2011-07-13

Crouzeix's Conjecture and the GMRES Algorithm

Sarah McBride Luo
Brigham Young University - Provo

Crouzeix's Conjecture and the GMRES Algorithm

Sarah M. Luo

A thesis submitted to the faculty of
Brigham Young University
in partial fulfillment of the requirements for the degree of

Master of Science

Jeffrey Humpherys, Chair
Wayne Barrett
Shue-Sum Chow
Christopher P. Grant

Department of Mathematics

Brigham Young University

August 2011

Abstract

Crouzeix's Conjecture and the GMRES Algorithm

Sarah M. Luo

Department of Mathematics

Master of Science

This thesis explores the connection between Crouzeix's conjecture and the convergence of the GMRES algorithm. GMRES is a popular iterative method for solving linear systems and is one of the many *Krylov methods*. Despite its popularity, the convergence of GMRES is not completely understood. While the spectrum can in some cases be a good indicator of convergence, it has been shown that in general, the spectrum does not provide sufficient information to fully explain the behavior of GMRES iterations. Other sets associated with a matrix that can also help predict convergence are the pseudospectrum and the numerical range. This work focuses on convergence bounds obtained by considering the latter. In particular, it focuses on the application of Crouzeix's conjecture, which relates the norm of a matrix polynomial to the size of that polynomial over the numerical range, to describing GMRES convergence.

Keywords: GMRES, Michel Crouzeix, Faber Polynomials, Complex Approximation, Krylov Subspace, Convergence, Iterative Methods, Linear Systems

Acknowledgments

I would first like to thank my adviser for all of the time and help he gave to me not only for this thesis, but throughout all my years in the BYU Math department. His encouragement and support were more than I could have asked for.

I also want to express thanks to my committee: Wayne Barrett, Shue-Sum Chow, and Chris Grant for their help, suggestions, and taking the time to be involved in this work.

My gratitude also goes to the department for accepting me into the Master's program and supporting me throughout my time here. I particularly want to thank Lonette Stoddard for her guidance and her efforts in my behalf, as well as all she does for the department as a whole.

Most importantly, I thank my parents for raising me in the right way and providing me with the many wonderful opportunities I have had in my life. Their tireless and unending support have been invaluable to me.

I also want to thank my husband, Yi, for all his encouragement and the many sacrifices he makes for me. I am truly blessed to have him in my life.

Contents

List of Tables

List of Figures

Chapter 1. Introduction

A popular strategy for solving systems of linear equations is to implement an iterative method. While there are many such methods, one that has received particular attention is the GMRES algorithm. GMRES stands for "generalized minimal residuals" and is favorable due to its ability to handle non-Hermitian and/or indefinite linear systems. Predecessors to GMRES, such as the conjugate gradient method or MINRES, are restricted to Hermitian problems, or problems where the Hermitian part is positive definite. Regardless of the type of problem a particular method is suited for, in all cases a prevailing question is how long will the method take to converge? A particular method may give excellent approximations, but if it takes an inordinate amount of time to produce that approximation, then for all intents and purposes it is useless to us. What is needed is a way to predict how many iterations will be necessary in order for the method to produce the desired result. While occasionally it is useful to consider the right hand side b, usually it is the properties of the matrix A that determine the behavior of the iterative method. An appreciable amount of research has been done in this area, leading to convergence bounds in terms of the spectrum, the psuedospectrum, and the numerical range of A. In this work, we focus on the latter.

Given a linear system $Ax = b$, and an initial guess x_0, GMRES works by producing approximations which lie in the affine Krylov subspace given by

$$x_0 + K_n(A, r_0) = x_0 + \text{Span}\{r_0, Ar_0, A^2r_0, ..., A^{n-1}r_0\},$$

where $r_0 = b - Ax_0$ is the initial residual. As will be shown below, the k^{th} residual satisfies the equation

$$||r_k|| = ||p(A)r_0||,$$

where $p(z)$ is a k^{th} degree polynomial with complex coefficients, satisfying $p(0) = 1$ and $||\cdot||$ denotes the 2-norm. Note that this implies

$$\frac{||r_k||}{||r_0||} \leq ||p(A)||.$$

Thus the task of approximating the rate of convergence can be undertaken by considering the quantity $||p(A)||$. As will be shown in Chapter 4, if we assume that A is diagonalizable, then we have the classical result

$$||p(A)|| \leq \kappa(V) \sup_{\lambda \in \sigma(A)} |p(\lambda)|, \tag{1.0.1}$$

where $\sigma(A)$ denotes the spectrum of A, V is a matrix of eigenvectors of A, and $\kappa(V)$ is the condition number of V. This relation allows us to restate our matrix approximation problem as a scalar approximation problem. However, in the language of Trefethen [47], there is a constant "gap" between these two problems, namely $\kappa(V)$. In the case of normal matrices, this bound is sharp [21] and in the case that A is not normal, but V is well-conditioned, then (1.0.1) can still provide good bounds. However, many matrices that arise in practice do not have well-conditioned eigenvector matrices and, as is shown in [24], the convergence of GMRES in general depends on more than just the spectrum. Getting around these issues is where the Crouzeix conjecture comes into play.

Crouzeix's conjecture relates the size of a polynomial of a matrix to the size of that polynomial over the numerical range of the matrix. The numerical range of a matrix is a convex subset of the complex plane, consisting of all Rayleigh quotients

$$W(A) = \left\{ \frac{x^*Ax}{x^*x} \mid x \neq 0, x \in \mathbb{C}^n \right\}.$$

Crouzeix's conjecture can then be concisely stated as follows

$$||p(A)|| \leq 2 \sup_{z \in W(A)} |p(z)|. \tag{1.0.2}$$

While this conjecture itself is somewhat new, similar inequalities have a history dating back many decades. The first was perhaps von-Neumann's inequality

$$||p(A)|| \leq \sup_{z \in D_{||A||}} |p(z)|,$$

where $D_{||A||}$ is the disk centered at 0 of radius $||A||$. Later came the following result of Badea [7], which is based on a result established by Ando in 1973 [1]:

$$||p(A)|| \leq 2 \sup_{z \in D_{w(A)}} |p(z)|,$$

where $D_{w(A)}$ is a disk centered at 0 of with radius equal to the numerical radius (see Section 2.4). As for results more directly related to the Crouzeix conjecture, there is the result of Delyon and Delyon [12] which states that for any bounded, open, convex set $K \subset \mathbb{C}$, there exists a finite constant $C(K)$, depending only on the set K such that

$$||p(A)|| \leq C(K) \sup_{z \in K} |p(z)|. \tag{1.0.3}$$

Letting $C(K)$ denote the best constant such that this equation holds, Crouzeix and his collaborators have derived approximations to $C(K)$ for particular sets K, including sectors, disks, and parabolic domains; see [2] and the references therein. Of particular interest is the result that for $W(A) \subset K$, $C(K) = 2$ for the case of a 2×2 matrix [2] and that (1.0.3) holds with 11.08 in place of $C(K)$ for general matrices A [8].

The Crouzeix conjecture can be of help in the analysis of GMRES in the following way:

simply replace the constant $\kappa(V)$ in (1.0.1) by 2, and the set $\sigma(A)$ by $W(A)$ (which, as we will see below, contains the spectrum of A). The advantages of this are three-fold. First, we have rid ourselves of the quantity $\kappa(V)$ and by so doing, have eliminated any need for our matrix to be near normal. Second, rather than restricting ourselves to the spectrum, which we know is insufficient in general, we are now considering the numerical range, which has many nice properties including convexity and compactness[1]. Furthermore, the numerical range has applications in many areas including operator theory, dilation theory, C^*-algebras, and factorization of matrix polynomials (to name a few), and thus is a promising set to consider when seeking information about a matrix. Lastly, to obtain a good bound, all we need to do now is answer the question of how big can $|p(z)|$ be over the numerical range. In other words, we now have a problem which falls into the category of approximation theory, a field where substantial work has been done. To answer this question, there are a number of paths we could consider, including Pick-Nevanlinna interpolation, interpolation in Fejér points, least squares approximation and orthogonal polynomials, and estimates via Faber polynomials. Here, we investigate some recent results obtained by considering the Faber polynomials. To our knowledge, these results are the best given so far in the context of GMRES analysis.

Chapter 2. Numerical Range

In this chapter, we define the numerical range of a matrix operator and give some of its key properties used in this thesis. See the references for a more comprehensive treatment. Let M_n denote the set of $n \times n$ matrices over the complex numbers. The numerical range of a matrix $A \in M_n$ is defined as follows:

[1]It should be noted that compactness does not hold in general when A is an infinite dimensional operator. However, in this work, we only consider finite dimensional $n \times n$ matrices, and so we will always assume $W(A)$ is compact.

Definition 2.0.1. The *numerical range* of $A \in M_n$, is the subset $W(A) \subset \mathbb{C}$, given by

$$W(A) = \{\langle Ax, x\rangle \mid x \in \mathbb{C}^n, ||x|| = 1\}, \tag{2.0.1}$$

where $||\cdot||$ denotes the 2-norm.

Note that $W(A)$ is the continuous image of a compact set, and is thus itself a compact set in $\mathbb{C}$. As we will show, the numerical range of a linear operator is a convex set. This is a consequence of the *Toeplitz-Hausdorff Theorem*. We first review some basic properties of the numerical range.

2.1 Basic Properties

We start with some results on the invariance of $W(A)$.

Proposition 2.1.1. *Let $A \in M_n$. Then the following properties hold.*

(i) For any $\alpha, \beta \in \mathbb{C}$, we have that $W(\alpha A + \beta I) = \alpha W(A) + \beta$.

(ii) $W(U^*AU) = W(A)$ for any unitary $U \in M_n$.

(iii) If $k \in \{1, ..., n-1\}$ and $X \in \mathbb{C}^{n\times k}$ satisfies $X^*X = I_k$, where I_k denotes the $k \times k$ identity matrix, then $W(X^*AX) \subset W(A)$.

Proof. To show (i), we calculate

$$\begin{aligned} W(\alpha A + \beta I) &= \{x^*(\alpha A + \beta I)x \mid x \in \mathbb{C}^n, ||x|| = 1\} \\ &= \{\alpha x^*Ax + \beta x^*x \mid x \in \mathbb{C}^n, ||x|| = 1\} \\ &= \{\alpha x^*Ax + \beta \mid x \in \mathbb{C}^n, ||x|| = 1\} \\ &= \alpha\{x^*Ax \mid x \in \mathbb{C}^n, ||x|| = 1\} + \beta \\ &= \alpha W(A) + \beta. \end{aligned}$$

To show (ii), let $\lambda \in W(U^*AU)$. Then there exists a unit vector $x \in \mathbb{C}^n$ such that $\langle U^*AUx, x\rangle = \lambda$. Since U is self-adjoint, we can write $\langle AUx, Ux\rangle = \lambda$. Now let $y = Ux$. Since multiplying by a unitary matrix preserves norm, that is, $||y||^2 = y^*y = x^*U^*Ux = x^*x = ||x||^2$, we have that $y \in \mathbb{C}^n$ is also a unit vector. Thus $\langle Ay, y\rangle = \lambda$ so $\lambda \in W(A)$.
To show the reverse inclusion, note that $W(A) = W(UU^*AUU^*) \subset W(U^*AU)$ by what was just shown. Thus $W(A) = W(U^*AU)$.
For (iii), let $\lambda \in W(X^*AX)$. Then there exists a unit vector $y \in \mathbb{C}^k$ such that $\langle X^*AXy, y\rangle = \lambda$. Note that $||Xy||^2 = y^*X^*Xy = y^*y = 1$. Thus setting $v = Xy$ yields $v^*Av = \langle Av, v\rangle = \lambda$. Hence $W(X^*AX) \subset W(A)$. □

The following example illustrates point (i) of Proposition 2.1.1.

Example 2.1.2. Let $A \in M_3$ be defined as

$$A = \begin{bmatrix} 1 & 3 & 6 \\ 2 & 8 & 9 \\ -4 & 8 & -1 \end{bmatrix}.$$

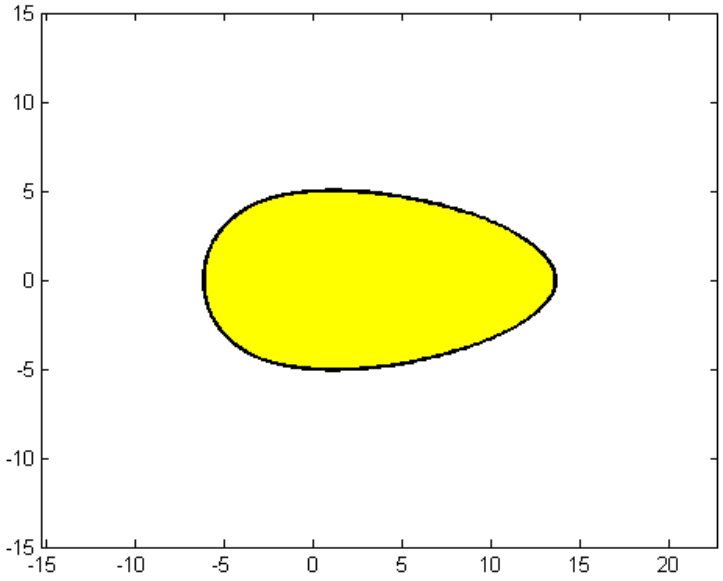
The numerical range of this matrix is in Figure 2.1(a). Now let $a = .5$ and $b = 10 + 8i$. The numerical range of $aA + bI$ is in Figure 2.1(b) along with the numerical range of A for comparison. The smaller egg-shaped region is the numerical range of $aA + bI$. Notice that $W(aA+bI)$ is indeed the same as $aW(A)+b$ and that the numerical ranges of A and $aA+bI$ have the same basic shape and geometric properties.

Next we show two important properties of $W(A)$.

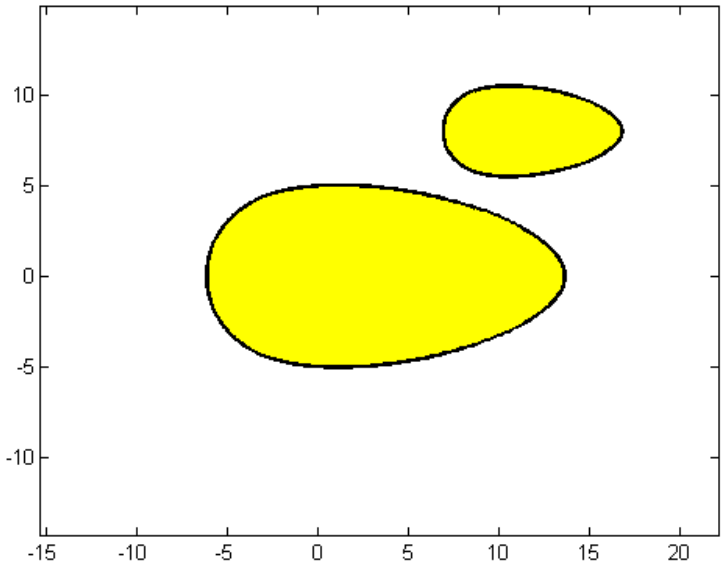
Proposition 2.1.3. *Let $A, B \in M_n$. Then*

(i) $W(A^*) = \{\overline{\lambda} \mid \lambda \in W(A)\} = \overline{W(A)}$.

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



(a) The numerical range of A as in Example 2.1.2



(b) The numerical ranges of A and $aA + bI$ as in Example 2.1.2

Figure 2.1: Example of Scalar and Translational Invariance

Proof. For (i), we have the following:

$$W(A^*) = \{\langle A^*x, x\rangle \mid ||x|| = 1\} = \{\langle x, Ax\rangle \mid ||x|| = 1\}$$
$$= \{\overline{\langle Ax, x\rangle} \mid ||x|| = 1\} = \overline{W(A)}$$

For (ii), let $\gamma \in W(B + A)$ with unit vector x satisfying $x^*(B + A)x = \gamma$. Let $\gamma_B = x^*Bx$ and $\gamma_A = x^*Ax$. Then $\gamma_B \in W(B)$ and $\gamma_A \in W(A)$ and $\gamma = x^*(B + A)x = x^*Bx + x^*Ax = \gamma_B + \gamma_A$. So γ is the sum of an element in $W(B)$ and an element in $W(A)$, thus $\gamma \in W(B) + W(A)$. □

The next two results will be very useful in showing how to sketch the numerical range as well as proving its convexity, but first we need the following lemma:

Lemma 2.1.4. *Let $A \in M_n$. If $\langle Ax, x\rangle = 0$ for all $x \in \mathbb{C}^n$, then $A = 0$.*

Proof. First suppose that A is Hermitian. Then for any $x, y \in \mathbb{C}^n$, we have that $\langle x, Ay\rangle = \langle Ax, y\rangle$. By the hypothesis, we also have that $\langle A(x + y), x + y\rangle = 0$. Therefore,

$$0 = \langle Ax, x\rangle + \langle Ay, x\rangle + \langle Ax, y\rangle + \langle Ay, y\rangle = \langle Ax, y\rangle + \langle y, Ax\rangle .$$

Letting $y = Ax$ then yields $0 = 2||Ax||^2$, which implies $Ax = 0$ for all $x \in \mathbb{C}^n$. Thus $A = 0$. Now let $A \in M_n$ be arbitrary. If we let

$$H = \frac{A + A^*}{2} \quad \text{and} \quad K = \frac{A - A^*}{2i},$$

then $A = H + iK$ with H and K both Hermitian (we call H the *Hermitian part of A* and iK the *skew-Hermitian part of A*). Thus, if $\langle Ax, x\rangle = x^*Ax = 0$ for all $x \in \mathbb{C}^n$, then since

$(x^*Ax)^* = x^*A^*x = 0$ also, we have

$$x^*Ax = x^*Hx + ix^*Kx = 0, \quad \text{and}$$
$$x^*A^*x = x^*Hx - ix^*Kx = 0,$$

for all $x \in \mathbb{C}^n$. Adding these two equations gives us that $2x^*Hx = 0$ for all $x \in \mathbb{C}^n$ and so by the result above, $H = 0$. Similarly, $K = 0$ and so $A = 0$. □

Proposition 2.1.5. *The numerical range of a Hermitian matrix $A \in \mathbb{C}^n$ is an interval $[\lambda_1, \lambda_n] \subset \mathbb{R}$ where λ_1 is the smallest eigenvalue of A and λ_n is the largest eigenvalue of A. Moreover, the set $L_A(\lambda_n) = \{x \in \mathbb{C}^n : ||x|| = 1, x^*Ax = \lambda_n\}$ is the set of all unit eigenvectors of A corresponding to λ_n and similarly for λ_1. We also have that if $W(A) \subset \mathbb{R}$, then A is Hermitian.*

Proof. Let $A \in M_n$ be Hermitian. Then there exists a set of n orthonormal eigenvectors of A, denoted $\{x_1, ..., x_n\}$, with corresponding eigenvalues $\lambda_1, ..., \lambda_n$, which are arranged so that $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$. Let $x = c_1x_1 + \cdots + c_nx_n$ be a unit vector in $\mathbb{C}^n$ (so $|c_1|^2 + \cdots + |c_n|^2 = 1$). Using the fact that eigenvalues of A are all real, we have that

$$\begin{aligned} x^*Ax &= \lambda_1|c_1|^2 + \lambda_2|c_2|^2 + \cdots + \lambda_n|c_n|^2 \\ &\leq \lambda_n(|c_1|^2 + |c_2|^2 + \cdots + |c_n|^2) = \lambda_n. \end{aligned}$$

The first line of this equation implies that $x^*Ax \in \mathbb{R}$ for all unit vectors $x \in \mathbb{C}^n$ and hence $W(A) \subset \mathbb{R}$. Furthermore, it also shows that for all $z \in W(A)$, $z \leq \lambda_n$. Similarly, we can show that $x^*Ax \geq \lambda_1$ for all unit vectors $x \in \mathbb{C}^n$, and thus $W(A) \subset [\lambda_1, \lambda_n]$. We still need to show that $W(A) = [\lambda_1, \lambda_n]$, i.e. for all $c \in [\lambda_1, \lambda_n]$, there exists a unit vector $x \in \mathbb{C}^n$ such that $x^*Ax = c$. To do this let $x_s = \sqrt{s}x_1 + \sqrt{1-s}x_n$ for $0 \leq s \leq 1$. Then $||x_s|| = 1$ and $x_s^*Ax_s = s\lambda_1 + (1-s)\lambda_n$. So given any $c \in [\lambda_1, \lambda_n]$, we can find a unit vector x_s such that $x_s^*Ax_s = c$ by choosing an appropriate s. Thus $W(A) = [\lambda_1, \lambda_n]$.

For the second assertion, we claim that $x^*Ax = \lambda_n$ if and only if x is a unit eigenvector of A corresponding to λ_n. The reverse implication is clear. For the forward direction, we prove the contrapositive. Suppose x is not an eigenvector of A corresponding to λ_n. Then x cannot be a linear combination of eigenvectors corresponding to λ_n either (for such a vector is, in fact, an eigenvector corresponding to λ_n), so in the representation $x = c_1x_1 + \cdots c_nx_n$, we must have that $c_j \neq 0$ for some j where $\lambda_j \neq \lambda_n$. Since λ_n is the maximum eigenvalue, this means that $\lambda_j < \lambda_n$ so in this case, the inequality above is strict, i.e. $x^*Ax < \lambda_n$. This proves the second assertion for λ_n. The proof for λ_1 is similar.

Finally, to show the last statement, let $A \in M_n$, not necessarily Hermitian, and suppose $x^*Ax \in \mathbb{R}$ for all unit vectors $x \in \mathbb{C}^n$. Then $x^*Ax = \overline{x^*Ax} = x^*A^*x$ for all unit vectors $x \in \mathbb{C}^n$. Rearranging the terms we get

$$x^*Ax - x^*A^*x = 0 \Longrightarrow x^*(A - A^*)x = 0 \quad \text{for all unit vectors } x \in \mathbb{C}^n.$$

By Lemma 2.1.4, $A - A^* = 0$ and hence $A = A^*$. □

Figure 2.2 is a plot of the numerical range for the Hermitian matrix $H = A^*A$, where A is as in Example 2.1.2. The stars indicate the location of the eigenvalues of H.

Proposition 2.1.6. *For all $A \in M_n$, let $H(A) = (A + A^*)/2$ and $iK(A) = (A - A^*)/2$ denote the Hermitian and skew-Hermitian parts of A, respectively. Then*

$$Re\ (W(A)) = W(H(A)) \quad \textit{and} \quad Im\ (W(A)) = W(K(A)),$$

where

$$Re\ (W(A)) = \{Re\ z \mid z \in W(A)\} \quad \textit{and} \quad Im\ (W(A)) = \{Im\ z \mid z \in W(A)\}.$$

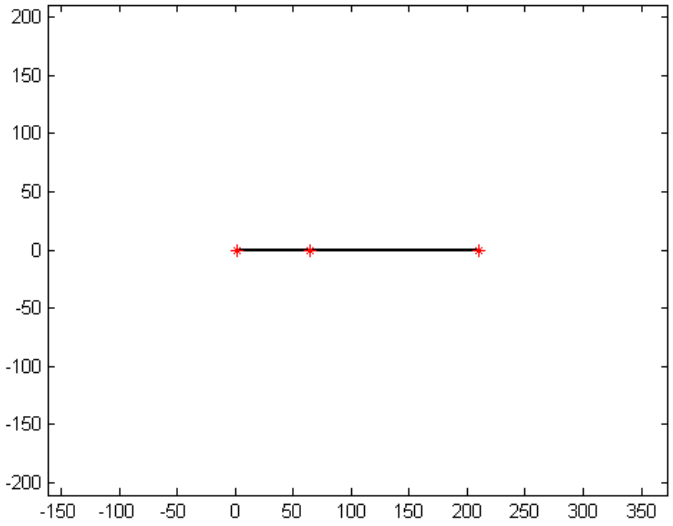


Figure 2.2: The numerical range of a Hermitian matrix

Proof. For all unit vectors $x \in \mathbb{C}^n$, we have that

$$x^*H(A)x = x^*\frac{1}{2}(A + A^*)x = \frac{1}{2}(x^*Ax + x^*A^*x)$$
$$\frac{1}{2}(x^*Ax + (x^*Ax)^*) = \frac{1}{2}(x^*Ax + \overline{x^*Ax}) = \text{Re } (x^*Ax).$$

So every point of $W(H(A))$ is of the form Re z for some $z \in W(A)$ and conversely. Similarly, for $K(A)$, we have that for all unit vectors $x \in \mathbb{C}^n$,

$$x^*K(A)x = x^*\frac{1}{2i}(A - A^*)x = \frac{1}{2i}(x^*Ax - x^*A^*x)$$
$$\frac{1}{2i}(x^*Ax - (x^*Ax)^*) = \frac{1}{2i}(x^*Ax - \overline{x^*Ax})$$
$$= \frac{1}{2i}(2i \text{ Im } (x^*Ax)) = \text{Im } (x^*Ax).$$

□

2.2 Convexity

One of the most significant properties of the numerical range is the fact that for any $A \in M_n$, the numerical range of A is convex. This fact was proved by Toeplitz and Hausdorff. Toeplitz showed that the boundary of the numerical range is a convex curve and later, Hausdorff showed that the numerical range is itself convex (see [44] and [28]). Thus this theorem has been named the *Toeplitz-Hausdorff Theorem*. There are various different proofs of this theorem. We present two of the more common ones below.

Theorem 2.2.1. *(Toeplitz-Hausdorff) Let $A \in M_n$. Then $W(A) \subset \mathbb{C}$ is convex.*

For the first proof, we need the preliminary result stating that for a 2×2 matrix, the numerical range is an elliptical disk whose foci are the eigenvalues of the matrix. There are several different ways of proving this fact ([26] contains two different proofs). We present here the proof provided in [32].

2.2.1 The Numerical Range of a 2×2 Matrix.

We will need the following lemma:

Lemma 2.2.2. *Given any $A \in M_2$, there exists a unitary $U \in M_2$ such that the two main diagonal entries of U^*AU are equal.*

Proof. Without loss of generality, we can suppose that tr $A = 0$. To see why, simply replace A with $A - \frac{1}{2}(\text{tr } A)I = A - \alpha I$. Suppose there exists a unitary matrix $U \in M_2$ such that the two main diagonal entries of $U^*(A - \alpha I)U$ are equal. Then if the (1,1) entry and the (2,2) entry of U^*AU are a'_{11} and a'_{22}, respectively, we would have that $a'_{11} - \alpha = a'_{22} - \alpha$ and so $a'_{11} = a'_{22}$. Thus we can suppose that tr $A = 0$, and our task is reduced to finding a unitary matrix $U \in M_2$ such that the two main diagonal entries of U^*AU are zero.

In order to do this, it suffices to show that there exists a nonzero $w \in \mathbb{C}^2$ such that $w^*Aw = 0$. This is because if we normalize w and set it as the first column of a unitary matrix W, we

will have

$$W^*AW = \begin{bmatrix} 0 & \times \\ \times & \times \end{bmatrix},$$

and since tr $(W^*AW) =$ tr $A = 0$, it must follow that the (2,2) entry is zero also.

To construct the vector w, first note that since tr $A = 0$, it is easily verified that the eigenvalues of A are $\pm\lambda$, for some $\lambda \in \mathbb{C}$. Let x and y be the normalized eigenvectors for $-\lambda$ and λ, respectively. If $\lambda = 0$, note that we can simply take $w = x$. Otherwise, let $w = e^{i\theta}x + y$. Since x and y are independent, w is nonzero for all $\theta \in \mathbb{R}$, and

$$\begin{aligned} w^*Aw &= (e^{-i\theta}x^* + y^*)A(e^{i\theta}x + y) = (e^{-i\theta}x^* + y^*)(-e^{i\theta}\lambda x + \lambda y) \\ &= \lambda(e^{-i\theta}x^*y - e^{i\theta}y^*x) = 2i\lambda \text{Im } (e^{-i\theta}x^*y). \end{aligned}$$

The result then follows by picking θ so that $e^{-i\theta}x^*y$ is real. □

Continuing along the same vein, let $A \in M_2$ and set $\alpha = (-1/2)$tr A. By Proposition 2.1.1, it suffices to consider $W(A + \alpha I)$. Further, tr $(A + \alpha I) = 0$ and by the preceding lemma, we can suppose that the two main diagonal entries are both zero. At this point, we have shown that we only need to consider matrices of the form $\begin{bmatrix} 0 & c \\ d & 0 \end{bmatrix}$, where $c, d \in \mathbb{C}$. However, we can simplify this further by noting that

$$\begin{bmatrix} 1 & 0 \\ 0 & e^{-i\theta} \end{bmatrix} \begin{bmatrix} 0 & c \\ d & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & e^{i\theta} \end{bmatrix} = \begin{bmatrix} 0 & ce^{i\theta} \\ de^{-i\theta} & 0 \end{bmatrix}.$$

Now, if $c = |c|e^{i\theta_1}$ and $d = |d|e^{i\theta_2}$, let $\theta = (1/2)(\theta_2 - \theta_1)$. Then the above matrix product equals

$$e^{i\phi} \begin{bmatrix} 0 & |c| \\ |d| & 0 \end{bmatrix}, \quad \phi = \frac{1}{2}(\theta_1 + \theta_2).$$

So by unitary and scalar invariance we only need to consider matrices of the form

$$\begin{bmatrix} 0 & a \\ b & 0 \end{bmatrix}, \quad a, b \geq 0. \tag{2.2.1}$$

We are now ready to prove

Lemma 2.2.3. *Let $A \in M_2$. Then $W(A)$ is an elliptical disk whose foci are the eigenvalues of A.*

Proof. By the above results, we can assume that A is of the form (2.2.1). Without loss of generality, suppose $a \geq b \geq 0$. Let $z \in \mathbb{C}^2$ be an arbitrary unit vector. The goal is to show that all numbers of the form z^*Az form an elliptical disk with the desired properties. Note that $(e^{i\theta}z)^*A(e^{i\theta}z) = z^*Az$ for all $\theta \in \mathbb{R}$ and so given any unit vector $z \in \mathbb{C}^2$, we can suppose that the first component of z is real and nonnegative. Since z is a unit vector, this means that z, with the first component real and nonnegative, has the form $z = (t, e^{i\theta}(1-t^2)^{1/2})^T$, where $t \in [0,1]$ and $\theta \in [0, 2\pi]$. Therefore,

$$\begin{aligned} z^*Az &= (t, e^{-i\theta}\sqrt{1-t^2}) \begin{bmatrix} 0 & a \\ b & 0 \end{bmatrix} \begin{pmatrix} t \\ e^{i\theta}\sqrt{1-t^2} \end{pmatrix} = (t, e^{-i\theta}\sqrt{1-t^2}) \begin{pmatrix} ae^{i\theta}\sqrt{1-t^2} \\ bt \end{pmatrix} \\ &= tae^{i\theta}\sqrt{1-t^2} + tbe^{-i\theta}\sqrt{1-t^2} = t\sqrt{1-t^2}((a+b)\cos(\theta) + i(a-b)\sin(\theta)). \end{aligned}$$

Letting θ vary from 0 to 2π, the point $(a+b)\cos(\theta) + i(a-b)\sin(\theta)$ traces out an ellipse E with center $(0,0)$. (Note that the ellipse could be degenerate, as would be the case if A were Hermitian.) As t varies from 0 to 1, the term $t\sqrt{1-t^2}$ varies from 0 to $1/2$ and back to 0. This shows that every point in the interior of $(1/2)E$ is attained for some z. Lastly, by considering the angles $\theta = 0$ and $\theta = \pi/2$, we see that the major axis of the ellipse extends from $-(a+b)/2$ to $(a+b)/2$ along the real axis and the minor axis extends from $i(b-a)/2$ to $i(a-b)/2$ along the imaginary axis. Thus in Cartesian coordinates, the ellipse E can be

represented by the equation $\frac{4x^2}{(a+b)^2} + \frac{4y^2}{(a-b)^2} = 1$. So the distance from the center to the foci is $[(1/4)(a+b)^2 - (1/4)(a-b)^2]^{1/2} = \sqrt{ab}$, which means the foci are given by $\pm\sqrt{ab}$, which are precisely the eigenvalues of A. □

Figure 2.2.1 provides an example of $W(A)$ for the 2 by 2 matrix $A = \begin{bmatrix} 4+2i & 1 \\ 3i & 7 \end{bmatrix}$.

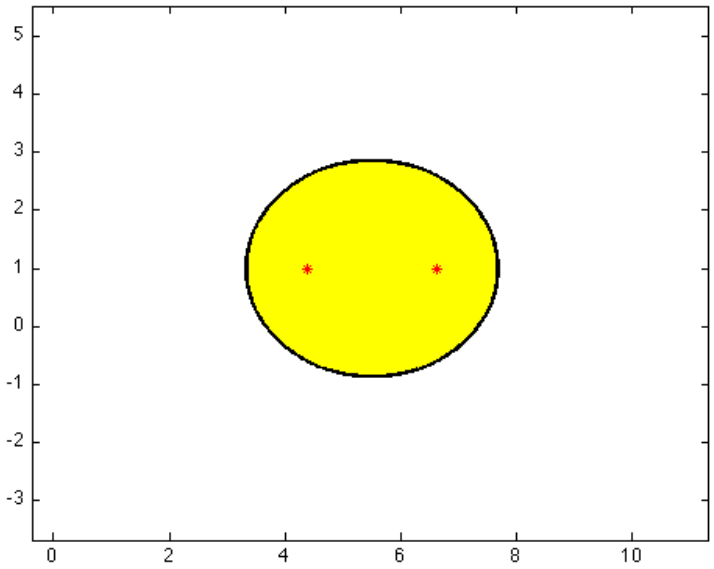


Figure 2.3: The numerical range of a 2 by 2 matrix

2.2.2 Convexity of $W(A)$ for a general $A \in M_n$.

Proof 1 of Theorem 2.2.1. Since we know the result is true in the 2×2 case, suppose $A \in M_n$, where $n > 2$. Let $\gamma, \mu \in W(A)$. We need to show that the line segment connecting γ and μ, denoted $[\gamma, \mu]$, is contained in $W(A)$. Let $x, y \in \mathbb{C}^n$ be unit vectors such that $\gamma = x^*Ax$ and $\mu = y^*Ay$. Let $X \in \mathbb{C}^{n\times2}$ be such that the column space of X contains x and y and $X^*X = I_2$. Then there exists unit vectors $v, w \in \mathbb{C}^2$ such that $Xv = x$ and $Xw = y$. Thus $\gamma = x^*Ax = v^*X^*AXv$ and $\mu = y^*Ay = w^*XAXw$ which means that $\gamma, \mu \in W(X^*AX)$.

Note that X^*AX is a 2 by 2 matrix, thus by Lemma 2.2.3, $W(X^*AX)$ is an ellipse. Since an ellipse is convex, we have that $[\gamma, \mu] \subset W(X^*AX)$. But $W(X^*AX) \subset W(A)$ by Proposition 2.1.1. Thus $W(A)$ contains $[\gamma, \mu]$ which shows that $W(A)$ is convex. □

For the second proof, we also begin with a lemma. This lemma as well as the proof following it are both due to [40].

Lemma 2.2.4. *Let $H \in M_n$ be a Hermitian matrix and let $\gamma \in W(H)$. Then the set $L_H(\gamma) = \{x \in \mathbb{C}^n \mid ||x|| = 1, x^*Hx = \gamma\}$ is path connected.*

Proof. Recall that any Hermitian matrix H is unitarily similar to a diagonal matrix with the eigenvalues of H along the diagonal. By Proposition 2.1.1 we can then assume without loss of generality that H is a real diagonal matrix and we can also assume that $\gamma = 0$. Denote the eigenvalues of H by λ_i, $i = 1, ..., n$. Then, the numerical range of H has the following form

$$W(H) = \left\{ \sum_{j=1}^{n} \lambda_j |x_j|^2 : x_1, ..., x_n \in \mathbb{C}, \ \sum_{j=1}^{n} |x_j|^2 = 1 \right\}.$$

Suppose that the vectors x and y are in $L_H(0)$. We need to show that there is a continuous path in $L_H(0)$ connecting x and y. Note that for any $z \in L_H(0)$, each entry of z is a complex number, and so we can write

$$z = \begin{pmatrix} r_1 e^{i\theta_1} & r_2 e^{i\theta_2} & \cdots & r_n e^{i\theta_n} \end{pmatrix}^T,$$

where $r_j \geq 0$ and $\theta_j \in [0, 2\pi)$, $j = 1, 2, ..., n$. We can connect any such z to the real vector $(r_1 \ r_2 \ \cdots \ r_n)^T$ by the continuous curve

$$z(t) = \begin{pmatrix} r_1 e^{i\theta_1(1-t)} & r_2 e^{i\theta_2(1-t)} & \cdots & r_n e^{i\theta_n(1-t)} \end{pmatrix}^T \quad t \in [0, 1].$$

This curve is also in $L_H(0)$ since $z \in L_H(0)$ implies

$$z(t)^* H z(t) = \sum_{j=1}^{n} \lambda_j |r_j e^{i\theta_j(1-t)}|^2 = \sum_{j=1}^{n} \lambda_j |r_j|^2 = 0.$$

Therefore, we can assume that x and y are both real with nonnegative entries. Now consider the continuous curve

$$u(t) = (u_j(t)) = \left(\sqrt{(1-t)x_j^2 + ty_j^2}\right) \quad t \in [0,1].$$

This curve satisfies $u(0) = x$ and $u(1) = y$. Also, $u(t) \in L_H(0) \cap \mathbb{R}^n$ since all the entries are real and

$$u(t)^* H u(t) = \sum_{j=1}^{n} \lambda_j((1-t)x_j^2 + ty_j^2) = (1-t)\sum_{j=1}^{n} \lambda_j x_j^2 + t\sum_{j=1}^{n} \lambda_j y_j^2 = 0.$$

The last equality follows from the fact that both x and y are in $L_H(0)$. This completes the proof. □

Proof 2 of Theorem 2.2.1. Let γ, μ be two distinct points in $W(A)$. Again by Proposition 2.1.1, we can assume that $\gamma = 0$ and $\mu = 1$. Let $x, y \in \mathbb{C}^n$ be two unit vectors such that $0 = x^*Ax$ and $1 = y^*Ay$. Let $A = H + iK$ where $H = (A + A^*)/2$ is the Hermitian part of A and $iK = (A - A^*)/2$ is the skew-Hermitian part of A. Thus $K = (A - A^*)/2i$ is Hermitian. Since $K = K^*$, by Lemma 2.2.4, the set $L_K(0)$ is path connected. Note that since x^*Ax and y^*Ay have zero imaginary part, x and y are both in $L_K(0)$ (see Proposition 2.1.6). Thus there is a continuous vector function $z(t) : [0,1] \to L_K(0)$ such that $z(0) = x$ and $z(1) = y$. Therefore, the function

$$z(t)^* A z(t) = z(t)^* H z(t) + z(t)^* K z(t) = z(t)^* H z(t)$$

is real and continuous with respect to t. We also have that $z(0)^*Az(0) = x^*Ax = 0$ and $z(1)^*Az(1) = y^*Ay = 1$. So $z(t)^*Az(t)$ takes on every value in $[0,1]$ and so $[0,1] \subset W(A)$ which shows that $W(A)$ is convex. □

The next proposition shows one of the applications of the Toeplitz-Hausdorff theorem.

Proposition 2.2.5. *For any matrix $A \in M_n$, $W(A)$ contains the convex hull of the eigenvalues of A, denoted $co(\sigma(A))$. Moreover, if A is normal, then $W(A) = co(\sigma(A))$.*

Proof. Assume $Ax = \lambda x$ with $\lambda \in \sigma(A)$ and $||x|| = 1$. Thus $x^*Ax = \lambda x^*x = \lambda$. So $\sigma(A) \subset W(A)$. The fact that the convex hull of $\sigma(A)$ is contained in $W(A)$ then follows from Theorem 2.2.1.

To show the second assertion, suppose A is normal. Then A is unitarily diagonalizable, i.e. $A = U^*DU$ for some unitary matrix U and $D = \text{diag}(\lambda_1, ..., \lambda_n)$ where $(\lambda_i)_{i=1}^n$ are the eigenvalues of A. By the unitary invariance property shown in Proposition 2.1.1, we have that $W(A) = W(D)$. Now let $x \in \mathbb{C}^n$ be a unit vector. Then

$$x^*Dx = \sum_{i=1}^{n} \lambda_i |x_i|^2.$$

Since x is a unit vector, $\sum_{i=1}^n |x_i|^2 = 1$. So we see that $W(D)$ is the set of all convex combinations of the eigenvalues of A. Thus $W(A) = W(D) = \text{co}(\sigma(A))$. □

Note that by Proposition 2.1.3 and Proposition 2.2.5 we have for any two $A, S \in M_n$, that $\sigma(A+S) \subset W(A+S) \subset W(A) + W(S)$. So while in general, $\sigma(A+S)$ is unrelated to $\sigma(A)$ and $\sigma(S)$, we can use the numerical range to say something about where the eigenvalues of $A+S$ are located in the complex plane.

Figure 2.4 shows the numerical ranges of the matrix of Example 2.1.2 and of the normal

matrix given by

$$N = \begin{bmatrix} 1 & 1+i & 0 \\ 0 & 1 & 1+i \\ 1+i & 0 & 1 \end{bmatrix}.$$

Note that in both cases, the numerical range contains the convex hull of the eigenvalues and in the normal case, the numerical range equals the convex hull of the eigenvalues.

One of the consequences of knowing the numerical range is convex is the advantage it provides in sketching it. Since we know it is convex, we only need to determine the boundary of $W(A)$ and then just shade in the interior. This idea provides a nice segue into the next section, which deals with the boundary of the numerical range. We pick up the issue of sketching the numerical range in Section 2.5.

2.3 Numerical Boundary

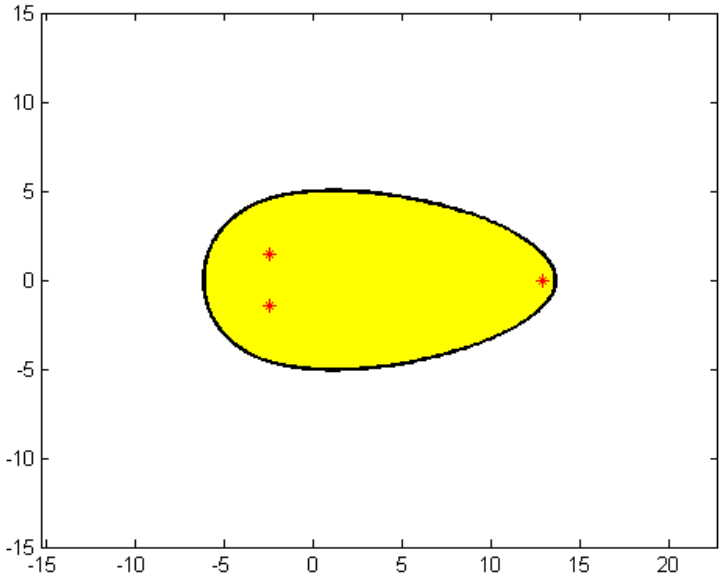
For any matrix $A \in M_n$, $W(A)$ is a compact subset of $\mathbb{C}$. Thus it natural to want to know what can be said about the boundary of $W(A)$. We will denote the *boundary* of the numerical range by $\partial W(A)$. The following is a result dealing with the case where $W(A)$ has empty interior.

Proposition 2.3.1. *Let $A \in M_n$. Then*

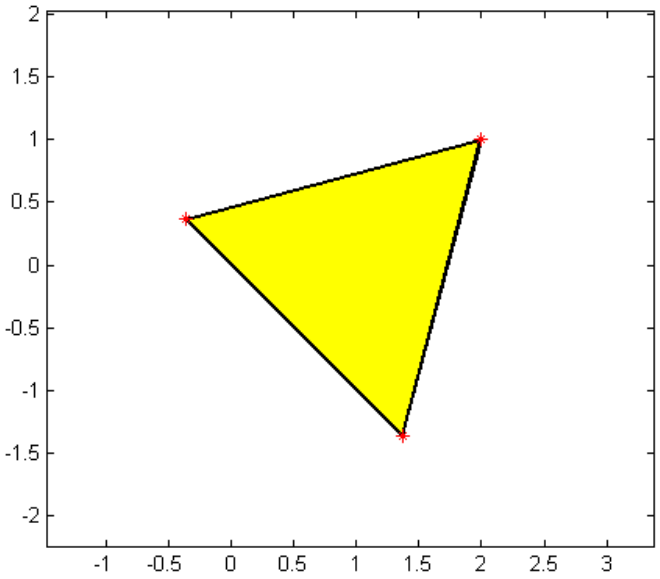
(i) $W(A) = \{\mu\}$ for some $\mu \in \mathbb{C}$ if and only if $A = \mu I$.

(ii) $W(A)$ has empty interior (meaning $W(A) = \partial W(A)$ is a line segment) if and only if $\frac{1}{n} tr\ A \in \partial W(A)$.

Proof. (i) Suppose that $W(A) = \{\mu\}$ for some $\mu \in \mathbb{C}$. Thus for all unit vectors $x \in \mathbb{C}^n$ we



(a) The numerical range of A as defined in Example 2.1.2 with eigenvalues



(b) The numerical range of a normal matrix

Figure 2.4: Example of Spectral Inclusion for a Non-normal and Normal matrix

have that $x^*Ax = \mu$. Therefore,

$$x^*Ax = \mu x^*x \Rightarrow x^*Ax - \mu x^*x = 0 \Rightarrow x^*(A - \mu I)x = 0.$$

Since this holds for all unit vectors $x \in \mathbb{C}^n$, by Lemma 2.1.4 we must have that $A - \mu I = 0$ and so $A = \mu I$.

Conversely, If $A = \mu I$ for some $\mu \in \mathbb{C}$, then for all unit vectors $x \in \mathbb{C}^n$, we have

$$x^*Ax = x^*\mu Ix = \mu x^*x = \mu.$$

Thus $W(A) = \{\mu\}$.

(ii) For the second assertion, suppose that $W(A)$ has empty interior. Then, since $W(A)$ is convex, $W(A) = \partial W(A)$ is a line segment. So $W(A) = r(t) = c_1t + (1-t)c_2$ for some $c_1, c_2 \in \mathbb{C}$, $t \in [0,1]$. By Proposition 2.1.1, we can without loss of generality assume that $c_1 = 1$ and $c_2 = 0$ so $W(A) = [0,1]$. By Proposition 2.1.5, we can conclude that A is Hermitian. Then A is unitarily similar to a diagonal matrix $D = \text{diag}(\lambda_1, ..., \lambda_n)$ where λ_i, $i = 1, ..., n$ are the eigenvalues of A. So again by Proposition 2.1.1, we have that $W(A) = W(D)$. Let $x \in \mathbb{C}^n$ and $x = (\frac{1}{\sqrt{n}} \ \cdots \ \frac{1}{\sqrt{n}})^T$. Then $||x|| = 1$ and

$$x^*Dx = \sum_{i=1}^{n} \frac{1}{n}\lambda_i = \frac{1}{n}\text{tr } D = \frac{1}{n}\text{tr } A.$$

Therefore, $\frac{1}{n}\text{tr } A \in W(A) = \partial W(A)$.

For the reverse direction, let $\zeta = \frac{1}{n}\text{tr } A$ and suppose $\frac{1}{n}\text{tr } A \in \partial W(A)$. Let $B = A - \zeta I$. Then a direct calculation shows that $\frac{1}{n}\text{tr } B = 0$ and so $0 \in W(B)$. By the hypothesis, and Proposition 2.1.1 we can further say that $0 \in \partial W(B)$. Since $W(B)$ is convex, there exists $\theta \in \mathbb{R}$ such that $W(e^{i\theta}B)$ is contained in the closed upper half plane, $UHP = \{z \in \mathbb{C} \mid$

Im $z \geq 0\}$. Let $C = e^{i\theta}B$. At this point, we have the following:

$$W(C) \subset UHP, \quad 0 \in \partial W(C), \quad \text{and} \quad \frac{1}{n}\text{tr } C = 0.$$

Note that $c_{ii} \in W(C)$ and Im $c_{ii} \geq 0$ for $1 \leq i \leq n$. But we must have that $\frac{1}{n}\sum \text{Im } c_{ii} = 0$ and so each $c_{ii} \in \mathbb{R}$. Let $i, j \in \{1, ..., n\}$ be arbitrary indices and let Γ_{ij} denote the 2×2 principal submatrix lying in the rows and columns of C indexed by i, j. Let λ_1, λ_2 denote the eigenvalues of Γ_{ij}. By Proposition 2.1.1, $W(\Gamma_{ij}) \subset W(C) \subset UHP$. Therefore, $\lambda_1, \lambda_2 \in UHP$. But Im $(\lambda_1) + \text{Im}(\lambda_2) = \text{Im}(\lambda_1 + \lambda_2) = \text{Im}(\text{tr}\Gamma_{ij}) = 0$. Therefore, λ_1, λ_2 are real. Since $W(\Gamma_{ij})$ is an ellipse with foci λ_1, λ_2 and has the real axis as its major axis, we must have that $W(\Gamma_{ij}) \subset \mathbb{R}$. By Proposition 2.1.5, we see that Γ_{ij} is Hermitian. Since i, j were arbitrary, it follows that C is also Hermitian and so $W(C)$ is an interval in the real line and thus has empty interior. By the invariance properties of the numerical range, this implies that $W(A)$ has empty interior. □

The proof of (ii) in the above proposition was adapted from a proof in [32], Section 1.6, where an equivalent, yet more general result is presented.

We now turn our attention to the *corners* of the numerical range. In order to talk about the corners, we need the following definition.

Definition 2.3.2. A point $\mu \in W(A)$ is called a *corner* of $W(A)$ if there exist angles θ_1 and θ_2 satisfying $0 \leq \theta_1 < \theta_2 < 2\pi$ such that Re $e^{i\theta}\mu = \max\{\text{Re } \beta \mid \beta \in W(e^{i\theta}A)\}$ for all $\theta \in (\theta_1, \theta_2)$.

We have the following result concerning the corners of the numerical range.

Proposition 2.3.3. *For any $A \in M_n$, if μ is a corner of $W(A)$, then μ is an eigenvalue of A.*

Proof. Let μ be a corner of $W(A)$. There then exists θ_1 and θ_2 as in the definition such that

$$\text{Re } e^{i\theta}\mu = \max\{\text{Re } \beta \mid \beta \in W(e^{i\theta}A)\} \quad \text{for all } \theta \in (\theta_1, \theta_2).$$

By Proposition 2.1.5 and Proposition 2.1.6, we have that this quantity is the same as the largest eigenvalue of the Hermitian part of $e^{i\theta}A$, $\theta \in (\theta_1, \theta_2)$. Thus for all such θ there exists a unit vector x_θ, such that $x_\theta^* A x_\theta = \mu$ and

$$x_\theta^* H(e^{i\theta}A)x_\theta = \lambda_n(H(e^{i\theta}A)) = \text{Re } e^{i\theta}\mu.$$

In fact, we have that the same vector x_θ works for all θ in the interval (θ_1, θ_2). To see why, suppose $\theta' \in (\theta_1, \theta_2)$ is different from θ. Note that

$$\begin{aligned} x_\theta^* H(e^{i\theta'}A)x_\theta &= \frac{x_\theta^* e^{i\theta'}Ax_\theta + x_\theta^* e^{-i\theta'}A^* x_\theta}{2} = \frac{1}{2}(e^{i\theta'}x_\theta^* A x_\theta + \overline{e^{i\theta'}x_\theta^* A x_\theta}) \\ &= \text{Re } x_\theta^* e^{i\theta'}Ax_\theta = \text{Re } e^{i\theta'}\mu = \lambda_n(H(e^{i\theta'}A)), \end{aligned}$$

where the last equality follows from the fact that $\theta' \in (\theta_1, \theta_2)$. From this, we can conclude by Proposition 2.1.5 that $H(e^{i\theta'}A)x_\theta = \lambda_n(H(e^{i\theta'}A))x_\theta = \text{Re } e^{i\theta'}\mu x_\theta$. So from now on, we will simply write x instead of x_θ.

Now let $\lambda_\theta = \text{Re } e^{i\theta}\mu$. Since x is independent of θ, we can take the derivative of the eigenvector equation $H(e^{i\theta}A)x = \lambda_\theta x$ with respect to θ to obtain

$$H(ie^{i\theta}A)x = \lambda'_\theta x,$$

It is easily verified that this is the same as

$$iK(e^{i\theta}A)x = -i\lambda'_\theta x.$$

Now if we add this last equation to the eigenvector equation $H(e^{i\theta}A)x = \lambda_\theta x$, we get

$$e^{i\theta}Ax = (\lambda_\theta - i\lambda'_\theta)x \qquad \text{or} \qquad Ax = e^{-i\theta}(\lambda_\theta - i\lambda'_\theta)x.$$

Thus $e^{-i\theta}(\lambda_\theta - i\lambda'_\theta)$ is an eigenvalue of A. But

$$x^*Ax = e^{-i\theta}(\lambda_\theta - i\lambda'_\theta),$$

which must equal μ since these last two equations hold for all $\theta \in (\theta_1, \theta_2)$. Thus μ is an eigenvalue of A. □

2.4 Numerical Radius

Recall that the spectral radius of a matrix A is given by $r(A) = \sup\{|\lambda| : \lambda \in \sigma(A)\}$. Similar to the spectral radius is the numerical radius, which has the following definition.

Definition 2.4.1. The *numerical radius* of an operator $A \in M_n$, is given by

$$w(A) = \sup\{|\lambda| \mid \lambda \in W(A)\}.$$

Remark 2.4.2. This definition immediately implies that $w(A) \geq 0$, where equality holds if and only if $W(A) = \{0\}$ which, by Proposition 2.3.1, is true if and only if $A = 0$. So the numerical radius satisfies one of the requirements for a norm on M_n. Next, we also have that

$$w(zA) = \sup\{|\langle zAx, x\rangle| : x \in \mathbb{C}^n, ||x|| = 1\} = |z|\{|\langle Ax, x\rangle| : x \in \mathbb{C}^n, ||x|| = 1\} = |z|w(A),$$

and Proposition 2.1.3 shows that the numerical radius also satisfies the triangle inequality. Thus the numerical radius is a norm on M_n. By Proposition 2.2.5, we also have that $r(A) \leq$

$w(A)$ for all $A \in M_n$.

Next, we introduce some basic results on the numerical radius. Since all norms on finite dimensional vector spaces are equivalent, we have that the numerical radius is equivalent to the matrix 2-norm of A. This next result states this more precisely.

Theorem 2.4.3. *Let $||\cdot||$ denote the matrix 2-norm of an operator $A \in M_n$. Then $w(A) \leq ||A|| \leq 2w(A)$.*

Proof. Let $\mu = \langle Ax, x\rangle$ where $||x|| = 1$. Then by the Cauchy-Schwarz inequality, we have

$$|\mu| = |\langle Ax, x\rangle| \leq ||Ax|| \leq ||A||.$$

Since this is true for all $\mu \in W(A)$, we have that $w(A) \leq ||A||$.
For the other inequality, first note that for any nonzero $x \in \mathbb{C}^n$, we have that

$$\langle Ax, x\rangle = \left\langle \frac{Ax}{||x||}, \frac{x}{||x||} \right\rangle ||x||^2 \leq w(A)||x||^2. \tag{2.4.1}$$

We will also make use of the following polarization identity:

$$\begin{aligned} 4\langle Ax, y\rangle = &\langle A(x+y), x+y\rangle - \langle A(x-y), x-y\rangle \\ &+ i\langle A(x+iy), x+iy\rangle - i\langle A(x-iy), x-iy\rangle. \end{aligned} \tag{2.4.2}$$

Now applying (2.4.1) to (2.4.2), we get that

$$\begin{aligned} 4|\langle Ax, y\rangle| &\leq w(A)\Big[||x+y||^2 + ||x-y||^2 + ||x+iy||^2 + ||x-iy||^2\Big] \\ &= \Big[||x||^2 + \langle y, x\rangle + \langle x, y\rangle + ||y||^2 + ||x||^2 - \langle x, y\rangle - \langle y, x\rangle + ||y||^2 \\ &+ ||x||^2 + i\langle y, x\rangle - i\langle x, y\rangle + ||y||^2 + ||x||^2 - i\langle y, x\rangle + i\langle x, y\rangle + ||y||^2\Big] \\ &= 4w(A)\Big[||x||^2 + ||y||^2\Big]. \end{aligned}$$

Since x and y were arbitrary, we can pick $||x|| = ||y|| = 1$ so that

$$| \langle Ax, y \rangle | \leq 2w(A).$$

Now let $y = Ax/||Ax||$. Then

$$\frac{| \langle Ax, Ax \rangle |}{||Ax||} \leq 2w(A).$$

Hence $||Ax|| \leq 2w(A)$. Taking the supremum over all $x \in \mathbb{C}^n$, with $||x|| = 1$ implies $||A|| \leq 2w(A)$. □

The following result deals with one of the extreme cases of Theorem 2.4.3, namely when the numerical radius equals the norm of A.

Theorem 2.4.4. *If $w(A) = ||A||$, then $r(A) = ||A||$.*

Proof. Since $w(A) = ||A||$, we can write

$$\sup_{||x||=1} | \langle Ax, x \rangle | = ||A||.$$

Since $W(A)$ is compact, there exists a unit vector $x \in \mathbb{C}^n$ such that this supremum is attained, that is, $| \langle Ax, x \rangle | = ||A||$ for this particular x. But, by the Cauchy-Schwarz inequality,

$$||A|| = | \langle Ax, x \rangle | \leq ||Ax|| \leq ||A||.$$

So we must have equalities throughout which implies $Ax = \lambda x$ for some $\lambda \in \mathbb{C}$. Thus $\lambda \in \sigma(A)$ and so $r(A) \geq |\lambda| = | \langle Ax, x \rangle | = ||A||$. But since in general, $r(A) \leq w(A) \leq ||A||$, this implies that $r(A) = ||A||$. □

This next theorem deals with the other extreme case of Theorem 2.4.3, which is $w(A) = \frac{1}{2}||A||$. First, let $R(A) = \{Ax \mid x \in \mathbb{C}^n\}$ denote the *range* of a matrix A and $N(A) = \{x \in$

$\mathbb{C}^n \mid Ax = 0\}$ denote the *nullspace* of A. Then we have the following:

Theorem 2.4.5. *If $R(A) \perp R(A^*)$, then $w(A) = \frac{1}{2}||A||$.*

Proof. Let x be a unit vector in $\mathbb{C}^n$. We can write x as $x_1 + x_2$ where $x_1 \in N(A)$ and $x_2 \in R(A^*)$, since $R(A^*) = N(A)^\perp$ by the fundamental theorem of linear algebra. So $Ax_1 = 0$ and since $R(A) \perp R(A^*)$, we also have that $\langle Ax_2, x_2\rangle = 0$. Therefore,

$$\begin{aligned}\langle Ax, x\rangle &= \langle A(x_1 + x_2), x_1 + x_2\rangle \\ &= \langle Ax_1, x_1\rangle + \langle Ax_1, x_2\rangle + \langle Ax_2, x_1\rangle + \langle Ax_2, x_2\rangle \\ &= \langle Ax_2, x_1\rangle .\end{aligned}$$

This implies that

$$|\langle Ax, x\rangle| \leq ||A||\ ||x_2||\ ||x_1|| \leq \frac{||A||}{2}(||x_1||^2 + ||x_2||^2) = \frac{||A||}{2},$$

where the second inequality follows from Young's inequality, or the fact that $(a-b)^2 \geq 0$ for any $a, b \in \mathbb{R}$.

Now since x is arbitrary, we can take the supremum on the left hand side, which yields

$$w(A) \leq \frac{||A||}{2},$$

and since $\frac{||A||}{2} \leq w(A)$ by Theorem 2.4.3, we get that $w(A) = \frac{1}{2}||A||$. □

Remark 2.4.6. As far as we are aware, there is no result indicating whether or not the converse of Theorem 2.4.5 is true.

Example 2.4.7. This example illustrates the result of Theorem 2.4.5. Consider the follow-

ing:

$$A = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}, \quad A^* = \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix}.$$

Then

$$R(A) = \text{span}\left\{ \begin{pmatrix} 1 \\ 0 \end{pmatrix} \right\} \quad \text{and} \quad R(A^*) = \text{span}\left\{ \begin{pmatrix} 0 \\ 1 \end{pmatrix} \right\},$$

so $R(A) \perp R(A^*)$. Note that $||A|| = 1$. To determine $w(A)$, let $(x_1, x_2)^T$ be a unit vector in $\mathbb{C}^2$. Then

$$|x^*Ax| = |\overline{x_1}x_2| = |x_1||x_2|.$$

By Young's inequality, we have $|x_1|^2 + |x_2|^2 = 1 \geq 2|x_1||x_2|$, so $\frac{1}{2} \geq |x_1||x_2|$. Since this bound is attained when $(x_1, x_2)^T = (\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}})^T$, we have $w(A) = \frac{1}{2}$. Hence $w(A) = \frac{1}{2}||A||$.

The last result we show here is the well-known *power inequality*.

Theorem 2.4.8. *Let $A \in M_n$. Then for any positive integer m, we have that $w(A^m) \leq w(A)^m$.*

Before proving this theorem, we prove the following lemmata.

Lemma 2.4.9. *If $w(A) \leq 1$ implies $w(A^m) \leq 1$ for all $m \in \mathbb{N}$, then $w(A^m) \leq w(A)^m$ for all $A \in M_n$.*

Proof. Let $w(A) = c$, for some $c \geq 0$. If $c = 0$, then $A = 0$ and the result holds trivially, so suppose $c > 0$. Then $\frac{1}{c}w(A) = w\left(\frac{1}{c}A\right)$ by Remark 2.4.2. Letting $B = \frac{1}{c}A$, we have by hypothesis that $w(B) \leq 1$ and so $w(B^m) \leq 1$. Therefore, $w\left(\frac{1}{c^m}A^m\right) = \frac{1}{c^m}w(A^m) \leq 1$ which implies $w(A^m) \leq c^m = w(A)^m$. □

Lemma 2.4.10. *Let $A \in M_n$ and $z \in \mathbb{C}$ with $|z| < 1$. Then the following are equivalent:*

(i) $w(A) \leq 1$.

(ii) Re $\langle (I - zA)x, x\rangle \geq 0$ *for all* $x \in \mathbb{C}^n$.

(iii) Re $\langle (I - zA)^{-1}y, y\rangle \geq 0$ *for all* $y \in \mathbb{C}^n$*, provided* $I - zA$ *is invertible.*

Proof. (i) $\Leftrightarrow$ (ii) Suppose that $w(A) \leq 1$. Recall that $|\langle Ax, x\rangle| \leq w(A)||x||^2$ for all $x \in \mathbb{C}^n$ (see (2.4.1)) and Re $z \leq |z|$ for all $z \in \mathbb{C}$. Now let z be any number in $\mathbb{C}$ with $|z| < 1$. Then,

$$\begin{aligned}\text{Re } \langle (I - zA)x, x\rangle &= ||x||^2 - \text{Re } \langle zAx, x\rangle \geq ||x||^2 - |z|\,|\langle Ax, x\rangle| \\ &\geq ||x||^2 - |z|w(A)||x||^2 \geq ||x||^2(1 - |z|) \geq 0,\end{aligned}$$

where the penultimate inequality follows from $w(A) \leq 1$. Conversely, we suppose that Re $\langle (I - zA)x, x\rangle \geq 0$ for all $|z| < 1$. Simplifying, we have that $||x||^2 \geq$ Re $\langle zAx, x\rangle$. Writing $z = te^{i\theta}$ and letting $t \to 1$, we get

$$\text{Re } \left\langle e^{i\theta}Ax, x\right\rangle \leq ||x||^2.$$

This implies that $w(A) \leq 1$. To see why, suppose that $\mu \in W(A)$. Then $\mu = re^{-i\theta}$, where $r > 0$ and $0 \leq \theta < 2\pi$. Also, $\mu = \langle Ay, y\rangle$ for some unit vector $y \in \mathbb{C}^n$. So we have that $|\mu| = r$ and $re^{-i\theta} = \langle Ay, y\rangle$. Thus $r = e^{i\theta}\langle Ay, y\rangle \in \mathbb{R}$ so $r = \text{Re } e^{i\theta}\langle Ay, y\rangle \leq ||y||^2 \leq 1$, by the above equation. Since $r = |\mu|$, and $\mu \in W(A)$ was arbitrary, we have that $w(A) \leq 1$.
(ii) $\Leftrightarrow$ (iii) Suppose $z \in C$, $|z| < 1$ is such that $I - zA$ is invertible. Then $x = (I - zA)^{-1}y$ for some $y \in \mathbb{C}^n$. Therefore, by plugging in $(I - zA)^{-1}y$ for x, we get

$$\begin{aligned}&\text{Re } \langle (I - zA)x, x\rangle \geq 0 \quad \text{for all } x \in \mathbb{C}^n, \\ &\Leftrightarrow \text{Re } \left\langle y, (I - zA)^{-1}y\right\rangle \geq 0 \quad \text{for all } y \in \mathbb{C}^n, \\ &\Leftrightarrow \text{Re } \left\langle (I - zA)^{-1}y, y\right\rangle \geq 0 \quad \text{for all } y \in \mathbb{C}^n.\end{aligned} \tag{2.4.3}$$

This completes the proof. □

Proof of Theorem 2.4.8. Assume $w(A) \leq 1$. By Lemma 2.4.9, it suffices to show that this implies $w(A^m) \leq 1$ for all $m \in \mathbb{N}$. To do this, we use Lemma 2.4.10 (iii). The invertibility of $I - zA$ for $|z| < 1$ follows from the fact that $r(A) \leq w(A) \leq 1$. Furthermore, $r(A) \leq 1$ implies $r(A^m) \leq 1$ for all $m \in \mathbb{N}$, so by similar reasoning, $I - z^m A^m$ is invertible.
So to prove the theorem, it is sufficient to show that for all $x \in \mathbb{C}^n$

$$\text{Re } \left\langle (I - z^m A^m)^{-1} x, x \right\rangle \geq 0 \quad \text{where } z \in \mathbb{C}, |z| < 1,$$

since this condition will imply that $w(A^m) \leq 1$. To do this, we use the following identity:

$$(I - z^m A^m)^{-1} = \frac{1}{m} \sum_{k=0}^{m-1} (I - \omega^k z A)^{-1}, \tag{2.4.4}$$

where ω is a primitive m^{th} root of unity[1]. Note that for $0 \leq k \leq m-1$, $|\omega^k| = 1$ so $|\omega^k z| < 1$ and since $w(A) \leq 1$, we have, by Lemma 2.4.10, that

$$\text{Re } \left\langle (I - \omega^k z A)^{-1} x, x \right\rangle \geq 0 \quad \text{for all } x \in \mathbb{C}^n, |z| < 1, k = 0, ..., m-1.$$

Therefore,

$$\begin{aligned} \text{Re } \left\langle (I - z^m A^m)^{-1} x, x \right\rangle &= \text{Re } \left\langle \frac{1}{m} \sum_{k=0}^{m-1} (I - \omega^k z A)^{-1} x, x \right\rangle \\ &= \frac{1}{m} \sum_{k=1}^{m-1} \text{Re } \left\langle (I - \omega^k z A) x, x \right\rangle \geq 0 \quad \text{for all } x \in \mathbb{C}^n, |z| < 1. \end{aligned}$$

By Lemma 2.4.10, this implies that $w(A^m) \leq 1$. Thus $w(A^m) \leq w(A)^m$ by Lemma 2.4.9. $\square$

[1]For an explanation of (2.4.4), see the solution to Problem 176 in [27].

2.5 Sketching the Numerical Range

Many of the ideas and results of this section are taken from [6]. The following lemma will be of aid in showing how to sketch the numerical range. It deals with the situation where the numerical range lies on or to left of a line Re $z = \mu$. But first we need the following terminology:

Definition 2.5.1. Let $A \in M_n$. Then A can be written as $A = H + iK$ where H is the Hermitian part of A and iK is the skew-Hermitian part of A. Let μ be an eigenvalue of H. Let $E_\mu = \{u \in \mathbb{C}^n \mid Hu = \mu u\}$ be the eigenspace of H corresponding to the eigenvalue μ. Let P denote the orthogonal projection of $\mathbb{C}^n$ onto E_μ and consider the linear transformation PKP. Now suppose $\{q_1, ..., q_j\}$, $1 \leq j \leq n$ is an orthonormal basis for E_μ. If Q is the matrix whose columns are $q_1, ..., q_j$, then the $j \times j$ matrix Q^*KQ is the *restriction* of PKP to E_μ with respect to the columns $q_1, ..., q_j$.

Lemma 2.5.2. *Let $A \in M_n$ and write $A = H + iK$ where $H = H^* = (A + A^*)/2$ and $K = K^* = (A - A^*)/2i$. If μ is a real number such that Re $\langle Ax, x\rangle \leq \mu$ for every unit vector $x \in \mathbb{C}^n$, then only one of the following situations holds:*

(i) $W(A)$ does not intersect the line Re $z = \mu$.

(ii) $W(A) \cap \{z \mid$ Re $z = \mu\} = \mu + iW(Q^*KQ)$ where Q^*KQ is defined as in Definition 2.5.1. Moreover, the set $\mu + iW(Q^*KQ)$ is a point or a line segment.

Proof. If $W(A)$ does not intersect Re $z = \mu$, then there is nothing to show. So suppose the intersection is not empty. By Proposition 2.1.6, we have that if $\gamma \in W(A)$, then Re $\gamma \in W(H)$ and Im $\gamma \in W(K)$. Thus since μ is the maximum of Re $W(A)$, we have that μ is the maximum value of $W(H)$. By Proposition 2.1.5, this means that $\mu = \lambda_n$, the largest eigenvalue of H. Furthermore, we also showed in Proposition 2.1.5 that if $x \in \mathbb{C}^n$ is a unit

vector that satisfies $\mu = \text{Re } \langle Ax, x\rangle = \langle Hx, x\rangle$, then $Hx = \mu x$. In other words,

$$W(A) \cap \{z \mid \text{Re } z = \mu\} = \{\mu + i \langle Kx, x\rangle \mid Hx = \mu x,\ ||x|| = 1\}.$$

Now note that if x is such that $Hx = \mu x$, then $x \in E_\mu$, where E_μ is defined as in Definition 2.5.1. Further, if $\dim(E_\mu) = j$, then let $\{q_1, ..., q_j\}$ be an orthonormal basis for E_μ and let Q be defined as in Definition 2.5.1. Then $x = Qu$ for some $u \in \mathbb{C}^j$ and $\langle Kx, x\rangle = \langle KQu, Qu\rangle = \langle Q^*KQu, u\rangle$. Thus any $x \in E_\mu$ yields a point in $W(Q^*KQ)$. Conversely, if $z \in W(Q^*KQ)$, then $z = \langle Q^*KQu, u\rangle$ for some $u \in \mathbb{C}^j$. But $\langle Q^*KQu, u\rangle = \langle KQu, Qu\rangle = \langle Kx, x\rangle$, where $x = Qu$. It follows that $x \in \text{span}\{q_1, ..., q_j\} = E_\mu$ and so $\langle Ax, x\rangle = \mu + i \langle Kx, x\rangle$. Finally, note that Q^*KQ is Hermitian, and therefore its numerical range is a point or a line segment by Proposition 2.1.5

□

The following remark will be useful in determining which points of $W(A)$ lie on the boundary.

Remark 2.5.3. Consider $e^{-it}W(A)$ and recall that this is the same as $W(e^{-it}A)$ by Proposition 2.1.1. Now compute the matrices H_t and K_t such that $e^{-it}A = H_t + iK_t$ and let μ_t be the maximum eigenvalue of H_t. If x_t is a unit eigenvector of H_t corresponding to μ_t, then the line $\{a + bi : a, b \in \mathbb{R},\ a = \mu_t\}$ is a supporting line of $W(e^{-it}A)$ at the point $x_t^*(e^{-it}A)x_t$. Then, we have that the line

$$e^{it}\{a + bi : a, b \in \mathbb{R},\ a = \mu_t\} = \{c + di : c, d \in \mathbb{R},\ c = \mu_t \cos t - b \sin t\}.$$

is a supporting line of $W(A)$ at the point $x_t^*Ax_t$. So we can conclude that $x_t^*Ax_t$ is in the boundary of $W(A)$.

Using the above ideas, we can outline an algorithm to sketch $W(A)$. Let $t \in [0, 2\pi)$ and consider $e^{-it}A = H_t + iK_t$. Let μ_t be the maximum eigenvalue of H_t, with corresponding

eigenspace E_{μ_t}. Let $\{q_1, .., q_j\}$ be an orthonormal basis for E_{μ_t} and let Q_t be the matrix whose columns are the vectors $\{q_1, ..., q_j\}$. If dim $E_{\mu_t} = 1$, then $W(Q^*KQ)$ is a point and we can simply take $q_1^* A q_1$ as the point in the boundary of $W(A)$ for this value of t. If $j > 1$, the next step is to form the matrix $Q_t^* K_t Q_t$ and compute $W(Q_t^* K_t Q_t)$. Then by Lemma 2.5.2, the set $\mu_t + W(Q_t^* K_t Q_t)$ is a line segment. By Remark 2.5.3, the line segment $e^{it}(\mu_t + W(Q_t^* K_t Q_t))$ makes up the part of the boundary of $W(A)$ that intersects the supporting line $e^{it}\{a + bi : a, b \in \mathbb{R},\ a = \mu_t\}$. If we do this for sufficiently many values of t, we can get a good approximation to the boundary of the numerical range. By the convexity of the numerical range, the last step is to just fill in the region described by the boundary. The following theorem describes this process in more detail.

Theorem 2.5.4. *Let $A \in M_n$. For $0 \leq t < 2\pi$, let H_t and K_t be Hermitian matrices so that $e^{-it}A = H_t + iK_t$ and let P_t be the orthogonal projection of $\mathbb{C}^n$ onto the eigenspace of H_t corresponding to μ_t, the largest eigenvalue of H_t. Denote this eigenspace by E_{μ_t}. Let $Q_t = [q_1|\cdots|q_j]$ where dim$(E_{\mu_t}) = j$, $1 \leq j \leq n$ and $\{q_1, ..., q_j\}$ is an orthonormal basis for E_{μ_t}. Let v_t^+ and v_t^- be unit eigenvectors of $Q_t^* K_t Q_t$ corresponding to the greatest and least eigenvalues of $Q_t^* K_t Q_t$. Then*

(i) $Q_t v_t^+ = x_t^+$ and $Q_t v_t^- = x_t^-$ are eigenvectors of H_t corresponding to μ_t.

(ii) The numbers $\langle Ax_t^+, x_t^+ \rangle$ and $\langle Ax_t^-, x_t^- \rangle$ are in the boundary of $W(A)$ and $W(A)$ is the convex hull of these numbers.

Proof. Note that for any $v \in \mathbb{C}^j$, the vector $Q_t v$ is a linear combination of the columns of Q_t and thus is in E_{μ_t}. So in particular, Qv_t^+ and Qv_t^- are in E_{μ_t} and therefore are eigenvectors of H_t corresponding to μ_t. This proves (i). Next, Lemma 2.5.2 shows that for each $t \in [0, 2\pi)$,

$$\text{Re}\left(\langle e^{-it}Ax, x\rangle\right) \leq \mu_t$$

for all $x \in \mathbb{C}^n$, $||x|| = 1$. Now let x_t be a eigenvector of H_t corresponding to μ_t. Then Re $\langle e^{-it}Ax_t, x_t\rangle = \mu_t$, which implies that $\langle Ax_t, x_t\rangle$ must be on the boundary of the numerical range, by Remark 2.5.3. Note that since $x_t^+ = Qv_t^+$ and $x_t^- = Qv_t^-$ are both in E_{μ_t}, that this result also holds for $\left\langle Ax_t^+, x_t^+\right\rangle$ and $\left\langle Ax_t^-, x_t^-\right\rangle$.

Since the numerical range is convex, and each number of the form $\left\langle Ax_t^+, x_t^+\right\rangle$ or $\left\langle Ax_t^-, x_t^-\right\rangle$ as described above is in the numerical range, then the convex hull of these numbers is in $W(A)$. Denote this convex hull by C. Now if $\gamma \in W(A)$ is not in C then there is a line, denote it by L, separating γ from C. Let θ be so that $e^{-i\theta}L$ is vertical and so that C lies to the left of $e^{-i\theta}L$. But then γ must be so that Re $e^{-i\theta}\gamma > \mu_\theta$ which is a contradiction. Moreover, if $\gamma \in W(A)$ has real part equal to μ_θ, then by Lemma 2.5.2, the imaginary part of γ must be in $W(Q_\theta^* K_\theta Q_\theta)$. In other words, such a γ is on the line segment connecting $\left\langle Ax_\theta^+, x_\theta^+\right\rangle$ and $\left\langle Ax_\theta^-, x_\theta^-\right\rangle$. This shows that $W(A)$ is indeed equal to C. □

Code implementing this algorithm can be found in Appendix A. The following example provides some insight into how this algorithm works.

Example 2.5.5. Let the matrix A be given by

$$A = \begin{bmatrix} -5 & 0 & i \\ -4i & -5-3i & 0 \\ i & 4i & -5 \end{bmatrix}.$$

The numerical range of this matrix is plotted in Figure 2.5. We will sketch out the process of determining the structure of the boundary of $W(A)$ by tracing out the algorithm for one particular value of t from the range $[0, 2\pi)$. Let $t = \pi/2$ and consider $A_t = e^{-it}A$. The numerical range $W(A_t) = W(e^{-it}A) = e^{-it}W(A)$ and thus $W(A_t)$ is just a clockwise rotation of $W(A)$ through an angle of $\pi/2$ about the origin (see Figure 2.5(b)).

The next step is to calculate the matrix H_t, such that $A_t = H_t + iK_t$, and consider its

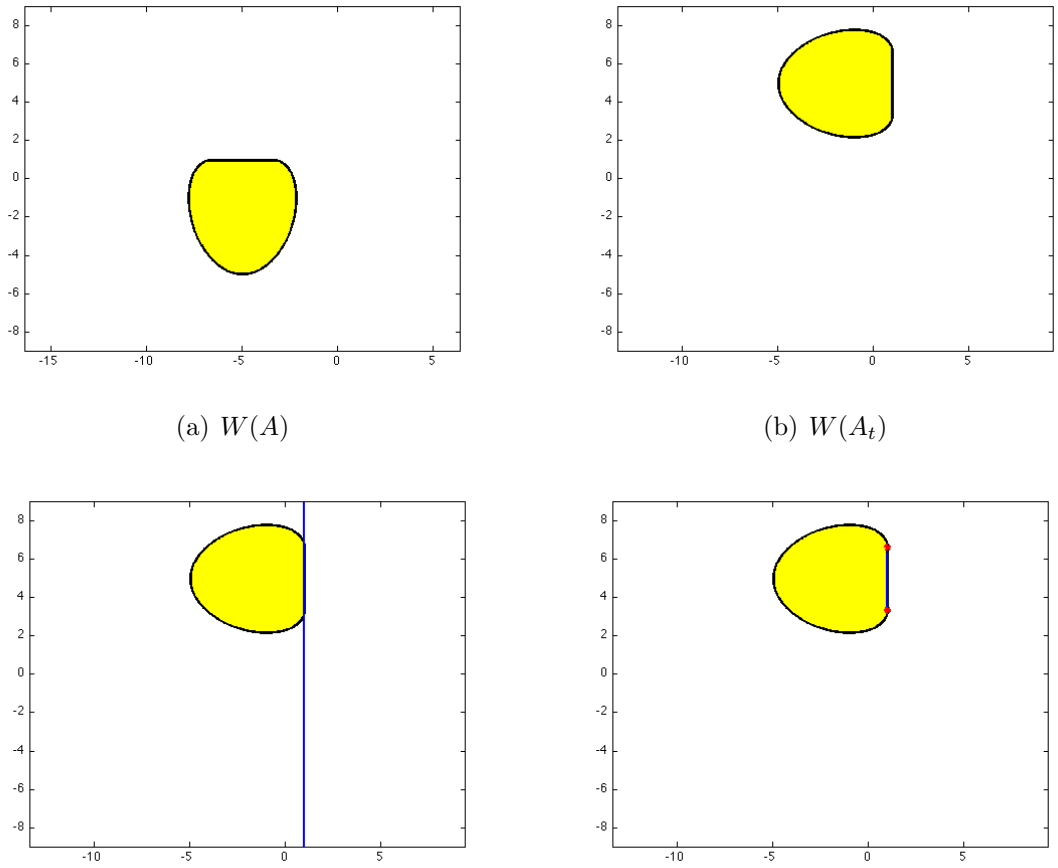
maximum eigenvalue with corresponding eigenvector. H_t is given by

$$H_t = \begin{bmatrix} 0 & -2 & 1 \\ -2 & -3 & 2 \\ 1 & 2 & 0 \end{bmatrix}$$

with eigenvalues 1,1,-5. So the maximum eigenvalue of H_t is 1 (with algebraic multiplicity 2). Let x_t be any one of the corresponding eigenvectors. According to the above results, this means that the line $L = \{1 + bi \mid b \in \mathbb{R}\}$ should be a supporting line of $W(A_t)$ at the point $x_t A_t x_t$. This can be seen in Figure 2.5(c). Now since the eigenspace of H_t corresponding to 1 is two dimensional, this means that the set $1 + W(Q_t^* K_t Q_t)$ is a line segment and moreover, by the above remarks, we can conclude that the part of $W(A_t)$ that intersects L is given by the segment connecting the points $1 + \lambda_- i$ and $1 + \lambda_+ i$ where λ_- and λ_+ are the least and greatest eigenvalues of $Q_t^* K_t Q_t$, respectively. Calculating these numbers, we see that $L \cap W(A_t)$ is the line segment connecting the points $1 + 3.3670i$ and $1 + 6.6330i$. Denote this line segment by ℓ. This can be seen in Figure 2.5(d) At this point we have completely determined the part of the boundary of $W(A_t)$ that intersects the supporting line L. By Remark 2.5.3, this means that $e^{(i\pi/2)}\ell$ lies on the boundary of $W(A)$. This can be seen in Figure 2.6.

The preceding example dealt with the case where the eigenspace E_{μ_t} was two dimensional. Most of the time, however, E_{μ_t} is one dimensional. In this case the supporting line L intersects $W(A_t)$ at a single point, given by $x_t^* A_t x_t$, where x_t is the eigenvector corresponding to μ_t. In this case, there is no need to calculate the matrix $Q_t^* K_t Q_t$ and we can simply note that the point $e^{it} x_t^* A_t x_t$ lies in the boundary of $W(A)$.

2.5.1 Accuracy. While the process described above can be a good way to compute the numerical range, it does not provide any measure of accuracy. We can introduce such a



(a) $W(A)$ (b) $W(A_t)$

(c) Supporting line of $W(A_t)$ (d) Highlighted section of $W(A_t)$

Figure 2.5: Figures for Example 2.5.5

measure by considering interior and exterior approximating polygons. An interior polygon Q is formed by connecting the points in the boundary of $W(A)$, which are given by $e^{it}x_tA_tx_t$. An exterior polygon P is obtained by taking the points of intersection of the supporting lines at these points. See Figure 2.7. The boundary of $W(A)$ is the solid line and the vertices of the interior and exterior approximating polygons are marked by dots and squares, respectively. Now, the vertices $\{p_k\}_{k=1}^n$ of P can be determined as follows: Let $\{\theta_k\}_{k=1}^n$ denote the rotation angles used to compute the points in the boundary of $W(A)$, in increasing order. Given two

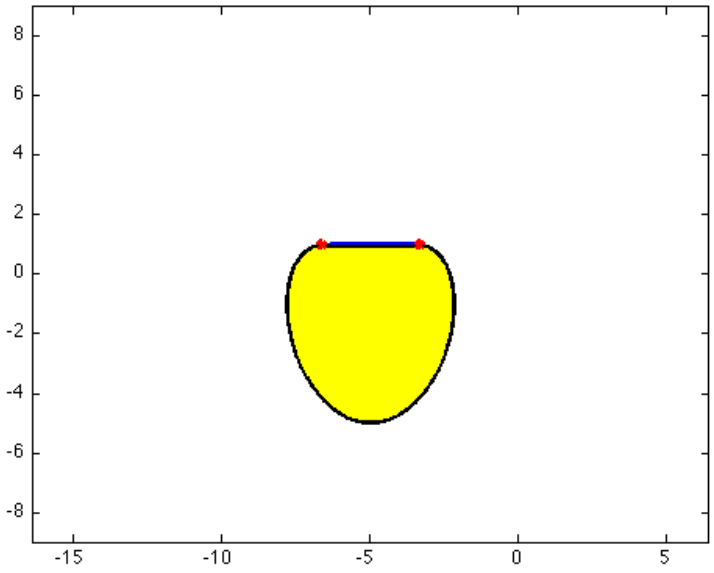


Figure 2.6: A section of the boundary of $W(A)$

consecutive angles θ_j, and θ_{j+1}, we have two supporting lines of $\partial W(A)$, given by

$$e^{i\theta_j}\{a+bi \mid a,b\in\mathbb{R}, a=\mu_{\theta_j}\} \quad \text{and} \quad e^{i\theta_{j+1}}\{c+di \mid c,d\in\mathbb{R}, c=\mu_{\theta_{j+1}}\},$$

where μ_{θ_k} equals the maximum eigenvalue of the Hermitian part of $e^{-i\theta_k}A$, $k=1,2,...,n$. We find the point of intersection of these two lines by equating their real and imaginary parts. The result is

$$p_j = e^{i\theta_j}\left(\mu_{\theta_j} + \frac{\mu_{\theta_j}\cos(\theta_{j+1}-\theta_j) - \mu_{\theta_{j+1}}}{\sin(\theta_{j+1}-\theta_j)}\right)$$

Given a polygon with vertices $p_k = \{x_k + iy_k\}_{k=1}^n$ in counter-clockwise order, the area of this polygon can be computed by the following formula:

$$\frac{1}{2}\text{Im}\sum_{k=1}^{n}\bar{p}_k p_{k+1},$$

where $p_{n+1} = p_1$. If the vertices are given in clockwise order, then the result of this equation will be negative, but correct in absolute value. Thus, if the exterior polygon P has vertices $\{p_k\}_{k=1}^n$ and the interior polygon Q has vertices $\{q_k\}_{k=1}^n$, then the difference in area is

$$\frac{1}{2}\text{Im}\left[\sum_{k=1}^{n} \bar{p}_k p_{k+1} - \bar{q}_k q_{k+1}\right].$$

By specifying this quantity to be small, we can therefore have a measure of how accurate our numerical approximation to $W(A)$ is. Code computing the numerical range to within a specified tolerance can be found in Appendix B. This code is based on a method by Higham [31].

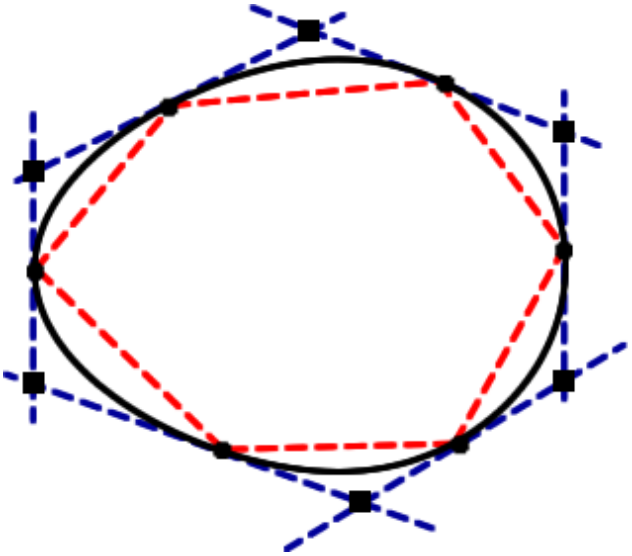


Figure 2.7: Interior and Exterior Approximating Polygons

Chapter 3. QR Factorization, Arnoldi, and GMRES

We now introduce the GMRES method. The acronym stands for for "generalized minimal residuals" and the method was introduced by Saad and Schultz in 1985 [41]. GMRES is notable for its ability to effectively solve systems where the coefficient matrix is not Hermitian. It is one of a family of iterative methods known as *Krylov methods*. These methods, including GMRES, are used to solve the following linear system of equations

$$Ax = b, \quad A \in M_n, \quad b \in \mathbb{C}^n, \tag{3.0.1}$$

where we assume the coefficient matrix A is nonsingular. Since Krylov methods are iterative, they must begin with an initial guess, x_0, which is often taken to be 0. At step m of the iteration, a Krylov method will produce an approximate solution x_m, which lies in the affine Krylov subspace generated by the initial residual $r_0 = b - Ax_0$. This affine subspace is written as

$$x_0 + K_m(A, r_0) = \text{span}\{r_0, Ar_0, A^2r_0, ..., A^{m-1}r_0\}.$$

We will abbreviate $K_m(A, r_0)$ as K_m when there is no confusion. Also, let $\mathcal{K}_m$ denote the *Krylov matrix*, which is given by

$$\mathcal{K}_m = \left[\begin{array}{c|c|c|c} r_0 & Ar_0 & \cdots & A^{m-1}r_0 \end{array}\right]$$

Krylov methods can be very powerful and useful. For instance, if the matrix in question is quite large, it is impractical to solve (3.0.1) via Gaussian elimination or by some factorization method due to the computational complexity. In some cases, we may not even have direct access to the matrix, that is, it only exists as a "black box" routine, which when given a

vector x, returns Ax. In these cases, Krylov methods are the tool of choice.

A natural question at this point is why do we search for a solution within a Krylov subspace? To answer this question, let $A \in M_n$, with A invertible, and let $p_A(t) = c_n t^n + \cdots c_1 t + c_0$, denote the characteristic polynomial of A (note that the invertibility of A guarantees that $c_0 \neq 0$). By the Cayley-Hamilton theorem, $p_A(A) = 0$. Therefore,

$$c_n A^n + c_{n-1} A^{n-1} + \cdots + c_1 A + c_0 I = 0.$$

Multiplying both sides by A^{-1} and solving for A^{-1}, we get

$$A^{-1} = \frac{-c_n A^{n-1} - c_{n-1} A^{n-2} - \cdots - c_1 I}{c_0},$$

and so

$$A^{-1} r_0 = -\frac{c_n}{c_0} A^{n-1} r_0 - \frac{c_{n-1}}{c_0} A^{n-2} r_0 - \cdots - \frac{c_1}{c_0} r_0 \in K_n. \tag{3.0.2}$$

Since

$$x = x_0 + A^{-1} r_0 = x_0 + A^{-1}(b - A x_0) = A^{-1} b,$$

we see that the solution x of (3.0.1) can be written as a vector in the affine Krylov subspace $x_0 + K_n$. So it makes sense to search for a solution within this space.

We should note the above results do not apply to singular systems. For a discussion on how Krylov methods apply in this case, see [33].

Returning to the topic of GMRES, we first begin with a review of the QR factorization. This provides an introduction to the Arnoldi iteration which is the basis for GMRES.

3.1 QR Factorization

Let $A \in M_{m,n}$ where $m \geq n$ and A full rank. The goal of QR factorization is compute an $m \times n$ matrix $\hat{Q}$ with orthonormal columns and an $n \times n$ upper triangular matrix $\hat{R}$ such that $A = \hat{Q}\hat{R}$. If $m > n$, then $\hat{Q}$ is not a square matrix, and this is called the *reduced* QR factorization of A. To get what is known as the *full* QR factorization, we append an additional $m - n$ orthonormal columns to $\hat{Q}$ and $m - n$ rows of zeros to $\hat{R}$, so that $A = QR$ where Q is an $m \times m$ unitary matrix and R is an $m \times n$ upper triangular matrix. If $m = n$, then it follows that the reduced and full QR factorizations coincide. Whenever $m \geq n$, it can be shown that every $A \in M_{m,n}$ with full rank has a unique QR factorization. For further details on these ideas, see [48, Chapter 7]. Thus we can simplify the following discussion by assuming that A is a square $n \times n$ matrix. The matrix formula $A = QR$ can be written

$$\left[\begin{array}{c|c|c|c} a_1 & a_2 & \cdots & a_n \end{array}\right] = \left[\begin{array}{c|c|c|c} q_1 & q_2 & \cdots & q_n \end{array}\right] \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1n} \\ & r_{22} & & \vdots \\ & & \ddots & \vdots \\ & & & r_{nn} \end{bmatrix}. \tag{3.1.1}$$

Note that this implies that each column of A can be written as a linear combination of the columns of Q, i.e.

$$\begin{aligned} a_1 &= r_{11}q_1, \\ a_2 &= r_{12}q_1 + r_{22}q_2, \\ a_3 &= r_{13}q_3 + r_{23}q_2 + r_{33}q_3, \\ &\vdots \\ a_n &= r_{1n}q_1 + r_{2n}q_2 + \cdots + r_{nn}q_n. \end{aligned} \tag{3.1.2}$$

The vectors $\{q_j\}$ and coefficients r_{ij} are determined via Gram-Schmidt orthogonalization.

We begin by setting $q_1 = a_1/||a_1||$ and $r_{11} = ||a_1||$. Then for $j = 2, ..., n$, we have

$$\begin{aligned} p_j &= a_j - \sum_{i=1}^{j-1} r_{ij} q_i, \quad \text{where } r_{ij} = q_i^* a_j, \\ q_j &= p_j / r_{jj}, \quad \text{where } r_{jj} = ||p_j||. \end{aligned} \tag{3.1.3}$$

The QR algorithm can be stated concisely as follows:

1: **for** $j = 1$ to n **do**
2: $\quad v_j = a_j$
3: $\quad$ **for** $i = 1$ to $j - 1$ **do**
4: $\quad\quad r_{ij} = q_i^* a_j$
5: $\quad\quad v_j = v_j - r_{ij} q_i$
6: $\quad$ **end for**
7: $\quad r_{jj} = ||v_j||^2$
8: $\quad q_j = v_j / r_{jj}$
9: **end for**

However, due to rounding errors in the computer, this algorithm typically yields vectors $q_1, q_2, ..., q_n$ that are not quite orthogonal. Thus, in this form, the Gram-Schmidt algorithm is numerically unstable. Fortunately, a small modification will ensure stability; the resulting algorithm is called the *modified Gram-Schmidt* algorithm. The difference in the modified version is that each time we form the scalar r_{ij}, we take the inner product of q_i and the newly modified vector v_j instead of using a_j each time (see line 4 in the above listing). In other words, rather than subtracting out the components of a_j in the direction of $q_1, ..., q_{j-1}$ all at once (as is done in (3.1.3)), they are subtracted out one at a time. This amounts to replacing the inner loop of the above algorithm with the following:

for $i = 1$ to $j - 1$ **do**

$r_{ij} = q_i^* v_j$

$v_j = v_j - r_{ij} q_i$

end for

These two methods are mathematically equivalent, yet the latter turns out to be numerically stable. Lastly, if A is rank deficient, then for some j we will get that $v_j = 0$. In this case, we can simply pick q_j to be any vector orthogonal to the previous $q_1, ..., q_{j-1}$ and keep going. For example, if $v_3 = 0$, then this means that a_3 is a linear combination of q_1 and q_2. Thus we can take q_3 to be an arbitrary unit vector orthogonal to $\text{Span}\{q_1, q_2\}$. To ensure that the relation $A = QR$ is still valid, simply set $r_{33} = 0$.

Next we will see how QR factorization can be used to form an iterative method to compute a reduction $A = QHQ^*$ where Q is an $n \times n$ unitary matrix and H is an $n \times n$ upper Hessenberg matrix.

3.2 Arnoldi Iteration

An upper Hessenberg matrix is a matrix where all entries below the first subdiagonal are zero. For example, the matrix

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ 0 & a_{32} & a_{33} & a_{34} \\ 0 & 0 & a_{43} & a_{44} \\ 0 & 0 & 0 & a_{54} \end{bmatrix}$$

is an upper Hessenberg matrix. Note that this class of matrices includes square and non-square matrices. Upper Hessenberg matrices have applications in eigenvalue algorithms and solving systems of equations (see [48, Chapters 25, 26, 33, and 35]). Given a matrix

$A \in M_n$, we can reduce the entire matrix to one that is in upper Hessenberg form via a series of Householder reflections, yielding the relation $Q^*AQ = H$, where Q is unitary, and H upper Hessenberg (one could also use Givens rotations, but this method requires more work than applying Householder reflections, see [48, Exercise 10.4]). However, in many applications, n is large or infinite and it is not practical (or even needed) to compute a full upper Hessenberg reduction. Furthermore, to obtain the matrix Q, the reduction must be carried to completion. What is advantageous about Arnoldi iteration is that we obtain the columns of Q and H one at a time. This way, we can stop the iteration whenever we please with a partial Hessenberg reduction. This is useful in the context of an iterative algorithm to solve (3.0.1), since these algorithms are typically never run to completion.

To get an idea of how the reduction $Q^*AQ = H$ is obtained, we consider the first m columns of $AQ = QH$. Let $Q_m = [q_1|q_2|\cdots|q_m]$ be the matrix whose columns are the first m columns of Q. Due to the Hessenburg structure of H, it is clear that the j^{th} column of AQ_m requires a linear combination of the first $j+1$ columns of Q, with coefficients from H. So we have that

$$AQ_m = Q_{m+1}H_{m+1,m}, \tag{3.2.1}$$

where $H_{m+1,m}$ is the $(m+1) \times m$ upper-left section of H, which is also an upper Hessenburg matrix:

$$H_{m+1,m} = \begin{bmatrix} h_{11} & h_{12} & \cdots & h_{1m} \\ h_{21} & h_{22} & & \vdots \\ & \ddots & \ddots & \\ & & h_{m,m-1} & h_{mm} \\ & & & h_{m+1,m} \end{bmatrix}.$$

Now, the m^{th} column of (3.2.1) is given by

$$Aq_m = h_{1m}q_1 + \cdots + h_{mm}q_m + h_{m+1,m}q_{m+1}. \tag{3.2.2}$$

Rearranging the above equation yields

$$q_{m+1} = \frac{Aq_m - h_{1m}q_1 - \cdots - h_{mm}q_m}{h_{m+1,m}},$$

assuming that $h_{m+1,m}$ is not 0. Thus at each step in the Arnoldi process, we obtain a column of Q_m as well as a column of H_m. Since we require the $\{q_i\}$ to be orthonormal, the h_{ij} are determined by using modified Gram-Schmidt orthogonalization.

Like many iterative algorithms, we need to supply a starting point for the Arnoldi iteration, so let $y \in \mathbb{C}^n$ be arbitrary. We then have the following. Note that the inner loop is the same as in the modified QR algorithm.

1: $y =$ arbitrary, $q_1 = y/||y||$
2: **for** $i = 1, 2, 3, ...$ **do**
3: $\quad v = Aq_i$
4: $\quad$ **for** $j = 1$ to i **do**
5: $\quad\quad h_{ji} = q_j^* v$
6: $\quad\quad v = v - h_{ji}q_j$
7: $\quad$ **end for**
8: $\quad h_{i+1,i} = ||v||$
9: $\quad q_{i+1} = v/h_{i+1,i}$
10: **end for**

By examining this algorithm carefully, it is evident that the $\{q_i\}$ span the successive Krylov subspaces. Thus one interpretation of the Arnoldi iteration is that it produces orthonormal

bases of these spaces. In other words,

$$K_m = \text{Span}\{y, Ay, ..., A^{m-1}y\} = \text{Span}\{q_1, q_2, ..., q_m\} \subset \mathbb{C}^n. \tag{3.2.3}$$

This idea will be important later in understanding how GMRES works.

The only issue we have not addressed is what if $h_{i+1,i} = 0$? Physically, this means that $Aq_i \in \langle q_1, ..., q_i \rangle$. In this case, as in the QR factorization, we simply pick q_{i+1} to be any vector orthogonal to $\{q_1, ..., q_i\}$, set $h_{i+1,i} = 0$, and keep going. However, in the context of GMRES, this is the indication that the solution has been found. We will return to this idea below.

3.3 GMRES

To see how Arnoldi iteration is used to solve the system $Ax = b$, denote the true solution to this problem by $\hat{x}$. Recall from the discussion above that $\hat{x}$ can be written as a linear combination of Krylov vectors, plus a shift, as in (3.0.2). The idea of GMRES is to approximate the true solution $\hat{x}$ by the vector $x_m \in x_0 + K_m$ that minimizes the 2-norm of the residual $r_m = b - Ax_m$. The vector x_0 is the initial guess, which can be nonzero, but in most cases is taken to be 0. Since the goal is to minimize $||r_m||$, we see that at the heart of GMRES, we have an ordinary least squares problem. Further, note that since $x_m \in x_0 + K_m$, we can write $x_m = x_0 + \mathcal{K}_m c$ for some $c \in \mathbb{C}^m$. Then our least squares problem has the form

$$\min_{x_m} ||Ax_m - b|| = \min_{c} ||A(x_0 + \mathcal{K}_m c) - b|| = \min_{c} ||A\mathcal{K}_m c - r_0|| \tag{3.3.1}$$

There are many ways we could solve this least squares problem. A naive approach would be to solve the normal equations:

$$(A\mathcal{K}_m)^*(A\mathcal{K}_m)c = (A\mathcal{K}_m)^* r_0.$$

We could also solve (3.3.1) by directly applying QR or SVD factorizations. Each of these methods is summarized in Chapter 11 of [48]. There it is shown that each of these require $O(n^3)$ operations, which is the same amount of work used in solving the system directly via Gaussian elimination. In practical applications, this is too expensive. So instead, we use the Arnoldi iteration and exploit the Hessenburg form of the matrix $H_{m+1,m}$. At each step m, we construct a matrix Q_m whose columns are orthonormal and span K_m. So we write $x_0 + Q_m y = x_m$ instead of $x_0 + \mathcal{K}_m c = x_m$ and (3.3.1) becomes

$$\min_y ||AQ_m y - r_0||.$$

Now, at the m^{th} step of the Arnoldi iteration, we have the the relation (3.2.1). Therefore,

$$\begin{aligned}\min_y ||AQ_m y - r_0|| &= \min_y ||Q_{m+1}H_{m+1,m}y - r_0|| \\ &= \min_y ||H_{m+1,m}y - Q_{m+1}^* r_0|| = \min_y || \ H_{m+1,m}y - ||r_0||e_1 \ ||,\end{aligned}$$

where e_1 denotes the first unit $m+1$ vector. The penultimate equality holds since both $Q_{m+1}H_m y$ and r_0 are in the column space of Q_{m+1} (since we take $q_1 = r_0/||r_0||$). So multiplying by Q_{m+1}^* does not change the norm[1]. The last equality holds since $r_0 = ||r_0||q_1$.

Once y is found, the approximate solution is then given by $x_m = x_0 + Q_m y$. The following is a listing of the GMRES algorithm.

1: Choose x_0, $\quad r_0 = b - Ax_0$, $\quad$ Initialize $q_1 = r_0/||r_0||$, $\quad Q_1 = q_1$.

[1] If $x = Q_{m+1}c$, then it is easily verified that $||x|| = ||c||$. Also, $Q_{m+1}^* x = c$ and so $||Q_{m+1}^* x|| = ||c|| = ||x||$.

2: **for** $j = 1, 2, 3, ...$ **do**

3: Orthogonalize Aq_j against the previous q_i, form $H_{j+1,j}$ and set $Q_{j+1} = [Q_j|q_{j+1}]$.

4: Solve $\min_y || H_j y - ||r_0||e_1 ||$, call the solution y_j.

5: Set $x_j = x_0 + Q_j y_j$.

6: **end for**

At line 4, note that we still have to solve a least squares problem. However, due to the Hessenburg structure of H_j, this can be done in much shorter time than the original problem would require, see [23, Section 2.4] for details.

As mentioned above, GMRES has found the solution to the system $Ax = b$ when the vector v in the Arnoldi iteration equals zero. If it so happens that $r_0 = 0$, then $Q_1 = 0$ and it follows that x_0 is the solution. Now suppose $||r_0|| > 0$ and for some $j > 0$ we have that $v = 0$. In this case, the last row of $H_{j+1,j}$ is zero. Let H_j be $H_{j+1,j}$ without its last row and note that H_j is $n \times n$. Then (3.2.1) simplifies to

$$AQ_j = Q_j H_j. \tag{3.3.2}$$

This equation implies that $Aq_j \in \text{Span}\{q_1, q_2, ..., q_j\}$, which means that $Aq_j \in K_j$ by (3.2.3). From this we can infer that $AK_j \subset K_j$, that is, K_j is an invariant subspace of A. It follows that any eigenvalue of H_j is an eigenvalue of A. Since A is nonsingular, we get that H_j is also nonsingular, so the least squares problem in line 4 now reduces to a nonsingular linear system $H_j y = ||r_0||e_1$. Denote the solution by y_j. Then by (3.3.2), we get

$$A(x_0 + Q_j y_j) = Ax_0 + Q_j H_j y_j = Ax_0 + ||r_0||Q_j e_1 = Ax_0 + r_0 = b$$

and so $x_j = x_0 + Q_j y_j$ is the solution.

The above explanation assumed that GMRES is run to completion. In practice, however,

the process is stopped as soon as an iterate satisfies whatever convergence criterion is set by the user.

Remark 3.3.1. The m^{th} iterate in GMRES, x_m, is taken from the affine Krylov subspace $x_0 + K_m(A, r_0)$, and thus we have

$$x_m = x_0 + c_0 r_0 + c_1 A r_0 + c_2 A^2 r_0 + \cdots + c_{m-1} A^{m-1} r_0.$$

If we let $q_m(z) = c_0 + c_1 z + c_2 z^2 + \cdots + c_{m-1} z^{m-1}$, then we can write

$$x_m = x_0 + q_m(A) r_0.$$

The corresponding residual is given by

$$r_m = b - A x_m = r_0 - A q_m(A) r_0 = (I - A q_m(A)) r_0 = p_m(A) r_0,$$

where $p_m(z) = 1 - z q_m(z)$. Let P_m denote the space of all polynomials with degree less than or equal to m and normalized to equal 1 at the origin. Since at each step $m = 1, 2, 3, ...$, GMRES minimizes the norm of the residual, r_m, we have

$$||r_m|| = \min_{p_m \in P_m} ||p_m(A) r_0||. \tag{3.3.3}$$

In other words, GMRES chooses the coefficients $c_0, ..., c_{m-1}$ so that the quantity $||p_m(A) r_0||$ is minimized over the Krylov subspace $K_m(A, r_0)$.

Chapter 4. Convergence Bounds for GMRES

There is no lack of existing research in the analysis of GMRES convergence and the convergence of related methods, such as CG and Orthomin [4, 5, 16, 17, 20, 41]. Recall from the above discussion GMRES solves the following approximation problem

$$||r_k|| = \min_{p_k \in P_k} ||p_k(A)r_0||,$$

where P_k is the set of all polynomials of degree k or less over $\mathbb{C}$ satisfying $p(0) = 1$. The initial residual can complicate the analysis, so we remove it via the bound

$$||r_k|| = \min_{p_k \in P_k} ||p_k(A)r_0|| \leq \min_{p_k \in P_k} ||p_k(A)||\ ||r_0||. \tag{4.0.1}$$

So now the task of approximating the k^{th} relative residual can be stated in terms of the following matrix approximation problem:

$$\frac{||r_k||}{||r_0||} \leq \min_{p_k \in P_k} ||p_k(A)||. \tag{4.0.2}$$

It may seem that we have taken a big jump in employing this bound. Perhaps there is no r_0 such that equality is attained in (4.0.1). Indeed, in [46], it is shown that for some nonnormal problems, this is precisely the case. However, such problems are thought to be rare in practice [45]. Thus we are content with considering $||p_k(A)||$, and estimating this quantity is the typical strategy in determining the size of the k^{th} relative residual. Greenbaum and Trefethen have named this problem the *Ideal GMRES* problem [25]. In this chapter, we briefly review some previous results done in approximating the right-hand side of (4.0.2), and then introduce how we feel the Crouzeix conjecture could be used to obtain new, and

hopefully better, convergence bounds.

4.1 Previous Results

If we assume A is normal, i.e. $A = UDU^*$, then we have

$$\min_{p_k \in P_k} ||p_k(A)|| = \min_{p_k \in P_k} ||Up_k(D)U^*|| = \min_{p_k \in P_k} \max_{\lambda \in \sigma(A)} |p_k(\lambda)|.$$

This equation can also be applied to non-normal matrices. By the spectral mapping theorem, for any $p \in P_k$, $p(\sigma(A)) = \sigma(p(A))$. Therefore,

$$\max_{\lambda \in \sigma(A)} |p_k(\lambda)| = r(p_k(A)) \leq ||p_k(A)||.$$

So we obtain a lower bound for the Ideal GMRES problem. On the other hand, for a general $A \in M_n$, if we suppose that A is diagonalizable, i.e. $A = VDV^{-1}$, then we have the upper bound

$$\frac{||r_k||}{||r_0||} = \min_{p_k \in P_k} ||Vp_k(D)V^{-1}|| \leq \kappa(V) \min_{p_k \in P_k} ||p_k(D)||,$$

where $\kappa(V) = ||V||||V^{-1}||$ denotes the condition number of V. Therefore,

$$\frac{||r_k||}{||r_0||} \leq \kappa(V) \min_{p_k \in P_k} \max_{i=1,...,n} |p_k(\lambda_i)|. \tag{4.1.1}$$

If A is normal, then $\kappa(V) = 1$, and the lower and upper bounds for Ideal GMRES are equal. In this case, this bound is sharp, meaning that for each k, there exists a $p_k \in P_k$ and an $i \in \{1, ..., n\}$ such that the inequality is an equality [21]. Furthermore, this also shows that for normal matrices, the question of GMRES convergence can be answered by only considering the spectrum. Now, all that we need to be concerned with is how small can the remaining quantity on the right hand side of (4.1.1) be? For a normal matrix A, if

the eigenvalues are tightly clustered about some nonzero point $c \in \mathbb{C}$, then we can consider the polynomial $p_k(z) = (1 - z/c)^k$. Note that the norm of this polynomial is small at points close to c, and so this implies that GMRES will converge quickly in this case. On the other hand, if the eigenvalues are spread all around the origin, then we have a worst case scenario. This is because we cannot have a polynomial $p(z)$ which satisfies $p(0) = 1$ and $|p(z)| < 1$ on some curve surrounding the origin (this follows from the maximum principle). This implies that the residual will not be reduced until step n, which is when the iteration has run to completion. If A is not normal, but V is well-conditioned, then (4.1.1) can still give a fairly good estimate on the size of the relative residual, and the intuition just given concerning eigenvalue distribution can still be a good indicator of the convergence behavior of GMRES. However, for general matrices, the convergence rate of GMRES does not depend on eigenvalues alone. In the paper by Greenbaum, Pták, and Strakoš [24] it is shown that given any non-increasing curve, there is a problem whose GMRES residuals plotted against the iterations is given by that curve. Furthermore, that problem can be chosen to have any eigenvalues. This means that, in general, we need to consider more than merely the spectrum of A.

Another method of estimating $||p(A)||$ has been introduced by Trefethen [47, 49] and involves using the pseudospectrum of A. Given $\varepsilon > 0$, the ε-pseudospectrum of A is defined to be the set

$$\Lambda_\varepsilon = \{z \in \mathbb{C} \mid ||(zI - A)^{-1}|| \geq 1/\varepsilon\}.$$

The points $z \in \mathbb{C}$ for which $||(zI-A)^{-1}|| = 1/\varepsilon$ make up the boundary of the ε-pseudospectrum. A bound on $||p(A)||$ can be obtained as follows. For any polynomial p, we can write

$$p(A) = \frac{1}{2\pi i} \int_\gamma p(z)(zI - A)^{-1} dz,$$

where γ is some simple closed curve containing the spectrum of A in its interior. Let $L(\gamma)$

denote the length of γ. Then by taking norms, we have

$$||p(A)|| \leq \frac{L(\gamma)}{2\pi} \sup_{z \in \gamma} ||p(z)(zI - A)^{-1}||.$$

Now, if we replace γ with the boundary of the ε-pseudospectrum, denoted γ_ε, for some $\varepsilon > 0$, then this last inequality now becomes

$$||p(A)|| \leq \frac{L(\gamma_\varepsilon)}{2\pi\varepsilon} \sup_{z \in \gamma_\varepsilon} |p(z)|.$$

Thus for GMRES, we have

$$\frac{||r_k||}{||r_0||} \leq \frac{L(\gamma_\varepsilon)}{2\pi\varepsilon} \sup_{z \in \gamma_\varepsilon} |p(z)|.$$

Psuedospectral bounds are dealt with in more detail in [49]. Note that this psuedospectral bound can be adjusted by varying the parameter ε. A large ε yields a small $L(\gamma_\varepsilon)/2\pi\varepsilon$, but the approximation of the quantity $|p(z)|$ is over a large domain, which may include the origin. This can be problematic, since $p(0) = 1$. Thus the maximum on the boundary must be larger than 1 and can potentially be much larger. A small ε will lead to a large constant $L(\gamma_\varepsilon)/2\pi\varepsilon$, but with the approximation over a small domain. This feature of the psuedospectral bound can be used to describe different phases of GMRES convergence. In his book, [49], Trefethen describes that for nonnormal GMRES problems, one often observes initial stagnation followed by more rapid convergence later. Thus, large values of the parameter ε can be descriptive for the early stage, and small values of ε can be descriptive of the later stage. However, these bounds fail to describe the intermittent convergence behavior, and sometimes can fail to be very descriptive at all [18, 22]. Thus we hope to gain greater insight by deriving bounds obtained by considering the numerical range.

Some work preceding Crouzeix's conjecture in the area of GMRES analysis involving the numerical range includes the papers by Eiermann [14, 15]. For a simple example, consider

the case $0 \notin W(A)$ and $W(A) \subset D = \{z \in \mathbb{C} \mid |z - c| \leq s\}$, where D also does not contain zero. Again, let $p_k(z) = (1 - z/c)^k$. By Proposition 2.1.1, we have

$$W(I - (1/c)A) = 1 - (1/c)W(A) \subset \{z \in \mathbb{C} \mid |z| \leq s/|c|\}.$$

From this, we have that the numerical radius $w(I - (1/c)A)$ is bounded by $s/|c|$. Then, by the power inequality (Theorem 2.4.8),

$$w(p_k(A)) = w((I - (1/c)A)^k) \leq w(I - (1/c)A)^k \leq (s/|c|)^k,$$

and using Theorem 2.4.3, we have

$$||p_k(A)|| \leq 2w(p_k(A)) = 2w((I - (1/c)A)^k) \leq 2(s/|c|)^k.$$

So for GMRES, we have the inequality

$$\frac{||r_k||}{||r_0||} \leq 2(s/|c|)^k.$$

What this tells us is that if $0 \notin W(A)$, and if $W(A)$ is small and positioned far away from the origin, then we can expect GMRES to converge quickly. So while this can be a good bound, it has some severe restrictions. For instance, if $s \geq |c|$, then this bound will be a gross overestimate of the actual norm of the relative residual. Also it requires that $0 \notin W(A)$, which we cannot guarantee for all nonsingular matrices.

Crouzeix's conjecture offers a possible alternative method for using the numerical range for GMRES analysis.

4.2 Crouzeix's Conjecture Applied to GMRES

Recall that by Proposition 2.2.5 the numerical range of A contains the spectrum of A, i.e. $\sigma(A) \subset W(A)$. Therefore, for diagonalizable A,

$$\frac{||r_k||}{||r_0||} \leq \min_{p_k \in P_k} ||p_k(A)|| = \kappa(V) \min_{p_k \in P_k} \max_{\lambda \in \sigma(A)} |p(\lambda)| \leq \kappa(V) \min_{p_k \in P_k} \sup_{z \in W(A)} |p(z)|,$$

Now, if the Crouzeix conjecture (1.0.2) is true, we will have

$$\frac{||r_k||}{||r_0||} \leq \min_{p_k \in P_k} ||p(A)|| \leq 2 \min_{p_k \in P_k} \sup_{z \in W(A)} |p(z)|. \tag{4.2.1}$$

One obvious advantage here is that we no longer have to worry about the condition number $\kappa(V)$, which varies for each matrix. Furthermore, this bound can be applied to any matrix, not just those with well-conditioned eigenvector matrices. Also, we are now considering the size of $|p(z)|$ over the numerical range rather than the spectrum, which we know is insufficient in general. The only issue now is to estimate the minimal supremum norm of a k degree polynomial over the numerical range. In some cases, the solution of this problem is well-known. These include symmetric (positive definite and indefinite) and some certain nonsymmetric problems, see [47] and the references therein. The challenge lies in estimating this quantity for general problems. There are several avenues to explore along this vein. These include minimum norm Pick-Nevanlinna interpolation, least squares approximation, orthogonal polynomials, and interpolation in Fejér points. In this work, we examine bounds obtained by considering the Faber polynomials, which will be introduced in the next chapter.

Chapter 5. Conformal Maps and Faber Polynomials

The purpose of this chapter is to give the necessary background in conformal maps and Faber polynomials to facilitate the discussion on convergence bounds for GMRES. Conformal maps have many applications in areas such as numerical analysis, mesh generation, electrostatics, and fluid mechanics. The interested reader may consult [13] for more information on the applications of these maps. Faber polynomials were introduced in 1903 by Faber [19] to solve the problem of approximating a function, analytic in a given region, by polynomials which do not depend on the function to be approximated. We will start with a brief review of Laurent series.

5.1 Laurent Series

In this section, we will provide some basic results and notation that will be of use in the following sections. We assume the reader is already familiar with these results and refer to other sources for some of the proofs.

Theorem 5.1.1. *Let f be analytic in the annulus $r < |z - z_0| < R$. Then f can be expressed as the sum of two series*

$$f(z) = \sum_{j=0}^{\infty} a_j (z - z_0)^j + \sum_{j=1}^{\infty} a_{-j} (z - z_0)^{-j} \tag{5.1.1}$$

both series converging absolutely in the annulus, and converging uniformly in any closed subannulus. The coefficients a_j are given by

$$a_j = \frac{1}{2\pi i} \int_C \frac{f(\zeta)}{(\zeta - z_0)^{j+1}} d\zeta, \qquad j = 0, \pm 1, \pm 2, ...$$

where C is any positively-oriented simple closed contour lying in the annulus and containing z_0 in its interior. The result also holds for $r = 0$, $R = \infty$, or both.

The representation (5.1.1) is called the Laurent series, or Laurent expansion, of f about z_0. It turns out that any pointwise convergent expansion of this form equals the Laurent expansion, i.e., the Laurent expansion is unique. A proof of this along with Theorem 5.1.1 can be found in [38].

If f is analytic on a region that contains the punctured disk $0 < |z - z_0| < R$, then the point z_0 is called an *isolated singularity*. In this case, Theorem 5.1.1 is valid on $0 < |z - z_0| < R$ and in particular, we can write

$$f(z) = \sum_{j=-\infty}^{\infty} a_j (z - z_0)^j, \quad 0 < |z - z_0| < R. \tag{5.1.2}$$

One of the advantages of the Laurent expansion is that we can use it to classify the singularities of the function it represents. Recall the following definition:

Definition 5.1.2. Let f be analytic in a region A that contains a deleted neighborhood of the point z_0, so that z_0 is an isolated singularity. Let the Laurent expansion of f be given by (5.1.2).

(i) If $a_j = 0$ for all integers $j < 0$, we say that z_0 is a *removable singularity* of f.

(ii) If all but a finite number of the a_j, $j < 0$, are zero, then z_0 is called a *pole* of f. If k is the largest integer such that $a_{-k} \neq 0$, then we say that z_0 is a *pole of order k*. If $k = 1$, we say that z_0 is a *simple pole*.

(iii) If an infinite number of the a_j, $j < 0$, are nonzero, then z_0 is called an *essential singularity* of f.

(iv) The coefficient a_{-1} is the *residue* of f at z_0.

Now let $\hat{\mathbb{C}} = \mathbb{C} \cup \{\infty\}$ denote the extended complex plane, that is, the complex plane $\mathbb{C}$ along with the point at infinity. When dealing with the extended complex plane, we define $1/0 = \infty$ and $1/\infty = 0$. It turns out that for a function analytic in a neighborhood of infinity we can also define a Laurent expansion at infinity, with a result analogous to Theorem 5.1.1. First, let

$$D(\infty, r) = \{z \mid z \in \mathbb{C}, |z| > r\} \quad \text{and} \quad D'(\infty, r) = D(\infty, r) \cup \{\infty\}$$

be the punctured disk with center at infinity and the disk centered at infinity, respectively. Similarly, define $D(0, 1)$ to be the open unit disk centered at 0. We define a *neighborhood of infinity* to be $\mathcal{O} \cup \{\infty\}$ where $\mathcal{O}$ is open set containing $D(\infty, r)$ for some $r > 0$.

Theorem 5.1.3. *Let $\mathcal{O}$ be a neighborhood of infinity containing $D(\infty, R)$, $R \geq 0$ and let $f : \mathcal{O} \backslash \{\infty\} \to \mathbb{C}$ be analytic on $\mathcal{O} \backslash \{\infty\}$. Then there exists a series representation of f, given by*

$$f(z) = \sum_{k=0}^{\infty} b_k z^{-k} + \sum_{k=1}^{\infty} c_k z^k, \tag{5.1.3}$$

where both series converge absolutely on $D(\infty, R)$ and uniformly on any compact subset of $D(\infty, R)$. The coefficients are given by

$$b_k = \frac{1}{2\pi i} \int_{\gamma_\rho} f(z) z^{k-1} dz, \quad k = 0, 1, 2, ...$$
$$c_k = \frac{1}{2\pi i} \int_{\gamma_\rho} \frac{f(z)}{z^{k+1}} dz, \quad k = 1, 2, ...$$

where γ_ρ is any positively-oriented circle centered at zero with radius $\rho > R$.

A similar uniqueness result as in Theorem 5.1.1 also holds in this case. A proof can be found in [11].

Since the map $w := 1/z$ maps the point at infinity to 0, we can use the composite

function $g(w) := f(1/w)$ to classify the singularities of f at infinity.

(i) $f(z)$ has a removable singularity at ∞ if $f(1/w)$ has a removable singularity at $w = 0$.

(ii) $f(z)$ has a pole of order k at ∞ if $f(1/w)$ has a pole of order k at $w = 0$.

(iii) $f(z)$ has an essential singularity at ∞ if $f(1/w)$ has an essential singularity at $w = 0$.

The final result we need is Cauchy's coefficient estimate. This provides a bound on the coefficients of a convergent power series.

Proposition 5.1.4. *Let f be analytic on a region E containing the disk $\{z \mid |z - z_0| < \rho(f)\}$ with a power series representation*

$$f(z) = \sum_{k=0}^{\infty} a_k (z - z_0)^k,$$

with radius of convergence $\rho(f)$. Then for any $0 \leq \rho < \rho(f)$, we have

$$|a_k| \leq \frac{\sup_{|z|=\rho} |f(z)|}{\rho^k}.$$

Proof. If $\rho = 0$, then $|a_k| \leq \infty$ and there is nothing to show. Let γ be a circle of radius $0 < \rho < \rho(f)$ centered at z_0. Recall from Cauchy's integral formula that

$$f^{(k)}(z_0) = \frac{k!}{2\pi i} \int_\gamma \frac{f(z)}{(z - z_0)^{k+1}} dz.$$

Also, since f is analytic on and inside γ, it is continuous there and so for some $M > 0$ we have

$$\sup_{\{z \ \mid \ |z - z_0| \leq \rho\}} |f(z)| \leq M$$

Therefore,

$$|f^{(k)}(z_0)| \leq \frac{k!}{2\pi} \int_\gamma \left| \frac{f(z)}{(z - z_0)^{k+1}} \right| dz \leq \frac{k!}{2\pi} \frac{M}{\rho^{k+1}} 2\pi\rho = \frac{k! M}{\rho^k},$$

since the length of γ is $2\pi\rho$. But from the power series representation for f, we have

$$f^{(k)}(z_0) = k!a_k,$$

Thus

$$|a_k| \leq \frac{M}{\rho^k}.$$

□

Corollary 5.1.5. *Let f be analytic on a neighborhood of infinity $\mathcal{O}$ with Laurent expansion at infinity given by*

$$f(z) = \sum_{k=0}^{\infty} a_k z^{-k},$$

converging on $D(\infty, R(f)) \subset \mathcal{O}$, where $R(f) > 0$. Then for any $\sigma > 0$ satisfying $0 < 1/\sigma < 1/R(f)$, we have

$$|a_k| \leq \sigma^k \sup_{|z|=\sigma} |f(z)|$$

Proof. Let

$$g(w) = f(1/w) = \sum_{k=0}^{\infty} a_k w^k.$$

This is a Laurent expansion about 0, convergent for all $|w| < 1/R(f)$. By Proposition 5.1.4, for any σ satisfying $0 < 1/\sigma < 1/R(f)$, we have

$$|a_k| \leq \frac{\sup_{|w|=1/\sigma} |g(w)|}{(1/\sigma)^k}.$$

But $w = 1/z$, so

$$|a_k| \leq \sigma^k \sup_{|z|=\sigma} |f(z)|$$

□

Example 5.1.6. Let $\mathcal{O}$ be a neighborhood of ∞ containing $D(\infty, R)$ for some $R > 0$, and

let f be analytic on $\mathcal{O}\backslash\{\infty\}$ and have a simple pole at ∞. By Theorem 5.1.3, we can write f in the form of (5.1.3). Since f has a simple pole at infinity, it follows that $f(1/w)$ has a simple pole at zero, so in particular, we have that

$$f(z) = \sum_{k=0}^{\infty} b_k z^{-k} + c_1 z, \quad z \in D(\infty, R).$$

Since f converges almost uniformly on $D(\infty, R)$, we can differentiate term-by-term to obtain

$$f'(z) = \sum_{k=0}^{\infty} -k b_k z^{-k-1} + c_1, \quad z \in D(\infty, R).$$

Thus $f'(z)$ has a removable singularity at ∞. To determine c_1, we again use the reciprocation $z = 1/w$. First, we have

$$f'(1/w) = \sum_{k=0}^{\infty} -k b_k w^{k+1} + c_1.$$

Also,

$$\frac{f(1/w)}{1/w} = \sum_{k=0}^{\infty} b_k w^{k+1} + c_1.$$

Both of these series are clearly continuous and uniformly convergent in any closed subdisk of $D(0, 1/R)$. Therefore,

$$\lim_{w\to 0} f'(1/w) = \lim_{w\to 0} \frac{f(1/w)}{1/w} = c_1 = \lim_{z\to\infty} \frac{f(z)}{z} = \lim_{z\to\infty} f'(z).$$

Thus we can define $f'(\infty) = c_1$, which is finite. So the Laurent expansion at infinity for f can be written

$$f = f'(\infty) z + b_0 + b_1 z^{-1} + b_2 z^{-2} + \cdots, \quad z \in D(\infty, R) \tag{5.1.4}$$

5.2 Conformal Maps

Faber polynomials are generated from specific analytic functions, which are *conformal mappings*. It turns out that any analytic function is conformal at a point where its derivative is nonzero, but the ones from which the Faber polynomials are derived have specific series representations and it is helpful to see how these forms arise.

Definition 5.2.1. A map $f : E \to C$ is *conformal at* z_0 if there exists a $\theta \in [0, 2\pi)$ and an $r > 0$ such that for any curve $\gamma(t) \in E$ that is differentiable at $t = t_0$, $\gamma(t_0) = z_0$, and $\gamma'(t_0) \neq 0$, the curve $\sigma(t) = f(\gamma(t))$ is differentiable at $t = t_0$ and, setting $u = \sigma'(t_0)$ and $v = \gamma'(t_0)$, we have $|u| = r|v|$ and $\arg u = (\arg v + \theta) \mathrm{mod}\ 2\pi$. A map is called *conformal* if it conformal at every point.

What this definition says is that a conformal map stretches and rotates tangent vectors to curves. Therefore, such a map will also preserve angles between curves.

Example 5.2.2. Let $z_0 = 1 + i$ be fixed and consider the curves $\gamma_1(t) = \sqrt{2}e^{it}$ and $\gamma_2(t) = 1 + it^2$. Let $f(z) = z^2$. We will show that the angles between the tangents to γ_1 and γ_2 at z_0 are preserved under this map. First, note that $\gamma_1(\pi/4) = 1 + i$ and $\gamma_2(1) = 1 + i$. Also,

$$\gamma_1'(t) = \sqrt{2}ie^{it} \quad \text{and} \quad \gamma_2'(t) = 1 + 2it,$$

so

$$\gamma_1'(\pi/4) = i - 1 \quad \text{and} \quad \gamma_2'(1) = 1 + 2i.$$

These latter values specify the direction of the tangent vectors, i.e. if T_1 and T_2 are the tangents to γ_1 and γ_2 at $z_0 = 1 + i$ respectively, then

$$T_1(t) = 1 + i + (i - 1)t \quad \text{and} \quad T_2(t) = 1 + i + (1 + 2i)t,$$

where $\arg T_1 = 3\pi/4$ and $\arg T_2 = \tan^{-1}(2) \approx 1.107$. The angle between the two vectors is

given by the difference of these two arguments and is approximately 1.249 radians.

Now consider the images of γ_1 and γ_2 under $f(z) = z^2$. We have

$$\mu_1(t) = f(\gamma_1(t)) = 2e^{2it} \quad \text{and} \quad \mu_2(t) = f(\gamma_2(t)) = (t + it^2)^2.$$

Using the chain rule, we obtain

$$\mu_1'(t) = 4ie^{2it} \quad \text{and} \quad \mu_2'(t) = 2(t + it^2)(1 + 2it),$$

and so

$$\mu_1'(\pi/4) = -4 \quad \text{and} \quad \mu_2'(1) = -2 + 6i.$$

Therefore, if S_1 and S_2 are the tangents to μ_1 and μ_2 at $f(z_0) = 2i$, respectively, then arg $S_1 = \pi$ and arg $S_2 = \tan^{-1}(-3) + \pi$. The difference between these angles is $-\tan^{-1}(-3) \approx 1.249$ radians. So we see that the angle between these two tangents has been preserved under the map f.

We mentioned above that any analytic function with a nonzero derivative at a point z_0 is conformal at z_0. Here is that claim stated as a proposition with proof.

Proposition 5.2.3. *An analytic function f is conformal at every point z_0 for which $f'(z_0) \neq 0$.*

Proof. Using the notation of Definition 5.2.1, we have

$$u = \sigma'(0) = f'(\gamma(0))\gamma'(0) = f'(z_0)v.$$

Therefore

$$\arg u = (\arg \mathrm{v} + \arg f'(z_0)) \bmod 2\pi \quad \text{and} \quad |u| = |f'(z_0)||v|.$$

So taking $\theta = \arg f'(z_0)$ and $r = |f'(z_0)|$ in Definition 5.2.1 gives the result. □

Note that this result implies, by the Inverse Function Theorem, that any map that is conformal at a point z_0 is also one-to-one in a neighborhood of z_0. When speaking of functions of a complex variable, we often say that a one-to-one function is *univalent*.

Example 5.2.4. The function $f(z) = z^4$ has derivative $f'(z) = 4z^3$ and so by Proposition 5.2.3 it is conformal at every point in $\mathbb{C}$ except for the origin. Let $R = \{z \in \mathbb{C} \mid \text{Re } z \geq 1 \text{ and Im } z \geq 1\}$. Figure 5.1(a) contains a picture of the region R, with grid lines that meet at right angles. Figure 5.1(b) is the image of this region under the map $f(z) = z^4$. A close up of the origin is shown in Figure 5.1(c). It can be seen in these figures that the right angles have been preserved.

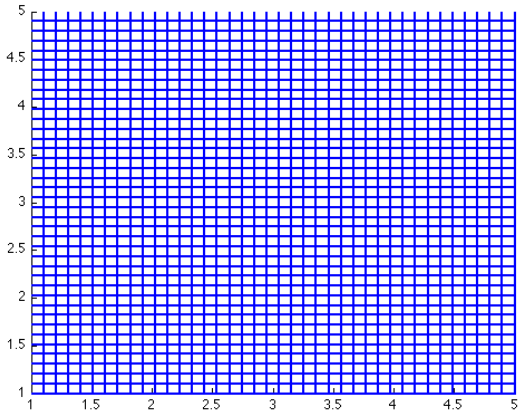
Example 5.2.5. The function $f(z) = \frac{1}{2}(z + (1/z))$ is known as the *Joukowski Map* (Add citation). It maps a circle with radius larger than 1 conformally onto an ellipse [39]. In Figure 5.2, we have shown a portion of the exterior of the unit disk, namely the region $C = \{z \in \mathbb{C} \mid 0.1 \leq |z| \leq 2\}$, and its image under the Joukowski Map.

The next theorem is the famous *Riemann Mapping Theorem.*

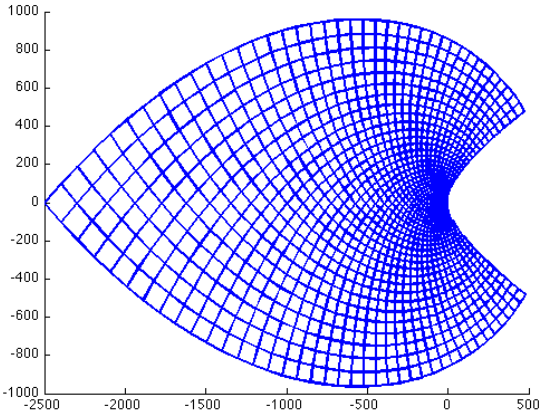
Theorem 5.2.6. *Let E be a simply connected, open, and proper subset of $\mathbb{C}$. Then there exists a one-to-one conformal map f which maps E onto the open unit disk. For a fixed point $z_0 \in E$, we can also require that $f(z_0) = 0$ and $f'(z_0) > 0$. Under these additional specifications, the mapping f is unique.*

The proof is quite involved and can be found in [11]. The Riemann mapping theorem can be used to prove the existence of conformal mappings between certain regions other than those specified in the statement of the theorem. The case we are concerned with is stated in the following theorem.

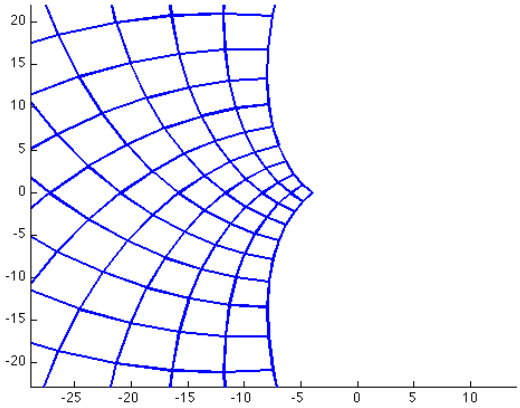
Theorem 5.2.7. *Let $E \subset \mathbb{C}$ be a compact, simply connected set containing more than one point, so that $E^c = \hat{\mathbb{C}} \backslash E$ is connected in $\hat{\mathbb{C}}$. There exists a unique conformal map $\psi : D'(\infty, 1) \to \hat{\mathbb{C}}$ which has the following properties:*



(a) The region R with grid lines

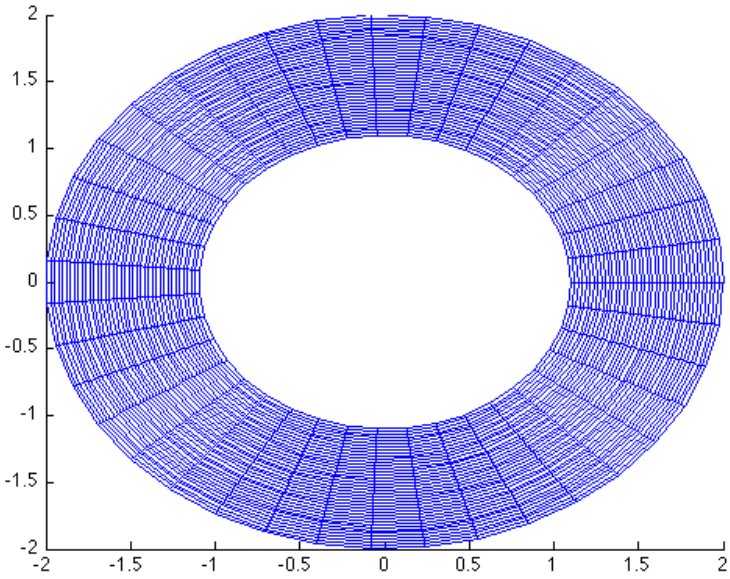


(b) The image of R of Example 5.2.4 under the map $f(z) = z^4$

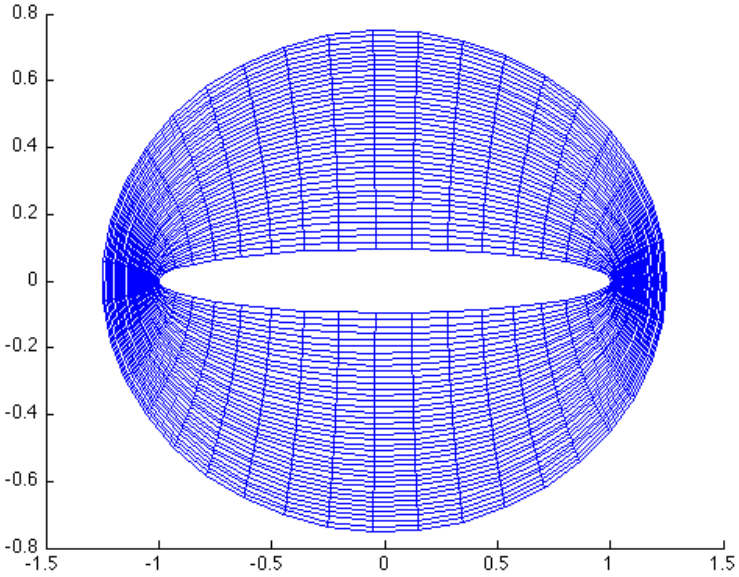


(c) Close up of the origin

Figure 5.1: Example of a Conformal Map



(a) A portion of the exterior of the unit disk



(b) The image of the region C of example Example 5.2.5 under the Joukowski Map

Figure 5.2: Example of the Joukowski Map

(i) ψ *is univalent on* $D'(\infty, 1)$ *and analytic on* $D(\infty, 1)$.

(ii) ψ *has a simple pole at* ∞ *with* $\psi'(\infty) > 0$.

(iii) ψ *maps* $D(\infty, 1)$ *onto* $E^c \backslash \{\infty\}$.

Furthermore, the inverse function $\phi = \psi^{-1}$ *also has a simple pole at* ∞ *with* $\phi'(\infty) = 1/\psi'(\infty) > 0$.

Proof. Let $a \in E$ and consider the mapping $\zeta = h(z) = 1/(z - a)$. This map takes $z = \infty$ to $\zeta = 0$ and $z = a$ to $\zeta = \infty$. Under this definition, this map is a homeomorphism of $\hat{\mathbb{C}}$ onto $\hat{\mathbb{C}}$ with inverse $z = h^{-1}(\zeta) = a + 1/\zeta$. Define B to be the image of E^c under the map h. Thus $0 \in B$, but $\infty \notin B$, since $a \notin E^c$. Since the inverse is continuous, and E^c is open and connected, it follows that B itself is open and connected. Furthermore, the preimage of $\hat{\mathbb{C}} \backslash B$ under h is the connected set E and so $\hat{\mathbb{C}}/B$ is connected. This shows that B is in fact simply connected, so the Riemann Mapping Theorem applies.

Let $\eta = f(\zeta)$ be the unique conformal mapping guaranteed by Theorem 5.2.6, which maps B onto $D(0, 1)$, with $f(0) = 0$ and $f'(0) > 0$. The inverse function $f^{-1} : D(0, 1) \to B$ is also a conformal mapping with $f^{-1}(0) = 0$ and $f'^{-1}(0) = 1/f'(0) > 0$. Note that this implies the zero of f^{-1} at $\eta = 0$ is a simple zero.

Now define the function $z = X(\eta) = a + [1/f^{-1}(\eta)]$. By the properties of f^{-1}, $X(\eta)$ is analytic and univalent on $D(0, 1) \backslash \{0\}$ (the punctured unit disk), and has a simple pole at $\eta = 0$. Also, by comparing the form of X with that of h^{-1} above, we have that X maps $D(0, 1)$ onto E^c. In particular, $\eta = 0$ maps to $z = \infty$. Now let $w = 1/\eta$. The resulting function, given by $z = \psi(w) = X(1/w)$, is analytic and univalent in $D(\infty, 1)$ and has a simple pole at $w = \infty$. By Example 5.1.6, we can compute $\psi'(\infty)$ as follows:

$$\psi'(\infty) = \lim_{w \to \infty} \frac{\psi(w)}{w} = \lim_{\eta \to 0} \eta X(\eta) = \lim_{\eta \to 0} [\eta a + (\eta / f^{-1}(\eta))] = f'(0) > 0,$$

where the last equality follows from an application of L'Hospital's rule. Let $w = \phi(z) = \psi^{-1}(z)$. Since $\psi'(\infty) > 0$, we can conclude that for large enough $R > 0$, $\phi(z)$ is analytic in $D(\infty, R)$. The same also holds for $\phi(z)/z$ and so

$$\lim_{z\to\infty} \frac{\phi(z)}{z} = \lim_{w\to\infty} \frac{w}{\psi(w)} = 1/\psi'(\infty).$$

This implies that ϕ has a simple pole at ∞ (see [11, Proposition 10.7.3]). So again by Example 5.1.6, we have that $\phi'(\infty) = 1/\psi'(\infty)$.

The last issue that needs addressing is that of uniqueness. Since $a \in E$ was chosen arbitrarily, we need to show that the resulting function ψ does not depend upon the choice of a. So let ψ_1 be another function, analytic on $D(\infty, 1)$ and univalent on $D'(\infty, 1)$, with range E^c, and having a simple pole at ∞ with $\psi_1'(\infty) > 0$. Let $\phi_1 = \psi_1^{-1}$, and consider the function

$$\eta = F(\zeta) = \frac{1}{\phi_1(\psi(1/\zeta))}, \quad \zeta \in D(0, 1).$$

The function $\phi_1(\psi(1/\zeta))$ is analytic in $D(0, 1)$ and has modulus which becomes arbitrarily large as ζ approaches zero, hence F is bounded and analytic on $D(0, 1)$ and so $\zeta = 0$ is a removable singularity of this function. Defining $F(0) = 0$ yields a function analytic and univalent on $D(0, 1)$ and maps $D(0, 1)$ onto itself. So by Proposition 9.8.1 in [11], $F(\zeta) = e^{i\alpha}\zeta$ in $D(0, 1)$ for some $\alpha \in \mathbb{R}$. It follows that $\phi_1(\psi(1/\zeta)) = e^{-i\alpha}/\zeta$ and so $\psi(1/\zeta) = \psi_1(e^{-i\alpha}/\zeta)$, for $\zeta \in D(0, 1)$. Now if we let $w = 1/\zeta$, we have $\psi(w) = \psi_1(e^{-i\alpha}w)$ where $|w| > 1$. Appealing one more time to Example 5.1.6, we have

$$\psi'(\infty) = \lim_{w\to\infty} \frac{\psi(w)}{w} = \lim_{w\to\infty} \frac{\psi_1(e^{-i\alpha}w)}{e^{-i\alpha}w} e^{-i\alpha} = \psi_1'(\infty)e^{-i\alpha}.$$

But both $\psi'(\infty)$ and $\psi_1'(\infty)$ are real and positive, and since $|e^{-i\alpha}| = 1$, it follows that $e^{-i\alpha} = 1$ and so $\psi(w) = \psi_1(w)$, $|w| > 1$. □

In the following chapters, we will have need to extend the mapping of Theorem 5.2.7 continuously to the boundary of E, in the case where E is also convex. In order to do this, we need the Osgood-Caratheodory theorem, which is stated here. The proof can be found in [29].

Theorem 5.2.8. *Let A_1 and A_2 be two bounded, open, and simply connected subsets of the complex plane whose boundaries ∂A_1 and ∂A_2 are simple, continuous closed curves. Then any conformal map from A_1 to A_2 can be extended to a continuous map of $A_1 \cup \partial A_1$ one-to-one and onto $A_2 \cup \partial A_2$.*

Corollary 5.2.9. *Using the same hypotheses and notation of Theorem 5.2.7, we can extend the map $\phi : E^c \to D(\infty, 1)$ continuously to the boundary of E so that $\phi : E^c \cup \partial E \to \overline{D(\infty, 1)}$ is continuous, one-to-one, and onto.*

Proof. First note that there exists an $a \in \mathbb{C}$ such that $0 \in E + a$ (if $0 \in E$, then take $a = 0$). For $z \in E^c$, map z to w via the map $w = 1/(z + a)$. Thus $w \in \Omega$, where Ω is a bounded, simply connected, open set and by the convexity of E we have that the boundary of Ω is a simple closed, continuous curve. Define $F(w) = 1/(\phi(\frac{1}{w} - a))$. Then it follows that $F(w)$ is a conformal map from Ω one-to-one and onto $D(0, 1)$. Thus by the Osgood-Caratheodory theorem, we can extend F to a continuous map from $\Omega \cup \partial\Omega$ one-to-one and onto $\overline{D(0, 1)}$. Since the boundary of E corresponds to the boundary of Ω under the map $z \mapsto 1/(z+a)$, for each $w \in \partial\Omega$, we have that $F(w) = 1/\phi(\frac{1}{w} - a) = 1/\phi(z) \in \partial\overline{D(0, 1)}$, where $z \in \partial E$. Finally, if $1/\phi(z) \in \partial\overline{D(0, 1)}$, then $|1/\phi(z)| = 1$ and so $|\phi(z)| = 1$, which implies $\phi(z) \in \partial\overline{D(0, 1)}$ as well. Thus the mapping ϕ of Theorem 5.2.7 can be continuously extended to the boundary of E. □

5.3 Faber Polynomials

We are now ready to derive the Faber polynomials. As mentioned above, the discovery of these polynomials was motivated by the desire to express an analytic function in the form $\sum_{n=0}^{\infty} a_n p_n(z)$, on a given region where the coefficients a_n depend on f, but the polynomials only depend on the region in question. Thus we see that the theory of Faber polynomials extends that of Laurent series to regions more general than just a disk or an annulus.

This discussion follows the treatment given in [10]. For an alternative derivation, see [29]. Let $E \subset \mathbb{C}$ be compact and such that E^c is simply connected in $\hat{\mathbb{C}}$. By Theorem 5.2.7, there exists a ψ which maps the exterior of the closed unit disk conformally onto E^c and is analytic and univalent on $D(\infty, 1)$. By Example 5.1.6, we can write

$$\psi(w) = bw + b_0 + \frac{b_1}{w} + \frac{b_2}{w^2} + \cdots, \quad |w| > 1, \tag{5.3.1}$$

where $b = \psi'(\infty) > 0$. The inverse map $\phi : E^c \to D(\infty, 1)$ exists and also by Theorem 5.2.7 and Example 5.1.6, we can conclude that

$$\phi(z) = \frac{z}{b} + c_0 + \frac{c_1}{z} + \frac{c_2}{z^2} + \cdots, \tag{5.3.2}$$

for z outside a sufficiently large circle. Then the n^{th} Faber polynomial associated with ψ (or E) is defined to be the polynomial part of the Laurent expansion at infinity of $[\phi(z)]^n$ (from here on, we will simply write $\phi(z)^n$ to keep the notation uncluttered). By (5.3.2), it is evident that this is a polynomial of degree n with leading term $(z/b)^n$. The following result is of central importance in showing how the Faber polynomials are used in analytic function approximation.

Proposition 5.3.1. *Let $R > 1$ and let $C_R = \{z \mid z = \psi(s), |s| = R\}$. Then if $p_n(z)$,*

$n = 0, 1, 2, ...,$ *are the Faber polynomials for* ψ*, then*

$$\frac{s\psi'(s)}{\psi(s) - z} = \sum_{n=0}^{\infty} \frac{p_n(z)}{s^n}$$

where $z \in$ *Int* C_R *and the series converges absolutely and uniformly for all* $|s| \geq R$*.*

Proof. The current situation is illustrated by Figure 5.3.

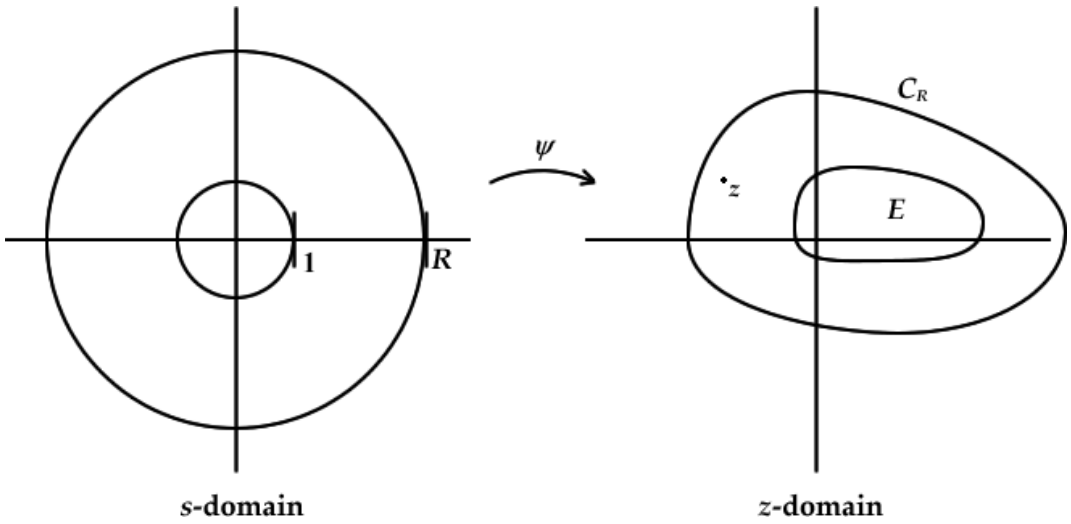


Figure 5.3: Figure for Proposition 5.3.1

Consider the following integral

$$\frac{1}{2\pi i} \int_{C_R} \frac{\phi(t)^n}{t - z} dt.$$

We can deform C_R into a circle C^*, of radius large enough so that the Laurent expansion at infinity of $\phi(z)$ (and thus also $\phi(z)^n$) converges uniformly for z outside C^*. Then we can integrate term-by-term and obtain

$$\begin{aligned} \frac{1}{2\pi i} \int_{C_R} \frac{\phi(t)^n}{t - z} dt = \frac{1}{2\pi i} \int_{C^*} \frac{\phi(t)^n}{t - z} dt = \frac{1}{2\pi i} \int_{C^*} \frac{1}{t - z} (p_n(t) + \mathcal{O}(1/t)) dt \\ = \frac{1}{2\pi i} \int_{C^*} \frac{p_n(t)}{t - z} dt + \frac{1}{2\pi i} \int_{C^*} \frac{\mathcal{O}(1/t)}{t - z} dt, \end{aligned} \tag{5.3.3}$$

where $p_n(t)$ is a polynomial in t of degree n. Since $p_n(t)$ is analytic on and in the interior of C^*, we have that the first integral equals $p_n(z)$ by the Cauchy integral formula. We claim that the second integral equals zero. To see this, first note that the highest degree term in the numerator of this integral is $1/t$. So each term making up the second integral is of the form

$$\frac{1}{2\pi i}\int_{C^*}\frac{1/t^k}{t-z}dt, \quad k \geq 1,$$

up to a multiplicative constant. Second, since $z \in \text{Int } C_R$, we also have that $z \in \text{Int } C^*$, and so $|t| > |z|$. Therefore, we can rewrite this last equation as

$$\frac{1}{2\pi i}\int_{C^*}\frac{1}{t^{k+1}}\frac{1}{1-\frac{z}{t}}dt = \frac{1}{2\pi i}\int_{C^*}\frac{1}{t^{k+1}}\sum_{n=0}^{\infty}\left(\frac{z}{t}\right)^n dt.$$

The highest order term is $1/t^2$ since $k \geq 1$ and so each of these terms must vanish. Therefore, the second integral in the last member of (5.3.3) is indeed equal to zero. At this point we have established the relation

$$p_n(z) = \frac{1}{2\pi i}\int_{C_R}\frac{\phi(t)^n}{t-z}dt, \quad n = 1,2,3,..., \quad z \in \text{Int } C_R, \tag{5.3.4}$$

where $p_n(z)$ is a polynomial of degree n. Now if we let $t = \psi(s)$, we get

$$p_n(z) = \frac{1}{2\pi i}\int_{|s|=R}\frac{s^n\psi'(s)}{\psi(s)-z}ds, \quad n = 1,2,3,..., \quad z \in \text{Int } C_R. \tag{5.3.5}$$

Now, since $z \in \text{Int } C_R$, it follows that $\psi(s) - z$ does not vanish for any $|s| = R$ and so the function $s\psi'(s)/(\psi(s) - z)$ is analytic for all such s. Also, the fact that $z \in \text{Int } C_R$ implies that we can find a positive constant $R' < R$ such that $z \in \text{Int } C_{R'}$ and so Theorem 5.1.3 gives us that the Laurent expansion at infinity of this function is not only valid for $|s| \geq R$, but that it converges absolutely for all such s. Uniform convergence for $|s| \geq R$ also follows

from Theorem 5.1.3 and will be shown explicitly Section 5.3.1. Furthermore, by L'Hospital's Rule, this function has value 1 at $s = \infty$ and therefore the Laurent expansion is of the following form:

$$\frac{s\psi'(s)}{\psi(s) - z} = 1 + \frac{p_1}{s} + \frac{p_2}{s^2} + \cdots, \quad |s| \geq R, \quad z \in \text{Int } C_R. \tag{5.3.6}$$

By Cauchy's integral formula, the coefficients p_n in the above expression can be obtained by multiplying both sides by s^{n-1} and integrating on $|s| = R$. But this yields the same result as in (5.3.5), and so the coefficients p_n in (5.3.6) must be the Faber polynomials associated with ψ. □

The expression on the left hand side of (5.3.6) is regarded as a generating function for the Faber polynomials and its uses will be seen below. Another useful result is obtained by letting $z = \psi(w)$ and considering $p_n(\psi(w)) = F_n(w)$. First choose R_1 so that $0 < R_1 < R$ and $R_1 < |w| < R$. This implies the image $\psi(w)$ lies in the region bounded by C_R and C_{R_1}, see Figure 5.4.

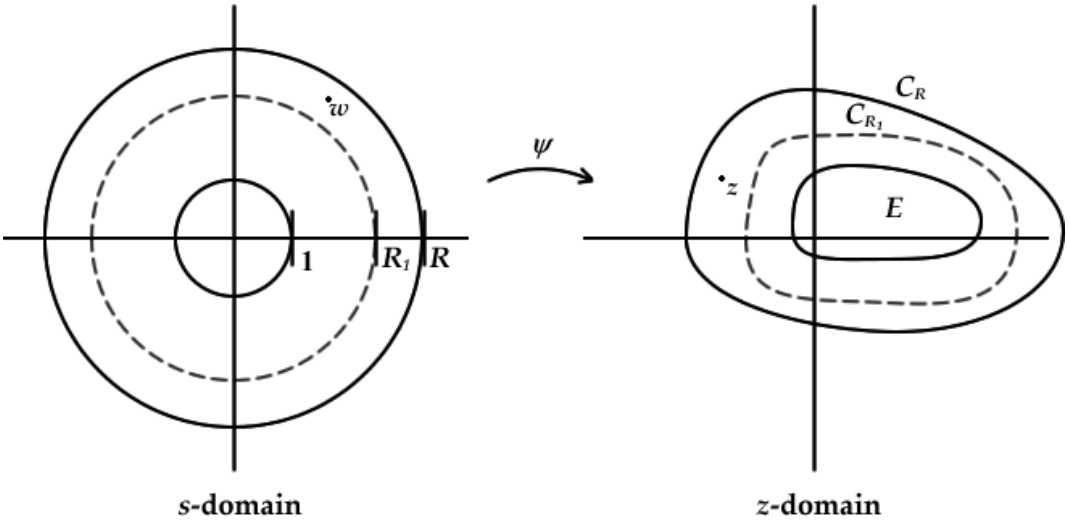


Figure 5.4: The point z with pre-image w

Denote the closed region bounded by the circles with radii R and R_1 by Ω. The integrand

of (5.3.5) can now be written as

$$\frac{s^n\psi'(s)}{\psi(s)-\psi(w)}.$$

This function, regarded as a function of s, is analytic on Ω except for an isolated singularity at $s=w$. Since the highest positive power in the expansion (5.3.1) is 1, it is easily verified that this singularity is a simple pole. The residue at this singularity is given by

$$\lim_{s\to w}(s-w)\frac{s^n\psi'(s)}{\psi(s)-\psi(w)} = \lim_{s\to w}\frac{(s-w)}{\psi(s)-\psi(w)}s^n\psi'(s) = \frac{1}{\psi'(w)}w^n\psi'(w) = w^n.$$

We now wish to use the residue theorem to evaluate the integral in (5.3.5). First, note that this integral is equal to the expression

$$\begin{aligned} p_n(\psi(w)) = &\frac{1}{2\pi i}\int_{|s|=R}\frac{s^n\psi'(s)}{\psi(s)-\psi(w)}ds \\ &-\frac{1}{2\pi i}\int_{|s|=R_1}\frac{s^n\psi'(s)}{\psi(s)-\psi(w)}ds + \frac{1}{2\pi i}\int_{|s|=R_1}\frac{s^n\psi'(s)}{\psi(s)-\psi(w)}ds. \end{aligned}$$

The first two integrals are integrals over the outer and inner boundaries of the annulus formed by the circles with radii R and R_1, respectively. We can continuously deform this annulus into a circle centered at the singularity of the integrand, namely w, as in Figure 5.5.

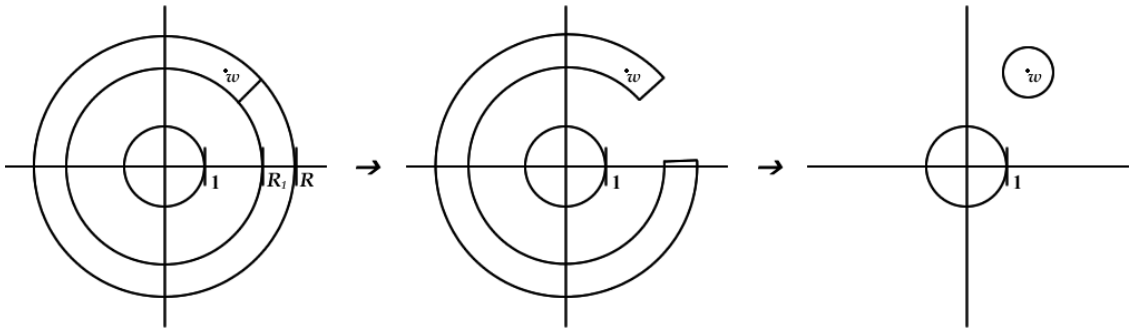


Figure 5.5: Deformation of $\partial\Omega$ into a circle about w

So the first two integrals are equal to the residue at the point w, which was calculated

above. Therefore,

$$F_n(w) = w^n + \frac{1}{2\pi i}\int_{|s|=R_1} \frac{s^n \psi'(s)}{\psi(s) - \psi(w)} ds.$$

Since s is restricted to have modulus R_1, the integral in the above equation is an analytic function of w for $|w| > R_1$. Also, it maps $w = \infty$ to $w = 0$. Therefore, by Theorem 5.1.3, we have that this integral has a Laurent expansion at infinity which converges for *at least* all $|w| > R$ (since $R > R_1$) and is of the form

$$F_n(w) = w^n + \sum_{k=1}^{\infty} a_{nk} w^{-k}. \tag{5.3.7}$$

It turns out we can be more precise about the convergence properties of this series. Since R and R_1, satisfying $R > R_1 > 1$, were chosen arbitrarily, we can further conclude that this series converges for all $|w| > 1$. Furthermore, since the series has no positive powers, if we let $w = 1/z$ we obtain a series valid for all $|z| < 1$, given by $\sum_{k=1}^{\infty} a_{nk} z^k$. The function represented by this series is analytic in the whole unit disk, including zero, and thus is uniformly convergent on any closed subdisk. Thus the infinite series in (5.3.7) is uniformly convergent for $|w| > R_1 > 1$ (since $|w|$ must always be greater than R_1). Since R_1 can be chosen arbitrarily close to 1, what this tells us is that for $|w| > 1$, the Faber polynomials, regarded as functions of w, can be written $F_n(w) = w^n + O(1/w)$.

The coefficients a_{nk} are called the *Faber coefficients* of E. It turns out that they also have an associated generating function. Choose the variables t and w so that $|t| > |w| > 1$. Then

$$\begin{aligned}
\frac{\partial}{\partial t} \ln \frac{\psi(t) - \psi(w)}{b(t-w)} &= \frac{\psi'(t)}{\psi(t) - \psi(w)} - \frac{1}{t-w} \\
&= \frac{1}{t} + F_1(w)\frac{1}{t^2} + F_2(w)\frac{1}{t^3} + \cdots - \frac{1}{t} - \frac{w}{t^2} - \frac{w^2}{t^3} - \cdots \\
&= \sum_{n=1}^{\infty} (F_n(w) - w^n) t^{-n-1} = \sum_{n=1}^{\infty} \left(\sum_{k=1}^{\infty} a_{nk} w^{-k} \right) t^{-(n+1)}
\end{aligned} \tag{5.3.8}$$

The second equality follows from (5.3.6) and the expansion of $1/(t-w)$ in a geometric series. The third and fourth equalities come from combining like terms and using (5.3.7). Note the the series in the parentheses is the same as the series appearing in (5.3.7) and thus converges for all $|w| > 1$. Fixing such a w, we see that the final series in (5.3.8) is a power series in t which converges as long as $|t| > |w| > 1$. Now choosing a positive constant $B > 1$, this power series will converge uniformly for $|t| \geq B > |w|$. Thus as long as $|t| > B$, we can integrate and differentiate this series term-by-term. In particular,

$$-\sum_{n=1}^{\infty} n^{-1} \left(\sum_{k=1}^{\infty} a_{nk} w^{-k} \right) t^{-n}$$

is an antiderivative of the last member of (5.3.8) and hence is analytic in t, convergent for $|t| > |w| > 1$, and uniformly convergent for $|t| \geq B > |w|$. By (5.3.8), this series is equal to $\ln((\psi(t) - \psi(w))/d(t-w)$ up to a constant, which depends on w since we integrated with respect to t. However, if we choose a branch of the natural log such that $\ln 1 = 0$, we can show that the constant must be zero. We have that

$$\ln \frac{\psi(t) - \psi(w)}{d(t-w)} + C(w) = -\sum_{n=1}^{\infty} n^{-1} \left(\sum_{k=1}^{\infty} a_{nk} w^{-k} \right) t^{-n},$$

where $C(w)$ denotes the arbitrary constant. Now if we let t approach infinity, the natural log term vanishes (apply L'Hospital's rule and (5.3.1)), as does the right hand side since all the powers of t are negative. Hence, $C(w) = 0$. Therefore,

$$\ln \frac{\psi(t) - \psi(w)}{d(t-w)} = -\sum_{n=1}^{\infty} n^{-1} \left(\sum_{k=1}^{\infty} a_{nk} w^{-k} \right) t^{-n}, \tag{5.3.9}$$

Since $\psi(w)$ is univalent for $|w| > 1$, the expression $(\psi(t) - \psi(w))/(t-w)$ can never vanish in the domain $M = \{t \mid |t| > 1\} \times \{w \mid |w| > 1\}$. For the case where $t = w$, L'Hosptial's

Rule gives us that this expression equals $\psi'(w)$ there, which is also nonzero for all $|w| > 1$ by conformality. Thus $(\psi(t) - \psi(w))/(t - w)$ is an analytic function of two complex variables in M, with a nonzero derivative for $|w| > 1$. So taking the log of this function is also analytic for $|w| > 1$ and therefore, the series on the right hand side of (5.3.9) must be its power series expansion about (∞, ∞), valid for $|t| > 1$ and $|w| > 1$. This also implies we can ignore the restriction $|t| > |w|$ used in deriving this relation when using (5.3.9).

Example 5.3.2. Let E be the closed disk defined by $\{z \mid |z - a| \leq d\}$. Then determining the conformal maps and the Faber polynomials is rather straightforward. The map $\psi : D(0,1)^c \to E^c$ is obtained by a magnification and then a translation. We have

$$\psi(w) = dw + a \quad \text{and} \quad \psi^{-1}(z) = \phi(z) = (z - a)/d.$$

Then $\phi(z)^n$ is itself a polynomial and so the Faber polynomials are simply

$$p_n(z) = [(z - a)/d]^n \quad \text{and} \quad F_n(w) = p_n(\psi(w)) = w^n.$$

Example 5.3.3. Let $E = [-1, 1]$. Here we will revisit the Joukowski map (see Example 5.2.5). Recall this map is given by

$$z = \psi(w) = \frac{1}{2}(w + w^{-1}).$$

If we solve for w in terms of z, we obtain

$$w = \phi(z) = z \pm \sqrt{z^2 - 1}.$$

This means that each z has two pre-images under this map. Since we are interested in mappings between the exterior of the unit disk to the exterior of $[-1, 1]$, we choose the

solution $w = z + \sqrt{z^2 - 1}$, which always has modulus greater than or equal to 1. The Faber polynomials could be obtained by computing the polynomial part of

$$\phi(z)^n = (z + \sqrt{z^2 - 1})^n.$$

However, this can be done more simply by recalling that

$$T_n(z) = T_n\left(\frac{1}{2}(w + w^{-1})\right) = \frac{1}{2}(w^n + w^{-n}),$$

where $T_n(z)$ is the n^{th} Chebyshev polynomial of the first kind (see [39] for details). We also had the result

$$w^n - F_n(w) = O(1/w).$$

Therefore,

$$2T_n(z) = (w^n + w^{-n}) = w^n + O(1/w) = F_n(w).$$

Thus the Faber polynomials are twice the Chebyshev polynomials of the first kind.

Example 5.3.4. Let E be the closed region bounded by the ellipse given by $(\text{Re } z/a)^2 + (\text{Im } z/b)^2 = 1$, where $a > b > 0$. The conformal map from $D(0,1)^c$ to E^c is given by

$$\psi(w) = \frac{a+b}{2}w + \frac{a-b}{2w} = cPw + \frac{c}{Pw},$$

where $c = \sqrt{(a^2 - b^2)}/2$ and $P = [(a+b)/(a-b)]^{1/2}$. Writing the mapping using these variables will help in the derivation of the Faber polynomials. Note that $P > 1$ and $\psi'(1/P) = 0$, so ψ is conformal (in particular, analytic and univalent) for $|w| > 1/P$. To determine the Faber polynomials, $p_n(z)$, we use the generating function given in (5.3.8). By plugging in

$\psi(w) = cPw + c/Pw$ into the left hand side and simplifying, one obtains the expression

$$\frac{1}{P^2t^2w(1-(1/P^2tw))} = \frac{1}{P^2t^2w}\sum_{n=0}^{\infty} P^{-2n}w^{-n}t^{-n} = \sum_{n=1}^{\infty} P^{-2n}w^{-n}t^{-n-1}.$$

This means that the coefficient of t^{-n-1} in the right hand side of (5.3.8) is simply $P^{-2n}w^{-n}$. Thus $F_n(w) = w^n + P^{-2n}w^{-n}$. Finally, if we solve $z = \psi(w)$ for w in terms of z and plug in for w, we obtain

$$p_n(z) = 2\sum_{k=0}^{n/2} \binom{n}{2k} z^{n-2k}(z^2 - 4c^2)^k, \quad n = 1, 2, ...$$

5.3.1 The Faber Series. The motivation that led to the discovery of the Faber polynomials was finding a way to approximate functions analytic in a given region by a series of polynomials. Let E be a simply connected, compact set in $\mathbb{C}$ and let D denote the closed unit disk. Let h be analytic in some open set containing E, and let ϕ be the conformal map guaranteed by Theorem 5.2.7 which maps E^c onto D^c, with $\phi(\infty) = \infty$ and $\phi'(\infty) > 0$. As shown above, ϕ has a Laurent expansion of the form

$$\phi(z) = a_1 z + z_0 + a_{-1}z^{-1} + a_{-2}z^{-2} + \cdots,$$

where $\phi'(\infty) = a_1$. Denote the inverse of ϕ by ψ, which has a Laurent expansion of the same form, i.e.

$$\psi(w) = b_1 w + b_0 + b_{-1}w^{-1} + b_{-2}w^{-2} + \cdots$$

where $b_1 = a_1^{-1}$.

Now let $\rho > 1$ and recall that ρe^{it} is a circle of radius ρ surrounding D. It follows that $z = \psi(\rho e^{it})$ is a curve surrounding E. Denote this curve by Γ_ρ. Since h is analytic in some open set containing E, by taking ρ sufficiently close to 1, we have that h is analytic inside

and on Γ_ρ and thus Cauchy's formula holds for $z \in E$, i.e.

$$h(z) = \frac{1}{2\pi i} \int_{\Gamma_\rho} \frac{h(t)}{t-z} dt.$$

If we substitute $t = \psi(w)$, where $|w| = \rho$, we get

$$h(z) = \frac{1}{2\pi i} \int_{|w|=\rho} \frac{h(\psi(w))\psi'(w)}{\psi(w)-z} dw. \tag{5.3.10}$$

For each $z \in E$, $\psi(w) \neq z$ for all $w \in D^c$, so the function $w \mapsto \frac{\psi'(w)}{\psi(w)-z}$ is analytic for $|w| > 1$. From Proposition 5.3.1, we have that

$$\frac{\psi'(w)}{\psi(w)-z} = \sum_{n=0}^{\infty} p_n(z) w^{-n-1}, \quad |w| > 1 \tag{5.3.11}$$

where the $p_n(z)$ are the Faber polynomials for E. We now want to show that this series converges uniformly for all $z \in E$ and $|w| \geq \rho > 1$. If we choose $\sigma > 0$ such that $1/\sigma < 1$, then by Corollary 5.1.5, we have

$$|p_n(z)| \leq \sigma^{n+1} \sup_{|w|=\sigma} \left| \frac{\psi'(w)}{\psi(w)-z} \right| = \frac{\mu(\sigma)}{\delta(\sigma)} \sigma^{n+1}, \quad z \in E.$$

where $\mu(\sigma) = \sup_{|w|=\sigma} |\psi'(w)|$ and $\delta(\sigma)$ is the distance from Γ_σ to E. Thus we have a bound for the $p_n(z)$, which depends only on the set E and σ. Now, given ρ as introduced above, if we also require $1/\rho < 1/\sigma$, then $1 < \sigma < \rho$ and so

$$\sum_{n=0}^{\infty} |p_n(z) w^{-n-1}| \leq \sum_{n=0}^{\infty} |p_n(z)| \rho^{-n-1} \leq \sum_{n=0}^{\infty} \frac{\mu(\sigma)}{\delta(\sigma)} \left(\frac{\sigma}{\rho}\right)^{n+1} < \infty$$

Thus (5.3.11) converges uniformly for $z \in E$, and $|w| \geq \rho$, as desired. Now if we substitute

(5.3.11) into (5.3.10), we can integrate term by term, obtaining the following:

$$h(z) = \frac{1}{2\pi i}\int_{|w|=\rho} \frac{h(\psi(w))\psi'(w)}{\psi(w)-z}dw = \frac{1}{2\pi i}\int_{|w|=\rho} h(\psi(w))\sum_{n=0}^{\infty} p_n(z)w^{-n-1}dw.$$
$$= \frac{1}{2\pi i}\sum_{n=0}^{\infty} p_n(z)\int_{|w|=\rho} h(\psi(w))w^{-n-1}dw = \sum_{n=0}^{\infty} c_n p_n(z)$$

where $c_n = \frac{1}{2\pi i}\int_{|w|=\rho} h(\psi(w))w^{-n-1}dw$. We have proven the following theorem:

Theorem 5.3.5. *Let h be analytic on a simply connected, compact set E and let $\{p_n(z)\}$ be the Faber polynomials associated with E. Let ψ be the conformal map which maps the complement of the closed unit disk onto the complement of E and let*

$$c_n = \frac{1}{2\pi i}\int_{|w|=\rho} h(\psi(w))w^{-n-1}dw, \qquad n = 0, 1, 2, ...$$

where $\rho > 1$ is such that h is continuous on and analytic inside the image of $|w| = \rho$ under ψ. Then the representation

$$h(z) = \sum_{n=0}^{\infty} c_n p_n(z)$$

holds uniformly for $z \in E$.

Next we will see how Faber polynomials are used, along with Crouzeix's results, in establishing convergence bounds on the convergence of GMRES.

Chapter 6. Crouzeix's Conjecture and GMRES

Here we focus on the work presented in the paper by Beckermann, et. al. [4]. This paper presents a result that improves upon the estimate given in the Ph.D. thesis by H.C. Elman [17]. In his thesis, for a matrix A with a positive definite Hermitian part M, Elman presented the following bound for the k^{th} relative residual of the generalized conjugate residual method

(see [17, Theorem 5.4]).

$$\frac{||r_k||}{||r_0||} \leq \sin^k(\beta), \quad \text{where } \cos(\beta) := \frac{\lambda_{\min}(M)}{||A||}, \quad \text{where } \beta \in [0, \pi/2). \tag{6.0.1}$$

The conjugate residual method (also known as MINRES) is an algorithm for Hermitian indefinite matrices which minimizes the 2-norm of the residual over the associated Krylov subspace. Generalized conjugate residual extends this method to non-symmetric matrices with positive definite Hermitian part. It gives the same result as the GMRES method for this subclass of matrices, and thus any bound shown for this method is also valid for GMRES applied to these matrices. See Chapter 5 in [17] for more details. Returning to (6.0.1), note that since M is positive definite, its numerical range is positive real, and by Proposition 2.1.5 and Proposition 2.1.6, the smallest eigenvalue of M bounds the real part of $W(A)$ from below. Also, note that by Theorem 2.4.3, $W(A)$ is contained in the disk of radius $||A||$. So the angle β can be visualized as the angle between the line connecting the origin and $\lambda_{\min}(M)$ and a line of length $||A||$, see Figure 6.1. Note that β also provides a way of assessing how close $W(A)$ is to 0. If β is close to $\pi/2$, then $W(A)$ is close to 0, and if β is small, then $W(A)$ is far away from the origin. Thus a small β is desirable, since when the numerical range is far from 0, GMRES converges quickly, as indicated by (6.0.1).

We can extend the result in (6.0.1) to any matrix with $0 \notin W(A)$. Since $W(A)$ is convex, there exists a $t \in \mathbb{R}$ such that $\text{dist}(0, W(A)) = \lambda_{\min}((e^{it}A + (e^{it}A)^*)/2)$. Since the norm of the k^{th} relative residual is unaffected by multiplying A by a number of modulus 1, (6.0.1) now becomes

$$\frac{||r_k||}{||r_0||} \leq \sin^k(\beta), \quad \text{where } \cos(\beta) := \frac{\text{dist}(0, W(A))}{||A||}. \tag{6.0.2}$$

for all matrices with $0 \notin W(A)$. We will use this version of Elman's estimate to compare with the results given in [4].

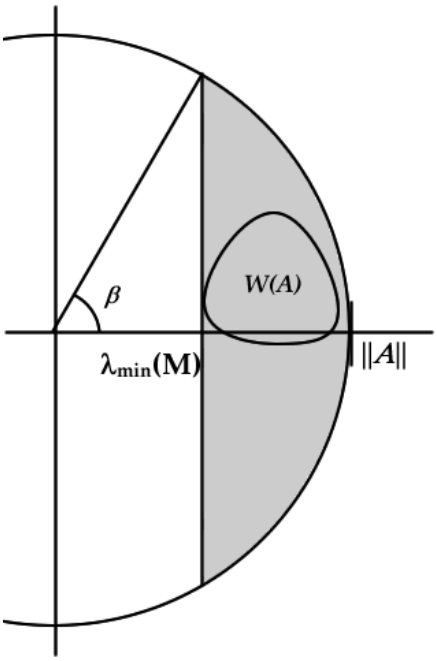


Figure 6.1: The angle β of (6.0.1)

6.1 Beckermann, et. al.'s Convergence Bound for GMRES

The main result of Beckermann, et. al. is given in the following theorem.

Theorem 6.1.1. *Let A be a square matrix with $0 \notin W(A)$ and define $\beta \in [0, \pi/2)$ as in (6.0.2). Then for the k^{th} relative residual of GMRES, we have*

$$\frac{||r_k||}{||r_0||} \leq (2 + 2/\sqrt{3})(2 + \gamma_\beta)\gamma_\beta^k, \quad k = 1, 2, 3, ... \tag{6.1.1}$$

where

$$\gamma_\beta = 2 \sin\left(\frac{\beta}{4 - 2\beta/\pi}\right) < \sin(\beta). \tag{6.1.2}$$

We will prove this theorem by breaking the bulk of the work into smaller results. First, we introduce some needed notation. For a compact set $K \subset \mathbb{C}$, let

$$E_k(K) = \min\{||p||_K \mid p \text{ a polynomial with degree } \leq k, p(0) = 1\},$$

where $||\cdot||_K$ denotes the maximum norm on K. Next, let $\alpha \in [0, \pi]$ and define

$$S_\alpha = \{z \in C \mid \arg z \in [0, \alpha]\}.$$

The set S_α is called a *sector*. Crouzeix showed in [9] that

$$||p(A)|| \leq (2 + 2/\sqrt{3}) \sup_{\substack{z \in S_\alpha \\ W(A) \subset S_\alpha}} |p(z)|. \tag{6.1.3}$$

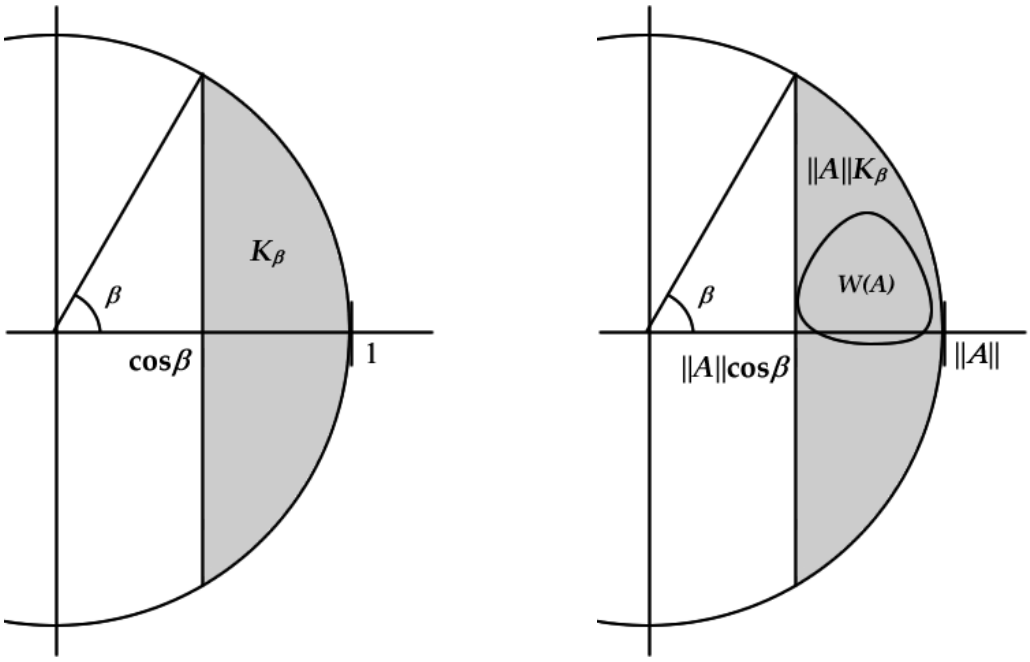
To prepare for the following lemma, recall by the remarks above, that the k^{th} relative residual of GMRES is unchanged when multiplying A by a complex number of modulus 1. So without loss of generality, we can suppose that the point of $W(A)$ closest to 0 is real positive. Then, by (6.0.2), we have that $W(A) \subset \{z \mid \text{Re } z \geq ||A|| \cos(\beta)\}$. Define K_β to be the intersection between the half plane $\{z \mid \text{Re } z \geq \cos(\beta)\}$ and the closed unit disk. Since $W(A)$ is also contained in the closed disk with radius $||A||$, we then have that $W(A) \subset ||A|| K_\beta$; see Figure 6.2. The importance of this is that now we have a convex, compact set which contains $W(A)$ and is easy to work with in the sense that we can explicitly construct a conformal map involving this region, as will be seen below. Using these ideas, we have the following estimate for the quantity $E_k(K_\beta)$.

Lemma 6.1.2. *Let $k \geq 1$ and $\beta \in [0, \pi/2)$. Then*

$$\gamma_\beta^k < E_k(K_\beta) \leq \min\left\{2 + \gamma_\beta, \frac{2}{1 - \gamma_\beta^{k+1}}\right\}\gamma_\beta^k, \tag{6.1.4}$$

where γ_β is as in (6.1.2).

Proof. Let K be a convex, compact set containing more than one point, with $0 \notin K$. We will first prove a similar result for this set and then extend it to prove (6.1.4). Let ϕ denote the conformal map from the exterior of K to the exterior of the closed unit disk with the



(a) The set K_β (b) The set $||A||K_\beta$ containing $W(A)$

Figure 6.2: The sets K_β and $||A||K_\beta$

properties $\phi(\infty) = \infty$ and $\phi'(\infty) > 0$ (see Theorem 5.2.7). We will show that for $k \geq 1$,

$$\gamma^k \leq E_k(K) \leq \min\left\{2 + \gamma, \frac{2}{1 - \gamma^{k+1}}\right\}\gamma^k, \quad \text{where } \gamma = 1/|\phi(0)|. \tag{6.1.5}$$

First we show the left hand inequality. Let p be an arbitrary polynomial of degree less than or equal to k satisfying $p(0) = 1$. Recall that by Corollary 5.2.9, we can continuously extend ϕ to the boundary of K, and by the definition of ϕ, we have $|\phi(z)| = 1$ for all $z \in \partial K$. Next, consider the function p/ϕ^k. We want to argue that the modulus of this function on $\hat{\mathbb{C}}/K$ is bounded by its modulus on ∂K. To do so, we will map $\hat{\mathbb{C}}\backslash K$ conformally onto a bounded set, Ω. Since $0 \notin K$, there exists a scalar $a \neq 0$ such that $0 \in K + a$. So let the map be

defined by $z \mapsto w$, where $w = 1/(z+a)$. Now note that if $\phi(z) = bz + b_0 + \sum_{n=1}^{\infty} b_n z^{-n}$, then

$$\begin{aligned}\phi\Big(\frac{1}{w} - a\Big) &= b\Big(\frac{1}{w} - a\Big) + b_0 + \sum_{n=1}^{\infty} b_n \frac{w^n}{(1-aw)^n} \\ &= \frac{1}{w}\left(b'w + b + w\sum_{n=1}^{\infty} b_n \frac{w^n}{(1-aw)^n}\right),\end{aligned}$$

where $b' = (b_0 - ab)$. Using this, we have the following result for all $z \in \hat{\mathbb{C}}\backslash K$,

$$\begin{aligned}\frac{p(z)}{\phi(z)^k} &= \frac{(1/w^k)q(w)}{\phi(\frac{1}{w} - a)^k} = \frac{(1/w^k)q(w)}{(1/w^k)(b'w + b + w\sum_{n=1}^{\infty} b_n \frac{w^n}{(1-aw)^n})^k} \\ &= \frac{q(w)}{(b'w + b + w\sum_{n=1}^{\infty} b_n \frac{w^n}{(1-aw)^n})^k},\end{aligned}$$

where q is a polynomial of degree less than or equal to k. Furthermore, the denominator never vanishes since ϕ maps to the exterior of the closed unit disk. Thus the right hand side of the last equation is an analytic function of w on the bounded set Ω and thus attains its maximum modulus on $\partial\Omega$. Since the boundary of Ω corresponds to the boundary of K under the map $z \mapsto w$, we can conclude that p/ϕ^k does indeed attain its maximum modulus on ∂K. We then have the following.

$$\gamma^k = \frac{|p(0)|}{|\phi(0)|^k} \leq \frac{||p||_{\partial K}}{||\phi||_{\partial K}^k} = ||p||_{\partial K} = ||p||_K. \tag{6.1.6}$$

Since $||\phi||_{\partial K}^k = 1$. Since p was arbitrary, this shows the left hand side of (6.1.5). An equality is attained if and only if ϕ^k is a polynomial equal to p. A result in [37] shows that ϕ^k is a polynomial if and only if K is a lemniscate, which is not the case for K_β, so we take the inequality to be strict in (6.1.4).

For the second inequality in (6.1.5), the goal is to find a polynomial whose maximum norm over K can be bounded by something of the form of the right hand side in this

equation. We will use the Faber polynomial of degree k, denoted F_k. Recall that this is the polynomial part of the Laurent expansion at infinity for ϕ^k. We will use the result of Kovari and Pommerenke given in [35], which states that

$$\delta_k := ||F_k - \phi^k||_{\partial K} \leq 1 \tag{6.1.7}$$

for general convex sets K. Consider the function $\phi F_k - \phi^{k+1}$. If we perform the same mapping as above, i.e. $z \mapsto w = 1/(z+a)$, we get, after some simplification,

$$\begin{aligned}
&\phi(z)F_k(z) - \phi(z)^{k+1} \\
&= \frac{(b'w + b + w\sum_{n=1}^{\infty} b_n \frac{w^n}{(1-aw)^n})G_k(w) - (b'w + b + w\sum_{n=1}^{\infty} b_n \frac{w^n}{(1-aw)^n})^{k+1}}{w^{k+1}} = \frac{f(w)}{g(w)},
\end{aligned}$$

where $G_k(w) = w^k F_k((1/w) - a)$. We want to show this expression is analytic for all $w \in \Omega$. The only point of concern is $w = 0$. Since $G_k(0) = b$, we have that $f(0) = 0$ and since the degree of the numerator is $k+1$, it can be shown that the first $k+1$ derivatives of $f(w)$ are zero also. By applying L'Hospital's rule, we see that this function is analytic at 0. Hence, $|f(w)/g(w)|$ attains its maximum on $\partial\Omega$, which implies $|\phi(z)F_k(z) - \phi^{k+1}(z)|$ has its maximum on ∂K. Therefore,

$$|\phi(0)||F_k(0) - \phi(0)^k| \leq ||\phi F_k - \phi^{k+1}||_{\partial K} \leq ||\phi||_{\partial K}||F_k - \phi^k||_{\partial K} = \delta_k, \tag{6.1.8}$$

since $||\phi||_{\partial K} = 1$, as before.

Now let $v \in [0,1]$ and define

$$p_v(z) = F_k(z) + v(\phi(0)^k - F_k(0)).$$

Then

$$\begin{aligned}|p_v(0)| &= |F_k(0) + v(\phi(0)^k - F_k(0))| = |F_k(0) - \phi(0)^k + \phi(0)^k + v(\phi(0)^k - F_k(0))| \\ &\geq |\phi(0)|^k - |F_k(0) - \phi(0)^k - v(F_k(0) - \phi(0)^k)| = |\phi(0)|^k - (1-v)|F_k(0) - \phi(0)^k| \\ &\geq |\phi(0)|^k - (1-v)\frac{\delta_k}{|\phi(0)|}.\end{aligned}$$

Also,

$$\begin{aligned}||p_v||_K = ||p_v||_{\partial K} &\leq ||\phi^k||_{\partial K} + ||\phi^k - F_k||_{\partial K} + v|\phi(0)^k - F_k(0)| \\ &\leq 1 + \delta_k + v\frac{\delta_k}{|\phi(0)|}.\end{aligned}$$

These latter two equations both follow from (6.1.7) and (6.1.8). Therefore,

$$E_k(K) \leq \min_{v\in[0,1]} \frac{||p_v||_K}{p_v(0)} \leq \min_{v\in[0,1]} \frac{1 + \delta_k + v(\delta_k/|\phi(0)|)}{|\phi(0)|^k - (1-v)(\delta_k/|\phi(0)|)} = \frac{1 + \delta_k(1+v\gamma)}{1 - (1-v)\delta_k\gamma^{k+1}}\gamma^k.$$

To determine the minimum of the right hand side over $v \in [0,1]$, we can simply take the derivative with respect to v and determine the critical points. It turns out that given expression has no critical points in the unit interval, and thus the minimum must occur at either $v = 0$ or $v = 1$. Since plugging in $v = 0$ yields $\frac{2}{1-\gamma^{k+1}}\gamma^k$ and $v = 1$ yields $(2+\gamma)\gamma^k$, we have

$$E_k(K) \leq \min\left\{2 + \gamma, \frac{2}{1-\gamma^{k+1}}\right\}\gamma^k.$$

This proves (6.1.5).

To extend this result to K_β, we only need to show that the conformal map for K_β satisfies $1/|\phi(0)| = \gamma_\beta$. It turns out that we can verify this directly by noting in this case

that $\phi = T_3 \circ T_2 \circ T_1$ where

$$T_1(z) = \frac{z - e^{i\beta}}{z - e^{-i\beta}}, \quad T_2(z) = (e^{i(\pi-\beta)}z)^{\pi/(2\pi-\beta)}, \quad T_3(z) = \frac{z - \overline{T_2(1)}}{z - T_2(1)}. \tag{6.1.9}$$

By direct substitution, we have

$$\frac{1}{|\phi(0)|} = \left| \frac{\exp\left(i\pi\frac{\pi+\beta}{2\pi-\beta}\right) - \exp\left(i\pi\frac{\pi-\beta}{2\pi-\beta}\right)}{\exp\left(i\pi\frac{\pi+\beta}{2\pi-\beta}\right) - \exp\left(-i\pi\frac{\pi-\beta}{2\pi-\beta}\right)} \right|$$

We can factor the terms $\exp\left(\frac{i\pi^2}{2\pi-\beta}\right)$ and $\exp\left(\frac{i\pi\beta}{2\pi/\beta}\right)$ out of the numerator and and denominator respectively. The moduli of these numbers are both 1, so we now have

$$\begin{aligned}\frac{1}{|\phi(0)|} &= \left| \frac{\exp\left(i\frac{\beta\pi}{2\pi-\beta}\right) - \exp\left(-i\frac{\beta\pi}{2\pi-\beta}\right)}{\exp\left(i\frac{\pi^2}{2\pi-\beta}\right) - \exp\left(-i\frac{\pi^2}{2\pi-\beta}\right)} \right| = \frac{\sin\left(\frac{\beta\pi}{2\pi-\beta}\right)}{\sin\left(\frac{\pi^2}{2\pi-\beta}\right)} \\ &= \frac{\sin\left(\frac{2\beta}{4-2\beta/\pi}\right)}{\sin\left(\frac{4\pi^2}{2(4\pi-2\beta)}\right)} = \frac{\sin\left(\frac{2\beta}{4-2\beta/\pi}\right)}{\sin\left(\frac{\pi}{2} + \frac{\beta}{4-2\beta/\pi}\right)},\end{aligned}$$

where the last equality comes from adding and subtracting $2\beta\pi$ in the numerator of the sine argument in the denominator and simplifying. Now if we let $u = \beta/(4 - 2\beta/\pi)$, we have

$$\frac{1}{|\phi(0)|} = \frac{\sin(2u)}{\cos(u)} = \frac{2\sin(u)\cos(u)}{\cos(u)} = 2\sin(u) = \gamma_\beta.$$

This completes the proof. □

We now want to compare the asymptotic convergence factor provided by Elman in (6.0.1) with the one just derived. Since $\gamma_\beta^k < E_k(K_\beta)$ by the left hand inequality of (6.1.4), we can

do this by considering this inequality for $k = 1$. Note that

$$E_1(K_\beta) = \min_{a\in\mathbb{C}} \sup_{z\in K_\beta} |az+1| = \min_{a\in\mathbb{C}} |a| \sup_{z\in K_\beta} |z+(1/a)| = \min_{b\in\mathbb{C}} \frac{1}{|b|} \sup_{z\in K_\beta} |z+b|,$$

where $b = 1/a$, $a \neq 0$. First consider the problem of finding the supremum of $|z+b|$ on K_β. Since the function $z+b$ is analytic, and K_β is compact, it follows from the maximum modulus principle that this supremum is attained on the boundary of K_β. Now if $\arg b \in [-\beta, \beta]$, then the maximum modulus will be attained at the point z on the unit disk that satisfies $\arg z = \arg b$. This follows from the fact that at no point on the unit circle is the curvature zero. Also note that a shift such as this will push K_β away from the origin, and so the *minimal* supremum norm cannot be attained if $\arg b \in [-\beta, \beta]$. So now let $\arg b \in (\beta, -\beta+2\pi)$. Now note that the part of the boundary of K_β that has real part equal to $\cos(\beta)$ is a straight line segment so if the maximum modulus occurs on the line segment, it must occur at the endpoints. Thus, the maximum modulus must be attained at a point of the form $e^{i\theta}$, where $\theta \in [-\beta, \beta]$. Writing $b = re^{i\phi}$, where $r > 0$ and $\phi \in (\beta, -\beta+2\pi)$, we have

$$\sup_{z\in K_\beta} |z+b|^2 = \sup_{\theta\in[-\beta,\beta]} |e^{i\theta} + re^{i\phi}|^2 = \sup_{\theta\in[-\beta,\beta]} 1 + r^2 + 2r\cos(\theta-\phi).$$

Taking the derivative with respect to θ and setting it equal to zero, we see that there is a critical point whenever $\theta = \phi$ or when $\theta - \phi = c\pi$, $c \in \mathbb{Z}$. The former case cannot happen since $\theta \in [-\beta, \beta]$ and $\phi \in (\beta, -\beta+2\pi)$. In the latter case, by examining the sign of the derivative on either side of the point $\theta = c\pi + \phi$, we see that such a point is a local minimum point. So the maximum must occur on the endpoints of the interval, namely $\theta = \pm\beta$. To determine $\arg b$, note that since we have already deduced that the supremum norm is attained at $e^{\pm\beta i}$, it follows from geometric considerations that this supremum norm is minimized only when $\arg b = \pi$. Now we need to determine the optimal value of r so that the supremum norm is

minimized. By symmetry, it suffices to consider the case where $\theta = \beta$. We have

$$\min_{b\in\mathbb{C}} \frac{1}{|b|}|b + e^{i\beta}|^2 = \min_{r>0} |1 + \frac{e^{i\beta}}{re^{i\pi}}|^2 = \min_{r>0} |1 - e^{i\beta}/r|^2 = \min_{r>0} 1 + \frac{1}{r^2} - \frac{e^{-i\beta}}{r} - \frac{e^{i\beta}}{r}.$$

Taking the derivative with respect to r and setting it equal to zero, we get a critical point when $r = 1/\cos(\beta)$. The second derivative test confirms this is a global maximum, and so $b = -1/\cos(\beta)$. Thus $a = -\cos(\beta)$.

The next step is to evaluate $|az + 1|$ for $a = -\cos(\beta)$ and $z = e^{-i\beta}$. We have

$$\begin{aligned} |az + 1| &= |-\cos(\beta)(e^{i\beta}) + 1| = \left|\frac{-e^{i\beta} - e^{-i\beta}}{2}e^{i\beta} + 1\right| = \left|\frac{e^{i2\beta} + 1}{2} - 1\right| \\ &= \left|e^{i\beta}\frac{e^{i\beta} - e^{-i\beta}}{2}\right| = |\sin(\beta)| = \sin(\beta), \end{aligned}$$

where the last equality follows from the fact that $\beta \in [0, \pi/2)$.

By this last argument, we have that the asymptotic convergence factor $\sin(\beta)$ provided by Elman in (6.0.1) is equal to $E_1(K_\beta)$. By Lemma 6.1.2, the asymptotic convergence factor γ_β of Theorem 6.1.1 is strictly less than this, thus we have obtained an result asymptotically sharper than (6.0.1).

Next, we have

Lemma 6.1.3. *Let $\beta \in [0, \pi/2)$ be as in* (6.0.2)*. Then for any nonzero polynomial p, we have*

$$||p(A)|| \leq (2 + 2/\sqrt{3})||p||_K, \quad \textit{where } K = ||A||K_\beta. \tag{6.1.10}$$

Proof. Choose an angle $\alpha \in (\beta, \pi/2)$. We will use the linear fractional transformation given by

$$r(z) = \frac{||A||e^{i\alpha} - z}{z - ||A||e^{-i\alpha}}. \tag{6.1.11}$$

This map satisfies $r(||A||K_\alpha) = S_\alpha$. Let $f = p \circ r^{-1}$. Note that $r^{-1}(z) = (z||A||e^{-i\alpha} +$

$||A||e^{i\alpha})(1+z)$, and so the only poles of f are at $z=-1$ which is not in S_α. Note that if $W(r(A)) \subset S_\alpha$, then we can apply (6.1.3) as follows:

$$||p(A)|| = ||f(r(A))|| \leq (2+2/\sqrt{3})||f||_{S_\alpha} = (2+2/\sqrt{3})||p\circ r^{-1}||_{S_\alpha} = (2+2\sqrt{3})||p||_K$$

So we need only show the relation $W(r(A)) \subset S_\alpha$. Let y be a nonzero vector, and for convenience, define $\tilde{y} = (A - ||A||e^{-i\alpha}I)^{-1}y$, which is also nonzero. Then

$$\begin{aligned} d := \frac{(r(A)y,y)}{(\tilde{y},\tilde{y})} &= \frac{((||A||e^{i\alpha}I - A)\tilde{y}, (A-||A||e^{-i\alpha}I)\tilde{y})}{(\tilde{y},\tilde{y})} = \frac{((A-||A||e^{-i\alpha}I)^*(||A||e^{i\alpha}-A)\tilde{y},\tilde{y})}{(\tilde{y},\tilde{y})} \\ &= ||A||e^{i\alpha}\frac{(A^*\tilde{y},\tilde{y})}{(\tilde{y},\tilde{y})} - \frac{(A^*A\tilde{y},\tilde{y})}{(\tilde{y},\tilde{y})} - ||A||^2e^{2i\alpha} + ||A||e^{i\alpha}\frac{(A\tilde{y},\tilde{y})}{(\tilde{y},\tilde{y})} \\ &= -||A||^2e^{2i\alpha} - \frac{||A\tilde{y}||^2}{||\tilde{y}||^2} + 2||A||e^{i\alpha}\text{Re}\left(\frac{(A\tilde{y},\tilde{y})}{(\tilde{y},\tilde{y})}\right) \end{aligned}$$

Next, we have two key observations. The first is that

$$\text{Im }(d) = -||A||^2\sin(2\alpha) + 2||A||\text{Re}\left(\frac{(A\tilde{y},\tilde{y})}{(\tilde{y},\tilde{y})}\right)\sin(\alpha) \geq 2||A||^2\sin(\alpha)(-\cos(\alpha)+\cos(\beta)) > 0.$$

The first inequality follows from the expression on the right hand side of (6.0.1), since $\lambda_{\min}(M)$ bounds the real part of $W(A)$ from below. The second inequality holds since $\alpha > \beta$. Next, we also have that

$$\text{Im }(e^{-i\alpha}d) = -||A||^2\sin(\alpha) + \frac{||A\tilde{y}||^2}{||\tilde{y}||^2}\sin(\alpha) = \sin(\alpha)\Big(-||A||^2 + \frac{||A\tilde{y}||^2}{||\tilde{y}||^2}\Big) < 0$$

What these two equations tell us is that the imaginary part of d is positive, but a clockwise rotation of d through the angle α results in a number with a negative imaginary part. This implies that $d \in S_\alpha$, see Figure 6.3

Lastly, recall that $d = (r(A)y,y)/(\tilde{y},\tilde{y})$, and therefore $\frac{||\tilde{y}||^2}{||y||^2}d$ is a point in $W(r(A))$. Since a real positive scaling of any point in S_α is also in S_α, we have that $W(r(A)) \subset S_\alpha$. $\square$

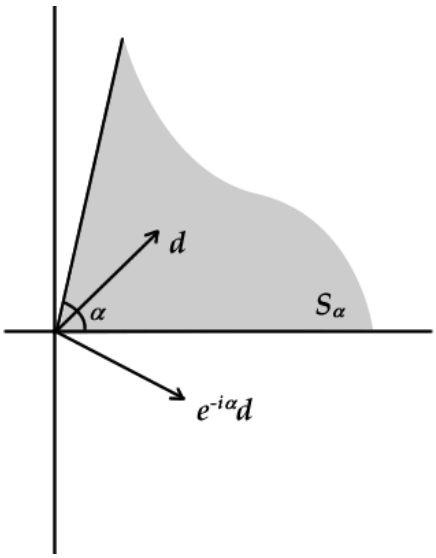


Figure 6.3: The point d in the set S_α

Now to apply this result to the set K_β, we need to show that the quantity $E_k(K)$ is invariant under a scaling of the set K. Let $p(z)$ be an arbitrary polynomial of degree less than or equal to k, $p(0) = 1$ and let $q(z) = p(cz)$ for some constant $c \in \mathbb{C}$. Then $\deg q = \deg p$, $p(0) = q(0) = 1$ and $||q||_K = ||p||_{cK}$. Therefore, $E_k(cK) = E_k(K)$. Then, from Lemma 6.1.3, we have the following:

$$\min\{||p(A)|| \mid \deg p \leq k, p(0) = 1\} \leq (2 + 2/\sqrt{3})E_k(||A||K_\beta) = (2 + 2/\sqrt{3})E_k(K_\beta).$$

Applying Lemma 6.1.2, we then have

$$\begin{aligned}\frac{||r_k||}{||r_0||} &\leq \min\{||p(A)|| \mid \deg p \leq k, p(0) = 1\} \leq (2 + 2/\sqrt{3})E_k(K_\beta) \\ &\leq (2 + 2/\sqrt{3}) \min\left\{2 + \gamma_\beta, \frac{2}{1 - \gamma_\beta^{k+1}}\right\}\gamma_\beta^k \leq (2 + 2/\sqrt{3})(2 + \gamma_\beta)\gamma_\beta^k.\end{aligned}$$

Since $\gamma_\beta < 1$, we have that $2/(1 - \gamma_\beta^{k+1}) \leq 2 + \gamma_\beta$ for large k, which explains the choice of $2 + \gamma_\beta$ in the above equation. The shrewd reader will have noted that the proof of (6.1.5)

does not depend on the particular shape of K. Thus we can state the following more general result:

Corollary 6.1.4. *Let $K \subset \mathbb{C}$ be a compact, convex set not containing 0. Let A be a matrix satisfying $W(A) \subset ||A||K$. Then for the k^{th} relative residual of GMRES, we have*

$$\frac{||r_k||}{||r_0||} \leq (2+\gamma)C(K)\gamma^k < (2+\gamma)C(K)E_k(K),$$

where $C(K)$ is as in (1.0.3) *and $\gamma = 1/|\phi(0)|$, where ϕ is the conformal map from $\hat{\mathbb{C}}\backslash K$ to $\hat{\mathbb{C}}\backslash\overline{D(0,1)}$ satisfying $\phi(\infty) = \infty$, $\phi'(\infty) > 0$.*

The second inequality in the statement of the corollary follows from the first inequality in (6.1.4).

6.2 Beckermann's Improvements

Not long after the above results were established, Beckermann independently published an improvement [3]. As we will see, his new bound is sharper than that given in Theorem 6.1.1, and *does not* use the Crouzeix conjecture. We do not imply that this undermines the importance of the Crouzeix conjecture, only that the best GMRES bounds obtained via Faber polynomials (that we are aware of) are not dependent on it. Despite this, we still present these results here as the method of proof closely resembles the strategy employed by Crouzeix in some of his papers [2, 8].

Theorem 6.2.1. *Let $A \in M_n$, and let $E \subset \mathbb{C}$ be convex, compact, and satisfy $W(A) \subset E$. Let F_n^E be the associated Faber polynomial. Then*

$$||F_n^E(A)|| \leq 2, \quad n \geq 1.$$

Proof. Let ϕ be the conformal map from E^c to $\overline{D(0,1)}^c$, with $\phi(\infty) = \infty$, $\phi'(\infty) > 0$. By the definition of F_n^E, we have that $\phi^n - F_n^E$ and $1/\phi^n$ are analytic in $\overline{\mathbb{C}}\backslash E$ and vanish at infinity. Combining this fact with (5.3.4) we have

$$\frac{1}{2\pi i}\int_{|\phi(z)|=r}\frac{\phi(z)^n}{z-\zeta}dz = \begin{cases} F_n^E(\zeta), & n \geq 0 \\ 0, & n < 0. \end{cases}$$

where $r > 1$ and ζ is some point in the interior of the curve defined by $|\phi(z)| = r$. Since $\sigma(A) \subset W(A) \subset E$, we can write

$$\frac{1}{2\pi i}\int_{|\phi(z)|=r}\phi(z)^n(zI-A)^{-1}dz = \begin{cases} F_n^E(A), & n \geq 0 \\ 0, & n < 0. \end{cases} \tag{6.2.1}$$

Since we can make r arbitrarily close to 1, we have, for $n \geq 1$,

$$0 = \frac{1}{2\pi i}\int_{\partial E}\phi(z)^{-n}(zI-A)^{-1}dz = \frac{1}{2\pi i}\int_{\partial E}\overline{\phi(z)}^n(zI-A)^{-1}dz. \tag{6.2.2}$$

Now note that since E is convex, the outward unit normal exists almost everywhere for $z \in \partial E$. Thus, if the boundary of E has a parameterization $z(t) = a(t) + ib(t)$, then the outward normal is given by $b'(t) - ia'(t) = \frac{1}{i}(a'(t) + ib'(t))$. In terms of differentials, we can write $\nu(z) = \frac{1}{i}\frac{dz}{|dz|}$ for the outward unit normal. Then, combining (6.2.1) and the adjoint of (6.2.2), we have

$$\begin{aligned} F_n^E(A) &= \frac{1}{2\pi i}\int_{\partial E}\phi(z)^n(zI-A)^{-1}dz - \frac{i}{2\pi}\int_{\partial E}\phi(z)^n(\overline{z}I-A^*)^{-1}d\overline{z} \\ &= \frac{1}{2\pi}\int_{\partial E}\phi(z)^n\nu(z)(zI-A)^{-1}|dz| + \frac{1}{2\pi}\phi(z)^n\overline{\nu}(z)(\overline{z}I-A^*)^{-1}|dz| \\ &= \int_{\partial E}\phi(z)^n\mu(z,A)|dz|, \end{aligned}$$

where

$$\mu(z, A) = \frac{1}{2\pi}(\nu(z)(zI - A)^{-1} + \overline{\nu}(z)(\overline{z}I - A^*)^{-1}).$$

It is shown in Lemma 6.2.2 that

$$\int_{\partial E} \mu(z, A)|dz| = 2I.$$

Therefore,

$$||F_n^E(A)|| \leq \max_{\zeta \in \partial E} |\phi(\zeta)^n| \sup_{y \in \mathbb{C}^n, ||y||=1} \int_{\partial E} (\mu(z, A)y, y)|dz| = 2.$$

Since ϕ has norm 1 on the boundary of E by definition. □

Lemma 6.2.2. *Let $A \in M_n$, $E \subset \mathbb{C}$ be a convex, compact, and positively oriented set satisfying $W(A) \subset E$. Then*

$$\int_{\partial E} \mu(\sigma, A)ds = 2I,$$

where $\sigma = \sigma(s)$ is a function of arclength which represents a point in ∂E and

$$\mu(\sigma, z) = \frac{1}{2\pi}(\frac{\nu(\sigma)}{\sigma - z} - \frac{\overline{\nu(\sigma)}}{\bar{\sigma} - \bar{z}}),$$
$$\text{and } \mu(\sigma, A) = \frac{1}{2\pi}(\nu(\sigma)(\sigma I - A)^{-1} - \overline{\nu(\sigma)}(\bar{\sigma} I - A^*)^{-1}),$$

where $\nu(\sigma) = \frac{1}{i}\frac{d\sigma}{ds}$ is the outward normal to E at the point $\sigma \in \partial E$.

Proof. We have that

$$\int_{\partial E} \mu(\sigma, A)ds = \frac{1}{2\pi}\int_{\partial E} \nu(\sigma)(\sigma I - A)^{-1}ds + \frac{1}{2\pi}\int_{\partial E} \overline{\nu(\sigma)}(\sigma I - A^*)^{-1}ds. \tag{6.2.3}$$

We first consider the first integral. Let $\zeta = \sigma(s)$. Then $d\zeta = \sigma'(s)ds = i\nu ds$. So this integral becomes

$$\frac{1}{2\pi i}\int_{\partial E} (\zeta - A)^{-1}d\zeta.$$

We can evaluate this integral as long as ζ is in the resolvent set of A. Otherwise, $\zeta \in \sigma(A)$. However, the spectrum of A is a set of measure zero and therefore these points do not affect the value of the integral. So by the Cauchy formula, we get

$$\frac{1}{2\pi i}\int_{\partial E}(\zeta I - A)^{-1}d\zeta = I.$$

To evaluate the second integral, again let $\zeta = \bar{\sigma}(s)$. Then $d\zeta = \bar{\sigma}'(s)ds = \frac{\bar{\nu}}{i}ds$. So we get

$$\frac{1}{2\pi}\int_{\partial E}\overline{\nu(\sigma)}(\sigma I - A^*)^{-1}ds = \frac{i}{2\pi}\int_{\partial \overline{E}}(\zeta I - A^*)^{-1}d\zeta = -\frac{1}{2\pi i}\int_{\partial \overline{E}}(\zeta I - A^*)^{-1}d\zeta.$$

Note that $\partial\overline{E}$ is negatively oriented, so we have

$$-\frac{1}{2\pi i}\int_{\partial \overline{E}}(\zeta I - A^*)^{-1}d\zeta = -(-I) = I.$$

Plugging these values into (6.2.3), we obtain

$$\int_{\partial E}\mu(\sigma, A)ds = 2I.$$

□

The next result is a general application to the convergence analysis of GMRES.

Theorem 6.2.3. *Let E be a convex, compact subset of the complex plane that does not contain zero, and satisfies $W(A) \subset E$. Then for the k^{th} relative residual of GMRES, we have*

$$\frac{||r_k||}{||r_0||} \leq \min\left\{2 + \gamma_E, \frac{2}{1 - \gamma_E^{k+1}}\right\}\gamma_E^k, \tag{6.2.4}$$

where $\gamma_E = 1/|\phi(0)|$, ϕ being the conformal map as in Theorem 6.2.1.

Proof. The proof of this theorem is similar to the proof of Lemma 6.1.2. We will show that

$$\min\left\{\frac{||p(A)||}{|p(0)|} \mid \deg p \leq k\right\} \leq \frac{2}{|F_k^E(0)|} \leq \frac{2}{1-\gamma_E^{k+1}}\gamma_E^k. \tag{6.2.5}$$

The first inequality follows directly from Theorem 6.2.1. To get an estimate for $|F_k^E(0)|$, we use the maximum principle applied to $\phi(\phi^k - F_k^E)$ as before. We have

$$\begin{aligned}
|\phi(0)||F_k^E(0)| &\geq |\phi(0)|^{k+1} - |\phi(0)(\phi(0)^k - F_k^E(0)| \\
&\geq |\phi(0)|^{k+1} - \max_{z\in\partial E}|\phi(z)(\phi(z)^k - F_k^E(z)| \\
&= |\phi(0)|^{k+1} - \max_{z\in\partial E}|\phi(z)^k - F_k^E(z)| \geq |\phi(0)|^{k+1} - 1.
\end{aligned}$$

The first inequality comes from adding and subtracting $\phi(0)^k$ and the last inequality follows from (6.1.7). From this we have

$$\frac{1}{F_k(0)} \leq \frac{1}{\gamma_E(\gamma_E^{k+1}-1)} = \frac{\gamma_E^k}{1-\gamma_E^{k+1}}.$$

This gives us one of the bounds in (6.2.4). The other follows from the fact that $\min\{1, \frac{2\gamma_E^k}{1-\gamma_E^{k+1}}\} \leq (2+\gamma_E)\gamma_E^k$. □

This final corollary is an application of Theorem 6.2.3 to a specific E.

Corollary 6.2.4. *Let $0 \notin W(A)$. Then for the k^{th} relative residual of GMRES we have*

$$\frac{||r_k||}{||r_0||} \leq (2+\gamma)\gamma^k, \quad \textit{where} \quad \gamma = 2\sin(\frac{\beta'}{4-2\beta'/\pi}) < \sin(\beta'),$$

where $\beta' \in (0, \pi/2)$ is defined by $\cos(\beta') = dist(0, W(A))/w(A)$.

Proof. The proof is relatively simple. By multiplying A by some complex number of modulus 1, we have that the element of $W(A)$ closest to 0 is real positive and the GMRES residual

is the same. Define E to be the set $\{z \mid \text{Re } z \geq \text{dist}(0, W(A)), |z| \leq w(A)\}$. Then the result follows from Theorem 6.2.3 where the conformal map ϕ is taken to be the same as the one utilized in Lemma 6.1.2. □

In comparing the result of Theorem 6.2.3 with that of Theorem 6.1.1, we note two improvements. One is that we no longer have the constant $2 + 2/\sqrt{3}$, and the other is that the angle β' is now slightly smaller than before, due to the fact that $w(A) \leq ||A||$ (compare the definition of β in (6.0.2) to the β' in Corollary 6.2.4). The first of these changes is quite significant as any scaling factor greater than 1 will only dull the accuracy of any GMRES bound. The second change makes this new bound asymptotically sharper, as it results in a change in the asymptotic convergence factor.

Chapter 7. Numerical Experiments

Here we present some numerical experiments comparing the bounds derived above. The first bound is the one derived by Beckermann, et. al. [4] stated in Theorem 6.1.1. The second is Beckermann's improvement of Theorem 6.2.3 [3]. The third bound we compute is a bound where the asymptotic convergence factor γ is taken to be $\gamma = 1/|\phi(0)|$, where ϕ is the conformal map of the exterior of an approximating polygon to $W(A)$ to the exterior of the unit disk. This value is computed using the Schwarz-Christoffel toolbox [43]. Lastly, we also included Elman's orginal bound given in (6.0.1). We do this because the asymptotic convergence factor of Theorem 6.1.1 is only *asymptotically* sharper than Elman's, and thus Elman's bound can be more descriptive for the earlier iterations.

We should note that for the figures of this section, the numerical range was computed using a modified version of the m-file `fv.m`, which is found in the Matrix Computation Toolbox by Higham [31]. See Appendix B.

Iteration	Residual	Beckermann, et. al	Improved Beckermann	Conf. Map	Elman
1	0.191504	3.433643	1.088396	0.438420	0.707213
2	0.040493	1.528442	0.484476	0.087395	0.500150
3	0.006804	0.680366	0.215654	0.017421	0.353712
4	0.001220	0.302856	0.095994	0.003473	0.250150
5	0.000216	0.134813	0.042729	0.000692	0.176909
6	0.000039	0.060010	0.019020	0.000138	0.125112

Table 7.1: Numerical Results for the Shifted Toeplitz Matrix

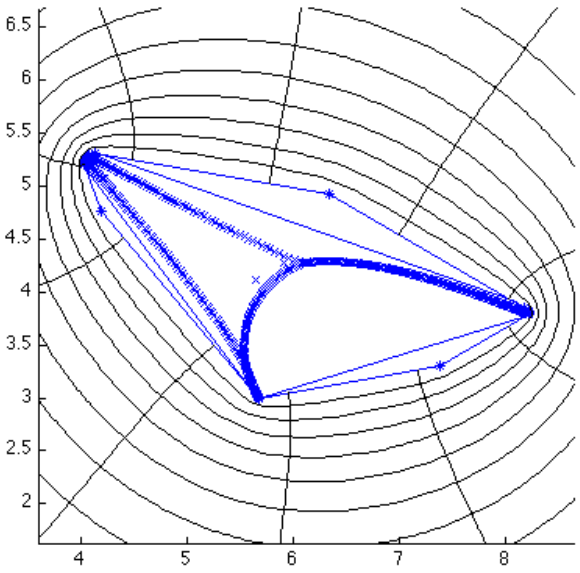
β	45.008570
β'	45.007772
γ_B	0.445137
γ_{imp}	0.445128
γ_C	0.199341
γ_E	0.707213

Table 7.2: Asymptotic Convergence Factors for the Shifted Toeplitz Matrix

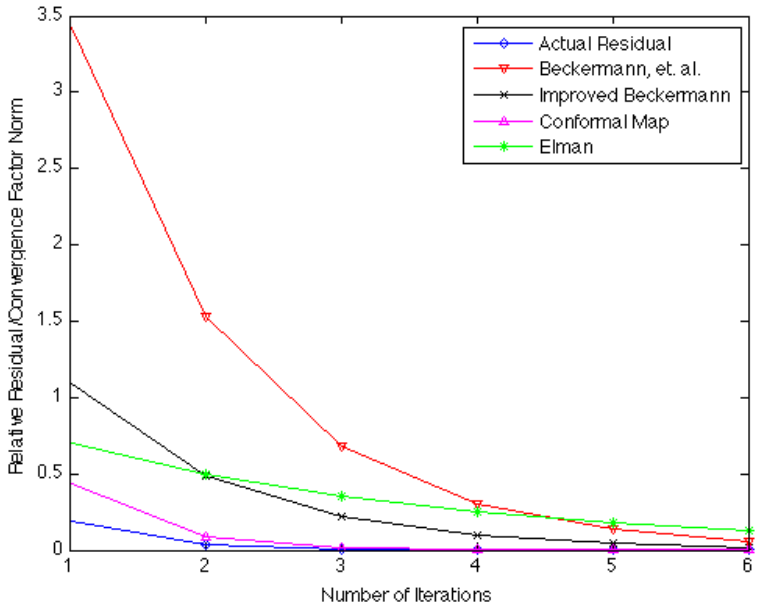
The first example is a 500×500 Toeplitz matrix, shifted so that $0 \notin W(A)$. The entries used to form the matrix were taken from the uniform distribution on the interval $[-1, 1]$. The right hand side was similarly generated. Figure 7.1(a) contains a plot of the numerical range with outer approximating polygon and conformal map lines computed by the Schwarz-Christoffel toolbox. The 'x's' indicate the location of the eigenvalues. Figure 7.1(b) is a plot of the various bounds along with the actual convergence curve of GMRES. Actual numerical values are recorded in Table 7.1.

Let γ_B, γ_{imp}, γ_C, and γ_E denote the asymptotic convergence factors for Beckermann, et. al., improved Beckermann, the one obtained via a conformal map, and Elman's, respectively. Table 7.2 lists the various convergence factors (denoted by γ) for each bound, as well as the angles β and β' for the Beckermann, et. al. bound and the improved Beckermann bound, respectively.

We see that the angles β and β' do not differ much, neither do the two Beckermann



(a) Numerical Range and Outer Approximating Polygon



(b) GMRES Convergence Curves

Figure 7.1: Numerical Experiment with a Shifted Toeplitz Matrix

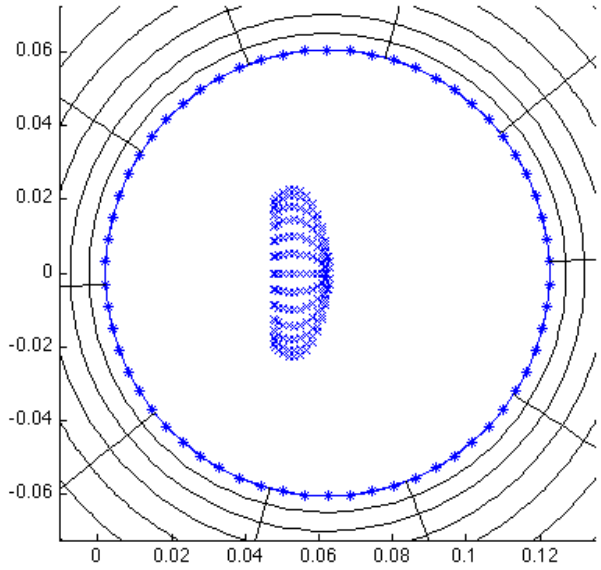
β	89.046713
β'	89.041155
γ_B	0.987210
γ_{imp}	0.987136
γ_C	0.969270
γ_E	0.999862

Table 7.3: Asymptotic Convergence Factors for the Convection-Diffusion Matrix

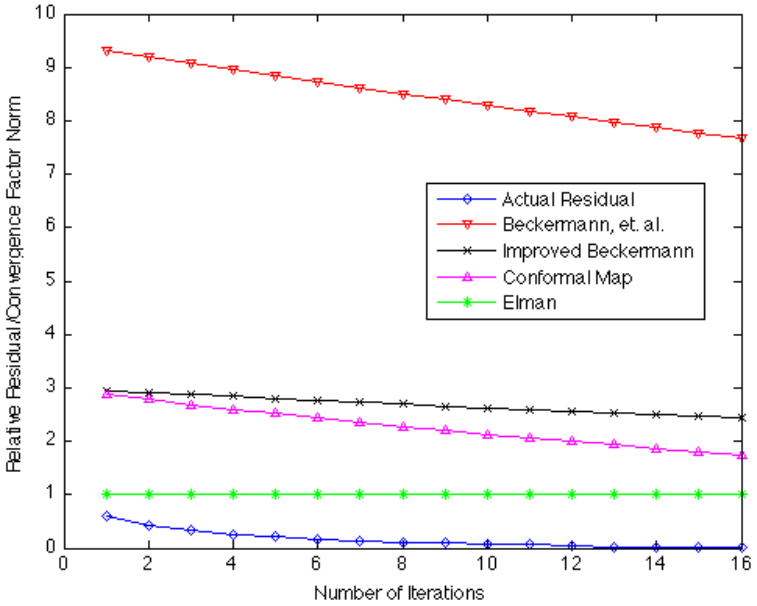
convergence factors. However, as indicated in Figure 7.1(b), Beckermann's improved bound greatly outperforms his original bound. This is due to the factor of $(2 + 2/\sqrt{3})$ which was eliminated in his improvement. The bound generated by the conformal map is quite promising, and seems to imply that if we can get an accurate enough approximation to the boundary of $W(A)$, and if we can compute the conformal map of the resulting polygon, then perhaps we can get descriptive estimates on the residuals.

Now we turn to an example, where all of these bounds do not give descriptive results. This example is derived from using a matrix which arises in the solution of the convection-diffusion problem, as described in [36]. The matrix generated was 169×169 and the righthand side vector was randomly generated as in the case for the Toeplitz matrix. Figure 7.2 contains the results.

This problem required 16 iterations and so we omit a table analogous to Table 7.1. By looking at Table 7.3, we see that the angles β and β' are both close to 90 degrees, which as mentioned earlier, indicates that $W(A)$ is close to the origin. This partially explains why this problem took longer to converge than the previous one. Also note that all of the convergence factors are close to 1, which is why none of them accurately describe the actual convergence behavior. This example shows that there are cases where bounds obtained from the numerical range can fail to be helpful. The take-away is that in order to fully understand GMRES, we may have to consider other characteristics of the matrix such as the spectrum and psuedospectrum together with the numerical range. A detailed work which



(a) Numerical Range and Outer Approximating Polygon



(b) GMRES Convergence Curves

Figure 7.2: Numerical Experiment with a Matrix Resulting from a Discretization of the Convection-Diffusion Equation

Matrix Order	Maximum Constant	Matrix Order	Maximum Constant
3	1.271	12	1.117
4	1.390	13	1.076
5	1.285	14	1.097
6	1.11	15	1.029
7	1.095	16	1.047
8	1.135	17	1.043
9	1.096	18	1.018
10	1.117	19	1.051
11	1.078	20	1.021

Table 7.4: Tests for the Crouzeix Constant

contains several examples comparing bounds obtained via the spectrum, pseudospectrum and numerical range can be found in [18].

No one has yet succeeded in proving or disproving Crouzeix's conjecture. The best constant found so far is 11.08 [8]. However, numerical experiments seem to confirm that the best constant is indeed 2. By considering the 2×2 matrix $A = \begin{pmatrix} 0 & 2 \\ 0 & 0 \end{pmatrix}$ and the polynomial $p(z) = z$, we see that the best constant cannot be any less than 2. This is because the numerical range of A is the unit disk and $||A|| = 2$. Table 7.4 lists some Crouzeix constants for matrices of orders 3 through 20. This experiment was done by randomly generating 20 matrices of each order and for each degree 1 through 10, computing $||p(A)|| / \sup_{z \in W(A)} |p(z)|$ for 20 randomly generated polynomials. The denominator was computed by taking the maximum of $|p(z)|$ over an approximation to the boundary of $W(A)$. The maximum Crouzeix constant for each matrix order is listed in Table 7.4.

Some interesting observations from these numbers are the fact that the maximum Crouzeix constant seems to decrease as the matrix order increases. Also, the maximum Crouzeix constant was, in most cases, obtained for a 1 or 2 degree polynomial. This would seem to indicate that if a counterexample is to be found, it probably lies in considering a matrix of small dimension with a low degree polynomial.

Chapter 8. Conclusion

Considering the existing research done in the area of GMRES convergence, the Crouzeix conjecture offers an attractive way to quantify the behavior of these iterations. Despite the improvements made by Beckermann that do not use the conjecture, the numerical experiments show there is still much room for improvement in finding a truly descriptive bound. Trefethen's work on psuedospectral bounds shows that there are different phases of GMRES convergence, which may lead some to think that a combination of bounds obtained from the spectrum, psuedospectrum, and numerical range are necessary to truly accurately describe GMRES convergence. However, we feel that there is much investigation to be done in the area of numerical range bounds itself. One major obstacle is that these bounds require the computation of the numerical range, which for large matrices (such as those larger than 1000×1000), can take some time to compute. An efficient algorithm for computing the numerical range of a matrix is thus in order. Secondly, as mentioned in Chapter 4, there are many ways we can seek to bound the quantity $\sup_{z \in W(A)} |p(z)|$. As far as we know, there is no major work focused on this effort in the context of GMRES analysis. Perhaps looking into these possibilities will provide further insight. On the other hand, if $0 \in W(A)$, then the Crouzeix conjecture will not be helpful, since by previous remarks, any polynomial satisfying $p(0) = 1$ at the origin must also satisfy $|p(z)| \geq 1$ on the boundary of $W(A)$. This suggests that it is also insufficient to solely consider bounds obtained from the numerical range.

However, regardless of how well we can approximate $\sup_{z \in W(A)} |p(z)|$, the question of whether or not Crouzeix's conjecture is true still remains to be answered. Numerical experiments seem to indicate that it is true, but we must careful in drawing conclusions from such experiments. By only considering randomly generated matrices, we are quite possibly excluding certain matrices with special structures that may provide a counterexample. Be that

as it may, we currently have little intuition about what type matrices could possibly prove Crouzeix's conjecture wrong. Thus this still remains an area which holds many research possibilities and opportunities for greater understanding.

Bibliography

[1] T. Ando. Structure of operators with numerical radius one. *Acta Sci. Math. (Szeged)*, 34:11–15, 1973.

[2] Catalin Badea, Michel Crouzeix, and Bernard Delyon. Convex domains and K-spectral sets. *Math. Z.*, 252(2):345–365, 2006.

[3] B. Beckermann. Image numerique, GMRES et polynomes de faber. *C. R. Acad. Sci. Paris, Ser. I*, 340:855–860, 2005.

[4] B. Beckermann, S. A. Goreinov, and E. E. Tyrtyshnikov. Some remarks on the Elman estimate for GMRES. *SIAM J. Matrix Anal. Appl.*, 27(3):772–778 (electronic), 2005.

[5] S. L. Campbell, I. C. F. Ipsen, C. T. Kelley, and C. D. Meyer. GMRES and the minimal polynomial. *BIT*, 36(4):664–675, 1996.

[6] C. Cowen and E. Harel. An Effective Algorithm for Computing the Numerical Range. August 1995.

[7] Michel Crouzeix. Bounds for analytical functions of matrices. *Integral Equations Operator Theory*, 48(4):461–477, 2004.

[8] Michel Crouzeix. Numerical range and functional calculus in Hilbert space. *J. Funct. Anal.*, 244(2):668–690, 2007.

[9] Michel Crouzeix and Bernard Delyon. Some estimates for analytic functions of strip or sectorial operators. *Arch. Math. (Basel)*, 81(5):559–566, 2003.

[10] J. H. Curtiss. Faber polynomials and the Faber series. *Amer. Math. Monthly*, 78:577–596, 1971.

[11] John Hamilton Curtiss. *Introduction to functions of a complex variable*, volume 44 of *Monographs and Textbooks in Pure and Applied Math.* Marcel Dekker Inc., New York, 1978. With a foreword by E. F. Beckenbach.

[12] Bernard Delyon and François Delyon. Generalization of von Neumann's spectral sets and integral representation of operators. *Bull. Soc. Math. France*, 127(1):25–41, 1999.

[13] Tobin A. Driscoll and Lloyd N. Trefethen. *Schwarz-Christoffel mapping*, volume 8 of *Cambridge Monographs on Applied and Computational Mathematics*. Cambridge University Press, Cambridge, 2002.

[14] Michael Eiermann. On semiiterative methods generated by Faber polynomials. *Numer. Math.*, 56(2-3):139–156, 1989.

[15] Michael Eiermann. Fields of values and iterative methods. *Linear Algebra Appl.*, 180:167–197, 1993.

[16] Stanley C. Eisenstat, Howard C. Elman, and Martin H. Schultz. Variational iterative methods for nonsymmetric systems of linear equations. *SIAM J. Numer. Anal.*, 20(2):345–357, 1983.

[17] H.C. Elman. *Iterative Methods for Large, Sparse, Nonsymmetric Systems of Linear Equations*. PhD thesis, Yale University, 1982.

[18] Mark Embree. How Descriptive are GMRES Convergence Bounds? Technical report, 1999.

[19] Georg Faber. Über polynomische Entwickelungen. *Math. Ann.*, 57(3):389–408, 1903.

[20] Nabil Gmati and Bernard Philippe. Comments on the GMRES convergence for preconditioned systems. In *Large-scale scientific computing*, volume 4818 of *Lecture Notes in Comput. Sci.*, pages 40–51. Springer, Berlin, 2008.

[21] A. Greenbaum and L. Gurvits. Max-min properties of matrix factor norms. *SIAM J. Sci. Comput.*, 15(2):348–358, 1994. Iterative methods in numerical linear algebra (Copper Mountain Resort, CO, 1992).

[22] A. Greenbaum and Strakoš, Z. Matrices that generate the same Krylov residual spaces. In *Recent Advances in Iterative Methods*, pages 95–118. Springer-Verlag, Berlin, New York, 1994.

[23] Anne Greenbaum. *Iterative methods for solving linear systems*, volume 17 of *Frontiers in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1997.

[24] Anne Greenbaum, Vlastimil Pták, and Zdeněk Strakoš. Any nonincreasing convergence curve is possible for GMRES. *SIAM J. Matrix Anal. Appl.*, 17(3):465–469, 1996.

[25] Anne Greenbaum and Lloyd N. Trefethen. GMRES/CR and Arnoldi/Lanczos as matrix approximation problems. *SIAM J. Sci. Comput.*, 15(2):359–368, 1994. Iterative methods in numerical linear algebra (Copper Mountain Resort, CO, 1992).

[26] Karl E. Gustafson and Duggirala K. M. Rao. *Numerical range*. Universitext. Springer-Verlag, New York, 1997. The field of values of linear operators and matrices.

[27] Paul R. Halmos. *A Hilbert space problem book*. D. Van Nostrand Co., Inc., Princeton, N.J.-Toronto, Ont.-London, 1967.

[28] Felix Hausdorff. Der Wertvorrat einer Bilinearform. *Math. Z.*, 3(1):314–316, 1919.

[29] Peter Henrici. *Applied and computational complex analysis. Vol. 3.* Pure and Applied Mathematics (New York). John Wiley & Sons Inc., New York, 1986. Discrete Fourier analysis—Cauchy integrals—construction of conformal maps—univalent functions, A Wiley-Interscience Publication.

[30] Peter Henrici. *Applied and computational complex analysis. Vol. 1.* Wiley Classics Library. John Wiley & Sons Inc., New York, 1988. Power series—integration—conformal mapping—location of zeros, Reprint of the 1974 original, A Wiley-Interscience Publication.

[31] Nicholas J. Higham. The Matrix Computation Toolbox. `http://www.ma.man.ac.uk/~higham/mctoolbox`.

[32] Roger A. Horn and Charles R. Johnson. *Topics in Matrix Analysis*. Cambridge University Press, Cambridge, 1991.

[33] Ilse C. F. Ipsen and Carl D. Meyer. The idea behind Krylov methods. *Amer. Math. Monthly*, 105(10):889–899, 1998.

[34] Tosio Kato. Some mapping theorems for the numerical range. *Proc. Japan Acad.*, 41:652–655, 1965.

[35] T. Kövari and Ch. Pommerenke. On Faber polynomials and Faber expansions. *Math. Z.*, 99:193–206, 1967.

[36] J. Liesen and Z. Strakoš. GMRES convergence analysis for a convection-diffusion model problem. *SIAM J. Sci. Comput.*, 26(6):1989–2009 (electronic), 2005.

[37] A. I. Markushevich. *Theory of functions of a complex variable. Vol. III.* Revised English edition, translated and edited by Richard A. Silverman. Prentice-Hall Inc., Englewood Cliffs, N.J., 1967.

[38] Jerrold E. Marsden and Michael J. Hoffman. *Basic complex analysis*. W. H. Freeman and Company, New York, second edition, 1987.

[39] J. C. Mason and D. C. Handscomb. *Chebyshev polynomials*. Chapman & Hall/CRC, Boca Raton, FL, 2003.

[40] Panayiotis J. Psarrakos and Michael J. Tsatsomeros. Numerical range: (in) a matrix nutshell. May 2002.

[41] Youcef Saad and Martin H. Schultz. GMRES: a generalized minimal residual algorithm for solving nonsymmetric linear systems. *SIAM J. Sci. Statist. Comput.*, 7(3):856–869, 1986.

[42] E.B. Saff and A.D. Snider. *Fundamentals of Complex Analysis*. Pearson Education, Inc., New Jersey, 2003.

[43] DRISCOLL T. A matlab toolbox for schwarz-christoffel mapping. *ACM Trans. on Mathematical Software*, 22:168–186, 1996.

[44] Otto Toeplitz. Das algebraische Analogon zu einem Satze von Fejér. *Math. Z.*, 2(1-2):187–197, 1918.

[45] Kim C Toh. Matrix approximation problems and nonsymmetric iterative methods. Technical report, Ithaca, NY, USA, 1996.

[46] Kim-Chuan Toh. GMRES vs. ideal GMRES. *SIAM J. Matrix Anal. Appl.*, 18(1):30–36, 1997.

[47] L. N. Trefethen. Approximation theory and numerical linear algebra. In *Algorithms for approximation, II (Shrivenham, 1988)*, pages 336–360. Chapman and Hall, London, 1990.

[48] Lloyd N. Trefethen and David Bau, III. *Numerical linear algebra*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1997.

[49] Lloyd N. Trefethen and Mark Embree. *Spectra and pseudospectra*. Princeton University Press, Princeton, NJ, 2005. The behavior of nonnormal matrices and operators.

Appendix A. Cowen's Numerical Range Code

This is code used to produce the figures in Chapter 2.

```
%
% This script finds the numerical range of an n x n matrix by
% finding the real and imaginary parts of rotates of the matrix
% and finding the associated boundary point of that rotate by
% finding the largest eigenvalue of the real part and using the
% corresponding eigenvector?s contribution to the numerical range.
% Multiplicity of the largest eigenvalue, as occurs in a normal
% matrix, is handled by plotting the end points of the corresponding
% segment in the boundary of the numerical range.
%
function w = getNumRange(A,plotrange,ploteigvals)
close all;
nm=ceil(norm(full(A)));
th= 0:.01:6.29;
k=1;
w=zeros(1,length(th));
for j=1:length(th)
    Ath=(exp(1i*(-th(j))))*A;
    Hth=(Ath+Ath')/2;
    [r e]=eigs(Hth);
    e=real(diag(e));
```

```
    %e=round(e);
    m=max(e);
s=find(e==m);
if size(s,1)==1
    w(k)=r(:,s)'*A*r(:,s);
%
%   This is the point of the numerical range contributed by
%   v_t=r(:,s) when the eigenspace of Hth (H_t) is one dimensional.
%
else
    Kth=1i*(Hth-Ath);
    pKp=r(:,s)'*Kth*r(:,s);
    %
    % The matrix Q described above is r(:,s)
    %
    [rr ee]=eig(pKp);
    ee=real(diag(ee));

    mm=min(ee);
    sm=find(ee==mm);
    w(k)=rr(:,sm(1,:))'*r(:,s)'*A*r(:,s)*rr(:,sm(1,:));
    %
    %   This is the point of the numerical range contributed by
    % v_t^- = r(:,s)*rr(:,sm(:,1))
    %
    k=k+1;
```

```
    mM=max(ee);
    sM=find(ee==mM);
    w(k)=rr(:,sM(1,:))'*r(:,s)'*A*r(:,s)*rr(:,sM(1,:));
    %
    %   This is the point of the numerical range contributed by v_t^+
    %
end
    k=k+1;
end
if(plotrange)
figure
H = fill(real(w),imag(w),'y');
set(H, 'LineWidth', 2);
if ploteigvals
    hold on
eigvals = eigs(A);
plot(real(eigvals),imag(eigvals),'r*');
end
axis([-nm,nm,-nm,nm]);
axis('equal');


end


end
```

Appendix B. Higham's Numerical Range Code

This code was used to produce the figures in Chapter 7.

```
function [f, e,vertices] = fvpolyarea(B, tol,plotme)
%FV     Field of values (or numerical range).
%       FV(A, NK, THMAX) evaluates and plots the field of values of the
%       NK largest leading principal submatrices of A, using THMAX
%       equally spaced angles in the complex plane.
%       The defaults are NK = 1 and THMAX = 16.
%       (For a 'publication quality' picture, set THMAX higher, say 32.)
%       The eigenvalues of A are displayed as 'x'.
%       Alternative usage: [F, E] = FV(A, NK, THMAX, 1) suppresses the
%       plot and returns the field of values plot data in F, with A's
%       eigenvalues in E.   Note that NORM(F,INF) approximates the
%       numerical radius,
%                 max {abs(z): z is in the field of values of A}.

%       Theory:
%       Field of values FV(A) = set of all Rayleigh quotients. FV(A) is a
%       convex set containing the eigenvalues of A.  When A is normal FV(A) is
%       the convex hull of the eigenvalues of A (but not vice versa).
%               z = x'Ax/(x'x),  z' = x'A'x/(x'x)
%               => REAL(z) = x'Hx/(x'x),   H = (A+A')/2
%       so      MIN(EIG(H)) <= REAL(z) <= MAX(EIG(H)),
```

```
%       with equality for x = corresponding eigenvectors of H.  For these x,
%       RQ(A,x) is on the boundary of FV(A).
%
%       Based on an original routine by A. Ruhe.
%
%       References:
%       R. A. Horn and C. R. Johnson, Topics in Matrix Analysis, Cambridge
%            University Press, 1991; sec. 1.5.
%       A. S. Householder, The Theory of Matrices in Numerical Analysis,
%            Blaisdell, New York, 1964; sec. 3.3.
%       C. R. Johnson, Numerical determination of the field of values of a
%            general complex matrix, SIAM J. Numer. Anal., 15 (1978),
%            pp. 595-602.

%close all
%figure
%hold on
%if nargin < 2 | isempty(nk), nk = 1; end
%if nargin < 3 | isempty(thmax), thmax = 16; end
thmax = 2;
areaDiff = tol*3;

while areaDiff > tol
thmax = 2*thmax;
thmax = thmax - 1;  % Because code below uses thmax + 1 angles.
```

```
iu = sqrt(-1);
[n, p] = size(B);
if n ~= p, error('Matrix must be square.'), end
f = [];
z = zeros(2*thmax+1,1);
evals = zeros(2*thmax+1,1);
vertices = zeros(2*thmax,1);
inc = pi/thmax;
e = eig(B);

% Filter out cases where B is Hermitian or skew-Hermitian, for efficiency.
if isequal(B,B')

   f = [min(e) max(e)];

elseif isequal(B,-B')

   e = imag(e);
   f = [min(e) max(e)];
   e = iu*e; f = iu*f;

else

    %for m = 1:nk

      % ns = n+1-m;
```

```
%A = B(1:ns, 1:ns);
A=B;

for i = 0:thmax
   th = i/thmax*pi;
   Ath = exp(iu*th)*A;               % Rotate A through angle th.
   H = 0.5*(Ath + Ath');             % Hermitian part of rotated A.
   [X, D] = eig(H);
   [lmbh, k] = sort(real(diag(D)));
   evals(1+i) = lmbh(1);
   evals(1+i+thmax) = lmbh(end);
   z(1+i) = rq(A,X(:,k(1)));         % RQ's of A corr. to eigenvalues of H
   z(1+i+thmax) = rq(A,X(:,k(n)));  % with smallest/largest real part.
   if i >= 1  % we can start to compute vertices
       theta1 = (i-1)/thmax*pi;
     v = (evals(i)*cos(inc) - evals(i+1))/sin(inc);
     vertices(i) = exp(-1i*theta1)*(evals(i) + 1i*v);

     theta2 = (i-1)/thmax*pi;
     v = (evals(i+thmax)*cos(inc) - evals(i+1+thmax))/sin(inc);
     vertices(i+thmax) = exp(-1i*theta2)*(evals(i+thmax) + 1i*v);

   end
end
% now calculate difference in the area of the polygons
p=vertices;
```

```
        p = [p; p(1,:)];
        q = [z(1:end-1,:); z(1,:)];

        s1=0;
        s2=0;
        for k=1:length(p)-1
         s1=s1+q(k)'*q(k+1);
         s2=s2+p(k)'*p(k+1);
        end

        areaDiff=0.5*imag(s1-s2);

end % end while
if plotme
    figure, plot(real(q),imag(q),'r*');
    hold on
  plot(real(vertices),imag(vertices),'b*');
end
        f = [f; z];

    % Next line ensures boundary is 'joined up' (needed for orthogonal matrices).
    %f = [f; f(1,:)];
    f(end,:) = f(1,:);

end
if thmax == 0; f = e; end
```

```
if plotme

   ax = cpltaxes(f);
   plot(real(f), imag(f))      % Plot the field of values
   axis(ax);
   axis('square');

   hold on
   plot(real(e), imag(e), 'x')    % Plot the eigenvalues too.
   hold off

end

function z = rq(A,x)
%RQ      Rayleigh quotient.
%       RQ(A,x) is the Rayleigh quotient of A and x, x'*A*x/(x'*x).

z = x'*A*x/(x'*x);
```