Operators
and
Matrices
Volume 11, Number 1 (2017), 219–240
doi:10.7153/oam-11-15

ON THE NORMALIZED NUMERICAL RANGE

ILYA M. SPITKOVSKY AND ANDREI-FLORIAN STOICA

In memory of Professor Leiba Rodman

(*Communicated by Y.-T. Poon*)

Abstract. The normalized numerical range of an operator A is defined as the set $F_N(A)$ of all the values $\langle Ax,x\rangle / \|Ax\|$ attained by unit vectors $x \notin \ker A$. We prove that $F_N(A)$ is simply connected, establish conditions for it to be star-shaped with the center at zero, to be open, closed, and to have empty interior. For some classes of operators (weighted shifts, isometries, essentially Hermitian) the complete description of $F_N(A)$ is obtained.

1. Introduction

Let $\mathscr{H}$ be a Hilbert space with the scalar product denoted by $\langle .,. \rangle$ and the induced norm $\|x\| = \langle x,x\rangle^{1/2}$. Denote by $S(\mathscr{H})$ the unit sphere of $\mathscr{H}$, and by $[\mathscr{H}]$ the algebra of all bounded linear operators on $\mathscr{H}$. We are not excluding the case $\dim\mathscr{H} = n < \infty$, in which $\mathscr{H}$ is identified with the column space $\mathbb{C}^n$ and $[\mathscr{H}]$ with the algebra $\mathbb{C}^{n\times n}$ of n-by-n matrices with complex entries. For a subspace $\mathscr{L}$ of $\mathscr{H}$, the *compression* of A onto $\mathscr{L}$ is the operator $B \in [\mathscr{L}]$ defined by $Bx = PAx$, where P is the orthoprojection of $\mathscr{H}$ onto $\mathscr{L}$.

We will use the standard notation $\mathbb{D}$ and $\mathbb{T}$ for the open unit disk $\{z\colon\ |z| < 1\}$ and the unit circle $\{z\colon\ |z| = 1\}$, respectively. For any set X in $\mathbb{C}$ or $\mathscr{H}$, $\operatorname{int} X$, ∂X and $\overline{X}$ denote its interior, boundary, and closure, in this order. In particular, $\overline{\mathbb{D}}$ is the closed unit disk $\mathbb{D} \cup \mathbb{T}$.

The (classical) *numerical range*, also known as the *field of values*, or the *Hausdorff set* of $A \in [\mathscr{H}]$, is defined as

$$F(A) = \{\langle Ax,x\rangle / \|x\|^2 : x \in \mathscr{H},\ x \neq 0\}. \tag{1.1}$$

Monographs [11] and [13, Chapter 1] are standard references for the properties of $F(A)$ and history of the subject. We will be recalling the needed facts about $F(A)$ as we go.

Mathematics subject classification (2010): 15A60, 47A12, 47B15.

Keywords and phrases: Numerical range, normalized numerical range, partial isometry, essentially Hermitian operator, weighted shift.

The first author was supported in part by Faculty Research funding from the Division of Science and Mathematics, New York University Abu Dhabi, and the Plumeri Award for Faculty Excellence from the College of William and Mary.

Several generalizations of the numerical range are also known (c-numerical range, q-numerical range, higher rank numerical range, to name a few). In this paper, we consider the *normalized* numerical range, NNR for short, introduced in [1] as

$$F_N(A) = \left\{ \frac{\langle Ax, x\rangle}{\|x\|\,\|Ax\|} : x \in \mathscr{H},\ Ax \neq 0 \right\}. \tag{1.2}$$

Since then, several properties of NNR were established by Gevorgyan, see [6]–[10]. Note that $F_N(A) = \emptyset$ if and only if $A = 0$. So, in what follows we suppose without saying that $A \neq 0$.

Denote by $F(A,\theta)$ and $F_N(A,\theta)$ the intersection of the ray

$$\ell_\theta := \{\rho e^{i\theta} : \rho > 0\} \tag{1.3}$$

with $F(A)$ and $F_N(A)$, respectively. A direct comparison of the definitions (1.1) and (1.2) reveals that $F(A,\theta)$ and $F_N(A,\theta)$ are non-empty only simultaneously. Let

$$\Theta(A) := \{e^{i\theta} : F_N(A,\theta) \neq \emptyset\} = \{e^{i\theta} : F(A,\theta) \neq \emptyset\}. \tag{1.4}$$

Recall that $F(A)$ is a convex set (the classical Toeplitz-Hausdorff theorem). Therefore, $\Theta(A)$ is either a point, two opposing points, an arc of $\mathbb{T}$ with the length not exceeding π, or the whole $\mathbb{T}$. The latter case occurs if and only if zero lies in the interior of $F(A)$. We will denote by $\Theta_o(A)$ the relative interior of $\Theta(A)$. Of course, in the first two cases $\Theta_o(A) = \emptyset$, while for $\Theta(A) = \mathbb{T}$, $\Theta_o(A) = \mathbb{T}$ as well.

As was already observed in [1], and more explicitly mentioned in [6], the Cauchy-Schwarz inequality immediately implies the following

PROPOSITION 1.1. *For any* $(0 \neq)A \in [\mathscr{H}]$,

$$F_N(A) \subset \overline{\mathbb{D}} \text{ and } F_N(A) \cap \mathbb{T} = \{\operatorname{sgn}\lambda : 0 \neq \lambda \in \sigma_p(A)\}. \tag{1.5}$$

Here $\sigma_p(A)$ stands for the point spectrum (that is, the set of the eigenvalues) of A, and $\operatorname{sgn} z = z/|z|$ for non-zero $z \in \mathbb{C}$. Below it will sometimes be convenient to also use $\operatorname{sgn} 0$; by convention we set it equal to one.

Points of $\overline{F_N(A)} \cap \mathbb{T}$ are associated with the approximate spectrum $\sigma_{ap}(A)$, as was observed in [7, Proposition 7]. This relation is somewhat more delicate than (1.5).

PROPOSITION 1.2. *If* $\lambda \in \sigma_{ap}(A) \setminus \{0\}$, *then* $\operatorname{sgn}\lambda \in \overline{F_N(A)}$. *Almost (but not quite) conversely, if* $e^{i\theta} \in \overline{F_N(A)}$, *then* $\{\rho e^{i\theta} : \rho \geqslant 0\} \cap \sigma_{ap}(A) \neq \emptyset$.

If $\dim \mathscr{H} = 1$, then $A = a$ is just a number, $F(a) = \{a\}$, while $F_N(a) = \{\operatorname{sgn} a\}$ for $a \neq 0$. We therefore suppose in what follows that $\dim \mathscr{H} \geqslant 2$.

Another obvious property of the NNR is its unitary invariance: $F_N(A) = F_N(U^*AU)$ for any unitary U. This is similar to the $F(A)$ behavior, and is proved in the exactly same way. On the other hand, $F_N(cA) = (\operatorname{sgn} c)F_N(A)$ for any non-zero scalar c (in particular, $F_N(cA) = F_N(A)$ if $c > 0$), while $F(cA) = cF(A)$. The translation property $F(A + cI) = c + F(A)$ however does not have any obvious analogue for F_N.

Section 2 contains some preliminary results on NNR of operators with non-zero kernels, rank one operators being the leading particular case. Next two sections deal with topological properties: the simply connectedness of $F_N(A)$ is proved in Section 3, and a description of its interior and boundary are provided in Section 4. Section 5 is devoted to some special classes of operators: weighted shifts, isometries, and (essentially) Hermitian operators. As a result, the criterion for $F_N(A)$ to have empty interior is obtained. Simplifications occurring in the finite dimensional setting are discussed in final Section 6.

2. Non-injective A

Consider $A \in [\mathscr{H}]$ with a non-trivial kernel $\ker A$. Using the decomposition $\mathscr{H} = \overline{\operatorname{Im} A^*} \oplus \ker A$, write A in the block matrix form

$$A = \begin{bmatrix} B & 0 \\ C & 0 \end{bmatrix}, \tag{2.1}$$

and parameterize a unit vector $x \in \mathscr{H}$ as

$$\begin{bmatrix} \sqrt{t}u \\ \sqrt{1-t}v \end{bmatrix}, \text{ where } t \in [0,1],\ u \in S(\overline{\operatorname{Im} A^*}), \text{ and } v \in S(\ker A). \tag{2.2}$$

Then $Ax = \sqrt{t}\begin{bmatrix} Bu \\ Cu \end{bmatrix}$, $\|Ax\| = \sqrt{t}\sqrt{\|Bu\|^2 + \|Cu\|^2}$, and thus

$$\begin{aligned} F_N(A) &= \left\{ \frac{\sqrt{t}\langle Bu,u\rangle + \sqrt{1-t}\langle Cu,v\rangle}{\sqrt{\|Bu\|^2+\|Cu\|^2}} : 0 < t \leqslant 1,\ \|u\| = \|v\| = 1 \right\} \\ &= \left\{ \frac{\sqrt{t}\langle Bu,u\rangle + \sqrt{1-t}\,\|Cu\|\,\zeta}{\sqrt{\|Bu\|^2+\|Cu\|^2}} : 0 < t \leqslant 1,\ \|u\| = 1,\ \zeta \in \overline{\mathbb{D}} \right\}. \end{aligned} \tag{2.3}$$

Formula (2.3) takes its simplest form if $C = 0$, that is, when[1]

$$\overline{\operatorname{Im} A} = \overline{\operatorname{Im} A^*}, \text{ or equivalently, } \ker A = \ker A^*. \tag{2.4}$$

Namely, if (2.4) holds for a non-injective A, then

$$F_N(A) = \left\{ \frac{\sqrt{t}\langle Bu,u\rangle}{\|Bu\|} : 0 < t \leqslant 1,\ \|u\| = 1 \right\} = \{\tau z \colon 0 < \tau \leqslant 1,\ z \in F_N(B)\}. \tag{2.5}$$

This can be restated as follows.

[1] Note that under an additional requirement that $\operatorname{Im} A$ (equivalently, $\operatorname{Im} A^*$) is closed, operators satisfying (2.4) are sometimes called *range-Hermitian*, or *EP* operators, see e.g. [4] and references therein.

Proposition 2.1. *Let A be such that $\ker A = \ker A^* \neq \{0\}$. Then the normalized numerical range of A is obtained from the normalized numerical range of its compression B onto $\overline{\operatorname{Im} A^*}$ by connecting each point z of $F_N(B)$ with the origin via a line segment $(0,z]$. In particular, $F_N(B) \subset F_N(A)$, while $0 \in F_N(A)$ if and only if $0 \in F(B)$.*

It is worth clarifying that in the setting of Proposition 2.1 the operator B is injective, and so $0 \in F_N(B)$ if and only if $0 \in F(B)$.

To state our next result, let us denote by $G(r_1,r_2)$ the subset of $\mathbb{C}$ bounded in the left half-plane by the semicircle (centered at the origin) of the radius r_1 and in the right half-plane by the elliptical arc with the vertical and horizontal semi-axes of the length r_1 and r_2, respectively. More specifically, for $0 < r_1 < r_2$ the elliptical arc is included in $G(r_1,r_2)$ while the circular one is not; their common points $\pm ir_1$ are not in the set either. When $r_1 = 0, r_2 > 0$, the set $G(0,r_2)$ degenerates into the interval $(0,r_2]$ of the real line. Finally, by convention $G(r,r) = r\mathbb{D}$ if $r > 0$, and $G(0,0) = \{0\}$.

Proposition 2.2. *Let A be a rank one operator. Then*

$$F_N(A) = \operatorname{sgn}(\operatorname{trace} A) G(r,1), \text{ where } r = \sqrt{1 - (|\operatorname{trace} A| / \|A\|)^2}.$$

Proof. For rank one operators, the block B in (2.1) is actually a scalar, thus equal $\operatorname{trace} A$, and $\|A\|^2 = \|C\|^2 + |B|^2$. Consequently, in (2.3)

$$\langle Bu, u\rangle = B, \ \|Bu\| = |B|, \text{ and } \|Cu\| = \|C\|$$

are in fact independent of u. Thus, in the case under consideration (2.3) simplifies to

$$F_N(A) = \left\{ \frac{\sqrt{t}B + \sqrt{1-t}\,\|C\|\,\zeta}{\sqrt{|B|^2 + \|C\|^2}} : 0 < t \leqslant 1, \ \zeta \in \overline{\mathbb{D}} \right\}.$$

In other words, $F_N(A)$ coincides with the NNR of the 2-by-2 matrix $\begin{bmatrix} B & 0 \\ \|C\| & 0 \end{bmatrix}$. The result for $B \neq 0$ now follows from [10, Section 3], while the case $B = 0$ is rather straightforward (though also mentioned in [9, Proposition 3]). □

Returning to multidimensional $\operatorname{Im} A$, for a unit vector $u \in \overline{\operatorname{Im} A^*}$ set

$$r_1(u) = \frac{\|Cu\|}{\sqrt{\|Bu\|^2 + \|Cu\|^2}}, \quad r_2(u) = \sqrt{\frac{|\langle Bu, u\rangle|^2 + \|Cu\|^2}{\|Bu\|^2 + \|Cu\|^2}}, \tag{2.6}$$

and

$$r(u) = \frac{r_1(u)}{r_2(u)} = \frac{\|Cu\|}{\sqrt{|\langle Bu, u\rangle|^2 + \|Cu\|^2}} \quad \text{if} \quad r_2(u) \neq 0.$$

Now observe that

$$\begin{aligned}&\left\{\frac{\sqrt{t}\langle Bu,u\rangle+\sqrt{1-t}\,\|Cu\|\,\zeta}{\sqrt{\|Bu\|^2+\|Cu\|^2}}:0<t\leqslant 1,\ \zeta\in\overline{\mathbb{D}}\right\}\\ &=r_2(u)F_N\left(\begin{bmatrix}\langle Bu,u\rangle & 0\\ \|Cu\| & 0\end{bmatrix}\right)=\operatorname{sgn}\langle Bu,u\rangle r_2(u)G(r(u),1)\\ &=\operatorname{sgn}\langle Bu,u\rangle G(r_1(u),r_2(u)).\end{aligned}\tag{2.7}$$

Combining (2.3) with (2.7), we immediately arrive at

THEOREM 2.3. *Let an operator A with a non-trivial* $\ker A$ *be represented as* (2.1). *Then*

$$F_N(A)=\bigcup_{u\in S(\overline{\operatorname{Im}A^*})}\operatorname{sgn}\langle Bu,u\rangle G(r_1(u),r_2(u)),\tag{2.8}$$

with $r_1(u),r_2(u)$ defined by (2.6).

Of course, Proposition 2.1 follows also from Theorem 2.3, since $r_1(u)\equiv 0$, and so all $G(r_1(u),r_2(u))$ turn into line segments stemming from zero.

On the other hand, if $\operatorname{Im}A\not\perp\ker A$, then $Cu\neq 0$ for at least one vector u, the respective value of $r_1(u)$ is then positive, and $G(r_1(u),r_2(u))$ contains the disk $r_1(u)\mathbb{D}$. Due to (2.8), then

$$F_N(A)\supset r(A)\mathbb{D},\ \text{where } r(A)=\sup\{r_1(u)\colon u\in\operatorname{Im}A^*\}.\tag{2.9}$$

From (2.6) we easily obtain

$$r(A)=\begin{cases}1 & \text{if } B \text{ is not invertible}\\ \rho/\sqrt{1+\rho^2}, & \text{where } \rho=\left\|CB^{-1}\right\| \text{ otherwise.}\end{cases}\tag{2.10}$$

To provide an alternative characterization of $r(A)$ recall that the minimal angle $\angle(\mathscr{M},\mathscr{N})$ between two subspaces of the Hilbert space is defined as $\cos^{-1}\sup\{\langle x,y\rangle\colon x\in S(\mathscr{M}),y\in S(\mathscr{N})\}$. A direct computation shows that for $\mathscr{M}=\operatorname{Im}A$, $\mathscr{N}=\ker A$ the supremum involved equals $r(A)$, and so

$$r(A)=\cos\angle(\operatorname{Im}A,\ker A).$$

In particular, $r(A)=1$ if and only if the minimal angle between $\operatorname{Im}A$ and $\ker A$ is zero.

COROLLARY 2.1. *Let the minimal angle between* $\operatorname{Im}A$ *and* $\ker A$ *be zero. Then* $F_N(A)=\mathbb{D}\cup\{\operatorname{sgn}\lambda\colon 0\neq\lambda\in\sigma_p(A)\}$.

Proof. Due to Proposition 1.1, we only need to show that $F_N(A)\supset\mathbb{D}$. This follows from (2.8), combined with $r(A)=1$. □

We will call $\mathscr{C}_A:=r(A)\mathbb{T}$ the *critical circle* of A. Of course, $\mathscr{C}_A=\{0\}$ if $\operatorname{Im}A\perp\ker A$ and is a proper circle otherwise. Using Proposition 2.1 for $r(A)=0$ and formula (2.9) for $r(A)>0$, we conclude:

COROLLARY 2.2. *If A is not injective, then $0 \in \overline{F_N(A)}$. If in addition* $\operatorname{Im} A$ *is not orthogonal to* $\ker A$*, then moreover $0 \in \operatorname{int} F_N(A)$ while $\mathscr{C}_A \subset \overline{F_N(A)}$.*

3. Shape of NNR: simply connectedness

Corollary 2.1 and Proposition 2.2 above show that for some classes of operators A, $F_N(A)$ is convex along with $F(A)$; more examples of this kind are available in Section 5. Proposition 2.2 implies in particular that $F_N(A)$ is convex for all non-invertible $A \in \mathbb{C}^{2\times 2}$. Moreover, convexity of $F_N(A)$ persists for some invertible 2-by-2 matrices, in particular for those with the spectrum $\sigma(A) = \{\lambda\}$, or $\{\lambda, -\lambda\}$, $\lambda \neq 0$. Indeed, for such matrices $F_N(A)$ is an elliptical disk [9, Propositions 4 and 6], degenerating into line segments when A is normal.

Nevertheless, in general $F_N(A)$ is not convex. The simplest example is delivered by

$$A = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} =: \operatorname{diag}[\lambda_1, \lambda_2] \quad \text{with} \quad \lambda_1, \lambda_2 \neq 0,\ \lambda_1/\lambda_2 \notin \mathbb{R}, \tag{3.1}$$

considered in [8, 9]. Since this example is important for some other reasons as well, for convenience of reference we provide here the explicit description of $F_N(A)$ for A given by (3.1). It is a slight rewording of the statement from [8]. The following notation will be useful.

Given $\theta_1, \theta_2, k \in \mathbb{R}$, denote by $\Gamma(\theta_1, \theta_2, k)$ the arc of the quadratic

$$\begin{aligned}(\sin^2\theta_1 + \sin^2\theta_2 - k\sin\theta_1\sin\theta_2)x^2 + (k\sin(\theta_1+\theta_2) - \sin 2\theta_1 - \sin 2\theta_2)xy \\ + (\cos^2\theta_1 + \cos^2\theta_2 - k\cos\theta_1\cos\theta_2)y^2 = \sin^2(\theta_1 - \theta_2),\end{aligned} \tag{3.2}$$

connecting the points $e^{i\theta_1}$ and $e^{i\theta_2}$.

PROPOSITION 3.1. *Let A be given by* (3.1). *Then $F_N(A) = \Gamma(\theta_1, \theta_2, k)$, where $\theta_j = \arg \lambda_j$ and $k = |\lambda_1/\lambda_2| + |\lambda_2/\lambda_1|$.*

If $|\lambda_1| = |\lambda_2|$, then $k = 2$, and $\Gamma(\theta_1, \theta_2, k)$ degenerates into a chord of $\mathbb{T}$; thus, the convexity still holds. However, whenever $|\lambda_1| \neq |\lambda_2|$, we have $k > 2$, and $\Gamma(\theta_1, \theta_2, k)$ is an arc of a hyperbola. The set $F_N(A)$ is then not convex, and not even star-shaped. Note that for A as in (3.1), $0 \notin F(A)$, and thus also $0 \notin F_N(A)$.

The next statement shows that $F_N(A)$ is star-shaped with respect to the origin whenever an obvious necessary condition $0 \in F_N(A)$ holds.

THEOREM 3.2. *If $A \in [\mathscr{H}]$ is such that $0 \in F_N(A)$, then $F_N(A)$ is star-shaped with respect to the origin.*

Formally speaking, Theorem 3.2 is a corollary of the following

THEOREM 3.3. *For any $A \in [\mathscr{H}]$ and any line ℓ passing through the origin, the intersection $F_N(A) \cap \ell$ is connected.*

Conversely, under condition $0 \in F_N(A)$ the statement of Theorem 3.3 is equivalent to that of Theorem 3.2. We will thus proceed by proving both theorems in parallel.

Proof. (i) Suppose first that A is injective, and consider any line ℓ passing through the origin. If $F_N(A) \cap \ell$ is a singleton or empty, it is of course connected. If it contains more than one point, pick any two arbitrarily, say $z_0, z_1 \in F_N(A) \cap \ell$, $z_0 \neq z_1$. By definition, there exist unit vectors $x_0, x_1 \in \mathscr{H}$ such that

$$z_j = \langle Ax_j, x_j \rangle / \|Ax_j\|, \quad j = 0, 1.$$

Note that x_0, x_1 are linearly independent, since otherwise z_0 and z_1 would coincide. As was shown in the proof of convexity of the classical numerical range in [12], it is possible to replace one of the vectors x_j by its product with an appropriate unimodular factor in such a way that

$$\langle Ax(t), x(t) \rangle \in \ell \quad \text{for} \quad x(t) = tx_0 + (1-t)x_1 \quad \text{and all} \quad t \in [0,1]. \tag{3.3}$$

But then also

$$\phi(t) := \frac{\langle Ax(t), x(t) \rangle}{\|x(t)\| \, \|Ax(t)\|} \in \ell, \quad t \in [0,1]. \tag{3.4}$$

The function ϕ defined by (3.4) is continuous on $[0,1]$ – this is the place in the proof where the injectivity of A is invoked, – and takes the values in $F_N(A) \cap \ell$. Since $\phi(j) = z_j$, $j = 0, 1$, the whole interval $[z_0, z_1]$ of ℓ lies in $F_N(A)$. This proves Theorem 3.3 (and thus Theorem 3.2 as well) for injective A.

(ii) Let now A be such that $\ker A \not\subset \operatorname{Im} A$. Then $0 \in F_N(A)$ by Corollary 2.2, and we need to show that the statement of Theorem 3.2 holds. To this end, observe that the sets $G(r_1, r_2)$ are star-shaped with respect to the origin if $r_1 > 0$, become star-shaped with respect to the origin when joined by 0 if $r_1 = 0$, and remain such under any rotation around the origin. So, the union in the right hand side of (2.8) in our setting is also star-shaped.

(iii) It remains to consider the case when $\ker A = \ker A^* \neq \{0\}$. According to Proposition 2.1, $F_N(A)$ has the form $\bigcup_{z \in F_N(B)} \langle 0, z]$, with B being injective (it is actually the compression of A onto $\ker A^\perp$, but this detail is not important right now). The notation indicates that the zero endpoint is included if $0 \in F_N(B)$ and excluded otherwise. The statement of Theorem 3.2 is now immediate. To take care of Theorem 3.3 we thus only need to consider the case $0 \notin F_N(B)$. Equivalently, $0 \notin F(B)$, implying that for any $z_1, z_2 \in F(B)$ the ratio z_1/z_2 is not a real negative number. The same is then true for any two points in $F_N(B)$. Consequently, the intersection of a line ℓ passing through the origin with $F_N(A)$ is either empty, or an interval with an endpoint zero (not included).

This exhausts all the possibilities, so the proof is complete. □

To illustrate the applicability of Theorem 3.3, here is a useful addition to Proposition 2.1.

COROLLARY 3.1. *Let A satisfy* (2.4)*, and let $0 \in F(B)$, where B is the compression of A onto the closure of its range. Then $F_N(A) = F_N(B)$.*

Proof. This is a tautology if $\ker A = \{0\}$. Otherwise, invoke the description of $F_N(A)$ from Proposition 2.1, observing that when $0 \in F(B)$ the line segments connecting it with points in $F_N(B)$ all lie in $F_N(B)$. □

It is a simple consequence of Theorem 3.3 that for every point $\alpha \notin F_N(A)$ there is a ray emanating from α not intersecting $F_N(A)$. Indeed, let $\ell(\alpha)$ be the line passing through α and the origin (any of these lines if $\alpha = 0$). Then $\ell(\alpha) \setminus F_N(A)$ is either the whole line or the union of two rays, one of which contains α. Such sets, if in addition connected, were called *ray connected* in [2]. It was also observed there that complements of these sets are connected, and therefore the sets themselves are simply connected. The connectedness (and even path-connectedness) of NNR was established in [6, Proposition 7]. So, we have

COROLLARY 3.2. *For any $A \in [\mathscr{H}]$, the set $F_N(A)$ is simply connected.*

Theorem 3.3 has another interesting consequence. Recall that a regular numerical range has the property that $F(B) \subset F(A)$ whenever B is a compression of A. If B is in fact a restriction of A onto its invariant subspace, the inclusion holds for the NNR as well. This was used in the considerations of Section 2 for A, B as in (2.1). However, it fails in general. Moreover, the following criterion holds.

THEOREM 3.4. *An operator A possesses the property $F_N(B) \subset F_N(A)$ for all its compressions B if and only if $\sigma_p(A)$ intersects with every ray ℓ_θ passing through $\Theta(A)$.*

Note that, though we agreed to consider only non-zero operators, the statement is vacuously correct for $A = 0$.

Proof. Necessity. For $e^{i\theta} \in \Theta(A)$, pick any $z \in F(A, \theta)$, a unit vector x such that $\langle Ax, x\rangle = z$, and let B_z be the compression of A onto the span of x. Then $F_N(B_z) = e^{i\theta}$, an so $e^{i\theta} \in F_N(A)$. By Proposition 1.1, there exists $\lambda \in \sigma_p(A)$ with $\operatorname{sgn}\lambda = e^{i\theta}$.

Sufficiency. Let B be the compression of A onto a subspace L. Take a unit vector $x \in L$ for which $Bx \neq 0$; then of course $Ax \neq 0$ as well; moreover, $\|Ax\| \geqslant \|Bx\|$. Since $\langle Bx, x\rangle = \langle Ax, x\rangle$, the point $\langle Bx, x\rangle / \|Bx\|$ of $F_N(B)$ lies on the interval with the endpoints $z = \langle Ax, x\rangle / \|Ax\|$ and $\operatorname{sgn} z$. But $z \in F_N(A)$, and so we are given that there exists $\lambda \in \sigma_p(A)$ with $\operatorname{sgn}\lambda = \operatorname{sgn} z$. Using Proposition 1.1 again, we conclude that $\operatorname{sgn} z \in F_N(A)$. Finally, Theorem 3.3 implies that $\langle Bx, x\rangle / \|Bx\| \in F_N(A)$. □

COROLLARY 3.3. *For a Hermitian A the property $F_N(B) \subset F_N(A)$ holds for all its compressions B if and only if either A is positive semi-definite with some positive eigenvalues, or negative semi-definite with some negative eigenvalues, or indefinite having both positive and negative eigenvalues.*

A similar result obviously holds for scalar multiples of Hermitian operators. On the other hand, in a separable Hilbert space no other operators can possibly satisfy conditions of Theorem 3.4. Indeed, the point spectrum in this setting is at most countable.

4. Shape of NNR: boundary, interior, closedness

4.1. Auxiliary considerations

Recall the definition of $F_N(A,\theta)$ as the intersection of $F_N(A)$ with the ray (1.3). By Theorem 3.3, $F_N(A,\theta)$ is an interval, possibly degenerating into a point, for any $e^{i\theta} \in \Theta(A)$. Denote its endpoints by $\gamma(\theta)$ and $\Gamma(\theta)$, with $|\gamma(\theta)| \leqslant |\Gamma(\theta)|$. Of course, $|\Gamma(\theta)| > 0$ whenever $e^{i\theta} \in \Theta(A)$, and $\gamma(\theta) \equiv 0$ if $0 \in \overline{F_N(A)}$. According to [7, Proposition 5]

$$0 \in \overline{F_N(A)} \Longleftrightarrow 0 \in \overline{F(A)}, \tag{4.1}$$

and these inclusions hold, in particular, when $\Theta(A) = \mathbb{T}$ or A is not injective. So, in these cases we have $(0 =) |\gamma(\theta)| < |\Gamma(\theta)|$. Our next statement shows that the strict inequality holds whenever $F(A,\theta)$ is a non-degenerate interval, even if $\gamma(\theta) > 0$.

PROPOSITION 4.1. *If $F(A,\theta)$ has positive length, then so does $F_N(A,\theta)$.*

Proof. We start by slightly modifying the approach of part (i) in the proof of Theorem 3.3. Namely, we again pick two distinct points z_0, z_1, but this time in $F(A,\theta)$, not in $F_N(A)$. Then, introduce unit vectors $x_0, x_1 \in \mathscr{H}$ for which $z_j = \langle Ax_j, x_j\rangle$, $j = 0,1$. Rotate one of the vectors to enforce (3.3). The values of the function ϕ defined by (3.4) will then lie in $F_N(A,\theta)$, and we only need to make sure that this function is not constant on $[0,1]$. To this end, observe that

$$\|x(t)\|^2 = t^2 + (1-t)^2 + 2t(1-t)\operatorname{Re}\langle x_0, x_1\rangle$$

is a quadratic polynomial in t with distinct roots, and not a scalar multiple of

$$\langle Ax(t), x(t)\rangle = z_0 t^2 + z_1(1-t)^2 + t(1-t)(\langle Ax_0, x_1\rangle + \langle Ax_1, x_0\rangle).$$

So, ϕ^2 is a rational function of t with a non-empty set of poles, and thus not constant on any interval. The latter is therefore true for ϕ as well. □

We now turn to the description of boundary and interior point of $F_N(A)$. For $\Theta(A)$ being a single point or two opposing points, the situation is simple: due to Theorem 3.3, $F_N(A)$ is an interval, and so $\operatorname{int} F_N(A) = \emptyset$, $\partial F_N(A) = \overline{F_N(A)}$. Note that this situation occurs if and only if A is a scalar multiple of an Hermitian operator. The explicit location of the endpoints of $F_N(A)$, and criteria for them (not) to lie in $F_N(A)$, follow from the discussion in Section 5.3 below. The remaining results of this section are formally correct (though trivial) for such A, and in their proofs we therefore silently suppose that $\Theta(A)$ is an arc of positive length.

LEMMA 4.2. *For any $A \in [\mathscr{H}]$:*

(i) *$|\Gamma|$ and $|\gamma|$ are lower (resp., upper) semi-continuous functions of θ for $e^{i\theta} \in \Theta(A)$;*

(ii) *If $F_N(A)$ is closed, then Γ and γ are continuous functions of θ;*

(iii) *If $e^{i\theta} \in \Theta_o(A)$, then $(\gamma(\theta), \Gamma(\theta)) \subset \operatorname{int} F_N(A)$.*

Proof. Given $e^{i\theta_0} \in \Theta(A)$, choose $e^{i\theta_1} \in \Theta(A)$ different from $\pm e^{i\theta_0}$. Pick now $z_j \in F_N(A,\theta_j)$, $j=0,1$. Proceeding as in the proof of Proposition 4.1, we generate a curve $\eta \subset F_N(A)$ connecting z_0 with z_1 and intersecting $F_N(A,\theta)$ at exactly one point $\eta(\theta)$ for θ lying between θ_0 and θ_1. If $e^{i\theta_0} \in \Theta_o(A)$ this construction can be applied to both one sided neighborhoods. Either way, we obtain a continuous curve $\eta \subset F_N(A)$, with $\eta(\theta)$ defined correctly on some neighborhood $\mathscr{O}$ of $e^{i\theta_0}$ in $\Theta(A)$, and such that $\eta(\theta_0) = z_0$.

Obviously, $|\Gamma(\theta)| \geqslant |\eta(\theta)| \geqslant |\gamma(\theta)|$ for $e^{i\theta} \in \mathscr{O}$. Passing to the limit when $\theta \to \theta_0$, we thus obtain:

$$\liminf_{\theta\to\theta_0} |\Gamma(\theta)| \geqslant |z_0| \geqslant \limsup_{\theta\to\theta_0} |\gamma(\theta)|.$$

Since z_0 is an arbitrary point of $F_N(A,\theta_0)$, the last inequality can be strengthened to

$$\liminf_{\theta\to\theta_0} |\Gamma(\theta)| \geqslant |\Gamma(\theta_0)|, \quad \limsup_{\theta\to\theta_0} |\gamma(\theta)| \leqslant |\gamma(\theta_0)|. \tag{4.2}$$

This proves statement (i).

To prove (ii), consider again $e^{i\theta_0} \in \Theta(A)$ and suppose that for some sequence $e^{i\theta_k} \in \Theta(A)$ converging to $e^{i\theta_0}$ we have $\Gamma(\theta_k) \to z$. Since $F_N(A)$ is closed, it contains all the points $\Gamma(\theta_k)$, and thus $z \in F_N(A)$ as well. At the same time $z \in \ell_{\theta_0}$, and so $|z| \leqslant |\Gamma(\theta_0)|$. Comparing this with the first inequality in (4.2), we conclude that $|z| = |\Gamma(\theta_0)|$, and thus simply $z = \Gamma(\theta_0)$. The proof for γ is similar, using the second inequality in (4.2).

When proving (iii), we need only to consider the case when $\gamma(\theta) \neq \Gamma(\theta)$, since the interval in question is void otherwise. Let us repeat the construction above using two distinct points, say $z_\pm \in (\gamma(\theta_0), \Gamma(\theta_0))$ with $|z_+| > |z_-|$, in place of z_0. The resulting continuous functions $\eta_\pm$ will therefore satisfy the inequality $|\eta_+(\theta)| > |\eta_-(\theta)|$ on some neighborhood (θ', θ'') of θ_0.

By Corollary 3.2, the domain bounded by the curves η_+, η_- and the rays $\ell_{\theta'}, \ell_{\theta''}$ lies in $F_N(A)$. Consequently, $(z_-, z_+) \subset \operatorname{int} F_N(A)$. Due to the freedom in choosing $z_\pm$, the latter inclusion yields (iii). □

4.2. Boundary and interior

Denote $\gamma(A) = \{\gamma(\theta)\colon e^{i\theta} \in \Theta(A)\}$ and $\Gamma(A) = \{\Gamma(\theta)\colon e^{i\theta} \in \Theta(A)\}$. Also, let $e^{i\theta_\pm}$ be the endpoints of $\Theta(A)$ when it differs from $\mathbb{T}$.

THEOREM 4.3. *For any $A \in [\mathscr{H}]$,*

$$\Gamma(A) \subset \partial F_N(A). \tag{4.3}$$

If in addition $\Theta(A) = \mathbb{T}$, then

$$\operatorname{int} F_N(A) = \cup_{\theta=0}^{\pi} (\Gamma(\theta), \Gamma(\theta+\pi)). \tag{4.4}$$

Otherwise, we also have

$$F_N(A,\theta_+) \cup F_N(A,\theta_-) \cup \gamma(A) \subset \partial F_N(A), \tag{4.5}$$

while

$$\operatorname{int} F_N(A) = \cup_{e^{i\theta}\in\Theta_0(A)}(\gamma(\theta),\Gamma(\theta)). \tag{4.6}$$

Proof. We will first treat the statement concerning the boundary points.

If θ is such that $\gamma(\theta)=\Gamma(\theta)$, then $F_N(A,\theta)$ collapses to a singleton. This singleton lies in $\partial F_N(A)$, since it is the only point of $F_N(A,\theta)$.

Let now $\gamma(\theta)\neq\Gamma(\theta)$. Since $(\gamma(\theta),\Gamma(\theta))\subset F_N(A)$, both $\gamma(\theta)$ and $\Gamma(\theta)$ lie in $\overline{F_N(A)}$. On the other hand, $t\Gamma(\theta)\notin F_N(A)$ for $t>1$, and so $\Gamma(\theta)\in\partial F_N(A)$. This proves (4.3).

Similarly, $t\gamma(\theta)\notin F_N(A)$ for $t<1$, and so $\gamma(\theta)\in\partial F_N(A)$ *if* $\gamma(\theta)\neq 0$. For $\gamma(\theta)=0$ this reasoning is not applicable. But $\Theta(A)\neq\mathbb{T}$ implies that $F_N(A)$ lies to one side of some line passing through the origin, and $0\notin\operatorname{int}F_N(A)$ because of that. Since the inclusions $F_N(A,\theta_\pm)\subset\partial F_N(A)$ are obvious, (4.5) also holds.

Consider now the interior of $F_N(A)$. If $\Theta(A)$ is a proper subarc of $\mathbb{T}$, (4.5) implies that the left hand side of (4.6) is contained in its right hand side. The converse follows from part (iii) of Lemma 4.2. Finally, for $\Theta(A)=\mathbb{T}$ we have $\gamma(\theta)\equiv 0$, and so $\cup_{\theta=0}^{2\pi}(\gamma(\theta),\Gamma(\theta))$ equals the right hand side of (4.4) with the origin deleted. Since $\Gamma(A)$ lies in the boundary, to complete the proof we only need to show that $0\in\operatorname{int}F_N(A)$. But $0\in F_N(A)$ by Theorem 3.3, while the positive-valued function $|\Gamma(\theta)|$, being lower semi-continuous due to part (i) of Lemma 4.2, attains its infimum and thus is bounded away from zero by some $\rho>0$. So, $\rho\mathbb{D}\subset F_N(A)$, and we are done. □

Comparing formulas (4.4) and (4.6) we see that $0\in\operatorname{int}F_N(A)$ if and only if $\Theta(A)=\mathbb{T}$. Since the latter, in turn, is equivalent to $0\in\operatorname{int}F(A)$, we arrive at the following

COROLLARY 4.1. *For any $A\in[\mathscr{H}]$, $0\in\operatorname{int}F_N(A)$ if and only if $0\in\operatorname{int}F(A)$.*

Recalling (4.1), we conclude from Corollary 4.1 that also

$$0\in\partial F_N(A)\Longleftrightarrow 0\in\partial F(A). \tag{4.7}$$

Theorem 4.3 when combined with part (ii) of Lemma 4.2 easily yields a complete description of $\partial F_N(A)$ provided that $F_N(A)$ is closed.

THEOREM 4.4. *Let $F_N(A)$ be closed. Then*

$$\partial F_N(A)=\begin{cases}\Gamma(A) & \text{if } 0\in\operatorname{int}F(A),\\ \Gamma(A)\cup\gamma(A)\cup F_N(A,\theta_+)\cup F_N(A,\theta_-) & \text{otherwise.}\end{cases} \tag{4.8}$$

Proof. The right hand side of (4.8) always lies in $\partial F_N(A)$ according to (4.3) and (4.5). When $F_N(A)$ is closed, due to Lemma 4.2(ii), it is actually a simple closed curve. Moreover, $\operatorname{int}F_N(A)$ coincides with its interior region due to (4.4),(4.6). By the Jordan curve theorem, it implies the equality in (4.8). □

For $0\in\partial F(A)$ due to (4.7) we have $\gamma(A)=\{0\}$. If in addition $F_N(A)$ is closed, then also $0\in F_N(A,\theta_\pm)$. So, in this case $\gamma(A)$ can be dropped in the right hand side of (4.8).

4.3. (Non)closedness

Prompted by Theorem 4.8 in particular, it is natural to seek verifiable tests for $F_N(A)$ to be closed. A necessary condition follows immediately from Corollary 2.2. Namely:

LEMMA 4.5. *Let $A \in [\mathscr{H}]$ be non-injective and such that $F_N(A)$ is closed. Then $\mathscr{C}_A \subset F_N(A)$.*

We will show in Section 6 that in finite dimensional setting this condition is also sufficient. For now, we will use it to single out a class of operators with non-closed $F_N(A)$.

THEOREM 4.6. *Let in representation (2.1) of A the block C be non-zero while $0 \notin \operatorname{int} F(B)$. Then $F_N(A)$ is not closed.*

Proof. We will use the description of $F_N(A)$ provided by (2.8). The intersection of the sets $\operatorname{sgn}\langle Bu,u\rangle G(r_1(u),r_2(u))$ with the critical circle $\mathscr{C}_A$ is an open arc of $\mathscr{C}_A$ centered at $\operatorname{sgn}\langle Bu,u\rangle \rho/\sqrt{\rho^2+1}$ and of radian measure not exceeding π. Due to the condition imposed on $F(B)$, their union will not cover the whole $\mathscr{C}_A$. □

EXAMPLE 1. Let in (2.1) $B = \begin{bmatrix} 1 & 0 \\ 0 & i \end{bmatrix}$, $C \neq 0$.

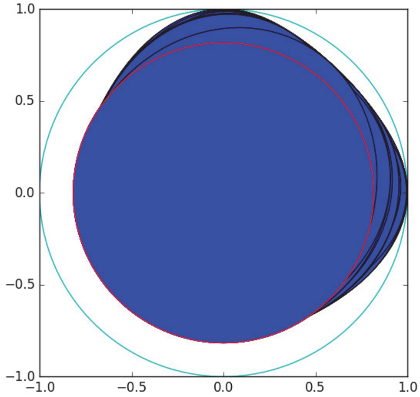

Figure 1: *The normalized numerical range obtained for $C = [1,-1]$. The black lines are the elliptical arcs corresponding to the G-sets, and the red circle is the critical circle.*

Then $\rho = \|C\|$, and the portion of $\mathscr{C}_A$ located in the third quadrant lies in $\partial F_N(A)$ but not in $F_N(A)$. The set $F_N(A)$ is neither open nor closed.

5. Special classes of operators

5.1. Weighted shifts

In this subsection we suppose that $\mathscr{H}$ is infinite dimensional but separable. Suppose that for some choice of an orthonormal basis $\{e_j\}$ of $\mathscr{H}$ and $\alpha_j \in \mathbb{C}$ we have $Te_j = \alpha_j e_{j+1}$. Then T extends to a linear bounded operator on $\mathscr{H}$ if and only if the sequence α_j is bounded; it is called a *right* (resp. *bilateral*) *weighted shift* if the basis is labeled by $\mathbb{N}$ (resp., $\mathbb{Z}$). An operator is a *left weighted shift* if its adjoint is a right weighted shift; right and left shifts are also called *unilateral*.

It is a well known and simple observation that the shifts with the weight sequences $\{\alpha_j\}$ and $\{|\alpha_j|\}$ are unitarily similar. An immediate (and also known) consequence of this observation is that T and ωT are unitarily similar for any $\omega \in \mathbb{T}$, and so the spectra (along with their point, approximate, and residual parts) of weighted shifts, as well as the numerical ranges, are circularly symmetric. The same is of course true for NNR.

For any shift T both $F(T)$ and $F_N(T)$ are therefore circular disks, open or closed, and centered at the origin. For $F(T)$ it follows from its a priori convexity, while for $F_N(T)$ Theorem 3.3 does the job.

Computing the radius of $F(A)$ and deciding whether this set is open or closed is a non-trivial task, see e.g. [17, 16], or more recent [19]. For $F_N(T)$ the situation is often simpler. To describe it, recall that the spectral radius of a weighted shift T with the weight sequence $\{\alpha_j\}$ equals

$$r(T) = \lim_{n\to\infty} \sup_k |\alpha_{k+1}\cdots\alpha_{k+n}|^{1/n}.$$

Denote also

$$r_1(T) = \liminf_{n\to\infty} |\alpha_{m+1}\cdots\alpha_{m+n}|^{1/n},$$

with m being the smallest index for which $\alpha_j \neq 0$ if $j > m$. By convention, $r_1(T) = 0$ if such m does not exist.

THEOREM 5.1. *Let T be a unilateral weighted shift for which either $r(T) > 0$ or at least one zero weight is preceded by non-zero ones. Then $F_N(T)$ is the open unit disk $\mathbb{D}$ if T is a left shift or $r_1(T) = 0$, and the closed unit disk $\overline{\mathbb{D}}$ otherwise.*

Proof. If say $\alpha_m = 0$ while not all $\alpha_1,\dots,\alpha_{m-1}$ are zeros, then the span $\mathscr{L}$ of $\{e_1,\dots,e_m\}$ is invariant under T, and the restriction T_0 of T onto $\mathscr{L}$ is nilpotent but non-zero. But then $\ker T \cap \operatorname{Im} T \neq \{0\}$, by Corollary 2.1 $F_N(T_0) = \mathbb{D}$, and so $F_N(T) \supset \mathbb{D}$. (See Corollary 6.1 for some additional information on the finite-dimensional nilpotent case.) If $r(T) > 0$ (that is, the operator T is not quasinilpotent), the same inclusion holds by Proposition 1.2 since the circle $\{z\colon |z| = r(T)\}$ lies in $\sigma_{ap}(T)$. Due to Proposition 1.1, $F_N(T)$ is therefore the open or closed unit disk, depending on whether or not the set $\sigma_p(T)\setminus\{0\}$ is empty. It remains to recall that the right weighted shift never has non-zero eigenvalues, while for the left weighted shift they exist if and only if $r_2(T) > 0$. □

Passing to bilateral shifts, observe first of all that in case of at least one zero weight T splits into the direct sum of two unilateral shifts. So, the description of $F_N(T)$ in this case can be easily derived from Theorem 5.1 when applicable. We will therefore concentrate on the situation of non-zero weights only. Denote

$$r_+ = \limsup_{n\to\infty} |\alpha_1 \cdots \alpha_n|^{1/n}, \quad r_- = \liminf_{n\to\infty} |\alpha_{-1} \cdots \alpha_{-n}|^{1/n}.$$

THEOREM 5.2. *Let T be a bilateral weighted shift with all non-zero weights α_j and $r(T) > 0$. Then $F_N(T)$ is the closed unit disk if*

(i) *$r_+ < r_-$, or*

(ii) *$r_+ = r_- := r_0 > 0$ and*

$$\sum_{n=1}^{\infty} \frac{|\alpha_1 \cdots \alpha_n|^2}{r_0^{2n}} + \sum_{n=1}^{\infty} \frac{r_0^{2n}}{|\alpha_{-1} \cdots \alpha_{-n}|^2} < \infty, \tag{5.1}$$

and the open unit disk otherwise.

Proof. As in the setting of Theorem 5.1, condition $r(T) > 0$ guarantees non-emptyness of $\sigma_{ap}(T)$. Due to its circularity, Proposition 1.2 implies that $F_N(T) \supset \mathbb{D}$. In addition, (i)$\vee$(ii) is equivalent to $\sigma_p(T) \setminus \{0\}$ being non-empty, and so Proposition 1.1 finishes the job. □

Consider in particular the *unweighted* right shift S and bilateral shift B, for which $\alpha_j \equiv 1$. We then have $r = r_1 = r_0 = r_+ = r_- = 1$, while convergence condition (5.1) fails. So, $F_N(S) = F_N(B) = \mathbb{D}$ and $F_N(S^*) = \overline{\mathbb{D}}$. The latter equality was established in [6].

5.2. Isometries

If $A \in [\mathscr{H}]$ is an isometry, then by definition $\|Ax\| = \|x\|$ for all $x \in \mathscr{H}$, and so $F_N(A) = F(A)$. In order to describe this set explicitly, recall now the Wold decomposition (see e.g. [18, Theorem 1.1]). According to the latter, A is unitarily similar to the direct sum of several (possibly, none) unweighted right shifts S and (also, possibly missing) unitary U. More specifically, there are no S-summands if and only if A itself is unitary; on the other hand, absence of U corresponds to a *pure isometry* A, that is, the case $\bigcap_{k=1}^{\infty} \operatorname{Im} A^k = \{0\}$.

THEOREM 5.3. (i) *For a unitary operator U,*

$$F_N(U) = \operatorname{int} \operatorname{conv} \sigma(U) \cup \sigma_p(U) \cup \Delta(U),$$

where conv *denotes the convex hull and $\Delta(U)$ stands for the union of the line segments connecting the points of $\sigma_p(U)$, if any, the arcs of $\mathbb{T}$ between which are disjoint with $\sigma(U)$;*

(ii) *If an isometry A is not unitary, then $F_N(A) = \mathbb{D} \cup \sigma_p(A)$.*

Proof. Part (i) follows directly from the description of numerical ranges for normal operators from [5]. For (ii), we utilize the fact that A has an invariant subspace on which it coincides with S, and so $F_N(A) \supset F_N(S) = \mathbb{D}$. Also, $\sigma_p(A) \subset \mathbb{T}$. It remains to invoke (1.5). □

Note that in the finite dimensional case any isometry is unitary, so the second option does not materialize, and $F_N(A) = \operatorname{conv} \sigma(A)$ is closed. At another extreme, pure isometries do not have eigenvectors, and so for them $F_N(A) = \mathbb{D}$ is open. Of course, there are plenty of unitary operators with the empty point spectrum, for which therefore the NNR is open as well.

5.3. Hermitian operators

Let $A \in [\mathscr{H}]$ be Hermitian, that is, $A = A^*$. The spectrum of A is real; denote

$$m = \min \sigma(A) \text{ and } M = \max \sigma(A). \tag{5.2}$$

Then $F(A)$ is the interval with the endpoints m and M, and so $F_N(A)$ also is an interval in $\mathbb{R}$. Next statement provides its explicit description.

THEOREM 5.4. *For a Hermitian A, in the notation* (5.2) *we have:*

(i) *The interval $F_N(A)$ has the endpoints* 1 *and* $\frac{2\sqrt{mM}}{m+M}$ *if A is positive semi-definite (that is, $m \geqslant 0$); -1 and $-\frac{2\sqrt{mM}}{m+M}$ if A is negative semi-definite (that is, $M \leqslant 0$), and ± 1 if A is indefinite ($mM < 0$).*

Moreover:

(ii) *The endpoint ± 1 belongs to $F_N(A)$ if and only if A has at least one positive (resp., negative) eigenvalue,*

while

(iii) $\pm\frac{2\sqrt{mM}}{m+M} \in F_N(A)$ *if and only if $mM > 0$ and $m, M \in \sigma_p(A)$.*

Proof. For Hermitian operators $\sigma(A) = \sigma_{ap}(A)$, and so by Proposition 1.2 one is an endpoint of $F_N(A)$ if and only if $M > 0$, while by (1.5) this endpoint belongs to $F_N(A)$ if and only if $(0, M] \cap \sigma_p(A) \neq \emptyset$. Similarly, -1 is an endpoint of $F_N(A)$ if and only if $m < 0$, and $-1 \in F_N(A)$ if and only if $[m, 0) \cap \sigma_p(A) \neq \emptyset$. This proves (ii) and the portion of (i) concerning the indefinite case and the outmost endpoint in the semi-definite case $mM \geqslant 0$. It remains thus to prove (iii) and the formula in (i) for the second endpoint when $mM \geqslant 0$. It suffices to consider a positive semi-definite A only; the negative semi-definite case will then follow by passing from A to $-A$.

So, we need only to establish that, when $0 \leqslant m \leqslant M$,

$$\inf \frac{\langle Ax, x\rangle}{\|Ax\| \, \|x\|} = \frac{2\sqrt{mM}}{m+M}, \tag{5.3}$$

and this infimum is attained if and only if $0 < m, M \in \sigma_p(A)$.

If A is uniformly positive, that is, $m > 0$, the equality (5.3) appears as Problem 33 in [3], with a reference to [14] (for the latter, see also its English translation [15]). For the sake of self-containment note that the method of [14, 15] amounts to invoking the spectral decomposition $A = \int_{\sigma(A)} \lambda \, dE_\lambda$ of A, due to which $\frac{\langle Ax,x\rangle}{\|Ax\|\|x\|}$ can be rewritten as

$$\left(\int_{\sigma(A)} \lambda \, d\langle E_\lambda x, x\rangle\right) \Big/ \left(\int_{\sigma(A)} \lambda^2 \, d\langle E_\lambda x, x\rangle \int_{\sigma(A)} d\langle E_\lambda x, x\rangle\right)^{1/2}, \tag{5.4}$$

and further construction of a sequence $\{x_k\} \subset \mathscr{H}$ minimizing (5.4).

It is not hard to see that the requirement $m \neq 0$ is redundant, and equality (5.3) holds for $M > m = 0$ as well. Moreover, if $m, M \in \sigma_p(A)$, f_m, f_M are eigenvectors corresponding to $m(>0)$ and M, respectively, then the infimum is attained on vectors collinear to $\sqrt{M} f_m + \sqrt{m} f_M$, and is not attained in all other cases. □

5.4. Essentially Hermitian operators

By definition, and operator A is *essentially Hermitian* if it is a linear combination of a Hermitian operator and the identity: $A = aH + bI$, where $H = H^*$, $a, b \in \mathbb{C}$. Essentially Hermitian operators are in fact normal, with the spectrum lying in a straight line: $\sigma(A) \subset \ell$. Equivalently, A is essentially Hermitian if and only if $F(A) \subset \ell$, that is, $\operatorname{int} F(A) = \emptyset$. If $0 \in \ell$, then A is just a scalar multiple of some Hermitian operator B: $A = cB$. Consequently, $F_N(A) = (\operatorname{sgn} c) F_N(B)$, and Theorem 5.4 suffices to describe it fully. So, only the case $0 \notin \ell$ needs to be investigated further.

Denote the endpoints of $\sigma(A)$ by λ_0 and μ_0. The set $[\lambda_0, \mu_0] \setminus \sigma(A)$ is relatively open in ℓ, and thus consists of at most countably many disjoint open intervals. Label them (λ_j, μ_j), $j \in J$ for some $J \subset \mathbb{N}$. Recall also the notation $\Gamma(\theta_1, \theta_2, k)$ introduced in Section 3 for an arc of the curve (3.2).

THEOREM 5.5. *Let $A \in [\mathscr{H}]$ be essentially Hermitian, with the spectrum $\sigma(A) = [\lambda_0, \mu_0] \setminus \bigcup_{j \in J} (\lambda_j, \mu_j)$ as described above. Then:*

(i) *The boundary of $F_N(A)$ is described by the formula*

$$\partial F_N(A) = \cup_{j \in J \cup \{0\}} \Gamma(\arg \lambda_j, \arg \mu_j, k_j) \bigcup \{\operatorname{sgn} \zeta : \zeta \in \sigma(A)\}, \tag{5.5}$$

where $k_j = |\lambda_j/\mu_j| + |\mu_j/\lambda_j|$. Moreover,

(ii) *The points of $\Gamma(\arg \lambda_j, \arg \mu_j, k_j) \setminus \mathbb{T}$ belong to $F_N(A)$ if and only if $\lambda_j, \mu_j \in \sigma_p(A)$, while $\operatorname{sgn} \zeta \in F_N(A)$ if and only if $\zeta \in \sigma_p(A)$.*

Proof. Recall the notation $\gamma(\theta), \Gamma(\theta)$ for the endpoints of the intersection $F_N(A, \theta)$ of $F_N(A)$ with the ray (1.3). To prove (i) we just need to show that

$$\gamma(\theta) \in \Gamma(\arg \lambda_0, \mu_0, k_0), \tag{5.6}$$

while

$$\Gamma(\theta) \in \begin{cases} \Gamma(\arg \lambda_j, \arg \mu_j, k_j) \\ \mathbb{T} \end{cases} \quad \text{if } \ell_\theta \text{ intersects with } \begin{cases} (\lambda_j, \mu_j) \\ \sigma(A). \end{cases} \tag{5.7}$$

Since the statement of the theorem is invariant under scaling of A by any non-zero number, we may without loss of generality suppose that the line ℓ containing $\sigma(A)$ is $\{z\colon \operatorname{Im} z = 1\}$, that is, $A = H + iI$ for some Hermitian H. Then for $\|x\| = 1$ we have

$$\frac{\langle Ax, x\rangle}{\|Ax\|} = \frac{\langle Hx, x\rangle + i}{(1 + \|Hx\|^2)^{1/2}},$$

and $\gamma(\theta)$, $\Gamma(\theta)$ can be found by maximizing (resp., minimizing) $\|Hx\|$ under the constraint $\langle Hx, x\rangle = \cot\theta$.

Invoking the spectral decomposition $H = \int_{\sigma(H)} \lambda\, dE_\lambda$ of H, this becomes an extremal problem for $\int_{\sigma(H)} \lambda^2\, d\langle E_\lambda x, x\rangle$ under the constraint

$$\int_{\sigma(H)} \lambda\, d\langle E_\lambda x, x\rangle = \cot\theta \left(\int_{\sigma(H)} d\langle E_\lambda x, x\rangle \right)^{1/2}.$$

An approach similar to the one outlined in the proof of Theorem 5.4 reveals that the maximizing sequence of vectors is achieved by concentrating the measure $\langle dE_\lambda x, x\rangle$ at the endpoints of $\sigma(H)$, while for the minimizing sequence the concentration occurs as close to $\cot\theta$ as the geometry of $\sigma(H) = \sigma(A) - i$ allows.

So, $\gamma(\theta)$ is the same as the intersection of ℓ_θ with $F_N(\operatorname{diag}[\lambda_0, \mu_0])$. Due to Proposition 3.1, this proves (5.6). If $\cot\theta \in \sigma(H)$, then ℓ_θ intersects $\sigma(A)$ at the point $\cot\theta + i$, and the bottom line of (5.7) follows from Proposition 1.2. Finally, for $\cot\theta \notin \sigma(H)$ the ray ℓ_θ passes through one of the intervals (λ_j, μ_j). Consequently, $\Gamma(\theta)$ is the intersection of ℓ_θ with $F_N(\operatorname{diag}[\lambda_j, \mu_j])$, and another use of Proposition 3.1 proves the upper line of (5.7).

Statement (ii) concerning point of $\mathbb{T}$ follows from (1.5). As for the points of the arcs $\Gamma(\arg\lambda_j, \arg\mu_j, k_j)$, we just need to observe that the extrema of $\int_{\sigma(H)} \lambda^2\, d\langle E_\lambda x, x\rangle$ are attained if and only if $\lambda_j, \mu_j \in \sigma_p(A)$. □

5.5. Empty interior

Combining several already obtained results, we can now settle the question when $F_N(A)$ has empty interior.

THEOREM 5.6. *The interior of $F_N(A)$ is empty if and only if A is either*
(i) *a scalar multiple of a Hermitian operator, or*
(ii) *essentially Hermitian, with the spectrum consisting of two points.*

Proof. Sufficiency. For Hermitian A, according to Theorem 5.4, the set $F_N(A)$ is an interval in $\mathbb{R}$. For scalar multiples of Hermitian operators, the result follows by rotation. In particular, it applies to essentially Hermitian operators with the spectrum $\sigma(A) = \{\lambda_1, \lambda_2\}$ if one of λ_j is zero or $\lambda_1/\lambda_2 \in R$. In the remaining cases, the set $F_N(A)$ is the same as for A given by (3.1), which can also be seen from (5.5), and so coincides with an hyperbolic arc described by (3.2).

Necessity. We need only to consider the case when A is not a scalar multiple of an Hermitian operator, that is, $\Theta(A)$ is a non-trivial arc. By Theorem 4.3, in order

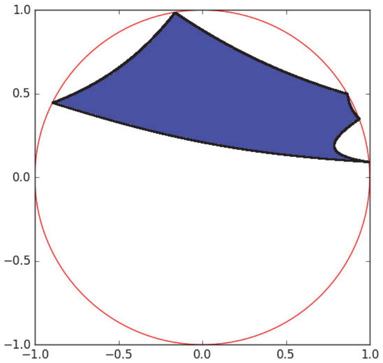

Figure 2: *The normalized numerical range of the essentially Hermitian 5 by 5 matrix: diag* $(-6+3i,-0.5+3i,5.2+3i,8+3i,32+3i)$

for $\operatorname{int} F_N(A)$ to be empty we then must have $\Theta(A) \neq \mathbb{T}$ and $\gamma(\theta) = \Gamma(\theta)$ for all $\theta \in \Theta_o(A)$, that is, $\operatorname{int} F(A) = \emptyset$. But then A is essentially Hermitian. It remains to observe that, by Theorem 5.5, the curve (5.5) is a boundary of a non-empty domain as long as $\sigma(A)$ contains at least one point besides λ_0, μ_0. □

6. Finite dimensional case

Naturally, things simplify in a finite dimensional setting.

6.1. Unit disk containment

Recall Corollary 2.1, according to which $F_N(A) \supset \mathbb{D}$ if A is not injective and the minimal angle between $\operatorname{Im} A$ and $\ker A$ is zero. This condition is by no means necessary, as examples of weighted shifts and non-unitary isometries show. However, these examples are typically infinite-dimensional. In finite dimensions the zero angle condition holds only if $\operatorname{Im} A \cap \ker A \neq \{0\}$, and actually becomes necessary for $F_N(A)$ to contain $\mathbb{D}$. Even more can be said.

THEOREM 6.1. *For $A \in \mathbb{C}^{n\times n}$, the following statements are equivalent:*
(i) $F_N(A) \supset \mathbb{D}$,
(ii) *the closure of $F_N(A)$ contains at least one point of $\mathbb{T} \setminus F_N(A)$,*
(iii) $\operatorname{Im} A \cap \ker A \neq \{0\}$.

Proof. Due to (1.5), the set $\mathbb{T} \cap F_N(A)$ is (at most) finite, and so the implication (i) $\Longrightarrow$ (ii) is obvious. As was already discussed, (iii) $\Longrightarrow$ (i) always holds. It remains therefore to show that (ii) $\Longrightarrow$ (iii).

Suppose $\lambda \in \mathbb{T}$ is a limit point of $F_N(A)$. By definition, there exists a sequence of unit vectors x_n such that $\langle Ax_n, x_n\rangle / \|Ax_n\| \to \lambda$. Denote $y_n = Ax_n$ and $z_n = y_n / \|y_n\|$. Passing to subsequences if needed, we may without loss of generality suppose that $x_n \to x$ and $z_n \to z$. Then of course $\langle z, x\rangle = \lambda$ and, since $\|z\| = \|x\| = 1$, $z = \lambda x$. On the other hand, $Ax_n = y_n = z_n \|Ax_n\|$, and so $Ax = z\|Ax\| = \lambda \|Ax\| x$. This means that $\lambda \|Ax\|$ is an eigenvalue of A, while x is the respective eigenvector. But then $\lambda \in F_N(A)$ if $\|Ax\| \neq 0$, and $x \in \ker A$ otherwise. Since at the same time $x(= \lambda^{-1} z) \in \operatorname{Im} A$, this completes the proof. □

COROLLARY 6.1. *For $A \in \mathbb{C}^{n\times n}$, the following statements are equivalent:*
(i) *$F_N(A) = \mathbb{D}$,*
(ii) *$F_N(A)$ is open,*
(iii) *A is nilpotent but different from 0.*

Proof. Implication (i) $\Longrightarrow$ (ii) is obvious. The contrapositive of (ii) $\Longrightarrow$ (iii) is seen from (1.5): every non-zero eigenvalue of A yields a point of $F_N(A)$ which lies in $\mathbb{T}$ and thus on the boundary of $F_N(A)$. Finally, if A is non-zero and nilpotent, then claim (iii) of Theorem 6.1 holds and, according to implication (iii) $\Longrightarrow$ (i) of this theorem, $F_N(A) \supset \mathbb{D}$. Combining this with (1.5) yields $F_N(A) = \mathbb{D}$. □

For Jordan blocks J with zero eigenvalue the equality $F_N(J) = \mathbb{D}$ was observed in [9].

6.2. When is $F_N(A)$ closed?

It is well known (and obvious) that for $A \in \mathbb{C}^{n\times n}$ the set $F(A)$ is closed, simply because it is the range of a continuous function $x \mapsto \langle Ax, x\rangle$ on the unit sphere of $\mathbb{C}^n$. If A is invertible, then $x \mapsto \langle Ax, x\rangle / \|Ax\|$ is also continuous, and $F_N(A)$ is closed, for exactly the same reason. This is in complete agreement with Theorem 5.5, the finite dimensional version of which can be stated as follows.

THEOREM 6.2. *Let $A \in \mathbb{C}^{n\times n}$ be essentially Hermitian, with the spectrum lying on the line ℓ not passing through the origin and the eigenvalues $\lambda_1, \dots, \lambda_n$ labeled in the order of appearance on ℓ. Then $F_N(A)$ is a closed domain bounded by the curve $\bigcup_{j=1}^{n} \Gamma(\arg\lambda_j, \arg\lambda_{j+1}, k_j)$.*

Here $k_j = |\lambda_j/\lambda_{j+1}| + |\lambda_{j+1}/\lambda_j|$, and by convention $\lambda_{n+1} = \lambda_1$.

For non-invertible A, Proposition 2.2 shows that $F_N(A)$ may not be closed. Moreover, it is easy to see from there that for $n = 2$ the set $F_N(A)$ is closed if and only if A is invertible.

The situation is slightly more complicated if $n > 2$. Let us first settle the case $\operatorname{Im} A \perp \ker A$.

THEOREM 6.3. *Let $A \in \mathbb{C}^{n\times n}$ be range-Hermitian, with $\ker A \neq \{0\}$. Then $F_N(A)$ is closed if and only if $0 \in F(B)$, where B is the compression of A onto its range.*

Proof. By Proposition 2.1, $0 \in \partial F_N(A)$, and $0 \in F_N(A)$ if and only if $0 \in F(B)$. This proves *necessity*. For *sufficiency* just observe that by Corollary 3.1 we have $F_N(A) = F_N(B)$, and in finite dimensional setting B is invertible. □

So, already for $n = 3$ it is easy to construct non-invertible A with closed $F_N(A)$. Consider, in particular,

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

Then $B = \begin{bmatrix} 1 & k \\ 0 & 1 \end{bmatrix}$, and $F_N(A) = F_N(B)$ is thus closed whenever $0 \in F(B)$, that is $|k| \geqslant 2$. In fact, $F_N(B)$ in this case is a closed elliptical disk a complete description of which is given in [9, Proposition 4]. Even simpler examples of this kind are delivered by singular Hermitian matrices having at least one positive and one negative eigenvalue; once again, the smallest size for which this is possible is $n = 3$. On the other hand, for semi-definite singular Hermitian matrices A, as well as scalar multiples thereof, $F_N(A)$ is a half open interval of length one with an endpoint zero (not included).

Finally, let $\operatorname{Im} A \not\subseteq \ker A$.

THEOREM 6.4. *Let $A \in \mathbb{C}^{n \times n}$ be represented as (2.1) with $C \neq 0$. Then $F_N(A)$ is closed if and only if it contains the circle $\mathscr{C}_A$ of A.*

Proof. Necessity follows from Lemma 4.5.

To prove *sufficiency*, we need only to establish that $F_N(A)$ contains all its limit points with absolute value bigger than $\rho/\sqrt{\rho^2+1}$. Suppose z is such a point, and $\langle Ax_n, x_n\rangle / \|Ax_n\| \to z$ for some sequence of unit vectors x_n. Formulas (2.3) and the definition of $r(A)$ then imply that in the representations (2.2) of x_n the respective values of t_n must be bounded away from zero. Selecting a subsequence of $\{x_n\}$ converging to some unit vector x, we then conclude that $Ax \neq 0$. Consequently, $z = \langle Ax, x\rangle / \|Ax\| \in F_N(A)$. □

COROLLARY 6.2. *If in (2.8) B is such that $0 \in \operatorname{int} F(B)$, then for sufficiently small C the set $F_N(A)$ is closed.*

Indeed, according to (2.6) the sets $G(r_1(u), r_2(u))$ depend continuously on C. Since 0 is an interior point of $F_N(A) (= F_N(B))$ for $C = 0$, this property persists for sufficiently small C. On the other hand, the radius of the critical circle also is a continuous function of C, equal zero at $C = 0$. Consequently, $\mathscr{C}_A \subset F_N(A)$ for all C in some neighborhood of 0.

EXAMPLE 2. Let in (2.1) $B = \begin{bmatrix} 1 & 0 \\ b & 1 \end{bmatrix}$, $C = [c \ \ 0]$ for $b > 2$ and sufficiently small $c > 0$.

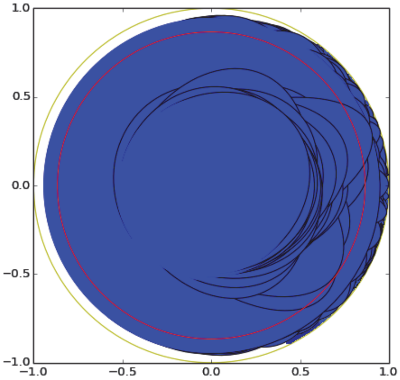

Figure 3: *The normalized numerical range obtained for* $b=5$, $c=3$. *It is closed and thus contains the critical circle, depicted with red.*

6.3. Boundary of $F_N(A)$

In Section 4 we introduced $\Gamma(\theta)$ and $\gamma(\theta)$ as the endpoints of the interval $F_N(A,\theta)$ and proved their continuity as functions of θ provided that $F_N(A)$ was closed. We will show here that in finite dimensional setting the latter restriction is redundant, which in turn yields the complete description of $\partial F_N(A)$.

THEOREM 6.5. *Let* $A\in\mathbb{C}^{n\times n}$. *Then* Γ *and* γ *are continuous functions of* θ, *and the boundary of* $F_N(A)$ *is given by* (4.8).

Proof. Due to Lemma 4.2 and Theorem 4.3, we need only consider the case when $F_N(A)$ is not closed, and thus $\ker A\neq\{0\}$. Then of course $0\in F(A)$, due to (4.1) implying that $0\in\overline{F_N(A)}$. So, $\gamma(A)=\{0\}$, and the function γ is continuous in θ in a trivial way.

When proving the continuity of Γ, let us make use of the representation (2.1). If A is range Hermitian, we may in addition suppose that $0\notin F(B)$, since otherwise $F_N(A)$ is closed by Theorem 6.3. From Proposition 2.1 we see that the Γ functions for A and B coincide, thus implying the desired continuity. Moreover, $\partial F_N(A)$ is indeed given by the second line of (4.8).

Finally, let $\ker A\not\perp \operatorname{Im} A$. By Corollary 2.2, $0\in\operatorname{int}F_N(A)$, and so for any sequence $\theta_k\to\theta_0$ with $\Gamma(\theta_k)\to z$ we have $|z|\geqslant r(A)$, with $r(A)$ given by (2.9). From the definition of Γ it follows that we can find $z_k\in F_N(A,\theta_k)$ converging to z. As in the proof of Theorem 6.4, we then conclude that $z\in F_N(A)$, and so $|\Gamma(\theta_0)|\geqslant\lim|\Gamma(\theta_k)|$. As in the proof of Lemma 4.2, from here and the first inequality in (4.2) it follows that $\Gamma(\theta_0)=\lim\Gamma(\theta_k)$. So, Γ is a continuous function of θ in this case as well, and the closed curve $\Gamma(A)$ is indeed the boundary of $F_N(A)$. □

6.4. Compressions

In finite dimensions, $\sigma(A)=\sigma_p(A)$. So, Theorem 3.4 and its Corollary 3.3 immediately yield

THEOREM 6.6. *Given $A \in \mathbb{C}^{n\times n}$, the inclusion $F_N(B) \subset F_N(A)$ holds for all its compressions B if and only if A is a scalar multiple of a Hermitian matrix.*

REFERENCES

[1] W. AUZINGER, *Sectorial operators and normalized numerical range*, Appl. Numer. Math. **45** (4): 367–388, 2003.

[2] D. COREY, C. JOHNSON, R. KIRK, B. LINS, AND I. M. SPITKOVSKY, *The product field of values*, Linear Algebra Appl. **438**: 2155–2173, 2013.

[3] JU. L. DALECKII AND M. G. KREIN, *Stability of solutions of differential equations in Banach space*, American Mathematical Society, Providence, R. I., 1974.

[4] D. S. DJORDJEVIĆ, *Characterizations of normal, hyponormal and EP operators*, J. Math. Anal. Appl. **329** (2): 1181–1190, 2007.

[5] E. DURSZT, *On the numerical range of normal operators*, Acta Sci. Math. (Szeged), **25**: 262–265, 1964.

[6] L. Z. GEVORGYAN, *On the convergence rate of iterations and the normalized numerical range of an operator*, Math. Sci. Res. J. **8** (1): 16–26, 2004.

[7] L. Z. GEVORGYAN, *On some properties of the normalized numerical range*, Izv. Nats. Akad. Nauk Armenii Mat. **41** (1): 41–48, 2006.

[8] L. Z. GEVORGYAN, *An example of the normalized numerical range*, Armenian J. Math. **1** (1): 50–33, 2009.

[9] L. Z. GEVORGYAN, *Normalized numerical ranges of some operators*, Operators and Matrices **3** (1): 145–153, 2009.

[10] L. Z. GEVORGYAN, *Normalized numerical ranges of some complex 2×2 matrices*, Izv. Nats. Akad. Nauk Armenii Mat. **46** (5): 41–52, 2011.

[11] K. E. GUSTAFSON AND D. K. M. RAO, *Numerical Range. The Field of Values of Linear Operators and Matrices*, Springer, New York, 1997.

[12] P. HALMOS, *A Hilbert Space Problem Book*, Van Nostrand, Princeton, N. J., 1967.

[13] R. A. HORN AND C. R. JOHNSON, *Topics in matrix analysis*, Cambridge University Press, Cambridge, 1994. Corrected reprint of the 1991 original.

[14] L. V. KANTOROVICH, *Functional analysis and applied mathematics*, Uspehi Matem. Nauk (N. S.), **3** (6 (28)): 89–185, 1948.

[15] L. V. KANTOROVICH, *Functional analysis and applied mathematics*, NBS Rep. 1509. U. S. Department of Commerce, National Bureau of Standards, Los Angeles, Calif., 1952., translated by C. D. Benster.

[16] W. C. RIDGE, *Numerical range of a weighted shift with periodic weights*, Proc. Amer. Math. Soc. **55**: 107–110, 1976.

[17] Q. F. STOUT, *The numerical range of a weighted shift*, Proc. Amer. Math. Soc. **88**: 495–502, 1983.

[18] B. SZ.-NAGY, C. FOIAS, H. BERCOVICI, AND L. KÉRCHY, *Harmonic analysis of operators on Hilbert space*, Universitext, Springer, New York, second edition, 2010.

[19] K.-Z. WANG AND P. Y. WU, *Numerical ranges of weighted shifts*, J. Math. Anal. Appl. **381** (2): 897–909, 2011.

(Received June 3, 2016)

Ilya M. Spitkovsky
Division of Science
New York University Abu Dhabi (NYUAD)
Saadiyat Island, P. O. Box 129188 Abu Dhabi, UAE
e-mail: ims2@nyu.edu, imspitkovsky@gmail.com

Andrei-Florian Stoica
Division of Science
New York University Abu Dhabi (NYUAD)
Saadiyat Island, P. O. Box 129188 Abu Dhabi, UAE
e-mail: as8490@nyu.edu