Linear Algebra and its Applications 444 (2014) 254–262

Contents lists available at ScienceDirect

Linear Algebra and its Applications

www.elsevier.com/locate/laa

On the Gau–Wu number for some classes of matrices ☆

Kristin A. Camenga [a], Patrick X. Rault [b,*], Tsvetanka Sendova [c], Ilya M. Spitkovsky [d]

[a] *Department of Mathematics and Computer Science, Houghton College, Houghton, NY 14744, USA*
[b] *Department of Mathematics, State University of New York at Geneseo, Geneseo, NY 14454, USA*
[c] *Department of Mathematics, Michigan State University, East Lansing, MI 48824, USA*
[d] *Department of Mathematics, College of William and Mary, Williamsburg, VA 23187, USA*

A R T I C L E   I N F O

Article history:
Received 11 November 2013
Accepted 28 November 2013
Available online 13 December 2013
Submitted by A. Böttcher

MSC:
15A60

Keywords:
Numerical range
Toeplitz matrices
Almost normal matrices

A B S T R A C T

For a given $n \times n$ matrix A, let $k(A)$ stand for the maximal number of orthonormal vectors $\mathbf{x}_j$ such that the scalar products $\langle A\mathbf{x}_j, \mathbf{x}_j \rangle$ lie on the boundary of the numerical range $W(A)$. This number was recently introduced by Gau and Wu and we therefore call it the Gau–Wu number of the matrix A. We compute $k(A)$ for two classes of $n \times n$ matrices A. A simple and explicit expression for $k(A)$ for tridiagonal Toeplitz matrices A is derived. Furthermore, we prove that $k(A) = 2$ for every pure almost normal matrix A. Note that for every matrix A we have $k(A) \geqslant 2$, and for normal matrices A we have $k(A) = n$, so our results show that pure almost normal matrices are in fact as far from normal as possible with respect to the Gau–Wu number. Finally, matrices with maximal Gau–Wu number ($k(A) = n$) are considered.

☆ The work on this paper was prompted by the authors' discussions during a SQuaRE workshop at the American Institute of Mathematics in May 2013, supported by the NSF. The second author [P.X.R.] was partially supported by a SUNY Geneseo research development award. The fourth author [I.M.S.] has been partially supported by the William & Mary summer research grant, and is also thankful to the NYU Abu Dhabi for the perfect working environment he enjoyed during his stay there in the Fall semester of the 2013–2014 academic year when the work on the paper was completed.

* Corresponding author.
E-mail addresses: kristin.camenga@houghton.edu (K.A. Camenga), rault@geneseo.edu (P.X. Rault), tsendova@math.msu.edu (T. Sendova), ilya@math.wm.edu, imspitkovsky@gmail.com (I.M. Spitkovsky).

0024-3795/$ – see front matter
http://dx.doi.org/10.1016/j.laa.2013.11.045

1. Introduction

Let $\mathbb{C}^n$ and $M_n(\mathbb{C})$ stand for the standard n-dimensional complex space and the algebra of all $n \times n$ matrices with complex entries, respectively. Denote also by $\langle \cdot, \cdot \rangle$ the scalar product on $\mathbb{C}^n$, and by $\| \cdot \|$ the norm associated with it. The *numerical range* of $A \in M_n(\mathbb{C})$, defined as

$$W(A) = \{ \langle A\mathbf{x}, \mathbf{x} \rangle \colon \ \|\mathbf{x}\| = 1 \},$$

has been studied extensively since the pioneering work by Toeplitz [13] and Hausdorff [6] in which the convexity of $W(A)$ was established; see e.g. [5] or [7, Chapter 1] for a systematic and relatively up to date exposition of this and further results on the subject.

A recent development in the theory of numerical ranges was the introduction in [4] of $k(A)$, a new matrix invariant. $k(A)$ is defined as the maximal size of an orthonormal set $\{\mathbf{x}_1, \ldots, \mathbf{x}_k\} \subset \mathbb{C}^n$ such that the values $\langle A\mathbf{x}_j, \mathbf{x}_j \rangle$ lie on the boundary $\partial W(A)$ of $W(A)$. We will call $k(A)$ the *Gau–Wu number* of $A \in M_n(\mathbb{C})$. Obviously, $1 \leqslant k(A) \leqslant n$, and it is not hard to see [4, Lemma 4.1] that $k(A) \geqslant 2$ if $n \geqslant 2$. So $k(A) = 2$ for all $A \in M_2(\mathbb{C})$. Let S_n denote the set of matrices A with all eigenvalues in the open unit disk and with $I_n - A^*A$ having rank 1. Then the main result of [4] is the formula $k(A) = \lceil n/2 \rceil$, for A in S_n with $n \geqslant 3$.

Further results on $k(A)$ were obtained in [14]. Namely, the values of $k(A)$ are completely classified when $n = 3$ (Proposition 2.11). Conditions are also given that characterize weighted shift matrices $A \in M_n(\mathbb{C})$ for which $k(A) = n$ (Theorem 3.1).

In this paper, we consider $k(A)$ for yet some other classes of matrices. In particular, Section 2 treats the case of almost normal matrices (as defined by Ikramov [8]). Theorem 1 shows that $k(A) = 2$ for pure almost normal matrices, while the Gau–Wu number for general almost normal matrices is calculated in Theorem 3. In Section 3 the case of tridiagonal Toeplitz matrices is considered and the Gau–Wu number is explicitly calculated in Theorem 5. Section 4 is devoted to matrices with maximal Gau–Wu number, i.e. $k(A) = n$, for which we verify the conjecture in [9] by proving Theorem 6. For the numerical range of unitarily irreducible matrices of maximal Gau–Wu number, Theorem 7 claims that $\langle A\mathbf{x}_j, \mathbf{x}_j \rangle$ are concentrated on two parallel support lines.

2. Almost normal matrices

There are various generalizations of the notion of normal matrices. We here adopt Ikramov's in [8], according to which $A \in M_n(\mathbb{C})$ is *almost normal* if it has $n - 1$ orthogonal eigenvectors.

This notion of almost normality was further dealt with in [11], where in particular it was mentioned that every almost normal matrix is unitarily similar to $A_n \oplus A_a$, where the block A_n is normal while A_a is almost normal and unitarily irreducible. Recall that a matrix A is unitarily reducible if and only if $U^*AU = A_1 \oplus A_2$ for some unitary matrix U and for some lower dimensional matrices A_1 and A_2. Each of the blocks A_n and A_a in the decomposition of an almost normal matrix is defined up to a unitary similarity; for convenience of reference we will call them the normal and pure almost normal components of A. Note that each of the components may be void: if A_a disappears, then A is normal. At the other extreme are unitarily irreducible almost normal matrices, called *pure almost normal* in [11].

Theorem 1. *For every pure almost normal matrix A, $k(A) = 2$.*

Proof. Let A be a pure almost normal matrix. According to [11, Theorem 2.1], it is then unitarily similar to

$$\begin{bmatrix} \lambda_1 & 0 & \ldots & 0 & \beta_1 \\ 0 & \lambda_2 & \ldots & 0 & \beta_2 \\ \vdots & & \ddots & & \vdots \\ 0 & 0 & \ldots & \lambda_{n-1} & \beta_{n-1} \\ 0 & 0 & \ldots & 0 & \mu \end{bmatrix} \tag{1}$$

with $\beta_j \neq 0$ and distinct λ_j, $j = 1, \dots, n-1$. Without loss of generality, we may suppose that A itself equals (1). Therefore, the real part of the matrix $e^{i\theta}A$ is

$$\operatorname{Re}(e^{i\theta}A) = \begin{bmatrix} \operatorname{Re}(\lambda_1 e^{i\theta}) & 0 & \dots & 0 & e^{i\theta}\beta_1/2 \\ 0 & \operatorname{Re}(\lambda_2 e^{i\theta}) & \dots & 0 & e^{i\theta}\beta_2/2 \\ \vdots & & \ddots & & \vdots \\ 0 & 0 & \dots & \operatorname{Re}(\lambda_{n-1}e^{i\theta}) & e^{i\theta}\beta_{n-1}/2 \\ e^{-i\theta}\overline{\beta_1}/2 & e^{-i\theta}\overline{\beta_2}/2 & \dots & e^{-i\theta}\overline{\beta_{n-1}}/2 & \operatorname{Re}(\mu e^{i\theta}) \end{bmatrix}.$$

Since $\beta_j \neq 0$ for all j, the (j,n)-principal minor of the matrix $\operatorname{Re}(e^{i\theta}A - \lambda_j e^{i\theta}I_n)$ is negative. Due to the interlacing theorem, the maximal eigenvalue $\xi(\theta)$ of $\operatorname{Re}(e^{i\theta}A)$ is strictly bigger than $\operatorname{Re}(\lambda_j e^{i\theta})$ for all j. Consequently, the first $n-1$ diagonal entries of $\operatorname{Re}(e^{i\theta}A) - \xi(\theta)I_n$ are all non-zero, i.e. $\xi(\theta)$ is a simple eigenvalue.

This means that each supporting line of $W(A)$ contains exactly one point of $\partial W(A)$. Moreover, for the line lying to the right of $W(A)$ and forming the angle $\pi/2 - \theta$ with the positive real axis, this point is generated by the unit eigenvector of $\operatorname{Re}(e^{i\theta}A)$ corresponding to $\xi(\theta)$. A straightforward computation shows that this eigenvector is collinear with

$$\mathbf{x}(\theta) = \left[\frac{\beta_1 e^{i\theta}}{2(\xi(\theta) - \operatorname{Re}(\lambda_1 e^{i\theta}))}, \dots, \frac{\beta_{n-1} e^{i\theta}}{2(\xi(\theta) - \operatorname{Re}(\lambda_{n-1} e^{i\theta}))}, 1\right]^T.$$

So, the scalar product of two such vectors, corresponding to the angles θ_1 and θ_2, is evaluated as follows:

$$\langle \mathbf{x}(\theta_1), \mathbf{x}(\theta_2)\rangle = e^{i(\theta_1-\theta_2)} \sum_{j=1}^{n-1} \frac{|\beta_j|^2}{4(\xi(\theta_1) - \operatorname{Re}(\lambda_j e^{i\theta_1}))(\xi(\theta_2) - \operatorname{Re}(\lambda_j e^{i\theta_2}))} + 1. \tag{2}$$

Since each term of the sum in the right hand side of (2) is positive, for this scalar product to equal zero it is necessary that $\theta_1 - \theta_2 = \pi \mod 2\pi$. Obviously, this condition is also sufficient for orthogonality, because then $\mathbf{x}(\theta_1)$ and $\mathbf{x}(\theta_2)$ are eigenvectors corresponding to distinct eigenvalues of the same Hermitian matrix. Therefore, only two mutually orthogonal vectors generating boundary points of $W(A)$ can be picked simultaneously. □

Note that as a byproduct of the proof we have shown that for pure almost normal matrices A, every point of $\partial W(A)$ is *singularly generated*, i.e. its pre-image under the mapping $\mathbf{x} \mapsto \langle A\mathbf{x}, \mathbf{x}\rangle$ of the unit sphere in $\mathbb{C}^n$ is one-dimensional. Actually, this condition must hold for every matrix $A \in M_n(\mathbb{C})$ with $k(A) = 2$. The converse is obviously true for $n = 2$; it still holds for $n = 3$ but fails starting with $n = 4$ ([14], Corollary 2.12 and examples stemming from Theorem 3.10, respectively). A description of all matrices $A \in M_n(\mathbb{C})$ with $k(A) = 2$ remains an open problem for $n \geqslant 4$.

We now turn to the computation of $k(A)$ for general (that is, not necessarily pure) almost normal matrices A. The key ingredient here, besides Theorem 1, is one of Lee's results from [9] on the computation of $k(A)$ for unitarily reducible matrices A, i.e. for A which are unitarily similar to some reduction $A_1 \oplus A_2$. To formulate this result, we denote by n_j the size of the block A_j and introduce (also following [9]) $k_1(A_j)$ as the maximal number of orthonormal vectors $\mathbf{x}_{js} \in \mathbb{C}^{n_j}$ for which

$$\langle A_j \mathbf{x}_{js}, \mathbf{x}_{js}\rangle \in \partial W(A) \cap \partial W(A_j), \quad s = 1, \dots, n_j; \; j = 1, 2.$$

Theorem 2. *(See [9, Proposition 3.1].) Let A be unitarily similar to $A_1 \oplus A_2$, with at least one block being normal. Then*

$$k(A) = k_1(A_1) + k_1(A_2). \tag{3}$$

Before stating the result for almost normal matrices A, recall the decomposition $A_n \oplus A_a$ with normal A_n and pure almost normal A_a which can be achieved for such A via a unitary similarity.

Theorem 3. *Let A be almost normal. Then $k(A) = \ell_1 + \ell_2$, where ℓ_1 is the number of eigenvalues of A_n located on $\partial W(A)$, counting their multiplicities, and*

$$\ell_2 = \begin{cases} 0 & \textit{if } W(A_a) \textit{ lies in the interior of } W(A_n), \\ 2 & \textit{if there exist distinct parallel supporting lines of } W(A) \\ & \textit{passing through points of } W(A_a), \textit{ or} \\ 1 & \textit{otherwise}. \end{cases}$$

Note that the case $\ell_2 = 1$ occurs exactly when $W(A_a) \cap \partial W(A)$ is non-empty but does not contain points lying on distinct parallel supporting lines of $W(A)$.

Proof. For pure almost normal matrices $\ell_1 = 0$, $\ell_2 = 2$, and the result follows from Theorem 1. So, we need only to consider the case when the block A_n is actually present.

Recall that the numerical range of every normal matrix is the convex hull of its spectrum. Consequently, $P := \partial W(A_n)$ is a polygon. Also, $W(A)$ is the convex hull of $W(A_n)$ and $W(A_a)$, which due to the convexity of $W(A_a)$ implies that if $z \in \partial W(A)$ is in the relative interior of some edge of P then the whole edge lies in $\partial W(A)$. In other words, $P \cap \partial W(A)$ is the (possibly empty) union of some vertices and edges of P. Consider the subspace L spanned by all the eigenvectors of A_n corresponding to its eigenvalues lying in $P \cap \partial W(A)$. By construction, $\dim L = \ell_1$.

Since there is an orthogonal basis of L whose elements are eigenvectors of A_n, we see that $k_1(A_n) \geqslant \ell_1$. On the other hand, as each unit vector $\mathbf{x}$ for which $\langle A_n\mathbf{x}, \mathbf{x}\rangle \in P \cap \partial W(A)$ is a linear combination of (at most two) such eigenvectors, the converse inequality also holds. So, $k_1(A_n) = \ell_1$.

Passing on to the second block, A_a, observe that as was shown while proving Theorem 1, the unit vectors generating the points of $\partial W(A_a)$ are orthogonal only if the supporting lines of $W(A_a)$ passing through these points are parallel. Therefore, $k_1(A_a) = \ell_2$ and Theorem 2 implies the result. □

Corollary 4. *Let A be almost normal, with the numerical range of its pure almost normal component A_a lying in the interior of $W(A_n)$. Then $k(A)$ coincides with the number of the eigenvalues of A_n lying in $\partial W(A)$, counting their multiplicities.*

(Of course, $W(A) = W(A_n)$ in the setting of Corollary 4.) The result holds in particular for normal A, in which case it becomes [9, Proposition 2.1].

3. Tridiagonal Toeplitz matrices

As usual, let us denote the (i, j)-entry of $A \in M_n(\mathbb{C})$ by a_{ij}. Toeplitz matrices, by definition, have constant diagonals: $a_{ij} = a_{i+1,j+1}$ for $i, j = 1, \dots, n-1$. For tridiagonal matrices, on the other hand, $a_{ij} \neq 0$ only if $|i - j| \leqslant 1$. So, tridiagonal Toeplitz matrices are those of the form

$$T(a,b,c) = \begin{bmatrix} a & c & 0 & \dots & 0 \\ b & a & c & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & b & a & c \\ 0 & \dots & 0 & b & a \end{bmatrix}, \tag{4}$$

depending on three parameters $a, b, c \in \mathbb{C}$. Denoting $J = T(0, 0, 1)$, we may rewrite (4) as follows:

$$T(a,b,c) = aI + bJ^T + cJ.$$

If $bc = 0$, then $T(a, b, c)$ is a triangular matrix, with a being its only eigenvalue. For $b, c \neq 0$, on the other hand, the eigenvalues of $T(a, b, c)$ are all simple and given by the formula

$$\lambda_j = a + 2\sqrt{bc}\cos\left(\frac{j\pi}{n+1}\right), \quad j = 1, \dots, n. \tag{5}$$

The respective eigenvectors are $\mathbf{x}_j = [x_1^{(j)}, \dots, x_n^{(j)}]^T$, with the entries given by

$$x_k^{(j)} = \left(\sqrt{\frac{b}{c}}\right)^k \sin\left(\frac{jk\pi}{n+1}\right), \quad k = 1, \dots, n. \tag{6}$$

For more details, see e.g. [10] or [1, Section 2.2].

Theorem 5. *Let $n \geqslant 3$ and A be an $n \times n$ tridiagonal Toeplitz matrix, as in (4). Then*

$$k(A) = \begin{cases} n & \text{if } |b| = |c|, \text{ or} \\ \lceil \frac{n}{2} \rceil & \text{otherwise}. \end{cases} \tag{7}$$

Proof. The normality criterion for tridiagonal matrices [3, Lemma 1] (or simply a straightforward computation) shows that $T(a, b, c)$ is normal under the condition $|b| = |c|$. Since all the eigenvalues (5) lie on a line, $W(T(a, b, c))$ is then a line segment. Thus, $\partial W(T(a, b, c)) = W(T(a, b, c))$, and all the eigenvalues lie on the boundary of the numerical range. This verifies the first line of (7).

The case $a = b = 0$, $c = 1$ is covered by [4, Theorem 4.4], because then $A = J$ belongs to the class S_n. Since the Gau–Wu number is invariant under shifting, scaling, and transposing, the result holds whenever exactly one of b, c is different from zero. So, it remains to consider the situation $b, c \neq 0$, $|b| \neq |c|$.

As in the proof of Theorem 1, we make use of the fact that the boundary points of $W(A)$ are generated by the eigenvectors of $\operatorname{Re}(e^{i\theta} A)$ corresponding to its maximal eigenvalues. For $A = T(a, b, c)$, the matrix

$$\operatorname{Re}\left(e^{i\theta} A\right) = T\left(\operatorname{Re}\left(e^{i\theta} a\right), w, \overline{w}\right), \quad \text{where } w = w(\theta) = \frac{e^{i\theta} b + e^{-i\theta} \bar{c}}{2},$$

is tridiagonal Toeplitz along with A. Thus the eigenvalues and eigenvectors of $\operatorname{Re}(e^{i\theta} A)$ can be computed explicitly for any θ by an appropriate change of notation in (5)–(6). The maximal eigenvalue $\lambda(\theta)$ corresponds to $j = 1$ and thus equals

$$\lambda(\theta) = \operatorname{Re}\left(e^{i\theta} a\right) + |w| \cos \frac{\pi}{n+1},$$

while the associated eigenvector $\mathbf{x}(\theta) = [x_1(\theta), \dots, x_n(\theta)]^T$ has the coordinates

$$x_k(\theta) = \phi^k \sin \frac{k\pi}{n+1}, \quad \text{where } \phi = \phi(\theta) = \frac{w}{|w|}. \tag{8}$$

Note that ϕ maps $[0, 2\pi)$ bijectively onto the unit circle $\mathbb{T}$. Therefore, in order to find the maximal number of pairwise orthogonal vectors $\mathbf{x}(\theta)$ for $\theta \in \mathbb{T}$, we may consider them as functions of $\phi \in \mathbb{T}$. From (8) we conclude that

$$\left\langle \mathbf{x}(\phi_1), \mathbf{x}(\phi_2) \right\rangle = \sum_{k=1}^{n} \zeta^k \sin^2 \frac{k\pi}{n+1}, \quad \text{where } \zeta = \frac{\phi_1}{\phi_2}.$$

Denoting by $\alpha = e^{2\pi i/(n+1)}$ the $(n+1)$st root of unity, one arrives at

$$\begin{aligned} \sum_{k=1}^{n} \zeta^k \sin^2 \frac{k\pi}{n+1} &= \sum_{k=1}^{n} \zeta^k \left[\frac{e^{ik\pi/(n+1)} - e^{-ik\pi/(n+1)}}{2i} \right]^2 = -\frac{1}{4} \sum_{k=0}^{n} \zeta^k \left(\alpha^k + \bar{\alpha}^k - 2\right) \\ &= -\frac{1}{4} \left[\frac{(\zeta\alpha)^{n+1} - 1}{\zeta\alpha - 1} + \frac{(\zeta\bar{\alpha})^{n+1} - 1}{\zeta\bar{\alpha} - 1} - 2\frac{\zeta^{n+1} - 1}{\zeta - 1} \right] \\ &= -\frac{\zeta(\zeta^{n+1} - 1)(\zeta + 1)(\alpha + \bar{\alpha} - 2)}{4(\zeta\alpha - 1)(\zeta\bar{\alpha} - 1)(\zeta - 1)} = \frac{\zeta(\zeta^{n+1} - 1)(\zeta + 1)}{(\zeta - \bar{\alpha})(\zeta - \alpha)(\zeta - 1)} \sin^2 \frac{\pi}{n+1}. \end{aligned}$$

So, in order for the set $\{\mathbf{x}(\phi_1), \dots, \mathbf{x}(\phi_k)\}$ to be orthogonal, the ratios of the distinct ϕ_j involved must all be among the roots of the polynomial $\frac{(\zeta^{n+1}-1)(\zeta+1)}{(\zeta-\bar{\alpha})(\zeta-\alpha)(\zeta-1)}$, which are -1 and $\alpha^2, \dots, \alpha^{n-1}$. Choosing $\phi_j = \alpha^{2j}$, $j = 1, \dots, \lceil n/2 \rceil$, we thus obtain an orthogonal set of the respective $\mathbf{x}(\phi_j)$'s. So, $k(A) \geqslant \lceil n/2 \rceil$.

To prove the reverse inequality, we consider separately the cases of odd and even n.

Case 1. Odd n. Then $-1 = \alpha^{(n+1)/2}$, and the set of admissible ratios for distinct ϕ_j's is simply $\{\alpha^2, \dots, \alpha^{n-1}\}$. Since it consists of powers of α but does not contain α itself, the arc distance between any two ϕ_j's must be at least $4\pi/(n+1)$. Therefore, the number of these points is limited by

$$\frac{2\pi}{4\pi/(n+1)} = \frac{n+1}{2} = \left\lceil \frac{n}{2} \right\rceil. \tag{9}$$

Case 2. Even n. Suppose an orthogonal set $\{\mathbf{x}(\phi_1), \dots, \mathbf{x}(\phi_k)\}$ contains two vectors with opposite ϕ_j's, that is, -1 is one of the ratios. Then this set cannot contain any other vectors. Indeed, if (without loss of generality) $\phi_2/\phi_1 = -1$, then for any other choice of $\phi \in \mathbb{T}$ at least one of the ratios ϕ/ϕ_1 and ϕ/ϕ_2 will not belong to the admissible set.

So, in order to achieve more than 2 vectors, all the ratios must lie within $\{\alpha^2, \dots, \alpha^{n-1}\}$. Now the reasoning of Case 1 can be repeated, with an obvious replacement of the left (and middle) side of (9) by its integer part. □

Observe that matrices (4) with $b, c \neq 0$ and $|b| \neq |c|$ for $n \geqslant 5$ deliver new examples of the situation when every boundary point of $W(A)$ is singularly generated while $k(A) > 2$.

As an example, for $n = 7$, consider the matrix (4) with $a = 5 + 4i$, $b = -1 + i$, $c = -3$. The set of four pairwise orthogonal vectors

$$\mathbf{x}_1 = \left[i \sin\frac{\pi}{8}, \frac{-1}{\sqrt{2}}, -i\cos\frac{\pi}{8}, 1, i\cos\frac{\pi}{8}, \frac{-1}{\sqrt{2}}, -i\sin\frac{\pi}{8} \right]^T,$$

$$\mathbf{x}_2 = \left[-\sin\frac{\pi}{8}, \frac{1}{\sqrt{2}}, -\cos\frac{\pi}{8}, 1, -\cos\frac{\pi}{8}, \frac{1}{\sqrt{2}}, -\sin\frac{\pi}{8} \right]^T,$$

$$\mathbf{x}_3 = \left[-i \sin\frac{\pi}{8}, \frac{-1}{\sqrt{2}}, i\cos\frac{\pi}{8}, 1, -i\cos\frac{\pi}{8}, \frac{-1}{\sqrt{2}}, i\sin\frac{\pi}{8} \right]^T,$$

$$\mathbf{x}_4 = \left[\sin\frac{\pi}{8}, \frac{1}{\sqrt{2}}, \cos\frac{\pi}{8}, 1, \cos\frac{\pi}{8}, \frac{1}{\sqrt{2}}, \sin\frac{\pi}{8} \right]^T$$

respectively generate the points

$$z_1 = \langle A\mathbf{x}_1, \mathbf{x}_1 \rangle = 5 + \frac{\sqrt{2+\sqrt{2}}}{2} + i\left(4 - \sqrt{2+\sqrt{2}}\right),$$

$$z_2 = \langle A\mathbf{x}_2, \mathbf{x}_2 \rangle = 5 + 2\sqrt{2+\sqrt{2}} + i\left(4 - \frac{\sqrt{2+\sqrt{2}}}{2}\right),$$

$$z_3 = \langle A\mathbf{x}_3, \mathbf{x}_3 \rangle = 5 - \frac{\sqrt{2+\sqrt{2}}}{2} + i\left(4 + \sqrt{2+\sqrt{2}}\right),$$

$$z_4 = \langle A\mathbf{x}_4, \mathbf{x}_4 \rangle = 5 - 2\sqrt{2+\sqrt{2}} + i\left(4 + \frac{\sqrt{2+\sqrt{2}}}{2}\right),$$

lying on the boundary of $W(A)$. These boundary points are graphed in Fig. 1. In agreement with Theorem 5, $k(A) = 4$.

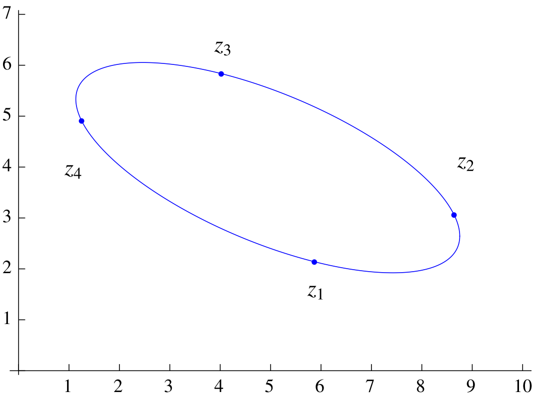

Fig. 1. An example of a 7×7 tridiagonal Toeplitz matrix, A, of the form (4) with $a=5+4i$, $b=-1+i$, $c=-3$; in particular $|b|\neq|c|$ and $k(A)=\lceil 7/2\rceil=4$, as asserted by Theorem 5.

4. Matrices with maximal Gau–Wu number

Let us return to the situation of A being unitarily similar to the direct sum $A_1\oplus A_2$, and recall the notation $k_1(A_j)$ introduced just before Theorem 2. Equality (3) is valid, according to Theorem 2, when one of the blocks is normal, but it actually holds under various other additional conditions as well, e.g., when $W(A_1)\cap W(A_2)=\emptyset$ or A_2-A_1 is a scalar multiple of the identity; see [9]. Note also that the inequality $k(A)\geqslant k_1(A_1)+k_1(A_2)$ is obvious.

It was conjectured in [9] that (3) holds without any additional conditions imposed on the blocks A_j. It is presently not known whether the conjecture is true. However, here is yet another case in which formula (3) is valid.

Theorem 6. *Let $A\in M_n(\mathbb{C})$ be such that $k(A)=n$. If, in addition, A is unitarily similar to the direct sum $A_1\oplus\cdots\oplus A_m$, then for each block $A_j\in M_{n_j}(\mathbb{C})$ there is an orthonormal basis $\{\mathbf{x}_{j,1},\dots,\mathbf{x}_{j,n_j}\}$ of $\mathbb{C}^{n_j}$ such that $\langle A_j\mathbf{x}_{j,i},\mathbf{x}_{j,i}\rangle\in\partial W(A)$ for $i=1,\dots,n_j$.*

Proof. According to [14, Theorem 2.7], A is unitarily similar to a direct sum of the blocks

$$B_j=\left[\begin{array}{ccc|ccc}\alpha_1^{(j)} & \dots & 0 & & & \\ \vdots & \ddots & \vdots & & e^{i\theta_j}D_j & \\ 0 & \dots & \alpha_{s_j}^{(j)} & & & \\ \hline & & & \beta_1^{(j)} & \dots & 0\\ & -e^{i\theta_j}D_j^* & & \vdots & \ddots & \vdots\\ & & & 0 & \dots & \beta_{t_j}^{(j)}\end{array}\right], \tag{10}$$

with the real parts of $e^{-i\theta_j}\alpha_i^{(j)}$ (resp. $e^{-i\theta_j}\beta_i^{(j)}$) all coinciding with the maximal (resp. minimal) eigenvalue of $\mathrm{Re}(e^{-i\theta_j}A)$, implying that the diagonal entries of B_j (generated by the respective standard basis) lie on the boundary of $W(A)$.

In [14], the angles θ_j were distinct and the blocks B_j could be unitarily reducible. Observe, however, that every reducing subspace L of a matrix (10) must be invariant under the complementary projections onto the first s_j and last t_j coordinates. Equivalently, $L=L_1\oplus L_2$, with L_1 (resp. L_2) lying in the span of the first s_j (resp. last t_j) vectors of the standard basis of $\mathbb{C}^{s_j+t_j}$. This follows from the invariance of L under

$$\frac{1}{2}\left(e^{-i\theta_j}B_j+e^{i\theta_j}B_j^*\right)=MI_{s_j}\oplus mI_{t_j}, \tag{11}$$

where M (resp. m) is the maximal (resp., minimal) eigenvalue of $\mathrm{Re}(e^{-i\theta_j}A)$, provided that $M \neq m$. The case $m = M$ is trivial, because then A is normal.

Moreover, L_1 and L_2 must be invariant under $\mathrm{diag}[\alpha_1^{(j)}, \ldots, \alpha_{s_j}^{(j)}]$ and $\mathrm{diag}[\beta_1^{(j)}, \ldots, \beta_{t_j}^{(j)}]$, respectively, while $D_j L_1 \subset L_2$, $D_j^* L_2 \subset L_1$. Therefore we can break down each matrix (10) into unitarily irreducible blocks of the same form while maintaining the diagonal entries. In other words, A is in fact unitarily similar to the direct sum of unitarily irreducible blocks (10), with the diagonal entries lying on $\partial W(A)$ and with not necessarily distinct θ_j.

Since the decomposition of any matrix into unitarily irreducible blocks under unitary similarity is unique up to the order of the blocks and their unitary similarities (see e.g. [12, Section 8]), the matrices A_j from the statement of the theorem must in turn be unitarily similar to direct sums of the blocks (10). Consequently, for each A_j there are exactly n_j orthonormal vectors which generate points on the boundary of $W(A)$. □

Finally, let us consider a unitarily irreducible matrix $A \in M_n(\mathbb{C})$ with $k(A) = n$. It must then be unitarily similar to just one block of the form (10). Therefore, from now on we will suppress the index j for θ, α_i, β_i, and D. Rotating by the corresponding angle θ will align all of the α_i on one vertical line, a supporting line on the right of the rotated numerical range. Similarly, all of the β_i will then lie on another vertical supporting line on the left. Therefore, the numerical range $W(A)$ of the originally given (that is, not subjected to the rotation) matrix A has two parallel supporting lines ℓ_1, ℓ_2 such that for some orthonormal basis $\{\mathbf{x}_j\colon\ j = 1, \ldots, n\}$ of $\mathbb{C}^n$,

$$\langle A\mathbf{x}_j, \mathbf{x}_j\rangle \in \ell_1 \cup \ell_2, \quad j = 1, \ldots, n. \tag{12}$$

For $n = 2$, naturally, every pair of supporting lines will satisfy (12), regardless of their direction. Starting with $n = 3$, the pair becomes unique and, moreover, at least one of the supporting lines will have to intersect $W(A)$ in a line segment (that is, contain a *flat portion* of $\partial W(A)$).

Theorem 7. *Let $A \in M_n(\mathbb{C})$, for some $n > 2$ with $k(A) = n$, be unitarily irreducible. Then there is exactly one pair of parallel lines supporting $W(A)$ for which (12) holds. Also, at least one of the intersections $\ell_j \cap W(A)$ is not a singleton.*

Proof. Since A is unitarily irreducible, (11) implies that $\mathrm{Re}(e^{-i\theta}A)$ is a linear combination of some orthogonal projection P and the identity I for θ determining the direction of the supporting lines in (12). Equivalently,

$$H\cos\theta + K\sin\theta = aI + bP \quad \text{for some } a, b \in \mathbb{R}, \tag{13}$$

where $H = \mathrm{Re}\,A$, $K = \mathrm{Im}\,A$. So, if there are two pairs of supporting lines with property (12), then (13) holds with θ replaced by two different (mod π) values θ_1, θ_2, and P respectively replaced by two (possibly, but not necessarily different) orthogonal projections P_1, P_2. Consequently, H, K, and therefore A itself, are linear combinations of I, P_1, P_2. It remains to observe that all matrices from an algebra generated by two orthogonal projections are unitarily similar to direct sums of at most two-dimensional blocks, see e.g. [2].

We now pass to the part of the proof concerning the flat portion. Suppose that both $\ell_1 \cap W(A)$ and $\ell_2 \cap W(A)$ are singletons. Without loss of generality, and for simplicity of notation, we may rotate A in order to make ℓ_j vertical. Then, choosing an orthonormal basis satisfying (12), we see that A is unitarily similar to $H + iK$, where

$$H = \begin{bmatrix} \lambda_1 I & 0 \\ 0 & \lambda_2 I \end{bmatrix}, \qquad K = \begin{bmatrix} \mu_1 I & Z \\ Z^* & \mu_2 I \end{bmatrix}.$$

Applying an additional block diagonal unitary similarity via $U = \mathrm{diag}[U_1, U_2]$, we may, without changing the diagonal blocks of H and K, replace Z by $U_1 Z U_2^*$. In particular, we can change Z to the middle factor of its singular value decomposition, thus making A unitarily similar to the direct sum of at most two-dimensional blocks (again). □

Examples of unitarily irreducible $A \in M_4(\mathbb{C})$ with two parallel flat portions on the boundary of $W(A)$ can be found e.g. in [3] (Example 20).

Acknowledgements

We are thankful to the anonymous referee for several remarks which helped to improve the exposition.

References

[1] A. Böttcher, S.M. Grudsky, Spectral Properties of Banded Toeplitz Matrices, SIAM, Philadelphia, 2005.
[2] A. Böttcher, I.M. Spitkovsky, A gentle guide to the basics of two projections theory, Linear Algebra Appl. 432 (6) (2010) 1412–1459.
[3] E. Brown, I. Spitkovsky, On flat portions on the boundary of the numerical range, Linear Algebra Appl. 390 (2004) 75–109.
[4] H.-L. Gau, P.Y. Wu, Numerical ranges and compressions of S_n-matrices, Oper. Matrices 7 (2) (2013) 465–476.
[5] K.E. Gustafson, D.K.M. Rao, Numerical Range. The Field of Values of Linear Operators and Matrices, Springer, New York, 1997.
[6] F. Hausdorff, Der Wertvorrat einer Bilinearform, Math. Z. 3 (1919) 314–316.
[7] R.A. Horn, C.R. Johnson, Topics in Matrix Analysis, Cambridge University Press, Cambridge, 1991.
[8] Kh.D. Ikramov, On almost normal matrices, Vestnik Moskov. Univ. Ser. XV Vychisl. Mat. Kibernet. 1 (2011) 5–9, 56.
[9] H.-Y. Lee, Diagonals and numerical ranges of direct sums of matrices, Linear Algebra Appl. 439 (2013) 2584–2597.
[10] G. Lombardi, R. Rebaudo, Eigenvalues and eigenvectors of a special class of band matrices, Rend. Istit. Mat. Univ. Trieste 20 (1) (1988) 113–128, 1989.
[11] T. Moran, I.M. Spitkovsky, On almost normal matrices, Textos Mat. 44 (2013) 131–144.
[12] H. Shapiro, A survey of canonical forms and invariants for unitary similarity, Linear Algebra Appl. 147 (1991) 101–167.
[13] O. Toeplitz, Das algebraische Analogon zu einem Satze von Fejér, Math. Z. 2 (1918) 187–197.
[14] K.-Z. Wang, P.Y. Wu, Diagonals and numerical ranges of weighted shift matrices, Linear Algebra Appl. 438 (1) (2013) 514–532.