2013-07-05

Bounding the Norm of Matrix Powers

Daniel Ammon Dowler
Brigham Young University - Provo

Bounding the Norm of Matrix Powers

Daniel A. Dowler

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 W. Barrett
Scott Glasgow

Department of Mathematics

Brigham Young University

July 2013

ABSTRACT

Bounding the Norm of Matrix Powers

Daniel A. Dowler
Department of Mathematics, BYU
Master of Science

In this paper I investigate properties of square complex matrices of the form $\mathbf{A}^k$, where $\mathbf{A}$ is also a complex matrix, and k is a nonnegative integer. I look at several ways of representing $\mathbf{A}^k$. In particular, I present an identity expressing the k^{th} power of the Schur form $\mathbf{T}$ of $\mathbf{A}$ in terms of the elements of $\mathbf{T}$, which can be used together with the Schur decomposition to provide an expression of $\mathbf{A}^k$. I also explain bounds on the norm of $\mathbf{A}^k$, including some based on the element-based expression of $\mathbf{T}^k$. Finally, I provide a detailed exposition of the most current form of the Kreiss Matrix Theorem.

Keywords: Matrix Powers, Matrix Norm Bounds, Matrix Power Bounds, Kreiss Matrix Theorem, Schur Decomposition, Schur Form

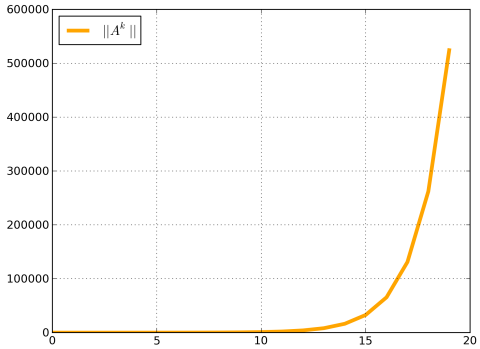
Contents

List of Figures

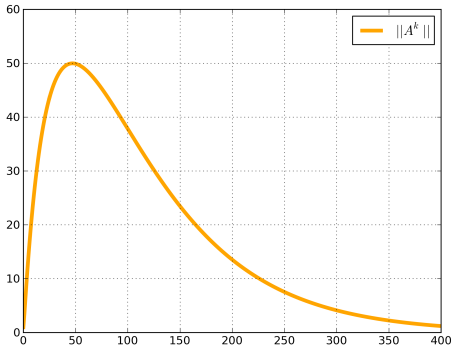
What happens when you take a square matrix, $\mathbf{A}$, raise it to an integer power k, and then take the norm, $||\mathbf{A}^k||$? Below there are four matrices and four corresponding graphs of the norm (2-norm) of their powers. Can you match each matrix to the correct graph of the norm of its powers? We will revisit this question again later in the thesis. For now, note how similar the matrices are, yet how differently the graphs of their norms behave.

Figure 1: Examples of Norms of Matrix Powers

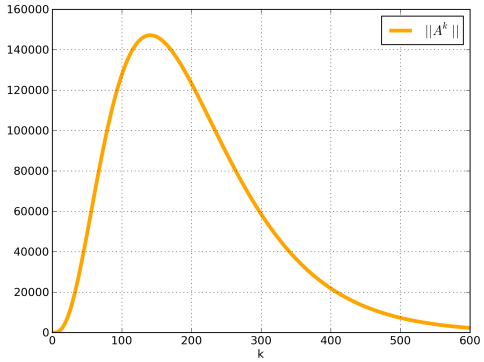
1.



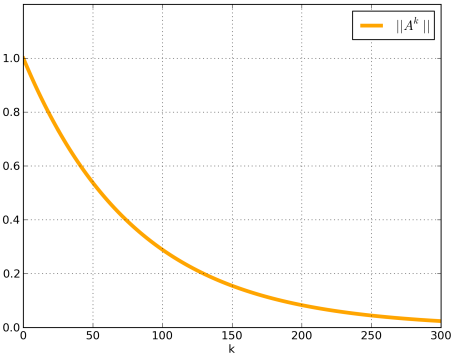
2.



3.



4.



$$\mathbf{V} = \begin{bmatrix} 0 & 2 & 0 & 0 \\ 2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 2 \end{bmatrix}$$

$$\mathbf{W} = \begin{bmatrix} \frac{20}{21} & 0 & 2 & 2 \\ 0 & \frac{40}{41} & 0 & 2 \\ 0 & 0 & \frac{60}{61} & 0 \\ 0 & 0 & 0 & \frac{80}{81} \end{bmatrix}$$

$$\mathbf{X} = \begin{bmatrix} \frac{20}{21} & 2 & 0 & 0 \\ 0 & \frac{40}{41} & 2 & 0 \\ 0 & 0 & \frac{60}{61} & 2 \\ 0 & 0 & 0 & \frac{80}{81} \end{bmatrix}$$

$$\mathbf{Y} = \begin{bmatrix} \frac{20}{21} & 0 & 0 & 0 \\ 0 & \frac{40}{41} & 0 & 0 \\ 0 & 0 & \frac{60}{61} & 0 \\ 0 & 0 & 0 & \frac{80}{81} \end{bmatrix}$$

Chapter 1. Introduction

Before we begin discussing norms of matrix powers in detail, perhaps we should begin by asking what happens when you take a square complex matrix $\mathbf{A}$, and multiply it by itself over and over again to get $\mathbf{A}^k$? The answer depends on various properties of $\mathbf{A}$. The eigenvalues of $\mathbf{A}$ will be of central importance. In particular, we shall see that if the spectral radius is greater than one, then at least some of the entries of $\mathbf{A}^k$ will diverge in magnitude toward infinity as the power k increases. On the other hand, if the spectral radius is less than one, all of the entries of $\mathbf{A}^k$ will converge to 0 as k tends to infinity. The resolvent of $\mathbf{A}$ will also be of great importance. Once this is defined, we will be able to express $\mathbf{A}^k$ as a complex integral. We will be able to use properties of the resolvent to derive the Kreiss Matrix Theorem (Theorem 3.17), a fundamental result of numerical linear algebra.

Furthermore, matrix normality—or lack thereof—will also play a key role. We will see that if $\mathbf{A}$ is a normal matrix, then both $\mathbf{A}$ and $\mathbf{A}^k$ may be decomposed into products of unitary and diagonal matrices; whereas if $\mathbf{A}$ is a nonnormal matrix, then $\mathbf{A}$ and $\mathbf{A}^k$ may only be decomposed as products of unitary and upper-triangular matrices. These results will follow from the Spectral Theorem (Theorem 1.4) and Schur's Unitary Triangularization Theorem (Theorem 1.5). The difference in these decompositions, leads to rather different approaches in determining the norm of $\mathbf{A}^k$. When $\mathbf{A}$ is normal, the norm of $\mathbf{A}^k$ is rather easy to compute; however, if $\mathbf{A}$ is nonnnormal, the difficulty of direct computation is often so high that we opt for finding a bound, rather than an exact value. As we will see, the process of developing bounds for the norm $||\mathbf{A}^k||$ in the case where $\mathbf{A}$ is nonnormal, is no trivial pursuit.

In fact, bounds for $||\mathbf{A}^k||$ are part of an active area of mathematical research. In this paper I present several new bounds based on entries from the Schur form of the matrix $\mathbf{A}$. I also provide a contiguous exposition of the most current version of the Kreiss Matrix Theorem (Theorem 3.17), whose proof was completed in 1991 with the addition of Spijker's Lemma (Lemma 3.21). The theorem gives a lower and upper bound for the supremum of the norm $||\mathbf{A}^k||$, for all nonnegative integers k. As the theorem (originally by Heinz-Otto Kreiss [3]) has gone through many iterations since its debut in 1962, a complete presentation including both the theorem and its proof in the most current form is lacking in the literature; yet, due to its importance to both theoretical and

applied mathematics, a unified exposition is quite worthwhile.

In Chapter 4 we revisit Figure 1, to analyze bounds on the norms of powers of the given matrices, and to finally answer the question of which graph goes with which matrix. Next, we discuss an example of where the norm of matrix powers is used to determine error in Gauss-Seidel iteration. Then we investigate the effects of nonnormality in a Markov process. Finally, we look at a random upper triangular matrix with a particular structure.

1.1 Acknowledgments

While I researched and wrote the material that follows, I have certainly been inspired, guided, and supported by so many others. Many thanks go to my advisor Jeffrey Humpherys for guiding me along the way. I would also like to thank my committee members Wayne Barrett and Scott Glasgow for their time and insights in evaluating my work. My wife, Marquita, and my wonderful children deserve special recognition for supporting me and loving me through the long hours of study and work that have gone into the pursuit of a master's degree. I want to thank all of my professors who have believed in me, and who have labored to teach me. There are many other friends, family members, and fellow students who have made my journey brighter. Thank you to all of you for your friendship and support.

1.2 Eigenvalues and Normal Matrices

We begin with an $N \times N$ matrix $\mathbf{A}$ with entries from $\mathbb{C}$, the set of complex numbers. We will denote the set of all such matrices by M_N, where the dimension N is from $\mathbb{N}$, the set of natural numbers. Let $\mathbf{A}^*$ denote the conjugate transpose of $\mathbf{A}$ ($\mathbf{A}^* = (\overline{\mathbf{A}})^T$). Likewise, given $\mathbf{u} \in \mathbb{C}^N$, where $\mathbb{C}^N$ denotes the set of complex N-tuples, let $\mathbf{u}^*$ denote its conjugate transpose. Given a matrix $\mathbf{A}$ in M_N, the value λ in $\mathbb{C}$ is an ***eigenvalue*** of $\mathbf{A}$ if and only if there exists a nonzero $N \times 1$ vector $\mathbf{x}$ in $\mathbb{C}^N$, such that $\mathbf{A}\mathbf{x} = \lambda\mathbf{x}$. The vector $\mathbf{x}$ is referred to as an ***eigenvector*** corresponding to λ.

The ***spectrum*** of the matrix $\mathbf{A}$ in M_N, denoted $\sigma(\mathbf{A})$, is the set of all eigenvalues of $\mathbf{A}$. The spectrum is a subset of $\mathbb{C}$, and it is nonempty. By rearranging the equation $\mathbf{A}\mathbf{x} = \lambda\mathbf{x}$, we may

write the spectrum as the set,

$$\sigma(\mathbf{A}) = \{\lambda \in \mathbb{C} : (\lambda\mathbf{I} - \mathbf{A})\mathbf{x} = \mathbf{0} \text{ for some } \mathbf{x} \in \mathbb{C}^N \backslash \{\mathbf{0}\}\}$$

or equivalently,

$$\sigma(\mathbf{A}) = \{\lambda \in \mathbb{C} : \det(\lambda\mathbf{I} - \mathbf{A}) = 0\}$$

Here, 'det' is the determinant operator. Note that $\sigma(\mathbf{A}) \neq \emptyset$, because each matrix in M_N has at least one eigenvalue. The ***spectral radius*** of $\mathbf{A} \in M_N$ is the largest magnitude attained by any eigenvalue of $\mathbf{A}$:

$$\rho(\mathbf{A}) := \sup\{|\lambda| : \lambda \in \sigma(\mathbf{A})\}$$

The ***characteristic polynomial*** p of $\mathbf{A} \in M_N$ is the complex polynomial defined by,

$$\mathrm{p}(z) := \det(z\mathbf{I} - \mathbf{A})$$

To see that this is actually a polynomial, a simple inductive proof is in order. If N is 1, then $\mathrm{p}(z) = z - a$, which is a polynomial in z of degree 1. Now inductively, suppose that we have shown that $\det(z\mathbf{I} - \mathbf{A})$ is in fact a polynomial of degree N, for some integer $N \geq 1$, and for all $\mathbf{A} \in M_N$. Choose $\mathbf{B} = (b_{mn}) \in M_{N+1}$. By cofactor expansion along the first row of $\mathbf{B}$,

$$\det(z\mathbf{I} - \mathbf{B}) = (z - b_{11})C_{11} - \sum_{j=1}^{N} b_{1j}C_{1j}$$

where C_{1j} is the $1j^{th}$ cofactor of $\mathbf{B}$, for $j = 1, 2, \ldots, N$. Recall that C_{1j} is defined as $(-1)^{1+j}$ times the determinant of the $N \times N$ matrix obtained from $\mathbf{B}$ by deleting the 1^{st} row and j^{th} column. Thus, C_{1j} is a polynomial in z of degree N, by the inductive hypothesis. In particular, $(z - b_{11})C_{11}$ is a polynomial of degree $(N + 1) \times (N + 1)$. Subtracting each of the other terms in the cofactor expansion will still yield a polynomial of degree $(N + 1) \times (N + 1)$. The roots of the characteristic polynomial of a matrix $\mathbf{A}$ are the eigenvalues of $\mathbf{A}$; this is rather straightforward to show based on the definition of the spectrum.

As with any other polynomial, we may find coefficeints $c_0, c_1, \ldots, c_N$, such that the characteristic polynomial p of the matrix $\mathbf{A} \in M_N$ may be written as $\mathrm{p}(z) = c_N z^N + c_{N-1} z^{N-1} + \ldots + c_1 z + c_0$.

Setting this equal to $\det(z\mathbf{I} - \mathbf{A})$, and evaluating both sides at $z = 0$ gives,

$$c_0 = \det(-\mathbf{A}) = (-1)^N \det(\mathbf{A}) \tag{1.1}$$

Though we will not prove it here, it turns out that the coefficient c_N is 1, which makes the characteristic polynomial monic, and thus unique. Another useful result involving the characteristic polynomial is the Cayley-Hamilton Theorem.

Theorem 1.1 (Cayley-Hamilton Theorem)**.** *Let $\boldsymbol{A} \in M_N$ have characteristic polynomial $p(x) = x^N + c_{N-1}x^{N-1} + \ldots + c_1 x + c_0$. Then,*

$$p(\boldsymbol{A}) = \boldsymbol{A}^N + c_{N-1}\boldsymbol{A}^{N-1} + \ldots + c_1\boldsymbol{A} + c_0\boldsymbol{I} = \boldsymbol{0}$$

The proof of the Cayley-Hamilton Theorem may be found in the book *Matrix Analysis*, by Horn and Johnson [2, p. 86].

Theorem 1.2. *Given $\boldsymbol{A} \in M_N$ with all eigenvalues equal to zero, $\boldsymbol{A}^N = \boldsymbol{0}$.*

Proof. Since the eigenvalues of $\mathbf{A}$ are zero, the characteristic polynomial of $\mathbf{A}$ is given by $\mathrm{p}(x) = x^N$. By the Cayley-Hamilton Theorem,

$$\mathbf{A}^N = \mathrm{p}(\mathbf{A}) = \mathbf{0}$$

□

A matrix $\mathbf{A} \in M_N$ is said to be ***normal*** if and only if $\mathbf{A}^*\mathbf{A} = \mathbf{A}\mathbf{A}^*$. All matrices without this property are said to be ***nonnormal***. A normal matrix $\mathbf{U}$ is ***unitary*** if $\mathbf{U}^*\mathbf{U} = \mathbf{U}\mathbf{U}^* = \mathbf{I}$, or equivalently, if the columns of $\mathbf{U}$ form an orthonormal set. Normal matrices are extremely useful, since they can be expressed in terms of their eigenvalues and eigenvectors as shown in Theorem 1.4.

Lemma 1.3. *If $\boldsymbol{A} \in M_N$ is normal and invertible, then $\boldsymbol{A}^{-1}$ is also normal.*

Proof. First, note that

$$\mathbf{A}\mathbf{A}^{-1} = \mathbf{I} \quad \Longrightarrow \quad (\mathbf{A}^{-1})^*\mathbf{A}^* = \mathbf{I}^* = \mathbf{I} \quad \Longrightarrow \quad (\mathbf{A}^{-1})^* = (\mathbf{A}^*)^{-1}$$

Thus the above, coupled with the normality of $\mathbf{A}$ implies that,

$$\mathbf{A}^{-1}\left(\mathbf{A}^{-1}\right)^{*} = \mathbf{A}^{-1}\left(\mathbf{A}^{*}\right)^{-1} = \left(\mathbf{A}^{*}\mathbf{A}\right)^{-1} = \left(\mathbf{A}\mathbf{A}^{*}\right)^{-1} = \left(\mathbf{A}^{*}\right)^{-1}\mathbf{A}^{-1} = \left(\mathbf{A}^{-1}\right)^{*}\mathbf{A}^{-1}$$

Therefore, $\mathbf{A}^{-1}$ is normal. □

A matrix $\mathbf{A} \in M_N$ is **Hermitian** if $\mathbf{A} = \mathbf{A}^{*}$ and it is **symmetric** if $\mathbf{A} = \mathbf{A}^{T}$. The reader can verify that real symmetric matrices (that is, symmetric matrices with entries from the reals) are Hermitian, and that Hermitian matrices are normal. Additionally, all Hermitian matrices have the convenience of real eigenvalues. Note that if λ is an eigenvalue of the matrix $\mathbf{A}$, with corresponding eigenvector $\mathbf{x}$, then $\mathbf{A}\mathbf{x} = \lambda\mathbf{x}$, which implies that

$$\mathbf{x}^{*}\mathbf{A}^{*} = \left(\mathbf{A}\mathbf{x}\right)^{*} = \left(\lambda\mathbf{x}\right)^{*} = \overline{\lambda}\mathbf{x}^{*}$$

In addition, if we assume that $\mathbf{A}$ is Hermitian, the following holds:

$$\overline{\lambda}\mathbf{x}^{*}\mathbf{x} = \mathbf{x}^{*}\mathbf{A}^{*}\mathbf{x} = \mathbf{x}^{*}\mathbf{A}\mathbf{x} = \mathbf{x}^{*}\lambda\mathbf{x} = \lambda\mathbf{x}^{*}\mathbf{x}.$$

Since eigenvectors are nonzero, $\mathbf{x}^{*}\mathbf{x}$ is also nonzero, which gives $\bar{\lambda} = \lambda$.

Regardless of whether $\mathbf{A} \in M_N$ is normal or not, it is easy to prove that the matrices $\mathbf{A}^{*}\mathbf{A}$ and $\mathbf{A}\mathbf{A}^{*}$ will always be Hermitian (and hence normal) with real eigenvalues. Moreover, they are positive semidefinite, and thus, the eigenvalues are nonnegative. In fact, $\mathbf{A}^{*}\mathbf{A}$ and $\mathbf{A}\mathbf{A}^{*}$ always have the same nonzero eigenvalues with the same multiplicities. Due to their importance, the square roots of the eigenvalues of $\mathbf{A}\mathbf{A}^{*}$ are referred to as ***singular values*** of the matrix $\mathbf{A}$. Since singular values are always real, they can be ordered; and we are often interested in the largest or smallest singular value.

Perhaps one of the most useful theorems in linear algebra regarding normal matrices, is the Spectral Theorem; this theorem allows us to decompose a normal matrix into a product of matrices that are very easy to work with.

Theorem 1.4 (The Spectral Theorem)**.** *If $\boldsymbol{A} \in M_n$ is normal, then $\boldsymbol{A}$ is unitarily diagonalizable. That is,*

$$\boldsymbol{A} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{U}^* \tag{1.2}$$

where $\boldsymbol{D}$ is a diagonal matrix with the eigenvalues of $\boldsymbol{A}$ along its diagonal. The columns of $\boldsymbol{U}$ form an orthonormal set of eigenvectors of $\boldsymbol{A}$. In fact, each column of $\boldsymbol{U}$ is an eigenvector of $\boldsymbol{A}$, which corresponds to the eigenvalue of $\boldsymbol{A}$ located in the equivalent column of $\boldsymbol{D}$.

Proof. See Horn and Johnson [2, p. 101] □

We will see shortly that this spectral decomposition will allow for computationally efficient matrix multiplication and norm evaluation.

The drawback of the Spectral Theorem is that it only applies to normal matrices. Nonnormal matrices are not unitarily diagonalizable. Schur's Unitary Triangularization Theorem (Theorem 1.5) provides a result which applies to all square matrices in M_N; although it is not quite as structurally convenient as the Spectral Theorem, it is more widely applicable.

Theorem 1.5 (Schur's Unitary Triangularization Theorem)**.** *Given a matrix $\boldsymbol{A} \in M_N$, with eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_N$, there exists a unitary matrix $\boldsymbol{U}$, and an upper-triangular matrix $\boldsymbol{T}$, such that,*

$$\boldsymbol{A} = \boldsymbol{U}\boldsymbol{T}\boldsymbol{U}^*$$

Furthermore, the diagonal entries of $\boldsymbol{T}$ are the eigenvalues of $\boldsymbol{A}$.

Proof. See Horn and Johnson [2, p. 79 - 80] □

This form of representing $\mathbf{A}$ is referred to as the Schur decomposition. The matrix $\mathbf{T}$ is called the Schur form of $\mathbf{A}$.

1.3 Norms

Informally, a ***matrix norm*** is an operator which assigns a nonnegative real number to each matrix it operates on, much like a ruler might be used to assign lengths to various physical objects. As such, a matrix norm is a function which takes a matrix as its argument, and it is endowed with several key properties which we outline more formally below.

The properties of a matrix norm are based on those of a ***vector norm***. A vector norm, $||\cdot|| : V \to (\mathbb{R}^+ \cup \{0\})$, takes a vector argument from a vector space V, and maps it to a nonnegative real value ($\mathbb{R}^+$ denotes the positive reals). All vector norms are characterized by four axioms. Given vectors $\mathbf{u}, \mathbf{v} \in V$, and scalar α,

Axiom 1: $||\mathbf{u}|| \geq 0$

Axiom 2: $||\mathbf{u}|| = 0$ if and only if $\mathbf{u} = \mathbf{0}$

Axiom 3: $||\alpha\mathbf{u}|| = |\alpha|\,||\mathbf{u}||$

Axiom 4: $||\mathbf{u} + \mathbf{v}|| \leq ||\mathbf{u}|| + ||\mathbf{v}||$

Observe that Axiom 4 is an analog of the triangle inequality. We can use it to obtain another useful norm identity: given vectors $\mathbf{u}$ and $\mathbf{v}$ in V, and a vector norm $||\cdot||$,

$$|\,||\mathbf{u}|| - ||\mathbf{v}||\,| \leq ||\mathbf{u} - \mathbf{v}||. \tag{1.3}$$

The proof of inequality (1.3) is straightforward using Axiom 4:

$$||\mathbf{u}|| = ||\mathbf{u} - \mathbf{v} + \mathbf{v}|| \leq ||\mathbf{u} - \mathbf{v}|| + ||\mathbf{v}||$$

Now, subtract $||\mathbf{v}||$ from both sides to get $||\mathbf{u}|| - ||\mathbf{v}|| \leq ||\mathbf{u} - \mathbf{v}||$. Interchanging the roles of $\mathbf{u}$ and $\mathbf{v}$ in the proof above justifies taking the absolute value.

If we replace $\mathbf{u}$ and $\mathbf{v}$ in Axioms 1–4 above by the matrices $\mathbf{A}$ and $\mathbf{B}$, then the axioms become the essential four axioms underpinning all matrix norms. Thus, a matrix norm is similar to a vector norm; however, in the case of a matrix norm, the domain is a space of matrices such as M_N. The same notation is often used for both vector norms and matrix norms. Matrix norms are often characterized by one additional axiom, called the Sub-Multiplicative Axiom: given matrices $\mathbf{A}$ and $\mathbf{B}$,

Axiom 5: $||\mathbf{AB}|| \leq ||\mathbf{A}||\,||\mathbf{B}||$.

All of the matrix norms that we will consider, will satisfy the Sub-Multiplicative Axiom.

The generality of the axioms above allows for endless variety in defining specific vector and matrix norms. We will only consider a few which are among the most commonly used. Let the component-wise representation of $\mathbf{u} \in \mathbb{C}^N$ and $\mathbf{A} \in M_N$ be $\mathbf{u} = (u_1, u_2, \ldots, u_N)^T$, and

$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1N} \\ a_{21} & a_{22} & \dots & a_{2N} \\ \vdots & & \ddots & \vdots \\ a_{N1} & a_{N2} & \dots & a_{NN} \end{pmatrix}$, respectively.

1.3.1 Vector Norms

The 1-norm: $$||\mathbf{u}||_1 := \sum_{n=1}^{N} |u_n|$$

. The 2-norm (Euclidean norm): $$||\mathbf{u}||_2 := \left(\sum_{n=1}^{N} |u_n|^2 \right)^{\frac{1}{2}}$$

The ∞-norm: $$||\mathbf{u}||_\infty := \max_{1 \le n \le N} |u_n|$$

To justify our calling the identities above norms, we must verify that they satisfy the four vector axioms. This is a straightforward exercise for Axioms 1 through 3, which we leave to the reader. Lemma 1.7 below verifies Axiom 4 for the 2-norm. First, we give the famous Cauchy-Schwartz inequality:

Lemma 1.6 (Cauchy-Schwartz Inequality)**.** *Given vectors* $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{C}^N$*,*

$$|\boldsymbol{u}^* \boldsymbol{v}| \le ||\boldsymbol{u}|| \, ||\boldsymbol{v}||$$

Proof. Based on properties of matrix and vector multiplication, it is easy to verify that $\mathbf{v}^*\mathbf{v} = ||\mathbf{v}||_2^2$; thus, this quantity is 0, if and only if $\mathbf{v} = \mathbf{0}$. If $\mathbf{v} = \mathbf{0}$, the inequality holds. Suppose $\mathbf{v} \neq \mathbf{0}$. Let $\mathbf{z} = \mathbf{u} - \frac{\mathbf{u}^*\mathbf{v}}{||\mathbf{v}||_2^2}\mathbf{v}$, and note that $\mathbf{z}^*\mathbf{v} = 0$. Taking the conjugate gives $\overline{\mathbf{z}^*\mathbf{v}} = 0$, which implies that $\mathbf{v}^*\mathbf{z} = 0$. Also, note that,

$$\mathbf{u} = \frac{\mathbf{u}^*\mathbf{v}}{||\mathbf{v}||_2^2}\mathbf{v} + \left(\mathbf{u} - \frac{\mathbf{u}^*\mathbf{v}}{||\mathbf{v}||_2^2}\mathbf{v} \right) = \frac{\mathbf{u}^*\mathbf{v}}{||\mathbf{v}||_2^2}\mathbf{v} + \mathbf{z}$$

This implies that,

$$\begin{aligned} ||\mathbf{u}||_2^2 &= \left|\left| \frac{\mathbf{u}^*\mathbf{v}}{||\mathbf{v}||_2^2}\mathbf{v} + \mathbf{z} \right|\right|_2^2 = \left(\frac{\mathbf{u}^*\mathbf{v}}{||\mathbf{v}||_2^2}\mathbf{v} + \mathbf{z} \right)^* \left(\frac{\mathbf{u}^*\mathbf{v}}{||\mathbf{v}||_2^2}\mathbf{v} + \mathbf{z} \right) \\ &= \left(\frac{\mathbf{u}^*\mathbf{v}}{||\mathbf{v}||_2^2} \right)^2 \mathbf{v}^*\mathbf{v} + \overline{\left(\frac{\mathbf{u}^*\mathbf{v}}{||\mathbf{v}||_2^2} \right)} \mathbf{v}^*\mathbf{z} + \left(\frac{\mathbf{u}^*\mathbf{v}}{||\mathbf{v}||_2^2} \right) \mathbf{z}^*\mathbf{v} + \mathbf{z}^*\mathbf{z} = \left(\frac{\mathbf{u}^*\mathbf{v}}{||\mathbf{v}||_2^2} \right)^2 \mathbf{v}^*\mathbf{v} + \mathbf{z}^*\mathbf{z} \\ &\ge \left(\frac{\mathbf{u}^*\mathbf{v}}{||\mathbf{v}||_2^2} \right)^2 \mathbf{v}^*\mathbf{v} = \frac{(\mathbf{u}^*\mathbf{v})^2}{||\mathbf{v}||_2^2} \end{aligned}$$

Rearranging terms we get,

$$||\mathbf{u}||_2^2||\mathbf{v}||_2^2 \geq (\mathbf{u}^*\mathbf{v})^2$$

Taking the square root of both sides finishes the proof. □

Lemma 1.7. *Given vectors* $\boldsymbol{u} = (u_1, u_2, \ldots, u_N)^T$ *and* $\boldsymbol{v} = (v_1, v_2, \ldots, v_N)^T$ *in* $\mathbb{C}^N$,

$$||\boldsymbol{u} + \boldsymbol{v}||_2 \leq ||\boldsymbol{u}||_2 + ||\boldsymbol{v}||_2.$$

Proof.

$$\begin{aligned} ||\mathbf{u}^2 + \mathbf{v}^2||_2^2 &= (\mathbf{u} + \mathbf{v})^*(\mathbf{u} + \mathbf{v}) = \mathbf{u}^*\mathbf{u} + \mathbf{u}^*\mathbf{v} + \mathbf{v}^*\mathbf{u} + \mathbf{v}^*\mathbf{v} \\ &\leq ||\mathbf{u}||_2^2 + 2\,||\mathbf{u}||_2\,||\mathbf{v}||_2 + ||\mathbf{v}||_2^2 = (||\mathbf{u}||_2 + ||\mathbf{v}||_2)^2 \end{aligned}$$

The inequality above follows by the Cauchy Schwartz inequality. Taking the square root finishes the proof. □

Norms are important because they provide metrics for comparing one vector to another. This allows us to extend many single-variable concepts, like convergence and continuity, to multiple dimension vector spaces. A vector $\mathbf{u}$ with $||\mathbf{u}|| = 1$ is referred to as a unit vector. Given a sequence of vectors $\{\mathbf{x}^{(k)}\}_{k=1}^{\infty}$ in a vector space V, and a single vector $\mathbf{x}$ in V, we say that the sequence $\{\mathbf{x}^{(k)}\}_{k=1}^{\infty}$ ***converges*** to $\mathbf{x}$ with respect to the norm, $||\cdot||$ if and only if the norm, $||\mathbf{x}^{(k)} - \mathbf{x}||$, approaches zero as k approaches infinity; that is, given some $\varepsilon > 0$, there exists $K \in \mathbb{N}$, such that $||\mathbf{x}^{(k)} - \mathbf{x}|| < \varepsilon$, whenever $k \geq K$. This is written more compactly as, $||\mathbf{x}^{(k)} - \mathbf{x}|| \to 0$ as $k \to \infty$. The vector $\mathbf{x}$ is then referred to as the ***limit*** of the sequence $\{\mathbf{x}^{(k)}\}_{k=1}^{\infty}$ with respect to the norm $||\cdot||$. A subset $W \subset V$ is ***closed*** (with respect to some particular norm) if and only if every sequence in W that converges, converges to a vector which is also contained in W. The subset W is ***bounded*** (also with respect to some particular norm) if and only if there exists some $M \in \mathbb{R}$, such that $M \geq ||\mathbf{x}||$, for every $\mathbf{x} \in W$. A subset of vectors $W \subset V$ is ***compact*** in V if and only if it is closed and bounded (actually, there is a more abstract definition for the more general idea of compactness in a topological space, but this is sufficient for vector spaces). A complex-valued function f defined on a vector space V is ***continuous at*** $\mathbf{x} \in V$ with respect to

the norm $||\cdot||$ if and only if for every sequence $\{\mathbf{x}^{(k)}\}_{k=0}^{\infty} \subset V$ such that $||\mathbf{x}^{(k)} - \mathbf{x}|| \to 0$, it follows that $|f(\mathbf{x}^{(k)}) - f(\mathbf{x})| \to 0$, as $n \to \infty$. If f is continuous at every vector of its domain, then we say f is ***continuous***. By replacing $\mathbf{u}$ and $\mathbf{v}$ in inequality (1.3) by a convergent sequence and its limit (respectively), we obtain a proof that vector norms themselves are continuous functions.

Given a matrix $\mathbf{A} \in M_N$ and any vector $\mathbf{x} \in \mathbb{C}^N$, since $\mathbf{Ax}$ is also a vector we may be interested in the function $f(\mathbf{x}) = ||\mathbf{Ax}||$. In fact, for all the vector norms we have discussed, this function is continuous! Lemma 1.9 will prove this for the 2-norm, and similar strategies can be used to prove continuity in the 1 and ∞-norm cases. First, we look at a useful lemma relating convergence of vector components to convergence of vectors.

Lemma 1.8. *The sequence $\{\boldsymbol{x}^{(k)}\}_{k=0}^{\infty} \subset \mathbb{C}^N$, converges to the vector $\boldsymbol{x} \in \mathbb{C}^N$ with respect to the 2-norm if and only if $x_j^{(k)}$ converges (in the single-variable sense) to x_j, for each $j = 1, 2, \ldots, N$, where $x_j^{(k)}$ and x_j are the j^{th} components of $\boldsymbol{x}^{(k)}$ and $\boldsymbol{x}$, respectively.*

Proof. First, we prove the forward direction. Fix $j \in \{1, 2, \ldots, N\}$, and suppose that $\{\mathbf{x}^{(k)}\}_{k=0}^{\infty} \subset \mathbb{C}^N$ converges to the vector $\mathbf{x} \in \mathbb{C}^N$. Then given some $\varepsilon > 0$, there exists $K \in \mathbb{N}$, such that for every $k \geq K$,

$$|x_j^{(k)} - x_j|^2 \leq \sum_{n=1}^{N} |x_n^{(k)} - x_n|^2 = ||\mathbf{x}^{(k)} - \mathbf{x}||_2^2 < \varepsilon^2$$

Taking the square root of boths sides shows that the sequence $\{x_j^{(k)}\}_{k=0}^{\infty}$ converges (in the single variable sense) to x_j. Since j is arbitrary, each component of the sequence $\{\mathbf{x}^{(k)}\}_{k=0}^{\infty}$ converges to each respective component of the vector $\mathbf{x}$.

Now, we prove the other direction. Suppose each component sequence $\{x_n^{(k)}\}_{k=0}^{\infty}$ converges to the respective component x_n for $n = 1, 2, \ldots, N$. By definition of convergence, given $\varepsilon > 0$, there exists $K_1, K_2, \ldots, K_N \in \mathbb{N}$, such that for each $n = 1, 2, \ldots, N$,

$$|x_n^{(k)} - x_n| < \frac{\varepsilon}{\sqrt{N}}, \quad \text{whenever } k \geq K_n$$

Let $K = \max\{K_1, K_2, \ldots, K_N\}$. Then, for every $k \geq K$,

$$||\mathbf{x}^{(k)} - \mathbf{x}||_2^2 = \sum_{n=1}^{N} |x_n^{(k)} - x_n|^2 < N\frac{\varepsilon^2}{N} = \varepsilon^2$$

Taking the square root finishes the proof. □

Now we are ready to discuss the continuity of $||\mathbf{Ax}||_2$.

Lemma 1.9. *Given a matrix $\boldsymbol{A} \in M_N$, the function $f : \mathbb{C}^N \to \mathbb{R} \cup \{0\}$, defined by $f(x) = ||\boldsymbol{Ax}||_2$ is continuous.*

Proof. Suppose that $\{\mathbf{x}^{(k)}\}_{k=0}^{\infty}$ is a sequence in $\mathbb{C}^N$ that converges to the vector $\mathbf{x}$. Choose $\varepsilon > 0$. Denote the mn^{th} entry of the matrix $\mathbf{A}$ by a_{mn}, and let $\alpha = \max_{1 \le m,n \le N} \{|a_{mn}|\}$. If $\alpha = 0$, then $\mathbf{A}$ is the zero matrix. In this case, $|f(\mathbf{x}^{(k)}) - f(\mathbf{x})| = |\,||\mathbf{Ax}^{(k)}||_2 - ||\mathbf{Ax}||_2| = 0 < \varepsilon$, for every $k \in \mathbb{N}$.

Suppose that $\alpha > 0$. As in the proof of Lemma 1.8, there exists $K_1, K_2, \ldots, K_N \in \mathbb{N}$, such that for each $n = 1, 2, \ldots, N$,

$$|x_n^{(k)} - x_n| < \frac{\varepsilon}{\alpha N}, \quad \text{whenever } k \ge K_n$$

Let $K = \max\{K_1, K_2, \ldots, K_N\}$. Then for $k \ge K$,

$$\begin{aligned}
|f(\mathbf{x}^{(k)}) - f(\mathbf{x})|^2 &= \left| ||\mathbf{Ax}^{(k)}||_2 - ||\mathbf{Ax}||_2 \right|^2 \le \left|\left| \mathbf{Ax}^{(k)} - \mathbf{Ax} \right|\right|_2^2 \\
&= \left|\left| \mathbf{A}\left(\mathbf{x}^{(k)} - \mathbf{x}\right) \right|\right|_2^2 = \sum_{m=1}^{N} \left| \sum_{n=1}^{N} a_{mn}\left(x_n^{(k)} - x_n\right) \right|^2 \\
&\le \sum_{m=1}^{N}\sum_{n=1}^{N} |a_{mn}|^2 \left| x_n^{(k)} - x_n \right|^2 < \sum_{m=1}^{N}\sum_{n=1}^{N} \alpha^2 \frac{\varepsilon^2}{\alpha^2 N^2} \\
&= \varepsilon^2
\end{aligned}$$

Taking the square root finishes the proof: $f(\mathbf{x})$ is continuous. □

1.3.2 Matrix Norms. Given any vector norm, $||\cdot||_v : \mathbb{C}^N \to (\mathbb{R}^+ \cup \{0\})$, one way to define a matrix norm $||\cdot|| : M_N \to \mathbb{R} \cup \{0\}$ is the following:

$$||\mathbf{A}|| := \sup_{||\mathbf{u}||=1} ||\mathbf{Au}||_v \tag{1.4}$$

where $\mathbf{A} \in M_N$. Matrix norms generated in this manner are referred to as ***operator norms***. Note that the set $B = \{\mathbf{x} : ||\mathbf{x}|| = 1\}$ is closed and bounded, and therefore, compact; thus, when

$||\cdot||_v$ is continuous, it attains a maximum value on B, and we are justified in replacing 'sup' in the definition above by 'max'. Since the properties of operator norms depend on the properties of the supremum, and on the properties of corresponding vector norms, operator norms pass Axioms 1–4 almost effortlessly.

In addition, all operator norms posses the sub-multiplicative property (Axiom 5), as the following lemma shows.

Lemma 1.10. *Given the operator norm* $||\cdot|| : \mathbb{C}^N \to \mathbb{R} \cup \{0\}$*, and matrices* $\boldsymbol{A}$ *and* $\boldsymbol{B}$ *in* M_N*,*

$$||\boldsymbol{AB}|| \leq ||\boldsymbol{A}||\,||\boldsymbol{B}||$$

Proof. Choose a unit vector $\mathbf{x} \in \mathbb{C}^N$. If $\mathbf{Bx} = \mathbf{0}$, then trivially,

$$0 = ||\mathbf{A0}|| = ||\mathbf{ABx}|| \leq ||\mathbf{A}||\,||\mathbf{B}||$$

If $\mathbf{Bx} \neq \mathbf{0}$, then observe that $\frac{1}{||\mathbf{Bx}||}\mathbf{Bx}$ is a unit vector. Then,

$$||\mathbf{ABx}|| = \frac{||\mathbf{Bx}||}{||\mathbf{Bx}||}\,||\mathbf{ABx}|| = ||\mathbf{Bx}||\,||\mathbf{A}\left(\frac{1}{||\mathbf{Bx}||}\mathbf{Bx}\right)||$$

Since both $\mathbf{x}$ and $\frac{1}{||\mathbf{Bx}||}\mathbf{Bx}$ are unit vectors, the definition of an operator norm as a supremum gives that $||\mathbf{A}||\,||\mathbf{B}||$ is greater than the right-hand side above. Since $\mathbf{x}$ is arbitrary, and $||\mathbf{AB}|| = \max_{||\mathbf{x}||=1} ||\mathbf{AB}||$, we have $||\mathbf{AB}|| \leq ||\mathbf{A}||\,||\mathbf{B}||$. □

The previous vector norms each produce a corresponding operator norm:

The 1-norm:	$||\mathbf{A}||_1 := \max_{||\mathbf{u}||=1} ||\mathbf{Au}||_1$
The 2-norm (spectral norm):	$||\mathbf{A}||_2 := \max_{||\mathbf{u}||=1} ||\mathbf{Au}||_2$
The ∞-norm:	$||\mathbf{A}||_\infty := \max_{||\mathbf{u}||=1} ||\mathbf{Au}||_\infty$

Note that we have overloaded the norm notation above, as is common practice in the field. In the equations above, $\mathbf{Au}$ is a vector, so on the right-hand side, $||\cdot||_1$ is a vector norm; whereas $\mathbf{A}$ is a matrix, so on the left-hand side, $||\cdot||_1$ is a matrix norm. Thus, the meaning of $||\cdot||_1$ is different depending on whether we put a vector or a matrix inside. In fact, subscripts are often dropped altogether when the norm in use is well understood, and also when discussing properties of norms

in general.

The matrix 2-norm is often referred to as the spectral norm because of its relationship to the eigenvalues of $\mathbf{A}$ and $\mathbf{A}^*\mathbf{A}$. The next few lemmas explain these concepts, as well as some other nice properties of the 2-norm.

Lemma 1.11. *Given a unitary matrix $\boldsymbol{U} = (u_{mn}) \in M_N$, and a vector $\boldsymbol{x} \in \mathbb{C}^N$,*

$$||\boldsymbol{Ux}||_2 = ||\boldsymbol{x}||_2$$

Proof.

$$||\mathbf{Ux}||_2^2 = (\mathbf{Ux})^* \mathbf{Ux} = \mathbf{x}^*\mathbf{U}^*\mathbf{Ux} = \mathbf{x}^*\mathbf{x} = ||\mathbf{x}||_2^2$$

Taking square roots finishes the proof. □

Observe that an immediate consequence of Lemma 1.11 is that the 2-norm of any unitary matrix is one.

Lemma 1.12. *If $\boldsymbol{A} \in M_N$ is normal, then $\rho(\boldsymbol{A}) = ||\boldsymbol{A}||_2$.*

Proof. By the Spectral Theorem (Theorem 1.4), we can write $\mathbf{A} = \mathbf{UDU}^*$, where $\mathbf{U}$ is unitary and

$\mathbf{D} = (d_{mn})$ is a diagonal matrix with the eigenvalues of $\mathbf{A}$ along its diagonal. Note that at least one of the eigenvalues of $\mathbf{A}$ will have magnitude equal to $\rho(\mathbf{A})$. Without loss of generality, suppose that this is true of the entry d_{11}. Let $\mathbf{e} = (e_1, e_2, \ldots, e_N)^T$ be the vector with 1 in the first entry and zeros thereafter. Then, by the definition of the spectral norm,

$$\begin{aligned} ||\mathbf{D}||_2 \geq ||\mathbf{De}||_2 &= \left|\left|\left(\sum_{n=1}^{N} d_{1n}e_n, \sum_{n=1}^{N} d_{2n}e_n, \ldots, \sum_{n=1}^{N} d_{Nn}e_n \right)^T\right|\right|_2 \\ &= \left|\left|(d_{11}, 0, \ldots, 0)^T\right|\right|_2 = \sqrt{|d_{11}|^2} = \rho(\mathbf{A}) \end{aligned} \tag{1.5}$$

In contrast to the above result, note that for every $\mathbf{x} = (x_1, x_2, \ldots, x_N)^T \in \mathbb{C}^N$ with $||\mathbf{x}||_2 = 1$,

$$\begin{aligned}
||\mathbf{Dx}||_2^2 &= \left| \sum_{n=1}^{N} d_{1n}x_n \right|^2 + \left| \sum_{n=1}^{N} d_{2n}x_n \right|^2 + \ldots + \left| \sum_{n=1}^{N} d_{Nn}x_n \right|^2 \\
&= |d_{11}x_1|^2 + |d_{22}x_2|^2 + \ldots + |d_{NN}x_N|^2 \\
&= |d_{11}|^2|x_1|^2 + |d_{22}|^2|x_2|^2 + \ldots + |d_{NN}|^2|x_N|^2 \\
&\leq \rho(\mathbf{A})^2|x_1|^2 + \rho(\mathbf{A})^2|x_2|^2 + \ldots + \rho(\mathbf{A})^2|x_N|^2 \\
&= \rho(\mathbf{A})^2 \sum_{n=1}^{N} |x_n|^2 = \rho(\mathbf{A})^2||\mathbf{x}||_2^2 = \rho(\mathbf{A})^2
\end{aligned}$$

Since $\mathbf{x}$ is arbitrary, we have $||\mathbf{D}||_2^2 \leq \rho(\mathbf{A})^2$, which implies $||\mathbf{D}||_2 \leq \rho(\mathbf{A})$. Combining this with inequality (1.5) proves that $||\mathbf{D}||_2 = \rho(\mathbf{A})$.

Now note that,

$$||\mathbf{A}||_2 = ||\mathbf{UDU}^*||_2 \leq ||\mathbf{U}||_2 \, ||\mathbf{D}||_2 \, ||\mathbf{U}^*||_2 = ||\mathbf{D}||_2 = \rho(\mathbf{A})$$

In contrast, suppose that the unit eigenvector corresponding to d_{11} is $\mathbf{u}$. Then,

$$\rho(\mathbf{A}) = |d_{11}| = |d_{11}| \, ||\mathbf{u}||_2 = ||d_{11}\mathbf{u}||_2 = ||\mathbf{Au}||_2 \leq ||\mathbf{A}||_2$$

By the last two inequalities, $\rho(\mathbf{A}) = ||\mathbf{A}||_2$. □

Although the above lemma does not hold in general for all matrices in M_N, a similar result involving the largest singular value does!

Lemma 1.13. *Given $\boldsymbol{A} \in M_N$ with largest singular value σ_{max},*

$$\sigma_{max} = ||\boldsymbol{A}||_2$$

Proof. Since the singular values of $\mathbf{A}$ are nonnegative, σ_{max}^2 is the eigenvalue of $\mathbf{A}^*\mathbf{A}$ of largest magnitude. Since $\mathbf{A}^*\mathbf{A}$ is normal, it can be decomposed as $\mathbf{A}^*\mathbf{A} = \mathbf{UDU}^* = \mathbf{V}^*\mathbf{DV}$, where $\mathbf{U} = \mathbf{V}^*$ is a unitary matrix, and $\mathbf{D}$ is a diagonal matrix with the squares of the singular values of $\mathbf{A}$ along its diagonal. Let $\mathbf{D}^{1/2}$ be the real matrix obtained from $\mathbf{D}$ by taking the positive square

root of each of its entries; because $\mathbf{D}$ is diagonal, $\mathbf{D}^{1/2}\mathbf{D}^{1/2} = \mathbf{D}$ and $\mathbf{D}^{1/2} = \left(\mathbf{D}^{1/2}\right)^*$. Note that given any unit vector $\mathbf{x}$, the vector $\mathbf{Vx}$ is also a unit vector by Lemma 1.11. Then,

$$\begin{aligned}||\mathbf{Ax}||_2^2 &= (\mathbf{Ax})^*\mathbf{Ax} = \mathbf{x}^*\mathbf{A}^*\mathbf{Ax} = \mathbf{x}^*\mathbf{V}^*\mathbf{DVx} = \mathbf{x}^*\mathbf{V}^*\mathbf{D}^{1/2}\mathbf{D}^{1/2}\mathbf{Vx} \\ &= (\mathbf{D}^{1/2}\mathbf{Vx})^*\mathbf{D}^{1/2}(\mathbf{Vx}) = ||\mathbf{D}^{1/2}(\mathbf{Vx})||_2^2 \leq ||\mathbf{D}^{1/2}||_2^2 = \sigma_{max}^2\end{aligned}$$

Since this is true for all unit vectors $\mathbf{x}$, we have $||\mathbf{A}||_2^2 \leq \sigma_{max}^2$, which implies that $||\mathbf{A}||_2 \leq \sigma_{max}$. Now note that since σ_{max}^2 is an eigenvalue of $\mathbf{A}^*\mathbf{A}$, there exists a corresponding unit eigenvector $\mathbf{y}$, such that $\mathbf{A}^*\mathbf{Ay} = \sigma_{max}^2\mathbf{y}$. Then,

$$||\mathbf{A}||_2^2 \geq ||\mathbf{Ay}||_2^2 = \mathbf{y}^*\mathbf{A}^*\mathbf{Ay} = \sigma_{max}^2\mathbf{y}^*\mathbf{y} = \sigma_{max}^2||\mathbf{y}||_2^2 = \sigma_{max}^2$$

Thus, $||\mathbf{A}||_2 \geq \sigma_{max}$. Combining this with the result from above yields: $||\mathbf{A}||_2 = \sigma_{max}$. □

Corollary 1.14. *Given a matrix $\boldsymbol{A} \in M_N$,*

$$||\boldsymbol{A}^*||_2 = ||\boldsymbol{A}||_2$$

Proof. Let σ_{max} be the largest singular value of $\mathbf{A}$. Then σ_{max}^2 is the largest eigenvalue of $\mathbf{A}^*\mathbf{A}$ and $\mathbf{AA}^*$, since their nonzero eigenvalues are the same. Then σ_{max} is the largest singular value of $\mathbf{A}^*$. By Lemma 1.13, $||\mathbf{A}^*||_2 = \sigma_{max} = ||\mathbf{A}||_2$. □

Another matrix norm that we will consider is the Frobenius norm, which is not an operator norm:

The Frobenius norm: $$||\mathbf{A}||_F := \left(\sum_{m,n=1}^{N} |a_{mn}|^2\right)^{\frac{1}{2}}$$

The Frobenius norm is often also written as:

$$||\mathbf{A}||_F = \sqrt{\text{trace}\,(\mathbf{AA}^*)} = \sqrt{\text{trace}\,(\mathbf{A}^*\mathbf{A})} \tag{1.6}$$

To see the equivalence of equation (1.6) to the Frobenius norm definition, note that the m^{th} diagonal

$(\mathbf{AA}^*)_{mm}$ of $\mathbf{AA}^*$ is just

$$(\mathbf{AA}^*)_{mm} = \sum_{j=1}^{N} (\mathbf{A})_{mj} (\mathbf{A}^*)_{jm} = \sum_{j=1}^{N} a_{mj}\overline{a}_{mj} = \sum_{j=1}^{N} |a_{mj}|^2$$

Summing over m from $m = 1$ to $m = N$ gives the trace of $\mathbf{AA}^*$, and it also gives the square of $||\mathbf{A}||_F$. Furthermore, note that the m^{th} diagonal entry of $\mathbf{A}^*\mathbf{A}$ is given by,

$$(\mathbf{A}^*\mathbf{A})_{mm} = \sum_{j=1}^{N} (\mathbf{A}^*)_{mj}(\mathbf{A})_{jm} = \sum_{j=1}^{N} \overline{a}_{jm}a_{jm} = \sum_{j=1}^{N} |a_{jm}|^2$$

Again, summing over all m gives trace$(\mathbf{A}^*\mathbf{A})$, as well as $||\mathbf{A}||_F^2$.

A matrix norm is called ***unitarily invariant*** if pre-multiplying on the left or post-multiplying on the right by a unitary matrix does not change the value of the norm; that is, the norm $||\cdot||$ is unitarily invariant if and only if given $\mathbf{A} \in M_N$ and a unitary matrix $\mathbf{U} \in M_N$, we have

$$||\mathbf{UA}|| = ||\mathbf{A}|| = ||\mathbf{AU}||$$

Unitary invariance will be a valuable property in some of the theorems that follow. Of the norms that we have considered, there are just two that are unitarily invariant: the Frobenius norm, and the spectral norm.

Lemma 1.15. *The Frobenius norm is unitarily invariant.*

Proof. Choose $\mathbf{A}$ and $\mathbf{U}$ from M_N, with $\mathbf{U}$ unitary. Then,

$$||\mathbf{UA}||_F = \sqrt{\text{trace}((\mathbf{UA})^*(\mathbf{UA}))} = \sqrt{\text{trace}(\mathbf{A}^*\mathbf{U}^*\mathbf{UA})} = \sqrt{\text{trace}(\mathbf{A}^*\mathbf{A})} = ||\mathbf{A}||_F$$

Similarly,

$$||\mathbf{AU}||_F = \sqrt{\text{trace}((\mathbf{AU})(\mathbf{AU})^*)} = \sqrt{\text{trace}(\mathbf{AUU}^*\mathbf{A}^*)} = \sqrt{\text{trace}(\mathbf{AA}^*)} = ||\mathbf{A}||_F$$

□

Lemma 1.16. *The 2-norm is unitarily invariant.*

Proof. Choose $\mathbf{A}$ and $\mathbf{U}$ from M_N, with $\mathbf{U}$ unitary. Let σ_{max} be the largest singular value of $\mathbf{A}$. By Lemma 1.13, $||\mathbf{A}||_2 = \sigma_{max}$. Then,

$$||\mathbf{UA}||_2 \leq ||\mathbf{U}||_2 \, ||\mathbf{A}||_2 = ||\mathbf{A}||_2 = \sigma_{max}$$

By a similar argument, $||\mathbf{AU}||_2 \leq \sigma_{max}$. In contrast, note that there exists unit vectors $\mathbf{x}$ and $\mathbf{y}$, such that $\mathbf{A}^*\mathbf{Ax} = \sigma_{max}^2\mathbf{x}$ and $\mathbf{AA}^*\mathbf{y} = \sigma_{max}^2\mathbf{y}$. Then,

$$||\mathbf{UA}||_2^2 \geq ||\mathbf{UAx}||_2^2 = \mathbf{x}^*\mathbf{A}^*\mathbf{U}^*\mathbf{UAx} = \mathbf{x}^*\mathbf{A}^*\mathbf{Ax} = \sigma_{max}^2\mathbf{x}^*\mathbf{x} = \sigma_{max}^2||\mathbf{x}||_2 = \sigma_{max}^2$$

and

$$\begin{aligned} ||\mathbf{AU}||_2^2 = ||(\mathbf{AU})^*||_2^2 \geq ||(\mathbf{AU})^*\mathbf{y}||_2^2 = \mathbf{y}^*\mathbf{AU}(\mathbf{AU})^*\mathbf{y} = \mathbf{y}^*\mathbf{AUU}^*\mathbf{A}^*\mathbf{y} = \mathbf{y}^*\mathbf{AA}^*\mathbf{y} \\ = \sigma_{max}^2\mathbf{y}^*\mathbf{y} = \sigma_{max}^2||\mathbf{y}||_2 = \sigma_{max}^2 \end{aligned}$$

Thus, $||\mathbf{UA}||_2 = \sigma_{max} = ||\mathbf{AU}||_2$, and the theorem result follows from $||\mathbf{A}||_2 = \sigma_{max}$. □

1.4 The Resolvent

The ***resolvent*** of a matrix $\mathbf{A} \in M_N$ is the function $R : \mathbb{C}\backslash\sigma(\mathbf{A}) \to M_N$ defined by

$$R(z) := (z\mathbf{I} - \mathbf{A})^{-1}$$

Note that $\sigma(\mathbf{A})$ is a closed set, which makes the complement $\mathbb{C}\backslash\sigma(\mathbf{A})$ open. From linear algebra and identities for matrix inverses, it follows that

$$R(z) = (z\mathbf{I} - \mathbf{A})^{-1} = \frac{1}{\det(z\mathbf{I} - \mathbf{A})}\text{adj}(z\mathbf{I} - \mathbf{A}) \tag{1.7}$$

where 'adj' refers to the adjugate of $(z\mathbf{I} - \mathbf{A})$. This form of the resolvent is insightful, because we will be able to use it to show that upon left and right multiplication by suitable vectors, the resolvent can be made into a complex rational function. Indeed, we have already shown above that,

$\det(z\mathbf{I} - \mathbf{A})$ is a polynomial in z of degree N. We now show that the adjugate of $(z\mathbf{I} - \mathbf{A})$ is a matrix polynomial in z.

Choose $z \in \mathbb{C}\backslash\sigma(\mathbf{A})$, and let $\mathrm{p}(\xi) = \xi^N + c_{N-1}\xi^{N-1} + \ldots + c_1\xi + c_0$ be the characteristic polynomial of $(z\mathbf{I} - \mathbf{A})$. Note that $\mathrm{p}(\xi) = \det(\xi\mathbf{I} - (z\mathbf{I} - A))$. Then $c_0 = (-1)^N\det(z\mathbf{I} - \mathbf{A})$, by equation (1.1). By the Cayley-Hamilton Theorem,

$$\begin{aligned} 0 &= \mathrm{p}(z\mathbf{I} - \mathbf{A}) \\ &= (z\mathbf{I} - \mathbf{A})^N + c_{N-1}(z\mathbf{I} - \mathbf{A})^{N-1} + \ldots + c_1(z\mathbf{I} - \mathbf{A}) + (-1)^N\det(z\mathbf{I} - \mathbf{A})\mathbf{I} \end{aligned}$$

Rearranging terms we get,

$$\begin{aligned} \mathbf{I} &= \frac{(-1)^N}{\det(z\mathbf{I} - \mathbf{A})}\left(-(z\mathbf{I} - \mathbf{A})^N - c_{N-1}(z\mathbf{I} - \mathbf{A})^{N-1} - \ldots - c_1(z\mathbf{I} - \mathbf{A})\right) \\ &= (z\mathbf{I} - \mathbf{A})\left(\frac{(-1)^N}{\det(z\mathbf{I} - \mathbf{A})}\left(-(z\mathbf{I} - \mathbf{A})^{N-1} - c_{N-1}(z\mathbf{I} - \mathbf{A})^{N-2} - \ldots - c_1\right)\right) \end{aligned}$$

Thus,

$$(z\mathbf{I} - \mathbf{A})^{-1} = \frac{(-1)^N}{\det(z\mathbf{I} - \mathbf{A})}\left(-(z\mathbf{I} - \mathbf{A})^{N-1} - c_{N-1}(z\mathbf{I} - \mathbf{A})^{N-2} - \ldots - c_1\right)$$

and by equation 1.7 we have,

$$\mathrm{adj}(z\mathbf{I} - \mathbf{A}) = (-1)^N\left(-(z\mathbf{I} - \mathbf{A})^{N-1} - c_{N-1}(z\mathbf{I} - \mathbf{A})^{N-2} - \ldots - c_1\right)$$

Therefore, since z is arbitrary, $\mathrm{adj}(z\mathbf{I} - \mathbf{A})$ is a matrix polynomial of degree $N - 1$ for each $z \in \mathbb{C}\backslash\sigma(\mathbf{A})$.

By pre and post multiplying $\mathrm{adj}(z\mathbf{I} - \mathbf{A})$ by $\mathbf{u}^*$ and $\mathbf{v}$ respectively, for some $\mathbf{u}, \mathbf{v} \in \mathbb{C}^N$, we create a complex polynomial in z of degree less than or equal to $N - 1$. In turn, this makes $\mathbf{u}^*R(z)\mathbf{v}$ a complex rational function, with numerator and denominator having degrees less than or equal to N. Since rational functions are continuous and analytic on their domains, the same holds true for $\mathbf{u}^*R(z)\mathbf{v}$, and for the resolvent $R(z)$ on the domain $\mathbb{C}\backslash\sigma(\mathbf{A})$. This fact will be integral to the proof of Spijker's Lemma 3.21 later on.

The next theorem shows that the norm of the resolvent of $\mathbf{A}$ evaluated at z, is bounded below

by one over the distance between z and the spectrum of $\mathbf{A}$.

Theorem 1.17. *Given $\boldsymbol{A} \in M_N$, $z \in \mathbb{C}\backslash\sigma(\boldsymbol{A})$, and matrix norm $||\cdot||$, we have:*

$$||(z\boldsymbol{I} - \boldsymbol{A})^{-1}|| \geq \frac{1}{\text{dist}(z, \sigma(\boldsymbol{A}))}$$

Proof. The fact that $\sigma(\mathbf{A})$ is closed, implies that $\text{dist}(z, \mathbf{A}) > 0$ for every $z \in \mathbb{C}\backslash\sigma(\mathbf{A})$. Also, note that $z\mathbf{I} - \mathbf{A}$ is invertible for each $z \in \mathbb{C}\backslash\sigma(\mathbf{A})$.

Choose $z \in \mathbb{C}\backslash\sigma(\mathbf{A})$. Let λ be an eigenvalue of $\mathbf{A}$, and let $\mathbf{v}$ be a corresponding orthonormal eigenvector. Then,

$$(z\mathbf{I} - \mathbf{A})\mathbf{v} = z\mathbf{v} - \mathbf{A}\mathbf{v} = z\mathbf{v} - \lambda\mathbf{v} = (z - \lambda)\mathbf{v}$$

Multiplying both sides of the equation by $\left(\frac{1}{z-\lambda}\right)(z\mathbf{I} - \mathbf{A})^{-1}$ we get:

$$\left(\frac{1}{z - \lambda}\right)\mathbf{v} = (z\mathbf{I} - \mathbf{A})^{-1}\mathbf{v}$$

Taking norms,

$$\left|\frac{1}{z - \lambda}\right| ||\mathbf{v}|| = \left|\left|\left(\frac{1}{z - \lambda}\right)\mathbf{v}\right|\right| = \left|\left|(z\mathbf{I} - \mathbf{A})^{-1}\mathbf{v}\right|\right| \leq \left|\left|(z\mathbf{I} - \mathbf{A})^{-1}\right|\right| ||\mathbf{v}||$$

Thus,

$$\left|\frac{1}{z - \lambda}\right| \leq \left|\left|(z\mathbf{I} - \mathbf{A})^{-1}\right|\right|$$

Since λ is arbitrary, the inequality holds for all $\lambda \in \sigma(\mathbf{A})$, and we may replace $\left|\frac{1}{z-\lambda}\right|$ with $\frac{1}{\text{dist}(z,\sigma(\mathbf{A}))}$ in the inequality above. □

Corollary 1.18. *If $\boldsymbol{A} \in M_N$ is normal and $z \in \mathbb{C}$, then equality holds:*

$$||(z\boldsymbol{I} - \boldsymbol{A})^{-1}||_2 = \frac{1}{\text{dist}(z, \sigma(\boldsymbol{A}))} \tag{1.8}$$

Proof. Choose $z \in \mathbb{C}$. Let the eigenvalues of $\mathbf{A}$ be $\lambda_1, \lambda_2, \cdots, \lambda_n$. It is easily verified that the eigenvalues of $z\mathbf{I} - \mathbf{A}$ are $z - \lambda_1, z - \lambda_2, \cdots, z - \lambda_n$. There is no loss of generality in assuming that the eigenvalues are labeled so that $|z - \lambda_1| \leq |z - \lambda_2| \leq \ldots \leq |z - \lambda_n|$. In this case we have that $\text{dist}(z, \sigma(\mathbf{A})) = |z - \lambda_1|$. By Theorem 1.4 there is a unitary matrix $\mathbf{U}$ such that $\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{U}^*$,

where $\mathbf{D}$ is a diagonal matrix with the eigenvalues of $\mathbf{A}$ along its diagonal. The following shows that $z\mathbf{I} - \mathbf{A}$ is normal:

$$(z\mathbf{I}-\mathbf{A})^*(z\mathbf{I}-\mathbf{A}) = |z|^2\mathbf{I}-\overline{z}\mathbf{I}^*\mathbf{A}-z\mathbf{A}^*\mathbf{I}+\mathbf{A}^*\mathbf{A} = |z|^2\mathbf{I}-\overline{z}\mathbf{A}\mathbf{I}^*-z\mathbf{I}\mathbf{A}^*+\mathbf{A}\mathbf{A}^* = (z\mathbf{I}-\mathbf{A})(z\mathbf{I}-\mathbf{A})^*$$

By Lemma 1.3, $(z\mathbf{I} - \mathbf{A})^{-1}$ is also normal. Then by Lemma 1.12,

$$\begin{aligned} ||(z\mathbf{I} - \mathbf{A})^{-1}||_2 &= \rho((z\mathbf{I} - \mathbf{A})^{-1}) \\ &= \rho((z\mathbf{U}\mathbf{U}^* - \mathbf{U}\mathbf{D}\mathbf{U}^*)^{-1}) \\ &= \rho(\mathbf{U}(z\mathbf{I} - \mathbf{D})^{-1}\mathbf{U}^*) \\ &= \rho((z\mathbf{I} - \mathbf{D})^{-1}) \\ &= \frac{1}{|z - \lambda_1|} \end{aligned}$$

The last equality comes since the diagonal (and thus eigenvalues) of $(z\mathbf{I}-\mathbf{D})^{-1}$ are $\frac{1}{z-\lambda_1}, \frac{1}{z-\lambda_2}, \cdots, \frac{1}{z-\lambda_n}$. Since z is arbitrary, the equality holds for all $z \in \mathbb{C}\backslash\sigma(\mathbf{A})$, and we may replace $\left|\frac{1}{z-\lambda}\right|$ with $\frac{1}{\text{dist}(z,\sigma(\mathbf{A}))}$. □

1.5 Pseudospectra

Eigenvalues and eigenvectors are extremely powerful tools in working with matrices. When one knows the eigenvalues of a square matrix $\mathbf{A}$, one can immediately determine the determinant, trace, and invertibility of $\mathbf{A}$. We have seen that if $\mathbf{A}$ is normal, and hence unitarily diagonalizable, then it can be expressed entirely in terms of it's eigenvalues and eigenvectors. And yet, there is a certain instability associated with eigenvalues. For example, a matrix $\mathbf{A}$ with $\lambda = 0$ as an eigenvalue is known to be singular, a highly undesirable property. In this case, the equation $\mathbf{A}\mathbf{x} = \mathbf{b}$ does not have a unique solution, and $\mathbf{A}^{-1}$ does not exist. However, an arbitrarily small perturbation of the entries of $\mathbf{A}$ can immediately change this. Consider $\mathbf{A}$ as the $N \times N$ identity matrix, with the entry in the first row and first column, a_{11}, switched to 0. Let $\mathbf{B}$ be the matrix obtained by changing a_{11} from zero to some arbitrarily small value ε. We succeed in turning a rank deficient, zero determinant, matrix into a full rank, nonzero determinant matrix. The solution to $\mathbf{B}\mathbf{x} = \mathbf{b}$ is now unique, and $\mathbf{B}^{-1}$ now exists! On the other hand, starting with $\mathbf{B}$ and reversing the step above

to obtain $\mathbf{A}$, we also reverse our success.

On closer inspection, the matrix $\mathbf{B}$ will have a determinant equal to ε; the spectral norm and the determinant of the inverse will be ε^{-1}, a potentially large number. This is a red flag that computational analysis done on such a matrix may run into problems. While the matrix $\mathbf{B}$ may have analytically desirable properties, it suffers the same problems as $\mathbf{A}$ in many practical applications.

Still, if perturbing matrix entries can have such a profound effect on matrix properties, one might legitimately wonder how this technique might be employed to solve—or at least better understand—problems involving matrices with undesirable properties. One might also call into question the accuracy of scientific conclusions where measurement error creates perturbation of matrix entries. Indeed, measurement error is ubiquitous in problems of applied mathematics, and understanding the effects of small perturbations on matrix entries is another step in knowing the quality of scientific results.

To understand how perturbing matrix entries can affect matrix properties, we must look beyond the spectrum. The ***ε-pseudospectrum*** of a matrix $\mathbf{A} \in M_N$, is the subset

$$\sigma_\varepsilon(\mathbf{A}) = \left\{ z \in \mathbb{C} : ||(z\mathbf{I} - \mathbf{A})^{-1}|| > \varepsilon^{-1} \right\}$$

where $\varepsilon > 0$, and $|| \cdot ||$ is some pre-determined norm. An element $z \in \sigma_\varepsilon(\mathbf{A})$ is referred to as an ***ε-pseudoeigenvalue*** of $\mathbf{A}$. The maximum modulus attained by the elements of $\sigma_\varepsilon(\mathbf{A})$, which we denote $\rho_\varepsilon(\mathbf{A})$, is called the ***$\varepsilon$-pseudospectral radius*** of $\mathbf{A}$. An $N \times N$ matrix $\mathbf{A}$ has an ε-pseudospectrum and an ε-pseudospectral radius for each $\varepsilon > 0$. In plural, the sets $\sigma_\varepsilon(\mathbf{A})$ are referred to as the ε-pseudospectra of $\mathbf{A}$. We see immediately that the ε-pseudospectra are nested; that is, given $0 < \varepsilon_1 < \varepsilon_2$, we have:

$$\sigma_{\varepsilon_1}(\mathbf{A}) \subseteq \sigma_{\varepsilon_2}(\mathbf{A}) \qquad (\text{since } \varepsilon_1^{-1} > \varepsilon_2^{-1})$$

Theorem 1.17 shows that, given $\lambda \in \sigma(\mathbf{A})$, the convention

$$\left|\left|(\lambda\mathbf{I} - \mathbf{A})^{-1}\right|\right| = \infty$$

is justified due to the following:

$$\infty = \lim_{z\to\lambda} \frac{1}{\operatorname{dist}(z,\sigma(\mathbf{A}))} \leq \lim_{z\to\lambda} \left|\left|(z\mathbf{I} - \mathbf{A})^{-1}\right|\right|$$

Then the following also holds:

$$\sigma(\mathbf{A}) = \bigcap_{\varepsilon>0} \sigma_\varepsilon(\mathbf{A})$$

The next theorem shows that an ε-pseudospectrum contains the set of points that are within ε of the spectrum.

Theorem 1.19. *Given $\boldsymbol{A} \in M_N$ and $\varepsilon > 0$,*

$$\sigma_\varepsilon(\boldsymbol{A}) \supseteq \sigma(\boldsymbol{A}) + \{z : |z| < \varepsilon\} \tag{1.9}$$

Proof.

$$\begin{aligned}
\sigma(\mathbf{A}) + \{z : |z| < \varepsilon\} &= \{\lambda + z : \lambda \in \sigma(\mathbf{A}), |z| < \varepsilon\} \\
&= \{\omega : \omega = \lambda + z, \lambda \in \sigma(\mathbf{A}), |z| < \varepsilon\} \\
&= \{\omega : \omega - \lambda = z, \lambda \in \sigma(\mathbf{A}), |z| < \varepsilon\} \\
&= \{\omega : \lambda \in \sigma(\mathbf{A}), |\omega - \lambda| < \varepsilon\} \\
&= \{\omega : \operatorname{dist}(\omega, \sigma(\mathbf{A})) < \varepsilon\} \\
&= \left\{\omega : \frac{1}{\operatorname{dist}(\omega, \sigma(\mathbf{A}))} > \varepsilon^{-1}\right\}
\end{aligned} \tag{1.10}$$

By Theorem 1.17 and the definition of the pseudospectrum, this last set is contained in $\sigma_\varepsilon(\mathbf{A})$. □

Corollary 1.20. *If the matrix $\boldsymbol{A} \in M_N$ is normal, then*

$$\sigma_\varepsilon(\boldsymbol{A}) = \sigma(\boldsymbol{A}) + \{z : |z| < \varepsilon\}$$

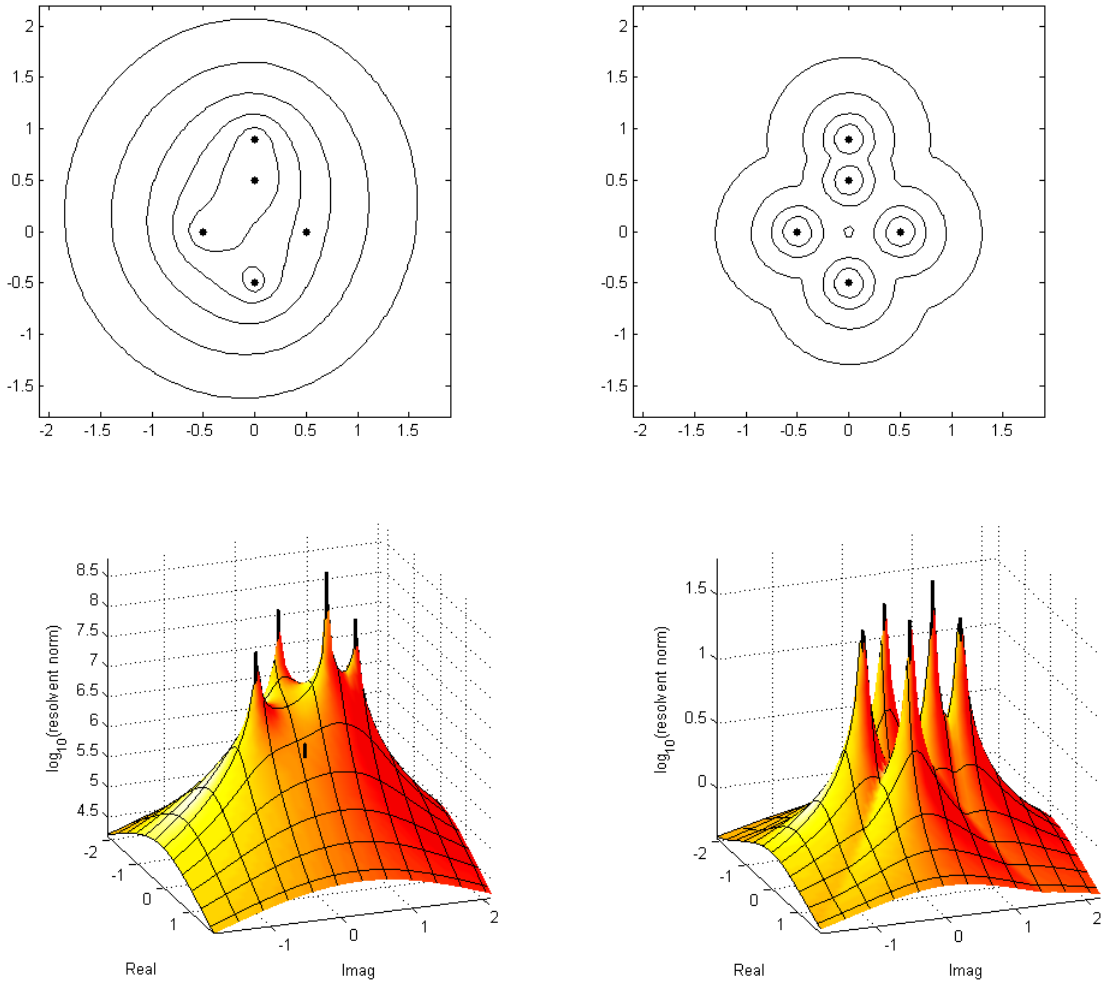
Proof. By Corollary 1.18, if $\mathbf{A}$ is normal, then $||(z\mathbf{I} - \mathbf{A})^{-1}||_2 = \frac{1}{\operatorname{dist}(z,\sigma(\mathbf{A}))}$. By substituting the right-hand side into equation 1.10 above, the corollary is proved. □

Since $\left|\left|(z\mathbf{I} - \mathbf{A})^{-1}\right|\right|$ is positive for $z \in \mathbb{C}\backslash\sigma(\mathbf{A})$, its graph is a surface over the complex plane

with vertical asymptotes at each of the eigenvalues of $\mathbf{A}$. When $\mathbf{A}$ is normal, the level curves of the norm of the resolvent are marked by circles and curves that are equidistant from the spectrum. However, in the cases where $\mathbf{A}$ is a nonnormal matrix, the level curves may take on many diverse forms. For example, the matrices

$$\mathbf{A} = \begin{bmatrix} \frac{1}{2} & 0 & 0 & 0 & 0 \\ 0 & \frac{i}{2} & 1000 & 0 & 0 \\ 0 & 0 & \frac{-i}{2} & 1 & 0 \\ 0 & 0 & 0 & \frac{-1}{2} & 1000 \\ 0 & 0 & 0 & 0 & \frac{9i}{10} \end{bmatrix} \quad \text{and} \quad \mathbf{D} = \begin{bmatrix} \frac{1}{2} & 0 & 0 & 0 & 0 \\ 0 & \frac{i}{2} & 0 & 0 & 0 \\ 0 & 0 & \frac{-i}{2} & 0 & 0 \\ 0 & 0 & 0 & \frac{-1}{2} & 0 \\ 0 & 0 & 0 & 0 & \frac{9i}{10} \end{bmatrix}$$

Figure 1.1: Pseudospectra and Resolvents of $\mathbf{A}$ and $\mathbf{D}$



have the same eigenvalues, but their ε-pseudospectra are quite different, particularly since $\mathbf{D}$ is normal, but $\mathbf{A}$ is not. This is evident in their respective graphs in Figure 1.5, where the eigenvalues are plotted as points (or poles in the 3D graphs). The interior of each closed contour in the 2D graphs is an ε-pseudospectrum for different values of ε.

Chapter 2. Expressing $\mathbf{A}^k$

In this chapter, we discuss three basic ways of expressing a square matrix $\mathbf{A} \in M_N$. If $\mathbf{A}$ is normal, then the Spectral Theorem (Theorem 1.4) comes to the rescue. If not, then we make use of Schur's Unitary Triangularization Theorem to construct a more complex expression. We also look at an integral expression of $\mathbf{A}^k$ involving the resolvent.

Suppose that $\mathbf{A} \in M_N$ is normal. Then by the Spectral Theorem (Theorem 1.4), there exists a diagonal matrix $\mathbf{D}$ and unitary matrix $\mathbf{U}$, such that $\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{U}^*$. A straightforward inductive proof on k, using the fact that $\mathbf{U}^*\mathbf{U} = \mathbf{I}$, shows that,

$$\mathbf{A}^k = \mathbf{U}\mathbf{D}^k\mathbf{U}^* \tag{2.1}$$

Since working with a diagonal matrix is very simple, this decomposition is invaluable when it can be applied.

For the more general case where $\mathbf{A} \in M_N$ is not assumed to be normal, we may still take advantage of the Schur unitary triangularization theorem (Theorem 1.5): there exists an upper-triangular matrix $\mathbf{T}$, and unitary matrix $\mathbf{U}$, such that $\mathbf{A} = \mathbf{U}\mathbf{T}\mathbf{U}^*$. Again, an inductive process yields the decomposition:

$$\mathbf{A}^k = \mathbf{U}\mathbf{T}^k\mathbf{U}^* \tag{2.2}$$

for nonnegative integers k. This decomposition can be much harder to work with when $\mathbf{T}$ is not diagonal. For this reason, we must develop a deeper understanding of the structure of the upper-triangular matrix $\mathbf{T}^k$, and its relationship to $\mathbf{T}$.

2.1 Expressing $\mathbf{T}^k$

For this section, we will assume that we have been given a square matrix $\mathbf{A}$, with Schur decomposition $\mathbf{A} = \mathbf{UTU}^*$. The following notation will be used: $\mathbf{T}^k = \left(t_{mn}^{(k)}\right)$, where $t_{mn}^{(k)}$ is the element from the m^{th} row and nth column of $\mathbf{T}^k$. In the special case that $k = 1$, we write

$$\mathbf{T} = (t_{mn}) = \begin{bmatrix} \lambda_1 & t_{12} & t_{13} & \dots & t_{1N} \\ 0 & \lambda_2 & t_{23} & \dots & t_{2N} \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & t_{N-1,N} \\ 0 & \dots & \dots & 0 & \lambda_N \end{bmatrix}$$

where $\lambda_1, \lambda_2, \dots, \lambda_N$ are the eigenvalues of $\mathbf{A}$.

To develop bounds for the norm of powers of an upper triangular matrix $||\mathbf{T}^k||$, we will find it useful to express the matrix $\mathbf{T}^k$ in terms of the entries of the matrix $\mathbf{T}$. This will be accomplished in Theorem 2.4. First, we set the stage for the theorem with three lemmata. The first lemma is a component-wise version of the statement, "the product of upper-triangular matrices is an upper-triangular matrix."

Lemma 2.1. *Given two $N \times N$ upper triangular matrices $\boldsymbol{V} = (v_{mn})$ and $\boldsymbol{W} = (w_{mn})$, the mn^{th} entry of VW is given by:*

$$(\boldsymbol{VW})_{mn} = \begin{cases} \displaystyle\sum_{m \le r \le n} v_{mr} w_{rn} & \text{if } m \le n \\ 0 & \text{if } m > n \end{cases}$$

Proof. Since $\mathbf{V}$ and $\mathbf{W}$ are upper triangular, $v_{mr} = 0$ for $r < m$ and $w_{rn} = 0$ for $r > n$. The proof follows readily from these facts and the identity from matrix multiplication,

$$(\mathbf{VW})_{mn} = \sum_{r=1}^{N} v_{mr} w_{rn} \qquad \text{for each} \quad m, n \in \{1, 2, \dots, N\}$$

□

The next lemma makes use of the multi-index notation $\Delta_j = (\delta_1, \delta_2, \dots, \delta_j)$, for $j = 0, 1, 2, \dots$.

Each component of Δ_j is considered to be a nonnegative integer, and we define the norm of Δ_j by $|\Delta_j| := \sum_{i=1}^{j} \delta_i$. Although the identity in the next lemma is quite complex, we will need it for simplification in Theorem 2.4.

Lemma 2.2. *Given an $N \times N$ upper triangular matrix $\boldsymbol{T} = (t_{mn})$, fix m and n with $m < n$. Let $\lambda_1, \lambda_2, \ldots, \lambda_N$ be the eigenvalues of $\boldsymbol{T}$. Then,*

$$t_{mn}\lambda_n^k + \sum_{m<r<n} t_{mr} \sum_{j=1}^{n-m} \sum_{\substack{r=\alpha_1<\alpha_2 \\ <\ldots<\alpha_{j+1}=n}} \left(\prod_{i=1}^{j} t_{\alpha_i,\alpha_{i+1}} \right) \sum_{|\Delta_{j+1}|=k-j} \prod_{l=1}^{j+1} \lambda_{\alpha_l}^{\delta_l} \tag{2.3}$$

$$= \sum_{j=1}^{n-m} \sum_{\substack{m=\alpha_1<\alpha_2 \\ <\ldots<\alpha_{j+1}=n}} \left(\prod_{i=1}^{j} t_{\alpha_i\alpha_{i+1}} \right) \sum_{|\Delta_{j+1}|=(k+1)-j} \prod_{l=1}^{j+1} \lambda_{\alpha_l}^{\delta_l} \Bigg|_{\delta_1=0} \tag{2.4}$$

Proof. The proof proceeds as follows. There are two terms added together in (2.3). First, we change the form of the second term, and then that of the first. Then we add the results together to obtain (2.4).

Beginning with the second term of (2.3), note that for each r with $m < r < n$, there are at most $n - m$ integers between r and n, including n but not r; this is because there are $n - m$ integers between m and n, not including m. There are $j + 1$ α's which must be chosen in the summation $\sum_{\substack{r=\alpha_1<\alpha_2 \\ <\ldots<\alpha_{j+1}=n}}$; thus, the summation returns 0 when $j = n - m$, since $j + 1$ α's are not possible in this case. Then we can replace $\sum_{j=1}^{n-m}$ by $\sum_{j=1}^{n-m-1}$ in the second term of (2.3):

$$\sum_{m<r<n} t_{mr} \sum_{j=1}^{n-m} \sum_{\substack{r=\alpha_1<\alpha_2 \\ <\ldots<\alpha_{j+1}=n}} \left(\prod_{i=1}^{j} t_{\alpha_i,\alpha_{i+1}} \right) \sum_{|\Delta_{j+1}|=k-j} \prod_{l=1}^{j+1} \lambda_{\alpha_l}^{\delta_l}$$

$$= \sum_{m<r<n} t_{mr} \sum_{j=1}^{n-m-1} \sum_{\substack{r=\alpha_1<\alpha_2 \\ <\ldots<\alpha_{j+1}=n}} \left(\prod_{i=1}^{j} t_{\alpha_i,\alpha_{i+1}} \right) \sum_{|\Delta_{j+1}|=k-j} \prod_{l=1}^{j+1} \lambda_{\alpha_l}^{\delta_l}$$

Rearranging the order of summation gives:

$$= \sum_{j=1}^{n-m-1} \sum_{m<r<n} \sum_{\substack{r=\alpha_1<\alpha_2 \\ <\ldots<\alpha_{j+1}=n}} \left(t_{mr} \prod_{i=1}^{j} t_{\alpha_i,\alpha_{i+1}} \right) \sum_{|\Delta_{j+1}|=k-j} \prod_{l=1}^{j+1} \lambda_{\alpha_l}^{\delta_l}$$

Now, change the bounds of the first summation to start at $j = 2$, and end at $n - m$:

$$= \sum_{j=2}^{n-m} \sum_{m<r<n} \sum_{\substack{r=\alpha_1<\alpha_2 \\ <\ldots<\alpha_j=n}} \left(t_{mr} \prod_{i=1}^{j-1} t_{\alpha_i,\alpha_{i+1}} \right) \sum_{|\Delta_j|=k-(j-1)} \prod_{l=1}^{j} \lambda_{\alpha_l}^{\delta_l}$$

Now, we wish to pass t_{mr} into the product $\prod_{i=1}^{j} t_{\alpha_i,\alpha_{i+1}}$. To accomplish this 1) we need to change α_1 to m, and 2) we need α_2 to range from $m + 1$ to $n - 1$; the second part will be provided for by the summation $\sum_{m<r<n}$. Also, note that for each j, when t_{mr} is passed into the product $\prod_{i=1}^{j} t_{\alpha_i,\alpha_{i+1}}$, the product will then have one more element; therefore, we will need j to begin at 2, and end at $n - m$. Finally, we must not ignore the last summation $\sum_{|\Delta_j|=k-j} \prod_{l=1}^{j} \lambda_{\alpha_l}^{\delta_l}$. Since α_1 will become m, but λ_m is not included in the original products, we must set $\delta_1 = 0$. Also, making $\alpha_1 = m$, means that we will have an additional element in each product of the summation; thus, we must change the product upper bound from j to $j + 1$, and Δ_j will again become Δ_{j+1}. However, the sum of the exponents will remain unchanged at $k - (j - 1) = (k + 1) - j$. Putting everything together we continue from above:

$$= \sum_{j=2}^{n-m} \sum_{\substack{m=\alpha_1<\alpha_2 \\ <\ldots<\alpha_{j+1}=n}} \left(\prod_{i=1}^{j} t_{\alpha_i,\alpha_{i+1}} \right) \sum_{|\Delta_{j+1}|=(k+1)-j} \prod_{l=1}^{j+1} \lambda_{\alpha_l}^{\delta_l} \Bigg|_{\delta_1=0} \tag{2.5}$$

Now that we have a new form of the second term of (2.3), let's turn our attention to the first term, $t_{mn}\lambda_n^k$. Note that,

$$t_{mn}\lambda_n^k = \sum_{m=\alpha_1<\alpha_2=n} \left(\prod_{i=1}^{1} t_{\alpha_i,\alpha_{i+1}} \right) \sum_{\Delta_2=(k+1)-1} \lambda_{\alpha_1}^{\delta_1} \lambda_{\alpha_2}^{\delta_2} \Bigg|_{\delta_1=0}$$

This is just the 1^{st} term of the 1^{st} sum (for $j = 1$) of (2.4) above. Adding it to (2.5) yields (2.4). □

One more lemma will prove useful in simplifying the work of Theorem 2.4.

Lemma 2.3.

$$\sum_{|\Delta_{j+1}|=(k+1)-j} \prod_{l=1}^{j+1} \lambda_{\alpha_l}^{\delta_l} \quad = \quad \sum_{|\Delta_{j+1}|=k-j} \lambda_{\alpha_1}^{\delta_1+1} \lambda_{\alpha_2}^{\delta_2} \cdots \lambda_{\alpha_{j+1}}^{\delta_{j+1}} + \sum_{|\Delta_{j+1}|=(k+1)-j} \prod_{l=1}^{j+1} \lambda_{\alpha_l}^{\delta_l} \Bigg|_{\delta_1=0}$$

Proof. Notice that the exponents of the λ's in both the first and second terms on the right sum to $(k+1)-j$, (since $|\Delta_{j+1}| = \delta_1 + \delta_2 + \ldots + \delta_{j+1}$). Now observe that the first term on the right is the sum of all products of the form $\lambda_{\alpha_1}^{\eta_1} \lambda_{\alpha_2}^{\eta_2} \cdots \lambda_{\alpha_{j+1}}^{\eta_{j+1}}$, where η_1 can be any value from 1 to $(k+1)-j$. In contrast, the second term on the right is the sum of all products of the same form, where η_1 is restricted to be 0. This complementary nature of the first and second terms on the right allows us to combine them into a single summation. □

Armed with Lemmas 2.1, 2.2, and 2.3, we are now ready for a major theorem regarding the values of the entries of $\mathbf{T}^k$.

Theorem 2.4. *Given $k \in \mathbb{N}$ and an upper-triangular matrix $\boldsymbol{T} = (t_{mn}) \in M_N$, the mn^{th} entry of $\boldsymbol{T}^k$ is given by,*

$$t_{mn}^{(k)} = \begin{cases} \lambda_m^k & \text{if } m = n \\ \displaystyle\sum_{j=1}^{n-m} \sum_{\substack{m=\alpha_1<\alpha_2 \\ <\ldots<\alpha_{j+1}=n}} \left(\prod_{i=1}^{j} t_{\alpha_i \alpha_{i+1}} \right) \sum_{|\Delta_{j+1}|=k-j} \prod_{l=1}^{j+1} \lambda_{\alpha_l}^{\delta_l} & \text{if } m < n \\ 0 & \text{if } m > n \end{cases}$$

which may be written in the slightly more compact form,

$$t_{mn}^{(k)} = \begin{cases} \lambda_m^k & \text{if } m = n \\ \displaystyle\sum_{j=1}^{n-m} \sum_{\substack{m=\alpha_1<\alpha_2 \\ <\ldots<\alpha_{j+1}=n}} \left(\prod_{i=1}^{j} t_{\alpha_i \alpha_{i+1}} \right) \sum_{|\Delta_{j+1}|=k-j} \prod_{l=1}^{j+1} \lambda_{\alpha_l}^{\delta_l} & \text{if } m \neq n \end{cases}$$

Remark. The fact that the summation $\displaystyle\sum_{\substack{m=\alpha_1<\alpha_2 \\ <\ldots<\alpha_{j+1}=n}}$ returns 0 if $m > n$ demonstrates the equivalence of the more compact form above to the former representation.

Proof. We proceed with an inductive proof. Initial step: suppose $k = 1$. Then, $t_{mn}^{(k)} = t_{mn}$, and if $m = n$, then $t_{mn} = \lambda_m$. If $m > n$, then $t_{mn} = 0$. If $m < n$, notice that

$$\sum_{|\Delta_{j+1}|=1-j} \prod_{l=1}^{j+1} \lambda_{\alpha_l}^{\delta_l} = \begin{cases} 1 & \text{if } j = 1 \\ 0 & \text{if } j > 1 \end{cases} \tag{2.6}$$

Working backwards from the theorem statement, we have,

$$\begin{aligned} &\sum_{j=1}^{n-m} \sum_{\substack{m=\alpha_1<\alpha_2 \\ <\ldots<\alpha_{j+1}=n}} \left(\prod_{i=1}^{j} t_{\alpha_i\alpha_{i+1}} \right) \sum_{|\Delta_{j+1}|=k-j} \prod_{l=1}^{j+1} \lambda_{\alpha_l}^{\delta_l} \\ &\quad = \sum_{j=1}^{n-m} \sum_{\substack{m=\alpha_1<\alpha_2 \\ <\ldots<\alpha_{j+1}=n}} \left(\prod_{i=1}^{j} t_{\alpha_i\alpha_{i+1}} \right) \sum_{|\Delta_{j+1}|=1-j} \prod_{l=1}^{j+1} \lambda_{\alpha_l}^{\delta_l} \qquad \text{(since } k = 1\text{)} \end{aligned}$$

By equation (2.6) we may replace the last summation in the expression above to get,

$$= \sum_{\substack{m=\alpha_1<\alpha_2 \\ <\ldots<\alpha_{j+1}=n}} \left(\prod_{i=1}^{j} t_{\alpha_i\alpha_{i+1}} \right) \Bigg|_{j=1} \quad = \quad t_{mn}$$

Inductive step: now suppose that the theorem holds for some $k \in \mathbb{N}$. We wish to show that,

$$t_{mn}^{(k+1)} = \begin{cases} \lambda_m^{k+1} & \text{if } m = n \\ \displaystyle\sum_{j=1}^{n-m} \sum_{\substack{m=\alpha_1<\alpha_2 \\ <\ldots<\alpha_{j+1}=n}} \left(\prod_{i=1}^{j} t_{\alpha_i\alpha_{i+1}} \right) \sum_{|\Delta_{j+1}|=(k+1)-j} \prod_{l=1}^{j+1} \lambda_{\alpha_l}^{\delta_l} & \text{if } m < n \\ 0 & \text{if } m > n \end{cases}$$

By properties of matrix multiplication, and since $\mathbf{T}^{k+1} = \mathbf{T}\mathbf{T}^k$, we have $t_{mn}^{(k+1)} = \sum_{r=1}^{N} t_{mr} t_{rn}^{(k)}$. If $m > n$ then Lemma 2.1 and the inductive step give $t_{mn}^{(k+1)} = 0$. If $m = n$ then Lemma 2.1 and the inductive step give,

$$t_{mn}^{(k+1)} = t_{mm}^{k+1} = \sum_{r=1}^{N} t_{mr} t_{rm}^{(k)} = \sum_{m \le r \le m} t_{mr} t_{rm}^{(k)} = t_{mm} t_{mm}^{(k)} = \lambda_m^{k+1}$$

Finally, if $m < n$ then by matrix multiplication,

$$t_{mn}^{(k+1)} \;=\; \sum_{r=1}^{N} t_{mr} t_{rn}^{(k)}$$

By Lemma 2.1 this becomes,

$$\sum_{m \le r \le n} t_{mr} t_{rn}^{(k)} \;=\; t_{mm} t_{mn}^{(k)} \;+\; t_{mn} t_{nn}^{(k)} \;+\; \sum_{m < r < n} t_{mr} t_{rn}^{(k)}$$

Now we use the initial step and the inductive step to replace t_{mm}, $t_{nn}^{(k)}$, and $t_{rn}^{(k)}$.

$$\lambda_m t_{mn}^{(k)} \;+\; t_{mn} \lambda_n^k \;+\; \sum_{m<r<n} t_{mr} \left(\sum_{j=1}^{n-m} \sum_{\substack{r=\alpha_1<\alpha_2 \\ <\ldots<\alpha_{j+1}=n}} \left(\prod_{i=1}^{j} t_{\alpha_i \alpha_{i+1}} \right) \sum_{|\Delta_{j+1}|=k-j} \prod_{l=1}^{j+1} \lambda_{\alpha_l}^{\delta_l} \right)$$

By Lemma 2.2, the last two terms can be combined to give:

$$\lambda_m t_{mn}^{(k)} \;+\; \sum_{j=1}^{n-m} \sum_{\substack{m=\alpha_1<\alpha_2 \\ <\ldots<\alpha_{j+1}=n}} \left(\prod_{i=1}^{j} t_{\alpha_i \alpha_{i+1}} \right) \sum_{|\Delta_{j+1}|=(k+1)-j} \prod_{l=1}^{j+1} \lambda_{\alpha_l}^{\delta_l} \Bigg|_{\delta_1=0}$$

Now we use the inductive step again to substitute for $t_{mn}^{(k)}$.

$$\begin{aligned} &\lambda_m \left[\sum_{j=1}^{n-m} \sum_{\substack{m=\alpha_1<\alpha_2 \\ <\ldots<\alpha_{j+1}=n}} \left(\prod_{i=1}^{j} t_{\alpha_i \alpha_{i+1}} \right) \sum_{|\Delta_{j+1}|=k-j} \prod_{l=1}^{j+1} \lambda_{\alpha_l}^{\delta_l} \right] \\ &\quad + \sum_{j=1}^{n-m} \sum_{\substack{m=\alpha_1<\alpha_2 \\ <\ldots<\alpha_{j+1}=n}} \left(\prod_{i=1}^{j} t_{\alpha_i \alpha_{i+1}} \right) \sum_{|\Delta_{j+1}|=(k+1)-j} \prod_{l=1}^{j+1} \lambda_{\alpha_l}^{\delta_l} \Bigg|_{\delta_1=0} \end{aligned}$$

Passing λ_m into the product of λ's gives:

$$\begin{aligned} &= \sum_{j=1}^{n-m} \sum_{\substack{m=\alpha_1<\alpha_2 \\ <\ldots<\alpha_{j+1}=n}} \left(\prod_{i=1}^{j} t_{\alpha_i \alpha_{i+1}} \right) \sum_{|\Delta_{j+1}|=k-j} \lambda_{\alpha_1}^{\delta_1+1} \lambda_{\alpha_2}^{\delta_2} \cdots \lambda_{\alpha_{j+1}}^{\delta_{j+1}} \\ &\quad + \sum_{j=1}^{n-m} \sum_{\substack{m=\alpha_1<\alpha_2 \\ <\ldots<\alpha_{j+1}=n}} \left(\prod_{i=1}^{j} t_{\alpha_i \alpha_{i+1}} \right) \sum_{|\Delta_{j+1}|=(k+1)-j} \prod_{l=1}^{j+1} \lambda_{\alpha_l}^{\delta_l} \Bigg|_{\delta_1=0} \end{aligned}$$

Rearranging terms we get:

$$= \sum_{j=1}^{n-m} \sum_{\substack{m=\alpha_1<\alpha_2 \\ <\ldots<\alpha_{j+1}=n}} \left(\prod_{i=1}^{j} t_{\alpha_i \alpha_{i+1}} \right) \left[\sum_{|\Delta_{j+1}|=k-j} \lambda_{\alpha_1}^{\delta_1+1} \lambda_{\alpha_2}^{\delta_2} \cdots \lambda_{\alpha_{j+1}}^{\delta_{j+1}} + \sum_{|\Delta_{j+1}|=(k+1)-j} \prod_{l=1}^{j+1} \lambda_{\alpha_l}^{\delta_l} \Big|_{\delta_1=0} \right]$$

Finally, by Lemma 2.3 we may combine the summations in brackets and finish the proof:

$$\sum_{j=1}^{n-m} \sum_{\substack{m=\alpha_1<\alpha_2 \\ <\ldots<\alpha_{j+1}=n}} \left(\prod_{i=1}^{j} t_{\alpha_i \alpha_{i+1}} \right) \sum_{|\Delta_{j+1}|=(k+1)-j} \prod_{l=1}^{j+1} \lambda_{\alpha_l}^{\delta_l}$$

□

Having obtained the mn^{th} term of the matrix $\mathbf{T}^k$ we may now write $\mathbf{T}^k$ entirely in terms of the entries of $\mathbf{T}$. For a 3×3 matrix this becomes:

$$\mathbf{T}^k = \begin{bmatrix} \lambda_1^k & t_{12} \sum_{|\Delta_2|=k-1} \lambda_1^{\delta_1} \lambda_2^{\delta_2} & t_{13} \sum_{|\Delta_2|=k-1} \lambda_1^{\delta_1} \lambda_3^{\delta_2} + t_{12} t_{23} \sum_{|\Delta_3|=k-2} \lambda_1^{\delta_1} \lambda_2^{\delta_2} \lambda_3^{\delta_3} \\ 0 & \lambda_2^k & t_{23} \sum_{|\Delta_2|=k-1} \lambda_2^{\delta_1} \lambda_3^{\delta_2} \\ 0 & 0 & \lambda_3^k \end{bmatrix} \tag{2.7}$$

In equation (2.7), notice that k shows up as an exponent in the diagonal entries, and also in each of the summations. As k gets large, this will create a very large number of terms for each of the summations; thus, it will be better to develop a new expression in which the number of terms in the summations depends only on the size of the matrix, rather than on k. In the next subsection we show how to do this in cases where the eigenvalues are either all equal, or all distinct.

2.1.1 Expressing $\sum_{|\Delta_{j+1}|=C} \prod_{l=1}^{j+1} \lambda_{\alpha_l}^{\delta_l}$. We will show shortly that the last summation in the expression for $t_{mn}^{(k)}$ in Theorem 2.4, when $m < n$, has $\binom{k}{j}$ terms, for each j. As k becomes large, computation of this summation becomes very costly, because the number of operations depends directly on the value of k. Therefore, it is advantageous to replace the last summation with a more practical alternative when possible. In the case that all of the eigenvalues are equal, or in the case that they are all distinct, we will derive alternative representations in which the number

of operations will depend only on the dimensions of the matrix $\mathbf{T}$, rather than on the power k.

First, consider the case in which all of the eigenvalues are equal. In this case,

$$\sum_{|\Delta_{j+1}|=k-j} \prod_{l=1}^{j+1} \lambda_{\alpha_l}^{\delta_l} = \sum_{|\Delta_{j+1}|=k-j} \lambda^{k-j} = \lambda^{k-j} \sum_{|\Delta_{j+1}|=k-j} 1 \tag{2.8}$$

where $\lambda = \lambda_{\alpha_1} = \lambda_{\alpha_2} = \ldots = \lambda_{\alpha_{j+1}}$. To evaluate the sum, $\sum_{|\Delta_{j+1}|=k-j} 1$, we need a lemma from combinatorics.

Lemma 2.5. *There are* $\binom{C+i-1}{i-1} = \binom{C+i-1}{C}$ *different sums of the form*

$$\delta_1 + \delta_2 + \ldots + \delta_i = C$$

where the δ's are nonnegative integers.

Proof. First, note that the two expressions are equal since,

$$\binom{C+i-1}{i-1} = \frac{(C+i-1)!}{C!(i-1)!} = \binom{C+i-1}{C}$$

To show that this is equal to the number of sums, consider dividing a line segment into i bins by means of $i-1$ different perpendicular notches, none of which occurs on the endpoints of the segment. Let δ_j correspond to the j^{th} bin for $j = 1, 2, ...$ Now select C distinct points on the line segment, none of which correspond with the positions of the $i-1$ notches. Count the number of points within each bin and let this be the value of the corresponding δ for that bin. By taking the sum of the δ's, we create one sum of the desired form. The problem of finding the total number of sums, is thus equivalent to the problem of finding the number of permutations of the C points and $i-1$ notches, without regard to the order of the notches among themselves, or to the order of the points among themselves; this is equal to $\frac{(C+i-1)!}{C!(i-1)!}$. □

Corollary 2.6. *There are* $\binom{k}{j}$ *different sums of the form*

$$\delta_1 + \delta_2 + \ldots + \delta_{j+1} = k - j$$

where the δ's are nonnegative integers.

Proof. This follows directly from Lemma 2.5, by substituting $k - j$ in for C, and $j + 1$ for i. $\square$

By Corollary 2.6 and equation (2.8), we have the following lemma:

Lemma 2.7. *Given* $\lambda = \lambda_{\alpha_1} = \lambda_{\alpha_2} = \ldots = \lambda_{\alpha_{j+1}}$,

$$\sum_{|\Delta_{j+1}|=k-j} \prod_{l=1}^{j+1} \lambda_{\alpha_l}^{\delta_l} = \sum_{|\Delta_{j+1}|=k-j} \lambda^{k-j} = \begin{cases} \binom{k}{j}\lambda^{k-j}, & \text{if } k \geq j \\ 0, & \text{if } k < j \end{cases}$$

The right-hand side of Lemma 2.7 is a simpler representation than the left-hand side, and therefore, it will be easier to code and faster to compute. In the case of equal eigenvalues, substituting the right-hand side for the last summation in Theorem 2.4 will greatly reduce the number of operations in computing the element $t_{mn}^{(k)}$. The 3×3 case on page 32, for $k > 2$ becomes,

$$\mathbf{T}^k = \begin{bmatrix} \lambda^k & t_{12}\binom{k}{1}\lambda^{k-1} & t_{13}\binom{k}{1}\lambda^{k-1} + t_{12}t_{23}\binom{k}{2}\lambda^{k-2} \\ 0 & \lambda^k & t_{23}\binom{k}{1}\lambda^{k-1} \\ 0 & 0 & \lambda^k \end{bmatrix} \tag{2.9}$$

Now, we derive a similar result in the case where all of the eigenvalues are distinct. This result is more complex, so we smooth the way with two preliminary lemmata.

Lemma 2.8. *Given* $j, C \in \mathbb{N}$ *and* $\lambda_1, \lambda_2, \ldots, \lambda_{j+1} \in \mathbb{C}$,

$$\sum_{|\Delta_{j+1}|=C} \prod_{l=1}^{j+1} \lambda_{\alpha_l}^{\delta_l} = \sum_{|\Delta_{j+1}|=C} \lambda_{\alpha_{j+1}}^{\delta_{j+1}} \prod_{l=1}^{j} \lambda_{\alpha_l}^{\delta_l} = \sum_{i=0}^{C} \lambda_{\alpha_{j+1}}^{C-i} \sum_{|\Delta_j|=i} \prod_{l=1}^{j} \lambda_{\alpha_l}^{\delta_l}$$

Proof. The first equality is rather trivial. We will focus on the second, which is also easily seen by

expanding part of the summation.

$$\begin{aligned}
&\sum_{|\Delta_{j+1}|=C} \lambda_{\alpha_{j+1}}^{\delta_{j+1}} \prod_{l=1}^{j} \lambda_{\alpha_l}^{\delta_l} \\
&\quad = \sum_{|\Delta_j|=C} \lambda_{\alpha_{j+1}}^{0} \prod_{l=1}^{j} \lambda_{\alpha_l}^{\delta_l} + \sum_{|\Delta_j|=C-1} \lambda_{\alpha_{j+1}} \prod_{l=1}^{j} \lambda_{\alpha_l}^{\delta_l} + \cdots + \sum_{|\Delta_j|=1} \lambda_{\alpha_{j+1}}^{C-1} \prod_{l=1}^{j} \lambda_{\alpha_l}^{\delta_l} + \lambda_{\alpha_{j+1}}^{C} \\
&\quad = \sum_{i=1}^{C} \lambda_{\alpha_{j+1}}^{C-i} \sum_{|\Delta_j|=i} \prod_{l=1}^{j} \lambda_{\alpha_l}^{\delta_l}
\end{aligned}$$

□

Lemma 2.9. *Given $j \in \mathbb{N}$ and $\lambda_1, \lambda_2, \ldots, \lambda_{j+1} \in \mathbb{C}$,*

$$-\lambda_{j+1}^{j-1} \prod_{h=1}^{j} (\lambda_{j+1} - \lambda_h)^{-1} = \sum_{r=1}^{j} \lambda_r^{j-1} \prod_{\substack{h=1 \\ h \neq r}}^{j+1} (\lambda_r - \lambda_h)^{-1}$$

Proof. Define the rational function $p(x)$ by replacing λ_{j+1} by the variable x in the left-hand side of the inequality above:

$$p(x) := -x^{j-1} \prod_{h=1}^{j} (x - \lambda_h)^{-1}$$

Now we can proceed with a proof by partial fraction decomposition. We must solve for the values of $c_1, c_2, \ldots, c_j$ in the following equation:

$$-x^{j-1} \prod_{h=1}^{j} (x - \lambda_h)^{-1} = \frac{c_1}{(x - \lambda_1)} + \frac{c_2}{(x - \lambda_2)} + \ldots + \frac{c_j}{(x - \lambda_j)} \tag{2.10}$$

By cross multiplying to get rid of fractions we get:

$$-x^{j-1} = c_1 \prod_{\substack{h=1 \\ h \neq 1}}^{j} (x - \lambda_h) + c_2 \prod_{\substack{h=1 \\ h \neq 2}}^{j} (x - \lambda_h) + \ldots + c_j \prod_{\substack{h=1 \\ h \neq j}}^{j} (x - \lambda_h)$$

Now replace x by λ_r, for each $r \in \{1, 2, \ldots, j\}$, and then solve for c_r to get:

$$c_1 = -\lambda_1^{j-1} \prod_{\substack{h=1 \\ h \neq 1}}^{j} (\lambda_1 - \lambda_h)^{-1}, \qquad c_2 = -\lambda_2^{j-1} \prod_{\substack{h=1 \\ h \neq 1}}^{j} (\lambda_2 - \lambda_h)^{-1},$$

$$, \ldots, \quad c_j = -\lambda_j^{j-1} \prod_{\substack{h=1 \\ h \neq 1}}^{j} (\lambda_j - \lambda_h)^{-1}$$

Plugging these values into equation (2.10) yields:

$$\frac{-x^{j-1}}{\prod_{h=1}^{j}(x - \lambda_h)} = -\lambda_1^{j-1} \left(\prod_{\substack{h=1 \\ h \neq 1}}^{j} (\lambda_1 - \lambda_h)^{-1} \right) (x - \lambda_1)^{-1} - \lambda_2^{j-1} \left(\prod_{\substack{h=1 \\ h \neq 2}}^{j} (\lambda_2 - \lambda_h)^{-1} \right) (x - \lambda_2)^{-1}$$

$$- \ldots - \quad \lambda_j^{j-1} \left(\prod_{\substack{h=1 \\ h \neq j}}^{j} (\lambda_j - \lambda_h)^{-1} \right) (x - \lambda_j)^{-1}$$

Finally, replace x by the value λ_{j+1} to finish the proof:

$$\frac{-\lambda_{j+1}^{j-1}}{\prod_{h=1}^{j}(\lambda_{j+1} - \lambda_h)} = -\lambda_1^{j-1} \left(\prod_{\substack{h=1 \\ h \neq 1}}^{j} (\lambda_1 - \lambda_h)^{-1} \right) (\lambda_{j+1} - \lambda_1)^{-1} - \lambda_2^{j-1} \left(\prod_{\substack{h=1 \\ h \neq 2}}^{j} (\lambda_2 - \lambda_h)^{-1} \right) (\lambda_{j+1} - \lambda_2)^{-1}$$

$$- \ldots - \lambda_j^{j-1} \left(\prod_{\substack{h=1 \\ h \neq j}}^{j} (\lambda_j - \lambda_h)^{-1} \right) (\lambda_{j+1} - \lambda_j)^{-1}$$

$$= \lambda_1^{j-1} \prod_{\substack{h=1 \\ h \neq 1}}^{j+1} (\lambda_1 - \lambda_h)^{-1} + \lambda_2^{j-1} \prod_{\substack{h=1 \\ h \neq 2}}^{j+1} (\lambda_2 - \lambda_h)^{-1} + \ldots + \lambda_j^{j-1} \prod_{\substack{h=1 \\ h \neq j}}^{j+1} (\lambda_j - \lambda_h)^{-1}$$

$$= \sum_{r=1}^{j} \lambda_r^{j-1} \prod_{\substack{h=1 \\ h \neq r}}^{j+1} (\lambda_r - \lambda_h)^{-1}$$

□

We are now ready to derive another form of the last summation in Theorem 2.4 for the case when all of the eigenvalues are distinct.

Theorem 2.10. *Given $C \in \mathbb{N}$, $j \in \mathbb{N}$, and $\lambda_1, \lambda_2, \ldots, \lambda_{j+1} \in \mathbb{C}$, with all of the λ's distinct,*

$$\sum_{|\Delta_{j+1}|=C} \prod_{r=1}^{j+1} \lambda_r^{\delta_r} = \sum_{r=1}^{j+1} \lambda_r^{C+j} \prod_{\substack{h=1 \\ h \neq r}}^{j+1} (\lambda_r - \lambda_h)^{-1} \tag{2.11}$$

Proof. We proceed with an inductive proof. Base case: suppose $j = 1$. Then,

$$\sum_{|\Delta_2|=C} \prod_{l=1}^{2} \lambda_{\alpha_l}^{\delta_l} = \sum_{|\Delta_2|=C} \lambda_{\alpha_1}^{\delta_1} \lambda_{\alpha_2}^{\delta_2} = \sum_{i=0}^{C} \lambda_{\alpha_1}^{C-i} \lambda_{\alpha_2}^{i} = \frac{\lambda_1^{C+1} - \lambda_2^{C+1}}{\lambda_1 - \lambda_2} = \frac{\lambda_1^{C+1}}{(\lambda_1 - \lambda_2)} + \frac{\lambda_2^{C+1}}{(\lambda_2 - \lambda_1)}$$

Inductive case: now suppose that the theorem statement holds for $j - 1$, for some $j \geq 2$. Then,

$$\begin{aligned}
\sum_{|\Delta_{j+1}|=C} \prod_{r=1}^{j+1} \lambda_r^{\delta_r} &= \sum_{i=0}^{C} \lambda_{j+1}^{C-i} \sum_{|\Delta_j|=i} \prod_{r=1}^{j} \lambda_r^{\delta_r} && \text{(by Lemma 2.8)} \\
&= \sum_{i=0}^{C} \lambda_{j+1}^{C-i} \sum_{r=1}^{j} \lambda_r^{i+j-1} \prod_{\substack{h=1 \\ h \neq r}}^{j} (\lambda_r - \lambda_h)^{-1} && \text{(by inductive step)} \\
&= \sum_{r=1}^{j} \prod_{\substack{h=1 \\ h \neq r}}^{j} (\lambda_r - \lambda_h)^{-1} \lambda_r^{j-1} \sum_{i=0}^{C} \lambda_{j+1}^{C-i} \lambda_r^{i} \\
&= \sum_{r=1}^{j} \prod_{\substack{h=1 \\ h \neq r}}^{j} (\lambda_r - \lambda_h)^{-1} \lambda_r^{j-1} \left(\frac{\lambda_r^{C+1} - \lambda_{j+1}^{C+1}}{\lambda_r - \lambda_{j+1}} \right) \\
&= \sum_{r=1}^{j} \prod_{\substack{h=1 \\ h \neq r}}^{j+1} (\lambda_r - \lambda_h)^{-1} \lambda_r^{j-1} \left(\lambda_r^{C+1} - \lambda_{j+1}^{C+1} \right) \\
&= \sum_{r=1}^{j} \lambda_r^{C+j} \prod_{\substack{h=1 \\ h \neq r}}^{j+1} (\lambda_r - \lambda_h)^{-1} - \lambda_{j+1}^{C+1} \sum_{r=1}^{j} \lambda_r^{j-1} \prod_{\substack{h=1 \\ h \neq r}}^{j+1} (\lambda_r - \lambda_h)^{-1}
\end{aligned}$$

Lemma 2.9 allows us to substitute for the last summation.

$$\begin{aligned}
&= \sum_{r=1}^{j} \lambda_r^{C+j} \prod_{\substack{h=1 \\ h \neq r}}^{j+1} (\lambda_r - \lambda_h)^{-1} - \lambda_{j+1}^{C+1} \left(-\lambda_{j+1}^{j-1} \prod_{h=1}^{j} (\lambda_{j+1} - \lambda_h)^{-1} \right) \\
&= \sum_{r=1}^{j+1} \lambda_r^{C+j} \prod_{\substack{h=1 \\ h \neq r}}^{j+1} (\lambda_r - \lambda_h)^{-1}
\end{aligned}$$

□

By replacing C in Theorem 2.10 by $k-j$, we can then substitute for the last summation in the right-hand side of Theorem 2.4. Note that this substitution is dependent upon $k > j$. If instead, $k < j$, then the summation $\sum_{|\Delta_{j+1}|=k-j}$ just returns zero; if $k = j$ then it returns 1. Again, the importance of this substitution arises from the fact that the number of terms to sum on the right-hand side of equation (2.11) depends only on j, while for the left-hand side, the number of operations depends on C. For large values of C, the right-hand side will be computationally more efficient. For the 3×3 case with $k > 3$ we have,

$$\mathbf{T}^k = \begin{bmatrix} \lambda_1^k & t_{12}\sum_{r=1}^{2}\lambda_r^k\prod_{\substack{h=1\\h\neq r}}^{2}(\lambda_r-\lambda_h)^{-1} & t_{13}\sum_{\substack{r=1\\r\neq 2}}^{3}\lambda_r^k\prod_{\substack{h=1\\h\neq r}}^{2}(\lambda_r-\lambda_h)^{-1} + t_{12}t_{23}\sum_{r=1}^{3}\lambda_r^k\prod_{\substack{h=1\\h\neq r}}^{2}(\lambda_r-\lambda_h)^{-1} \\ 0 & \lambda_2^k & t_{23}\sum_{r=1}^{2}\lambda_r^k\prod_{\substack{h=1\\h\neq r}}^{2}(\lambda_r-\lambda_h)^{-1} \\ 0 & 0 & \lambda_3^k \end{bmatrix}$$

The next lemma provides a bound for the sum of eigenvalues which we have been considering; it will be useful in developing upper bounds for the norm of matrix powers.

Lemma 2.11. *Given a nonzero upper-triangular matrix $\boldsymbol{T} \in M_N$ with eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_N$, and nonnegative integers k and j,*

$$\left| \sum_{|\Delta_{j+1}|=k-j} \prod_{l=1}^{j+1} \lambda_{\alpha_l}^{\delta_l} \right| \leq \sum_{|\Delta_{j+1}|=k-j} \prod_{l=1}^{j+1} |\lambda_{\alpha_l}^{\delta_l}| \leq \binom{k}{j} \rho(\boldsymbol{T})^{k-j}$$

Proof. The first inequality is trivial.

$$\left| \sum_{|\Delta_{j+1}|=k-j} \prod_{l=1}^{j+1} \lambda_{\alpha_l}^{\delta_l} \right| \leq \sum_{|\Delta_{j+1}|=k-j} \prod_{l=1}^{j+1} |\lambda_{\alpha_l}^{\delta_l}| = \sum_{|\Delta_{j+1}|=k-j} |\lambda_{\alpha_1}^{\delta_1}||\lambda_{\alpha_2}^{\delta_2}|\ldots|\lambda_{\alpha_{j+1}}^{\delta_{j+1}}|$$

$$\leq \sum_{|\Delta_{j+1}|=k-j} \rho(\mathbf{T})^{\delta_1+\delta_2+\ldots+\delta_{j+1}} = \sum_{|\Delta_{j+1}|=k-j} \rho(\mathbf{T})^{k-j} = \binom{k}{j}\rho(\mathbf{T})^{k-j}$$

The last two steps follow by the definition of $|\Delta_{j+1}|$, and by Corollary 2.6. Notice that this works

even for $k < j$. $\square$

Interestingly, if $\rho(\mathbf{T}) < 1$, the bound above approaches zero as k approaches infinity. The following lemma solidifies this concept.

Lemma 2.12. *Given a nonnegative integer j, and $0 \leq \rho < 1$,*

$$\lim_{k\to\infty} \binom{k}{j} \rho^{k-j} = 0$$

Proof. If $\rho = 0$, the lemma is trivial. Suppose that $\rho > 0$. If we replace k by the real-valued variable x, the limit on the left-hand side above is the same as

$$\lim_{x\to\infty} \frac{x(x-1)\cdots(x-j+1)}{j!} \rho^{x-j} = \lim_{x\to\infty} \frac{x(x-1)\cdots(x-j+1)}{j!\rho^{j-x}}$$

Observe that the numerator is a polynomial of degree j. Taking successive derivatives of the numerator yields new polynomials, each of degree one less than the last. Thus, the j^{th} derivative is a polynomial of degree 0; in fact, it is equal to $j!$, since the highest degree term from the original polynomial is x^j. The n^{th} derivative of the denominator is $(-1)^n j![\ln(\rho)]^n \rho^{j-x}$. Given these facts, it is permissible to apply L'Hopital's rule from calculus j times to get:

$$\lim_{x\to\infty} \frac{j!}{(-1)^j j![\ln(\rho)]^j \rho^{j-x}} = \lim_{x\to\infty} \frac{\rho^{x-j}}{(-1)^j [\ln(\rho)]^j} = 0$$

$\square$

Now that Theorem 2.4 has given us ways of expressing the powers of an upper-triangular matrix in terms of the entries of the original matrix, remember that we can combine this with Schur's Unitary Triangularization Theorem to provide a way to express the power of any square matrix, in terms of the entries of its upper-triangular Schur form.

2.2 $\mathbf{A}^k$ in terms of the Resolvent

In this section we look at a way of expressing the matrix power $\mathbf{A}^k$ in terms of an integral involving the resolvent of $\mathbf{A}$. This will be the substance of Theorem 2.16. As usual, we must develop a few

preliminary results. We start by showing that the resolvent of $\mathbf{A}$, $R(z) = (z\mathbf{I} - \mathbf{A})^{-1}$, can be expressed as a series involving $\mathbf{A}^k$.

Theorem 2.13. *For $z \in \mathbb{C}$ such that $|z| > ||\boldsymbol{A}||$,*

$$R(z) = \sum_{k=0}^{\infty} \frac{\boldsymbol{A}^k}{z^{k+1}}$$

Proof. First, note that for each $N \in \mathbb{N}$,

$$\begin{aligned} \mathbf{I} - (z^{-1}\mathbf{A})^{N+1} &= (\mathbf{I} - z^{-1}\mathbf{A})\left(\mathbf{I} + z^{-1}\mathbf{A} + (z^{-1}\mathbf{A})^2 + \ldots + (z^{-1}\mathbf{A})^N\right) \\ &= (\mathbf{I} - z^{-1}\mathbf{A}) \sum_{k=0}^{N} \frac{\mathbf{A}^k}{z^k} \end{aligned}$$

Taking the limit as $N \to \infty$ yields

$$\mathbf{I} = (\mathbf{I} - z^{-1}\mathbf{A}) \sum_{k=0}^{\infty} \frac{\mathbf{A}^k}{z^k}$$

multiplying by z gives

$$z\mathbf{I} = (z\mathbf{I} - \mathbf{A}) \sum_{k=0}^{\infty} \frac{\mathbf{A}^k}{z^k}$$

Then

$$R(z) = (z\mathbf{I} - \mathbf{A})^{-1} = z^{-1} \sum_{k=0}^{\infty} \frac{\mathbf{A}^k}{z^k} = \sum_{k=0}^{\infty} \frac{\mathbf{A}^k}{z^{k+1}}$$

□

Corollary 2.14. *If $|z| > ||\boldsymbol{A}||$, then*

$$\lim_{|z| \to \infty} ||R(z)|| = 0$$

Proof.

$$||R(z)|| = \left|\left| \sum_{k=0}^{\infty} \frac{\mathbf{A}^k}{z^{k+1}} \right|\right| \leq \sum_{k=0}^{\infty} \frac{||\mathbf{A}^k||}{|z^{k+1}|} = z^{-1} \sum_{k=0}^{\infty} \frac{||\mathbf{A}^k||}{|z^k|} \leq z^{-1} \sum_{k=0}^{\infty} \left(\frac{||\mathbf{A}||}{|z|} \right)^k = \frac{1}{|z| - ||\mathbf{A}||}$$

Taking the limit as $z \to \infty$ finishes the proof. □

Note that Corollary 2.14, together with the facts that $R(z)$ and the norm $||\cdot||$ are continuous, lead to the conclusion that the ε-pseudospectra $\sigma_\varepsilon(\mathbf{A})$ are bounded open sets.

Lemma 2.15. *Let Γ be a positively oriented Jordan Curve with 0 in its interior. For every $k \in \mathbb{N}$,*

$$\frac{1}{2\pi i}\int_\Gamma \frac{1}{z^{j-k+1}}\,dz = \begin{cases} 1 & \text{if } k = j \\ 0 & \text{if } k \neq j \end{cases} \qquad \text{for every } j \in \mathbb{R}$$

Proof. See Saff and Snider [5, p. 166] □

Theorem 2.16. *Given a positively oriented Jordan curve Γ which contains the ball centered at the origin of radius $\rho(\boldsymbol{A})$ in its interior, we have*

$$\boldsymbol{A}^k = \frac{1}{2\pi i}\int_\Gamma z^k R(z)\,dz$$

for every $k \in \mathbb{N}$.

Proof.

$$\begin{aligned} \frac{1}{2\pi i}\int_\Gamma z^k R(z)\,dz &= \frac{1}{2\pi i}\int_\Gamma z^k (z\mathbf{I} - \mathbf{A})^{-1}\,dz \\ &= \frac{1}{2\pi i}\int_\Gamma z^k \left(\sum_{j=0}^{\infty}\frac{\mathbf{A}^j}{z^{j+1}}\right)dz, \qquad \text{(by Theorem 2.13)} \\ &= \frac{1}{2\pi i}\int_\Gamma \sum_{j=0}^{\infty}\frac{\mathbf{A}^j}{z^{j-k+1}}\,dz \end{aligned}$$

Note that $R(z)$ is continuous on its domain and Γ is compact, which implies that $z^k R(z)$ is uniformly continuous on Γ; therefore, we are justified in switching the order of the sum and integral.

$$\begin{aligned} &= \frac{1}{2\pi i}\sum_{j=0}^{\infty}\mathbf{A}^j\int_\Gamma \frac{1}{z^{j-k+1}}\,dz \\ &= \mathbf{A}^k \qquad \text{(by Lemma 2.15)} \end{aligned}$$

□

Chapter 3. Behavior of $\left|\left|\mathbf{A}^k\right|\right|$ for $N \times N$ Matrices

Now that we have different ways of expressing the matrix power $\mathbf{A}^k$, it is natural to ask whether these different expressions can improve analysis of the norm, $||\mathbf{A}^k||$. In fact, we will see that the expressions of the last chapter can be used along with the spectral radius to determine bounds on norms; and in the case of the 2×2 matrix, an exact formula will be achieved.

Perhaps the most useful parameter in analyzing matrix norms quickly, is the spectral radius. Once we know the spectral radius $\rho(\mathbf{A})$ of a matrix $\mathbf{A}$, we instantly know the end behavior of norms such as $||\mathbf{A}^k||_2$ as $k \to \infty$. Theorem 3.1 will show that if $\rho(\mathbf{A}) > 1$, then $||\mathbf{A}^k||$ is guaranteed to diverge to ∞ as $k \to \infty$. In contrast, Corollary 3.3 will imply that if $\rho(\mathbf{A}) < 1$, then $||\mathbf{A}^k||$ converges to 0 as $k \to \infty$; consequently, it is also bounded above for all k.

Theorem 3.1. *For every nonnegative integer k,*

$$\rho(\boldsymbol{A})^k \leq ||\boldsymbol{A}^k||,$$

where $\rho(\boldsymbol{A})$ is the spectral radius of the $N \times N$ matrix $\boldsymbol{A}$, and $||\cdot||$ is any matrix norm satisfying the sub-multiplicative property.

Proof. Let λ be an eigenvalue of $\mathbf{A}$ with corresponding eigenvector $\mathbf{x}$. Then

$$|\lambda^k| \, ||\mathbf{x}|| = ||\lambda^k \mathbf{x}|| = ||\mathbf{A}^k \mathbf{x}|| \leq ||\mathbf{A}^k|| \, ||\mathbf{x}||$$

We can divide both sides by $||\mathbf{x}||$, which is nonzero since $\mathbf{x}$ is an eigenvector. Since λ is an arbitrary eigenvalue of $\mathbf{A}$, the relationship holds for $\rho(\mathbf{A})$. □

Corollary 3.2. *If $\boldsymbol{A}$ is normal, then for every nonnegative integer k,*

$$\rho(\boldsymbol{A})^k = ||\boldsymbol{A}^k||_2$$

Proof. By Theorem 1.4, we can write $\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{U}^*$, where $\mathbf{U}$ is unitary and $\mathbf{D}$ is a diagonal matrix with the eigenvalues of $\mathbf{A}$ along its diagonal. Then,

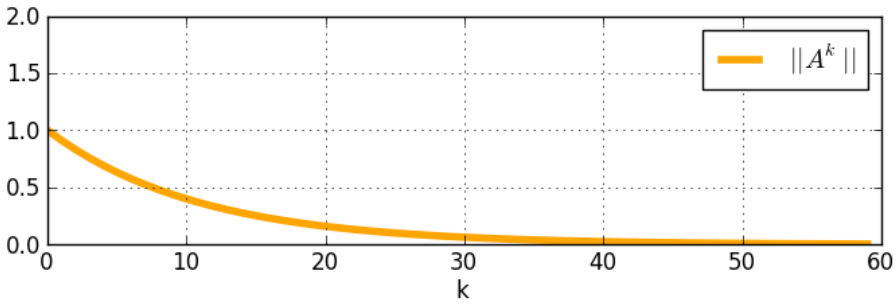
$$||\mathbf{A}^k||_2 = ||(\mathbf{UDU}^*)^k||_2 = ||\mathbf{UD}^k\mathbf{U}^*||_2 = \rho(\mathbf{UD}^k\mathbf{U}^*) = \rho(\mathbf{D}^k) = [\rho(\mathbf{D})]^k = [\rho(\mathbf{A})]^k$$

□

Corollary 3.2 is extremely useful in analyzing norms of normal matrices; it reduces a complex computation involving matrix norms and powers, into that of taking the power of a scalar value. Also, while normal matrices can take many diverse forms, all those with spectral radius less than one will display a graph similar to that of the symmetric (and normal) matrix $\mathbf{A}$ in Figure 3.1, with spectral radius $\rho(\mathbf{A}) \approx .91$:

Figure 3.1: The Norm of Powers of a Normal Matrix

$$\mathbf{A} = \begin{bmatrix} \frac{1}{2} & 0 & \frac{1}{10} & 0 \\ 0 & \frac{1}{2} & \frac{2}{5} & 0 \\ \frac{1}{10} & \frac{2}{5} & \frac{1}{2} & 0 \\ 0 & 0 & 0 & \frac{9}{10} \end{bmatrix}$$



Theorem 3.3. *Given any matrix $\boldsymbol{A}$ in M_N with spectral radius $\rho(\boldsymbol{A})$,*

$$\lim_{k\to\infty} \boldsymbol{A}^k = \boldsymbol{0} \qquad \textit{if and only if} \qquad \rho(\boldsymbol{A}) < 1$$

Proof. Suppose that $\lim_{k\to\infty} \mathbf{A}^k = \mathbf{0}$. Then for each eigenvalue λ and corresponding eigenvector $\mathbf{x}$ of $\mathbf{A}$,

$$\lim_{k\to\infty} \lambda^k\mathbf{x} = \lim_{k\to\infty} \mathbf{A}^k\mathbf{x} = \mathbf{0}$$

Since each eigenvector is nonzero, this implies that $\lim_{k\to\infty} \lambda^k = 0$, for each eigenvalue λ. Thus, $|\lambda| < 1$ for each eigenvalue λ. Therefore, $\rho(\mathbf{A}) < 1$.

Now suppose that $\rho(\mathbf{A}) < 1$. Let $\mathbf{A} = \mathbf{UTU}^*$ be the Schur decomposition of $\mathbf{A}$. Using the

same notation as in Theorem 2.4, if $m > n$, then $t_{mn}^{(k)} = 0$. If $m = n$, then

$$0 \leq \lim_{k\to\infty} |t_{mn}^{(k)}| = \lim_{k\to\infty} |\lambda_m|^k \leq \lim_{k\to\infty} |\rho(\mathbf{A})|^k = 0$$

Finally, if $m < n$, then

$$\begin{aligned}\lim_{k\to\infty} |t_{mn}^{(k)}| &= \lim_{k\to\infty} \left| \sum_{j=1}^{n-m} \sum_{\substack{m=\alpha_1<\alpha_2 \\ <\ldots<\alpha_{j+1}=n}} \left(\prod_{i=1}^{j} t_{\alpha_i\alpha_{i+1}} \right) \sum_{|\Delta_{j+1}|=k-j} \prod_{l=1}^{j+1} \lambda_{\alpha_l}^{\delta_l} \right| \\ &= \lim_{k\to\infty} \sum_{j=1}^{n-m} \sum_{\substack{m=\alpha_1<\alpha_2 \\ <\ldots<\alpha_{j+1}=n}} \left| \prod_{i=1}^{j} t_{\alpha_i\alpha_{i+1}} \right| \left| \sum_{|\Delta_{j+1}|=k-j} \prod_{l=1}^{j+1} \lambda_{\alpha_l}^{\delta_l} \right|\end{aligned}$$

Apply Lemma 2.11 and Lemma 2.12 to get:

$$\leq \lim_{k\to\infty} \sum_{j=1}^{n-m} \sum_{\substack{m=\alpha_1<\alpha_2 \\ <\ldots<\alpha_{j+1}=n}} \left| \prod_{i=1}^{j} t_{\alpha_i\alpha_{i+1}} \right| \binom{k}{j} \rho(\mathbf{T})^{k-j} = \sum_{j=1}^{n-m} \sum_{\substack{m=\alpha_1<\alpha_2 \\ <\ldots<\alpha_{j+1}=n}} \left| \prod_{i=1}^{j} t_{\alpha_i\alpha_{i+1}} \right| \left[\lim_{k\to\infty} \binom{k}{j} \rho(\mathbf{T})^{k-j} \right]$$

$$= 0$$

Since all of the entries of $\mathbf{T}^k$ approach 0 as $k \to \infty$, then $\lim_{k\to\infty} \mathbf{T}^k = 0$. Then $\lim_{k\to\infty} \mathbf{A}^k = 0$.

□

For an alternative proof of Theorem 3.3, see Horn and Johnson [2, p. 298]. The continuity of the norm leads directly to the following corollary.

Corollary 3.4. *Given any matrix $\boldsymbol{A} \in M_N$ with spectral radius $\rho(\boldsymbol{A})$,*

$$\lim_{k\to\infty} ||\boldsymbol{A}^k|| = 0 \qquad \textit{if and only if} \qquad \rho(\boldsymbol{A}) < 1$$

Regardless of whether $||\mathbf{A}^k||$ is bounded by a constant, the following theorem shows that its growth is at most exponential.

Theorem 3.5. *Given $\boldsymbol{A} \in M_N$, there exists $\gamma > 0$ and $M \geq 1$, such that for each nonnegative integer k,*

$$||\boldsymbol{A}^k|| \leq M\gamma^k$$

Proof. Choose $\varepsilon > 0$. Let $\tilde{\mathbf{A}} := (\rho(\mathbf{A}) + \varepsilon)^{-1}\mathbf{A}$. Observe that,

$$\rho(\tilde{\mathbf{A}}) = \frac{\rho(\mathbf{A})}{\rho(\mathbf{A}) + \varepsilon} < 1$$

By Corrolary 3.4 we have that $\lim_{k\to\infty} ||\tilde{\mathbf{A}}^k|| = 0$. Thus, there exists $N \geq 0$ such that for every $k \geq N$ we have $||\tilde{\mathbf{A}}^k|| < 1$. Then for $k \geq N$,

$$(\rho(\mathbf{A}) + \varepsilon)^{-k}||\mathbf{A}^k|| = ||(\rho(\mathbf{A}) + \varepsilon)^{-k}\mathbf{A}^k|| = ||\tilde{\mathbf{A}}^k|| < 1$$

$$||\mathbf{A}^k|| < (\rho(\mathbf{A}) + \varepsilon)^k \tag{3.1}$$

Let $\gamma = \rho(\mathbf{A}) + \varepsilon$, and let $M = \max_{1\leq k\leq N}\{1, \gamma^{-k}||\mathbf{A}^k||\}$. Then $M\gamma^k \geq ||\mathbf{A}^k||$ for every k, and we are done. □

Corollary 3.6.

$$\rho(\boldsymbol{A}) = \lim_{k\to\infty} ||\boldsymbol{A}^k||^{1/k}$$

Proof. From Theorem 3.1 and equation (3.1), given $\varepsilon > 0$, there exists $N \in \mathbb{N}$, such that for every $k \geq N$,

$$\rho(\mathbf{A})^k \leq ||\mathbf{A}^k||_2 \leq (\rho(\mathbf{A}) + \varepsilon)^k$$

Then for $k \geq N$,

$$\rho(\mathbf{A}) \leq ||\mathbf{A}^k||_2^{1/k} \leq (\rho(\mathbf{A}) + \varepsilon)$$

In fact, this is true for all $\varepsilon > 0$, thus

$$\lim_{k\to\infty} ||\mathbf{A}^k||_2^{1/k} = \rho(\mathbf{A})$$

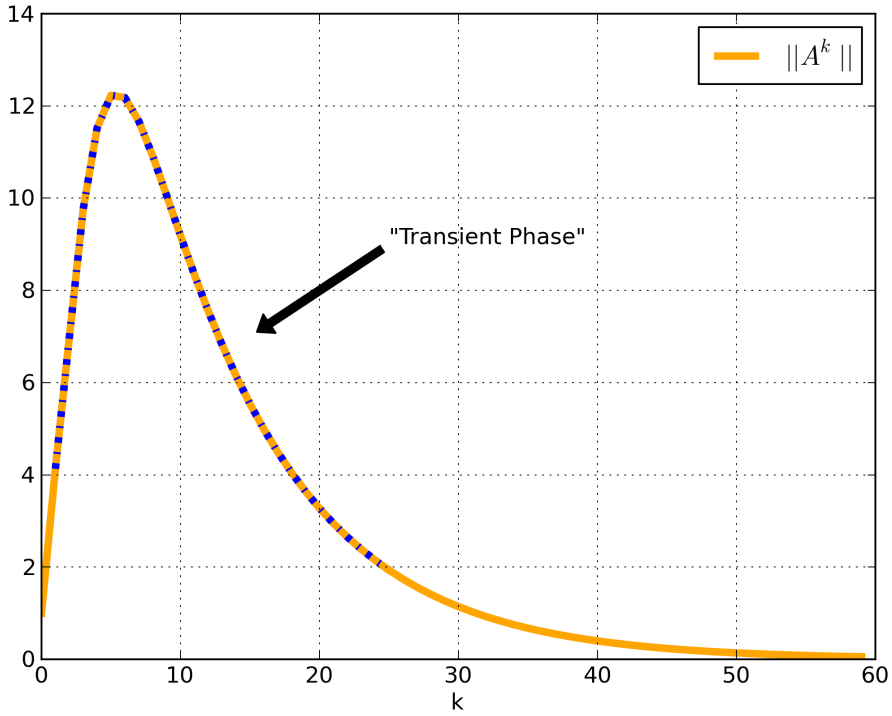
□

For a nonnormal matrix $\mathbf{A}$ with spectral radius $\rho(\mathbf{A}) < 1$, the shape of the graph of $||\mathbf{A}^k||_2$ may take the same form as seen in Figure 3.1 on page 43, or it may display a prominent ***transient phase***; in fact, the latter is more typical, as seen in the upper-triangular (and nonnormal) example

of Figure 3.2.

Figure 3.2: The Norm of Powers of a Nonnormal Matrix

$$\mathbf{A} = \begin{bmatrix} \frac{1}{2} & 0 & 1 & 0 \\ 0 & \frac{1}{2} & 0 & 0 \\ 0 & 0 & \frac{1}{2} & 4 \\ 0 & 0 & 0 & \frac{9}{10} \end{bmatrix}$$



While the peak of the transient phase for the 4×4 matrix above is quite manageable, in cases with larger dimensions, or in cases with larger off-diagonal entries, the transient phase can grow very large before peaking. Those interested in taking norms of matrix powers are presented with two major problems regarding computation: 1) larger numbers require more time for computation, and 2) if numbers get too large, a computer will no longer be able to store them in memory. If the answer that we desire requires iteration beyond the transient effect, then we may wish to have bounds that can tell us how large the norm may grow, before actually running a mathematical model involving the norm.

3.1 Bounds on $||\mathbf{T}^k||$

Now we look at bounds on the norm $||\mathbf{T}^k||$, where as before, $\mathbf{T}$ is an upper-triangular matrix, and k is a nonnegative integer. First, we look at the 2×2 case, where we are in fact able to develop an exact expression for the norm.

Theorem 3.7. *Given the* 2×2 *upper triangular matrix* $\boldsymbol{T} = \begin{pmatrix} \lambda_1 & t \\ 0 & \lambda_2 \end{pmatrix}$,

$$||\boldsymbol{T}^k||_2 = \left[\frac{1}{2}\left(b + \sqrt{b^2 - 4|\lambda_1\lambda_2|^{2k}}\right)\right]^{\frac{1}{2}} \qquad \textit{where } b = |\lambda_1|^{2k} + |\lambda_2|^{2k} + |t|^2\left|\sum_{j=1}^{k-1}\lambda_1^j\lambda_2^{k-1-j}\right|^2$$

Proof. By Theorem 2.4,

$$\mathbf{T}^k = \begin{bmatrix} \lambda_1^k & t\sum\limits_{\Delta_2 = k-1}\lambda_1^{\delta_1}\lambda_2^{\delta_2} \\ 0 & \lambda_2^k \end{bmatrix} = \begin{bmatrix} \lambda_1^k & t\sum\limits_{j=0}^{k-1}\lambda_1^j\lambda_2^{k-1-j} \\ 0 & \lambda_2^k \end{bmatrix}$$

Then,

$$\left(\mathbf{T}^k\right)^*\mathbf{T}^k = \begin{bmatrix} |\lambda_1|^{2k} & \overline{\lambda_1^k}\left(t\sum\limits_{j=0}^{k-1}\lambda_1^j\lambda_2^{k-1-j}\right) \\ \overline{\left(t\sum\limits_{j=0}^{k-1}\lambda_1^j\lambda_2^{k-1-j}\right)}\lambda_1^k & |\lambda_2|^{2k} + |t|^2\left|\sum\limits_{j=0}^{k-1}\lambda_1^j\lambda_2^{k-1-j}\right|^2 \end{bmatrix}$$

The characteristic polynomial of $\left(\mathbf{T}^k\right)^*\mathbf{T}^k$ is:

$$\begin{aligned} \mathrm{p}(\xi) &= \left(\xi - |\lambda_1|^{2k}\right)\left(\xi - \left(|\lambda_2|^{2k} + |t|^2\left|\sum_{j=0}^{k-1}\lambda_1^j\lambda_2^{k-1-j}\right|^2\right)\right) - |\lambda_1|^{2k}|t|^2\left|\sum_{j=0}^{k-1}\lambda_1^j\lambda_2^{k-1-j}\right|^2 \\ &= \xi^2 - \left(|\lambda_1|^{2k} + |\lambda_2|^{2k} + |t|^2\left|\sum_{j=0}^{k-1}\lambda_1^j\lambda_2^{k-1-j}\right|^2\right)\xi + |\lambda_1\lambda_2|^{2k} \\ &\quad + |\lambda_1|^{2k}|t|^2\left|\sum_{j=0}^{k-1}\lambda_1^j\lambda_2^{k-1-j}\right|^2 - |\lambda_1|^{2k}|t|^2\left|\sum_{j=0}^{k-1}\lambda_1^j\lambda_2^{k-1-j}\right|^2 \\ &= \xi^2 - \left(|\lambda_1|^{2k} + |\lambda_2|^{2k} + |t|^2\left|\sum_{j=0}^{k-1}\lambda_1^j\lambda_2^{k-1-j}\right|^2\right)\xi + |\lambda_1\lambda_2|^{2k} \end{aligned}$$

By the quadratic formula, the largest root is:

$$\xi_{max} = \frac{1}{2}\left[|\lambda_1|^{2k} + |\lambda_2|^{2k} + |t|^2 \left| \sum_{j=0}^{k-1} \lambda_1^j \lambda_2^{k-1-j} \right|^2 + \sqrt{\left(|\lambda_1|^{2k} + |\lambda_2|^{2k} + |t|^2 \left| \sum_{j=0}^{k-1} \lambda_1^j \lambda_2^{k-1-j} \right|^2 \right)^2 - 4|\lambda_1\lambda_2|^{2k}} \right]$$

Note that the arithmetic-geometric mean inequality ensures that the expression under the square root is nonnegative, since:

$$\frac{|\lambda_1|^{2k} + |\lambda_2|^{2k}}{2} \geq \sqrt{|\lambda_1\lambda_2|^{2k}} \qquad \text{implies that} \qquad \left(|\lambda_1|^{2k} + |\lambda_2|^{2k}\right)^2 \geq 4|\lambda_1\lambda_2|^{2k}$$

Observe that ξ_{max} is real, and it is the largest eigenvalue of $\left(\mathbf{T}^k\right)^* \mathbf{T}^k$; thus, $\sqrt{\xi_{max}}$ is the largest singular value of $\mathbf{T}^k$. By Lemma 1.13, we have $||\mathbf{T}^k||_2 = \sqrt{\xi_{max}}$; thus, taking the square root of the right-hand side above finishes the proof. □

Unfortunately, this method for finding $||\mathbf{T}^k||_2$ is not practical for matrices of dimension 3 and higher, as it relies on the quadratic formula to find the largest root of the characteristic polynomial. For higher dimensions, we will seek to bound the norm, rather than writing out an explicit formula for its exact value. The next theorem gives a lower bound for $||\mathbf{T}^k||_2$, where $\mathbf{T}$ is of arbitrary dimension. In fact, one can replace $||\cdot||_2$ by $||\cdot||_1$, $||\cdot||_\infty$, or $||\cdot||_F$ and the theorem below still holds.

Theorem 3.8. *Let $\boldsymbol{T} \in M_N$ be an upper-triangular matrix, and let k be a nonnegative integer. Let $\boldsymbol{T}$ be written entry-wise in the notation of Theorem 2.4. Then for $1 \leq m < n \leq N$,*

$$||\boldsymbol{T}^k||_2 \geq |t_{mn}^{(k)}| = \left| \sum_{j=1}^{n-m} \sum_{\substack{m=\alpha_1<\alpha_2 \\ <\ldots<\alpha_{j+1}=n}} \left(\prod_{i=1}^{j} t_{\alpha_i\alpha_{i+1}} \right) \sum_{|\Delta_{j+1}|=k-j} \prod_{l=1}^{j+1} \lambda_{\alpha_l}^{\delta_l} \right|$$

Proof. First, note that give any vector $\mathbf{x} = (x_1, x_2, \ldots, x_N)^T$, we have $|x_i|^2 \leq \sum_{j=1}^{N} |x_j|^2 = ||\mathbf{x}||_2^2$, for each $i = 1, 2, \ldots, N$. Let $\mathbf{e}^{(n)} = (e_1, e_2, \ldots, e_N)^T$ be the vector with 1 in the n^{th} coordinate,

and zeros otherwise. Fix m and n, with $1 \leq m < n \leq N$. Then,

$$|t_{mn}^{(k)}|^2 = \sum_{q=1}^{N} |t_{mq}^{(k)}|^2 e_q^2 = ||\mathbf{T}^k \mathbf{e}^{(n)}||_2^2 \leq ||\mathbf{T}^k||_2^2$$

Substituting the value for $t_{mn}^{(k)}$ given by Theorem 2.4, and taking the square root finishes the proof. □

In general, using the absolute value of one matrix entry to bound the norm of the matrix, provides a bound that is likely not very sharp; however, details about the structure of $\mathbf{T}^k$ provided by Theorem 2.4, can guide us to select a matrix entry for which Theorem 3.8 gives a surprisingly good bound. In chapter 4 we will see several examples where this is the case.

In order to develop and simplify upper bounds for $||\mathbf{T}^k||$, we will need results from combinatorics. Lemmas 3.9, 3.10, and 3.11 provide ways for working with combinations which will be useful in the proof of the next theorem.

Lemma 3.9. *Given positive integers n and m with $n > m$, then for each j with $1 \leq j \leq n - m$ there are $\binom{n-m-1}{j-1}$ distinct tuples $(\alpha_1, \alpha_2, \ldots, \alpha_{j+1})$, where*
$m = \alpha_1 < \alpha_2 < \ldots < \alpha_{j+1} = n$.

Proof. Since α_1 and α_{j+1} are fixed, there are $j - 1$ terms which must be selected in order from least to greatest. There are $n - m - 1$ terms between n and m (noninclusive). Since only one ordering of the terms is valid, this is the same as selecting $j - 1$ objects from $n - m - 1$ objects without regard to order; that is, it is the number of combinations of $n - m - 1$ objects, taken $j - 1$ at a time. □

Lemma 3.10 (Pascal's Identity)**.** *Given an integer N with $N \geq 1$, and given an integer q with $1 \leq q \leq N$,*

$$\binom{N}{q} = \binom{N-1}{q} + \binom{N-1}{q-1}$$

Proof.

$$\begin{aligned}\binom{N}{q} &= \frac{N!}{q!(N-q)!} = \frac{N! - q(N-1)! + q(N-1)!}{q!(N-q)!} = \frac{N! - q(N-1)!}{q!(N-q)!} + \frac{q(N-1)!}{q!(N-q)!} \\ &= \frac{N(N-1)! - q(N-1)!}{q!(N-q)!} + \frac{q(N-1)!}{q(q-1)!(N-q)!} = \frac{(N-q)(N-1)!}{q!(N-q)!} + \frac{(N-1)!}{(q-1)!(N-q)!}\end{aligned}$$

$$= \frac{(N-1)!}{q!(N-1-q)!} + \frac{(N-1)!}{(q-1)!(N-1-(q-1))!} = \binom{N-1}{q} + \binom{N-1}{q-1}$$

□

Lemma 3.11. *Given an integer N with $N \geq 1$, for each integer q with $1 \leq q \leq N$,*

$$\sum_{s=q}^{N} \binom{s-1}{q-1} = \binom{N}{q}$$

Proof. We proceed by induction. If $N = 1$, then we must have $q = 1$.

$$\sum_{s=q}^{N} \binom{s-1}{q-1} = \sum_{s=1}^{1} \binom{s-1}{0} = \binom{0}{0} = 1 = \binom{1}{1} = \binom{N}{q}$$

Now suppose that the lemma holds for some integer $N \geq 1$, and choose some q with $1 \leq q \leq N$. Then,

$$\begin{aligned}
\sum_{s=q}^{N+1} \binom{s-1}{q-1} &= \sum_{s=q}^{N} \binom{s-1}{q-1} + \binom{N}{q-1} \\
&= \binom{N}{q} + \binom{N}{q-1} && \text{(by the inductive step)} \\
&= \binom{N+1}{q} && \text{(by Lemma 3.10)}
\end{aligned}$$

□

We are now ready to derive an upper-bound.

Theorem 3.12. *Given a nonzero $N \times N$ upper-triangular matrix $\boldsymbol{T}$ with*

$$\rho(\boldsymbol{T}) = \max(|\lambda| : \lambda \in \sigma(\boldsymbol{T})) \qquad \text{and} \qquad M_{\boldsymbol{T}} = \max(|t_{ij}| : i < j)$$

for each $k = 0, 1, 2, \ldots$, we have

$$||\boldsymbol{T}^k||_2 \leq \sqrt{\sum_{l=0}^{N-1} \left[\sum_{j=0}^{l} \binom{l}{j} \binom{k}{j} M_{\boldsymbol{T}}^j \rho(\boldsymbol{T})^{k-j} \right]^2}$$

where we take $\binom{k}{j} = 0$ *for* $j > k$.

Proof. Given $\mathbf{x}$ with $||\mathbf{x}|| = 1$, we have

$$\begin{aligned}
||\mathbf{T}^k\mathbf{x}||_2^2 &= \sum_{m=1}^{N}\left|\sum_{r=1}^{N} t_{mr}^{(k)}x_r\right|^2 \\
&= \sum_{m=1}^{N}\left|\sum_{r=m}^{N} t_{mr}^{(k)}x_r\right|^2 && \text{(since } t_{mr}^{(k)} = 0 \text{ for } r < m\text{)} \\
&= \sum_{m=1}^{N}\left|\lambda_m^k x_m + \sum_{r=m+1}^{N} t_{mr}^{(k)}x_r\right|^2 && \text{(since } t_{mr}^{(k)} = \lambda_m^k \text{ for } r = m\text{)}
\end{aligned}$$

Now, using Theorem 2.4 we substitute for $t_{mr}^{(k)}$ in the second summation.

$$\begin{aligned}
||\mathbf{T}^k\mathbf{x}||_2^2 &= \sum_{m=1}^{N}\left|\lambda_m^k x_m + \sum_{r=m+1}^{N}\left(\sum_{j=1}^{r-m}\sum_{\substack{m=\alpha_1<\alpha_2\\<\ldots<\alpha_{j+1}=r}}\left(\prod_{i=1}^{j} t_{\alpha_i\alpha_{i+1}}\right)\sum_{|\Delta_{j+1}|=k-j}\prod_{l=1}^{j+1}\lambda_{\alpha_l}^{\delta_l}\right)x_r\right|^2 \\
&\leq \sum_{m=1}^{N}\left(|\lambda_m^k||x_m| + \sum_{r=m+1}^{N}\left(\sum_{j=1}^{r-m}\sum_{\substack{m=\alpha_1<\alpha_2\\<\ldots<\alpha_{j+1}=r}}\left(\prod_{i=1}^{j}|t_{\alpha_i\alpha_{i+1}}|\right)\sum_{|\Delta_{j+1}|=k-j}\prod_{l=1}^{j+1}|\lambda_{\alpha_l}^{\delta_l}|\right)|x_r|\right)^2
\end{aligned}$$

Observe that $|x_m|, |x_r| \leq 1$ for each r, so dropping these terms on the right will only increase the value of the overall sum.

$$||\mathbf{T}^k\mathbf{x}||_2^2 \leq \sum_{m=1}^{N}\left(|\lambda_m^k| + \sum_{r=m+1}^{N}\left(\sum_{j=1}^{r-m}\sum_{\substack{m=\alpha_1<\alpha_2\\<\ldots<\alpha_{j+1}=r}}\left(\prod_{i=1}^{j}|t_{\alpha_i\alpha_{i+1}}|\right)\sum_{|\Delta_{j+1}|=k-j}\prod_{l=1}^{j+1}|\lambda_{\alpha_l}^{\delta_l}|\right)\right)^2$$

Since $\mathbf{x}$ was arbitrary, and since $||\mathbf{T}^k||_2 = \sup_{||\mathbf{x}||=1}||\mathbf{T}^k\mathbf{x}||_2$, we have that,

$$||\mathbf{T}^k||_2^2 \leq \sum_{m=1}^{N}\left(|\lambda_m^k| + \sum_{r=m+1}^{N}\left(\sum_{j=1}^{r-m}\sum_{\substack{m=\alpha_1<\alpha_2\\<\ldots<\alpha_{j+1}=r}}\left(\prod_{i=1}^{j}|t_{\alpha_i\alpha_{i+1}}|\right)\sum_{|\Delta_{j+1}|=k-j}\prod_{l=1}^{j+1}|\lambda_{\alpha_l}^{\delta_l}|\right)\right)^2$$

By Lemma 2.11 we may substitute the expression $\binom{k}{j}\rho(\mathbf{T})^{k-j}$ for the last summation in the

inequality above; we may also replace $|t_{\alpha_i\alpha_{i+1}}|$ by $M_{\mathbf{T}}$, to get:

$$||\mathbf{T}^k||_2^2 \leq \sum_{m=1}^{N}\left(\rho(\mathbf{T})^k + \sum_{r=m+1}^{N}\left(\sum_{j=1}^{r-m}\sum_{\substack{m=\alpha_1<\alpha_2\\<\ldots<\alpha_{j+1}=r}}\left(\prod_{i=1}^{j}M_{\mathbf{T}}\right)\binom{k}{j}\rho(\mathbf{T})^{k-j}\right)\right)^2$$

Since $\prod_{i=1}^{j} M_{\mathbf{T}} = M_{\mathbf{T}}^j$, we can make this substitution and rearrange terms to get:

$$= \sum_{m=1}^{N}\left(\rho(\mathbf{T})^k + \sum_{r=m+1}^{N}\left(\sum_{j=1}^{r-m}M_{\mathbf{T}}^j\binom{k}{j}\rho(\mathbf{T})^{k-j}\sum_{\substack{m=\alpha_1<\alpha_2\\<\ldots<\alpha_{j+1}=r}}1\right)\right)^2$$

By Lemma 3.9 this is equal to,

$$\sum_{m=1}^{N}\left(\rho(\mathbf{T})^k + \sum_{r=m+1}^{N}\left(\sum_{j=1}^{r-m}M_{\mathbf{T}}^j\binom{k}{j}\rho(\mathbf{T})^{k-j}\binom{r-m-1}{j-1}\right)\right)^2$$

Now we expand the summation $\sum_{r=m+1}^{N}$ to get:

$$= \sum_{m=1}^{N}\left[\rho(\mathbf{T})^k + \left(\sum_{j=1}^{1}M_{\mathbf{T}}^j\binom{k}{j}\rho(\mathbf{T})^{k-j}\binom{0}{j-1}\right) + \left(\sum_{j=1}^{2}M_{\mathbf{T}}^j\binom{k}{j}\rho(\mathbf{T})^{k-j}\binom{1}{j-1}\right)\right.$$
$$\left. + \ldots + \left(\sum_{j=1}^{N-m}M_{\mathbf{T}}^j\binom{k}{j}\rho(\mathbf{T})^{k-j}\binom{N-m-1}{j-1}\right)\right]^2$$

Rearranging terms we get,

$$\sum_{m=1}^{N}\left[\rho(\mathbf{T})^k + \binom{k}{1}M_{\mathbf{T}}\,\rho(\mathbf{T})^{k-1}\left[\binom{0}{0} + \binom{1}{0} + \ldots + \binom{N-m-1}{0}\right]\right.$$
$$+ \binom{k}{2}M_{\mathbf{T}}^2\,\rho(\mathbf{T})^{k-2}\left[\binom{1}{1} + \binom{2}{1} + \ldots + \binom{N-m-1}{1}\right]$$
$$\left. + \ldots + \binom{k}{N-m}M_{\mathbf{T}}^{N-m}\,\rho(\mathbf{T})^{k-(N-m)}\left[\binom{N-m-1}{N-m-1}\right]\right]^2$$

By Lemma 3.11 this is equal to,

$$\sum_{m=1}^{N}\left[\rho(\mathbf{T})^k + \binom{k}{1}M_{\mathbf{T}}\,\rho(\mathbf{T})^{k-1}\binom{N-m}{1} + \binom{k}{2}M_{\mathbf{T}}^2\,\rho(\mathbf{T})^{k-2}\binom{N-m}{2}\right.$$
$$\left. + \ldots + \binom{k}{N-m}M_{\mathbf{T}}^{N-m}\,\rho(\mathbf{T})^{k-(N-m)}\left[\binom{N-m}{N-m}\right]\right]^2$$

Now since $\rho(\mathbf{T})^k = \binom{N-m}{0}\binom{k}{0}M_{\mathbf{T}}^0\,\rho(\mathbf{T})^k$, we have,

$$||\mathbf{T}^k||_2^2 \leq \sum_{m=1}^{N}\left[\sum_{j=0}^{N-m}\binom{N-m}{j}\binom{k}{j}M_{\mathbf{T}}^j\,\rho(\mathbf{T})^{k-j}\right]^2$$

Let $l = N - m$ and note that as m ranges from 1 to N, l ranges from $N-1$ to 0. If we sum the terms above backwards over l, we get:

$$||\mathbf{T}^k||_2^2 \leq \sum_{l=0}^{N-1}\left[\sum_{j=0}^{l}\binom{l}{j}\binom{k}{j}M_{\mathbf{T}}^j\,\rho(\mathbf{T})^{k-j}\right]^2$$

Taking the square root finishes the proof. □

The bound from Theorem 3.12 is not sharp, as many concessions were made in whittling down the initial expression for the norm to the more simplified statement of the theorem. However, similar logic may be applied to other norms to derive their respective bounds. In the next theorem, we present a bound for the Frobenius norm.

Theorem 3.13. *Given a nonzero upper-triangular matrix* $\boldsymbol{T}$ *with,*

$$\rho(\boldsymbol{T}) = \max(|\lambda| : \lambda \in \sigma(\boldsymbol{T})) \qquad \textit{and} \qquad M_{\boldsymbol{T}} = \max(|t_{ij}| : i < j),$$

for each $k = 0, 1, 2, \ldots,$

$$||\boldsymbol{T}^k||_F \leq \sqrt{N\rho(\boldsymbol{A})^{2k} + \sum_{\substack{m,n=1\\ m<n}}^{N}\left|\sum_{j=1}^{n-m}\binom{n-m-1}{j-1}\binom{k}{j}M_{\boldsymbol{T}}^j\,\rho(\boldsymbol{T})^{k-j}\right|^2}$$

where $\binom{k}{j} = 0$ *for* $j > k$.

Note that since $||\mathbf{A}||_2 \leq ||\mathbf{A}||_F$, for any matrix $\mathbf{A}$, the above bound also works for the 2-norm.

Proof. Again, we use the notation of Theorem 2.4.

$$\begin{aligned}
||\mathbf{T}^k||_F^2 &= \sum_{m,n=1}^{N} \left|t_{mn}^{(k)}\right|^2 = \sum_{\substack{m,n=1 \\ m\leq n}}^{N} \left|t_{mn}^{(k)}\right|^2 \qquad \text{(since } t_{mn}^{(k)} = 0 \text{ if } m > n\text{)} \\
&= \sum_{m=1}^{N} |\lambda_m|^{2k} + \sum_{\substack{m,n=1 \\ m<n}}^{N} \left| \sum_{j=1}^{n-m} \sum_{\substack{m=\alpha_1<\alpha_2 \\ <\ldots<\alpha_{j+1}=n}} \left(\prod_{i=1}^{j} t_{\alpha_i\alpha_{i+1}} \right) \sum_{|\Delta_{j+1}|=k-j} \prod_{l=1}^{j+1} \lambda_{\alpha_l}^{\delta_l} \right|^2 \\
&\leq N\rho(\mathbf{T})^{2k} + \sum_{\substack{m,n=1 \\ m<n}}^{N} \left| \sum_{j=1}^{n-m} \sum_{\substack{m=\alpha_1<\alpha_2 \\ <\ldots<\alpha_{j+1}=n}} M_{\mathbf{T}}^j \sum_{|\Delta_{j+1}|=k-j} \rho(\mathbf{T})^{k-j} \right|^2 \\
&= N\rho(\mathbf{T})^{2k} + \sum_{\substack{m,n=1 \\ m<n}}^{N} \left| \sum_{j=1}^{n-m} \binom{n-m-1}{j-1} M_{\mathbf{T}}^j \binom{k}{j} \rho(\mathbf{T})^{k-j} \right|^2 \qquad \text{(by Lemmas 2.7 and 3.9)}
\end{aligned}$$

The proof is completed by rearranging terms and taking the square root of both sides. □

Now we look at a similarly derived bound for the infinity norm.

Theorem 3.14.

$$||\boldsymbol{T}^k||_\infty \leq \sum_{j=0}^{N-1} \binom{N-1}{j} \binom{k}{j} M_{\boldsymbol{T}}^j \rho(\boldsymbol{T})^{k-j}$$

Proof.

$$\begin{aligned}
||\mathbf{T}^k||_\infty &= \max_{1\leq m\leq N} \sum_{n=1}^{N} \left|t_{mn}^{(k)}\right| = \max_{1\leq m\leq N} \sum_{n=m}^{N} \left|t_{mn}^{(k)}\right| \qquad \left(\text{since } \mathbf{T}^k \text{ is upper-triangular}\right) \\
&\leq \max_{1\leq m\leq N} \left\{ \rho(\mathbf{T})^k + \sum_{n=m+1}^{N} \left|t_{mn}^{(k)}\right| \right\} \qquad \left(\text{since } |t_{mm}^{(k)}| = |\lambda_m|^k \leq \rho(\mathbf{T})^k\right)
\end{aligned}$$

Now, use Theorem 2.4 to substitute for the values of $t_{mn}^{(k)}$.

$$= \max_{1 \le m \le N} \left\{ \rho(\mathbf{T})^k + \sum_{n=m+1}^{N} \left| \sum_{j=1}^{n-m} \sum_{\substack{m=\alpha_1<\alpha_2 \\ <\ldots<\alpha_{j+1}=n}} \left(\prod_{i=1}^{j} t_{\alpha_i \alpha_{i+1}} \right) \sum_{|\Delta_{j+1}|=k-j} \prod_{l=1}^{j+1} \lambda_{\alpha_l}^{\delta_l} \right| \right\}$$

$$\le \max_{1 \le m \le N} \left\{ \rho(\mathbf{T})^k + \sum_{n=m+1}^{N} \left| \sum_{j=1}^{n-m} \sum_{\substack{m=\alpha_1<\alpha_2 \\ <\ldots<\alpha_{j+1}=n}} M_{\mathbf{T}}^{j} \sum_{|\Delta_{j+1}|=k-j} \rho(\mathbf{T})^{k-j} \right| \right\}$$

By Lemmas 2.7 and 3.9 we get:

$$= \max_{1 \le m \le N} \left\{ \rho(\mathbf{T})^k + \sum_{n=m+1}^{N} \left| \sum_{j=1}^{n-m} \binom{n-m-1}{j-1} M_{\mathbf{T}}^{j} \binom{k}{j} \rho(\mathbf{T})^{k-j} \right| \right\}$$

$$\le \rho(\mathbf{T})^k + \sum_{n=2}^{N} \sum_{j=1}^{N-1} \binom{n-2}{j-1} M_{\mathbf{T}}^{j} \binom{k}{j} \rho(\mathbf{T})^{k-j} \tag{3.2}$$

In order to simplify further, we expand the first summation of the second term.

$$= \rho(\mathbf{T})^k + \sum_{j=1}^{1} \binom{0}{j-1} M_{\mathbf{T}}^{j} \binom{k}{j} \rho(\mathbf{T})^{k-j} + \sum_{j=1}^{2} \binom{1}{j-1} M_{\mathbf{T}}^{j} \binom{k}{j} \rho(\mathbf{T})^{k-j}$$
$$+ \ldots + \sum_{j=1}^{N-1} \binom{N-2}{j-1} M_{\mathbf{T}}^{j} \binom{k}{j} \rho(\mathbf{T})^{k-j}$$

Rearranging terms we get:

$$= \rho(\mathbf{T})^k + M_{\mathbf{T}} \binom{k}{1} \rho(\mathbf{T})^{k-1} \left[\binom{0}{0} + \binom{1}{0} + \ldots + \binom{N-2}{0} \right]$$
$$+ M_{\mathbf{T}}^2 \binom{k}{2} \rho(\mathbf{T})^{k-2} \left[\binom{1}{1} + \binom{2}{1} + \ldots + \binom{N-2}{1} \right]$$
$$+ \ldots + M_{\mathbf{T}}^{N-1} \binom{k}{N-1} \rho(\mathbf{T})^{k-(N-1)} \left[\binom{N-2}{N-2} \right]$$

By Lemma 3.11, this is equal to:

$$= \rho(\mathbf{T})^k + M_{\mathbf{T}} \binom{k}{1} \rho(\mathbf{T})^{k-1} \left[\binom{N-1}{1} \right] + M_{\mathbf{T}}^2 \binom{k}{2} \rho(\mathbf{T})^{k-2} \left[\binom{N-1}{2} \right]$$
$$+ \ldots + M_{\mathbf{T}}^{N-1} \binom{k}{N-1} \rho(\mathbf{T})^{k-(N-1)} \left[\binom{N-1}{N-1} \right]$$

$$= \rho(\mathbf{T})^k + \sum_{j=1}^{N-1} M_{\mathbf{T}}^j \binom{k}{j} \rho(\mathbf{T})^{k-j} \left[\binom{N-1}{j} \right]$$

$$= \sum_{j=0}^{N-1} M_{\mathbf{T}}^j \binom{k}{j} \rho(\mathbf{T})^{k-j} \left[\binom{N-1}{j} \right]$$

The last inequality follows since $\rho(\mathbf{T})^k = \binom{N-1}{0}\binom{k}{0} M_{\mathbf{T}}^0 \rho(\mathbf{T})^k$. Rearranging terms finishes the proof. □

Corollary 3.15.

$$||\boldsymbol{T}^k||_2 \leq \sqrt{N} \sum_{j=0}^{N-1} \binom{N-1}{j} \binom{k}{j} M_{\boldsymbol{T}}^j \rho(\boldsymbol{T})^{k-j}$$

Proof. The proof follows directly from the fact that $||\mathbf{T}^k||_2 \leq \sqrt{N}||\mathbf{T}^k||_\infty$. □

3.2 The Kreiss Matrix Theorem

The Kreiss Matrix Theorem gives both a lower and upper bound for the supremum of the norm of a matrix power; these bounds were originally developed in 1962 by Kreiss [3]. There have been at least eight improvements to the upper bound with the proof of the last improvement completed in 1991. See Trefethen and Embree [7, p. 177] for a list of these improvements, as well as for further references to their proofs. In this section we provide the proof of the last improvement. First, we need to discuss an important term behind the theorem. The **Kreiss constant** of the matrix $\mathbf{A} \in M_N$ is defined as

$$\mathcal{K}(\mathbf{A}) := \sup_{\varepsilon > 0} \frac{\rho_\varepsilon(\mathbf{A}) - 1}{\varepsilon}$$

Our first order of business in this section is to show that the Kreiss constant is equivalent to another expression, which we will use more in this paper.

Theorem 3.16. *Given $\boldsymbol{A} \in M_N$,*

$$\mathcal{K}(\boldsymbol{A}) = \sup_{\varepsilon > 0} \frac{\rho_\varepsilon(\boldsymbol{A}) - 1}{\varepsilon} = \sup_{|z| > 1} (|z| - 1)||(z\boldsymbol{I} - \boldsymbol{A})^{-1}||$$

Proof. Fix $\varepsilon > 0$. By Corollary 2.14, and since $R(z)$ is continuous on its domain, the set $\sigma_\varepsilon(\mathbf{A})$ is bounded; it is also easy to see that it is open. By definition of $\rho_\varepsilon(\mathbf{A})$, there exists a sequence

$\{z_m\} \in \sigma_\varepsilon(\mathbf{A})$, with $|z_m| \to \rho_\varepsilon(\mathbf{A})$ as $m \to \infty$. Note that $||(z_m\mathbf{I} - \mathbf{A})^{-1}|| > \varepsilon^{-1}$, for every m. Furthermore, since $\sigma_\varepsilon(\mathbf{A})$ is bounded, there is a subsequence $\{z_{m_j}\}$, which converges to a point z_0 in the closure of $\sigma_\varepsilon(\mathbf{A})$ as $j \to \infty$. Again, since $R(z)$ and the norm are continuous, and since the absolute value is continuous, we have

$$||(z_0\mathbf{I} - \mathbf{A})^{-1}|| \geq \varepsilon^{-1} \qquad \text{and} \qquad |z_0| = \rho_\varepsilon(\mathbf{A})$$

Thus,

$$(|z_0| - 1)||(z_0\mathbf{I} - \mathbf{A})^{-1}|| \geq \frac{\rho_\varepsilon(\mathbf{A}) - 1}{\varepsilon}$$

Now, take the supremum over all complex numbers:

$$\sup_z(|z| - 1)||(z\mathbf{I} - \mathbf{A})^{-1}|| \geq \frac{\rho_\varepsilon(\mathbf{A}) - 1}{\varepsilon}$$

If $|z| \leq 1$, then $(|z| - 1)||(z\mathbf{I} - \mathbf{A})^{-1}|| \leq 0$, and if $|z| > 1$, then $(|z| - 1)||(z\mathbf{I} - \mathbf{A})^{-1}|| > 0$. Thus, we may restrict the supremum above to complex numbers of modulus greater than one, since there are certainly values for z for which $(|z| - 1)||(z\mathbf{I} - \mathbf{A})^{-1}||$ is positive.

$$\sup_z(|z| - 1)||(z\mathbf{I} - \mathbf{A})^{-1}|| = \sup_{|z|>1}(|z| - 1)||(z\mathbf{I} - \mathbf{A})^{-1}||$$

Now we show that we can get the inequality in the other direction. Fix z with $|z| > 1$. Let $\varepsilon_0 = ||(z\mathbf{I} - \mathbf{A})^{-1}||$. Then z is on the boundary of $\sigma_{\varepsilon_0}(\mathbf{A})$. There exists a sequence $\{\omega_m\} \subseteq \sigma_{\varepsilon_0}(\mathbf{A})$, with $\omega_m \to z$. Since $\rho_{\varepsilon_0}(\mathbf{A}) \geq |\omega_m|$, for every m, we have $\rho_{\varepsilon_0}(\mathbf{A}) \geq |z|$. Then,

$$\frac{\rho_{\varepsilon_0}(\mathbf{A}) - 1}{\varepsilon_0} \geq (|z| - 1)||(z\mathbf{I} - \mathbf{A})^{-1}||$$

Therefore,

$$\sup_{\varepsilon>0}\frac{\rho_\varepsilon(\mathbf{A}) - 1}{\varepsilon} \geq (|z| - 1)||(z\mathbf{I} - \mathbf{A})^{-1}||$$

In fact, this is true for all z with $|z| > 1$; thus,

$$\sup_{\varepsilon > 0} \frac{\rho_\varepsilon(\mathbf{A}) - 1}{\varepsilon} = \sup_{|z|>1} (|z| - 1)||(z\mathbf{I} - \mathbf{A})^{-1}||$$

□

The Kreiss Matrix Theorem is stated below; it's proof will follow in two parts—the lower bound, and upper bound–which are the substance of Theorems 3.18 and 3.24, respectively. For simplicity, for the rest of this section we will assume that $||\cdot||$ is the spectral norm.

Theorem 3.17 (Kreiss Matrix Theorem)**.** *Given a matrix $\boldsymbol{A} \in M_N$,*

$$\mathcal{K}(\boldsymbol{A}) \leq \sup_{k \geq 0} ||\boldsymbol{A}^k|| \leq eN\mathcal{K}(\boldsymbol{A}) \tag{3.3}$$

In 1984, LeVeque and Trefethen [4] proved the above theorem with the factor eN replaced by $2eN$. They accompanied this by the conjecture that the factor of 2 was unnecessary. It wasn't until 1991 that Spijker's result [6] could be applied to Leveque and Trefethen's work to reduce the Kreiss upper bound to its present form.

Note that Theorem 3.17 is quite uninteresting when applied to normal matrices. Given a normal matrix $\mathbf{A}$ with $\rho(\mathbf{A}) < 1$,

$$\sup_{k \geq 0} ||\mathbf{A}^k|| = \sup_{k \geq 0} \rho(\mathbf{A})^k = 1$$

by Corollary 3.2. Thus, the graph of $||\mathbf{A}^k||$ starts at 1 and tends to 0 as $k \to \infty$. Also, note that for a normal matrix $\mathbf{A}$, Corollary 1.20, leads us to conclude that $\rho_\varepsilon(\mathbf{A}) = \rho(\mathbf{A}) + \varepsilon$. If in addition, we assume $\rho(\mathbf{A}) > 1$, then

$$\lim_{\varepsilon \to 0^+} \frac{\rho(\mathbf{A}) + \varepsilon - 1}{\varepsilon} = \infty$$

Thus,

$$\mathcal{K}(\mathbf{A}) = \sup_{\varepsilon > 0} \frac{\rho(\mathbf{A}) + \varepsilon - 1}{\varepsilon} = \infty$$

On the other hand, if we assume $\rho(\mathbf{A}) \leq 1$, then for each $\varepsilon > 0$,

$$\frac{\rho(\mathbf{A}) + \varepsilon - 1}{\varepsilon} \leq 1$$

Observe that,

$$\lim_{\varepsilon \to \infty} \frac{\rho(\mathbf{A}) + \varepsilon - 1}{\varepsilon} = 1$$

Thus we may conclude, $\mathcal{K}(\mathbf{A}) = 1$.

The Kreiss Matrix Theorem becomes more interesting when examining the norm of powers of nonnormal matrices, which tend to exhibit transient growth over an interval and then eventually approach zero as k gets large. In nonnormal cases, many values for $\mathcal{K}(\mathbf{A})$ are possible, and they may be used in the Kreiss Matrix Theorem to form bounds for the peak of the transient phase of the graph of $||\mathbf{A}^k||$.

We will proceed with the proof of the lower bound of the Kreiss Matrix Theorem. Then we will provide a few preliminary results, including Spijker's Lemma, which will pave the way for a proof of the Kreiss Matrix Theorem upper bound.

Theorem 3.18. *Given $\boldsymbol{A} \in M_N$,*

$$\sup_{k \geq 0} ||\boldsymbol{A}^k|| \geq \mathcal{K}(\boldsymbol{A})$$

Proof. The case where $\sup_{k \geq 0} ||\mathbf{A}^k|| = \infty$ is trivial. Thus, suppose that there exists $M < \infty$ with $\sup_{k \geq 0} ||\mathbf{A}^k|| = M$. By Theorem 3.1 we must have $\rho(\mathbf{A}) \leq 1$.

Choose $z \in \mathbb{C}$, with $|z| > 1$. Note that $z \notin \sigma(\mathbf{A})$ and so $R(z) < \infty$. Furthermore,

$$\begin{aligned} ||(z\mathbf{I} - \mathbf{A})^{-1}|| &= \left|\left| \sum_{k=0}^{\infty} \frac{\mathbf{A}^k}{z^{k+1}} \right|\right| \leq \sum_{k=0}^{\infty} \frac{||\mathbf{A}^k||}{|z^{k+1}|} \\ &\leq \frac{M}{|z|} \sum_{k=0}^{\infty} \left(\frac{1}{|z|}\right)^k = \frac{M}{|z|} \frac{1}{1 - 1/|z|} = \frac{M}{|z| - 1} \end{aligned}$$

Then,

$$(|z| - 1)||(z\mathbf{I} - \mathbf{A})^{-1}|| \leq M = \sup_{k \geq 0} ||\mathbf{A}^k||$$

Since z is arbitrary we have $\mathcal{K}(\mathbf{A}) \leq \sup_{k \geq 0} ||\mathbf{A}^k||$. □

In order to prove the upper bound of the Kreiss Matrix Theorem we will need a result by Spijker [6] obtained in 1991. Two preliminary lemmas will smooth the way.

Lemma 3.19. *Given $a, b \in \mathbb{R}$, and $\theta \in [0, 2\pi]$,*

$$|a\cos\theta + b\sin\theta| \leq \sqrt{a^2 + b^2}$$

Proof. This is true by the Cauchy Schwartz inequality (Lemma 1.6):

$$|a\cos\theta + b\sin\theta| = |(a,b)\cdot(\cos\theta,\sin\theta)| \leq \sqrt{a^2+b^2}\sqrt{\cos^2\theta+\sin^2\theta} = \sqrt{a^2+b^2}$$

□

Lemma 3.20. *Let $Q(z)$ be a complex polynomial of degree n. The restriction of $z^n\overline{Q(z)}$ to the circle $\mathbb{S} = \{z : |z| = d\}$, for some $d > 0$, is equivalent to the restriction to $\mathbb{S}$ of some polynomial of degree less than or equal to n.*

Proof. Let the expansion of Q be $Q(z) = q_n z^n + q_{n-1}z^{N-1} + \ldots + q_1 z + q_0$. Then,

$$\overline{Q(z)} = \overline{q}_n\,\overline{z}^n + \overline{q}_{n-1}\,\overline{z}^{N-1} + \ldots + \overline{q}_1\,\overline{z} + \overline{q}_0$$

Observe that $\mathbb{S}$ is parameterized by the function $z(t) = de^{it}$ as t ranges from 0 to 2π. Thus,

$$\overline{z(t)} = de^{-it} \qquad \text{and} \qquad (z(t))^{-1} = \frac{e^{-it}}{d}.$$

Thus, $\overline{z(t)} = d^2(z(t))^{-1}$, and on $[0, 2\pi]$, we have:

$$\begin{aligned}\overline{Q(z)} &= \overline{q}_n d^{2n}(z(t))^{-n} + \overline{q}_{n-1}d^{2(n-1)}(z(t))^{-(n-1)} + \ldots + \overline{q}_1 d^2(z(t))^{-1} + \overline{q}_0\\ z^n\overline{Q(z)} &= \overline{q}_n d^{2n} + \overline{q}_{n-1}d^{2(n-1)}z(t) + \ldots + \overline{q}_1 d^2 z(t)^{N-1} + \overline{q}_0 z^n\end{aligned}$$

□

Lemma 3.21 (Spijker's Lemma)**.** *Let $\mathbb{S}$ be a circle in the complex plane with radius d. Let $r(z) = P(z)/Q(z)$, where P and Q are polynomials over $\mathbb{C}$ of degree less than or equal to N, for some $N \in \mathbb{N}$, with $Q(z) \neq 0$ on $\mathbb{S}$. Then,*

$$\int_{\mathbb{S}} |r'(z)||dz| \leq 2\pi N \max_{z\in\mathbb{S}} |r(z)|.$$

Proof. First, parameterize $\mathbb{S}$ by the map z, where $z = z(t) = de^{it}$ for $0 \leq t \leq 2\pi$. On the domain $[0, 2\pi]$, define the complex valued function f, and real-valued functions g and h, according to the following:

$$f(t) = g(t) + ih(t) = r(de^{it})$$

Since r is a rational function, it is analytic on its domain, which includes $\mathbb{S}$. Thus, r is differentiable, and the following holds:

$$f'(t) = g'(t) + ih'(t) = die^{it}r'(de^{it})$$

This, coupled with the parameterization of $\mathbb{S}$ results in the two identities,

$$\frac{|f'(t)|}{d} = |r'(t)| \qquad \text{and} \qquad |dz| = d\,dt \tag{3.4}$$

Equations (3.4) can be used to turn the contour integral of $|r'(z)|$ into a definite integral of $|f'(t)|$.

$$\int_{\mathbb{S}} |r'(z)||dz| = \int_0^{2\pi} |f'(t)|\,dt \tag{3.5}$$

Since $f'(t)$ is complex-valued, we can write it in polar form as $f'(t) = |f'(t)|\cos\omega_t + i|f'(t)|\sin\omega_t$, where ω_t depends on t. Thus for each t, we have $g'(t) = |f'(t)|\cos\omega_t$ and $h'(t) = |f'(t)|\sin\omega_t$. Observe that for each t,

$$\begin{aligned}
\int_0^{2\pi} |g'(t)\cos\theta + h'(t)\sin\theta|\,d\theta &= \int_0^{2\pi} |(|f'(t)|\cos\omega_t)\cos\theta + (|f'(t)|\sin\omega_t)\sin\theta|\,d\theta \\
&= \int_0^{2\pi} |f'(t)||\cos\omega_t\cos\theta + \sin\omega_t\sin\theta|\,d\theta \\
&= |f'(t)|\int_0^{2\pi} |\cos(\omega_t - \theta)|\,d\theta = 4|f'(t)|
\end{aligned}$$

Dividing both sides by 4 and integrating with respect to t, we get

$$\frac{1}{4}\int_0^{2\pi}\left(\int_0^{2\pi} |g'(t)\cos\theta + h'(t)\sin\theta|\,d\theta\right)dt = \int_0^{2\pi} |f'(t)|\,dt \tag{3.6}$$

For now let's focus on the inner integral on the left of equation (3.6). Fix $\theta \in [0, 2\pi]$, and define

F_θ by

$$F_\theta(t) \;=\; g(t)\cos\theta \;+\; h(t)\sin\theta \tag{3.7}$$

Since $r(z)$ is a rational function, g and h, will be restrictions of rational functions to the interval $[0, 2\pi]$. Then g' and h' are also restrictions of rational functions. Thus, g' and h' are each either identically 0, or they have a finite number of zeros. Thus, if F'_θ has an infinite number of zeros on $[0, 2\pi]$, then F'_θ is in fact identical to 0 on the interval. In this case, since $F'_\theta(t) \;=\; g'(t)\cos\theta \;+\; h'(t)\sin\theta$, the integral (3.6) will just be 0, and by equation (3.5), the lemma is proved.

Now suppose without loss of generality that $F'_\theta(t)$ has finitely many zeros in $[0, 2\pi]$. Then, there exists an integer k and real values $t_0, t_1, \ldots, t_k$, with $t_0 \;=\; 0 \;<\; t_1 \;<\; t_2 \;<\; \ldots \;<\; t_k \;=\; 2\pi$, such that $|F'_\theta(t)| \;>\; 0$ on (t_{j-1}, t_j), for $j \;=\; 1, 2, \ldots, k$. Then, the sign of F'_θ does not change on the open intervals (t_{j-1}, t_j). If $F'_\theta(t) \;<\; 0$ on (t_{j-1}, t_j), then $-\,(F_\theta(t_j) \;-\; F_\theta(t_{j-1}) \;>\; 0$. This implies that

$$\int_{t_{j-1}}^{t_j} |F'_\theta(t)|\,dt \;=\; -\int_{t_{j-1}}^{t_j} F'_\theta(t)\,dt \;=\; -\,(F_\theta(t_j) \;-\; F_\theta(t_{j-1})) \;=\; |F_\theta(t_j) \;-\; F_\theta(t_{j-1})|$$

In contrast, if $F'_\theta(t) \;>\; 0$ on (t_{j-1}, t_j) then trivially,

$$\int_{t_{j-1}}^{t_j} |F'_\theta(t)|\,dt \;=\; |F_\theta(t_j) \;-\; F_\theta(t_{j-1})|$$

Since this is true for $j \;=\; 1, 2, \ldots, k$,

$$\int_0^{2\pi} |F'_\theta(t)|\,dt \;=\; \sum_{j=1}^{k} \int_{t_{j-1}}^{t_j} |F'_\theta(t)|\,dt \;=\; \sum_{j=1}^{k} |F_\theta(t_j) \;-\; F_\theta(t_{j-1})|$$

Now, for $j \;=\; 1, 2, \ldots, k$, define the sets,

$$B_j \;=\; \left\{ y \in \mathbb{R} : \min_{t\in[t_{j-1},t_j]} F_\theta(t) \;\leq\; y \;\leq\; \max_{t\in[t_{j-1},t_j]} F_\theta(t) \right\}$$

and let

$$a \;=\; \max_{0\,\leq\, t\,\leq\, 2\pi} |F_\theta(t)|$$

Observe that $B_j \subseteq [-a, a]$ for each j. Also observe that a value y^* is in B_j if and only if there exists $t \in [0, 2\pi]$, such that $F(t) \;=\; y^*$. For each j, let χ_{B_j} be the function that takes the value 1

on B_j, and 0 everywhere else:

$$\chi_{B_j}(y) \;=\; \begin{cases} 1, & \text{if } y \in B_j \\ 0, & \text{if } y \notin B_j \end{cases}$$

Then since F_θ is monotonic on $[t_{j-1}, t_j]$, the maxima and minima of the F_θ on $[t_{j-1}, t_j]$ occur at the endpoints; thus,

$$\int_{-a}^{a} \chi_{B_j}(y)dy \;=\; |F_\theta(t_j) \;-\; F_\theta(t_{j-1})|,$$

for $j \;=\; 1, 2, \ldots, k$. This leads to the following,

$$\begin{aligned} \int_0^{2\pi} |F'_\theta(t)|\, dt \;&=\; \sum_{j=1}^{k} |F_\theta(t_j) \;-\; F_\theta(t_{j-1})| \\ &= \sum_{j=1}^{k} \int_{-a}^{a} \chi_{B_j}(y)dy \\ &= \int_{-a}^{a} \sum_{j=1}^{k} \chi_{B_j}(y)dy \\ &\leq\; 2a \max_{y\in[\,-a,a]} \sum_{j=1}^{k} \chi_{B_j}(y) \end{aligned} \tag{3.8}$$

Now let y^* satisfy

$$\sum_{j=1}^{k} \chi_{B_j}(y^*) \;=\; \max_{y\in[\,-a,a]} \sum_{j=1}^{k} \chi_{B_j}(y) \tag{3.9}$$

Since F_θ is continuous on $\mathbb{S}$, there exists $t^* \in [0, 2\pi]$ such that $F_\theta(t^*) \;=\; y^*$. We want to show that there are at most $2N$ distinct values, $t \in [0, 2\pi]$, such that $F(t) \;=\; y^*$. Observe that,

$$\begin{aligned} 2F_\theta(t) \;&=\; 2(g(t)\cos\theta \;+\; h(t)\sin\theta) \\ &=\; 2(g(t)\cos\theta \;+\; h(t)\sin\theta) \;+\; ig(t)\sin\theta \;-\; ig(t)\sin\theta \;+\; ih(t)\cos\theta \;-\; ih(t)\cos\theta \\ &=\; (\cos\theta \;-\; i\sin\theta)(g(t) \;+\; ih(t)) \;+\; (\cos\theta \;+\; i\sin\theta)(g(t) \;-\; ih(t)) \\ &=\; \mathrm{e}^{-i\theta} r(z) \;+\; \mathrm{e}^{i\theta}\overline{r(z)} \end{aligned}$$

Since $F'_\theta(t)$ is not identically 0 on $[0, 2\pi]$, neither side above is constant as t (and $z \;=\; d\mathrm{e}^{it}$) varies.

If we replace $F_\theta(t)$ above by y^*, we obtain the equation

$$2y^* \;=\; \mathrm{e}^{-i\theta} r(z) \;+\; \mathrm{e}^{i\theta}\overline{r(z)} \tag{3.10}$$

which is <u>not</u> satisfied for all z in $\mathbb{S}$. Likewise, since $Q(z) \neq 0$ for all z in $\mathbb{S}$, we may multiply both sides of equation (3.10) by $Q(z)\overline{Q(z)}$ to get,

$$2y^* Q(z)\overline{Q(z)} \;=\; Q(z)\overline{Q(z)}\left[\mathrm{e}^{-i\theta} r(z) \;+\; \mathrm{e}^{i\theta}\overline{r(z)}\right]$$

This new equation is satisfied by the exact same values that satisfy equation (3.10). Substituting $r(z) = P(z)/Q(z)$ and rearranging terms gives,

$$\mathrm{e}^{-i\theta} P(z)\overline{Q(z)} \;+\; \mathrm{e}^{i\theta}\overline{P(z)}Q(z) \;-\; 2y^* Q(z)\overline{Q(z)} \;=\; 0 \tag{3.11}$$

Now multiply the above identity by z^N, and set $u(z)$ equal to the left-hand side:

$$u(z) \;=\; \mathrm{e}^{-i\theta} P(z) z^N \overline{Q(z)} \;+\; \mathrm{e}^{i\theta} z^N \overline{P(z)} Q(z) \;-\; 2y^* Q(z) z^N \overline{Q(z)} \tag{3.12}$$

By Lemma 3.20, $u(z)$ is the restriction of a polynomial to $\mathbb{S}$, of degree less than or equal to $2N$. Note that the polynomial cannot be of degree 0 because equation (3.12) is not satisfied for all z in $\mathbb{S}$. Therefore, u has at most $2N$ distinct complex zeros on $\mathbb{S}$. Since the map $t \mapsto d\mathrm{e}^{it}$ is bijective, and since equation (3.11) is equivalent to equation (3.10) for $z \in \mathbb{S}$, there are at most $2N$ distinct values t such that both $F(t) = y^*$, and $z = d\mathrm{e}^{it}$ satisfies equation (3.11). Thus, the set $\{F(t) : F(t) = y^*, t \in [0, 2\pi]\}$, has nonempty intersection with at most $2N$ distinct sets B_j. By equation (3.9) and the definition of χ_{B_j},

$$\max_{y \in [-a,a]} \sum_{j=1}^{k} \chi_{B_j}(y) \;=\; \sum_{j=1}^{k} \chi_{B_j}(y^*) \;\leq\; 2N \tag{3.13}$$

Now we put everything together, starting from equation (3.5).

$$\begin{aligned}
\int_{\mathbb{S}} |r'(z)||dz| &= \int_0^{2\pi} |f'(t)|\,dt \\
&= \frac{1}{4}\int_0^{2\pi}\left(\int_0^{2\pi} |g'(t)\cos\theta + h'(t)\sin\theta|\,d\theta\right)dt && \text{(by euation (3.6))} \\
&= \frac{1}{4}\int_0^{2\pi}\left(\int_0^{2\pi} |F'(t)|\,d\theta\right)dt \\
&\leq \frac{1}{4}(2\pi 2a 2N) && \text{(by equations (3.8) and (3.9))} \\
&= 2\pi N \max_{0\leq t\leq 2\pi} |F_\theta(t)| \\
&\leq 2\pi N \max_{z\in\mathbb{S}} |r(z)|
\end{aligned}$$

The last inequality follows from Lemma 3.19, since for every $t \in [0, 2\pi]$,

$$|F_\theta(t)| = |g(t)\cos\theta + h(t)\sin\theta| \leq \sqrt{g(t)^2 + h(t)^2} = |r(de^{it})|$$

□

Lemma 3.22. *Given a matrix $\boldsymbol{A} \in M_N$, let $R(z)$ be the resolvent of $\boldsymbol{A}$, and let $r(z) = \boldsymbol{u}^* R(z)\boldsymbol{v}$, for some unit vectors $\boldsymbol{u}$ and $\boldsymbol{v}$. Let $\Gamma = \left\{z : |z| = 1 + (k+1)^{-1}\right\}$, for some $k \in \mathbb{N}$. Then,*

$$\sup_{z\in\Gamma} |r(z)| \leq (k+1)\mathcal{K}(\boldsymbol{A})$$

Proof.

$$\begin{aligned}
\mathcal{K}(\mathbf{A}) &= \sup_{|z|>1} (|z|-1)||(z\mathbf{I}-\mathbf{A})^{-1}|| = \sup_{|z|>1} (|z|-1)||R(z)|| \\
&= \sup_{|z|>1} (|z|-1)|r(z)| \geq \sup_{z\in\Gamma}(|z|-1)|r(z)| = \sup_{z\in\Gamma}(k+1)^{-1}|r(z)|
\end{aligned}$$

The result follows by multiplying both sides by $(k+1)$. □

Lemma 3.23. *The sequence, $\{\left(1+\frac{1}{n}\right)^n\}_{n=1}^{\infty}$, who's limit is defined as the irrational number e, is monotonically increasing.*

Proof. Let f be defined $f(x) = x^{n+1}$. Then f is continuous and differentiable, and the derivative

of f is $f'(x) = (n+1)x^n$. Given numbers a and b with $0 \leq a < b$, by the mean value theorem there exists a value $c \in (a,b)$, such that,

$$\begin{aligned} b^{n+1} - a^{n+1} &= (n+1)c^n(b-a) < (n+1)b^n(b-a) = (n+1)b^{n+1} - (n+1)b^n a \\ -a^{n+1} &< nb^{n+1} - (n+1)b^n a = b^n(nb - (n+1)a) \\ a^{n+1} &> b^n((n+1)a - nb) \end{aligned} \tag{3.14}$$

This is true for all $0 \leq a < b$, so in particular, it is true for $a = 1 + \frac{1}{n+1}$, and $b = 1 + \frac{1}{n}$, for each $n = 1, 2, \ldots$. Substituting these values into inequality (3.14), we get,

$$\begin{aligned} \left(1 + \frac{1}{n+1}\right)^{n+1} &> \left(1 + \frac{1}{n}\right)^n \left((n+1)\left(1 + \frac{1}{n+1}\right) - n\left(1 + \frac{1}{n}\right)\right) \\ &= \left(1 + \frac{1}{n}\right)^n (n + 1 + 1 - n - 1) = \left(1 + \frac{1}{n}\right)^n \end{aligned}$$

□

Theorem 3.24. *Given a matrix $\boldsymbol{A} \in M_N$ with spectral radius $\rho(\boldsymbol{A}) < 1$,*

$$\sup_{k>0} ||\boldsymbol{A}^k|| \leq eN\mathcal{K}(\boldsymbol{A})$$

Proof. Let $\Gamma = \{z : |z| = d\}$, where $d = 1 + (k+1)^{-1}$. Define the rational function $r(z) = \mathbf{u}^* R(z)\mathbf{v}$, where $\mathbf{u}$ and $\mathbf{v}$ are unit vectors in $\mathbb{C}^N$, and $R(z)$ is the resolvent of $\mathbf{A}$. By Theorem 2.16,

$$\mathbf{u}^*\mathbf{A}^k\mathbf{v} = \mathbf{u}^*\left(\frac{1}{2\pi i}\int_\Gamma z^k R(z)\,dz\right)\mathbf{v} = \frac{1}{2\pi i}\int_\Gamma z^k \mathbf{u}^* R(z)\mathbf{v}\,dz = \frac{1}{2\pi i}\int_\Gamma z^k r(z)\,dz$$

If we parameterize Γ with the function $z(t) = d\mathrm{e}^{it}$, for $0 \leq t \leq 2\pi$, then the last integral becomes,

$$\frac{1}{2\pi i}\int_0^{2\pi} \left(d\mathrm{e}^{it}\right)^k r(d\mathrm{e}^{it}) i d\mathrm{e}^{it}\,dt = \frac{1}{2\pi i}\int_0^{2\pi} i d^{k+1}\mathrm{e}^{it(k+1)} r(d\mathrm{e}^{it})\,dt$$

Using integration by parts we obtain the following,

$$\frac{1}{2\pi i}\left[\frac{d^{k+1}}{i(k+1)}\mathrm{e}^{it(k+1)}r(d\mathrm{e}^{it})\Big|_0^{2\pi}\right] - \frac{1}{2\pi i}\int_0^{2\pi}\frac{d^{k+1}}{i(k+1)}\mathrm{e}^{it(k+1)}r'(d\mathrm{e}^{it})id\mathrm{e}^{it}\,dt$$
$$= 0 \quad + \quad \frac{1}{2\pi(k+1)}\int_0^{2\pi}\left(d\mathrm{e}^{it}\right)^{k+1}r'(d\mathrm{e}^{it})id\mathrm{e}^{it}\,dt$$

This is then equal to the contour integral,

$$\frac{1}{2\pi(k+1)}\int_\Gamma z^{k+1}r'(z)\,dz$$

Thus,

$$\mathbf{u}^*\mathbf{A}^k\mathbf{v} = \frac{1}{2\pi(k+1)}\int_\Gamma z^{k+1}r'(z)\,dz$$

Then,

$$\left|\mathbf{u}^*\mathbf{A}^k\mathbf{v}\right| \leq \frac{1}{2\pi(k+1)}\int_\Gamma |z|^{k+1}|r'(z)||dz|$$

Observe that $|z| = d = 1 + (k+1)^{-1}$, and thus $|z|^{k+1}$ is one term in the sequence $\{\left(1+\frac{1}{n}\right)^n\}_{n=1}^\infty$ who's limit defines the irrational number e. By Lemma 3.23, this sequence is increasing, and thus, $|z|^{k+1} < \mathrm{e}$. Continuing from above, we may replace $|z|^{k+1}$ by e.

$$\left|\mathbf{u}^*\mathbf{A}^k\mathbf{v}\right| \leq \frac{1}{2\pi(k+1)}\int_\Gamma \mathrm{e}|r'(z)||dz| = \frac{\mathrm{e}}{2\pi(k+1)}\int_\Gamma |r'(z)||dz|$$

By Spijker's Lemma (Lemma 3.21), we have,

$$\int_\Gamma |r'(z)||dz| \leq 2\pi N \sup_\Gamma |r(z)|$$

Then,

$$\left|\mathbf{u}^*\mathbf{A}^k\mathbf{v}\right| \leq \frac{\mathrm{e}}{2\pi(k+1)}2\pi N \sup_\Gamma |r(z)| = \frac{\mathrm{e}N}{(k+1)}\sup_\Gamma |r(z)|$$

By Lemma 3.22, $\sup_{z\in\Gamma}|r(z)| \leq (k+1)\mathcal{K}(\mathbf{A})$. Thus,

$$\left|\mathbf{u}^*\mathbf{A}^k\mathbf{v}\right| \leq \frac{\mathrm{e}N}{(k+1)}(k+1)\mathcal{K}(\mathbf{A}) = \mathrm{e}N\mathcal{K}(\mathbf{A})$$

Since the above is true for all $k \geq 0$, and since $||\mathbf{A}^k||_2 = |\mathbf{u}^*\mathbf{A}^k\mathbf{v}|$ (by Lemma 1.16), the theorem is proved. □

Chapter 4. Examples

In this chapter we investigate a few examples of where matrix powers and their norms are used in mathematics. First, we revisit the question asked on the first page of this thesis regarding the four different graphs and four different matrices (Figure 1). The figure is reprinted here for convenience. After that, we will look at an example of the Gauss-Seidel method, which is used in numerical linear algebra. Another example will investigate nonnormality in Markov processes, and the last example will look at a random upper-triangular matrix with a particular structure.

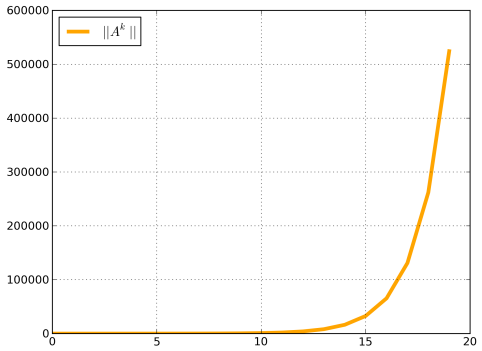
4.1 Revisiting Figure 1

We are now ready to revisit the question asked on the first page regarding the matrices in Figure 1. Matrix $\mathbf{V}$ has 2 as an eigenvalue, implying that the spectral radius is greater than one. Thus, by Theorem 3.1, The norm of powers of $\mathbf{V}$ will diverge to infinity. Graph number 1 is the only graph where this happens. Now observe that the matrix $\mathbf{Y}$ is symmetric, and therefore normal. Also, it is diagonal, and so the spectral radius is equal to the largest magnitude of the diagonal entries. Since this is $\left(\frac{80}{81}\right)$, Lemma 1.12 tells us that the graph of the norm of powers of $\mathbf{Y}$ will be equal to the graph of $\left(\frac{80}{81}\right)^k$; it will start at one, when $k = 0$, and then decrease exponentially as k increases. The only graph with these characteristics is graph number 4.

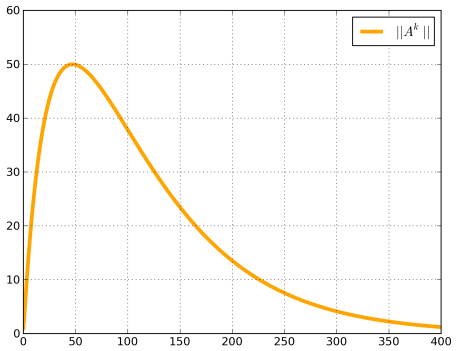
Now for the more subtle examples. Matrices $\mathbf{X} = (x_{mn})$ and $\mathbf{W} = (w_{mn})$ are nonnormal due to their upper-triangular forms; therefore, we might expect the graphs of the norms of their powers to have transient effects. Notice that graphs 2 and 3 look the same, but they are on very different scales: graph 2 peaks at around 50, while graph 3 peaks at a value well over 140,000. Let's take a

Figure 4.1: Figure 1 Revisited!

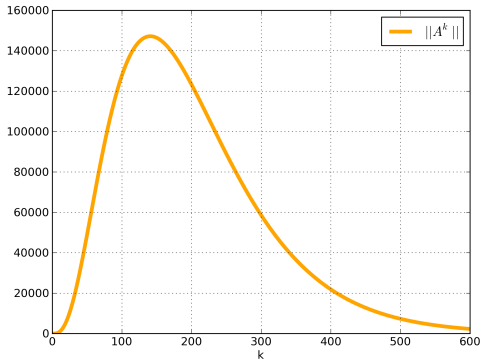
1.



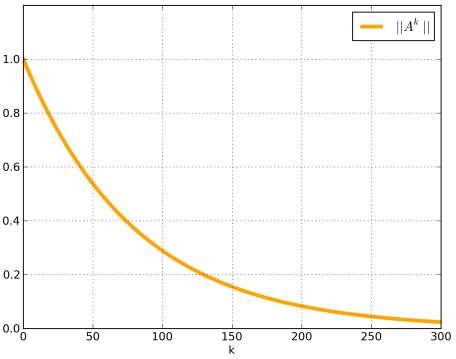
2.



3.



4.



$$\mathbf{V} = \begin{bmatrix} 0 & 2 & 0 & 0 \\ 2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 2 \end{bmatrix} \qquad \mathbf{W} = \begin{bmatrix} \frac{20}{21} & 0 & 2 & 2 \\ 0 & \frac{40}{41} & 0 & 2 \\ 0 & 0 & \frac{60}{61} & 0 \\ 0 & 0 & 0 & \frac{80}{81} \end{bmatrix}$$

$$\mathbf{X} = \begin{bmatrix} \frac{20}{21} & 2 & 0 & 0 \\ 0 & \frac{40}{41} & 2 & 0 \\ 0 & 0 & \frac{60}{61} & 2 \\ 0 & 0 & 0 & \frac{80}{81} \end{bmatrix} \qquad \mathbf{Y} = \begin{bmatrix} \frac{20}{21} & 0 & 0 & 0 \\ 0 & \frac{40}{41} & 0 & 0 \\ 0 & 0 & \frac{60}{61} & 0 \\ 0 & 0 & 0 & \frac{80}{81} \end{bmatrix}$$

look at bounds from Theorem 3.8.

$$\begin{aligned} ||\mathbf{W}^k||_2 &\geq |w_{2,4}^{(k)}| \\ &= w_{24} \sum_{|\Delta_2|=k-1} \left(\frac{40}{41}\right)^{\delta_1} \left(\frac{80}{81}\right)^{\delta_2} + w_{23}w_{34} \sum_{|\Delta_4|=k-3} \left(\frac{40}{41}\right)^{\delta_1} \left(\frac{60}{61}\right)^{\delta_3} \left(\frac{80}{81}\right)^{\delta_3} \end{aligned}$$

The equality follows from Theorem 2.4, and since the entries of $\mathbf{W}$ are all nonnegative. Observe that since $w_{23} = 0$, and since the eigenvalues are distinct, by Theorem 2.10 (with $C = k - j$) we

have that this is equal to:

$$2\left(\frac{\left(\frac{40}{41}\right)^k - \left(\frac{80}{81}\right)^k}{\frac{40}{41} - \frac{80}{81}}\right) + 0$$

This last expression can be graphed on a calculator to find that the bound reaches 40.59 at $k = 56$. This tells us that the graph of $||\mathbf{W}^k||_2$ is at least this high for the same value of k. Now let's look at the same bound for the matrix $\mathbf{X}$, but this time based on the $1,4$ entry of $\mathbf{X}^k$. Again, we take advantage of the fact that the eigenvalues are distinct to use Theorem 2.10.

$$\begin{aligned}
||\mathbf{X}^k||_2 &\geq |x_{1,4}^{(k)}| \\
&= x_{14} \sum_{|\Delta_2|=k-1} \left(\frac{20}{21}\right)^{\delta_1} \left(\frac{80}{81}\right)^{\delta_2} + x_{12}x_{24} \sum_{|\Delta_3|=k-2} \left(\frac{20}{21}\right)^{\delta_1} \left(\frac{40}{41}\right)^{\delta_2} \left(\frac{80}{81}\right)^{\delta_3} \\
&\quad + x_{23}x_{34} \sum_{|\Delta_3|=k-2} \left(\frac{40}{41}\right)^{\delta_1} \left(\frac{60}{61}\right)^{\delta_2} \left(\frac{80}{81}\right)^{\delta_3} + x_{12}x_{23}x_{34} \sum_{|\Delta_4|=k-3} \left(\frac{20}{21}\right)^{\delta_1} \left(\frac{40}{41}\right)^{\delta_2} \left(\frac{60}{61}\right)^{\delta_3} \left(\frac{80}{81}\right)^{\delta_4} \\
&= 0 + 0 + 0 + x_{12}x_{23}x_{34} \sum_{|\Delta_4|=k-3} \left(\frac{20}{21}\right)^{\delta_1} \left(\frac{40}{41}\right)^{\delta_2} \left(\frac{60}{61}\right)^{\delta_3} \left(\frac{80}{81}\right)^{\delta_4} \\
&= 8\left(\frac{\left(\frac{20}{21}\right)^k}{\left(\frac{20}{21} - \frac{40}{41}\right)\left(\frac{20}{21} - \frac{60}{61}\right)\left(\frac{20}{21} - \frac{80}{81}\right)} + \frac{\left(\frac{40}{41}\right)^k}{\left(\frac{40}{41} - \frac{20}{21}\right)\left(\frac{40}{41} - \frac{60}{61}\right)\left(\frac{40}{41} - \frac{80}{81}\right)}\right. \\
&\qquad \left. + \frac{\left(\frac{60}{61}\right)^k}{\left(\frac{60}{61} - \frac{20}{21}\right)\left(\frac{60}{61} - \frac{40}{41}\right)\left(\frac{60}{61} - \frac{80}{81}\right)} + \frac{\left(\frac{80}{81}\right)^k}{\left(\frac{80}{81} - \frac{20}{21}\right)\left(\frac{80}{81} - \frac{40}{41}\right)\left(\frac{80}{81} - \frac{60}{61}\right)}\right)
\end{aligned}$$

This latter expression has a maximum value of about $147,167$ at $k = 141$. we could have obtained an even simpler bound by noting that,

$$\begin{aligned}
x_{12}x_{23}x_{34} \sum_{|\Delta_4|=k-3} \left(\frac{20}{21}\right)^{\delta_1} \left(\frac{40}{41}\right)^{\delta_2} \left(\frac{60}{61}\right)^{\delta_3} \left(\frac{80}{81}\right)^{\delta_4} &> \sum_{|\Delta_4|=k-3} \left(\frac{20}{21}\right)^{\delta_1+\delta_2+\delta_3+\delta_4} \\
&= \binom{k}{3}\left(\frac{20}{21}\right)^{k-3} = \frac{k(k-1)(k-2)}{6}\left(\frac{20}{21}\right)^{k-3}
\end{aligned}$$

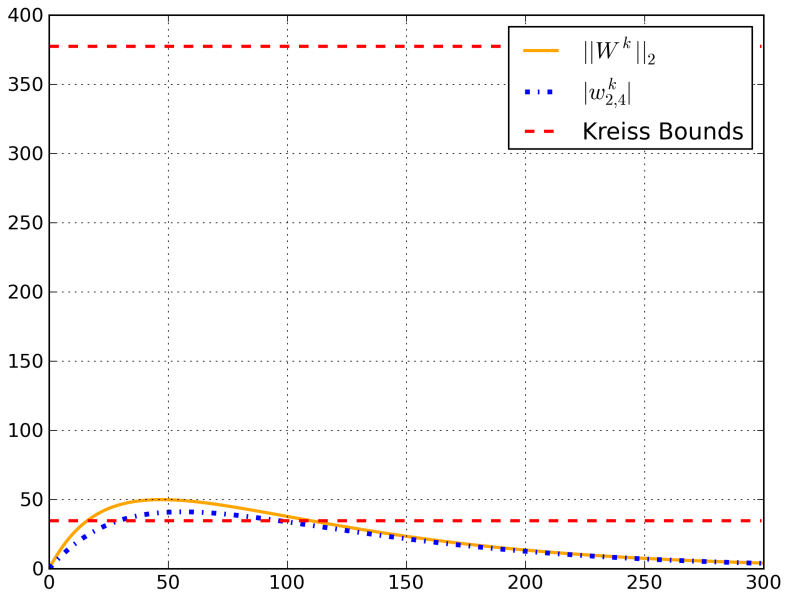
The maximum of this bound is about $17,007$, and it occurs at $k = 62$. Either of these bounds gives us enough information to exclude graph 2 as a choice for the graph of the norm of powers of the matrix $\mathbf{X}$. By the process of elimination, we conclude that graph 2 pertains to matrix $\mathbf{W}$, and graph 3 pertains to matrix $\mathbf{X}$.

What about other bounds? For $\mathbf{W}$, the Kreiss lower and upper bounds are 34.71 and 377.45, respectively. Thus, the lower bound from Theorem 3.8 with the entry $w_{2,4}$ beats the Kreiss lower bound! The maximum of the upper bound from Theorem 3.12 is:

$$||\mathbf{W}^k||_2 \leq \sum_{l=0}^{N-1} \sqrt{\sum_{j=0}^{l} \binom{l}{j}\binom{k}{j} 2^j \left(\frac{80}{81}\right)^{k-j}}$$

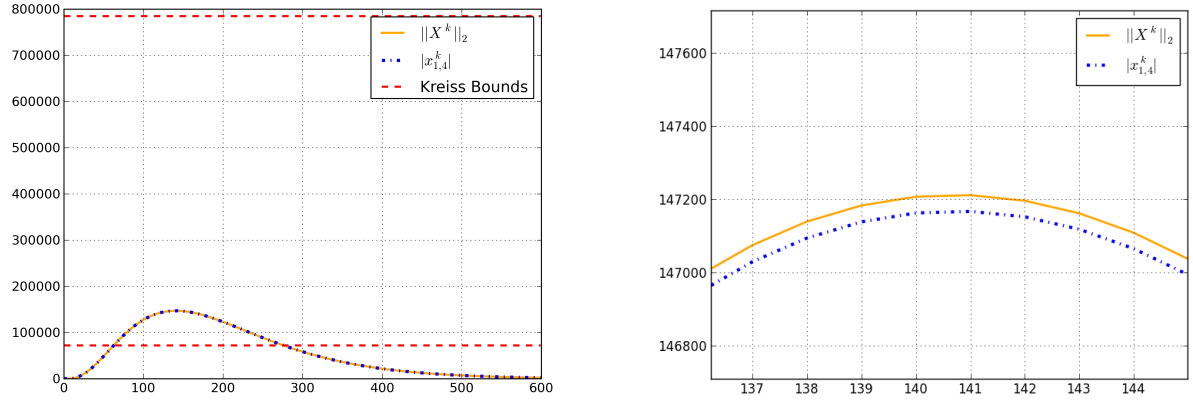
which is equal to $982,330$ for $k = 241$. For $\mathbf{X}$, the Kreiss lower and upper bounds are $72,210$ and $785,148$, respectively. Again, teh bound from Theorem 3.8 with entry $x_{1,4}$ is substantially better than the Kreiss lower bound. The upper bound based on Theorem 3.12 is exactly the same as for the previous matrix—$982,330$; this is due to the fact that the bound is obtained essentially by taking the norm after replacing all of the eigenvalues by the spectral radius, and all of the entries above the diagonal by the largest magnitude of the entries. Thus, this particular upper bound works much better for the matrix $\mathbf{X}$, than for the matrix $\mathbf{W}$.

Figure 4.2: Graph of $||\mathbf{W}^k||_2$



How do the bounds compare to actual values for the norm? The actual norm $||\mathbf{W}^k||_2$ peaks at a value of 49.98 when $k = 47$. In contrast, The actual norm $||\mathbf{X}^k||_2$ peaks at a value of $147,211.41$ when $k = 141$. This means that, among the bounds we have considered, the bounds obtained from

Figure 4.3: Graphs of $||\mathbf{X}^k||_2$



Theorem 3.8 are the closest to the actual norms for all values of k. A surprising amount of detail in the norm is captured by observing the upper-right corner entries of $\mathbf{W}^k$ and $\mathbf{X}^k$! In fact, in the latter case the lower bound appears to be equal to the norm, but upon zooming in we see that it is actually lower. Figures 4.2 and 4.3 show the graphs of the norm of powers for the matrices $\mathbf{W}$ and $\mathbf{X}$ with some of their bounds.

4.2 The Gauss-Seidel Method

The Gauss-Seidel method is an iterative method used in numerical linear algebra to solve matrix equations of the form $\mathbf{Ax} = \mathbf{b}$, where $\mathbf{A}$ is an $\mathbb{N} \times N$ matrix and $\mathbf{b}$ is an $N \times 1$ vector. Typically, it is most useful when N is large and $\mathbf{A}$ is sparse; in these cases, the Gauss-Seidel method tends to work better than Gaussian elimination techniques. The method proceeds by decomposing $\mathbf{A}$ into a sum of its diagonal, strictly upper-triangular, and strictly lower-triangular parts: $\mathbf{A} = \mathbf{D} + \mathbf{U} + \mathbf{L}$. If $\rho((\mathbf{D} + \mathbf{L})^{-1}\mathbf{U}) < 1$, then, given an initial vector $\mathbf{x}_0$, it has been shown that the recursive relation,

$$\mathbf{x}_{k+1} = (\mathbf{D} + \mathbf{L})^{-1}(\mathbf{b} - \mathbf{U}\mathbf{x}_k) \qquad \text{for } k = 0, 1, 2, \ldots$$

yields a sequence of vectors $\{\mathbf{x}_k\}_{k=0}^{\infty}$, which converges to the actual solution $\mathbf{x}$ of $\mathbf{Ax} = \mathbf{b}$. For background theory on why this is the case, consult Demmel [1, p. 282–294]. The error at the k^{th} step of iteration, $||\mathbf{x}^k - \mathbf{x}||_2$, is bounded by the product of $|| \left((\mathbf{D} + \mathbf{L})^{-1}\mathbf{U}\right)^k ||_2$ and the initial

error, $||\mathbf{x}_0 - \mathbf{x}||_2$; that is,

$$||\mathbf{x}^k - \mathbf{x}||_2 \leq ||\left((\mathbf{D} + \mathbf{L})^{-1}\mathbf{U}\right)^k||_2 ||\mathbf{x}_0 - \mathbf{x}||_2$$

In cases where $(\mathbf{D} + \mathbf{L})^{-1}\mathbf{U}$ is nonnormal, larger errors may result while k iterates through a transient phase, after which the error tends to zero as k tends to ∞.

The Gauss-Seidel method is often used in solving boundary value problems, such as those involving the famous Poisson equation:

$$\begin{cases} -u_{xx} = f, \quad \text{for } x \in [0,1] \\ u(0) = 0, \quad u(1) = 0 \end{cases}$$

One popular approach to solving this problem is to discretize the domain into $N+2$ equally spaced points (including the endpoints). Given a spacing of h, a function v evaluated at the n^{th} point can be represented with the condensed notation v_n for $n = 0, 1, 2, \ldots, N+1$. The finite difference scheme:

$$\begin{cases} -\frac{v_{n-1} - 2v_n + v_{n+1}}{h^2} = f, \quad \text{for } n = 1, 2, \ldots, N \\ v_0 = 0, \quad v_{N+1} = 0 \end{cases}$$

is in terms of the discrete function v, which—as soon as it can be solved for—will give an approximation of the actual solution u of the boundary value problem. Observe that the first line of the finite difference scheme gives N equations. The matrix form of these equations is:

$$\begin{bmatrix} -2 & 1 & & & \\ 1 & -2 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & -2 & 1 \\ & & & 1 & -2 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ \vdots \\ v_N \end{bmatrix} = -h^2 \begin{bmatrix} f_1 \\ f_2 \\ \vdots \\ \vdots \\ f_N \end{bmatrix}$$

which we may write more compactly as $\mathbf{Av} = \mathbf{b}$. Notice that we have taken advantage of the fact

that $v_0 = 0 = v_{N+1}$ in writing the matrix form. Decomposing $\mathbf{A}$ as indicated above gives,

$$\mathbf{D} + \mathbf{L} = \begin{bmatrix} -2 & & & & \\ 1 & -2 & & & \\ & \ddots & \ddots & & \\ & & 1 & -2 & \\ & & & 1 & -2 \end{bmatrix} \quad \text{and} \quad \mathbf{U} = \begin{bmatrix} 0 & 1 & & & \\ & 0 & 1 & & \\ & & \ddots & \ddots & \\ & & & 0 & 1 \\ & & & & 0 \end{bmatrix}$$

One can verify that,

$$-(\mathbf{D} + \mathbf{L})^{-1}\mathbf{U} = \begin{bmatrix} 0 & 2^{-1} & & & \\ 0 & 2^{-2} & 2^{-1} & & \\ \vdots & \vdots & \ddots & \ddots & \\ 0 & 2^{-(N-1)} & \cdots & 2^{-2} & 2^{-1} \\ 0 & 2^{-N} & \cdots & \cdots & 2^{-2} \end{bmatrix}$$

When N is even, the eigenvalues of this matrix are 0 and $\cos^2(\frac{j\pi}{N+1})$ for $j = 1, 2, \ldots, \frac{N}{2}$. Thus, the spectral radius is less than one, and so the norm $|| \left(\mathbf{D} + \mathbf{L}\right)^{-1}\mathbf{U}\big)^k ||_2$ will tend to 0 as $k \to \infty$.

Let's look at a specific example. Suppose now that $N = 40$, and further suppose that we are able to choose an initial vector $\mathbf{v}_0$ such that the initial error is bounded by 20. We may wish to know what error in our answer is possible while using the Gauss-Seidel method, given three-hundred steps of iteration. Determining that the spectral radius is 0.994, then by Lemma 3.1 we have that

$$0.994^k \leq || \left(\mathbf{D} + \mathbf{L}\right)^{-1}\mathbf{U}\big)^k ||_2$$

Thus, at $k = 300$ the bound for the error is at least $0.994^{300} \cdot 20 \approx 3.288$. As it turns out, the difference, $|| \left(\mathbf{D} + \mathbf{L}\right)^{-1}\mathbf{U}\big)^k ||_2 - 0.994^k$, never exceeds 0.0028 as $k \to \infty$, making 0.994^k a good choice for estimating the norm. While the matrix $(\mathbf{D} + \mathbf{L})^{-1}\mathbf{U}$ is nonnormal, it is "close" to normal in the sense that the norm of its power is close to the power of its spectral radius. The Kreiss lower and upper bounds on $\sup_{k \geq 0} || \left(\mathbf{D} + \mathbf{L}\right)^{-1}\mathbf{U}\big)^k ||_2$ are 1 and 108.73, respectively. The upper bound for $|| \left(\mathbf{D} + \mathbf{L}\right)^{-1}\mathbf{U}\big)^k ||_2$ based on Theorem 3.12 is approximately 1.666×10^{39}. Thus, the other bounds are not as useful in this example. Trefethen and Embree [7, p. 236] speculate that had the

matrix $(\mathbf{D} + \mathbf{L})^{-1}\mathbf{U}$ in the Poisson example exhibited stronger nonnormal characteristics, research on transient phases in norms of matrix powers may have begun as early as the 1950s.

4.3 Markov Chains

Markov Chains have been used extensively in probability theory to model mathematical systems which change over time; more specifically, systems which undergo a finite number of states, and the state at any one time can be predicted based on the state of the previous period. The transition matrix $\mathbf{P} = (p_{mn})$ of a Markov chain is a square matrix in which the element p_{mn} represents the probability that a system in state m at a specific time, will switch to a system in state n in the next period. A state vector $\mathbf{x}$ for an observation of a Markov process is a row vector in which the n^{th} component is the probability that the system is in the n^{th} state at the time of observation. Given an initial state vector $\mathbf{x}^{(0)}$, the state vector after k periods is given by,

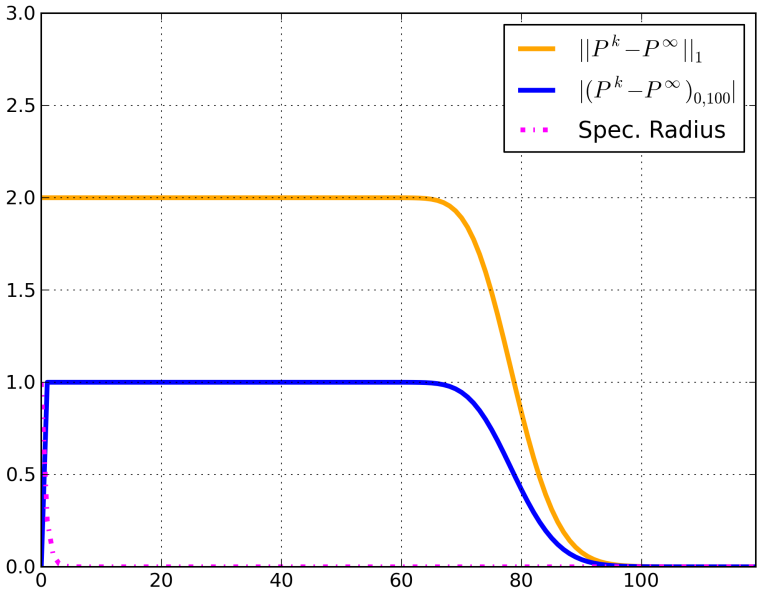
$$\mathbf{x}^k = \mathbf{x}^0\mathbf{P}^k$$

For example, imagine a line with one-hundred equally spaced nodes (counted left to right), onto which a particle is dropped. Given that it has landed on the 100^{th} node, suppose that the probability is 1 that the particle stays put in the next period. Given that the particle has landed on the 99^{th} node, suppose that the probabilities that the particle stays put, or moves one node to the right, are $\frac{1}{4}$ and $\frac{3}{4}$, respectively. Finally, given that the particle has landed on some node between the 1^{st} and the 98^{th}, suppose that the probabilities that the particle stays put, moves one node to the right, or moves two nodes to the right, are $\frac{1}{4}, \frac{1}{4}$, and $\frac{1}{2}$, respectively. The transition matrix for such a system is given by:

$$\mathbf{P} = \begin{bmatrix} \frac{1}{4} & \frac{1}{4} & \frac{1}{2} & & \\ & \ddots & \ddots & \ddots & \\ & & \frac{1}{4} & \frac{1}{4} & \frac{1}{2} \\ & & & \frac{1}{4} & \frac{3}{4} \\ & & & & 1 \end{bmatrix}$$

As it is only possible for the particle to stay put or move to the right, we may expect that given any

Figure 4.4: Nonnormality in Analysis of a Markov Chain



initial state vector, the limit of the Markov process will be a state vector in which all the entries are zero, except the last entry which is one:

$$\lim_{k\to\infty} \mathbf{x}^k = [0, 0, \ldots, 0, 1]$$

Taking the limit of powers of the transition matrix, $\lim_{k\to\infty} \mathbf{P}^k$, yields a matrix which we will call $\mathbf{P}^\infty$. The matrix $\mathbf{P}^\infty$ has a last column of all ones and all other entries equal to zero. This represents the reasonable expectation that if the experiment is run for many periods, the particle will be found at the 100^{th} node.

Now consider the idea of dropping many particles on the line. Given the previous conditions we can expect the particles to all converge to the right-most node over repeated observations. An interesting question lies in how quickly we can expect this to occur. One way to measure this idea is by taking the norm $||\mathbf{P}^k - \mathbf{P}^\infty||_\infty$ as k tends to ∞. Figure 4.4 shows the graph of this norm, along with that of the spectral radius, and a lower bound based on Theorem 3.8 using the entry from the 1^{th} row and 100^{th} column of the Schur form $\mathbf{T}$ of $\mathbf{P}^k - \mathbf{P}^\infty$. This bound hits a maximum of 1 at $k = 5$ and stays within $\frac{1}{2}$ of the actual norm from that point on. Perhaps even more importantly, the bound gets closer to the actual value of the norm as k increases. While the spectral radius is a

lower bound, it is not as useful because it goes so quickly to zero, without exhibiting the nonnormal behavior inherent in the graph of the norm.

As in the Gauss-Seidel example, the other bounds are less useful here as well. The Kreiss lower and upper bounds which we derived earlier apply to the spectral norm, and so we will not use them here. The bound based on Theorem 3.14 reaches 1.73×10^{35} at $k = 75$.

How can we interpret the graph? If we run our experiment for seventy periods or less, there are still likely to be many particles in transition to the right. As the number of periods increases up past one-hundred, we can expect most of the particles to have already arrived at the right-most node, with perhaps a few stragglers still in transition.

4.4 Random Matrices with ± 1 in the First Two Superdiagonals

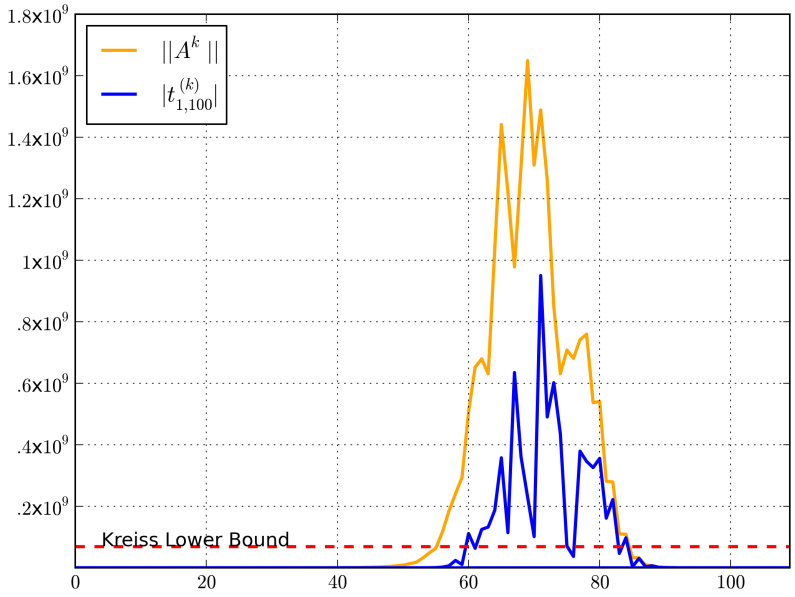
Random strictly upper-triangular matrices are nonnormal (with exception of the zero matrix), and thus it is impossible to unitarily diagonalize them. When taking the norm of their powers, we may run into significant transient effects before convergence; of course, we will still require that the spectral radius be less than one in order to obtain any type of convergence at all. Here, we look at a particular type of upper-triangular matrix in which the entries on the first two superdiagonals come from the $\{\pm 1\}$ distribution, with the probability of either value appearing in an entry equal to $\frac{1}{2}$. All of the other entries of the matrix will be set to 0.

$$\mathbf{A} = \begin{bmatrix} 0 & \pm 1 & \pm 1 & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & \pm 1 \\ & & & \ddots & \pm 1 \\ & & & & 0 \end{bmatrix}$$

Trefethen and Embree refer to this type of matrix as a random Fibonacci matrix [7, p. 351].

Right away, we recognize that such matrices have only one eigenvalue, 0. By Theorem 1.2, we may be certain that for powers $k \geq N$, where N is the matrix dimension, we will have $\mathbf{A}^k = \mathbf{0}$. The more interesting results happen for powers k with $0 < k < N$. Figure 4.5 shows the graph of one such random matrix, where $N = 100$. The norm of powers of $\mathbf{A}$ grows exceptionally large

Figure 4.5: A Random Matrix with ±1 in the first two Superdiagonals



for powers between 50 and 90. In fact, this norm peaks at about 1.65×10^9, when $k = 69$. Notice how jagged the graph is in contrast with the other examples; this is due to the sign changes, and the equal magnitude of the superdiagonal entries. The Kreiss lower and upper bounds for this graph are 6.92×10^7 and 1.88×10^{10}, respectively. Only the Kreiss lower bound makes it into the view of the graph as shown. The upper bound based on Theorem 3.12 reaches 1.16×10^{29}, and is also not shown in the figure. The lower bound based on Theorem 3.8 and the entry from the upper right corner of $\mathbf{A}^k$ provides the best approximation to the actual norm of all of the bounds considered—it peaks at about 0.95×10^9 when $k = 71$; it is rather interesting that one entry from $\mathbf{A}^k$ can account for so much of the transient effect in the graph of the norm.

Chapter 5. Summary

We have discussed several ways of representing a matrix power $\mathbf{A}^k$ where $\mathbf{A}$ is a square complex matrix and k is a nonnegative integer. If $\mathbf{A}$ is normal, the spectral decomposition (Theorem 1.4) allows us to decompose $\mathbf{A}$ into a product $\mathbf{U}\mathbf{D}^k\mathbf{U}^*$, where $\mathbf{U}$ is a unitary matrix and $\mathbf{D}$ is a diagonal matrix with the eigenvalues of $\mathbf{A}$ along its diagonal. If $\mathbf{A}$ is nonnormal, we can use the Schur decomposition (Theorem 1.5), which allows us to express $\mathbf{A}$ as the product $\mathbf{U}\mathbf{T}^k\mathbf{U}^*$, where $\mathbf{T}$ is an upper-triangular matrix with the eigenvalues of $\mathbf{A}$ along its diagonal, and $\mathbf{U}$ is again a unitary matrix.

A more in-depth investigation into the powers of the Schur form $\mathbf{T}$ of the matrix $\mathbf{A}$, leads to an expression of the mn^{th} entry $t_{mn}^{(k)}$ of the matrix $\mathbf{T}^k$ in terms of the entries of $\mathbf{T}$ (Theorem 2.4). In cases where the eigenvalues were either all equal or all distinct, we were able to substitute for part of the expression of $t_{mn}^{(k)}$, to get an expression in which the number of terms in the summation did not depend on k; these substitutions were based on Lemma 2.7 and Lemma 2.10, respectively. When possible in application, one should take advantage of these lemmas, rather than using the general expression in Theorem 2.4; these latter forms are easier to code and much faster to compute.

Theorem 2.4 allowed us to give a new proof that $\lim_{k\to\infty} \mathbf{A}^k = \mathbf{0}$ if and only if $\rho(\mathbf{A}) < 1$ (Lemma 3.3). We also used the expression of $t_{mn}^{(k)}$ from Theorem 2.4 to produce new lower (Theorem 3.8) and upper (Theorems 3.12, 3.13, and 3.14) bounds on norms of $\mathbf{A}^k$. In the case of the 2×2 matrix we were able to get an exact formula for the spectral norm of $\mathbf{T}^k$ in terms of the entries of $\mathbf{T}$ (Theorem 3.7).

Another representation of $\mathbf{A}^k$ came as an integral over a Jordan curve containing the ball centered at the origin of radius $\rho(\mathbf{A})$, and involving the resolvent of $\mathbf{A}$ (Theorem 2.16): $\frac{1}{2\pi i}\int_\Gamma z^k(z\mathbf{I} - \mathbf{A})^{-1}\,dz$. We used this expression, along with Spijker's Lemma (Lemma 3.21) to state and prove the most current form of the Kreiss Matrix Theorem (Theorem 3.17).

Armed with bounds new and old, we analyzed the four examples first introduced in Figure 1. We also looked at examples of where matrix norms are used in analysis of error in Gauss-Seidel iterations, in estimation of convergence time in Markov processes, and in observing transient effects of a random upper-triangular matrix. We saw that none of the bounds performs better than the others in all cases. We found that at times one bound may perform particularly well for all values

of k; while at other times this may be limited to a particular range of k, or in fact no bound may seem to work very well.

There are doubtlessly still many unanswered questions regarding how norms of matrix powers behave. We may have a high level of optimism that answers lie within reach, and they are certainly needed to advance our mathematical understanding. Matrices and their norms are used ubiquitously in fields of applied mathematics in building scientific models, and in measuring complex phenomenon. Therefore, better understanding of the properties of matrix powers, is very likely to lead to improved applications, which in turn will give us greater power to shape the world we live in.

Bibliography

[1] James W. Demmel. *Applied Numerical Linear Algebra*. SIAM, 3600 University City Science Center, Philadelphia, PA 19104–2688, 1997.

[2] Roger A. Horn and Charles R. Johnson. *Matrix Analysis*. Cambridge University Press, 40 West 20th Street, New York, NY 1001–4211, 1985.

[3] H.O. Kreiss. Uber die stabilitatsdefinition fur differenzengleichungen die partielle differentialgleichungen approximieren. *BIT*, 2:153–181, 1962.

[4] Randall J. LeVeque and Lloyd N. Trefethen. On the resolvent condition in the Kreiss matrix theorem. *BIT*, 24:584–591, 1984.

[5] E.B. Saff and A.D. Snider. Pearson Education Inc, Upper Saddle River, New Jersey 07458, 2003.

[6] M.N. Spijker. On a conjecture by LeVeque and Trefethen related to the Kreiss matrix theorem. *BIT*, 31:551–555, 1991.

[7] Lloyd N. Trefethen and Mark Embree. *Spectra and Pseudospectra: The Behavior of Nonnormal Matrices and Operators*. Princeton University Press, 41 William Street, Princeton, New Jersey 08540, 2005.