0\}_{j=r+s+t_1+1}^{r+s+t_1+t_2}. \end{aligned}$$

So, $k(A) = r + s + t_1 + t_2$, $k_1(B) \geq t_1$ and $k_1(C) \geq t_2$. Clearly, the inequality $k(A) \geq k_1(B) + k_1(C)$ holds. Now we are ready to prove Theorem 2.2.2.

Proof of Theorem 2.2.2. We need only prove that the reversed inequality $k_1(B) + k_1(C) \geq k(A)$ holds. First, we consider the case $r = 0$. Assume that $s = 0$. Then our assertion is obvious since

$$k_1(B) + k_1(C) \geq t_1 + t_2 = r + s + t_1 + t_2 = k(A).$$

Assume that $s = 1$, i.e., $z_1 = x_1 \oplus y_1 \in S$. Then $k_1(B) \geq t_1 + 1$ since the unit vector $(x_1/\|x_1\|) \oplus 0$ is clearly orthogonal to T_B and $\langle B(x_1/\|x_1\|), x_1/\|x_1\|\rangle$ is in $\partial W(B)$ by the convex combination

$$\langle Az_1, z_1\rangle = \|x_1\|^2 \left\langle B\frac{x_1}{\|x_1\|}, \frac{x_1}{\|x_1\|}\right\rangle + \|y_1\|^2 \left\langle C\frac{y_1}{\|y_1\|}, \frac{y_1}{\|y_1\|}\right\rangle \in (b, q).$$

Hence

$$k_1(B) + k_1(C) \geq (t_1 + 1) + t_2 = r + s + t_1 + t_2 = k(A).$$

Assume that $s = 2$, i.e., $z_1 = x_1 \oplus y_1$ and $z_2 = x_2 \oplus y_2 \in S$. If x_1 and x_2 are linearly independent, then by the Gram-Schmidt process, there are two unit vectors z_1' and z_2', where $z_j' = x_j' \oplus y_j'$ with $x_j' \neq 0$ for $j = 1, 2$ such that x_1' and x_2' are mutually orthogonal, and span $(\{z_1, z_2\})$ is equal to span $(\{z_1', z_2'\})$. Choosing the two unit vectors $(x_1'/\|x_1'\|) \oplus 0$ and $(x_2'/\|x_2'\|) \oplus 0$, we obtain that $k_1(B) \geq t_1 + 2$. Hence

$$k_1(B) + k_1(C) \geq (t_1 + 2) + t_2 = r + s + t_1 + t_2 = k(A).$$

On the other hand, if x_1 and x_2 are linearly dependent, say, $x_2 = \lambda x_1$ for some scalar λ, then we define a new unit vector

$$z_2' = \frac{z_2 - \lambda z_1}{\|z_2 - \lambda z_1\|} = 0 \oplus \frac{y_2 - \lambda y_1}{\|y_2 - \lambda y_1\|} \in \text{span}\left(\{z_1, z_2\}\right)$$

so that span $(\{z_1, z_2\}) = $ span $(\{z_1'\}) \oplus$ span $(\{z_2'\})$ for some unit vector $z_1' \equiv x_1' \oplus y_1'$, where z_1' and z_2' are mutually orthogonal. Clearly, $x_1' \neq 0$ for otherwise it leads to $x_1 = x_2 = 0$, which contradicts the definition of S. From the two unit vectors $(x_1' / \|x_1'\|) \oplus 0$ and z_2', we infer that $k_1(B) \geq t_1 + 1$ and $k_1(C) \geq t_2 + 1$. Hence

$$k_1(B) + k_1(C) \geq (t_1 + 1) + (t_2 + 1) = r + s + t_1 + t_2 = k(A).$$

Assume that $s \geq 3$, that is, $S = \{z_j = x_j \oplus y_j : x_j \neq 0 \text{ and } y_j \neq 0\}_{j=1}^{s}$. We consider the largest linearly independent subset of $\{x_j\}_{j=1}^{s}$ as follows. Without loss of generality, we may assume that this can be $\{x_j\}_{j=1}^{s}$, $\{x_1\}$ or $\{x_j\}_{j=1}^{l}$, where $1 < l < s$. For the first two cases, it can be done by applying similar arguments as for the case of $s = 2$. In the last case, since x_j is a linear combination of $x_1, ..., x_l$ for $j = l+1, ..., s$, it is easy to check that the unit vectors

$$\text{(1)} \qquad z_j' \equiv \frac{z_j - \Sigma_{i=1}^{l} a_i^{(j)} z_i}{\left\| z_j - \Sigma_{i=1}^{l} a_i^{(j)} z_i \right\|} = 0 \oplus \left(\frac{y_j - \Sigma_{i=1}^{l} a_i^{(j)} y_i}{\left\| y_j - \Sigma_{i=1}^{l} a_i^{(j)} y_i \right\|} \right), \; j = l+1, ..., s,$$

are linearly independent. Let $y_j' = \frac{y_j - \Sigma_{i=1}^{l} a_i^{(j)} y_i}{\left\| y_j - \Sigma_{i=1}^{l} a_i^{(j)} y_i \right\|}$ for $j = l+1, ..., s$. Since $F \equiv$ span $\left(\left\{ z_j' = 0 \oplus y_j' \right\}_{j=l+1}^{s} \right)$ is a subspace of the space $V \equiv$ span $\left(\{z_j\}_{j=1}^{s} \right)$, the orthogonal complement of F in V, called E, can be written as span $\left(\left\{ z_j' \equiv x_j' \oplus y_j' \right\}_{j=1}^{l} \right)$ for some unit vectors z_j', $j = 1, ..., l$. By (1), we see that $\{x_j'\}_{j=1}^{l}$ is linearly independent since $\{x_j\}_{j=1}^{l}$ is linearly independent. Hence we may assume that both $\{x_j'\}_{j=1}^{l}$ and $\left\{ y_j' \right\}_{j=l+1}^{s}$ are orthogonal subsets by the Gram-Schmidt process. This shows that $G_1 \equiv \left\{ \left(x_j' / \left\| x_j' \right\| \right) \oplus 0 \right\}_{j=1}^{l}$ and $G_2 \equiv \left\{ 0 \oplus y_j' \right\}_{j=l+1}^{s}$ are orthogonal to T_B and T_C, respectively. Since every vector v in G_1 (resp., G_2) is such that $\langle Av, v \rangle$ is in $\partial W(B)$ (resp., $\partial W(C)$), we obtain that $k_1(B) + k_1(C) \geq k(A)$ from $k_1(B) \geq t_1 + l$ and $k_1(C) \geq t_2 + s - l$. This completes the proof of the case $r = 0$.

Next, we prove for the case $r=1$. Obviously, it is sufficient to consider $s \geq 1$ since the case $r=1$, $s=0$ is the same as the case $r=0$, $s=1$. Assume that $s=1$, that is, $z_1 = x_1 \oplus x_2 \in R$ and $z_2 = x_2 \oplus y_2 \in S$. Then $k_1(B) \geq t_1 + 1$ and $k_1(C) \geq t_2 + 1$ since $(x_1/\|x_1\|) \oplus 0$ and $0 \oplus (y_2/\|y_2\|)$ are orthogonal to T_B and T_C, respectively. Moreover, $\langle B\left(x_1/\left\|x_1\right\|\right), x_1/\left\|x_1\right\|\rangle$ is in the boundary of $W(B)$ by the convex combination

$$\langle Az_1, z_1\rangle = \|x_1\|^2 \left\langle B\frac{x_1}{\|x_1\|}, \frac{x_1}{\|x_1\|}\right\rangle + \|y_1\|^2 \left\langle C\frac{y_1}{\|y_1\|}, \frac{y_1}{\|y_1\|}\right\rangle \in (a,p),$$

and $\langle C\left(y_2/\left\|y_2\right\|\right), y_2/\left\|y_2\right\|\rangle$ is in the boundary of $W(C)$ by the same arguments. Hence

$$k_1(B) + k_1(C) \geq (t_1+1) + (t_2+1) = r + s + t_1 + t_2 = k(A).$$

Assume that $s=2$. Then we have $R = \{z_1 = x_1 \oplus y_1 : x_1 \neq 0 \text{ and } y_1 \neq 0\}$ and $S = \{z_j = x_j \oplus y_j : x_j \neq 0 \text{ and } y_j \neq 0\}_{j=2}^{3}$. If $\{x_2, x_3\}$ is linearly independent, then we may assume that it is an orthogonal set by the Gram-Schmidt process. By the convex combination mentioned above, we infer from the three unit vectors $0 \oplus (y_1/\left\|y_1\right\|)$, $(x_2/\left\|x_2\right\|) \oplus 0$, and $(x_3/\left\|x_3\right\|) \oplus 0$ that $k_1(B) \geq t_1 + 2$ and $k_1(C) \geq t_2 + 1$. Hence

$$k_1(B) + k_1(C) \geq (t_1+2) + (t_1+1) = r + s + t_1 + t_2 = k(A).$$

On the other hand, if $\{x_2, x_3\}$ is linearly dependent, say, $x_2 = \lambda x_3$ for some scalar λ, then we define a new unit vector

$$z_2' = \frac{z_2 - \lambda z_3}{\|z_2 - \lambda z_3\|} = 0 \oplus \frac{y_2 - \lambda y_3}{\|y_2 - \lambda y_3\|} \in \operatorname{span}\left(\{z_2, z_3\}\right)$$

so that $\operatorname{span}\left(\{z_2, z_3\}\right) = \operatorname{span}\left(\{z_2'\}\right) \oplus \operatorname{span}\left(\{z_3'\}\right)$ for some unit vector $z_3' \equiv x_3' \oplus y_3'$, where z_2' is orthogonal to z_3'. Clearly, $x_3' \neq 0$ for otherwise it leads to $x_2 = x_3 = 0$, which contradicts the definition of S. From the three unit vectors $0 \oplus (y_1/\left\|y_1\right\|)$, $0 \oplus \left((y_2 - \lambda y_3)/\left\|y_2 - \lambda y_3\right\|\right)$, and $(x_3'/\left\|x_3'\right\|) \oplus 0$, we infer that $k_1(B) \geq t_1 + 1$ and $k_1(C) \geq t_2 + 2$. Hence

$$k_1(B) + k_1(C) \geq (t_1+1) + (t_2+2) = r + s + t_1 + t_2 = k(A).$$

Assume that $s \geq 3$, that is, $S = \{z_j = x_j \oplus y_j : x_j \neq 0 \text{ and } y_j \neq 0\}_{j=2}^{s+1}$, and $R = \{z_1 = x_1 \oplus y_1 : x_1 \neq 0 \text{ and } y_1 \neq 0\}$. We consider the largest linearly independent subset of $\{x_j\}_{j=2}^{s+1}$, which we may assume to be $\{x_j\}_{j=2}^{s+1}$, $\{x_2\}$ or $\{x_j\}_{j=2}^{l}$, where $2 < l < s+1$. These three largest subsets are similar to those considered under $r = 0$, $s \geq 3$. Indeed, we need only add the unit vector $0 \oplus (y_1/\|y_1\|)$ to every sub-case of the case $r = 0$, $s \geq 3$. Hence we have proved that the reversed inequality $k_1(B) + k_1(C) \geq k(A)$. This completes the proof of the case $r = 1$.

Let $r = 2$. With the help of the preceding discussions, we may assume that $s \geq 2$. Assume that $s = 2$, that is, $R = \{z_j = x_j \oplus y_j : x_j \neq 0 \text{ and } y_j \neq 0\}_{j=1}^{2}$ and $S = \{z_j = x_j \oplus y_j : x_j \neq 0 \text{ and } y_j \neq 0\}_{j=3}^{4}$. If $\{x_3, x_4\}$ is linearly independent, then we consider two cases as follows. First, we assume that $\{y_1, y_2\}$ is linearly independent. We may further assume that $\{x_3, x_4\}$ and $\{y_1, y_2\}$ are orthogonal subsets by the Gram-Schmidt process. Obviously, the two subsets $H_1 \equiv \{0 \oplus (y_1/\|y_1\|),\ 0 \oplus (y_2/\|y_2\|)\}$ and $H_2 \equiv \{(x_3/\|x_3\|) \oplus 0,\ (x_4/\|x_4\|) \oplus 0\}$ are orthogonal to T_C and T_B, respectively. Since every vector v in H_1 (resp., H_2) is such that $\langle Av, v\rangle$ is in the boundary of $W(C)$ (resp., $W(B)$), we infer, from $k_1(B) \geq t_1+2$ and $k_1(C) \geq t_2+2$, that $k_1(B)+k_1(C) \geq k(A)$. On the other hand, assume that $\{y_1, y_2\}$ is linearly dependent, say, $y_1 = \lambda y_2$ for some scalar λ. Then we define a new unit vector $z_1' = (z_1 - \lambda z_2)/\|z_1 - \lambda z_2\| = ((x_1-\lambda x_2)/\|x_1-\lambda x_2\|)\oplus 0$ so that span $(\{z_1, z_2\})$ = span $(\{z_1'\})\oplus$span $(\{z_2'\})$ for some unit vector $z_2' \equiv x_2' \oplus y_2'$, where z_1' and z_2' are mutually orthogonal. Clearly, $y_2' \neq 0$ for otherwise it leads to $y_1 = y_2 = 0$, which contradicts the definition of R. Moreover, we may assume that $\{x_3, x_4\}$ is an orthogonal subset by the Gram-Schmidt process. Hence $H_3 \equiv \{((x_1-\lambda x_2)/\|x_1-\lambda x_2\|) \oplus 0,\ (x_3/\|x_3\|) \oplus 0,\ (x_4/\|x_4\|) \oplus 0\}$ and $H_4 \equiv \{0 \oplus (y_2'/\|y_2'\|)\}$ are orthogonal to T_B and T_C, respectively. Since every vector v in H_3 (resp., H_4) is such that $\langle Av, v\rangle$ is in the boundary of $W(B)$ (resp., $W(C)$), we infer, from $k_1(B) \geq t_1+3$ and $k_1(C) \geq t_2+1$, that $k_1(B)+k_1(C) \geq k(A)$. On the other hand, if $\{x_3, x_4\}$ is linearly dependent, then we need only consider the case that $\{y_1, y_2\}$ is linearly dependent. So, we may assume that $y_1 = \lambda y_2$ and

$x_3 = \mu x_4$ for some scalars λ and μ. Define two new unit vectors

$$z_1' = \frac{z_1 - \lambda z_2}{\|z_1 - \lambda z_2\|} = \frac{x_1 - \lambda x_2}{\|x_1 - \lambda x_2\|} \oplus 0 \text{ and } z_3' = \frac{z_3 - \mu z_4}{\|z_3 - \mu z_4\|} = 0 \oplus \frac{y_3 - \mu y_4}{\|y_3 - \mu y_4\|}.$$

Then span $(\{z_1, z_2\})$ = span $(\{z_1'\})\oplus$span $(\{z_2'\})$ and span $(\{z_3, z_4\})$ = span $(\{z_3'\})\oplus$ span $(\{z_4'\})$ for some unit vectors $z_2' = x_2' \oplus y_2'$ and $z_4' = x_4' \oplus y_4'$, where z_2' (resp., z_4') is orthogonal to z_1' (resp., z_3'). Clearly, y_2' and x_4' are nonzero by the same argument as above. Hence $H_5 \equiv \{((x_1 - \lambda x_2) / \|x_1 - \lambda x_2\|) \oplus 0,\ (x_4'/ \|x_4'\|) \oplus 0\}$ and $H_6 \equiv \{0 \oplus (y_2'/ \|y_2'\|),\ 0 \oplus ((y_3 - \lambda y_4) / \|y_3 - \lambda y_4\|)\}$ are orthogonal to T_B and T_C, respectively. Since every vector v in H_5 (resp., H_6) is such that $\langle Av, v\rangle$ is in the boundary of $W(B)$ (resp., $W(C)$), we infer, from $k_1(B) \geq t_1+2$ and $k_1(C) \geq t_2+2$, that $k_1(B)+k_1(C) \geq k(A)$. Assume that $s \geq 3$, that is, $R = \{z_j = x_j \oplus y_j : x_j \neq 0 \text{ and } y_j \neq 0\}_{j=1}^{2}$, and $S = \{z_j = x_j \oplus y_j : x_j \neq 0 \text{ and } y_j \neq 0\}_{j=3}^{s+2}$. If $\{y_1, y_2\}$ is linearly independent, then we may assume that $\{y_1, y_2\}$ is orthogonal by the Gram-Schmidt process. In this case, we consider the largest linearly independent subset of $\{x_j\}_{j=3}^{s+2}$, which may be assumed to be $\{x_j\}_{j=3}^{s+2}$, $\{x_3\}$ or $\{x_j\}_{j=3}^{l}$ $(3 < l < s+2)$. Each of the three cases can be handled by applying similar arguments as for the cases of $r = 0$, $s \geq 2$. On the other hand, if $\{y_1, y_2\}$ is linearly dependent, say, $y_1 = \lambda y_2$ for some scalar λ, then we define a new unit vector $z_1' = ((x_1 - \lambda x_2)/\|x_1 - \lambda x_2\|) \oplus 0$ so that span $(\{z_1, z_2\})$ = span $(\{z_1'\})\oplus$span $(\{z_2'\})$ for some unit vector $z_2' = x_2' \oplus y_2'$, where z_1' and z_2' are mutually orthogonal. Clearly, y_2' is nonzero by the same argument as for the case of $r = 0$, $s = 2$. To complete the proof, it remains to consider the three cases mentioned above. By applying similar arguments again as for the cases of $r = 0$, $s \geq 2$, we obtain the reversed inequality $k_1(B) + k_1(C) \geq k(A)$. This completes the proof of the case $r = 2$.

Finally, assume that $r \geq 3$. It suffices to consider $s \geq 3$ since $s \leq 2$ has been proven if we exchange the roles of s and r. Hence $R = \{z_j = x_j \oplus y_j : x_j \neq 0 \text{ and } y_j \neq 0\}_{j=1}^{r}$ and $S = \{z_j = x_j \oplus y_j : x_j \neq 0 \text{ and } y_j \neq 0\}_{j=r+1}^{r+s}$. As mentioned previously, there are three cases by considering the largest linearly independent subset of $\{y_j\}_{j=1}^{r}$ (resp.,

$\{x_j\}_{j=r+1}^{r+s}$). Without loss of generality, we may assume that this subset is $\{y_j\}_{j=1}^{r}$, $\{y_1\}$ or $\{y_j\}_{j=1}^{l_1}$, where $1 < l_1 < r$, and $\{x_j\}_{j=r+1}^{r+s}$, $\{x_{r+1}\}$ or $\{x_j\}_{j=r+1}^{r+l_2}$, where $1 < l_2 < s$. There are a total of nine cases to be considered. Since each case is similar to the one under $r = 0$, $s \geq 1$, it follows that the reversed inequality $k_1(B) + k_1(C) \geq k(A)$ holds. This completes the proof of the case $r \geq 3$. ■

At the end of the section, we give a generalization of Theorem 2.2.2 under a slightly weaker condition on B and C. Let A be a matrix of the form $B \oplus C$. Since $W(A)$ is the convex hull of the union of $W(B)$ and $W(C)$, we consider two (disjoint) subsets of $\partial W(A)$ as follows: one is $\partial W(A) \setminus (\partial W(B) \cup \partial W(C)) \equiv \Gamma_1$, and the other is $\partial W(A) \cap \partial W(B) \cap \partial W(C) \equiv \Gamma_2$. Geometrically, Γ_1 consists of the line segments contained in $\partial W(A)$ but not in $\partial W(B) \cup \partial W(C)$. On the other hand, since the common boundaries of the three numerical ranges consist of line segments and points which are not in any line segments, every point of the latter can be regarded as a degenerate line segment. Hence Γ_2 consists of the (possibly degenerate) line segments contained in the common boundaries of the three numerical ranges. If $\Gamma \equiv \Gamma_1 \cup \Gamma_2$ consists of at most two (possibly degenerate) line segments, then we say that $W(A)$ has property Λ. Evidently, the disjointness of $W(B)$ and $W(C)$ implies that property Λ holds since Γ_1 consists of exactly two line segments and Γ_2 is empty.

Applying similar arguments as in the proof of Theorem 2.2.2, property Λ is enough to establish the equality $k(A) = k_1(B) + k_1(C)$. Hence we have the following theorem.

Theorem 2.2.6. *Let $A = B \oplus C$, where B and C are n-by-n and m-by-m matrices, respectively. If $W(A)$ has property Λ, then $k(A) = k_1(B) + k_1(C) \leq k(B) + k(C)$. In this case, $k(A) = k(B) + k(C)$ if and only if $k_1(B) = k(B)$ and $k_1(C) = k(C)$. In particular, $k(A) = m + n$ if and only if $k_1(B) = k(B) = n$ and $k_1(C) = k(C) = m$.*

2.3 Applications and discussions

The first application of our results in Section 2.2 is a generalization of Lemma 2.2.3. Indeed, we are able to determine the value of $k(A)$ for $A = B \oplus C$ with normal C.

Proposition 2.3.1. *Let $A = B \oplus C$, where C is an m-by-m normal matrix. Then $k(A) = k_1(B) + k_1(C)$. In this case, $k(A) = k(B) + k(C)$ if and only if $k_1(B) = k(B)$ and $k_1(C) = k(C)$. In particular, if $C = cI_m$ for some scalar c, then $k(A) = k_1(B) + k_1(cI_m)$.*

Proof. Let the normal C be unitarily similar to $\oplus_{j=1}^{m} [c_j]$. By [19, Lemma 2.9], we may assume that all the c_j's lie in $\partial W(A)$. This shows that $k_1(C) = m$ immediately. On the other hand, we also obtain $k(A) = k_1(B) + m$ by Lemma 2.2.3. Hence the asserted equality $k(A) = k_1(B) + k_1(C)$ has been proven. The remaining assertions hold trivially by this equality. ■

An easy corollary of Proposition 2.3.1 is to determine when $k(A)$ equals the size of A for a matrix $A = B \oplus C$ with normal C.

Corollary 2.3.2. *Let $A = B \oplus C$, where B is an n-by-n matrix and C is an m-by-m normal matrix. Then $k(A) = n + m$ if and only if $k_1(B) = n$ and $k_1(C) = m$. Assume, moreover, that $\dim H_\eta = 1$ for all $\eta \in \partial W(B)$. Then $k(A) = n + m$ if and only if $k_1(B) = n \leq 2$ and $k_1(C) = m$.*

Proof. By Proposition 2.3.1, it is clear that $k(A)$ equals the size of A if and only if $k_1(B)$ and $k_1(C)$ equal the sizes of B and C, respectively. In this case, the assumption on H_η implies that $k_1(B) = n \leq 2$ by [19, Proposition 2.10]. This completes the proof. ■

For a matrix A of the form $B \oplus C$, we recall the decomposition $\Gamma = \Gamma_1 \cup \Gamma_2$ at the end of Section 2.2, where $\Gamma_1 = \partial W(A) \setminus (\partial W(B) \cup \partial W(C))$ and $\Gamma_2 = \partial W(A) \cap \partial W(B) \cap \partial W(C)$. The next proposition gives a lower bound for $k(A)$.

Proposition 2.3.3. *Let $A = B \oplus C$ be an n-by-n $(n \geq 3)$ matrix. Then Γ is empty if and only if the numerical range of one summand is contained in the interior of the numerical range of the other. In particular, if Γ is nonempty, then $k(A) \geq 3$.*

Proof. If $\Gamma = \Gamma_1 \cup \Gamma_2$ is empty, then both Γ_1 and Γ_2 are empty. Since Γ_1 is empty, $\partial W(A)$ is contained in $\partial W(B) \cup \partial W(C)$. This implies that $W(B) \cap W(C)$ is nonempty, and thus $W(B) = W(C)$, $W(B) \subseteq \operatorname{int} W(C)$ or $W(C) \subseteq \operatorname{int} W(B)$. Moreover, $\Gamma_2 = \phi$ implies that $W(B) \neq W(C)$. With this, we conclude that either $W(B) \subseteq \operatorname{int} W(C)$ or $W(C) \subseteq \operatorname{int} W(B)$. The converse is obvious. Hence we have proved the first assertion. Let Γ be nonempty, that is, either Γ_1 or Γ_2 is nonempty. If Γ_1 is nonempty, then there is a line segment on the boundary of $W(A)$. This shows that $k(A) \geq 3$ by [19, Corollary 2.5]. On the other hand, if Γ_2 is nonempty, then there is a (possibly degenerate) line segment on the common boundaries of the three numerical ranges $W(A)$, $W(B)$ and $W(C)$. Using [19, Corollary 2.5] again, we may assume that the line segment is degenerate, say, to $\{\xi\}$. This implies immediately that $\dim_\xi H(A) \geq 2$. Thus $k(A) \geq 3$ by [19, Proposition 2.4]. ■

As an application, when A is reducible, the next corollary gives a necessary and sufficient condition for $k(A) = 2$.

Corollary 2.3.4. *Let $A = B \oplus C$ be an n-by-n $(n \geq 3)$ matrix. Then $k(A) = 2$ if and only if either $k(B) = 2$ and $W(C) \subseteq \operatorname{int} W(B)$, or $k(C) = 2$ and $W(B) \subseteq \operatorname{int} W(C)$.*

Proof. If $k(A) = 2$, then Proposition 2.3.3 shows that Γ is empty, and thus the numerical range of one summand, say, B is contained in the interior of the numerical range of C. Hence $k(C) = 2$ by [19, Lemma 2.9]. The converse is obvious by [19, Lemma 2.9] again. ■

The following proposition determines exactly when $k(A)$ equals the size of A for an irreducible matrix A. It is also stated in [2, Theorem 7] while the proof there is different from ours.

Proposition 2.3.5. *Let A be an n-by-n $(n \geq 3)$ irreducible matrix. Then $k(A) = n$ if and only if $\partial W(A)$ contains a line segment l and there are n points (not necessarily distinct) in $l \cup (\partial W(A) \cap L)$, where L is the supporting line parallel to l such that their corresponding unit vectors form an orthonormal basis for $\mathbb{C}^n$.*

Proof. We need only prove the necessity. Assume that A is an n-by-n $(n \geq 3)$ irreducible matrix with $k(A) = n$. If $\partial W(A)$ contains no line segment, then $\dim H_\xi = \dim E_{\xi,l} \leq n/2$ for all $\xi \in \partial W(A)$ by [19, Proposition 2.2]. If n is odd, say, $n = 2m+1$, then $\dim H_\xi = \dim E_{\xi,l} \leq m$ for all $\xi \in \partial W(A)$. Since $k(A) = n$, it follows from [19, Theorem 2.7] that A is reducible, which is absurd. If n is even, say, $n = 2m$, then $m \geq 2$ by our assumption that $n \geq 3$. Since $k(A) = n$ and $\partial W(A)$ contains no line segment, A is unitarily similar to a matrix of the form

$$\begin{bmatrix} \xi I_m & e^{i\theta} D \\ -e^{i\theta} D^* & \eta I_m \end{bmatrix}$$

by [19, Theorem 2.7], where $\dim H_\xi = \dim H_\eta = m$. Let $D = USV$ be the singular value decomposition of D, where U and V are unitary and $S = \mathrm{diag}\,(s_1, ..., s_m)$ is a diagonal matrix with $s_j \geq 0$, $j = 1, ..., m$. Then

$$\begin{bmatrix} U^* & 0 \\ 0 & V \end{bmatrix} \begin{bmatrix} \xi I_m & e^{i\theta} D \\ -e^{i\theta} D^* & \eta I_m \end{bmatrix} \begin{bmatrix} U & 0 \\ 0 & V^* \end{bmatrix} = \begin{bmatrix} \xi I_m & e^{i\theta} S \\ -e^{i\theta} S & \eta I_m \end{bmatrix}$$

and the latter is unitarily similar to

$$\bigoplus_{j=1}^{m} \begin{bmatrix} \xi & e^{i\theta} s_j \\ -e^{i\theta} s_j & \eta \end{bmatrix}.$$

This contradicts the irreducibility of A. Hence $\partial W(A)$ must contain a line segment. We then apply [19, Theorem 2.7] again to complete the proof. ■

An easy corollary of Proposition 2.3.5 is the following upper bound for $k(A)$. This was given in [19, Proposition 2.10]. Here we give a simpler proof.

Corollary 2.3.6. *If A is an n-by-n ($n \geq 3$) matrix with* $\dim H_\xi = 1$ *for all* $\xi \in \partial W(A)$, *then* $k(A) \leq n - 1$.

Proof. Assume that $k(A) = n$. It suffices to consider that A is reducible; this is because if otherwise, then Proposition 2.3.5 shows that $\partial W(A)$ contains a line segment, which contradicts the assumption on H_ξ. Let $A = B \oplus C$. Then our assumption on H_ξ implies that Γ is empty. By Proposition 2.3.3, we obtain that the numerical range of one summand is contained in the interior of the numerical range of the other summand. It follows from [19, Lemma 2.9] that the value of $k(A)$ equals $k(B)$ or $k(C)$. Thus $k(A) \leq n - 1$ as asserted. ■

We now combine Proposition 2.3.1, Corollary 2.3.2, Corollary 2.3.4, and Proposition 2.3.5 to determine the value of $k(A)$ for any 4-by-4 reducible matrix A. Corollary 2.3.4 shows exactly when the value of $k(A)$ equals two. By Proposition 2.3.1, Corollary 2.3.2 and Proposition 2.3.5, we get a necessary and sufficient condition for the value of $k(A)$ to be equal to four. In other words, the value of $k(A)$ can be determined completely for any 4-by-4 reducible matrix A. To do this, we note that a 4-by-4 reducible matrix A can be written, after a unitary similarity, as (i) $A = B \oplus [c]$, where B is a 3-by-3 irreducible matrix and c is a complex number, (ii) $A = B \oplus [c]$, where B is a 3-by-3 reducible matrix and c is a complex number, or (iii) $A = B \oplus C$, where B and C are 2-by-2 irreducible matrices. Proposition 2.3.7 below is to deal with case (i).

Recall that for a 3-by-3 irreducible matrix A, $W(A)$ is of one of the following shapes (cf. [9]): an elliptic disc, the convex hull of a heart-shaped region, in which case $\partial W(A)$ contains a line segment, and an oval region.

Proposition 2.3.7. *Let $A = B \oplus [c]$, where B is a 3-by-3 irreducible matrix and c is a complex number. Then $k(A) = 4$ if and only if $c \notin \operatorname{int} W(B)$ and $\{a_1, a_2, b\} \subseteq \partial W(A)$, where $W(B)$ is the convex hull of a heart-shaped region, in which case $\partial W(B)$ contains a line segment $[a_1, a_2]$ contained in the supporting line L_1 of $W(B)$ and L_2 is the supporting line of $W(B)$ passing through b and parallel to L_1.*

Proof. By Corollary 2.3.2, we see that $k(A) = 4$ is equivalent to $k_1(B) = 3$ and $k_1([c]) = 1$. Since a necessary and sufficient condition for $k_1([c]) = 1$ is that $c \notin \operatorname{int} W(B)$, it remains to show that $k_1(B) = 3$ if and only if $\{a_1, a_2, b\} \subseteq \partial W(A)$ and $W(B)$ satisfies the asserted properties. If $k_1(B) = 3$, then $k(B) = 3$. Hence it follows from Proposition 2.3.5 that $\partial W(A)$ contains $\{a_1, a_2, b\}$, and $W(B)$ is as asserted. The converse is trivial. ■

For case (ii), let $A = B \oplus [c]$, where B is a 3-by-3 reducible matrix. After a unitary similarity, B can be written as $C \oplus [b]$, where C is a 2-by-2 matrix, so that $k(A) = k_1(C) + k_1([b] \oplus [c])$ by Proposition 2.3.1. The following proposition gives a necessary and sufficient condition for $k(A)$ to be equal to four.

Proposition 2.3.8. *Let $A = C \oplus [b] \oplus [c]$, where C is a 2-by-2 matrix, and b and c are complex numbers. Then $k(A) = 4$ if and only if both b and c are in $\partial W(A)$ and $k_1(C) = 2$.*

Proof. By Corollary 2.3.2, it is obvious that $k(A) = 4$ if and only if $k_1(C) = 2$ and $k_1([b] \oplus [c]) = 2$. Moreover, it is also clear that $k_1([b] \oplus [c]) = 2$ is equivalent to both of b and c being in $\partial W(A)$. Hence the proof is complete. ■

To prove for case (iii), let $A = B \oplus C$, where B and C are 2-by-2 irreducible matrices. Since $W(A)$ is the convex hull of the union of the two elliptic discs $W(B)$ and $W(C)$, either $W(B)$ equals $W(C)$, or Γ consists of at most four (possibly degenerate) line segments. With this, we are now ready to give a necessary and sufficient condition for $k(A) = 4$.

Proposition 2.3.9. *Let $A = B \oplus C$, where B and C are 2-by-2 irreducible matrices. Then $k(A) = 4$ if and only if Γ consists of at least three line segments (including the possibly degenerate ones), or Γ consists of exactly two (possibly degenerate) line segments such that $k_1(B) = k_1(C) = 2$.*

Proof. If Γ consists of more than four (possibly degenerate) line segments, then the two elliptic discs $W(B)$ and $W(C)$ are identical. Hence $k(A) = 4$ by direct computations. If Γ consists of four or three (possibly degenerate) line segments, then the endpoints of the major axes of the two elliptic discs $W(B)$ and $W(C)$ are in $\partial W(A)$. Hence $k(A) = 4$. If Γ consists of exactly two (possibly degenerate) line segments such that $k_1(B) = k_1(C) = 2$, then $k(A) = 4$ by Theorem 2.2.6. Therefore we have proved the sufficient condition for $k(A) = 4$. Next assume that $k(A) = 4$ and either Γ consists of exactly two (possibly degenerate) line segments such that the equalities $k_1(B) = k_1(C) = 2$ fail, or Γ consists of at most one (possibly degenerate) line segment. Since property Λ holds in each case, we must have $k_1(B) = k_1(C) = 2$ by Theorem 2.2.6. This shows that we need only consider the latter. If Γ consists of exactly one (possibly degenerate) line segment, then Γ_1 is empty and Γ_2 is a singleton. Hence we may assume that $W(B)$ is contained in $W(C)$ and the intersection of $W(B)$ and $W(C)$ is Γ. This shows that $k_1(B) = 1$ and $k_1(C) = 2$, which is a contradiction. If Γ is empty, then it follows from Proposition 2.3.3 that the numerical range of one summand, say, B is contained in the interior of the numerical range of the other summand C. By Corollary 2.3.4 and [5, Lemma 4.1], we see that $k(A) = k(C) = 2$, which is absurd. This completes the proof. ∎

As a final application of Theorem 2.2.6, it is obvious that the convex hull of the union of $W(A)$ and $W(A+aI_n)$ has property Λ for any $a \neq 0$. Hence we obtain the following proposition.

Proposition 2.3.10. *Let A be an n-by-n matrix and a be a nonzero complex number. Then $k(A \oplus (A+aI_n)) = k_1(A) + k_1(A+aI_n)$. In this case, $k(A \oplus (A+aI_n)) = 2k(A)$ if and only if $k_1(A+aI_n) = k_1(A) = k(A)$.*

We conclude this paper by stating the following open questions concerning this topic. Is it true that the equality $k(A) = k_1(B) + k_1(C)$ holds for a matrix A of the form $B \oplus C$ even if property Λ fails? We note that although property Λ fails, the mentioned formula may still be correct (cf. Proposition 2.3.1). Another natural example of the failure of property Λ is that both $W(B)$ and $W(C)$ have the same numerical range. Is it true that $k(B \oplus C) = k(B) + k(C)$ in this case? In particular, can we determine the value of $k(A \oplus A)$ (cf. Proposition 2.3.10)? The following proposition gives a partial answer for $k(A \oplus A)$ if we assume, in addition, that $\dim H_\xi = 1$ for all $\xi \in \partial W(A)$.

Proposition 2.3.11. *If A is an n-by-n matrix with $\dim H_\xi = 1$ for all $\xi \in \partial W(A)$, then*

$$k\left(\bigoplus_{j=1}^{m} A\right) = m \cdot k(A).$$

Proof. Obviously, the inequality $k\left(\oplus_{j=1}^{m} A\right) \geq m \cdot k(A)$ holds. To prove the reversed inequality, we consider, for convenience, the case $m = 2$. Let $\xi_1 \in \partial W(A \oplus A)$. Then $\dim H_{\xi_1}(A \oplus A) = 2$ by our assumption on $H_\xi(A)$. Hence the subspace $H_{\xi_1}(A \oplus A)$ is spanned by the two unit vectors $x_1 \oplus 0$ and $0 \oplus x_1$, where $\xi_1 = \langle Ax_1, x_1 \rangle$. Let z_1 be a unit vector in $H_{\xi_1}(A \oplus A)$. Then $z_1 = (\alpha_1 x_1 \oplus \alpha_2 x_1) / \sqrt{|\alpha_1|^2 + |\alpha_2|^2}$, where α_1 and α_2 are in $\mathbb{C}$. Similarly for $\xi_2 \in \partial W(A \oplus A)$. That is, the subspace $H_{\xi_2}(A \oplus A)$

is spanned by the two unit vectors $x_2 \oplus 0$ and $0 \oplus x_2$, where $\xi_2 = \langle Ax_2, x_2 \rangle$. Moreover, if z_2 is a unit vector in $H_{\xi_2}(A \oplus A)$, then $z_2 = (\beta_1 x_2 \oplus \beta_2 x_2) / \sqrt{|\beta_1|^2 + |\beta_2|^2}$, where β_1 and β_2 are in $\mathbb{C}$. Obviously, the orthogonality of z_1 and z_2 is equivalent to $\left(\alpha_1 \bar{\beta}_1 + \alpha_2 \bar{\beta}_2\right) \langle x_1, x_2 \rangle = 0$, that is,

$$\left\langle \begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix}, \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} \right\rangle \langle x_1, x_2 \rangle = 0.$$

This shows that $k(A \oplus A) \leq 2k(A)$ immediately by the definition of $k(A)$.

For general m, a similar argument as above yields that

$$\left\langle \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_m \end{bmatrix}, \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_m \end{bmatrix} \right\rangle \langle x_1, x_2 \rangle = 0$$

for some scalars $\alpha_1, ..., \alpha_m$ and $\beta_1, ..., \beta_m$, where x_1 and x_2 are similarly defined. Since the dimension of $\mathbb{C}^m$ is m, the number of vectors of the form $[\alpha_1, ..., \alpha_m]^T$ which are orthogonal to each other is at most m. We infer from this and the above equality that the reversed inequality $k\left(\oplus_{j=1}^m A\right) \leq m \cdot k(A)$ holds. Therefore we have the asserted equality. ■

At the end of this section, we apply Proposition 2.3.11 to the quadratic matrices. Recall that an n-by-n quadratic matrix A is unitarily similar to a matrix of the form

$$aI_{n_1} \oplus bI_{n_2} \oplus \begin{bmatrix} aI_{n_3} & D \\ 0 & bI_{n_3} \end{bmatrix},$$

where $n_1, n_2, n_3 \geq 0$, $n_1 + n_2 + n_3 = n$, $D > 0$, and $a, b \in \sigma(A)$ (cf. [18, Theorem 2.1]).

Corollary 2.3.12. *If A is an n-by-n quadratic matrix of the above form and D is not missing, then* $k(A) = 2 \cdot \#\left(\{\lambda \in \sigma(D) : \lambda = \|D\|\}\right)$.

Proof. If $D > 0$, then D is unitarily similar to diag $(d_1, ..., d_{n_3})$, where $d_1 = \cdots = d_p = \|D\| \equiv d > d_{p+1} \geq \cdots \geq d_{n_3} \geq 0$ $(1 \leq p \leq n_3)$. Hence A is unitarily similar to a matrix of the form $aI_{n_1} \oplus bI_{n_2} \oplus_{j=1}^{p} B \oplus_{j=p+1}^{n_3} B_j$, where $n_1 + n_2 + 2n_3 = n$,

$$B \equiv \begin{bmatrix} a & d \\ 0 & b \end{bmatrix}, \text{ and } B_j \equiv \begin{bmatrix} a & d_j \\ 0 & b \end{bmatrix}, \ j = p+1, \ldots, n_3.$$

Since the set $\{a, b\}$ and all of the numerical ranges $W(B_j)$, $j = p+1, \ldots, m$, are contained in the interior of $W(B)$, it follows from [19, Lemma 2.9] that $k(A) = k(\oplus_{j=1}^{p} B)$. Since $\dim H_\xi(B) = 1$ for all $\xi \in \partial W(B)$, we have $k(A) = p \cdot k(B)$ by Proposition 2.3.11. Obviously, $k(B) = 2$ by [5, Lemma 4.1]. Thus $k(A) = 2p$ as asserted. ∎

We remark that in the preceding proof the equality $k(\oplus_{j=1}^{p} B) = 2p$ can also be established directly. Indeed, the inequality $k(\oplus_{j=1}^{p} B) \geq 2p$ holds trivially and we can infer from [5, Lemma 4.1] that $k(\oplus_{j=1}^{p} B) = 2p$.

3 Gau-Wu numbers of nonnegative matrices

3.1 Introduction

In Section 3.2 below, we first consider a matrix A of the form

$$\begin{bmatrix} 0 & A_1 & & 0 \\ & 0 & \ddots & \\ & & \ddots & A_{m-1} \\ A_m & & & 0 \end{bmatrix} \quad (m \geq 2),$$

where the diagonal zeros are zero square matrices. In this case, we obtain that $k(A)$ has a lower bound m (Proposition 3.2.1) if A has a boundary vector $x = \oplus_{j=1}^{m} x_k$, that is, $\langle Ax, x\rangle \in \partial W(A)$, with all component vectors x_j having the same norm $1/\sqrt{m}$. Next, we study a nonnegative matrix A of the above form with irreducible real part and $A_m = 0$. Proposition 3.2.3 yields that $k(A) \leq m - 1$. Moreover, with the help of [19], we are able to give necessary and sufficient conditions for such a matrix A with the value of $k(A)$ equal to $m-1$ (Theorem 3.2.4). Finally, we also consider a nonnegative matrix A of the above form with irreducible real part. Example 3.2.6 shows that no analogous results hold for such an A. In Section 3.3, we consider more special nonnegative matrices, namely, the doubly stochastic matrices. It can be proven that $k(A)$ equals 3 for any 3-by-3 doubly stochastic matrix (Proposition 3.3.2). Moreover, for a 4-by-4 doubly stochastic matrix A, we determine the value of $k(A)$ completely and give the description of its numerical range $W(A)$ (Propositions 3.3.4 and 3.3.5). For general n, we obtain the lower bound of $k(A)$ for an n-by-n doubly stochastic matrix A (Theorems 3.3.6 and 3.3.7). In particular, for an n-by-n irreducible doubly stochastic matrix A, we obtain a necessary and sufficient condition for $k(A)$ to be equal to this lower bound (Theorem 3.3.7).

We end this section by fixing some notations. For any finite matrix A, its trace, determinant, and spectral radius are denoted by $\operatorname{tr} A$, $\det A$, and $r(A)$, respectively. The number m of eigenvalues z of A with $|z| = r(A)$ is called the *index of imprimitivity* of A.

3.2 Nonnegative block shift matrix

We start by reviewing a couple of basic facts on a *block shift matrix*. Recall that a block shift matrix A is one of the form

$$\begin{bmatrix} 0_1 & A_1 & & 0 \\ & 0_2 & \ddots & \\ & & \ddots & A_{m-1} \\ A_m & & & 0_m \end{bmatrix} (m \geq 2),$$

where the diagonal zeros 0_j $(j = 1, ..., m)$ are zero square matrices. Let $\varphi = 2\pi/m$. Then it is easy to see that the numerical range $W(A)$ is an m-symmetric compact convex region since $U^*AU = e^{i\varphi}A$, where U is a unitary matrix of the form

$$\begin{bmatrix} e^{i\varphi}I_1 & 0 & & 0 \\ & e^{2i\varphi}I_2 & \ddots & \\ & & \ddots & 0 \\ 0 & & & e^{mi\varphi}I_m \end{bmatrix},$$

where the diagonal identity matrix I_j is of the same size as the corresponding 0_j $(j = 1, ..., m)$. Let $\langle Ax, x\rangle$ be a boundary point of $W(A)$, where $x = \oplus_{k=1}^m x_j$ is a unit vector. We define $x_{0\phi} = x$ and $x_{j\varphi} = \oplus_{k=1}^m e^{i(k-1)j\varphi}x_k$ for $j = 1, ..., m-1$. With these notations, we can give a lower bound for $k(A)$.

Proposition 3.2.1. *Let A be a block shift of the above form with the corresponding notations as above. Then $\|x_k\|$ is equal to $1/\sqrt{m}$ for all $k = 1, ..., m$ if and only if the vectors $x_{p\varphi}$, $0 \leq p \leq m-1$, are orthonormal. In this case, we have $k(A) \geq m$.*

Proof. Assume that $\langle x_{p\varphi}, x_{q\varphi}\rangle = 0$ for $0 \le p \ne q \le m-1$. This is equivalent to the equation

$$\|x_1\|^2 + e^{i(p-q)\varphi}\|x_2\|^2 + \cdots + e^{i(m-1)(p-q)\varphi}\|x_m\|^2 = 0$$

for $0 \le p \ne q \le m-1$. That is, $e^{i\varphi}, ..., e^{i(m-1)\varphi}$ are the roots of the polynomial

$$\|x_1\|^2 + \|x_2\|^2 t + \cdots + \|x_m\|^2 t^{m-1}.$$

Hence each $\|x_k\|$ is equal to $1/\sqrt{m}$ for $k = 1, ..., m$ by comparing the coefficients of the above polynomial with those of $\|x_m\|^2 \prod_{j=1}^{m-1}(t - e^{ij\varphi})$. Conversely, if $\|x_k\|$ is equal to $1/\sqrt{m}$ for all $k = 1, ..., m$, then it is a routine matter to check that $x_{p\varphi}$ and $x_{q\varphi}$ are orthonormal for $0 \le p \ne q \le m-1$. Clearly, in this case, $k(A)$ has a lower bound m. ■

Recall that the *numerical radius* $\omega(A)$ of a matrix A is the quantity $\max\{|z| : z \in W(A)\}$. For a nonnegative matrix with irreducible real part, [16, Lemma 1] says that, for $\omega(A)e^{i\theta}$ in $W(A)$, where θ is a real number with $e^{i\theta} \ne 1$, (a) if θ is an irrational multiple of 2π, then A is permutationally similar to a matrix of the form

(1)
$$\begin{bmatrix} 0 & A_1 & & 0 \\ & 0 & \ddots & \\ & & \ddots & A_{m-1} \\ 0 & & & 0 \end{bmatrix} (m \ge 2),$$

where the diagonal zeros are zero square matrices, and, in particular, $W(A)$ is a circular disc centered at the origin, and (b) if θ is a rational multiple of 2π, say, $\theta = 2\pi p/q$, where p and q are relatively prime integers and $q \ge 2$, then A is permutationally similar to

(2)
$$\begin{bmatrix} 0 & A_1 & & 0 \\ & 0 & \ddots & \\ & & \ddots & A_{q-1} \\ A_q & & & 0 \end{bmatrix} (m \ge 2),$$

and, in particular, $W(A) = e^{2\pi i/q}W(A)$.

The following lemma is a generalization of [19, Lemma 3.6], which is useful for the proof of Proposition 3.2.3. Recall that a vector x with positive components, denoted by $x \succ 0$, is called *positive*.

Lemma 3.2.2. *Let A be an n-by-n ($n \geq 2$) nonnegative matrix of the form* (1) *with irreducible real part and $m \geq 2$. Then the following hold:*

(a) $W(A) = \{z \in \mathbb{C} : |z| \leq \omega(A)\}$.

(b) *There is a unique positive vector $x = x_1 \oplus \cdots \oplus x_m \in \mathbb{C}^n$ such that $\langle Ax, x\rangle = \omega(A)$.*

(c) *For any $a = \omega(A)e^{i\theta}, \theta \in [0, 2\pi)$, in $\partial W(A)$, if $x_\theta = x_1 \oplus e^{i\theta}x_2 \oplus \cdots \oplus e^{i(m-1)\theta}x_m$, then $a = \langle Ax_\theta, x_\theta\rangle$ and H_a is generated by x_θ.*

(d) *Let $a_j = \omega(A)e^{i\theta_j}$ ($\theta_j \in [0, 2\pi)$), $j = 1, 2$, be two points in $\partial W(A)$ with the corresponding unit vector x_{θ_j}. Then x_{θ_1} and x_{θ_2} are orthogonal to each other if and only if $e^{i(\theta_1-\theta_2)}$ is a zero of the polynomial $\|x_1\|^2 + \|x_2\|^2 t + \cdots + \|x_m\|^2 t^{m-1}$.*

Proof. Since $U_\theta^* A U_\theta = e^{i\theta}A$ for any θ, where $U_\theta = \oplus_{k=1}^m e^{i(k-1)\theta} I_k$, that is, A is unitarily similar to $e^{i\theta}A$ for any θ, (a) follows immediately. (b) is a consequence of [11, Proposition 3.3]. To prove (c), note that

$$a = \omega(A)e^{i\theta} = \langle e^{i\theta}Ax, x\rangle = \langle U_\theta^* A U_\theta x, x\rangle = \langle A(U_\theta x), (U_\theta x)\rangle = \langle Ax_\theta, x_\theta\rangle,$$

which shows that x_θ is in H_a. That dim $H_a = 1$ is by [11, Corollary 3.10]. Thus H_a is generated by x_θ. (d) follows from the fact that $\langle x_{\theta_1}, x_{\theta_2}\rangle = \sum_{k=1}^m e^{i(k-1)(\theta_1-\theta_2)} x_k^2$. This completes the proof. ■

Thus, for a nonnegative matrix A of the form (1) with irreducible real part, $k(A)$ equals the maximum number of $\theta_1, \ldots, \theta_k$ in $[0, 2\pi)$ for which $e^{i(\theta_j-\theta_l)}$ is a zero of $p(t) \equiv \|x_1\|^2 + \|x_2\|^2 t + \cdots + \|x_m\|^2 t^{m-1}$ for all j and l, $1 \leq j \neq l \leq k$. If $m = 2$, then the polynomial $p(t)$ has degree one. Hence $k(A) = 2$ if $m = 2$. The following proposition says that $k(A) \leq m - 1$ if $m \geq 3$.

Proposition 3.2.3. *Let A be an n-by-n ($n \geq 2$) nonnegative matrix of the form* (1)*, where $m \geq 3$, with irreducible real part and $\omega(A) = 1$. Then $k(A) \leq m - 1$.*

Proof. From our assumptions on A and Lemma 3.2.2 (b), there is a unique vector $x = x_1 \oplus \cdots \oplus x_m \in \mathbb{C}^n$ with positive x_j for all j such that $\langle Ax, x\rangle = 1$. Letting $k(A) = k$, we may assume, by the proof of [19, Theorem 3.10] that $\theta_0 = 0, \theta_1 = 2\pi/k, ..., \theta_{k-1} = 2(k-1)\pi/k$, so that x_{θ_j} and x_{θ_l} are orthogonal to each other for all j and l, $0 \leq j \neq l \leq k-1$. From Lemma 3.2.2 (d), this yields that $k(A) \leq m$ since the degree of $p(t)$ is $m-1$. Assume that $k(A) = m$. If m is odd, then the degree of the polynomial $p(t)$ is equal to the even $m-1$. We note that -1 is a zero of $p(t)$ and the zeros of the real polynomial $p(t)$ appear in conjugate pairs, which is a contradiction. Hence $k(A) \leq m-1$ for odd m. On the other hand, if m is even, then $\|x_1\| = \cdots = \|x_m\| = 1/\sqrt{m}$ by examining the coefficients of $p(t)$. From the assumption that $\omega(A) = 1$, we have $(\mathrm{Re}\ A)x = x$ by [11, Proposition 3.3]. That is,

$$
\begin{array}{ll}
\text{(i)} & (A_1/2)x_2 = x_1, \\
\text{(ii)} & (A_j^T/2)x_j + (A_{j+1}/2)x_{j+2} = x_{j+1} \quad \text{for } 1 \leq j \leq m-2, \text{ and} \\
\text{(iii)} & (A_{m-1}^T/2)x_{m-1} = x_m.
\end{array}
$$

Taking the transpose of (i) and then multiplying x_1 from right on both sides, we obtain $x_2^T(A_1^T/2)x_1 = \|x_1\|^2$. Next, multiplying (ii) on both sides by x_j^T, we have $x_j^T(A_j/2)x_{j+1} = 0$ if j is even and $x_j^T(A_j/2)x_{j+1} = \|x_j\|^2$ if j is odd, where $2 \leq j \leq m$. Similarly, we multiply (iii) on both sides by x_m^T to get $x_{m-1}^T(A_{m-1}/2)x_m = \|x_{m-1}\|^2$. These are the same as

$$
\begin{array}{ll}
\langle A_j x_{j+1}, x_j\rangle = 0 & \text{for } j = 2, 4, ..., m-2, \text{ and} \\
\langle A_j x_{j+1}, x_j\rangle = \|x_j\|^2 & \text{for } j = 1, 3, ..., m-1.
\end{array}
$$

This implies that $A_2 = A_4 = \cdots = A_{m-2} = 0$, which contradicts the assumption of the irreducibility of the real part of A. Hence $k(A) \leq m-1$ for even m. We complete the proof. ∎

Our next result is a characterization of n-by-n ($n \geq 2$) nonnegative matrices A of the form (1) with irreducible real part for which $k(A) = m-1$, where $m \geq 3$.

Theorem 3.2.4. *Let A be an n-by-n ($n \geq 2$) nonnegative matrix of the form (1), where $m \geq 3$, with irreducible real part and $\omega(A) = 1$. Then the following are equivalent:*

(a) $k(A) = m-1$.

(b) *there is an $\alpha > 0$ satisfying $p(-\alpha) = 0$, where*

$$p(t) = \begin{cases} \frac{1}{2(m-1)} + \frac{1}{m-1}t + \cdots + \frac{1}{m-1}t^{m-2} + \frac{1}{2(m-1)}t^{m-1} & \text{if } m \text{ is even, or} \\ \frac{1}{2(1+\alpha)(m-1)} + \frac{1}{m-1}t + \cdots + \frac{1}{m-1}t^{m-2} + \frac{\alpha}{2(1+\alpha)(m-1)}t^{m-1} & \text{if } m \text{ is odd.} \end{cases}$$

(c) *If $x = x_1 \oplus \cdots \oplus x_m \in \mathbb{C}^n$ is the (unique) positive unit vector such that $\langle Ax, x\rangle = 1$, then its component vectors satisfy*

(i) $\|x_1\| = \|x_m\| = 1/\sqrt{2(m-1)}$ *and* $\|x_j\| = 1/\sqrt{m-1}$, $2 \leq j \leq m-1$, *if m is even,*

(ii) $\|x_1\| = 1/\sqrt{(1+\alpha)(m-1)}$, $\|x_m\| = \sqrt{\alpha/((1+\alpha)(m-1))}$, *and* $\|x_j\| = 1/\sqrt{m-1}$, $2 \leq j \leq m-1$ *for some $\alpha > 0$, if m is odd.*

Proof. If $k(A) = m-1$, then, for even m, zeros of the corresponding polynomial, denoted by $p_e(t)$, are exactly -1 and those points which are equally distributed over the unit circle by the proof of Proposition 3.2.3. In other words, the polynomial $(t+1)\prod_{j=1}^{m-2}(t-\omega^j)$, where $\omega = e^{2\pi i/(m-1)}$, has the same zeros as $p_e(t)$. This implies that $p_e(t) = \frac{1}{2(m-1)} + \frac{1}{m-1}t + \cdots + \frac{1}{m-1}t^{m-2} + \frac{1}{2(m-1)}t^{m-1}$ for even m since $\|x\| = \|x_1\|^2 + \cdots + \|x_m\|^2 = 1$. Conversely, if $p_e(t) = \frac{1}{2(m-1)} + \frac{1}{m-1}t + \cdots + \frac{1}{m-1}t^{m-2} + \frac{1}{2(m-1)}t^{m-1}$ for even m, then it is clear that $k(A) = m-1$. On the other hand, for odd m zeros of the corresponding polynomial, denoted by $p_o(t)$, are exactly $-\alpha$ and those points which are equally distributed over the unit circle and $-\alpha$ by the proof of Proposition 3.2.3. This α must be positive since the coefficients of $p_o(t)$ are nonnegative. That is, the polynomial $(t+\alpha)\prod_{j=1}^{m-2}(t-\omega^j)$, where $\omega = e^{2\pi i/(m-1)}$, has the same zeros as

$p_o(t)$. It follows that $p_o(t) = \frac{1}{2(1+\alpha)(m-1)} + \frac{1}{m-1}t + \cdots + \frac{1}{m-1}t^{m-2} + \frac{\alpha}{2(1+\alpha)(m-1)}t^{m-1}$ for odd m since $\|x\| = \|x_1\|^2 + \cdots + \|x_m\|^2 = 1$. The converse is obvious. This proves the equivalence of (a) and (b). The equivalence of (b) and (c) is obvious by the above arguments. ■

An easy consequence of Theorem 3.2.4 is that we can give a necessary condition for $k(A) = m - 1$ by dealing with the norms of blocks, which are similar to [19, Theorem 3.10]

Corollary 3.2.5. *Let A be an n-by-n $(n \geq 2)$ nonnegative matrix of the form* (1)*, where $m \geq 3$, with irreducible real part and $\omega(A) = 1$. If $k(A) = m - 1$, then either*

(a) *m is even, $\|A_1\| = \|A_{m-1}\| \geq \sqrt{2}$ and $\|A_2\| = \cdots = \|A_{m-2}\| \geq 1$, or*

(b) *m is odd, $\|A_1\| \geq 2/\sqrt{1+\alpha}$, $\|A_{2j}\| \geq 2\alpha/(1+\alpha)$, $\|A_{2j+1}\| \geq 2/(1+\alpha)$ for $1 \leq j \leq (n-3)/2$, and $\|A_{m-1}\| \geq 2\sqrt{\alpha/(1+\alpha)}$ for some $\alpha > 0$.*

Proof. Let $k(A) = m - 1$. If $\omega(A) = 1$, then $(\mathrm{Re}\, A)x = x$ by [11, Proposition 3.3] or, equivalently,

$$\begin{aligned}
&\text{(i)} \quad (A_1/2)x_2 = x_1, \\
&\text{(ii)} \quad (A_j^T/2)x_j + (A_{j+1}/2)x_{j+2} = x_{j+1} \quad \text{for } 1 \leq j \leq m-2, \text{ and} \\
&\text{(iii)} \quad (A_{m-1}^T/2)x_{m-1} = x_m.
\end{aligned}$$

Assume first that m is even. Then after some computations, we obtain that $\langle A_j x_{j+1}, x_j \rangle = 1/(m-1)$ for $1 \leq j \leq m-1$. This along with Theorem 3.2.4 (c) proves case (a). Similar arguments apply for odd m. We complete the proof. ■

According to the above discussions and [19, Section 3], it is natural to ask whether $k(A) \leq q$ holds for a matrix A of the form (2). The following gives a counterexample.

Example 3.2.6. Let

$$A = \begin{bmatrix} 0 & B & 0 \\ 0 & 0 & B \\ B & 0 & 0 \end{bmatrix}, \text{where } B = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}.$$

It is easy to check that A is a nonnegative normal matrix with irreducible real part and $\sigma(A) = \{1, \omega, \cdots, \omega^5\}$, where $\omega = e^{2\pi i/6}$. Hence $k(A) = 6 \nleq 3$ by Proposition 2.2.1.

In the next section, we consider the numerical ranges of certain special nonnegative matrices, namely, those of doubly stochastic matrices.

3.3 Doubly stochastic matrix

We recall that a nonnegative matrix is doubly stochastic if its row sums and column sums are all equal to one. Properties of such matrices were studied in [15]. The following lemma gives some basic properties of such matrices. We omit its easy proofs.

Lemma 3.3.1. *Let A be a doubly stochastic matrix. Then*

(a) 1 *is an eigenvalue of A with corresponding eigenvector* $[1, ..., 1]^T$,

(b) *the norm, spectral radius, and numerical radius of A are equal to* 1, *and*

(c) *A is permutationally similar to a direct sum of irreducible doubly stochastic matrices.*

For a matrix A of the form $B \oplus C$, we recall the decomposition $\Gamma = \Gamma_1 \cup \Gamma_2$ at the end of Section 2.2, where $\Gamma_1 = \partial W(A) \setminus (\partial W(B) \cup \partial W(C))$ and $\Gamma_2 = \partial W(A) \cap \partial W(B) \cap \partial W(C)$. Based on the properties in the above lemma and Proposition 2.3.3, we are able to determine the value of $k(A)$ for any 3-by-3 doubly stochastic matrix.

Proposition 3.3.2. *Let A be a 3-by-3 doubly stochastic matrix. Then $k(A) = 3$.*

Proof. Let A be a 3-by-3 doubly stochastic matrix. Then, by Lemma 3.3.1 (a) and (b), 1 is a reducing eigenvalue of A. This implies that Γ is nonempty. By Proposition 2.3.3, we obtain that $k(A) = 3$. ■

Recall also that the number m of eigenvalues z of A with $\|z\| = r(A)$ is called the index of imprimitivity of A, and is denoted by $m(A)$. The following result is shown in [15, Corollary 1.5 and Theorem 2.1], which is useful for our later work on a 4-by-4 reducible doubly stochastic matrix.

Proposition 3.3.3. *Let $A = [a_{ij}]_{i,j=1}^{3}$ be a 3-by-3 irreducible doubly stochastic matrix.*

(a) *If $m(A) = 1$, then the numerical range $W(A)$ is the convex hull of the point 1 and a compact convex set K contained in the open unit disc $\mathbb{D}$, K is either a (possibly degenerate) elliptic disc with foci $(\operatorname{tr} A - 1 \pm \sqrt{(\operatorname{tr} A - 1)^2 - 4 \det A})/2 \in \mathbb{R}$ and minor axis of length $\sqrt{\det A - \det \operatorname{Re} A}$, or a (possibly degenerate) elliptic disc with foci $(\operatorname{tr} A - 1 \pm \sqrt{(\operatorname{tr} A - 1)^2 - 4 \det A})/2 \in \mathbb{C}$ and minor axis of length $(\sqrt{3|a_{12} - a_{21}|^2 + (\operatorname{tr} A - 1)^2 - 4 \det A})/2$, and*

(b) *if $m(A) \geq 2$, then A is normal with the numerical range $W(A)$ the regular 3-polygon with vertices $e^{2\pi i/3}, 0 \leq j < 3$.*

From Proposition 3.3.2, we have proven that for any 3-by-3 doubly stochastic matrix A the value of $k(A)$ is always equal to its size. The following proposition indicates that this still holds for any 4-by-4 reducible doubly stochastic matrix. Note that any reducible doubly stochastic matrix is permutationally similar to a direct sum of irreducible doubly stochastic matrices by Lemma 3.3.1. Applying this result, we have the following proposition.

Proposition 3.3.4. *Let A be a 4-by-4 reducible doubly stochastic matrix. Then*

$k(A) = 4$. *Moreover, the following hold:*

(a) *If A is permutationally similar to a direct sum of two 2-by-2 irreducible doubly stochastic matrices H_1 and H_2, then H_1 and H_2 are Hermitian, $W(A) = [2a - 1, 1]$, where $a = (\operatorname{tr} A - \sqrt{(\operatorname{tr} A)^2 - 4(\det A + \operatorname{tr} A - 1)})/4$, with $0 \le a < 1$, and $\sigma(A)$ consists of $1, 1, 2a-1$ and $2b-1$, where a is defined above and $b = (\operatorname{tr} A + \sqrt{(\operatorname{tr} A)^2 - 4(\det A + \operatorname{tr} A - 1)})/4$, with $0 \le a \le b < 1$.*

(b) *If A is permutationally similar to a direct sum $[1] \oplus B$, where B is a 3-by-3 doubly stochastic matrix, then either*

(i) *B is reducible, $W(A) = [\operatorname{tr} A - 3, 1]$, where $2 \le \operatorname{tr} A \le 4$, and $\sigma(A) = \{1, 1, 1, \operatorname{tr} A - 3\}$, or*

(ii) *B is irreducible, $W(A) = W(B)$, and $\sigma(A) = \{1\} \cup \sigma(B)$, both of which were as described in Proposition 3.3.3.*

Proof. Since the proof is very similar to Proposition 3.3.3, we omit it. ∎

The next proposition is concerned with 4-by-4 irreducible doubly stochastic matrices. If the index of imprimitivity $m(A)$ equals one, then it shows that A is unitarily similar to a direct sum of [1] and a 3-by-3 matrix B, and the numerical range $W(B)$ is contained in the open unit disc $\mathbb{D}$ by [15, Theorem 1.2]. Hence we can describe the shape of $W(A)$ in terms of $W(B)$. Note that $W(B)$ has four possible shapes (cf. [9]). Moreover, if $m(A) \ge 2$, then $m(A) = 2$ or 4 by [13, p. 51].

Proposition 3.3.5. *Let A be a 4-by-4 irreducible doubly stochastic matrix.*

(a) *Assume that $m(A) = 1$. Then $k(A) = 4$ if and only if $A = [1] \oplus [\lambda] \oplus C$, where $\lambda\ (\ne 1) \in \partial W(A) \cap \mathbb{R}$ so that either $W(A)$ is a 4-polygon or $W(A)$ is the convex hull of the (closed) interval $[\lambda, 1]$ and the elliptic disc $W(C)$.*

(b) *If $m(A) \ge 2$, then $k(A) = 4$. More precisely, the following hold:*

(i) *If $m(A) = 4$, then A is normal with $W(A)$ the regular 4-polygon with vertices $e^{2\pi i/4}, 0 \le j < 4$.*

(ii) *If* $m(A)=2$*, then* $W(A)$ *is either the closed interval* $[-1,1]$*, in which case* A *is Hermitian with spectrum* $\{\pm 1, \pm\sqrt{\det A}\}$*, or the convex hull of the closed interval* $[-1,1]$ *and* $W(B)$*, in which case* A *is permutationally similar to* $\begin{bmatrix} 0 & A_1 \\ A_2 & 0 \end{bmatrix}$ *and* $W(B)$ *is the elliptic disc with foci* $\pm\sqrt{\det A}$ *and minor axis of length* $|\det A_1 - \det A_2|$.

Proof. Assume that $m(A)=1$. Then A is unitarily similar to the direct sum of $[1]$ and a 3-by-3 matrix B, and the numerical range $W(B)$ is contained in the open unit disc $\mathbb{D}$ by [15, Theorem 1.2]. Note that $W(A)$ is symmetric with respect to the x-axis since A is nonnegative. If $k(A)=4$ and B is reducible, say, $B = C \oplus [\lambda]$, then λ $(\neq 1)$ is in $\partial W(A) \cap \mathbb{R}$ and $k_1(C)=2$ by Proposition 2.3.8. Hence it follows that $W(A)$ has the asserted shapes. On the other hand, if $k(A)=4$ and B is irreducible, then $W(B)$ is the heart-shaped region which is symmetric with respect to the x-axis via $A \succeq 0$. However, this cannot happen by Proposition 2.3.7. Hence we have proven the necessity for $k(A)=4$. The converse is trivial.

Assume that $m(A) \geq 2$. Then $m(A)=2$ or 4 by [13, p. 51]. If $m(A)=4$, then case (i) holds trivially. Hence $k(A)=4$ by Proposition 2.2.1. On the other hand, if $m(A)=2$, then -1 and 1 are eigenvalues of A by the Perron–Frobenius Theorem (cf. [11, Theorem 15.5.1]). Since -1 and 1 are corners in the boundary of $W(A)$, both are reducing eigenvalues by [3, Theorem 1]. Hence A is unitarily similar to a direct sum of $\operatorname{diag}(1,-1)$ and a 2-by-2 matrix B, which shows that $W(A)$ is the convex hull of the closed interval $[-1,1]$ and $W(B)$. If $W(B)$ is contained in $[-1,1]$, then it is obvious that B is Hermitian and so is A. In this case, the value $k(A)=4$ holds obviously by Proposition 2.2.1. Therefore we may assume that $W(B)$ is not contained in $[-1,1]$. This implies that $W(B)$ is an elliptic disc. Furthermore, since $W(A)$ is symmetric with respect to the x-axis, $W(B)$ is also symmetric to the x-axis. Thus $W(A)$ has four line segments, called $L_1, \ldots, L_4$, on its boundary so that L_1 is parallel to L_2, and L_3 is parallel to L_4. This shows that $k(A)=4$ by [19, Corollary 2.5]. For the remaining proof of case (ii), we only need to compute $\operatorname{tr} A$ and $\det A$ directly. Hence we complete the proof of case (b). ∎

Recall that the index of imprimitivity of A is the number of eigenvalues z of A with $|z| = r(A)$. By Lemma 3.3.1, for a doubly stochastic matrix A an eigenvalue with absolute value one is a reducing eigenvalue. This implies that $k(A)$ has the lower bound $m(A)$. Hence we may combine Propositions 3.3.2 and 3.3.4 to give the following result on an n-by-n ($n \geq 3$) reducible doubly stochastic matrix.

Theorem 3.3.6. *Let A be an n-by-n ($n \geq 3$) reducible doubly stochastic matrix. If $n = 3$ or 4, then $k(A) = n$; otherwise, $k(A) \geq \max\{m(A), 4\}$.*

Our final result is on n-by-n ($n \geq 3$) irreducible doubly stochastic matrices.

Theorem 3.3.7. *Let A be an n-by-n ($n \geq 3$) irreducible doubly stochastic matrix.*

(a) *If $m(A) = 1$, then $k(A) \geq 3$.*

(b) *Assume that $m(A) \geq 2$.*

(i) *If n is a prime, then $k(A) = n$. In this case, A is normal with its numerical range $W(A)$ the regular n-polygon with vertices $e^{2\pi i/n}, 0 \leq j < n$.*

(ii) *If n is not a prime, then $k(A) \geq \max\{m(A), 3\}$. Moreover, $k(A) = m(A)$ if and only if $m(A) \geq 3$, the numerical range $W(A)$ is the $m(A)$-regular polygon with vertices $e^{2\pi i/m(A)}, 0 \leq j < m(A)$, and the dimension of H_a equals 2 for any nonextreme boundary point a of $W(A)$.*

Proof. Part (a) is obvious. So we assume that $m(A) \geq 2$ in the following. If n is a prime, then A is normal and its numerical range $W(A)$ is the n-polygon with vertices $e^{2\pi i/n}, 0 \leq j < n$ by [15, Corollary 1.5]. Hence $k(A) = n$ by Proposition 2.2.1. This completes the proof of (i). To show (ii), we note that $m(A)$ is exactly the number of reducing eigenvalues of A since A is a doubly stochastic matrix with $\omega(A) = 1$. This implies that $k(A) \geq m(A)$. In addition, it is obvious that $k(A) \geq 3$ for any n-

by-n ($n \geq 3$) doubly stochastic matrix. Hence $k(A) \geq \max\{m(A), 3\}$. Assume that $k(A) = m(A) \equiv m$. Then $m(A) \geq 3$ by the preceding arguments. Let $\omega_m = e^{2\pi i/m}$. Since each $\omega_m^j, 0 \leq j < n$, is a reducing eigenvalue of A, the matrix A is unitarily similar to $\begin{bmatrix} B & 0 \\ 0 & C \end{bmatrix}$, where $C = \operatorname{diag}(1, \omega_m, ..., \omega_m^{m-1})$. Since $k(A) = m(A)$, $W(B)$ is contained in the interior of $W(C)$. This proves the necessity of our assertion. The convers is obvious. ■

We end this section by stating a natural question on $k(A) = n$ for an n-by-n ($n \geq 3$) irreducible doubly stochastic matrix A. Is it true that $k(A) = n$ if and only if $n \geq 3$, the numerical range $W(A)$ is the n-regular polygon with vertices $e^{2\pi i/n}, 0 \leq j < n$, and the dimension of H_a equals 2 for any nonextreme boundary point a of $W(A)$? Obviously, the sufficiency for $k(A) = n$ holds since A must be normal. Nevertheless, the necessity fails. For example, let

$$A = \begin{bmatrix} 0 & 0 & 1/3 & 2/3 \\ 0 & 0 & 2/3 & 1/3 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}.$$

Then A is an irreducible doubly stochastic matrix with $m(A) = 2$ since its spectrum is $\{1, -1, \sqrt{3}i/3, -\sqrt{3}i/3\}$. In addition, A is clearly not Hermitian. By Proposition 3.3.5 (b) (ii), we have $k(A) = 4$. However, $W(A)$ is not the regular 4-polygon with vertices $e^{2\pi i/4}, 0 \leq j < 4$.

References

[1] K. A. Camenga, P. X. Rault, T. Sendova and I. M. Spitkovsky, Numerical range of some doubly stochastic matrces, Appl. Math. Comput., 221 (2013), 40–47.

[2] K. A. Camenga, P. X. Rault, T. Sendova and I. M. Spitkovsky, On the Gau-Wu number for some classes of matrices, Linear Algebra Appl., 444 (2014), 254–262.

[3] W. F. Donoghue, Jr., On the numerical range of a bounded operator, Michigan Math. J., 4 (1957), 261–263.

[4] M. R. Embry, The numerical range of an operator, Pacific J. Math., 32 (1970), 647–650.

[5] H.-L. Gau and P. Y. Wu, Numerical ranges and compressions of S_n-matrices, Oper. Matrices, 7 (2013), 465–476.

[6] E. Gutkin, The Toeplitz–Hausdorff theorem revisted: relating linear algebra and geometry, Math. Intelligencer, 26 (2004), no. 1, 8–14.

[7] F. Hausdorff, Der Wertevorrat einer Bilinearform, Math. Z., 3 (1919), 314–316.

[8] R. A. Horn and C. R. Johnson, Topics in Matrix Analysis, Cambridge Univ. Press, Cambridge, 1991.

[9] D. S. Keeler, L. Rodman and I. M. Spitkovsky, The numerical range of 3×3 matrices, Linear Algebra Appl., 252 (1997), 115–139.

[10] H.-Y Lee, Diagonals and numerical ranges of direct sums of matrices, Linear Algebra Appl., 439 (2013), 2584–2597.

[11] C.-K. Li, B.-S. Tam and P. Y. Wu, The numerical range of a nonnegative matrix, Linear Algebra Appl., 350 (2002), 1–23.

[12] P. Lancaster and M. Tismenetsky, The Theory of Matrices with Applications, 2nd ed., Academic Press, New York, 1985.

[13] H. Minc, Nonnegative Matrices, Wiley, 1988.

[14] T. Moran and I. M. Spitkovsky, On almost normal matrices, Textos Mat., 44 (2013), 131–144.

[15] P. Nylen and T.-Y. Tam, Numerical range of a doubly stochastic matrix, Linear Algebra Appl., 153 (1991), 161–176.

[16] B.-S. Tam and S. Yang, On matrices whose numerical ranges have circular or weak circular symmetry, Linear Algebra Appl., 302–303 (1999) 193–221.

[17] O. Toeplitz, Das algebraische Analogon zu einem Satze von Fejér, Math. Z., 2 (1918), 187–197.

[18] S.-H. Tso and P. Y. Wu, Matrical ranges of quadratic operators, Rocky Mountain J. Math., 29 (1999), 1139–1152.

[19] K.-Z. Wang and P. Y. Wu, Diagonals and numerical ranges of weighted shift matrices, Linear Algebra Appl., 438 (2013), 514–532.