university of groningen / faculty of science and engineering / mathematics and applied mathematics

The Numerical Range of Linear Operators on Hilbert Spaces

Bachelor's Project Mathematics

July 2019

Student: Boris Luttikhuizen s2718669

First supervisor: Dr. A.E. Sterk

Second assessor: Dr. ir. R. Luppes

Abstract

The numerical range of bounded linear operators on complex seperable Hilbert spaces can be used to make many useful conclusions. After discussing the definition and the basic properties of the numerical range, we will study several applications in operator theory and numerical methods. We will construct and compare two different methods that approximate the numerical range of an operator in MatLab. We will also discuss the relevance of the numerical range in discrete stability and in the convergence of the steepest descent method. We conclude the paper with a more recent appearance in quantum computing.

Contents

1 Introduction

In this report we set out to discuss the numerical range of linear operators on Hilbert spaces. In order to do this we should first familiarize ourselves with the definition of numerical range.

Definition 1. *The numerical range of an operator T on a complex Hilbert space H is the subset of the complex numbers given by*

$$W(T) = \{\langle Tx, x\rangle, x \in H, ||x|| = 1\}.$$

It follows directly from the definition of the inner product that

$$W(\alpha I + \beta T) = \alpha + \beta W(T) \text{ for } \alpha, \beta \in \mathbb{C},$$

$$W(T^*) = \{\overline{\lambda}, \lambda \in W(T)\},$$

$$W(U^*TU) = W(T) \text{ for any unitary } U.$$

In discussing the properties of the numerical range, we will follow Gustafson and Rao [9] and consider bounded linear operators on a complex and separable Hilbert space H. This does not mean that all of these properties do not hold for different variations of Hilbert spaces and operators. The question that must now be asked is: why are we interested in the numerical range of linear operators? To formulate an answer to this question we must first discuss some of the most important properties of the numerical range.

1.1 Properties and Applications

The most fundamental property of the numerical range is its convexity. Other important properties are that its closure contains the spectrum of the operator and that the numerical radius provides a norm equivalent to the operator norm [9]. From these properties it might already become clear that the numerical range can maybe be used in eigenvalue approximation, which plays, for example, a big role in the construction of an initial guess for iterative methods that estimate eigenvalues.

Let us now, as an example, estimate the eigenvalues of the following matrix in two different ways.

$$A = \begin{bmatrix} 1 & 2 \\ 1 & -1 \end{bmatrix}$$

The most common way to approximate the eigenvalues of this matrix would be to use Gershgorin cirles. The theorems concerning the Gershgorin circles are stated below and can be found accompanied by their proofs in the book on numerical mathematics by Quarteroni, Sacco and Saleri [10].

Theorem 1. *(Gershgorin circles). Let $A \in \mathbb{C}^{n\times n}$. Then*

$$\sigma(A) \subset S_R = \bigcup_{i=1}^{n} R_i, \qquad R_i = \{z \in \mathbb{C} : |z - a_{ii}| \leq \sum_{j=1, j\neq i}^{n} |a_{ij}|\}.$$

The sets R_i are called Gershgorin row circles.

We know that the spectrum of a matrix is invariant under transposition, so we can also take the column sum instead of the row sum in the set above to obtain the Gershgorin column circles [10].

$$\sigma(A) \subset S_C = \bigcup_{i=1}^{n} C_i, \qquad C_i = \{z \in \mathbb{C} : |z - a_{ii}| \leq \sum_{i=1, i \neq j}^{n} |a_{ij}|\}$$

The following property follows directly from the inclusion of the spectrum of A in both S_R and S_C.

Property 1. *For a given matrix $A \in \mathbb{C}^{n\times}$,*

$$\forall \lambda \in \sigma(A), \qquad \lambda \in S_R \bigcap S_C.$$

If we now apply this to the matrix in our example, we get the following row and column circles

$$\begin{aligned} R_1 &= \{z \in \mathbb{C} : |z - 1| \leq 2\}, & C_1 &= \{z \in \mathbb{C} : |z - 1| \leq 1\}, \\ R_2 &= \{z \in \mathbb{C} : |z + 1| \leq 1\}, & C_2 &= \{z \in \mathbb{C} : |z + 1| \leq 2\}. \end{aligned}$$

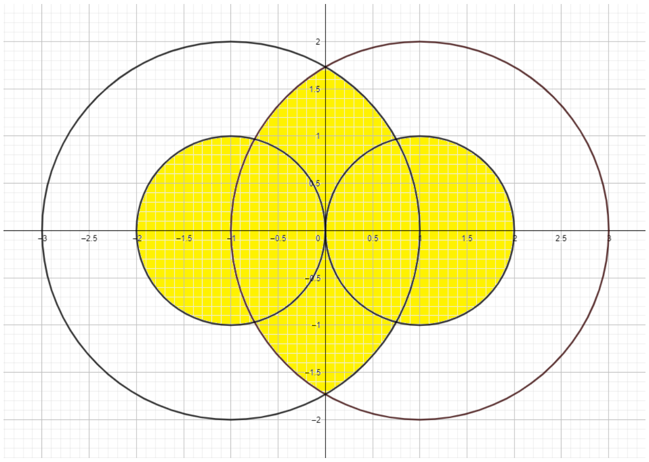

Figure 1: Gershgorin circles of matrix A in $\mathbb{C}^2$.

Combining the Gershgorin circles with property 1 tells us that the eigenvalues of A must be in the yellow area of Figure 1. If we want to use the numerical range instead, we make use of the following lemma, which can be found in the book by Gustafson and Rao [9].

Lemma 1. (Ellipse Lemma). Let T be an operator on a two-dimensional space. Then $W(T)$ is an ellipse whose foci are the eigenvalues of T.

To make use of this lemma, it is easier to first compute the Schur Decomposition of our matrix A in MatLab. We do this because it makes it easier to find the ellipse that represents the numerical range of our matrix. Normally it would not make sense to do this, because the decomposition already gives you the eigenvalues of the matrix on the diagonal. The upper-triangular matrix that follows from the Schur Decomposition of A is given by

$$S = \begin{bmatrix} \sqrt{3} & 1 \\ 0 & -\sqrt{3} \end{bmatrix}.$$

We can use this in combination with the approach of example 3 in Gustafson and Rao [9]. Let $u = (f, g)$ be a unit vector in $\mathbb{C}^2$, $f = e^{i\alpha}\cos\theta$, $g = e^{i\beta}\sin\theta$, $\alpha, \beta \in \left[0, \frac{\pi}{2}\right]$, $\theta \in [0, 2\pi)$. Straight-forward computations will then give us that

$$Su = \left(\sqrt{3}e^{i\alpha}\cos\theta + e^{i\beta}\sin\theta, -\sqrt{3}e^{i\beta}\sin\theta\right),$$

and

$$\langle Su, u\rangle = \sqrt{3}\left(\cos^2\theta - \sin^2\theta\right) + e^{i(\beta-\alpha)}\sin\theta\cos\theta =: x + iy.$$

Separation of the real and imaginary parts of $\langle Su, u\rangle$ yields

$$x = \sqrt{3}\cos 2\theta + \frac{1}{2}\sin 2\theta\cos(\beta - \alpha),$$

$$y = \frac{1}{2}\sin(\beta - \alpha)\sin 2\theta.$$

From which we obtain the following equality

$$(x - \sqrt{3}\cos 2\theta)^2 + y^2 = \frac{1}{4}\sin^2 2\theta.$$

This is a family of circles, we can represent these circles as

$$(x - \sqrt{3}\cos\phi)^2 + y^2 = \frac{1}{4}\sin^2\phi, \quad 0 \le \phi \le \pi.$$

If we now differentiate this with respect to ϕ we get that

$$(x - \sqrt{3}\cos\phi)\sqrt{3} = \frac{1}{4}\cos\phi.$$

We can eliminate ϕ by combining these two equalities. By doing this we obtain the formula for the ellipse in Figure 2, which is

$$\frac{x^2}{3 + \frac{1}{4}} + \frac{y^2}{\frac{1}{4}} = 1.$$

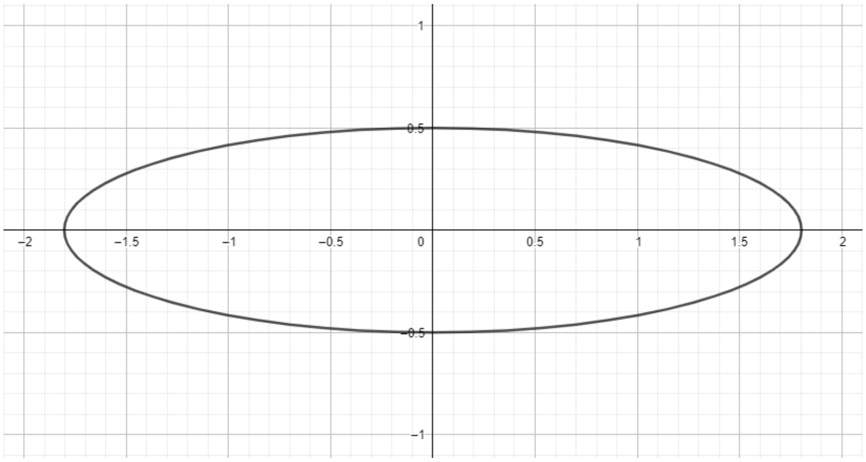

Figure 2: Numerical Range of matrix A in $\mathbb{C}^2$.

We saw in the lemma above that the foci of this ellipse are the eigenvalues of the matrix S and thus of the matrix A. Whilst the foci of an ellipse cannot be extracted from a plot directly (we can of course obtain them from the ellipse formula that we constructed from the Schur Decomposition), it is clear that one could make a more narrowed down guess of the eigenvalues using the plot of the numerical range instead of Gershgorin circles in this case.

The example above already shows us how the numerical range could be of interest for various fields of mathematics dealing with eigenvalues. To illustrate another application of the numerical range we introduce a notion closely related to it, which is that of numerical radius [9].

Definition 4. The numerical radius $w(T)$ of an operator T on a Hilbert space H is given by

$$w(T) = \sup\{|\lambda|, \lambda \in W(T)\}.$$

Gustafson and Rao prove in their book that the numerical radius is an equivalent norm to the operator norm. We should note that this is not the case in a real Hilbert space [9]. This equivalence can be used to construct many useful theorems about the operator norm. We have now seen an application of the numerical range in numerical mathematics and in operator theory, but there are many more applications of the numerical range in various fields of mathematics. An interesting and more recent example is the use of numerical ranges in quantum information processing [6].

1.2 Research Goals

In this report we set out to study the properties and applications of the numerical range of linear operators on Hilbert spaces. The main focus of the article will be to give a comprehensive explanation of what the numerical range is. We will do this by studying concrete examples of the numerical ranges of both finite and infinite dimensional operators. We will also discuss several applications of the numerical range in different fields of mathematics, of which some have been mentioned in the previous section, to show how important the numerical range is for both theoretical and applied mathematics.

2 Properties and Applications of the Numerical Range

Now that we have made ourselves familiar with the definition and the possible applications, we will use this chapter to really start our investigation into the properties and applications of the numerical range. All the results that are presented in this chapter are taken from Gustafson and Rao [9], unless stated otherwise. Some of the proofs have however been slightly modified or explained more explicitly.

2.1 Elliptic Range

We will begin this section by going back to the lemma that we used to compute the numerical range of our example in the introduction. This time, however, it will be accompanied by its proof.

Lemma 1. *(Ellipse lemma). Let T be an operator on a two-dimensional space. Then $W(T)$ is an ellipse whose foci are the eigenvalues of T.*

Proof. As mentioned in the introduction we can use a unitary transformation (Schur decomposition) to bring any operator T to the form

$$S = \begin{bmatrix} \lambda_1 & a \\ 0 & \lambda_2 \end{bmatrix},$$

where λ_1 and λ_2 are the eigenvalues of T. We are allowed to use this transformation, because the numerical range is invariant under unitary transformations. To prove this lemma we consider the three possible cases.

If $\lambda_1 = \lambda_2 = \lambda$, we have

$$T - \lambda = \begin{bmatrix} 0 & a \\ 0 & 0 \end{bmatrix}, \quad W(T-\lambda) = \left\{ z : |z| \leq \frac{|a|}{2} \right\},$$

where the expression for $W(T-\lambda)$ requires some explanation. If we take a unit vector $x = (f, g)$, we get that $Tx = (ag, 0)$ and $\langle Tx, x \rangle = ag\overline{f}$. If we now look at the magnitude of these vectors, we see that $|\langle Tx, x\rangle| = |a||g||f| \leq \frac{|a|}{2}(|f|^2 + |g|^2) = \frac{|a|}{2}$. Where the last equality follows from the fact that x is a unit vector. This inequality explains the expression for the numerical range above. It now follows immediately from the fact that $W(\alpha I+\beta T) = \alpha+\beta W(T)$, that the $W(T)$ is a circle of radius $\frac{|a|}{2}$ centered at λ.

If $\lambda_1 \neq \lambda_2$ and $a = 0$, we have

$$T = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix}$$

Let $x = (f, g)$ again be a unit vector, then $\langle Tx, x\rangle = \lambda_1|f|^2 + \lambda_2|g|^2$. We can show that this is a set of convex combinations of λ_1 and λ_2. To do this set $t = |f|^2$, it then immediately follows from the fact that x is a unit vector that $|g|^2 = 1 - |f|^2 = 1 - t$. We can thus write the combination above as

$$\langle Tx, x\rangle = t\lambda_1 + (1 - t)\lambda_2$$

Where $t + (1 - t) = 1$ and $t, (1 - t) \geq 0$, so $W(T)$ is indeed the set of convex combinations of λ_1 and λ_2. The $W(T)$ must therefore be the segment joining the two eigenvalues.

If $\lambda_1 \neq \lambda_2$ and $a \neq 0$, we have

$$T - \frac{\lambda_1 + \lambda_2}{2} = \begin{bmatrix} \frac{\lambda_1-\lambda_2}{2} & a \\ 0 & \frac{\lambda_2-\lambda_1}{2} \end{bmatrix}.$$

Let us now use Euler's formula to obtain $\frac{\lambda_1-\lambda_2}{2} =: re^{i\theta}$. We can multiply our matrix with $e^{-i\theta}$ to obtain

$$e^{-i\theta} \begin{bmatrix} \frac{\lambda_1-\lambda_2}{2} & a \\ 0 & \frac{\lambda_2-\lambda_1}{2} \end{bmatrix} = \begin{bmatrix} r & ae^{-i\theta} \\ 0 & -r \end{bmatrix}.$$

We saw how to find the ellipse for a matrix of this form in the example in the introduction. It will be an ellipse with its center at $(0, 0)$, minor axis $|a|$ and foci $(r, 0)$ and $(-r, 0)$. It follows from the properties of $W(T)$ that we need to shift this ellipse with $\frac{\lambda_1+\lambda_2}{2}$ and rotate it with an angle θ with respect to the real axis to obtain $W(T)$. This clearly does not change the fact that it is an ellipse. □

One of the nice consequences of the fact that the numerical range of an operator on a two-dimensional space is an ellipse is that its numerical range is thus convex. The following theorem tells us that we can generalize this convexity to operators on higher-dimensional spaces.

Theorem 2. *(Toeplitz-Hausdorff). The numerical range of an operator is convex.*

Proof. To prove this, we want to show that for any two elements of $W(T)$ the segment connecting them is contained in $W(T)$. If $\alpha, \beta \in W(T)$, then $\alpha = \langle Tf, f\rangle$ and $\beta = \langle Tg, g\rangle$, for some unit vectors f and g in H. If we now consider $V = span\{f, g\}$ and let E be the orthogonal projection of H on V, we get that $Ef = f$ and $Eg = g$. The elements in the $W(ETE)$ generated by f and g are then given by

$$\langle ETEf, f\rangle = \langle Tf, f\rangle = \alpha, \quad \langle ETEg, g\rangle = \langle Tg, g\rangle = \beta.$$

Because ETE is an operator on a two-dimensional space, we know that $W(ETE)$ is an ellipse. Therefore $W(ETE)$ is clearly convex, because of this the path between α and β is contained in $W(ETE)$. The final observation that we need to make is that $W(ETE) \subset W(T)$, which concludes the proof. □

A direct consequence of the Toeplitz Hausdorff theorem is that $W(T)$ is the union of two-dimensional numerical ranges, which are ellipses. Recall that we showed in the proof of the ellipse lemma that these ellipses could also, for example, be a line.

A good example of an operator on a higher-dimensional space is the left shift on ℓ_2, where ℓ_2 is the Hilbert space of square summable sequences [9]. Let $f = (f_1, f_2, \dots) \in \ell_2$, with $||f|| = 1$. If we call the unilateral shift T we get that

$$Tf = (f_2, f_3, \dots) \quad \Rightarrow \quad \langle Tf, f\rangle = f_1\overline{f_2} + f_2\overline{f_3} + f_3\overline{f_4} + \cdots$$

If we compute the magnitude of this element of $W(T)$ we get that

$$\begin{aligned} |\langle Tf, f\rangle| &\leq |f_1||f_2| + |f_2||f_3| + |f_3||f_4| + \cdots \\ &\leq \frac{1}{2}\left(|f_1|^2 + 2|f_2|^2 + 2|f_3|^2 + \cdots\right) \\ &\leq \frac{1}{2}\left(2 - |f_1|^2\right). \end{aligned}$$

If $|f_1| \neq 0$, we get that $|\langle Tf, f\rangle| < 1$. If it is zero and f contains a finite number of nonzero elements, we can rearrange the terms to obtain the same inequality. We can therefore conclude that $W(T)$ is a subset of the open unit disk. In fact, as we will see, it actually is the unit disk. We can represent any point in the unit disk using Euler's formula as $z = re^{i\theta}$, where $0 \leq r < 1$ and $\theta \in [0, 2\pi)$. If we take the following sequence

$$f = \left(\sqrt{1-r^2}, r\sqrt{1-r^2}e^{-i\theta}, r^2\sqrt{1-r^2}e^{-2i\theta}, \dots\right),$$

we get that

$$||f||^2 = 1 - r^2 + r^2(1-r^2) + r^4(1-r^2) + \cdots = 1.$$

Therefore we know that $\langle Tf, f\rangle \in W(T)$. To show that it is in fact equal to z, we compute it

$$\langle Tf, f\rangle = r(1-r^2)e^{i\theta} + r^3(1-r^2)e^{i\theta} + \cdots = re^{i\theta} = z.$$

This shows that the open unit disk is contained in $W(T)$, so the numerical range of T must in fact be the open unit disk. This example shows us that for this particular operator the convexity property indeed holds.

2.1.1 Spectral Inclusion

In this section we will study how the convexity of the numerical range leads to some interesting results. As we saw in the introduction an important property of the numerical range is that it can be used to bound the spectrum $\sigma(T)$.

Theorem 3. *(Spectral inclusion) The spectrum of an operator is contained in the closure of its numerical range.*

Proof. To prove this, we use the fact that the boundary of the spectrum is contained in the approximate point spectrum σ_{app}. The point spectrum of an operator is given by

$$\sigma_{app} := \{\lambda \in \mathbb{C} : \exists \text{ a sequence of unit } \{f_n\} : ||(T - \lambda I)f_n|| \to 0\}.$$

Because of the convexity of $W(T)$, we only need to show that $\sigma_{app} \subset W(T)$. Let us take $\lambda \in \sigma_{app}(T)$ and a sequence $\{f_n\}$ of unit vectors that satisfies

$$||(T - \lambda I)f_n|| \to 0.$$

The Cauchy-Schwarz inequality then gives us that

$$|\langle (T - \lambda I)f_n, f_n\rangle| \leq ||(T - \lambda I)f_n|| \cdot ||f_n|| = ||(T - \lambda I)f_n|| \to 0.$$

From the inequality above we can conclude that $\langle Tf_n, f_n\rangle \to \lambda$, so $\lambda \in \overline{W(T)}$. □

This alone is already an interesting result, but the consequences of this result might be of an even greater use. One of the consequences is that it allows us to say something about the spectrum of the sum of two operators, whereas we are normally not able to do this using just the spectrum of the two operators.

$$\sigma(A + B) \subset W(A + B) \subset W(A) + W(B).$$

Another useful result of the spectral inclusion in operator theory is that it gives us a characterization of selfadjoint operators.

Theorem 4. *T is selfadjoint if and only if $W(T)$ is real.*

Proof. If T is selfadjoint, we have that

$$\langle Tx, x\rangle = \langle x, Tx\rangle = \overline{\langle Tx, x\rangle},$$

for any $x \in H$. This can only be the case if $\langle Tx, x\rangle$ is real and thus $W(T)$ is real. Let us now assume that $W(T)$ is real. If this is the case we have that

$$\langle Tx, x\rangle - \langle x, Tx\rangle = 0 = \langle (T - T^*)x, x\rangle = 0 \quad \forall x \in H \text{ with } ||x|| = 1,$$

which implies that $W(T - T^*) = \{0\}$. This implies that the numerical radius, which was introduced in the introduction, is zero. We will prove in the next section that the numerical range is an equivalent norm to the operator norm, this has as a consequence that

$$||T - T^*|| = 0 \quad \Rightarrow \quad T = T^*.$$

It should be notes that this equivalence of norms does not generalize to real Hilbert spaces. We will show this in the next section using an example. □

The numerical range does not only provide us with a useful characterization of selfadjoint operators, it also allows us to say something about the operator norm of these operators.

Theorem 5. *Let T be selfadjoint and $W(T)$ be the real interval $[m, M]$. Then $||T|| = \sup\{|m|, |M|\}$.*

Proof. If $w(T) = \sup\{|m|, |M|\}$ we have that for any real non-zero λ that

$$\begin{aligned} 4||Tx||^2 &= \langle T(\lambda x + \lambda^{-1}Tx), \lambda x + \lambda^{-1}Tx\rangle \\ &-\langle T(\lambda x - \lambda^{-1}Tx), \lambda x - \lambda^{-1}Tx\rangle, \\ &\leq w(T)\left(||\lambda x + \lambda^{-1}Tx||^2 + ||\lambda x - \lambda^{-1}Tx||^2\right) \\ &= 2w(T)\left(\lambda^2||x||^2 + \lambda^{-2}||Tx||^2\right). \end{aligned}$$

Because this holds for abritrary non-zero λ, we can take $\lambda^2 = \frac{||Tx||}{||x||}$. Doing this yields that

$$\begin{aligned} 4||Tx||^2 &\leq 2w(T)\left(\frac{||Tx||}{||x||}||x||^2 + \frac{||x||}{||Tx||}||Tx||^2\right), \\ &= 4w(T)||Tx||||x||. \end{aligned}$$

Which is equivalent to the following inequality

$$||Tx|| \leq w(T)||x||.$$

From which we can conclude that $||T|| = w(T) = \sup\{|m|, |M|\}$. □

The next theorem considers a special case in which we can not only bound the spectrum of an operator, but also find actual eigenvalues of the operator.

Theorem 6. *Let $\overline{W(T)} = [m, M]$. Then $m, M \in \sigma(T)$.*

Proof. Because $m \in \overline{W(T)}$, we know that there exist a sequence of unit vectors $\{f_n\}$, such that $\langle Tf_n, f_n\rangle \to m$. We can use this and the fact that $|f_n| = 1$ to conclude that $||\langle(T - m)f_n, f_n\rangle|| = ||(T - m)^{\frac{1}{2}}f_n||^2 \to 0$. Moreover $||(T - m)f_n|| \to 0$, therefore $m \in \sigma_{app}(T) \subset \sigma(T)$. □

2.1.2 Numerical Radius

In the penultimate proof of the last section we used the supremum over the magnitude of the elements of the numerical range. This was not totally arbitrary, in fact, we already discussed the following definition in the introduction.

Definition 2. *The numerical radius $w(T)$ of an operator T on H is given by*

$$w(T) = sup\{|\lambda|, \lambda \in W(T)\}.$$

In the proof in the last section we used the numerical radius to find the operator norm of a selfadjoint operator. At the time that might have been quite surprising, but the following theorem from the book by Gustafson and Rao [9] should explain why we used it.

Theorem 7. *(Equivalent norm). $w(T) \leq ||T|| \leq 2w(T)$.*

Proof. Before we prove the equivalence we should show that $w(T)$ is a norm. Let us start by observing that

$$w(T) \geq 0.$$

This is an immediate consequence of the definition. Next we want to show that

$$w(T) = 0 \iff T = 0.$$

It clearly follows from the definition that $|\langle Tx, x\rangle| \leq w(T)||x||^2$, for any $x \in H$. This means that $|\langle Tx, x\rangle| = 0$, when $w(T) = 0$. Let us now observe that

$$|\langle Tx, x\rangle| = 0 \quad \Rightarrow \quad \langle Tx, x\rangle = 0 \quad \text{for any } x \in H.$$

We now use *corollary 6.18* in the lecture notes of de Snoo and Sterk [1], which tells us that the statement above can only hold if $T = 0$. The next thing we need to show is that

$$w(\alpha T) = |\alpha| \cdot w(T), \qquad \forall \alpha \in \mathbb{C}.$$

This is due to the fact that $|\langle \alpha Tx, x\rangle| = |\alpha| \cdot |\langle Tx, x\rangle|$, for all $x \in H$. The final property that $w(T)$ needs to satisfy is that

$$w(T + V) \leq w(T) + w(V),$$

which follows from the fact that $W(T + V) \subset W(T) + W(V)$.

To prove the equivalence, we take an element $\lambda \in W(T)$, i.e. $\lambda = \langle Tx, x\rangle$ with $||x|| = 1$. The Cauchy-Schwarz inequality then gives us that

$$|\lambda| \leq |\langle Tx, x\rangle| \leq ||Tx|| \cdot ||x|| = ||Tx|| \leq ||T||.$$

Because λ was chosen arbitrarily, this proves the first inequality. The second inequality is proven using the polarization identity, which can be found in section 2.4 of the lecture notes by de Snoo and Sterk [1].

$$4\langle Tx, y\rangle = \langle T(x+y), x+y\rangle - \langle T(x-y), x-y\rangle$$

$$+i\langle T(x+iy), x+iy\rangle - i\langle T(x-iy), x-iy\rangle$$

We now know that

$$4|\langle Tx, y\rangle| \leq |\langle T(x+y), x+y\rangle| + |\langle T(x-y), x-y\rangle|$$

$$+|i\langle T(x+iy), x+iy\rangle| + |i\langle T(x-iy), x-iy\rangle|.$$

Which simplifies to

$$4|\langle Tx, y\rangle| \leq w(T)\left(||x+y||^2 + ||x-y||^2 + ||x+iy||^2 + ||x-iy||^2\right),$$

$$= 4w(T)\left(||x||^2 + ||y||^2\right) = 8w(T).$$

The second inequality has therefore been proven. □

This equivalence of norms can be used to make many useful conclusions about bounded linear operators on a complex and separable Hilbert space. We should however be careful, for example, the implication that $T = 0$ if the numerical radius is zero is not necessarily true in a real Hilbert space. To show this, let us follow along with Gustafson and Rao and consider the following example on the real Hilbert space $H = \mathbb{R}^2$.

$$T = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}$$

If we take a unit vector $x = (f, g)$, we get that $Tx = (-g, f)$ and $\langle Tx, x\rangle = 0$. This implies that $w(T) = 0$. The operator norm is on the other hand is defined as follows [1]

$$||T|| = \sup_{x \neq 0} \frac{||Tx||}{||x||}$$

and for $x = (1, 0)$ we get that $\frac{||Tx||}{||x||} = 1$, which clearly implies that $||T|| > 0$.

A definition that is somewhat analogous to that of the numerical radius is that of the spectral radius.

Definition 3. *The spectral radius of an operator T is given by*

$$r(T) = sup\{|\lambda|, \lambda \in \sigma(T)\}.$$

We already saw in the last section that in some cases $w(T) = ||T||$, the next theorem shows that we also obtain the spectral radius when this is the case.

Theorem 8. *If $w(T) = ||T||$, then*

$$r(T) = ||T||.$$

Proof. Because of the basic properties of the numerical range, we are allowed to assume that $w(T) = ||T|| = 1$. When this is the case there exists a sequence of unit vectors $\{f_n\}$ such that $\langle Tf_n, f_n\rangle \to \lambda \in W(T)$, where $|\lambda| = 1$. Using the Cauchy-Schwarz inequality and the definition of the operator norm we get that

$$|\langle Tf_n, f_n\rangle| \leq ||Tf_n|| \cdot ||f_n|| = ||Tf_n|| \leq ||T|| = 1.$$

Combining this with the fact that $\langle Tf_n, f_n\rangle \to \lambda$, we can conclude that $||Tf_n|| \to 1$. Now to show that $\lambda \in \sigma_{app}(T)$ consider

$$||(T - \lambda I)f_n)||^2 = ||Tf_n||^2 - \langle Tf_n, \lambda f_n\rangle - \langle \lambda Tf_n, f_n\rangle + ||f_n||^2 \to 0,$$

where the convergence is caused by convergence of the components and the equality is caused by the linearity of the inner product. By definition, the convergence above tells us that $\lambda \in \sigma_{app}(T) \subset \sigma(T)$, so $r(T) = 1$. □

Next we will look at a theorem that allows us to find certain points of the point spectrum of T. To do this we should first define what the point spectrum of an operator actually is.

Definition 4. *The point spectrum of an operator T is given by*

$$\sigma_p = \{\lambda \in \sigma(T) : Tf = \lambda f \text{ for some } f \in H\}.$$

We are now ready to discus the theorem stated below.

Theorem 9. *If $\lambda \in W(T)$, $|\lambda| = ||T||$, then $\lambda \in \sigma_p(T)$.*

Proof. If $\lambda \in W(T)$, we know that there exists a unit vector $f \in H$, such that

$$\lambda = \langle Tf, f\rangle \to ||T|| = |\lambda| = |\langle Tf, f\rangle| \leq ||Tf|| \cdot ||f|| = ||Tf|| \leq ||T||.$$

Because of the squeeze property above, we have that

$$|\langle Tf, f\rangle| = ||Tf|| \cdot ||f|| \to Tf = \mu f \text{ for some } \mu \in \mathbb{C}.$$

However, we already knew that $\lambda = \langle Tf, f\rangle$, therefore

$$\lambda = \langle Tf, f\rangle = \langle \mu f, f\rangle = \mu.$$

This shows that $Tf = \lambda f$ and thus completes our proof. □

For the next theorem let us first recall that the range of an operator is defined as follows.

Definition 5. *The range of a linear operator T on H is defined as*

$$R(T) = \{Tf, f \in H\}.$$

We can use this definition to conclude something about the operator norm in the special case where the range of an operator is orthogonal to the range of its adjoint.

Theorem 10. *If $R(T) \perp R(T^*)$, then $w(T) = \frac{1}{2}||T||$.*

Before we prove this theorem, notice that this equality is equivalent to the extreme condition for the norm equivalence. This is therefore the maximal possible difference between the two norms.

Proof. In this proof we use that in Hilbert spaces

$$N(T) = \overline{R(T^*)}^{\perp}.$$

We can now apply this to a unit vector $f \in H$. Let $f = f_1 + f_2$, where $f_1 \in N(T)$ and $f_2 \in \overline{R(T^*)}$. For the corresponding element in $W(T)$ we now have that

$$\langle Tf, f\rangle = \langle T(f_1 + f_2), f_1 + f_2\rangle = \langle Tf_2, f_1\rangle.$$

To see this realize that $Tf_1 = 0$ and $\langle Tf_2, f_2\rangle = \langle f_2, T^* f_2\rangle = 0$, because $f_2 \in \overline{R(T^*)}$. We can use this to create the following string of inequalities

$$|\langle Tf, f\rangle| = |\langle Tf_2, f_1\rangle| \leq ||Tf_2|| \cdot ||f_1|| \leq ||T|| \cdot ||f_2|| \cdot ||f_1||$$

$$\leq \frac{||T||}{2}\left(||f_1||^2 + ||f_2||^2\right) = \frac{||T||}{2}.$$

Because this holds for any element $f \in H$, it also holds that

$$w(T) \leq \frac{||T||}{2} \leq w(T) \quad \Rightarrow \quad w(T) = \frac{||T||}{2}$$

The last inequality above is due to the norm equivalence. □

2.1.3 Normal Operators

We saw in the last section that for a selfadjoint operator T we have

$$r(T) = w(T) = ||T||$$

In this section we will see a property for normal operators that is somewhat analogous to it. We will also study several interesting inferences that can be made using the numerical range of normal operators. Let us start by recalling the definition of a normal operator T.

Definition 6. *A normal operator T on H is an operator that satisfies*

$$T^*T = TT^*.$$

The first theorem of this section shows us that we can sometimes directly infer the type of operator from its numerical range.

Theorem 11. *If $W(T)$ is a line segment, then T is normal.*

Proof. We will prove this by showing that the operator $e^{-i\theta}(T-\alpha I)$ is selfadjoint. It follows from the fact that T is then obtained by a rotation and a shift that T is normal.

Let α be a point on the line segment $W(T)$ and let θ be the inclination of this line segment. By rotating and shifting the line segment, we obtain that $W(e^{-i\theta}(T-\alpha I))$ is on the real line. We saw in the last section that this is a classification of selfadjoint operators and this therefore completes our proof. □

Let us study the theory above using an example. We saw in the proof of lemma 1 that the numerical range of the following operator is given by the line segment between $(1,i)$ and $(-1,-i)$ in the complex plane.

$$T = \begin{bmatrix} 1+i & 0 \\ 0 & -1-i \end{bmatrix}$$

It should then follow that the operator is normal. The adjoint of the matrix is given by the complex conjugate of its transpose.

$$T^* = \begin{bmatrix} 1-i & 0 \\ 0 & -1+i \end{bmatrix}$$

We can now compute TT^* and T^*T.

$$TT^* = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix} = T^*T$$

This shows that T is indeed normal.

We will now repeat and prove the above mentioned generalization of theorem 8 in the last section to normal operators.

Theorem 12. *If T is normal, then $||T^n|| = ||T||^n$, $n = 1, 2, \ldots$. Moreover, then*

$$r(T) = ||T||.$$

Proof. We will prove the first statement using induction. For the base step take any unit vector $x \in H$, we then have that

$$||Tx||^2 = \langle Tx, Tx\rangle = \langle T^*Tx, x\rangle = |\langle T^*Tx, x\rangle| \leq ||T^*Tx|| \cdot ||x|| = ||T^*Tx||$$

We now want to show that for a normal operator $||T^*Tx|| = ||T^2x||$, because it would then follow from the above inequality that $||T||^2 \leq ||T^2||$.

$$||T^2x||^2 = \langle T^2x, T^2x\rangle = \langle T^*T^2x, Tx\rangle$$

For the next step we use that T is normal.

$$= \langle TT^*Tx, Tx\rangle = \langle T^*Tx, T^*Tx\rangle = ||T^*Tx||^2$$

By taking the square root of both sides we obtain that $||T^*Tx|| = ||T^2x||$, so we may indeed conclude that $||T||^2 \leq ||T^2||$. Moreover, it is always true for bounded operators that $||T^2|| \leq ||T||^2$, so $||T^2|| = ||T||^2$.

By a similar argument as before, it holds that

$$||T^n x||^2 = \langle T^*T^n x, T^{n-1}x\rangle \leq ||T^{n+1}x|| \cdot ||T^n x||.$$

Combining this with the fact that $||T^2|| = ||T||^2$ and applying induction gives us that $||T^n|| = ||T||^n$ [9]. For the second statement we should recall that [9]

$$r(T) = \lim_{n\to\infty} ||T^n||^{1/n}$$

Combining this with what we just proved, we obtain that

$$r(T) = \lim_{n\to\infty} ||T^n||^{1/n} = ||T||.$$

□

The result above can be very useful when trying to determine the spectral radius of the power of a normal operator. To discus the next theorem, we first need to remind ourselves of the definition of the resolvent set of an operator.

Definition 7. *The resolvent set of an operator T is given by*

$$\rho(T) = \{\lambda \in H : \lambda \notin \sigma(T)\}.$$

We will state and prove this theorem, so that we can use it to prove the theorem that follows it. The theorem that follows it is an interesting geometrical result considering the spectrum of a normal operator. We first need to recall the definition of the distance between a point and a set [1].

Definition 8. *Let V be any set in a normed linear space $(X, ||\cdot||)$. Then the distance between x and V is given by*

$$d(x, V) = \inf\{||x - v|| : v \in V\}.$$

Now that we are familiar with this definition, we can study the following theorem.

Theorem 13. *Let z be any complex number in the resolvent set of a normal operator T. Then*

$$||(T - zI)x|| \geq d(z, \sigma(T)) \textit{ for } x \in H, ||x|| = 1.$$

Proof. We know that $(T - zI)$ is invertible, because $z \in \rho(T)$. We also know that $(T - zI)^{-1}$ is normal, because $(T - zI)$ is normal. Since it is normal, we have by the last theorem that

$$||(T - zI)^{-1}|| = r((T - zI)^{-1}).$$

The spectral mapping theorem tells us that

$$\sigma((T - zI)^{-1}) = \{(\lambda - z)^{-1} : \lambda \in \sigma(T)\}.$$

It follows that

$$||(T - zI)^{-1}|| = r((T - zI)^{-1}) = \frac{1}{d(z, \sigma(T))}.$$

We can now use this to conclude that, for any $x \in H$ with $||x|| = 1$,

$$d(z, \sigma(T)) = ||(T - zI)^{-1}||^{-1} \leq ||(T - zI)x||.$$

□

Theorem 14. *The closure of the numerical range of a normal operator is the convex hull of its spectrum.*

Proof. We can always scale and shift the spectrum in such a way that it is contained in a closed half-plane of $\mathbb{C}$. This means that proving this theorem is equivalent to proving that any half-plane containing $\sigma(T)$ also contains $W(T)$. Let us now assume that $\sigma(T)$ is contained in the left half-plane of $\mathbb{C}$ and that the imaginary axis is a supporting line of the convex hull of $\sigma(T)$, i.e. the line that is tangent to the convex hull and that does not split it. Let us now assume that $a + bi \in W(T)$, with $a > 0$. If this is possible it would contradict what we wanted to show. Let us set $Tx = (a + bi)x + y$, with x and y orthogonal. We then get that $\langle Tx, x\rangle = a + bi$. Notice that any positive real number c is in the resolvent set of T. We thus have by the inclusion in the half-plane and the theorem above that

$$c \leq d(c, \sigma(T)) \leq ||(T - cI)x|| = ||(a + ib - c)x + y||.$$

If we now square these terms we get that

$$\begin{aligned} c^2 \leq ||(a + ib - c)x + y||^2 &= (a - c)^2 + b^2 + ||y||^2 \\ &= a^2 - 2ac + c^2 + b^2 + ||y||^2. \end{aligned}$$

This can only be the case if $2ac \leq a^2 + b^2 + ||y||^2$, because we can take any positive real number c this can not always be the case. Therefore the theorem has been proven by contradiction. □

In the next theorem we make an important discovery concerning the closure of the numerical range of a normal operator and its eigenvalues. To do this we first have to define what an extreme point is [9].

Definition 9. *A point z is an extreme point of a set S if $z \in S$ and there is a closed half-plane containing z and no other element of S.*

We now use this definition in the following theorem.

Theorem 15. *The extreme points of the closure of the numerical range $W(T)$ of a normal operator T are eigenvalues of T if and only if $W(T)$ is closed.*

Proof. Let us first assume that $W(T)$ is closed. As before we can shift the numerical range to simplify our proof. Let $W(T) \subset \{\lambda : Im\lambda \geq 0$ and let the extreme point z be the origin. If we now consider $\langle Tx, x\rangle = 0 \in W(T)$, we have that

$$\langle Tx, x\rangle = \overline{\langle Tx, x\rangle} \quad \Rightarrow \quad \langle (T - T^*)x, x\rangle = 0.$$

We know that the operator $\frac{1}{i}(T - T^*)$ is positive, since it is selfadjoint and for any $x \in H$ it can be shown that

$$\langle \frac{1}{i}(T - T^*)x, x\rangle \geq 0.$$

These claims follow from lemma 1 in the short paper by J.G. Stampfli [7]. Therefore it must be the case that $(T - T^*)x = 0$, which implies that $x \in \{f : Tf = T^*f\} = N$. Combining this with the fact that T is normal gives us that

$$T^*Tx = TT^*x = TTx.$$

From which we can conclude that N is invariant under T and that the restriction of T on N is selfadjoint. Using the fact that the numerical range of a selfadjoint operator is real, we get that $W(T|_N) \subset W(T) \bigcup \mathbb{R} = \{0\}$. Because of this, $T|_N = 0 \rightarrow Tx = 0$, from which we can conclude that 0 is an eigenvalue of T.

Now we want to prove that it can only be the case if $W(T)$ is closed. We know that the $\overline{W(T)}$ is compact and convex. It is clear that it is also the convex hull of its extreme points. We assume that they are eigenvalues of T in the theorem so it must also be contained in the hull of the pointspectrum, i.e.

$$\overline{W(T)} \subset co(\sigma_p(T)) \subset co(W(T)) = W(T),$$

which completes the proof. □

If we check this theorem for the example that was introduced earlier this section, we see that $W(T)$ is closed and that the extreme points of the line are indeed the eigenvalues of the matrix.

2.1.4 Numerical Boundary

Because of the convexity of the numerical range, it could be of great interest to know which points of the numerical range are on the boundary. To get a bit more insight into the boundary of the numerical range we introduce two theorems considering the extreme points of $W(T)$ in this section.

Definition 10. *For any complex number z, let M_z be the subset of H given by*

$$M_z = \{x \in H : \langle Tx, x\rangle = z||x||^2\}.$$

It is clear that this set does not have to be linear. The following theorem gives us a condition that guarantees linearity of M_z.

Theorem 16. *If $z \in W(T)$ is an extreme point of $W(T)$, then M_z is linear.*

Proof. Let L be the line of support of $W(T)$ at z, i.e. the line that contains z and does not split $W(T)$. The existence of this line tells us that there exists some rotation $e^{-i\theta}$, such that by a rotation and a shift of z to the origin we get that $Re\langle e^{-i\theta}(T - zI)x, x\rangle \geq 0$, for any $x \in H$. Consequently,

$$M_z = \{x \in H : Re\langle e^{-i\theta}(T - zI)x, x\rangle = 0\}.$$

Let us now check that it is indeed linear. For $x, y \in M_z$ it follows from straightforward computation that

$$\langle e^{-i\theta}(T - zI)(x + y), (x + y)\rangle = -\langle e^{-i\theta}(T - zI)(-x + y), (-x + y)\rangle.$$

We showed that $Re\langle e^{-i\theta}(T - zI)x, x\rangle \geq 0$, for any $x \in H$, so in the equation above the real part on the left is greater or equal to zero and on the right it is smaller or equal to zero, due to the minus sign. This can only be the case if both terms are purely imaginary. This shows that $x + y \in M_z$ and thus concludes the proof. □

The relation in the theorem above is also true in the opposite direction, as is illustrated in the following theorem.

Theorem 17. *$z \in W(T)$ is an extreme point if M_z is linear.*

We will not prove this theorem in this paper, but the proof can be found in the book by Gustafson and Rao [9]. Notice that these two theories combined give us a characterization of the extreme points of the numerical range.

2.2 Mapping Theorems

In this section we will study if it is possible to construct mapping theorems for the numerical range and radius. We would of course like to find something analogous to the spectral mapping theorem, but we will see that we are limited by the convexity of the numerical range. It is however not completely impossible to relate the numerical ranges and radii of operators of the form $f(T)$ to those of T. We start this section with an important example that illustrates the nontrivial relation between $W(T)$ and $W(f(T))$.

Let $H = \mathbb{C}^2$ and $f(T) = T^2$ for

$$T = \begin{bmatrix} 4+1 & 4i \\ 4i & 16+4i \end{bmatrix}.$$

If we take $u = (x, y) \in \mathbb{C}^2$, we get that

$$Tu = ((4i + 1)x + 4iy, 4ix + (16 + 4i)y)$$

and

$$\langle Tu, u\rangle = (4 + i)|x|^2 + (16 + 4i)|y|^2 = 4|x|^2 + 16|y|^2 + i(|x|^2 + 4|y|^2).$$

So we can conclude that the real part is always larger than the imaginary part. However if we consider $u = (1, 0)$, we get

$$T^2u = (8i - 1, 80i - 20)\,.$$

This gives us that

$$\langle T^2u, u\rangle = 8i - 1.$$

Obviously the conclusion that the real part is always larger than the imaginary part no longer holds for T^2.

2.2.1 Radius Mapping

As mentioned before, it is not possible to relate $W(T)$ to $W(f(T))$ in general, but for some special cases we can make interesting conclusions. In this subsection we will look at the special cases where it is possible to find bounds for the numerical radius of an operator of the form $f(T)$. The first theorem we look at considers the case where $f(T) = T^n$.

Theorem 18. *(Power inequality) Let T be an operator and $w(T) \leq 1$. Then $w(T^n) \leq 1$, $n = 1, 2, 3, \cdots$*

Proof. We start the proof by observing that, for any $z \in \mathbb{C}$ with $|z| < 1$, we have

$$Re\langle(I - zT)x, x\rangle = ||x||^2 - Re\langle zTx, x\rangle \geq ||x||^2(1 - |z|) \geq 0.$$

Where we used the linearity in the first argument of the inner product, $|z| < 1$ and the fact that $w(T) \leq 1$. Let us now observe that if it is true that $Re\langle(I - zT)x, x\rangle \geq 0$, for all $z \in \mathbb{C}$ with $|z| < 1$, we can take $z = te^{i\alpha}$ and let $t \to 1$. If we do this we obtain that

$$0 \leq Re\langle(I - e^{i\alpha}T)x, x\rangle = ||x||^2 - Re\langle e^{i\alpha}Tx, x\rangle,$$

this can only be the case if $||x||^2 \geq Re\langle e^{i\alpha}Tx, x\rangle$. Because this relation holds for all α, we can always choose a rotation α of $\langle Tx, x\rangle$, such that $\langle e^{i\alpha}Tx, x\rangle = |\langle Tx, x\rangle|$. If we then apply this equality to the elements of $W(T)$ we get that $w(T) \leq 1$.

It follows that whenever $I - zT$ is invertible, $Re\langle(I - zT)x, x\rangle \geq 0$, for all $x \in H$, if and only if it holds that $Re\langle(I - zt)^{-1}y, y\rangle \geq 0$, for all $y \in H$. The invertibility allows us to take $x = (I - zT)^{-1}y$ to obtain this very inequality. We should also notice that $r(T) \leq 1$ by the spectral inclusion quality and that

$I - zT$ is invertible. The invertibility follows from *Theorem 4.36* in the lecture notes by Sterk and de Snoo [1].

Now that we have shown these equivalent conditions we can prove our theorem by showing that

$$Re\langle (I - z^n T)^{-1} y, y\rangle \geq 0 \text{ with } z \in \mathbb{C}, |z| < 1 \text{ for all } y \in H.$$

To do this we need the following identity

$$(I - z^n T)^{-1} = \frac{1}{n}\left((I - zt)^{-1} + (I - \omega z^2 T)^{-}1 + \cdots + (I - \omega^{n-1} z^n T)^{-1}\right),$$

where ω is a primative $n-$th root of 1. It follows that $|\omega^{n-1} z^n| < 1$. We use this to conclude that $Re\langle (I - \omega^{n-1} z^n T)^{-1} y, y\rangle \geq 0$ for any $y \in H$. Let us now see that

$$Re\langle (I - z^n T)^{-1} y, y\rangle = Re\langle (\frac{1}{n}((I - zt)^{-1} + (I - \omega z^2 T)^{-}1 + \cdots$$

$$+ (I - \omega^{n-1} z^n T)^{-1}))y, y\rangle \geq 0,$$

which completes our proof. □

A very simple example of this theorem would be the following matrix acting on $\mathbb{C}^2$.

$$T = \begin{bmatrix} 0 & 2 \\ 0 & 0 \end{bmatrix}$$

We saw in the proof of lemma 1 that the numerical range of T is given by a circle of radius $\frac{2}{2}$ centered at $(0, 0)$, i.e. the unit circle. We can thus conclude that $w(T) = 1$. We also know that $T^2 = 0$ and therefore $w(T^n) = 0$ for $n = 2, 3, \cdots$, so the theorem does indeed hold for this operator.

A necessary condition for the theorem above is that $w(T) \leq 1$, we saw in the proof that it can be useful to use equivalent conditions instead. This is why Gustafson and Rao introduce the following theorem.

Theorem 19. *(Power dilation) $w(T) \leq 1$ if and only if $T^n = 2PU^nP$, for $n = 1, 2, 3, \cdots$, where U is a unitary operator on a Hilbert space $K \supset H$ and P is the projection of K on H.*

Proof. As mentioned before by the spectral inclusion property we have that $w(T) \leq 1 \Rightarrow r(T) \leq 1$ and $(I - zT)$ is invertible for $|z| \leq 1$. The function $F(z) = (I - zT)^{-1}$ is holomorphic for $|z| < 1$, moreover $F(0) = I^{-1} = I$ and $Re\langle F(z)x, x\rangle \geq 0$. It is explained in a paper by Sz.-Nagy and Foias [2], that it follows from a theorem of Riesz that there exists a Hilbert space $K \supset H$ and a unitary operator U in K, such that

$$F(z) = P(I_K + zU)(I_K - zU)^{-1} \quad \text{for any } z \text{ with } \quad |z| < 1,$$

where P is the projection of K on H. They then tell us that, because

$$(I_K + zU)(I_K - zU)^{-1} = I + 2zU + \cdots + 2z^n U^n + \cdots \quad |z| < 1,$$

we get that

$$\begin{aligned} F(z) &= P\left(I + 2zU + \cdots + 2z^n U^n + \cdots\right) \\ &= I_H + zT + z^2T^2 + \cdots \quad . \end{aligned}$$

Using this equality one can obtain by computing the coefficients that

$$T^n = 2PU^nP \quad \text{for} \quad n = 1, 2, 3, \ldots \quad .$$

We know that the series $I_H + zT + z^2T^2 + \cdots$ converges to $(I_H - zT)^{-1}$ for $|z| < 1$. The fact that $T^n = 2PU^nP$ implies that the term inside the projection also converges for $|z| < 1$, because the T^n are bounded. Combining these facts we get that the term inside the projection, i.e.

$$I_H + zT + z^2T^2 + \cdots \quad ,$$

converges to $(I - zT)^{-1}$. Let us now consider the following inner product

$$\langle (I_K + zU)(I_K - zU)^{-1}y, y\rangle.$$

If we take $y = (I_K - zU)^{-1}x$, we get that

$$Re\langle (I_K + zU)x, (I_K - zU)x\rangle = (1 - |z|^2)||x||^2 \geq 0.$$

The inequality above implies that $Re\langle F(z)x, x\rangle \geq 0$ for $|z| < 1$. This was shown to be an equivalent condition for $w(T) \leq 1$ in the last proof and thus concludes this proof. □

As mentioned before, the main point of this theorem is to give an equivalent condition for $w(T) \leq 1$. We will now use this equivalent condition in the proof of the next theorem.

Theorem 20. *Let f be analytic inside the unit disk and continuous on the boundary, with $f(0) = 0$. If $|f(z)| \leq 1$ for $|z| \leq 1$ and $w(T) \leq 1$, then $w(f(T)) \leq 1$.*

Proof. Gustafson and Rao start their proof by showing that $\lim_{r\to 1} f(rT)$ exists. Notice that f inside the unit disk is given by a convergent power series, because it is analytic.

$$f(rT) = \sum a_n r^n T^n = 2P(\sum a_n r^n U^n)P$$

The last equality is caused by the previous theorem. If we now take $U = \int e^{i\lambda} dE(\lambda)$, we get that

$$= 2P\left(\int \sum a_n r^n e^{in\lambda} dE(\lambda)\right)P.$$

It follows from the power series representation of f that this is equal to

$$= 2P\left(\int f(re^{i\lambda})dE(\lambda)\right)P.$$

Observe that $|re^{i\lambda}| \leq 1$, so f is continuous. The continuity allows us to conclude the following.

$$\lim_{r\to 1} f(rT) = 2P\left(\int f(e^{i\lambda})dE(\lambda)\right)P = 2Pf(U)P$$

It now follows from the fact that P is a projection, that

$$f(T)^n = 2Pf(U)^nP.$$

If $f(U)$ is a unitary operator the above equality in combination with the last theorem show that $w(f(T)) \leq 1$. It follows from U and the criteria for f in the theorem that $f(U)$ is a contraction. B. Nagy has shown that every contraction in a Hilbert space has a unitary dilation [3]. □

The previous theorem dealt with a function f that was analytic inside the unit disk, we will now look at some theorems that relate the numerical range $W(f(T))$ to $W(T)$ when f is analytic in some other regions.

Theorem 21. *Let $f(z)$ be holomorphic on $\{z : |z| \leq 1\} = D$, and map D into $\{z : Rez \geq 0\} = P$. If $W(T) \subset \mathbb{S}^1$, then $W(f(T)) \subset P - Ref(0)$.*

Proof. To prove this theorem we use the following expression for $f(z)$ from a paper by Stone [12].

$$f(z) = iImf(0) + \frac{1}{2\pi}\int_0^{2\pi}(Ref(e^{it}))\frac{e^{it}+z}{e^{it}-z}dt$$

If we now rewrite the fraction in the last expression as follows

$$\frac{e^{it}+z}{e^{it}-z} = \frac{2e^{it}+z-e^{it}}{e^{it}-z} = \frac{2e^{it}}{e^{it}-z} - 1$$

$$= \frac{2}{1-\frac{z}{e^{it}}} - 1 = 2(1-e^{-it}z)^{-1} - 1.$$

We can now substitute this back into our expression for $f(z)$

$$f(z) = i\,\mathrm{Im}\,f(0) + \frac{1}{2\pi}\int_0^{2\pi}\left[\mathrm{Re}\,f\left(e^{it}\right)\right]\left[2\left(1-e^{-it}z\right)^{-1} - 1\right]dt,$$

$$= iImf(o) - \frac{1}{2\pi}\int_0^{2\pi} Re(f(e^{it}))dt + \frac{1}{\pi}\int_0^{2\pi}\left[\mathrm{Re}\,f\left(e^{it}\right)\right]\left(1-e^{-it}z\right)^{-1}dt,$$

$$= -f(0) + \frac{1}{\pi}\int_0^{2\pi}\left[\mathrm{Re}\,f\left(e^{it}\right)\right]\left(1-e^{-it}z\right)^{-1}dt.$$

Let us now apply this function to our operator A and take a look at the real part, as that is what we are most interested in for this theorem.

$$Ref(A) = -Re\ f(0) + \frac{1}{\pi}\int_0^{2\pi} \left[\mathrm{Re}\, f\left(e^{it}\right)\right] Re\left[\left(I - e^{-it}A\right)^{-1}\right] dt$$

It follows from the fact that $e^{it} \in D$, that $Ref(e^{it}) \geq 0$. The fact that $W(A) \subset D$ implies that $Re(1 - e^{-it}A) \geq 0$. We can then conclude that $Re\left[\left(I - e^{-it}A\right)^{-1}\right] \geq 0$, by *lemma 4* from the text by Kato [14]. This means that the integral above is positive. Because of this we have that $Re\ f(A) \geq -Re\ f(0)$, which completes the proof. □

2.2.2 Operator Products

We have already seen that the numerical range behaves well under addition of operators, but we have not made many claims about the product of operators. In this section we will study a special case where we can do exactly that.

Theorem 22. *If* $0 \notin \overline{W(A)}$, *then*

$$\sigma\left(A^{-1}B\right) \subset \frac{\overline{W(B)}}{\overline{W(A)}} = \left\{\frac{\lambda}{\mu}, \quad \lambda \in \overline{W(B)}, \mu \in \overline{W(A)}\right\}$$

Proof. We start by realizing that $0 \notin \sigma(A)$, this is due to the inclusion of the spectrum in $\overline{W(T)}$. The spectrum of A are the λ, such that $T - \lambda I$ is not invertible. Because 0 is not in the spectrum, we can conclude that A is invertible. Let us consider $\lambda \in \sigma(A^{-1}B)$. This has as a consequence that $0 \in \sigma(A^{-1}B - \lambda I)$. Moreover, we know that

$$A^{-1}B - \lambda I = A^{-1}(B - \lambda A).$$

We can conclude from this equality that if $\lambda \in \sigma(A^{-1}B)$ then $0 \in \sigma(B - \lambda A)$. The spectral inclusion now gives us that

$$0 \in \overline{W(B) - \lambda W(A)} \quad \Rightarrow \quad 0 \in \overline{W(B)} - \lambda\overline{W(A)},$$

which means that there exist $x \in \overline{W(A)}$ and $y \in \overline{W(B)}$, such that $\lambda = \frac{y}{x} \in \frac{\overline{W(B)}}{\overline{W(A)}}$. □

The next special case that we will consider is commuting operators, i.e. $AB = BA$.

Theorem 23. *Let A be nonnegative, selfadjoint operator and $AB = BA$. Then $W(AB) \subset W(A)W(B)$.*

Proof. Because A is nonnegative we can take the nonnegative square root $A^{1/2}$. Using this square root we obtain that

$$\langle ABx, x\rangle = \langle BA^{1/2}, A^{1/2}x\rangle. \tag{1}$$

Let us now define $g = \frac{A^{1/2}}{||A^{1/2}||}$, notice that this is a unit vector. We then get that

$$= ||A^{1/2}||^2 \langle Bg, g\rangle = \langle A^{1/2}x, A^{1/2}x\rangle\langle Bg, g\rangle = \langle Ax, x\rangle\langle Bg, g\rangle,$$

where $\langle Ax, x\rangle \in W(A)$ and $\langle Bx, x\rangle \in W(B)$. □

When we study the numerical radius instead of the numerical range, we can prove the following less restricted theorem.

Theorem 24. *It is always the case that*

$$w(AB) \leq 4w(A)w(B).$$

When AB=BA, it always hold that

$$w(AB) \leq 2w(A)w(B).$$

Proof. The first claim follows almost directly from the equivalence of norms, that was proved in an earlier section.

$$w(AB) \leq ||AB|| \leq ||A|| \cdot ||B|| \leq 4w(A)w(B)$$

Let us now look at the commuting case. We assume here that $w(A) = w(B) = 1$, because we can always scale them. Let us first use a different representation of AB, which allows us to use the triangle inequality.

$$w(AB) = w\left(\frac{1}{4}\left[(A+B)^2 - (A-B)^2\right]\right)$$

Next we use the fact that $w(T)$ is a norm.

$$\leq \frac{1}{4}\left[w((A+B)^2) + w((A-B)^2)\right]$$

We can now use theorem 18.

$$\leq \frac{1}{4}\left[w((A+B))^2 + w((A-B))^2\right]$$

Now we apply the subadditivity of the numerical radius.

$$\leq \frac{1}{4}\left[(w(A) + W(B))^2 + (w(A) - w(B))^2\right] = 2$$

It follows from our assumption $w(A) = w(B) = 1$ that this completes the proof. □

2.3 Finite Dimensions

In this paper we consider several examples of operators T acting on the finite-dimensional space $\mathbb{C}^n$, it is therefore interesting to study several simplifications and consequences when considering operators on this space. The first interesting result is the following theorem

Theorem 25. *The numerical range of any matrix A on $\mathbb{C}^n$ is compact and the numerical radius in attained.*

Proof. We should first of all observe that the function that maps x to its corresponding element in $W(T)$ is continuous on the compact set $\{x \in \ : ||x|| = 1\}$. We therefore have that the numerical radius is attained by an element in $W(T)$. □

Another result that is somewhat analogous to some of the results that we showed before is the following theorem that considers matrices for which the numerical range is a real interval $[m, M]$.

Theorem 26. *The numerical range of a symmetric matrix A is the real interval $[m, M]$, where m and M are the smallest and largest eigenvalue of A respectively.*

Proof. We know that $W(A)$ is convex and we saw in the last theorem that it is also compact. A theorem in an earlier section showed us that if $\overline{W(A)} = [m, M]$, then $m, M \in \sigma(T)$. We now have that $\overline{W(A)} = W(A) = [m, M]$, thus m and M are indeed in the spectrum of A. Moreover, we know that the entire spectrum is contained in $W(A)$, thus m and M must be the smallest and largest eigenvalue respectively. □

The last theorem showed us the strong connection between the eigenvalues of a symmetric matrix and its eigenvalues, the next theorem proves a similar result for unitary matrices.

Theorem 27. *The numerical range of a unitary matrix A is a polygon inscribed in the unit circle.*

Proof. A unitary matrix is an example of a normal matrix, it follows from theorem 14 that the closure of its numerical range is the convex hull of its spectrum. This means that we can prove the statement by showing that the eigenvalues are on the unit circle. Let us now find the λ such that $Ax = \lambda x$. To find these lambda we study the norm of Ax and λx:

$$||Ax||^2 = \langle Ax, Ax \rangle = \langle A^*Ax, x \rangle = \langle Ix, x \rangle = ||x||^2,$$

whilst on the other hand

$$||\lambda x||^2 = |\lambda|^2 ||x||^2.$$

Consequently the equality $Ax = \lambda x$ can only be preserved if $|\lambda|^2 = 1$. This implies that the eigenvalues of A are indeed on the unit circle. □

2.4 Operator Trigonometry

In this section we will discuss angles of operators. The definitions and results in this section have various interesting results, of which some will be discussed in the next section.

2.4.1 Operator Angles

We start this section by giving the original definition of the cosine of an operator. The definition that we give here should remind us of the definition of the dot product in a Euclidean space.

Definition 11. *The real cosine of T in H is defined as*

$$\cos T = \inf_{x \in H} \frac{Re\langle Tx, x\rangle}{||Tx|| \cdot ||x||}, \quad x \neq 0, Tx \neq 0.$$

Clearly this is a real quantity, the imaginary part and the total cosine (perhaps cosine magnitude would have been a better name) are defined in a similar manner. To see this let us look at the definition of the total cosine

$$|\cos|T = \inf_{x \in H} \frac{|\langle Tx, x\rangle|}{||Tx|| \cdot ||x||}, \quad x \neq 0, Tx \neq 0.$$

The angle $\phi(T)$ is determined by the definition of the real cosine. The geometrical interpretation of this angle is that it measures the maximum turning effect of an operator T on H. Let us as an example consider the rotation matrix on $\mathbb{C}^2$. The rotation matrix for a rotation of $90-$degrees is given by T below.

$$T = \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix}$$

If we take a vector $x = (f, g)$ we get that

$$Tx = (-f, -g) \quad \text{and} \quad \langle Tx, x\rangle = -(|f|^2 + |g|^2).$$

We can now use this to compute the real cosine that was defined at the beginning of this section.

$$\cos T = \inf_{x \in H} \frac{Re\{-(|f|^2 + |g|^2)\}}{\sqrt{|f|^2 + |g|^2}\sqrt{|f|^2 + |g|^2}} = -1,$$

which means that $\phi(T) = 90°$. We can conclude that the operator angle agrees with the rotation angle of the matrix.

The way in which $\cos T$ was constructed has the following direct consequences for the angle $\phi(T)$.

Property 2. *For an operator T we have that*

$$\phi(T) = \phi(T^{-1}) = \phi(cT) = \phi(cT^{-1}).$$

The definition of the total cosine also allows us to find a lower and upper bound using the operator norm and the numerical radii. The theorem below illustrates how we can do this.

Theorem 28. *For any operator T, its cosine is bounded by the upper and lower numerical radii $m(T)$ and $w(T)$:*

$$\frac{m(T)}{||T||} \leq |\cos|T \leq \frac{w(T)}{||T||}.$$

Proof. The proof is based on the basic property from analysis that for positive sequences a_n and b_n we have that

$$\inf(a_n)\inf(b_n) \leq \inf(a_n b_n) \leq \sup(a_n)\inf(b_n).$$

If we set $a_n = |\langle Tx, x\rangle|$ and $b_n = \frac{1}{||Tx||}$ the theorem follows directly from this property above when considering unit vectors only. □

Calculating the angle of an operator using the definition requires to calculate the infimum of the real parts of the elements of the numerical range, this can sometimes be tricky. The following theorem shows that it is sometimes not necessary to use the definition.

Theorem 29. *For T a strongly positive $(m > 0)$ selfadjoint operator,*

$$\cos T = \frac{2\sqrt{mM}}{m+M},$$

where $m = m(T)$ and $M = w(T)$.

Proof. To prove this we will use the Kantorovich inequality

$$\max_{||y||=1}\{\langle y, Ty\rangle\langle y, T^{-1}y\rangle\} = \frac{1}{4}\left(\sqrt{\frac{M}{m}} + \sqrt{\frac{m}{M}}\right)^2.$$

Next we will have to change the minimization in the definition of the cosine to a maximization. To do this observe that

$$\left(\min_x \frac{\langle Tx, x\rangle}{||Tx|| \cdot ||x||}\right)^{-2} = \left(\max_x \frac{||Tx||||x||}{\langle Tx, x\rangle}\right)^2 = \max_x \frac{||Tx||^2||x||^2}{\langle Tx, x\rangle^2}.$$

To simplify our expression even further we need a smart choice of x. The one given by Gustafson and Rao is $\tilde{x} = \langle Tx, x\rangle^{-1/2}x$. We then obtain by substitution that

$$\max_{\tilde{x}} \frac{||Tx||^2||x||^2}{\langle Tx, x\rangle^2} = \max_x \frac{\langle Tx, x\rangle^{-2}||Tx||^2||x||^2}{\langle Tx, x\rangle^2} = \max_{\langle Tx, x\rangle=1} ||Tx||^2||x||^2$$

$$= \max_{\langle Tx, x\rangle=1} \langle T^{1/2}x, T^{3/2}x\rangle\langle T^{1/2}x, T^{-1/2}x\rangle.$$

The maximization over $\langle Tx, x\rangle$ feels a bit unfamiliar. To get rid of it we take $y = T^{1/2}x$, such that $||y||^2 = \langle Tx, x\rangle = 1$. Using this as our maximization turns the one above into the Kantorovich inequality.

$$\max_{||y||=1} \langle y, Ty\rangle\langle y, T^{-1}y\rangle = \frac{1}{4}\left(\frac{M}{m} + \frac{m}{M} + 2\right) = \frac{M^2 + m^2 + 2mM}{4mM}$$

We started of with a power of -2, so $\cos T$ is the inverse square root of the equated solution.

$$\cos T = \frac{2\sqrt{mM}}{M+m}$$

□

We will use this theorem in an upcoming section on the study of the convergence of iterative methods. It will show that the convergence of the steepest descent method is related to the maximum turning effect of its operator.

2.4.2 Operator deviations

In the book by Gustafson and Rao [9] the deviation of an operator is discussed, which is equivalent to the operator angle $\phi(T)$.

Definition 12. *The deviation of an operator T is given by*

$$dev(T) = \sup_{x\in H} \phi(Tx, x),$$

where $\phi(Tx, x)$, $0 \leq \phi \leq \pi$, is defined by

$$cos(\phi(Tx, x)) = \frac{Re\langle Tx, x\rangle}{||Tx|| \cdot ||x||}.$$

Notice that this definition is indeed equivalent to that of the operator angle. Gustafson and Rao then introduce a lemma from a text by Krein [11]. They point out that the theorem was stated without a proof in the original text. Because the proof has no real additional value to what we are still going to discuss, we choose to omit it as well.

Lemma 2. *Let x,y,z be three unit vectors in a Hilbert space. Define the angles ϕ_{xy},ϕ_{yz},ϕ_{xz} by $\cos\phi_{xy} = Re\langle x, y\rangle$, $\cos\phi_{yz} = Re\langle y, z\rangle$, $\cos\phi_{xz} = Re\langle x, z\rangle$, respectively, with $o \leq \phi_{xy}, \phi_{yz}, \phi_{xz} \leq \pi$. Then*

$$\phi_{xz} \leq \phi_{xy} + \phi_{yz}.$$

The result of this theorem is very interesting however. The result in combination with the equivalence with the operator angle can now be used to prove the following theorem.

Theorem 30. *Let A and B be bounded invertible operators in a Hilbert space. Then*

$$\phi(AB) \leq \phi(A) + \phi(B).$$

Proof. If follows from lemma 2 that

$$\phi(ABx, x) \leq \phi(ABx, A^{-1}x) + \phi(A^{-1}x, x)$$

$$= \phi(Bx, x) + \phi(Ax, x).$$

The inequality will be preserved if we take the suprema of the seperate terms. Doing this will give us that

$$dev(AB) \leq dev(B) + dev(A)$$

or equivalently

$$\phi(AB) \leq \phi(A) + \phi(B),$$

which completes the proof. □

This result should not be very surprising considering the geometrical interpretation. Geometrically this theorem tells us that the maximum turning effect of AB is less or equal to the the maximum turning effect of B followed by A.

2.4.3 Antieigenvalue Theory

Gustafson and Rao [9] mention in the preface of their book that they think that the theory of antieigenvalues will eventually become a standard chapter in linear algebra [9] and that it has a large variety of applications. These claims should spark our interest, let us therefore define what an antieigenvalue is. The first antieigenvalue is really just a different interpretation of the real cosine of an operator A, namely

Definition 13. *The first antieigenvalue of A is given by*

$$\mu_1(A) = \inf_{\substack{x \in D(A) \\ x \neq 0, Ax \neq 0}} \frac{Re\langle Ax, x\rangle}{||Ax|| \cdot ||x||},$$

The higher antieigenvalues are defined as

$$\mu_n(A) = \inf_{\substack{x \in D(A) \\ x \perp \{x_1, \cdots, x_{n-1}\}}} \frac{Re\langle Ax, x\rangle}{||Ax|| \cdot ||x||},$$

where the x_k are the corresponding antieigenvectors of A. It should be noted that the higher antieigenvalues are only well defined if it is assumed that the previous antieigenvalues are attained by their corresponding antieigenvectors [8].

So in contrast to the normal combination of eigenvalue and eigenvector, which are the vectors for which A does not turn at all, these antieigenvalues denote the critical turning effects of A. A direct consequence of this definition and theorem 29 is that for a strongly positive (i.e. $m(T) > 0$), selfadjoint operator T we have that

$$\mu_1(T) = \frac{2\sqrt{\lambda_{min}\lambda_{max}}}{\lambda_{min} + \lambda_{max}},$$

where $\lambda_{min} = m(T)$ and $\lambda_{max} = M(T)$, the lower and upper bound of $W(T)$ respectively. Let us as an example consider the following matrix.

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

Clearly this matrix is selfadjoint. We have also seen that $W(T)$ is given by the line segment $[1, 2]$. This means that $\lambda_{min} = 1$ and λ_{max}=2. We can conclude from the fact that $\lambda_{min} > 0$ that the matrix is also strongly positive. Substitution in the formula above gives us that

$$\mu_1 = \frac{2\sqrt{1 \cdot 2}}{1+2} = \frac{2\sqrt{2}}{3}.$$

This example illustrates that in some cases there are strong relations between the eigenvalues of an operator and the antieigenvalues of an operator. In the book by Gustafson and Rao several more examples are given of perticular cases where the two can be related.

3 Approximations and Numerical Methods

In the preceding chapter we have seen a lot of theoretical properties of the numerical range. We will use this chapter to dive a little deeper into the applications of the numerical range in, for example, fluid dynamics using the before mentioned theoretical properties.

3.1 Approximating the Numerical Range

Before studying some of the applications, we do some numerical analysis of $W(T)$ ourselves. For an operator on the complex numbers it is possible to create an algorithm that approximates its numerical range. We are going to do this for an operator T using random points with the MatLab script given below.

```
function [area,conv]=generalnmr2(T,nit)
l=length(T);
for i=1:nit
    x(:,i)=-1 + 2.*rand(1,l)+1i.*(-1+2.*rand(1,l));
    x(:,i)=x(:,i)/norm(x(:,i));
    n(i)=dot(T*x(:,i),x(:,i));
    if i>2
        C=convhull(real(n),imag(n));
        area(i)=polyarea(real(n(C)),imag(n(C)));
        conv(i)=area(i)/area(i-1);
    end
end

figure(1)
hold on
scatter(real(n),imag (n));
plot(real(n(C)),imag(n(C)));

figure(2)
plot(area);

end
```

This function takes as an input a operator T and a number of iterations *nit*. It creates a random unit vector and computes the corresponding element of $W(T)$ in each iteration. It also approximates the convex hull corresponding to this points and its area. We saw in the previous chapter that the numerical range is convex, because the algorithm keeps adding points to a convex set we know that the area is strictly increasing. By studying the convergence of the area, we can try to say something about how close we are getting to the actual numerical range. We should note that the *rand* command in MatLab creates pseudorandom numbers using a uniform distribution. It is therefore possible for the script to keep creating points inside of the convex hull, whilst there are still points of $W(T)$ outside of the hull to be considered.

To study the effectiveness of our script we go back to the example in the introduction.

$$T = \begin{bmatrix} 1 & 2 \\ 1 & -1 \end{bmatrix}$$

We choose this operator, because we have an exact equation for the numerical range of this operator. The code produces two plots, one of them shows a scatter plot of the randomly generated elements in $W(T)$ together its convex hull generated by the MatLab command *convhull*, the other plot illustrates how the area grows with respect to the number of iterations. If we run this code for our operator T and $nit = 1000$ it returns the following two plots.

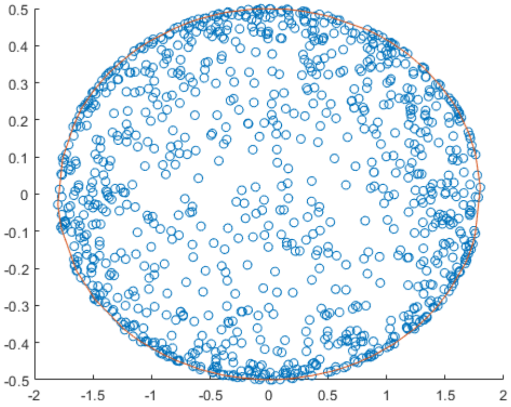

Figure 3: Scatter plot of the Numerical Range of matrix A in $\mathbb{C}^2$.

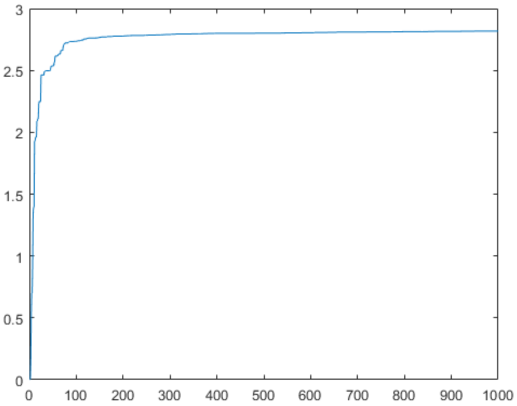

Figure 4: Area of convex hull of matrix A against i.

When we compare the convex hull in figure 3 to the ellipse in figure 2 in the introduction, we see that for $nit = 1000$ we get a relatively good estimate of the ellipse. The decrease in growth of the area in figure 4 suggests that we are indeed closing in on the true numerical range. To check if this is the case we will compare the computed area for $nit = 1000$ with the true one computed using the ellipse formula that we found in the introduction. It follows from the ellipse formula that the area is given by

$$Area = \pi * \sqrt{3.25} * \sqrt{0.25} = 2.8318\ldots.$$

On the other hand the area given by the algorithm is $2.8172\ldots$, so we are indeed closing in on the theoretical value of the area. As mentioned before the problem with this method is that it is possible for the script to keep creating points inside the true numerical range or even points that it has already generated before. A closer look at figure 3 reveals that there were indeed a lot of points generated in the interior of the numerical range. This gives rise to the question whether or not it is possible to generate these points more efficiently. The first thing that comes to mind is to discretize the complex unit n-sphere. This becomes complicated to code very fast, but for our example it should still be possible. Let us consider the following parametrization of the complex unit-sphere, which was also used to study this example in the introduction. Let $f = e^{i\alpha}\cos\theta$, $g = e^{i\beta}\sin\theta$, $\alpha, \beta \in \left[0, \frac{\pi}{2}\right]$, $\theta \in [0, 2\pi)$. Then we can create our unit vectors by summing over α, β and θ. The script below does exactly that.

```
function [area,conv]=generalnmr3(T,nit)
h=2*pi/nit;
f=0.5*pi/nit;
for j=1:nit
    alpha=f*j;
    for k=1:nit
        beta=f*k;
        for l=1:nit
            theta=h*l;
            x=[cos(theta)*exp(1i*alpha) ; sin(theta)*exp(1i*beta) ];
            n(j,k,l)=dot(T*x,x);
            q=nnz(n);
            if q>2
                z=reshape(n,numel(n),1);
                C=convhull(real(z),imag(z));
                area(q)=polyarea(real(z(C)),imag(z(C)));
                conv(q)=area(q)/area(q-1);
            end
        end
     end
end

z=reshape(n,q,1);
figure(1)
hold on
scatter(real(z),imag(z));
plot(real(z(C)),imag(z(C)));

figure(2)
plot(area);

end
```

This script produces the same plots as the previous one. The plots generated for $nit = 25$ by applying this script to our operator T are given below.

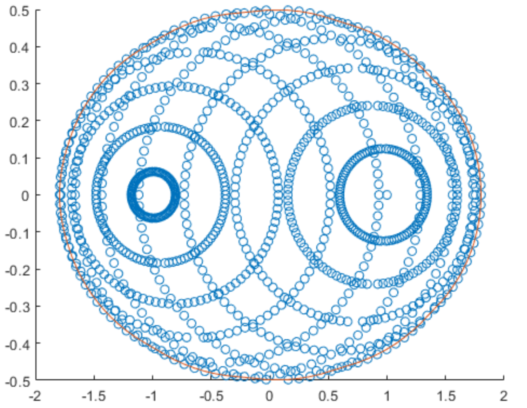

Figure 5: Convex hull of matrix A.

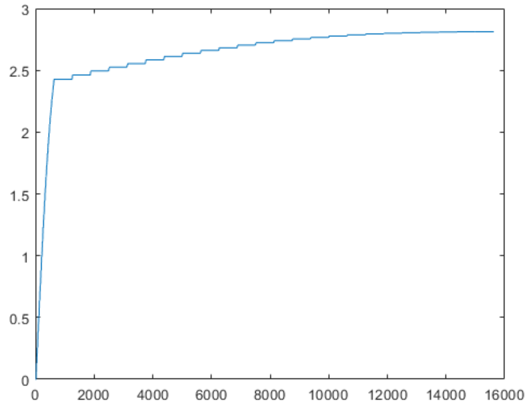

Figure 6: Area of convex hull of matrix A against i.

When we study figure 5, we see that the approximation of the numerical range of T generated by this algorithm is quite similar to the one generated by the other algorithm. The points in the interior of the convex hull look a lot more interesting though. We see that the points generated in $W(T)$ are families of circles. This is not completely surprising, in fact, if we go back to the derivation of the numerical range of our operator in the introduction, we see that we proved that the numerical range is a family of circles that depend on

θ. Whilst this dependence on θ is not something that we can immediately infer from the figure, the fact that it is a union of circles is as mentioned very clear. When we study figure 6, we see that the area remains constant at a lot of points in a repeating pattern. This is due to the fact that large parts of the circles are not part of the convex hull, so when we sum over these parts of the circles the area does not increase. An important note is that, whilst both scripts take *nit* as an input, the first creates *nit* elements in the numerical range and the second nit^3 elements. The second method looks like a more efficient method, but it is in fact significantly more computationally expensive to obtain a similar accuracy. In fact for $nit = 25$ we create $25^3 = 15625 >> 1000$ elements in $W(T)$ to obtain an area of $2.8121\ldots$. This is still less accurate then the method using 1000 random points.

3.2 Convergence of Iterative Methods

Another interesting application of the numerical range in numerical analysis is its appearance in the study of the convergence of iterative methods. In particular, the convergence of the steepest descent method for solving $Ax = b$ has a convergence closely related to the numerical range. This relation is given by the following theorem in Gustafson and Rao [9].

Theorem 31. *(Trigonometric convergence). In quadratic steepest descent, for any initial point x_0, there holds at every step k*

$$E_A(x_{k+1}) \leq (\sin^2 A)E_A(x_k).$$

Where the sin of an operator is determined using the definition of the cosine of an operator in the section on operator trigonometry.

$$\sin^2 A = 1 - \cos^2 A$$

Proof. The steepest descent method is an iterative method that can be used to minimize a function f. If we consider the quadratic case where

$$f(x) = \frac{\langle x, Ax\rangle}{2} - \langle x, b\rangle,$$

for a symmetric, positive definite matrix A with eigenvalues $0 < m = \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n = M$, where $m = m(T)$ and $M = w(T)$. Minimizing this is is equivalent to finding the solution of $Ax = b$. A general form of the steepest descent method is given by [9]

$$x_{k+1} = x_k - \alpha_k \nabla f(x_k)^T.$$

So with the f that we chose this becomes

$$x_{k+1} = x_k - \frac{||Ax_k - b||^2(Ax_k - b)}{\langle A(Ax_k - b)), Ax_k - b\rangle}.$$

If we now define the error as

$$E_A(x) = \frac{\langle (x - x^*), A(x - x^*) \rangle}{2} = f(x) + \frac{\langle x^*, Ax^* \rangle}{2}.$$

It then follows from the earlier introduced Kantorovich inequality that

$$E_A(x_{k+1}) \leq \left(1 - \frac{4\lambda_1 \lambda_n}{(\lambda_n + \lambda_1)^2}\right) E_A(x_k).$$

The proof of this inequality is given in the section on the quadratic form of the steepest descent method in the book on linear and nonlinear programming by Luenberger [5]. If now use that $\lambda_1 = m$ and $\lambda_n = M$, we get that

$$E_A(x_{k+1}) \leq (1 - \frac{4mM}{(m + M)^2}) E_A(x_k).$$

We can now apply theorem 29

$$= (1 - \cos^2(A)) E_A(x_k) = (\sin^2 A) E_A(x_k),$$

which completes the proof. □

The geometric interpretation of this theorem is that the angle $\phi(A)$ determines the convergence when solving $Ax = b$ using quadratic steepest descent. This is due to the fact that steepest descent method cannot converge faster than the maximum distance between x and Ax, which after normalization is represented by $\sin A$. In the book by Gustafson an Rao [9] it is also shown that the convergence of the conjugate gradient method is determined by $\sin(A^{1/2})$, but we will not prove this in this paper.

3.3 Discrete Stability

Another application of the numerical range is its relevance in the study of initial value problems. An initial value problem is a system of equations of the form

$$\begin{cases} \frac{d}{dt} u(t) = Au & t > 0, \\ u(0) = u_0. \end{cases}$$

Gustafson and Rao state that if A is an m-dissipative operator, then the solution $u(t)$ is generally given by $u(t) = e^{At} u_0$ for u_0 in the domain of A [9]. When solving a system like this we would like to approximate the continuous terms by discrete ones. A good example of this is the explicit Euler method, which is discussed in both [10] and [9]. To illustrate the relevance of the numerical range, we follow along with the example in [9] and apply the explicit Euler method to the heat equation

$$\begin{cases} \frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2}, \\ u(0) = f(x). \end{cases}$$

The discretization of the heat equation above using explicit Euler is given by [9]

$$\frac{\partial u}{\partial t} \cong \frac{u(x,t+\Delta t)-u(x,t)}{\Delta t},$$

$$\frac{\partial^2 u}{\partial x^2} \cong \frac{u(x+\Delta x,t)-2u(x,t)+u(x-\Delta x,t)}{(\Delta x)^2}.$$

Let us consider the coordinates (x_i, t_j) in space-time, where $x_i = i\Delta x$ and $t_j = j\Delta t$. If we now call the discrete solution $U_{i,j} = U(x_i, t_j)$, the Euler explicit scheme is given by [9]

$$U_{i,j+1} = U_{i,j} + \frac{\Delta t}{(\Delta x)^2}\left(U_{i+1,j} - 2U_{i,j} + U_{i-1,j}\right).$$

We can simplify this by representing it as a sequence of matrix iterations

$$U^{j+1} = C(\Delta t)U^j,$$

where

$$C(\Delta t) = I + (\Delta t)A_D$$

and A_D denotes the discretized version of $\frac{\partial^2 u}{\partial x^2}$ given above. We saw that $U(\delta t) = e^{A\delta t}$ was the semigroup that solved the original continuous problem, here $C(\Delta t)$ is the semigroup for the discrete problem. It is important to notice that $C(\Delta t)$ is equal to the first two terms of the Taylor expansion of $e^{A_D\delta t}$. Gustafson and Rao also show that it is a semigroup up to first order as follows [9]

$$C(\Delta t)C(\Delta t)u = (I + 2\Delta t A_D + (\Delta t)^2 A_D^2)u = C(2\Delta t)u + (\Delta t)^2 A_D^2 u.$$

For the context of this paper it is not that relevant to discuss semigroup generators, so we will use the result of Gustafson and Rao without justification. The infinitesimal generator of the semigroup for the heat equation is given by

$$Au = \frac{\partial^2 u}{\partial x^2} = s - \lim_{\Delta t \to 0} \frac{\left(e^{A\Delta t} - I\right)}{\Delta t} u,$$

for all $u \in D(A)$. We may now also write

$$A_D u = s - \lim_{\Delta t \to 0} \frac{(C(\Delta t) - I)}{\Delta t} u.$$

If everything has gone according to plan the following consistency condition should hold

$$||(A_D - A)u(t)|| \to \quad \text{as} \quad \Delta t \to 0,$$

for all solutions $u(t)$. Often it is not easy to prove that a discretization satisfies this condition. An important result concerning this problem is the Lax equivalence theorem [9]: the discrete solutions converge to the continuous solution if and only if the scheme is stable. When studying our example stability means that the successive iterates of the operators $C(\Delta t)$ are uniformly bounded [9],

$$||(C(\Delta t)^n|| \leq M,$$

for all small $0 < \Delta t \leq \tau$, on the interval $0 \leq n\Delta t \leq T$, $n = 1, 2, 3, \dots$. If this is not the case the discrete solution might expand uncontrollably when $\Delta t \to 0$.

This stability condition can be studied using the numerical range by considering the fact that the numerical radius is an equivalent norm to the operator norm. If we then consider the power-boundedness theorem we can look for M using the numerical radius of $C(\delta t)$.

3.3.1 Fluid Dynamics

Having studies the applications of the numerical range in discrete stability in the last section, we can now look at what it has to offer in Fluid dynamics. To do this we take a look at the Navier-Stokes equations for velocities u [9]

$$\begin{cases} \frac{du}{dt} = k(u)\Delta u - u \cdot \nabla u - \nabla p \quad t > 0, \\ u(0) = u_0. \end{cases}$$

To study the use of the numerical range in fluid dynamics we consider the one-dimensional model from Gustafson and Rao [9]

$$\begin{cases} u_t + vu_x - ku_{xx} = 0 \quad 0 < x < 1 \quad t > 0, \\ u(x,0) = f(x,0) \quad \text{given} \quad 0 < x < 1, \\ u(0,t) = u(1,t) = 0 \qquad t > 0. \end{cases}$$

For the purpose of illustration, Gustafson and Rao linearize this problem by assuming that v and k are positive constants. Just like in the last section we give a matrix representation of the explicit Euler discretization of this problem

$$\frac{U_{i,i+1} - U_{i,j}}{\Delta t} + v\frac{U_{i,j} - U_{i,j-1}}{\Delta x} - k\frac{U_{i-1,j} - 2U_{i,j} + U_{i+1,j}}{(\Delta x)^2} = 0,$$

which can be rewritten as

$$U_{i,j+1} = \left(\frac{k + v\Delta x}{(\Delta x)^2}\Delta t\right) U_{i-1,j} + \left(1 - \frac{2k + v\Delta x}{(\Delta x)^2}\Delta t\right) Ui,j + \left(\frac{k\Delta t}{(\Delta x)^2}\right) U_{i+1,j}.$$

We want to show that $U_{i,j+1}$ is a convex combination of $U_{i-1,j}$, $U_{i,j}$ and $U_{i+1,j}$. First of all we should notice that

$$\left(\frac{k + v\Delta x}{(\Delta x)^2}\Delta t\right) + \left(1 - \frac{2k + v\Delta x}{(\Delta x)^2}\Delta t\right) + \left(\frac{k\Delta t}{(\Delta x)^2}\right) = 1.$$

The second condition that needs to be satisfied is that all three the terms above are non-negative. This is clearly the case when

$$\frac{1}{2k + v\Delta x} \leq \frac{\Delta t}{(\Delta x)^2}.$$

We will accept this limitation on Δt, so that we can use the convexity whilst studying the stability. It follows from the convexity that $U_{i,j+1}$ is bounded uniformly independent of Δt and Δx. This is caused by the fact that the convex combination is bounded by the largest and the smallest component in the sum. These components are then again bounded by the largest and smallest component of the previous time step. Therefore the bound eventually only depends on the initial data. This allows us to take a similar approach to the last section and try to find a uniform bound to the operator $C(\Delta t)$, it can then eventually be proven that the convergence to the true solution for a sufficiently small grid size [9].

3.4 Pseudo Eigenvalues

Pseudo eigenvalues have been introduced to deal with the fact that eigenvalues are very sensitive to small perturbations. A lot of algorithms and results in numerical analysis depend strongly on the eigenvalues of operators. In matrix terms a perturbation to a matrix A could be expressed as adding a perturbation matrix B. Gustafson and Rao point out that the fact that, whilst we can not say much about the spectrum of the sum of two matrices, we did show that

$$W(A+B) \subset W(A) + W(B).$$

To study the the application of numerical range in the study of perturbed operators we define the following two notions.

Definition 14. *The augmented numerical range is given by*

$$W_\epsilon(A) = W(A) + \Delta_\epsilon,$$

where Δ_ϵ denotes the closed disk of radius ϵ.

Definition 15. *The pseudo-spectrum of A is given by*

$$\sigma_\epsilon(A) = \{\lambda \in \mathbb{C} : \lambda \in \sigma(A+E) \text{ for some } E \text{ with } ||E|| \leq \epsilon\}.$$

We will show later in this section that the region $\sigma_\epsilon(A)$ is in between the augmented numerical range and the spectrum of the operator. Gustafson and Rao then state and prove the following theorem [9].

Theorem 32. *For a given $n \times n$ matrix A and $\epsilon \geq 0$ the following are equivalent*

(i) *λ is an ϵ-pseudo-eigenvalue of A,*

(ii) *$||(\lambda I - A)x|| \leq \epsilon$ for some $||x|| = 1$,*

(iii) *$||(\lambda I - A)^{-1}||^{-1} \leq \epsilon$,*

(iv) *the smallest singular value of $\lambda I - A$ is smaller or equal than ϵ.*

Proof. Let us assume that λ is an pseudo eigenvalue of A. The definition gives us that there exist a perturbation matrix E with $||E|| \leq \epsilon$, such that $\lambda \in \sigma(A+E)$. This means that there must exist a unit vector x, such that $(A+E)x = \lambda x$. We can now conclude that using the definition of the operator norm that

$$||(\lambda I - A)x)|| = ||Ax - \lambda x|| = ||Ex|| \leq \epsilon,$$

which proves the equivalence to the second claim. To prove the next claim we use the result from functional analysis that $||(\lambda I - A)^{-1}||^{-1} \leq ||(\lambda I - A)||$. Clearly it follows from the last claim that $||(\lambda I - A)|| \leq \epsilon$. To see the equivalence to the next claim, notice that the eigenvalues of $(\lambda I - A)^{-1}$ are $\frac{1}{\mu}$, for $\mu \in \sigma((\lambda I - A)^{-1})$. Combining this with the fact the operator norm computes the largest singular value shows the equivalence (this was shown in, for example, the lecture notes from Dahleh [13]). We now consider the singular value decomposition, to show that this claim is again equivalent to the first.

$$\lambda I - A = U \sum V^* = \sum_{j=1}^{n} \sigma_j u_j v_j^*$$

We know that σ_n is the smallest singular value of $\lambda I - A$. Therefore we can take $E = \sigma_n u_n v_n^*$ and it will satisfy $||E|| \leq \epsilon$. We can now consider the matrix

$$\lambda I - (A + E) = \sum_{j=1}^{n-1} \sigma_j u_j v_j^*.$$

This matrix is singular, because of the construction of the singular value decomposition. Therefore λ has to be an eigenvalue of $A + E$. □

The next theorem shows that the pseudo-spectrum is a lot less sensitive to perturbation then the normal spectrum. It also proves the inclusion in the augmented numerical range.

Theorem 33. *(Pseudo-spectra stability). Let D be a perturbation of norm δ. Then*

$$\sigma_{\epsilon-\delta}(A + D) \subset \sigma_\epsilon(A) \subset \sigma_{\epsilon+\delta}(A + D),$$
$$\sigma_\epsilon(A) \subset W_\epsilon(A).$$

Proof. Let λ an pseudo eigenvalue of A. We then know that $\lambda \in \sigma(A+E)$, with $||E|| \leq \epsilon$. So λ is also an eigenvalue of $A + D + (E - D) = A + E$. We know that $||E - D|| \leq ||E|| + ||D|| = \epsilon + \delta$, which completes the proof of the first statement.

For the second statement we use that, for some unit vector x we have that

$$(A + E)x = \lambda x.$$

We can now use this to show that the distance from the numerical range from λ is less or equal to epsilon.

$$|\langle Ax, x\rangle - \lambda| = |\langle(\lambda - E)x, x\rangle| = || - Ex|| = ||Ex|| \leq \epsilon$$

This completes the proof. □

3.5 Quantum Computing

The final application that we will discus is that of quantum error correction. In the paper by Chi-Kwong Li and Yiu-Tung Poon [4] it is explained that for a noisy quantum channel, a quantum error correcting code exists if and only if the joint higher rank numerical range associated with the error operators of the channel is non-empty.

We will not discuss the content in too much depth, because it is a little too involved for our purposes. The main idea behind the paper is that the existence of an error correction depends on whether or not a generalization of the numerical range is non-empty. This generalization is given by the following definition [4].

Definition 16. *The joint $k-$numerical range of an $m-$tuple of matrices $A = (A_1, \dots, A_m)$ is given by*

$$\wedge_k(A) = \{(a_1, \dots, a_m) \in \mathbb{C}^{\gtrdot} : \exists P \in P_k \textit{ such that } PA_jP = a_jP \textit{ for } j = 1, \dots, m\}.$$

4 Conclusion

Throughout this paper we have seen numerous applications of the numerical range in different fields of mathematics. The application of the theory to examples showed us how the numerical range can be used when studying, for example, the spectrum or norm of an operator. The study of the approximation of $W(T)$ calls for further research, in this research it could be studied whether or not it is possible to approximate the convex hull in a more efficient manner. Some of the assumptions in the section on discrete stability are rather restrictive, so it would be interesting to study how much of an impact the numerical range can have in a less restrictive context. The more recent application of the numerical range in quantum computing shows us that there might still be a lot more to discover using the numerical range.

References

[1] A. Sterk, H. de Snoo. *Functional Analysis Lecture Notes*. Groningen, 2018.

[2] B. Sz-Nagy, C. Foias. *On certain classes of power-bounded operators in Hilbert space*. The American Mathematical Monthly, 1993.

[3] Béla Nagy. *On contractions in Hilbert space*. Acta Szeged, 1953.

[4] Chi-Kwong Li, Yiu-Tung Poon . *Quantum error correction and generalized numerical ranges.* 2008.

[5] David G. Luenberger, Yinyu Ye. *Linear and Nonlinear Programming.* Springer, 2008.

[6] David W. Kribs, Aron Pasieka, Martin Laforest, Ryan Colm, Marcus P. da Silva. *Research problems on numerical ranges in quantum computing.* Taylor & Francis, 2009.

[7] J.G. Stampfli. *Extreme points of the numerical range of a hyponormal operator*. 1966.

[8] Karl E. Gustafson. *Linear Algebra and its Applications, Pages 437-454*. Elsevier, 1994.

[9] Karl E. Gustafson, Duggirala K.M. Rao. *Numerical Range.* Springer-Verlag, New York, seventh edition, 1997.

[10] M. Quarteroni, R. Sacco, F. Saleri. *Numerical Mathematics.* Springer-Verlag, New York, 2000.

[11] M.G. Krein. *Angular Localization of the Spectrum of a Multiplicative Intregral in a Hilbert Space.* 1969.

[12] M.H. Stone. *Linear Transformations in Hilbert Space*. The American Mathematical Monthly, 1993.

[13] Mohammed Dahleh, Munther A. Dahleh, George Verghese. *Lectures on Dynamic Systems and Control.* Massachuasetts Intitute of Technology.

[14] T. Kato. *Some Mapping Theorems for the Numerical Range.* Proceedings of the Japan Academy, 1965.