On Condition Numbers of Companion Matrices

by
Michael Cox

A project submitted to the Department of
Mathematics and Statistics of McMaster University in conformity with
the requirements for Master of Science in Mathematics

McMaster University
Hamilton, Ontario, Canada

Abstract

Given a square matrix A, the condition number is an important invariant that bounds the changes in the solutions of a system of linear of equations, when we make perturbations to the system. De Téran, Dopico, and Pérez [1] studied the condition numbers of a family of companion matrices called Fiedler companion matrices. In this project, we give new proofs for some of their results, using the description of Fiedler companion matrices due to Eastman, Kim, Shader, and Vander Meulen [2]. We also study the condition numbers of striped companion matrices, and we also show in some specific cases that some striped companion matrices will have smaller condition numbers than any Fiedler companion matrix. All of these results use the Frobenius norm to find the condition number. In our last chapter we investigate the condition number of various companion matrices using the spectral norm.

Acknowledgements

Foremost I would like to express my thanks to Dr. Adam Van Tuyl and Dr. Kevin Vander Meulen for their constant guidance and patience throughout this project. Thank you for always giving me the direction I needed, and for always being available for whatever questions and concerns I had. Your passion and enthusiasm for mathematics has been inspiring. Thank you to the Department of Mathematics at McMaster University for providing an outstanding learning environment throughout my graduate and undergraduate studies here. Thank you as well to McMaster University for you financial support over the last 16 months. Finally, thank you to my friends and family here in Hamilton. Thank you for your constant love and support.

Contents

Chapter 1

Introduction

Whether it be a calculator or a computer, any tool used in doing arithmetic calculations has a finite amount of memory to store decimal places of a number. In making these calculations, suppose the finite number of decimal places that can be stored is n. If we make calculations with numbers which exceed that number of decimal places, the computer will do one of two things. It will either truncate the rest of the digits after the n digits, i.e., it will omit them, or it will round the number to n digits. Whenever a computer does either of these things to a number there will be a *roundoff error*, which could have a severe effect on calculations. The more calculations done with this roundoff error, the more and more the error could accumulate, potentially making the calculations rather inaccurate.

In mathematics, it is important to have a good understanding of the roundoff error when performing calculations. In this paper we explore the condition number which has been developed to know what kind of inaccuracies to expect when doing calculations on a system of linear equations $A\vec{x} = \vec{b}$.

When the software MATLAB calculates the roots of polynomials, it uses matrix techniques to find the eigenvalues of a matrix. In particular, for some polynomial $p(x)$, MATLAB uses companion matrices to the polynomial $p(x)$. Roughly, a companion matrix of $p(x)$ is a matrix A where $p(x)$ is the characteristic polynomial of A and the coefficients of $p(x)$ appear in A.

Example 1.1. Consider the polynomial $p(x) = x^4 + a_3x^3 + a_2x^2 + a_1x + a_0$. Now consider the matrix of the form

$$A = \begin{bmatrix} -a_3 & -a_2 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & -a_1 & 0 & -a_0 \\ 0 & 1 & 0 & 0 \end{bmatrix}.$$

The characteristic polynomial $\det(xI_n - A) = x^4 + a_3x^3 + a_2x^2 + a_1x + a_0 = p(x)$. We say A is a companion matrix to $p(x)$.

In order to understand what kind of innaccuracies to expect when we perform calculations with a system of linear equations, we conceptualize the "size" of A using matrix norms. For the majority of this project we will use the *Frobenius norm*. The norm of a matrix A will aid us in computing the *condition number*, denoted $\kappa(A)$, which gives us an upper bound on how sensitive a linear system will be to small changes. The linear systems for this project will use companion matrices. More detailed explanations of these concepts are discussed later in Chapter 2.

One of the most well known examples of a companion matrix is the *Frobenius* companion matrix. For some monic polynomial $p(x) = x^n + a_{n-1}x^{n-1} + a_{n-2}x^{n-2} + \cdots + a_1x + a_0$, the Frobenius companion matrix is a companion matrix to the polynomial $p(x)$ of the form

$$\begin{bmatrix} 0 & 1 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 1 & \cdots & 0 & 0 \\ 0 & 0 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & 1 \\ -a_0 & -a_1 & -a_2 & \cdots & -a_{n-2} & -a_{n-1} \end{bmatrix}.$$

Unfortunately, as the size of the square matrix n goes to infinity, the Frobenius companion matrix becomes nearly singular, so calculations using the inverse of a Frobenius companion matrix can be unreliable [1].

In 2003, Fiedler studied a new type of companion matrix, now called *Fiedler* companion matrices [4]. Fiedler showed that you can create $n \times n$ companion matrices by multiplying n matrices from a particular class of matrices together, and each permutation of the integers $\{0, \ldots, n-1\}$ would give a different Fiedler companion matrix.

We illustrate with a very small example.

Example 1.2. Suppose that $p(x) = x^3 + a_2x^2 + a_1x + a_0$. Define

$$M_0 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -a_0 \end{bmatrix}, M_1 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & -a_1 & 1 \\ 0 & 1 & 0 \end{bmatrix}, M_2 = \begin{bmatrix} -a_2 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

See chapter 2 for the construction of these matrices. Let $\sigma = (0, 2, 1)$. Then

$$M_\sigma = M_0M_2M_1 = \begin{bmatrix} -a_2 & -a_1 & 1 \\ 1 & 0 & 0 \\ 0 & -a_0 & 0 \end{bmatrix}.$$

The characteristic polynomial of M_σ is $x^3 + a_2x^2 + a_1x + a_0$.

De Téran, Dopico, and Pérez [1] extensively studied the condition numbers of Fiedler companion matrices in 2012. They discovered many results about upper and lower bounds of the condition numbers of Fiedler companion matrices, and were able to create explicit formulas for the norms of these matrices, as well as their inverses. They developed an ordering for all Fiedler matrices of a particular polynomial, according to increasing condition numbers, which also provides upper and lower bounds for the ratio of the condition numbers of any pair of Fiedler companion matrices.

A companion matrix A to a polynomial $p(x)$ can take various forms, as long as $\det(xI - A) = p(x)$. Eastman and Vander Meulen [3] gave a description of all companion matrices using a *lower Hessenberg form* of a companion matrix. An $n \times n$ matrix A is a sparse companion matrix of $p(x)$ is equivalent to a lower Hessenberg matrix. This representation of a companion matrix is the one that is used in the majority of this project. Fiedler companion matrices in lower Hessenberg form have the property that the coefficients of the characteristic polynomial $p(x)$ form a lattice path from the bottom left corner of the companion matrix to the main diagonal. We called the length of the first leg of the lattice path out of the bottom left corner of a Fiedler companion matrix in lower Hessenberg from the *initial step value*. We used lower Hessenberg form to give new proofs of some of De Téran et. al.'s results. One of the results is the structure of the inverse of a Fiedler companion matrix. The theorem is as follows:

Theorem 1.3 (Theorem 3.7)**.** *Let $p(x) = x^n + a_{n-1}x^{n-1} + a_{n-2}x^{n-2} + \cdots + a_1x + a_0$ be a polynomial over $\mathbb{R}$ with $n \geq 2$ and $a_0 \neq 0$. Let M be a Fiedler companion matrix to the polynomial $p(x)$ in lower Hessenberg form. Let t be the initial step value of M. Then:*

(a) M^{-1} has $t+1$ entries equal $-\frac{1}{a_0}, -\frac{a_1}{a_0}, \ldots, -\frac{a_t}{a_0}$, with exactly one copy of each;

(b) M^{-1} has $n-1-t$ entries equal to $a_{t+1}, a_{t+2}, \ldots, a_{n-1}$, with exactly one copy of each;

(c) M^{-1} has $n-1$ entries equal to 1; and

(d) the rest of the entries of M^{-1} are 0.

Inspired by the work of De Téran et. al. the main problem of this project is

Question 1.4. *What classes of companion matrices have smaller condition numbers than the Fiedler companion matrices?*

Ideally we would like to be able to compare the condition numbers of general companion matrices to Fiedler companion matrices. We focus mainly on one particular family of companion matrices known as *striped companion matrices.* A striped companion matrix A in lower Hessenberg form to a ploynomial $p(x) = x^n + a_{n-1}x^{n-1} + \cdots + a_1x + a_0$, has the property that the coefficients $-a_0, -a_1, \ldots, -a_{n-1}$ form horizontal stripes. We found some specific cases in which a striped companion matrix U to a polynomial $p(x)$ will always have a better condition number than any Fiedler companion matrix M to the same polynomial $p(x)$. We first discuss striped companion matrices where all of the stripes are the same size, and then moved on to the case where the stripes are not all the same size. One of our main theorems is for the case where the stripes are not the same size:

Theorem 1.5 (Theorem 4.9)**.** *Consider the monic polynomial:*

$$p(x) = x^n + a_{n-1}x^{n-1} + a_{n-1}x^{n-2} + \cdots + a_1x + 1,$$

where n is any positive integer. Consider any striped companion matrix $U = C_n(\boldsymbol{s})$ where $\boldsymbol{s} = (s_1, s_2, \ldots, s_r)$, $s_r < s_i \leq s_1$ for all $i \in \{2, \ldots, r-1\}$. Let a_{ℓ_j} for $j \in \{1, \ldots, r-1\}$

be the the nonzero coefficients of the polynomial $p(x)$ that fall in the column vector $\vec{u}$ of the matrix

$$U = \left[\begin{array}{c|c|c} \vec{0} & I_m & 0 \\ \hline \vec{u} & H & I_{n-m-1} \\ \hline -a_0 & \vec{y}^T & \vec{0}^T \end{array}\right]$$

and assume that each nonzero stripe has a nonzero entry in $\vec{u}$. If M is any lower Hessenberg Fiedler companion matrix to the polynomial $p(x)$ with initial step value $t = s_r$, then U will satisfy $\kappa_F(U) \leq \kappa_F(M)$ if the entries of U satisfy:

$$|a_k a_{\ell_j} - a_{k+\ell_j}| \leq |a_{k+\ell_j}| \text{ for } k \in \{1, \ldots, t\}.$$

In the second chapter we introduce all of the necessary definitions for the project. In chapter three, we reprove some of the results found by De Téran et. al. using the lower Hessenberg matrix form. For chapter four, we compare the condition numbers of striped companion matrices and Fiedler companion matrices. We discuss some cases, and the necessary condtions, in which the striped companion matrix will always have a better condition number than the Fiedler matrix. Finally, in the last chapter we briefly investigate the condition number of various companion matrices using singular values and another matrix norm called the *spectral norm*. In the last chapter of this project we revist some of the results discussed in De Téran et. al.'s paper [1], and some ways to find the singular values of a companion matrix A needed to find its spectral condition number.

Chapter 2

Background

This chapter will describe some relevant information from linear algebra and matrix analysis that is required for the project. First, we will discuss companion matrices and some of their properties. Second, we will talk about matrix norms, along with some of their properties and some specific examples. In the same section we present the idea of a condition number which uses matrix norms in their definition. Finally, we will describe the condition number for inversion with respect to the Frobenius norm.

2.1 Companion Matrices

The software MATLAB uses companion matrices to derive the roots of polynomials by applying matrix techniques used for finding the eigenvalues of a matrix [1]. In this section we define and discuss companion matrices, along with some of the different forms they can take, and their properties.

Definition 2.1. Let A be an $n \times n$ square matrix. The *characteristic polynomial* of A, denoted $p_A(x)$, is

$$p_A(x) = \det(xI_n - A).$$

For any monic polynomial, one can find a specific type of matrix in which all the coefficients of the polynomial are entries in the matrix, and furthermore, the monic polynomial is the characteristic polynomial to the matrix A.

Definition 2.2. Given $p(x) = x^n + a_{n-1}x^{n-1} + a_{n-2}x^{n-2} + \cdots + a_1x + a_0$, a *companion matrix* to $p(x)$ is an $n \times n$ matrix A over $\mathbb{R}[a_0, a_1, \ldots, a_{n-1}]$ such that the characteristic polynomial of A is $p(x)$. A companion matrix with entries $a_0, a_1, \ldots, a_{n-1}$, with $n - 1$ entries equal to one, and $n^2 - 2n + 1$ zero entries is called a *sparse companion matrix.*

When discussing the structure of a companion matrix C, we treat the a_i entries as formal variables, not real numbers. We say that A is a *realization* of a companion matrix C if A is obtained from C by replacing each of the variable entries by real numbers. We will often refer to a realization A as a companion matrix, but the context will make clear if A is a realization. In particular, in these cases, we will see that the coefficients are in $\mathbb{R}$.

Example 2.3. Consider the polynomial $p(x) = x^5 + a_4x^4 + a_3x^3 + a_2x^2 + a_1x + a_0$. Next consider a matrix of the form

$$A = \begin{bmatrix} -a_4 & -a_3 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & -a_2 & 0 & -a_1 & -a_0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix}.$$

Since $\det(xI - A) = p(x)$, A is a companion matrix. In fact, A is a sparse companion matrix.

Example 2.4. A companion matrix to $p(x)$ is not unique. For example, consider again the same polynomial $p(x) = x^5 + a_4x^4 + a_3x^3 + a_2x^2 + a_1x + +a_0$, but let

$$B = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & -a_4 & 1 & 0 \\ 0 & 0 & -a_3 & 0 & 1 \\ -a_0 & -a_1 & -a_2 & 0 & 0 \end{bmatrix}.$$

Since $\det(xI - B) = p(x)$, then B is also a companion matrix to $p(x)$.

Example 2.5. A matrix of the form

$$M = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ -a_3 & -a_4 & -a_5 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ -a_0 & -a_1 & -a_2 & 0 & 0 & 0 \end{bmatrix}.$$

has characteristic polynomial

$$p(x) = x^6 + a_5x^5 + a_4x^4 + a_3x^3 + a_2x^2 + a_1x + +a_0.$$

Note that M is a sparse companion matrix to $p(x)$.

Example 2.6. The matrix of the form

$$D = \begin{bmatrix} 0 & 0 & 1 & -a_1 & -a_0 \\ 0 & 0 & 1 & 0 & 0 \\ 1 & -1 & 0 & -a_2 & 0 \\ 0 & 1 & 0 & -a_4 & -a_3 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix}$$

has characteristic polynomial

$$p(x) = x^5 + a_4x^4 + a_3x^3 + a_1x + +a_0.$$

So D is a companion matrix to $p(x)$, but it is not sparse as it does not have $5^2 - 2(5) + 1 = 16$ nonzero entries.

For any monic polynomial, the following companion matrix is probably the most well-known example.

Definition 2.7. Given $p(x) = x^n + a_{n-1}x^{n-1} + a_{n-2}x^{n-2} + \cdots + a_1x + a_0$, the *Frobenius companion matrix* is a companion matrix to the polynomial $p(x)$ of the form

$$\begin{bmatrix} 0 & 1 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 1 & \cdots & 0 & 0 \\ 0 & 0 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & 1 \\ -a_0 & -a_1 & -a_2 & \cdots & -a_{n-2} & -a_{n-1} \end{bmatrix}.$$

Example 2.8. Consider the polynomial

$$p(x) = x^3 + a_2x^2 + a_1x + +a_0.$$

The matrix of the form

$$C = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ -a_0 & -a_1 & -a_2 \end{bmatrix}$$

is a Frobenius companion matrix to $p(x)$.

We now work towards a theorem characterizing sparse companion matrices.

Definition 2.9. A *permutation matrix* is an $n \times n$ matrix obtained by permuting the rows of the $n \times n$ identity matrix, and hence, $P^T = P^{-1}$.

If P is a permutation matrix, then $P^TP = I$, thus, if A is any matrix, then PAP^T has the same characteristic polynomial as A since PAP^T and A are similar matrices. So if A is a companion matrix to a polynomial $p(x)$, then so is PAP^T for any permutation matrix P. We say two matrices A and B are *equivalent* if there exists a permutation matrix P such that $B = PAP^T$.

Example 2.10. Consider the companion matrix C from Example 2.8, and the following permutation matrix

$$P = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix}. \text{ Then } P^{-1} = P^T = P.$$

We have that

$$B = PCP^T = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ -a_0 & -a_1 & -a_2 \end{bmatrix} \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix}$$

$$B = \begin{bmatrix} -a_2 & -a_1 & -a_0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}.$$

So C and B are equivalent.

Definition 2.11. Consider the $n \times n$ matrix:

$$A = \begin{bmatrix} a_{1,1} & a_{1,2} & a_{1,3} & a_{1,4} & \cdots & a_{1,n} \\ a_{2,1} & a_{2,2} & a_{2,3} & a_{2,4} & \cdots & a_{2,n} \\ a_{3,1} & a_{3,2} & a_{3,3} & a_{3,4} & \cdots & a_{3,n} \\ a_{4,1} & a_{4,2} & a_{4,3} & a_{4,4} & \cdots & a_{4,n} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ a_{n,1} & a_{n,2} & a_{n,3} & a_{n,4} & \cdots & a_{n,n} \end{bmatrix}.$$

The ith *subdiagonal* of A is the set of entries

$$\{a_{i+j-1,j} \mid j = 1, \dots, n-i\}.$$

The jth *superdiagonal* a of A is the set of entries

$$\{a_{i,j+i-1} \mid i = 1, \dots, n-j\}.$$

The *main diagonal* of A is the set of entries

$$\{a_{\ell,\ell} \mid \ell = 1, \dots, n\}.$$

Definition 2.12. A *lower Hessenberg matrix* $A = [a_{ij}]$ is a matrix that has zeros above the first superdiaganoal, that is, $a_{ij} = 0$ when $j > i+1$. A *lower Hessenberg companion matrix* is a lower Hessenberg matrix that is also a companion matrix.

The matrix in Example 2.4 is a lower Hessenberg companion matrix but the matrix in Example 2.3 is not. However for the remainder of this project we will primarily assume that our companion matrices are in lower Hessenberg form. The next theorem justifies why we can make this assumption. As we shall see, many properties of companion matrices are easily observable in this form. The next theorem characterizes the structure of the sparse companion matrices. This structure will be helpful when proving results about the condition numbers of sparse companion matrices.

Theorem 2.13. [2, Corollary 4.3] *Let $p(x) = x^n + a_{n-1}x^{n-1} + a_{n-2}x^{n-2} + \cdots + a_1 x + a_0$. An $n \times n$ matrix A is a sparse companion matrix of $p(x)$ if and only if there is a permutation matrix P such that PAP^T is equal to a lower Hessenberg matrix of the form:*

$$\left[\begin{array}{c|c} \vec{0} \;\; I_m & 0 \\ \hline R & \begin{matrix} I_{n-m-1} \\ \vec{0}^T \end{matrix} \end{array}\right] \tag{2.1}$$

where R is an $(n-m) \times (m+1)$ matrix with $-a_{n-1}$ in the top right corner, $-a_0$ in the bottom left corner of the matrix, $-a_{n-k}$ on the kth subdiagonal, and 0's elsewhere.

Example 2.14. Let $p(x) = x^5 + a_4x^4 + a_3x^3 + a_2x^2 + a_1x + a_0$. Consider the matrix of the form

$$A = \left[\begin{array}{cccc|c} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ \cline{1-4} 0 & 0 & -a_3 & -a_4 & 1 \\ -a_0 & -a_1 & -a_2 & 0 & 0 \end{array}\right].$$

We see that A satisfies all the conditions of Theorem 2.13, and hence A is a lower Hessenberg companion matrix to $p(x)$.

Example 2.15. Let $p(x) = x^5 + a_4x^4 + a_3x^3 + a_2x^2 + a_1x + a_0$, and consider the matrix of the form

$$M = \begin{bmatrix} 0 & 0 & -a_1 & 0 & -a_0 \\ 1 & 0 & 0 & -a_3 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & -a_4 & -a_2 \\ 0 & 0 & 1 & 0 & 0 \end{bmatrix}.$$

Since $\det(xI - M) = p(x)$, M is sparse companion matrix. Thus, PMP^T has the form described in Theorem 2.13 for some permutation matrix P. Eastman and Vander Meulen described an algorithm in [3] (see Algorithm 7.1) to determine an appropriate permutation matrix P. Indeed, taking

$$P = \begin{bmatrix} 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \end{bmatrix}, \text{ then}$$

$$PMP^T = \left[\begin{array}{ccc|cc} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ \hline -a_2 & 0 & -a_4 & 1 & 0 \\ 0 & 0 & -a_3 & 0 & 1 \\ -a_0 & -a_1 & 0 & 0 & 0 \end{array}\right].$$

In [4], Fiedler descibed a way to construct companion matrices for $p(x) = x^n + a_{n-1}x^{n-1} + \cdots + a_0$ using particular matrix products.

Definition 2.16. For $k \in \{1, 2, \ldots, n-1\}$, let

$$M_0 = \begin{bmatrix} I_{n-1} & 0 \\ 0 & -a_0 \end{bmatrix} \quad \text{and} \quad M_k = \begin{bmatrix} I_{n-k-1} & 0 & 0 \\ 0 & \begin{bmatrix} -a_k & 1 \\ 1 & 0 \end{bmatrix} & 0 \\ 0 & 0 & I_{k-1} \end{bmatrix}. \tag{2.2}$$

If $\sigma^{-1} = (i_1, \ldots, i_n)$ is any permutation of the integers $\{0, \ldots, n-1\}$, then the product of $M_\sigma = M_{i_1}M_{i_2}\ldots M_{i_n}$ is a companion matrix of $p(x) = x^n + a_{n-1}x^{n-1} + a_{n-2}x^{n-2} + \cdots + a_1x + a_0$. If a matrix is equivalent to one of these products, then it is called a *Fiedler companion matrix.*

Example 2.17. We show that the matrix in Example 2.3 is, in fact, a Fiedler companion matrix. Let

$$M_0 = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & -a_0 \end{bmatrix}, M_1 = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & -a_1 & 1 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix}, M_2 = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & -a_2 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix},$$

$$M_3 = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & -a_3 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}, \text{ and } M_4 = \begin{bmatrix} -a_4 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}.$$

Let $\sigma = (4, 2, 1, 0, 3)$. The corresponding matrix product $M_\sigma = M_4M_2M_1M_0M_3$ gives us the matrix in Example 2.3:

$$M_\sigma = M_4M_2M_1M_0M_3 = \begin{bmatrix} -a_4 & -a_3 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & -a_2 & 0 & -a_1 & -a_0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix}.$$

Therefore the matrix A of Example 2.3 is a Fiedler companion matrix.

Theorem 2.13 implies that any Fiedler companion matrix is equivelant to a lower Hessenberg companion matrix. In the next result we describe the lower Hessenberg form of the Fiedler companion matrices as characterized by Eastman and and Vander Meulen in [3].

Definition 2.18. Suppose that for all $k \in \{0, \ldots, n-2\}$, if $-a_k$ is in position (i, j) of a matrix A, then a_{k+1} is either in position $(i-1, j)$ or $(i, j+1)$. In this case, the entries $a_0, a_1, \ldots, a_{n-1}$ are said to form a *lattice path* in A.

The following theorem shows us that the entries $a_0, a_1, \ldots, a_{n-1}$ in every lower Hessenberg representation of a Fiedler companion matrix forms a lattice path.

Theorem 2.19. [2, Corollary 4.4] *A matrix M is an $n \times n$ Fiedler companion matrix if and only if M is equivalent to a lower Hessenberg matrix as in Theorem 2.13, such that the nonzero entries of R form a lattice path from the bottom left corner to the upper right corner of R.*

Example 2.20. We again consider the matrix from Example 2.3:

$$M_\sigma = \begin{bmatrix} -a_4 & -a_3 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & -a_2 & 0 & -a_1 & -a_0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix}.$$

We saw that we could use Algorithm 7.1 to find the appropriate P matrix to find an equivalent matrix in lower Hessenberg form. In particular, with

$$P = \begin{bmatrix} 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \end{bmatrix}, \quad \text{we have}$$

$$PM_\sigma P^T = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & \boxed{-a_3} & \boxed{-a_4} & 1 \\ \boxed{-a_0} & \boxed{-a_1} & \boxed{-a_2} & 0 & 0 \end{bmatrix}.$$

Notice the lattice path, highlighted with a box.

Example 2.21. Let $p(x) = x^5 + a_4x^4 + a_3x^3 + a_2x^2 + a_1x + a_0$. Consider the following companion matrix of the form

$$A = \left[\begin{array}{ccc|cc} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ \hline 0 & -a_3 & -a_4 & 1 & 0 \\ -a_1 & 0 & 0 & 0 & 1 \\ -a_0 & 0 & -a_2 & 0 & 0 \end{array}\right].$$

We know that this is a companion matrix from Theorem 2.13. However we see that the a_i's do not form a lattice path in A. So A is not a Fiedler companion matrix.

De Terán et. al [1] have introduced some definitions related to the product construction of the Fiedler matrices.

Definition 2.22. Let σ be a permutation of $\{0, \dots, n-1\}$.

- For $i = 0, \dots, n-2$, we say that σ has a *consecution* at i if $\sigma(i) < \sigma(i+1)$ and that σ has an *inversion* if $\sigma(i) > \sigma(i+1)$.
- The *consecution-inversion structure sequence* of σ, denoted CISS(σ), is the tuple $(c_0, j_0, c_1, j_1, \dots, c_\ell, j_\ell)$, where σ has c_0 consecutive consecutions at $0, \dots, c_0 - 1$. Next, it has j_0 consecutive inversions from $c_0, c_0+1, \dots, c_0+j_0-1$, and so on.
- The *reduced consecution-inversion structure sequence* of σ is the sequence obtained from CISS(σ) after removing the 0 entries.
- the number of initial consecutions or inversions of σ, denoted t_σ, is given by:

$$t_\sigma = \begin{cases} c_0 & \text{if } c_0 \neq 0 \\ j_0 & \text{if } c_0 = 0. \end{cases}$$

Example 2.23. Returning to Example 2.17, we used the permutation $\sigma = (4, 2, 1, 0, 3)$. That is

$$\sigma^{-1}(0) = 4, \quad \sigma^{-1}(1) = 2, \quad \sigma^{-1}(2) = 1, \quad \sigma^{-1}(3) = 0, \quad \sigma^{-1}(4) = 3.$$

So from Definition 2.22, there is an inversion at $i = 3$, because $\sigma(3) = 4 > \sigma(4) = 0$, and we have inversions at $i = 0, 1$ because:

$$\begin{aligned} \sigma(0) = 3 &> \sigma(1) = 2 \\ \sigma(1) = 2 &> \sigma(2) = 1 \\ \sigma(2) = 1 &> \sigma(3) = 4. \end{aligned}$$

So CISS$(\sigma) = (0, 2, 1, 0)$, and the reduced CISS$(\sigma) = (2, 1)$. Furthermore, we get that $t_\sigma = 2$.

As noted in Example 2.17, M_σ is equal to

$$\begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & -a_3 & -a_4 & 1 \\ -a_0 & -a_1 & -a_2 & 0 & 0 \end{bmatrix}.$$

As observed in [3], we can view CISS(σ) as a set of instructions on how to use right movements, and upwards movements, to traverse through the lattice path of the lower Hessenberg form of a Fiedler companion matrix. We can think of the nonzero numbers as the number of movements to make, either right or up, before you make a turn. For example, since CISS$(\sigma) = (0, 2, 1, 0)$ and the first number of the CISS(σ) is a zero then you start moving right. Two movements right before you turn upward, then one movement up before you turn right. Since the last number is zero.

Definition 2.24. Let $p(x) = x^n + a_{n-1}x^{n-1} + a_{n-2}x^{n-2} + \cdots + a_1 x + a_0$. If M is a Fiedler companion matrix in lower Hessenberg form to the polynomial $p(x)$, then the *initial step size* is the maximum number of nonzero entries in the left most column of M, or the maximum number of nonzero entries in the bottom row of M.

Example 2.25. The matrix of the form

$$M = \left[\begin{array}{ccccc} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & -a_3 & -a_4 & 1 \\ \multicolumn{3}{c}{\boxed{-a_0 \;\; -a_1 \;\; -a_2}} & 0 & 0 \end{array}\right]$$

has intial step size of 3.

We have already seen that M has the structure described in Theorem 2.13. We can further break down the matrix R in the following way:

$$R = \begin{bmatrix} \vec{u} & H \\ -a_0 & \vec{y}^T \end{bmatrix} = \left[\begin{array}{c|ccc} 0 & 0 & -a_3 & -a_4 \\ \hline -a_0 & -a_1 & -a_2 & 0 \end{array}\right].$$

This construction of the matrix R will help us to study sparse companion matrices. It is especially useful in finding the inverse of a sparse companion matrix in lower Hessenberg form. We can use the following lemma due to Vander Meulen and Vanderwoerd in line (13) of [9].

Lemma 2.26. [9] *Let $p(x) = x^n + a_{n-1}x^{n-1} + a_{n-2}x^{n-2} + \cdots + a_1 x + a_0$ be a polynomial over $\mathbb{R}$. Suppose that A is a companion matrix in lower Hessenberg form to $p(x)$. Then let*

$$A = \left[\begin{array}{c|c|c} \vec{0} & I_m & 0 \\ \hline \vec{u} & H & I_{n-m-1} \\ \hline -a_0 & \vec{y}^T & \vec{0}^T \end{array}\right]. \tag{2.3}$$

where $a_0 \neq 0$. Then

$$A^{-1} = \left[\begin{array}{c|c|c} \frac{1}{a_0}\vec{y}^T & \vec{0}^T & -\frac{1}{a_0} \\ \hline I_m & 0 & \vec{0} \\ \hline -\frac{1}{a_0}\vec{u}\vec{y}^T - H & I_{n-m-1} & \frac{1}{a_0}\vec{u} \end{array}\right]. \tag{2.4}$$

Proof. Let

$$A = \left[\begin{array}{c|c|c} \vec{0} & I_m & 0 \\ \hline \vec{u} & H & I_{n-m-1} \\ \hline -a_0 & \vec{y}^T & \vec{0}^T \end{array}\right] \text{ and } B = \left[\begin{array}{c|c|c} \frac{1}{a_0}\vec{y}^T & \vec{0}^T & -\frac{1}{a_0} \\ \hline I_m & 0 & \vec{0} \\ \hline -\frac{1}{a_0}\vec{u}\vec{y}^T - H & I_{n-m-1} & \frac{1}{a_0}\vec{u} \end{array}\right].$$

It a known result that if $AB = I_n$, then $AB = I_n = BA$. So it suffices to show $AB = I_n$. We can in fact consider both of these matrices as both 3×3 block matrices; multiplying A and B together gives

$$\begin{aligned} AB &= \left[\begin{array}{c|c|c} \vec{0} & I_m & 0 \\ \hline \vec{u} & H & I_{n-m-1} \\ \hline -a_0 & \vec{y}^T & \vec{0}^T \end{array}\right] \left[\begin{array}{c|c|c} \frac{1}{a_0}\vec{y}^T & \vec{0}^T & -\frac{1}{a_0} \\ \hline I_m & 0 & \vec{0} \\ \hline -\frac{1}{a_0}\vec{u}\vec{y}^T - H & I_{n-m-1} & \frac{1}{a_0}\vec{u} \end{array}\right] \\ &= \left[\begin{array}{c|c|c} I_m & 0 & 0 \\ \hline \frac{1}{a_0}\vec{u}\vec{y}^T - H + -\frac{1}{a_0}\vec{u}\vec{y}^T - H & I_{n-m-1} & -\frac{1}{a_0}\vec{u} + \frac{1}{a_0}\vec{u} \\ \hline -\vec{y}^T + \vec{y}^T & 0 & 1 \end{array}\right] \\ &= I_n. \end{aligned}$$

□

One other type of sparse companion matrix we will consider is the *striped* companion matrix introduced in [2]:

Definition 2.27. Let $\mathbf{s} = (s_1, s_2, \ldots, s_r)$ be an ordered r-tuple of positive integers that sum to n, with $s_1 \geq s_i$, for $2 \leq i \leq r$. The *striped* companion matrix $C_n(\mathbf{s})$ is the sparse companion matrix in lower Hessenberg form, with an $(n - s_1 + 1) \times (s_1)$ matrix R having r nonzero rows with the ith nonzero row of R having s_i entries in the first s_i positions for $1 \leq i \leq r$, and $s_{i+1} - 1$ rows of zeros immediately below it in R, $1 \leq i \leq r - 1$.

Example 2.28. Let $p(x) = x^6 + a_5x^5 + a_4x^4 + a_3x^3 + a_2x^2 + a_1x + a_0$, and consider the following 6×6 sparse companion matrices to $p(x)$ of the form:

$$C_6(2,2,2) = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ -a_4 & -a_5 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ -a_2 & -a_3 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ -a_0 & -a_1 & 0 & 0 & 0 & 0 \end{bmatrix} \text{ and } C_6(3,3) = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ -a_3 & -a_4 & -a_5 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ -a_0 & -a_1 & -a_2 & 0 & 0 & 0 \end{bmatrix}.$$

Note that the 'stripes' in a striped companion matrix do not always have the same number of nonzero entries, as in the example:

$$C_7(3,2,2) = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ -a_4 & -a_5 & -a_6 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ -a_2 & -a_3 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ -a_0 & -a_1 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}.$$

Note that each a_i falls on the appropriate subdiagonal, as required by Theorem 2.13.

2.2 Matrix Norms and Condition Numbers

A matrix norm is a way to measure the "size" of a matrix. We recall the definition given by Poole [8]:

Definition 2.29. A *matrix norm* on the set of $n \times n$ matrices $M_{n\times n}$ is a mapping that associates each $A \in M_{n\times n}$ with a real number $||A||$, called the *norm* of A, such that the following properties are satisfied for all $A, B \in M_{n\times n}$ and all scalars k:

- $||A|| = 0$ if and only if $A = 0$, and otherwise $||A|| > 0$;
- $||kA|| = |k| \cdot ||A||$;
- $||A + B|| \leq ||A|| + ||B||$; and
- $||AB|| \leq ||A|| \cdot ||B||$.

Definition 2.30. Let V be a vector space over a field $\mathbb{R}$. A function $||\cdot|| : V \to [0, +\infty)$ is a *vector norm* if for all $\vec{x}, \vec{y} \in V$,

- $||\vec{x}|| = 0$ if and only if $\vec{x} = 0$ and otherwise $||\vec{x}|| > 0$;
- $||k\vec{x}|| = |k| \cdot ||\vec{x}||$ for all scalars $k \in \mathbb{F}$; and
- $||\vec{x} + \vec{y}|| \leq ||\vec{x}|| + ||\vec{y}||$.

Definition 2.31. A matrix norm on $M_{n\times n}$ is said to be *compatible* with a vector norm $||\vec{x}||$ on $\mathbb{R}^n$ if, for all $n \times n$ matrices A and all vectors $\vec{x} \in \mathbb{R}^n$, we have

$$||A\vec{x}|| \leq ||A|| \cdot ||\vec{x}||.$$

Definition 2.32. Let $\vec{x} = (x_1, \cdots, x_n)$ be a vector in $\mathbb{R}^n$. The Euclidean norm for a vector is given by

$$||\vec{x}||_2 = \sqrt{x_1^2 + \cdots + x_n^2}.$$

If we change the vector space from $\mathbb{R}^n$ to $\mathbb{R}^{n\times n}$ we can define a similar norm.

Definition 2.33. Let $A = [a_{ij}]$ be a matrix in the set of all $n \times n$ matrices denoted $\mathbb{R}^{n\times n}$. The *Frobenius norm* of A is given by

$$||A||_F = \sqrt{\sum_{ij} a_{ij}^2}.$$

Note that if the matrix A was not square, the Frobenius norm could still be defined. For the sake of this paper we work with square matries, as companion matrices are square.

Lemma 2.34. *The Frobenius matrix norm $||\cdot||_F$ is compatible with the Euclidean vector norm $||\cdot||_2$. That is, for a matrix A, and a vector $\vec{x} \in \mathbb{R}^n$, we have*

$$||A\vec{x}||_2 \leq ||A||_F \cdot ||\vec{x}||_2.$$

Proof. It suffices to show that $||A\vec{x}||_2^2 \leq ||A||_F^2 \cdot ||\vec{x}||_2^2$. Since $A\vec{x}$ is a column vector by definition, we have that $[A\vec{x}]_i = \sum_{j=1}^n a_{ij}x_j$. So

$$||A\vec{x}||_2^2 = \sum_{i=1}^n \left(\sum_{j=1}^n a_{ij}x_j \right)^2.$$

But by the *Cauchy-Schwarz Inequality* [5, Theorem 5.1.4], we have that

$$\begin{aligned} \left(\sum_{j=1}^n a_{ij}x_j \right)^2 &\leq \left(\sum_{j=1}^n a_{ij}^2 \right) \left(\sum_{j=1}^n x_j^2 \right) \\ &= \left(\sum_{j=1}^n a_{ij}^2 \right) \cdot ||\vec{x}||_2^2. \end{aligned}$$

And so

$$||A\vec{x}||_2^2 \leq \sum_{i=1}^{n} \left(\sum_{j=1}^{n} a_{ij}^2 \right) \cdot ||\vec{x}||_2^2 = ||A||_F^2 \cdot ||\vec{x}||_2^2$$

□

Consider the linear system of equations $A\vec{x} = \vec{b}$. If $\vec{b}$ is nonzero and A is nonsingular, then there exists a unique solution $\vec{x}$, which is also nonzero. Consider the perturbed system $A\hat{x} = \vec{b} + \delta\vec{b}$ where we added a small vector to $\vec{b}$. Our hope is that the unique solution $\hat{x}$, to $A\hat{x} = \vec{b} + \delta\vec{b}$ will be close to $\vec{x}$. Let $\hat{x} = \vec{x} + \delta\vec{x}$. We would hope that if $\delta\vec{b}$ is small, then so is $\delta\vec{x}$. We quantify these relative terms of small by using a vector norm. The relative size of $\delta\vec{b}$ to $\vec{b}$ is then given by $||\delta\vec{b}||/||\vec{b}||$, and the relative size of $\delta\vec{x}$ to $\vec{x}$ is then given by $||\delta\vec{x}||/||\vec{x}||$. When $||\delta\vec{b}||/||\vec{b}||$ is small, we hope that $||\delta\vec{x}||/||\vec{x}||$ is as well. This may not happen, as we see in the next example.

Example 2.35. Suppose that we wanted to solve to following system of equations:

$$\begin{bmatrix} 1000 & 999 \\ 999 & 998 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1999 \\ 1997 \end{bmatrix}.$$

If we expand out the system we get:

$$\begin{aligned} 1000x_1 + 999x_2 &= 1999 \\ 999x_1 + 998x_2 &= 1997, \end{aligned}$$

and it is clear by observation that the unique solution to this system is given by $\vec{x} = [1 \ \ 1]^T$. Next consider the slightly perterbed system

$$\begin{bmatrix} 1000 & 999 \\ 999 & 998 \end{bmatrix} \begin{bmatrix} \hat{x_1} \\ \hat{x_2} \end{bmatrix} = \begin{bmatrix} 1998.99 \\ 1997.01 \end{bmatrix}.$$

So we have a new system

$$A\hat{x} = \vec{b} + \delta b, \quad \text{where} \ \ \delta\vec{b} = \begin{bmatrix} 0.01 \\ -0.01 \end{bmatrix}$$

After a simple MATLAB computation, we see that the new solution for this perturbed system is

$$\hat{x} = \begin{bmatrix} 20.97 \\ -18.99 \end{bmatrix}, \quad \text{which means} \ \ \delta\vec{x} = \begin{bmatrix} 19.97 \\ -19.99 \end{bmatrix}.$$

So as discussed, we hope that when $||\delta\vec{b}||/||\vec{b}||$ is small, that $||\delta\vec{x}||/||\vec{x}||$ is as well. For this calculation we will use the Euclidean norm. So

$$\begin{aligned} \frac{||\delta\vec{b}||}{||\vec{b}||} &= \frac{\sqrt{0.01^2 + (-0.01)^2}}{\sqrt{1999^2 + 1997^2}} \approx 5.005 \times 10^{-6}, \ \ \text{and} \\ \frac{||\delta\vec{x}||}{||\vec{x}||} &= \frac{\sqrt{19.97^2 + (-19.99)^2}}{\sqrt{1^2 + 1^2}} \approx 19.98. \end{aligned}$$

We had hoped that the relative sizes of $\delta\vec{b}$ to $\vec{b}$ would be comparable to $\delta\vec{x}$ to $\vec{x}$, but it turns out that $||\delta\vec{x}||/||\vec{x}||$ is about 4.0×10^6 times bigger than $||\delta\vec{b}||/||\vec{b}||$. So this matrix has the potential to give us poor accuracy when we make calculations with it.

But we can give a bound for $||\delta\vec{x}||/||\vec{x}||$ in terms of $||\delta\vec{b}||/||\vec{b}||$ in the following theorem of Watkins' book [10].

Theorem 2.36. [10, Theorem 2.2.4] *Fix a matrix norm $||\cdot||$, and let $||\cdot||$ be a compatible vector norm. Let A be a nonsingular matrix over $\mathbb{R}$. Let $\vec{x}$ be the solution of $A\vec{x} = \vec{b}$, and let $\hat{x} = \vec{x} + \delta\vec{x}$ be the solution to $A\hat{x} = \vec{b} + \delta\vec{b}$. Then*

$$\frac{||\delta\vec{x}||}{||\vec{x}||} \leq ||A|| \cdot ||A^{-1}|| \frac{||\delta\vec{b}||}{||\vec{b}||} \tag{2.5}$$

The factor $||A|| \cdot ||A^{-1}||$ is called the *condition number* with respect to the matrix norm, and it is a helpful tool to describe and study the sensitivity of a linear system to small perturbations in the system. Note, however, that it is not exclusive to the Frobenius norm, and that we can consider the condition number of a system with respect to any norm.

Definition 2.37. Fix any matrix norm $||\cdot||$. If A is an invertible (nonsingular) matrix with inverse A^{-1}, then the *condition number* of A, denoted $\kappa(A)$, is

$$\kappa(A) = ||A|| \cdot ||A^{-1}||.$$

We denote the condition number of a matrix A with respect to the Frobenius norm as

$$\kappa_F(A) = ||A||_F \cdot ||A^{-1}||_F.$$

We might also consider perturbing A in the system $A\vec{x} = \vec{b}$. Consider the two systems $A\vec{x} = \vec{b}$, and $(A + \delta A)\hat{x} = \vec{b}$, where $||\delta A||/||A||$ is small. We want to guarantee that $(A + \delta A)\hat{x} = \vec{b}$ indeed has a solution near to that of $A\vec{x} = \vec{b}$. We can consider a theorem similar to Theorem 2.36 in terms of perturbations to A.

Theorem 2.38. [10, Theorem 2.3.3] *Fix a matrix norm $||\cdot||$, and let $||\cdot||$ be a compatible vector norm. Let A be nonsingular, let $\vec{b} \neq 0$, and let $\vec{x}$ and $\hat{x} = \vec{x} + \delta\vec{x}$ be solutions of $A\vec{x} = \vec{b}$ and $(A + \delta A)\hat{x} = \vec{b}$, resepctively. Then*

$$\frac{||\delta\vec{x}||}{||\hat{x}||} \leq \kappa(A) \frac{||\delta A||}{||A||}.$$

Theorem 2.38 shows that if the condition number is small, then a small perturbation in A will not perturb the solution to $A\vec{x} = \vec{b}$ by a lot.

Example 2.39. Let

$$A = \begin{bmatrix} 400 & -201 \\ -800 & 401 \end{bmatrix}, \quad \text{and} \quad \vec{b} = \begin{bmatrix} 200 \\ -200 \end{bmatrix}.$$

So when we find a solution to the system $A\vec{x} = \vec{b}$ we get

$$\vec{x} = \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} -100 \\ -200 \end{bmatrix}.$$

But when we make a small change in the matrix A, i.e.

$$A = \begin{bmatrix} \mathbf{401} & -201 \\ -800 & 401 \end{bmatrix}$$

and calculate the solution again we get that

$$\hat{x} = \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 40,000 \\ 79,800 \end{bmatrix}.$$

So if $\hat{x} = \vec{x} + \delta x$, then

$$\delta x = \begin{bmatrix} 40,100 \\ 80,000 \end{bmatrix}.$$

Thus, if we take the ratio $||\delta x||$ and the original $||\vec{x}||$ we get

$$\frac{||\delta x||}{||\vec{x}||} = \frac{\sqrt{8,008,010,000}}{\sqrt{50,000}} \approx 400,$$

which is quite a large ratio for such a small change in the the matrix A.

Example 2.40. Return to the matrix of Example 2.39:

$$A = \begin{bmatrix} 400 & -201 \\ -800 & 401 \end{bmatrix}.$$

Using the Frobenius norm, we calculate $\kappa_F(A)$. From A, we compute A^{-1}

$$\begin{aligned} A^{-1} =& \frac{1}{(400)(401) - (-800)(-201)} \begin{bmatrix} 401 & 201 \\ 800 & 400 \end{bmatrix} \\ =& \frac{1}{-400} \begin{bmatrix} 401 & 201 \\ 800 & 400 \end{bmatrix} = \begin{bmatrix} -\frac{401}{400} & -\frac{201}{400} \\ -2 & -1 \end{bmatrix}. \end{aligned}$$

Then

$$||A||_F = \sqrt{(400)^2 + (-201)^2 + (-800)^2 + (401)^2} = \sqrt{1,001,202}$$

and

$$||A^{-1}||_F = \sqrt{\left(-\frac{401}{400}\right)^2 + \left(-\frac{201}{400}\right)^2 + (-2)^2 + (-1)^2} = \sqrt{6.2575125}.$$

Thus

$$\kappa_F(A) = ||A||_F \cdot ||A^{-1}||_F = \sqrt{1,001,202}\sqrt{6.2575125} = 2503.005.$$

This explains why the calculation with a small perturbation in Example 2.39 could give such a different result. If we fix the vector norm to be the Euclidean norm, and recall Theorem 2.38, it should be the case that

$$\frac{||\delta\vec{x}||}{||\hat{x}||} \leq \kappa_F(A)\frac{||\delta A||_F}{||A||_F}.$$

From the original example in Example 2.39, we had that

$$\delta\vec{x} = \begin{bmatrix} 40,100 \\ 80,000 \end{bmatrix}, \quad \hat{x} = \begin{bmatrix} 40,000 \\ 79,800 \end{bmatrix}, \quad A = \begin{bmatrix} 400 & -201 \\ -800 & 401 \end{bmatrix}, \quad \delta A = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}.$$

And so

$$\begin{aligned} \frac{||\delta\vec{x}||}{||\hat{x}||} &= \frac{\sqrt{40,100^2 + 80,000^2}}{\sqrt{40,000^2 + 79,800^2}} \approx 1.003 \text{ and} \\ \kappa_F(A)\frac{||\delta A||_F}{||A||_F} &= 2503 \times \frac{1}{\sqrt{1,001,202}} \approx 2.501. \end{aligned}$$

Thus, a small change in A results in a significantly different solution to $A\vec{x} = \vec{b}$. This becomes a significant issue when one designs algorithms with round-off errors in mind.

We recall that the relative size of δA to A is then given by $||\delta A||/||A||$, and the relative size of $\delta\vec{x}$ to $\vec{x}$ is then given by $||\delta\vec{x}||/||\vec{x}||$. Theorem 2.38 shows us that when we compare these numbers we can put an upper bound on it using the condition number. We use the following definition from Poole [8] to describe what large and small condition numbers mean to a system of linear equations.

Definition 2.41. A matrix A is *ill-conditioned* if small changes in its entries can produce large changes in the solutions to $A\vec{x} = \vec{b}$. If small changes in the entries of A produce only small changes in the solutions of $A\vec{x} = \vec{b}$, then A is called *well-conditioned.*

In other words, the larger the condition number, the larger the changes in the solutions to a system $A\vec{x} = \vec{b}$ can be when we make small changes to the matrix A. We saw that the matrix A in Example 2.39 gave large changes in the solutions to $A\vec{x} = \vec{b}$, and therefore is ill-conditioned. The following theorem gives us a lower bound on the condition number of any matrix with respect to any norm.

Theorem 2.42. [10, Proposition 2.2.7] *For any square matrix A, and any matrix norm $||\cdot||$,*

$$\kappa(A) \geq ||I||.$$

Proof. We have that $\kappa(A) = ||A|| \cdot ||A^{-1}||$, and from the axioms of a matrix norm it is the case that $||A|| \cdot ||A^{-1}|| \geq ||AA^{-1}|| = ||I||$. □

Theorem 2.42 implies that the closer a condition number is to $||I||$, the better conditioned it is.

Lemma 2.43. *Fix a matrix norm $||\cdot||$. If two companion matrices A and B are equivalent, then they have the same condition number, i.e., $\kappa(A) = \kappa(B)$.*

Proof. Suppose that $A = PBP^T$ for some permutation matrix P. Then A and B have the same entries. Since $A = PBP^T$, A^{-1} and B^{-1} have the same entries as well. Hence, $\kappa(A) = \kappa(B)$. □

2.3 Specific Matrix Norms and their Applications

In this section we introduce some specific types of matrix norms. First we introduce the Frobenius norm. Although it was introduced in the previous section, we now verify that it is a norm. In the last chapter we also discuss the spectral norm and how it will help us in the study of condition numbers.

Recall the definition of the *Frobenius norm* of a matrix A given in Defnition 2.33:

$$||A||_F = \sqrt{\sum_{ij} |a_{ij}|^2}.$$

It is straightforward to check that $||A||_F$ is a matrix norm.

- $||A||_F = 0$ if and only if $A = 0$ is certainly true.
- $$\begin{aligned} ||kA||_F &= \sqrt{\sum_{ij} |ka_{ij}|^2} = \sqrt{\sum_{ij} |k|^2 |a_{ij}|^2} = \sqrt{|k|^2 \sum_{ij} |a_{ij}|^2} = |k| \sqrt{\sum_{ij} |a_{ij}|^2} \\ &= |k| \cdot ||A||_F \end{aligned}$$
- It suffices to show that $||A+B||_F^2 \leq (||A||_F + ||B||_F)^2$. But for this part we'll need the *Cauchy-Schwarz Inequality* [5, Theorem 5.1.4]:

$$\left| \sum_{i=n}^{n} x_i y_i \right| \leq \sqrt{\sum_{i=n}^{n} x_i^2} \sqrt{\sum_{i=n}^{n} y_i^2}.$$

So with this we can solve

$$\begin{aligned}
||A+B||_F^2 &= \sum_{ij} |a_{ij}+b_{ij}|^2 \\
&= \sum_{ij} |a_{ij}|^2 + 2\sum_{ij} |a_{ij}b_{ij}| + \sum_{ij} |b_{ij}|^2 \\
&\leq \sum_{ij} |a_{ij}|^2 + 2\sqrt{\sum_{i,j} a_i^2}\sqrt{\sum_{i,j} b_i^2} + \sum_{ij} |b_{ij}|^2 \\
&= \left(\sqrt{\sum_{ij} |a_{ij}|^2} + \sqrt{\sum_{ij} |b_{ij}|^2}\right)^2 \\
&= (||A||_F + ||B||_F)^2.
\end{aligned}$$

- It suffices to show that $||AB||_F^2 \leq ||A||_F^2 \cdot ||B||_F^2$. Suppose we have $A \in \mathbb{R}^{m\times k}$, and we have $B \in \mathbb{R}^{k\times n}$.

$$\begin{aligned}
||AB||_F^2 &= \sum_{i=1}^{m}\sum_{j=1}^{n}(c_{ij}^2) \\
&= \sum_{i=1}^{m}\sum_{j=1}^{n}\left(\sum_{\ell=1}^{k}(a_{i\ell}b_{\ell j})^2\right) \\
&\leq \sum_{i=1}^{m}\sum_{j=1}^{n}\left(\sum_{\ell=1}^{k}a_{i\ell}^2\sum_{\ell=1}^{k}b_{\ell j}^2\right) \quad \text{(By C.S.)} \\
&= \left(\sum_{i=1}^{m}\sum_{\ell=1}^{k}a_{i\ell}^2\right)\left(\sum_{j=1}^{n}\sum_{\ell=1}^{k}b_{\ell j}^2\right) \\
&= ||A||_F^2 \cdot ||B||_F^2
\end{aligned}$$

Chapter 3

Using the Hessenberg form of a companion matrix

In this chapter, we review the results about condition numbers presented by De Terán et. al. [1]. In particular, De Terán et. al. determine condition numbers for Fiedler companion matrices using Fiedler's product construction. Our goal is to give new, simplified proofs of these results by using the Hessenberg structure of a Fiedler companion matrix. In fact, we extend the results of De Terán et. al. by presenting some corresponding results for the larger class of sparse companion matrices.

In [1] De Terán et. al. use a bijection to describe the structure of Fiedler matrices. The bijection corresponds to the order in which you multiply the matrices from Definition 2.16. The following theorem describes the entries of any Fiedler companion matrix.

Theorem 3.1. [1, Theorem 2.8] *Let $p(x) = x^n + a_{n-1}x^{n-1} + \cdots + a_0$ be a polynomial over $\mathbb{R}$ where $n \geq 2$. Let $\sigma = (i_1, \ldots, i_n)$ be a permutation of the integers $\{0, \ldots, n-1\}$. Let M_σ be the Fiedler companion of $p(x)$ associated to σ. Then*

(a) M_σ has n entries equal to $-a_0, -a_1, \ldots, -a_{n-1}$, with one copy of each.

(b) M_σ has $n-1$ entries equal to 1.

(c) The rest of the entries in M_σ are 0.

(d) If an entry equal to 1 of those in part (b) is at position (i, j), then either the rest of the entries in the ith row of M_σ are equal to 0, or the rest of the entries in the jth column of M_σ are equal to 0.

Example 3.2. To illustrate the last statement of the previous theorem, we revisit Example 2.14. If we let $p(x) = x^5 + a_4x^4 + a_3x^3 + a_2x^2 + a_1x + a_0$, and consider the companion matrix to $p(x)$ of the form

$$M_\sigma = \begin{bmatrix} -a_4 & -a_3 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & -a_2 & 0 & -a_1 & -a_0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix},$$

we see that for every position that M_σ has a 1, there are zeros either in the rest of the column or the rest of the row.

Theorem 3.1 $(a) - (c)$ is essentially noting that a Fiedler companion matrix is a sparse companion matrix. Theorem 3.1 (d) follows from (2.1). The Hessenberg form of these matrices was noted in Theorem 2.19. As such, Theorem 3.1 could be viewed as a corollary to Theorem 2.19. With Theorem 3.1, we can compute the Frobenius norm of any Fiedler companion matrix, which is independent of the permutation σ, and depends only on the companion matrix's characteristic polynomial $p(x)$.

Theorem 3.3. *Let $p(x) = x^n + a_{n-1}x^{n-1} + \cdots + a_1x + a_0$ be a polynomial over $\mathbb{R}$, where $n \geq 2$. Then every sparse companion matrix to $p(x)$ has the same Frobenius norm.*

Corollary 3.4. [1, Corollary 2.9] *Let $p(x) = x^n + a_{n-1}x^{n-1} + \cdots + a_0$ be a polynomial over $\mathbb{R}$, where $n \geq 2$. Let $\sigma = (i_1, \ldots, i_n)$ be a permutation of the integers $\{0, \ldots, n-1\}$, and let M_σ be the Fiedler companion of $p(x)$ associated to σ. Then*

$$||M_\sigma||_F = \sqrt{(n-1) + |a_0|^2 + |a_1|^2 + \cdots + |a_{n-1}|^2}.$$

Since the condition number of a matrix depends on the entries of M and M^{-1}, we need to consider the entries of M^{-1} as well. De Terán et. al. in [1] described the structure of the inverse of a Fiedler companion matrix, and determined all the entries of M^{-1} using initial consecutions and inversions, as introduced in Definition 2.22.

Theorem 3.5. [1, Theorem 3.2] *Let $p(x) = x^n + a_{n-1}x^{n-1} + a_{n-2}x^{n-2} + \cdots + a_1x + a_0$ be a polynomial over $\mathbb{R}$, with $n \geq 2$ and $a_0 \neq 0$, and let $\sigma = (i_1, \ldots, i_n)$ be a permutation of the integers $\{0, \ldots, n-1\}$. Let M_σ be the lower Hessenberg Fiedler companion matrix to $p(x)$, and let t_σ be number of initial consecutions or inversions of σ. Then:*

(a) M_σ^{-1} has $t_\sigma + 1$ entries equal to $-\frac{1}{a_0}, -\frac{a_1}{a_0}, \ldots, -\frac{a_{t_\sigma}}{a_0}$, with exactly one copy of each;

(b) M_σ^{-1} has $n-1-t_\sigma$ entries equal to $a_{t_\sigma+1}, a_{t_\sigma+2}, \ldots, a_{n-1}$, with exactly one copy of each;

(c) M_σ^{-1} has $n-1$ entries equal to 1; and

(d) the rest of the entries of M_σ^{-1} are 0.

In Theorem 3.5, the entries depend on t_σ, which in turn, depends on σ. However, we give an alternative proof to Theorem 3.5 by relating t_σ to the lattice path in the lower Hessenberg form of a Fiedler companion matrix.

Definition 3.6. Let $p(x) = x^n + a_{n-1}x^{n-1} + a_{n-2}x^{n-2} + \cdots + a_1 x + a_0$ be a monic polynomial. Let M be a Fiedler companion matrix to the polynomial $p(x)$ in lower Hessenberg form. The *initial step size* of the lattice path of M, is the number of coefficients other than a_0 in the row or column containing both a_0 and a_1.

And with this definition, we can prove a result similar to Theorem 3.5.

Theorem 3.7. *Let $p(x) = x^n + a_{n-1}x^{n-1} + a_{n-2}x^{n-2} + \cdots + a_1 x + a_0$ be a polynomial over $\mathbb{R}$ with $n \geq 2$ and $a_0 \neq 0$. Let M be a Fiedler companion matrix to the polynomial $p(x)$ in lower Hessenberg form. Let t be the initial step value of M. Then:*

(a) M^{-1} has $t+1$ entries equal $-\frac{1}{a_0}, -\frac{a_1}{a_0}, \ldots, -\frac{a_t}{a_0}$, with exactly one copy of each;

(b) M^{-1} has $n-1-t$ entries equal to $a_{t+1}, a_{t+2}, \ldots, a_{n-1}$, with exactly one copy of each;

(c) M^{-1} has $n-1$ entries equal to 1; and

(d) the rest of the entries of M^{-1} are 0.

Proof. Consider Lemma 2.26 which gives us the inverse of any companion matrix in lower Hessenberg form. Notice that if M is a Fiedler matrix, then the a_i entries of the matrix form a lattice path according to Theorem 2.19. Thus one of the vectors $\vec{u}$ or $\vec{y}^T$ in the lower Hessenberg form of the Fiedler matrix given in Lemma 2.26 will have to be a zero vector. Without loss of generality, let $\vec{y}^T$ be zero, which means that $-\frac{1}{a_0}\vec{u}\vec{y}^T - H = -H$. If the initial step value of F is t, then there will t nonzero elements in $\vec{u}$, which must take the form

$$\vec{u} = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ -a_t \\ \vdots \\ -a_1 \end{bmatrix}.$$

Thus, the inverse of this Fiedler matrix, we will have t entries that have a $\frac{1}{a_0}$ coefficient from the $\frac{1}{a_0}\vec{u}$ term in Lemma 2.26. Thus

$$M^{-1} = \left[\begin{array}{c|c|c} 0 & \vec{0}^T & -\frac{1}{a_0} \\ \hline I_m & 0 & \vec{0} \\ \hline -H & I_{n-m-1} & \frac{1}{a_0}\vec{u} \end{array}\right]. \tag{3.1}$$

The matrix M^{-1} will have precisely $t+1$ entries equal to $-\frac{1}{a_0}, -\frac{a_1}{a_0}, \ldots, -\frac{a_t}{a_0}$, and the remaining variables appear in H. All of the properties of M^{-1} in the theorem follow. □

With Theorem 3.7 we can recover De Terán et. al.'s computation for the norm of the inverse of a Fiedler matrix.

Corollary 3.8. [1, Corollary 3.3] *Let $p(x) = x^n + a_{n-1}x^{n-1} + a_{n-2}x^{n-2} + \cdots + a_1x + a_0$ be a polynomial over $\mathbb{R}$, with $n \geq 2$ and $a_0 \neq 0$. Let M be a Fiedler companion matrix to $p(x)$, with an initial step value of t. Then*

$$||M^{-1}||^2 = (n-1) + \frac{1 + |a_1|^2 + \cdots + |a_t|^2}{|a_0|^2} + |a_{t+1}|^2 + \cdots + |a_{n-1}|^2.$$

Proof. This follows directly from Theorem 3.7. □

With Corollary 3.8 and Theorem 3.4, we compute the condition number of a Fiedler matrix with respect to the Frobenius norm. The next theorem first appeared in De Terán et. al. [1], and we have re-expressed it using the initial step value.

Theorem 3.9. *Let $p(x) = x^n + a_{n-1}x^{n-1} + a_{n-2}x^{n-2} + \cdots + a_1x + a_0$ be a polynomial over $\mathbb{R}$, with $n \geq 2$ and $a_0 \neq 0$. Let M be a sparse lower Hessenberg Fiedler companion matrix to $p(x)$, with an initial step value of t. Then*

$$\kappa_F(M)^2 = ||M||_F^2 \cdot \left((n-1) + \frac{1 + |a_1|^2 + \cdots + |a_t|^2}{|a_0|^2} + |a_{t+1}|^2 + \ldots + |a_{n-1}|^2\right),$$

with

$$||M||_F^2 = (n-1) + |a_0|^2 + |a_1|^2 + \cdots |a_{n-1}|^2.$$

Corollary 3.10. [1, Corollary 4.3] *Let $p(x) = x^n + a_{n-1}x^{n-1} + a_{n-2}x^{n-2} + \cdots + a_1x + a_0$ be a polynomial over $\mathbb{R}$, with $n \geq 2$ and $a_0 \neq 0$. Let A and B be Fiedler companion matrices to the polynomial $p(x)$ in lower Hessenberg form. If the initial step value of both A and B is t, then $\kappa_F(A) = \kappa_F(B)$.*

Proof. We know that any two companion matrices to the polynomial $p(x)$ will have the same norm by Lemma 2.43. So we have then that $||A||_F = ||B||_F$. From Corollary 3.8, we also have that $||A^{-1}||_F = ||B^{-1}||_F$ when A and B have the same initial step size. Therefore

$$\kappa_F(A) = ||A||_F \cdot ||A^{-1}||_F = ||B||_F \cdot ||B^{-1}||_F = \kappa_F(B).$$

□

The following corollary is inspired by [1, Corollary 4.3], and gives the necessary conditions for two Fiedler companion matrices in lower Hessenberg form to have the same condition number.

Corollary 3.11. *Let $p(x) = x^n + a_{n-1}x^{n-1} + a_{n-2}x^{n-2} + \cdots + a_1x + a_0$ be a polynomial over $\mathbb{R}$ with $n \geq 2$ and $a_0 \neq 0$. Let M be any lower Hessenberg Fiedler companion matrix to $p(x)$, and define $S_t := \{M : M$ is a Fiedler companion matrix to $p(x)$ with initial step size $t\}$. We then define*

$$\kappa(t) := \kappa_F(M), \text{ for } M \in S_t.$$

Then

(a) if $|a_0| < 1$, then $\kappa(1) \leq \kappa(2) \leq \cdots \leq \kappa(n-1)$;

(b) if $|a_0| = 1$, then $\kappa(1) = \kappa(2) = \cdots = \kappa(n-1)$; and

(c) if $|a_0| > 1$, then $\kappa(1) \geq \kappa(2) \geq \cdots \geq \kappa(n-1)$.

Proof. If $|a_0| < 1$, then it is clear from Theorem 3.9 that $\kappa(M)$ increases as the number of coefficients divided by $|a_0|^2$ increases. It is the opposite but similar case when $a_0 > 1$. □

Theorem 3.12. *Let $p(x) = x^n + a_{n-1}x^{n-1} + a_{n-2}x^{n-2} + \cdots + a_1 x + a_0$ with $n \geq 2$. Let*

$$S_t := \{M \ : \ M \text{ is a Fiedler companion matrix with initial step size } t\},$$

i.e., S_t is the set of all lower Hessenberg Fiedler companion matrices M to the polynomial $p(x)$ with an intial step size of t, where $1 \leq t \leq n-1$. Then

$$|S_t| = \begin{cases} 2^{n-1-t}, & \text{if } t < n-1, \\ 2, & \text{if } t = n-1. \end{cases}$$

Proof. Recall the structure of a companion matrix in lower Hessenberg form from Lemma 2.26:

$$M = \left[\begin{array}{c|c|c} \vec{0} & I_m & 0 \\ \hline \vec{u} & H & I_{n-m-1} \\ \hline -a_0 & \vec{y}^T & \vec{0}^T \end{array}\right]. \tag{3.2}$$

If $t = n-1$, then M has the coefficients $-a_1, \ldots, -a_{n-1}$ either completely in $\vec{u}$ or completely in $\vec{y}^T$. If $t < n-1$, then there are $t-1$ coefficients that fall in $\vec{u}$ or in $\vec{y}^T$, and then $-a_t$ must fall in H. From there, each succeeding $-a_i$, for $t+1 \leq i \leq n-1$, can go either directly to the right, or directly above the previous entry since M has a lattice path. So there are 2^{n-t-2} ways to do that. We then multiply this by 2 since then $-a_1, \ldots, -a_{t-1}$ fall in either $\vec{u}$ or $\vec{y}^T$. □

The next result compares condition numbers of general companion matrices with some restrictions to Fiedler companion matrices.

Theorem 3.13. *Let $p(x) = x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0$ be a polynomial over $\mathbb{R}$, with $n \geq 2$, and $|a_0| > 1$, and let M be any Fiedler matrix in lower Hessenberg form, i.e.,*

$$M = \left[\begin{array}{c|c|c} \vec{0} & I_m & 0 \\ \hline \vec{u}_M & H_M & I_{n-m-1} \\ \hline -a_0 & \vec{y}_M^T & \vec{0}^T \end{array}\right]$$

where t is the initial step size of M. Suppose C is any companion matrix to $p(x)$ in lower Hessenberg form, i.e.,

$$C = \left[\begin{array}{c|c|c} \vec{0} & I_m & 0 \\ \hline \vec{u}_C & H_C & I_{n-m-1} \\ \hline -a_0 & \vec{y}_C^T & \vec{0}^T \end{array}\right]$$

where the initial number of nonzero elements in either $\vec{u}_C$ or $\vec{y}_C^T$ is also t. Suppose that $\kappa(M) \leq \kappa(C)$. Then if either $\vec{u}_C$ or $\vec{y}_C^T$ are zero vectors, and a_{n-1} is in the same position in both matrices, then:

$$1 \leq \frac{\kappa(C)}{\kappa(M)} \leq \kappa(M).$$

Proof. We want to show that

$$\frac{||C|| \cdot ||C^{-1}||}{||M|| \cdot ||M^{-1}||} \leq ||M|| \cdot ||M^{-1}||.$$

Since $||C|| = ||M||$, it suffices to show that

$$||C^{-1}|| \leq ||M|| \cdot ||M^{-1}||^2.$$

Without loss of generality, assume that $\vec{u}_C = \vec{0}$ in C. This means that

$$\vec{y}_C^T = (-a_1, \ldots, -a_{t-1}, 0, -a_{j_1}, \ldots, -a_{j_s}, 0, \ldots, 0).$$

In this vector, $-a_1, \ldots, -a_{t-1}$ are all nonzero, but $a_{j_1}, \ldots, a_{j_s}$ do not need to be nonzero. Consider then, C^{-1}

$$C^{-1} = \left[\begin{array}{c|c|c} \frac{1}{a_0}\vec{y}_C^T & \vec{0}^T & -\frac{1}{a_0} \\ \hline I_m & 0 & \vec{0} \\ \hline -\frac{1}{a_0}\vec{u}_C\vec{y}_C^T - H_C & I_{n-m-1} & \frac{1}{a_0}\vec{u}_C \end{array}\right].$$

Because $\vec{u}_C = \vec{0}$, the norm of C^{-1} would then be

$$\begin{aligned} ||C^{-1}|| &= (n-1) + \left(\frac{1}{a_0}\right)^2 + \sum_{i=1}^{t-1}\left|\frac{a_i}{a_0}\right|^2 + \sum_{i=1}^{s}\left|\frac{a_{j_i}}{a_0}\right|^2 + \sum_{i=1}^{\ell}|a_{k_i}|^2 \\ &= (n-1) + \left(\frac{1}{a_0}\right)^2 + \sum_{a_i \in \vec{y}_C^T}\left|\frac{a_i}{a_0}\right|^2 + \sum_{a_k \in H_C}|a_k|^2 \qquad (3.3) \end{aligned}$$

where $\{k_1, \ldots, k_\ell\} \sqcup \{1, \ldots, t-1, j_1, \ldots, j_s,\} = \{1, \ldots, n-1\}$. We also have that

$$\begin{aligned} ||M|| \cdot ||M^{-1}||^2 = &\left[(n-1) + \sum_{i=0}^{n-1}|a_i|^2\right] \\ &\times \left[(n-1) + \left(\frac{1}{a_0}\right)^2 + \sum_{i=1}^{t-1}\left|\frac{a_i}{a_0}\right|^2 + \sum_{j=t}^{n-1}|a_j|^2\right] \\ &\times \left[(n-1) + \left(\frac{1}{a_0}\right)^2 + \sum_{i=1}^{t-1}\left|\frac{a_i}{a_0}\right|^2 + \sum_{j=t}^{n-1}|a_j|^2\right]. \qquad (3.4) \end{aligned}$$

We want to show that $||C^{-1}|| \leq ||M|| \cdot ||M^{-1}||^2$. To do this we simply consider the four different summands in (3.3), and show that we can find unique terms in
$||M|| \cdot ||M^{-1}||^2$ that is greater than or equal to this summand.

- The summand $(n-1)$ in (3.3) is less than $(n-1)^3$ in (3.4).
- The summand $\left(\frac{1}{a_0}\right)^2$ in (3.3) is less than $(n-1)^2\left(\frac{1}{a_0}\right)^2$ in (3.4).
- The summand $\sum_{a_i \in \vec{y}_C^T} \left|\frac{a_i}{a_0}\right|^2$ in (3.3) is less than $\sum_{i=0}^{n-1} |a_i|^2(n-1)\left(\frac{1}{a_0}\right)^2$ in (3.4).
- Lastly, the summand $\sum_{a_k \in H_C} |a_k|^2$ in (3.3) is less than $(n-1)\sum_{j=t}^{n-1} |a_j|^2(n-1)$ in (3.4).

As a consequence, $||C^{-1}|| \leq ||M|| \cdot ||M^{-1}||^2$. □

Corollary 3.11 gives us a result about ascending and descending condition numbers for Fiedler companion matrices, which allows us put bounds on them in regards to there initial step values. Theorem 3.13 also uses the initial step size. The question left to ask is whether or not there exists a similar result for general companion matrices.

Chapter 4

Comparing Condition Numbers for Striped and Fiedler Cases

One of the main objectives of this project is to see if we can find other types companion matrices that have better condition numbers than the Fiedler companion matrices. In this chapter we explore striped companion matrices. We provide conditions on the coefficients of the characteristic polynomial so that a striped companion matrix has a better (smaller) condition number than any Fiedler companion matrices.

4.1 Striped Companion Matrices with the Same Sized Stripes

When $a_0 = 1$ we can explicitly find the inverse for a sparse companion matrix. If $p(x) = x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + 1$ and

$$A = \left[\begin{array}{c|c|c} \vec{0} & I_m & 0 \\ \hline \vec{u} & H & I_{n-m-1} \\ \hline -a_0 & \vec{y}^T & \vec{0}^T \end{array}\right]$$

is a companion matrix to A then by Lemma 2.26, we can write the inverse as

$$A^{-1} = \left[\begin{array}{c|c|c} \vec{y}^T & \vec{0}^T & -1 \\ \hline I_m & 0 & \vec{0} \\ \hline -\vec{u}\vec{y}^T - H & I_{n-m-1} & \vec{u} \end{array}\right]. \tag{4.1}$$

Recall Definition 2.27 of a striped companion matrix, and consider the following lemma:

Lemma 4.1. *Let $U = C_{(m+1)k}(k,k,\dots,k)$, $a_0 = 1$, and $a_1, \dots, a_{(m+1)k-1} \in \mathbb{R}$. Then the inverse for U is*

$$U^{-1} = \left[\begin{array}{cccc|c|c}
-a_1 & -a_2 & \dots & -a_{k-1} & \vec{0}^T & -1 \\ \hline
\multicolumn{4}{c|}{I_m} & 0 & \vec{0} \\ \hline
-a_1a_{mk}+a_{mk+1} & -a_2a_{mk}+a_{mk+2} & \dots & -a_{k-1}a_{mk}+a_{(m+1)k-1} & & -a_{mk} \\
0 & 0 & \dots & 0 & & 0 \\
\vdots & \vdots & \dots & \vdots & & \vdots \\
-a_1a_{2k}+a_{2k+1} & -a_2a_{2k}+a_{2k+2} & \dots & -a_{k-1}a_{2k}+a_{3k-1} & & -a_{2k} \\
0 & 0 & \dots & 0 & & 0 \\
\vdots & \vdots & \dots & \vdots & I_{n-m-1} & \vdots \\
0 & 0 & \dots & 0 & & 0 \\
-a_1a_k+a_{k+1} & -a_2a_k+a_{k+2} & \dots & -a_{k-1}a_k+a_{2k-1} & & -a_k \\
0 & 0 & \dots & 0 & & 0 \\
\vdots & \vdots & \dots & \vdots & & \vdots \\
0 & 0 & \dots & 0 & & 0
\end{array}\right].$$

Proof. Note U is the matrix

$$U = \left[\begin{array}{cccc|c}
\vec{0} & \multicolumn{3}{c|}{I_{k-1}} & 0 \\ \hline
-a_{mk} & -a_{mk+1} & \dots & -a_{(m+1)k-1} & \\
0 & 0 & \dots & 0 & \\
\vdots & \vdots & \dots & \vdots & \\
0 & 0 & \dots & 0 & \\
-a_{ik} & -a_{ik+1} & \dots & -a_{(i+1)k-1} & \\
\vdots & \vdots & \dots & \vdots & \\
0 & 0 & \dots & 0 & \\
-a_{2k} & -a_{2k+1} & \dots & -a_{3k-1} & I_{n-k-1} \\
0 & 0 & \dots & 0 & \\
\vdots & \vdots & \dots & \vdots & \\
0 & 0 & \dots & 0 & \\
-a_k & -a_{k+1} & \dots & -a_{2k-1} & \\
0 & 0 & \dots & 0 & \\
\vdots & \vdots & \dots & \vdots & \\
0 & 0 & \dots & 0 & \\
1 & -a_1 & \dots & -a_{k-1} & \vec{0}^T
\end{array}\right].$$

This result is a special case of Lemme 2.26. In particular, notice that

$$-\frac{1}{a_0}\vec{u}\vec{y}^T = \frac{-1}{a_0}\begin{bmatrix} -a_{mk} \\ 0 \\ \vdots \\ -a_{2k} \\ 0 \\ \vdots \\ 0 \\ -a_k \\ 0 \\ \vdots \\ 0 \end{bmatrix}\begin{bmatrix} -a_1 & -a_2 & \dots & -a_{k-1} \end{bmatrix} = \begin{bmatrix} \frac{-a_1 a_{mk}}{a_0} & \frac{-a_2 a_{mk}}{a_0} & \dots & \frac{-a_{k-1} a_{mk}}{a_0} \\ 0 & 0 & \dots & 0 \\ \vdots & \vdots & \dots & \vdots \\ 0 & 0 & \dots & 0 \\ \frac{-a_1 a_{2k}}{a_0} & \frac{-a_2 a_{2k}}{a_0} & \dots & \frac{-a_{k-1} a_{2k}}{a_0} \\ 0 & 0 & \dots & 0 \\ \vdots & \vdots & \dots & \vdots \\ 0 & 0 & \dots & 0 \\ \frac{-a_1 a_k}{a_0} & \frac{-a_2 a_k}{a_0} & \dots & \frac{-a_{k-1} a_k}{a_0} \\ 0 & 0 & \dots & 0 \\ \vdots & \vdots & \dots & \vdots \\ 0 & 0 & \dots & 0 \end{bmatrix}.$$

From this matrix, we subtract H:

$$-\frac{1}{a_0}\vec{u}\vec{y}^T - H = \begin{bmatrix} \frac{-a_1 a_{mk}+a_0 a_{mk+1}}{a_0} & \frac{-a_2 a_{mk+1}+a_0 a_{mk+2}}{a_0} & \dots & \frac{-a_{k-1} a_{mk}+a_0 a_{(m+1)k-1}}{a_0} \\ 0 & 0 & \dots & 0 \\ \vdots & \vdots & \dots & \vdots \\ 0 & 0 & \dots & 0 \\ \frac{-a_1 a_{2k}+a_0 a_{2k+1}}{a_0} & \frac{-a_2 a_{2k}+a_0 a_{2k+2}}{a_0} & \dots & \frac{-a_{k-1} a_{2k}+a_0 a_{3k-1}}{a_0} \\ 0 & 0 & \dots & 0 \\ \vdots & \vdots & \dots & \vdots \\ 0 & 0 & \dots & 0 \\ \frac{-a_1 a_k+a_0 a_{k+1}}{a_0} & \frac{-a_2 a_k+a_0 a_{k+2}}{a_0} & \dots & \frac{-a_{k-1} a_k+a_0 a_{2k-1}}{a_0} \\ 0 & 0 & \dots & 0 \\ \vdots & \vdots & \dots & \vdots \\ 0 & 0 & \dots & 0 \end{bmatrix}.$$

With this computation, we use Lemma 2.26 to compute U^{-1}:

$$U^{-1} = \frac{1}{a_0}\left[\begin{array}{cccc|c|c}
-a_1 & -a_2 & \cdots & -a_{k-1} & \vec{0}^T & -1 \\
\hline
\multicolumn{4}{c|}{a_0 I_m} & 0 & \vec{0} \\
\hline
-a_1a_{mk}+a_0a_{mk+1} & -a_2a_{mk+1}+a_0a_{mk+2} & \cdots & -a_{k-1}a_{mk}+a_0a_{(m+1)k-1} & & -a_{mk} \\
0 & 0 & \cdots & 0 & & 0 \\
\vdots & \vdots & \cdots & \vdots & & \vdots \\
-a_1a_{2k}+a_0a_{2k+1} & -a_2a_{2k}+a_0a_{2k+2} & \cdots & -a_{k-1}a_{2k}+a_0a_{3k-1} & & -a_{2k} \\
0 & 0 & \cdots & 0 & & 0 \\
\vdots & \vdots & \cdots & \vdots & a_0 I_{n-m-1} & \vdots \\
0 & 0 & \cdots & 0 & & 0 \\
-a_1a_k+a_0a_{k+1} & -a_2a_k+a_0a_{k+2} & \cdots & -a_{k-1}a_k+a_0a_{2k-1} & & -a_k \\
0 & 0 & \cdots & 0 & & 0 \\
\vdots & \vdots & \cdots & \vdots & & \vdots \\
0 & 0 & \cdots & 0 & & 0
\end{array}\right].$$

The conclusion follows from the fact that $a_0 = 1$. □

Our goal is to determine conditions for when striped companion matrices have a better (i.e., smaller) condition number than any Fiedler matrix. Suppose we have some characteristic polynomial $p_M(x) = x^n + a_{n-1}x^{n-1} + a_{n-2}x^{n-2} + \cdots + a_1x + a_0$ of a companion matrix M. Recall that the condition number of the matrix M is:

$$\kappa_F(M) = ||M||_F \cdot ||M^{-1}||_F.$$

By Theorem 3.3, every sparse companion matrix of $p_M(x)$ will have the same norm. When we compare the condition numbers of matrices, we must then focus on the norm of the inverse M^{-1}. From Lemma 2.26, the norm of the inverse of a companion matrix depends primarily on the term $-\frac{1}{a_0}\vec{u}\vec{y}^T - H$.

By Corollary 3.8, the norm of the inverse of a Fiedler matrix should be exactly the same as the norm of the original matrix if $|a_0| = 1$. We wish to see when non-Fiedler cases have smaller condition number then that of the Fiedler case when $|a_0| = 1$. We answer this question in a special case in the next theorem.

Theorem 4.2. *Consider the monic polynomial*

$$p(x) = x^n + a_{n-1}x^{n-1} + a_{n-1}x^{n-2} + \cdots + a_1x + a_0$$

with $a_0 = 1$, $a_1, \ldots, a_{n-1} \in \mathbb{R}$, and $n = (m+1)k$. Then there exists a striped companion matrix $U = C_{(m+1)k}(k, k, \ldots, k)$ for $p(x)$ such that $\kappa_F(U) \leq \kappa_F(M)$ for any Fiedler companion matrix M if and only if the coefficients of $p(x)$ satisfy:

$$\sum_{j=1}^{m}\left(\sum_{i=1}^{k-1}|a_ia_{jk} - a_{jk+i}|^2\right) \leq \sum_{j=1}^{m}\left(\sum_{i=1}^{k-1}|a_{jk+i}|^2\right).$$

Proof. Let $U = C_{M+1}(k, \ldots, k)$, and let M be a Fiedler companion matrix. By definition $\kappa_F(U) \leq \kappa_F(M)$ if and only if

$$||U|| \cdot ||U^{-1}|| \leq ||M|| \cdot ||M^{-1}||.$$

Any sparse companion matrix to the polynomial $p(x)$, whether it is a Fiedler companion matrix or a striped companion matrix, will have the same norm by Theorem 3.3. It suffices to show that $||U^{-1}|| \leq ||M^{-1}||$ if and only if the coefficients of $p(x)$ satisfy:

$$\sum_{j=1}^{m} \left(\sum_{i=1}^{k-1} |a_i a_{jk} - a_{jk+i}|^2 \right) \leq \sum_{j=1}^{m} \left(\sum_{i=1}^{k-1} |a_{jk+i}|^2 \right).$$

By Lemma 4.1, there are n ones in the matrix U^{-1}, so

$$\begin{aligned} ||U^{-1}||^2 = n &+ |a_1|^2 + |a_2|^2 + \cdots + |a_{k-1}|^2 \\ &+ |a_k|^2 + |a_{2k}|^2 + \cdots + |a_{mk}|^2 \\ &+ |a_{k+1} - a_1 a_k|^2 + |a_{k+2} - a_2 a_k|^2 + \cdots + |a_{2k-1} - a_{k-1} a_k|^2 + \cdots \\ &+ |a_{2k+1} - a_1 a_{2k}|^2 + |a_{2k+2} - a_2 a_{2k}|^2 + \cdots + |a_{3k-1} - a_{k-1} a_{2k}|^2 \\ &+ |a_{mk+1} - a_1 a_{mk}|^2 + |a_{mk+2} - a_2 a_{mk}|^2 + \cdots + |a_{m(k+1)-1} - a_{k-1} a_{mk}|^2. \end{aligned}$$

By Corollary 3.8, we can similarly write out the norm of the inverse of the Fiedler matrix M as:

$$\begin{aligned} ||M^{-1}||^2 = n &+ |a_1|^2 + |a_2|^2 + \cdots + |a_{k-1}|^2 \\ &+ |a_k|^2 + |a_{2k}|^2 + \cdots + |a_{mk}|^2 \\ &+ |a_{k+1}|^2 + |a_{k+2}|^2 + \cdots + |a_{2k-1}|^2 \\ &+ |a_{2k+1}|^2 + |a_{2k+2}|^2 + \cdots + |a_{3k-1}|^2 + \cdots \\ &+ |a_{mk+1}|^2 + |a_{mk+2}|^2 + \cdots + |a_{(m+1)k-1}|^2. \end{aligned}$$

Note that the first two rows of each equation are the same. Thus, when we compare them, we can leave those terms out. We can write the remaining terms in $||U^{-1}||$ as

$$\begin{aligned} \sum_{j=1}^{m} \left(\sum_{i=1}^{k-1} |a_i a_{jk} - a_{jk+i}|^2 \right) &= |a_{k+1} - a_1 a_k|^2 + |a_{k+2} - a_2 a_k|^2 + \cdots + |a_{2k-1} - a_{k-1} a_k|^2 \\ &+ |a_{2k+1} - a_1 a_{2k}|^2 + |a_{2k+2} - a_2 a_{2k}|^2 + \cdots + |a_{3k-1} - a_{k-1} a_{2k}|^2 \\ &+ |a_{mk+1} - a_1 a_{mk}|^2 + |a_{mk+2} - a_2 a_{mk}|^2 + \cdots + |a_{m(k+1)-1} - a_{k-1} a_{mk}|^2. \end{aligned}$$

Similarly, we can represent the remaining $||M^{-1}||^2$ terms as:

$$\begin{aligned} \sum_{j=1}^{m} \left(\sum_{i=1}^{k-1} |a_{jk+i}|^2 \right) &= |a_{k+1}|^2 + |a_{k+2}|^2 + \cdots + |a_{2k-1}|^2 \\ &+ |a_{2k+1}|^2 + |a_{2k+2}|^2 + \cdots + |a_{3k-1}|^2 \\ &+ |a_{mk+1}|^2 + |a_{mk+2}|^2 + \cdots + |a_{(m+1)k-1}|^2. \end{aligned}$$

Therefore there can exists a striped companion matrix U that has a better condition number than a Fiedler companion matrix of the same characteristic polynomial if and only if,

$$\sum_{j=1}^{m}\left(\sum_{i=1}^{k-1}|a_i a_{jk}-a_{jk+i}|^2\right)\leq\sum_{j=1}^{m}\left(\sum_{i=1}^{k-1}|a_{jk+i}|^2\right).$$

□

Corollary 4.3. *Suppose that we have a monic polynomial*

$$p(x)=x^n+a_{n-1}x^{n-1}+a_{n-1}x^{n-2}+\cdots+a_1x+a_0$$

with $a_0=1$, $a_1,\ldots,a_{n-1}\in\mathbb{R}$, and $n=m(k+1)$. If the coefficients of $p(x)$ satisfy

$$\begin{array}{ccc}
|a_1a_k-a_{k+1}|\leq|a_{k+1}| & |a_2a_k-a_{k+2}|\leq|a_{k+2}| & \ldots|a_{k-1}a_k-a_{2k-1}|\leq|a_{2k-1}| \\
|a_1a_{2k}-a_{2k+1}|\leq|a_{2k+1}| & |a_2a_{2k}-a_{2k+2}|\leq|a_{2k+2}| & \ldots|a_{k-1}a_{2k}-a_{3k-1}|\leq|a_{3k-1}| \\
|a_1a_{3k}-a_{3k+1}|\leq|a_{3k+1}| & |a_2a_{3k}-a_{3k+2}|\leq|a_{3k+2}| & \ldots|a_{k-1}a_{3k}-a_{4k-1}|\leq|a_{4k-1}| \\
\vdots & \vdots & \vdots \\
|a_1a_{mk}-a_{mk+1}|\leq|a_{mk+1}| & |a_2a_{mk}-a_{mk+2}|\leq|a_{mk+2}| & \ldots|a_{k-1}a_{mk}-a_{(m+1)k-1}|\leq|a_{(m+1)k-1}|,
\end{array}$$

then there exists a non-Fiedler striped companion matrix $U=C_n(k,\ldots,k)$, such that $\kappa_F(U)\leq\kappa_F(M)$ for any Fiedler companion matrix M.

Proof. The hypotheses of the corollary implies that $|a_ia_{jk}-a_{jk+i}|\leq|a_{jk+i}|$, for $1\leq j\leq m$ and $1\leq i\leq k-1$. We then apply Theorem 4.2, to finish the proof. □

Example 4.4. Consider the following characteristic polynomial:

$$p(x)=x^9+4x^8+6x^7+2x^6+5x^5+5x^4+3x^3+3x^2+2x+1.$$

We can find a striped companion matrix and a Fiedler companion matrix for this polynomial, for example,

$$U=\begin{bmatrix}
0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
-2 & -6 & -4 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\
-3 & -5 & -5 & 0 & 0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\
-1 & -2 & -3 & 0 & 0 & 0 & 0 & 0 & 0
\end{bmatrix}\quad\text{and}\quad M=\begin{bmatrix}
0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & -4 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & -6 & 0 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & -2 & 0 & 0 & 1 & 0 & 0 & 0 \\
0 & 0 & -5 & 0 & 0 & 0 & 1 & 0 & 0 \\
0 & 0 & -5 & 0 & 0 & 0 & 0 & 1 & 0 \\
0 & 0 & -3 & 0 & 0 & 0 & 0 & 0 & 1 \\
-1 & -2 & -3 & 0 & 0 & 0 & 0 & 0 & 0
\end{bmatrix}.$$

The two corresponding inverse matrices are:

$$U^{-1} = \begin{bmatrix} -2 & -3 & 0 & 0 & 0 & 0 & 0 & 0 & -1 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 2 & -2 & 1 & 0 & 0 & 0 & 0 & 0 & -2 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ -1 & -4 & 0 & 0 & 0 & 1 & 0 & 0 & -3 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix} \text{ and}$$

$$M^{-1} = \begin{bmatrix} -2 & -3 & 0 & 0 & 0 & 0 & 0 & 0 & -1 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 4 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 6 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 5 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 5 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 3 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix}.$$

From Corollary 4.3 we consider the following inequalities:

$$\begin{array}{ll} |a_1a_6 - a_7| \leq |a_7| & \qquad |a_2a_6 - a_8| \leq |a_8| \\ |a_1a_3 - a_4| \leq |a_4| & \qquad |a_2a_3 - a_5| \leq |a_5| . \end{array}$$

In our case, all these inequalities are satisfied since

$$(a_0, a_1, a_2, a_3, a_4, a_5, a_6, a_7, a_8) = (1, 2, 3, 3, 5, 5, 2, 6, 4).$$

Thus Corollary 4.3 implies that $\kappa_F(U) \leq \kappa_F(M)$. Indeed,

$$\kappa_F(U) = ||U|| \cdot ||U^{-1}|| = \sqrt{137} \cdot \sqrt{60} \approx 91 \text{ and}$$
$$\kappa_F(M) = ||M|| \cdot ||M^{-1}|| = \sqrt{137} \cdot \sqrt{137} = 137.$$

Example 4.5. Consider the following characteristic polynomial:

$$p(x) = x^9 + 6x^8 + 4x^7 + 2x^6 + 9x^5 + 6x^4 + 3x^3 + 3x^2 + 2x + 1.$$

We can find a striped and a Fiedler companion matrix for this polynomial, e.g.,

$$U = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ -2 & -4 & -6 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ -3 & -6 & -9 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ -1 & -2 & -3 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \quad \text{and} \quad M = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -6 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -4 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & -2 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & -9 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & -6 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & -3 & 0 & 0 & 0 & 0 & 0 & 1 \\ -1 & -2 & -3 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

We see that for this example, when we construct the striped companion matrix $U = C_9(3, 3, 3)$, the submatrix R from Theorem 2.13 has rank 1. We now find the inverses of both matrices:

$$U^{-1} = \begin{bmatrix} -2 & -3 & 0 & 0 & 0 & 0 & 0 & 0 & -1 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & -2 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & -3 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix} \quad \text{and}$$

$$M^{-1} = \begin{bmatrix} -2 & -3 & 0 & 0 & 0 & 0 & 0 & 0 & -1 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 6 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 4 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 9 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 6 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 3 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix}.$$

Then by Corollary 4.3 we consider the following inequalities. Notice that we have extreme cases on the left hand sides of our inequalities:

$$|a_1a_6 - a_7| = 0 \le |a_7| \qquad |a_2a_6 - a_8| = 0 \le |a_8|$$
$$|a_1a_3 - a_4| = 0 \le |a_4| \qquad |a_2a_3 - a_5| = 0 \le |a_5|$$

$$\text{where } (a_0, a_1, a_2, a_3, a_4, a_5, a_6, a_7, a_8) = (1, 2, 3, 3, 6, 9, 2, 4, 6).$$

So these inequalities are satisfied, since all the left hand sides are zero. In the following, we will see that this extreme case happens whenever a striped lower Hessenberg companion matrix of the form in Theorem 2.13 has a submatrix R of rank 1. Theorem 4.10 generalizes this result further. When we compute the condition numbers of both matrices we get the following

$$\begin{aligned}\kappa_F(U) &= ||U|| \cdot ||U^{-1}|| = \sqrt{204} \cdot \sqrt{35} \approx 85; \text{ and,}\\ \kappa_F(M) &= ||M|| \cdot ||M^{-1}|| = \sqrt{204} \cdot \sqrt{204} = 204.\end{aligned}$$

In this case the ratio $\dfrac{\kappa_F(M)}{\kappa_F(U)} \approx \dfrac{204}{85} = 2.4$. This suggests we can make the ratio $\dfrac{\kappa_F(M)}{\kappa_F(U)}$ quite large.

Example 4.6. Let's consider the striped matrix $C_6(2,2,2)$

$$C_6(2,2,2) = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0\\ -a_4 & -a_5 & 1 & 0 & 0 & 0\\ 0 & 0 & 0 & 1 & 0 & 0\\ -a_2 & -a_3 & 0 & 0 & 1 & 0\\ 0 & 0 & 0 & 0 & 0 & 1\\ -1 & -a_1 & 0 & 0 & 0 & 0 \end{bmatrix},$$

for $a_1, \ldots, a_5 \in \mathbb{R}$. When the following inequalites hold

$$\begin{aligned} |a_1a_2 - a_3| &\le a_3\\ |a_1a_3 - a_4| &\le a_4\\ |a_1a_4 - a_5| &\le a_5\ , \end{aligned}$$

we know that the striped companion matrix will have a condition number equal to or better than any Fiedler companion matrix. When the rank of the submatrix

$$\begin{bmatrix} -a_4 & -a_5\\ 0 & 0\\ -a_2 & -a_3\\ 0 & 0\\ -1 & -a_1 \end{bmatrix}$$

is 1, the inequalities above become

$$\begin{aligned} |a_1a_2 - a_3| &= 0 \le a_3\\ |a_1a_3 - a_4| &= 0 \le a_4\\ |a_1a_4 - a_5| &= 0 \le a_5. \end{aligned}$$

We can use this fact to demonstrate an interesting example. Consider when $(a_0, a_1, a_2, a_3, a_4, a_5) = (1, k, ck, ck^2, ck^2, ck^3)$. Then

$$C_6(2,2,2) = U = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ -ck^2 & -ck^3 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ -ck & -ck^2 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ -1 & -k & 0 & 0 & 0 & 0 \end{bmatrix} \quad \text{and} \quad U^{-1} = \begin{bmatrix} -k & 0 & 0 & 0 & 0 & -1 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ ck^3 - ck^3 & 1 & 0 & 0 & 0 & -ck^2 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ ck - ck & 0 & 0 & 1 & 0 & -ck \\ 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix}.$$

Thus

$$||U^{-1}||^2 = c^2k^4 + c^2k^2 + k^2 + 6.$$

But for any 6×6 Fiedler companion matrix M with $a_0 = 1$, we have

$$\begin{aligned} ||M^{-1}||^2 &= a_1^2 + a_2^2 + a_3^2 + a_4^2 + a_5^2 + 6 \ \ \text{(by Corollary 3.8)} \\ &= c^2k^6 + c^2k^4 + c^2k^4 + c^2k^2 + k^2 + 6. \end{aligned}$$

When we consider the ratio of the two condition numbers we get

$$\frac{||M^{-1}||^2}{||U^{-1}||^2} = \frac{c^2k^6 + c^2k^4 + c^2k^4 + c^2k^2 + k^2 + 6}{c^2k^4 + c^2k^2 + k^2 + 6} \approx k^2$$

for sufficiently large k. Since we used the condition numbers squared for simplicity,

$$\frac{\kappa(M)}{\kappa(U)} \approx k.$$

The next theorem generalizes this example.

Theorem 4.7. *Suppose that we have a monic polynomial* $p(x) = q(x) + c_1x^kq(x) + c_2x^{2k}q(x) + \cdots + c_mx^{mk}q(x) + x^{(m+1)k}$ *over* $\mathbb{R}$, *with*

$$q(x) = a_{k-1}x^{k-1} + a_{k-2}x^{k-2} + \cdots + a_1x + 1$$

and $n = (m+1)k$. *Let* U *be the striped companion matrix in lower Hessenberg form to* $p(x)$

given by:

$$U = C_n(k,k,\dots,k) = \left[\begin{array}{cccc|c}
\vec{0} & & I_{k-1} & & 0 \\ \hline
-c_m & -c_m a_1 & \dots & -c_m a_{k-1} & \\
0 & 0 & \dots & 0 & \\
\vdots & \vdots & \dots & \vdots & \\
0 & 0 & \dots & 0 & \\
-c_i & -c_i a_1 & \dots & -c_i a_{k-1} & \\
0 & 0 & \dots & 0 & \\
\vdots & \vdots & \dots & \vdots & \\
0 & 0 & \dots & 0 & I_{n-k-1} \\
-c_2 & -c_2 a_1 & \dots & -c_2 a_{k-1} & \\
0 & 0 & \dots & 0 & \\
\vdots & \vdots & \dots & \vdots & \\
0 & 0 & \dots & 0 & \\
-c_1 & -c_1 a_1 & \dots & -c_1 a_{k-1} & \\
0 & 0 & \dots & 0 & \\
\vdots & \vdots & \dots & \vdots & \\
0 & 0 & \dots & 0 & \\
-1 & -a_1 & \dots & -a_{k-1} & \vec{0}^T
\end{array}\right].$$

For any Fiedler companion matrix M to $p(x)$, we have that

$$\left[\frac{\kappa_F(M)}{\kappa_F(U)}\right]^2 = \frac{(1+c_1^2+\cdots+c_m^2)(a_1^2+\cdots+a_{k-1}^2)+(c_1^2+\cdots+c_m^2)+n}{(a_1^2+\cdots+a_{k-1}^2)+(c_1^2+\cdots+c_m^2)+n}.$$

Proof. Note that, by Lemma 2.26

$$U^{-1} = \left[\begin{array}{cccc|c|c}
-a_1 & -a_2 & \dots & -a_{k-1} & \vec{0}^T & -1 \\
\hline
\multicolumn{4}{c|}{I_k} & 0 & \vec{0} \\
\hline
c_m a_1 - c_m a_1 & c_m a_2 - c_m a_2 & \dots & c_m a_{k-1} - c_m a_{k-1} & & -c_m \\
0 & 0 & \dots & 0 & & 0 \\
\vdots & \vdots & \dots & \vdots & & \vdots \\
0 & 0 & \dots & 0 & & 0 \\
c_i a_1 - c_i a_1 & c_i a_2 - c_i a_2 & \dots & c_i a_{k-1} - c_i a_{k-1} & & -c_i \\
\vdots & \vdots & \dots & \vdots & & \vdots \\
0 & 0 & \dots & 0 & & 0 \\
c_2 a_1 - c_2 a_1 & c_2 a_2 - c_2 a_2 & \dots & c_2 a_{k-1} - c_2 a_{k-1} & I_{n-k-1} & -c_2 \\
0 & 0 & \dots & 0 & & 0 \\
\vdots & \vdots & \dots & \vdots & & \vdots \\
0 & 0 & \dots & 0 & & 0 \\
c_1 a_1 - c_1 a_1 & c_1 a_2 - c_1 a_2 & \dots & c_1 a_{k-1} - c_1 a_{k-1} & & -c_1 \\
0 & 0 & \dots & 0 & & 0 \\
\vdots & \vdots & \dots & \vdots & & \vdots \\
0 & 0 & \dots & 0 & & 0
\end{array}\right].$$

All the terms in the bottom left submatrix are zero, which leaves us with

$$U^{-1} = \left[\begin{array}{cccc|c|c}
-a_1 & -a_2 & \dots & -a_{k-1} & \vec{0}^T & -1 \\
\hline
\multicolumn{4}{c|}{I_k} & 0 & \vec{0} \\
\hline
0 & 0 & \dots & 0 & & -c_m \\
0 & 0 & \dots & 0 & & 0 \\
\vdots & \vdots & \dots & \vdots & & \vdots \\
0 & 0 & \dots & 0 & & 0 \\
0 & 0 & \dots & 0 & & -c_i \\
\vdots & \vdots & \dots & \vdots & & \vdots \\
0 & 0 & \dots & 0 & & 0 \\
0 & 0 & \dots & 0 & I_{n-k-1} & -c_2 \\
0 & 0 & \dots & 0 & & 0 \\
\vdots & \vdots & \dots & \vdots & & \vdots \\
0 & 0 & \dots & 0 & & 0 \\
0 & 0 & \dots & 0 & & -c_1 \\
0 & 0 & \dots & 0 & & 0 \\
\vdots & \vdots & \dots & \vdots & & \vdots \\
0 & 0 & \dots & 0 & & 0
\end{array}\right]$$

So computing the norm squared of the inverse, we get

$$||U^{-1}||^2 = a_1^2 + \dots + a_{k-1}^2 + c_1^2 + \dots + c_m^2 + n.$$

From Corollary 3.8

$$\begin{aligned}
||M^{-1}||^2 &= a_1^2 + \cdots + a_{k-1}^2 + \\
&\quad c_1^2 + c_1^2 a_1^2 + \cdots + c_1^2 a_{k-1}^2 + \\
&\quad c_2^2 + c_2^2 a_1^2 + \cdots + c_2^2 a_{k-1}^2 + \\
&\qquad\qquad \vdots \\
&\quad c_m^2 + c_m^2 a_1^2 + \cdots + c_m^2 a_{k-1}^2 + n \\
&= c_1^2 + \cdots + c_m^2 + \\
&\quad 1(a_1^2 + \cdots + a_{k-1}^2) + \\
&\quad c_1^2(a_1^2 + \cdots + a_{k-1}^2) + \\
&\quad c_2^2(a_1^2 + \cdots + a_{k-1}^2) + \\
&\qquad\qquad \vdots \\
&\quad c_m^2(a_1^2 + \cdots + a_{k-1}^2) + n \\
&= (1 + c_1^2 + \cdots + c_m^2)(a_1^2 + \cdots + a_{k-1}^2) + (c_1^2 + \cdots + c_m^2) + n.
\end{aligned}$$

And so,

$$\left[\frac{\kappa(M)}{\kappa(U)}\right]^2 = \frac{||M^{-1}||^2}{||U^{-1}||^2} = \frac{(1 + c_1^2 + \cdots + c_m^2)(a_1^2 + \cdots + a_{k-1}^2) + (c_1^2 + \cdots + c_m^2) + n}{(a_1^2 + \cdots + a_{k-1}^2) + (c_1^2 + \cdots + c_m^2) + n}.$$

□

Corollary 4.8. *Suppose that we have a monic polynomial* $p(x) = q(x) + c_1 x^k q(x) + c_2 x^{2k} q(x) + \cdots + c_m x^{mk} q(x) + x^{(m+1)k}$ *over* $\mathbb{R}$*, with*

$$q(x) = a_{k-1}x^{k-1} + a_{k-2}x^{k-2} + \cdots + a_1 x + 1$$

and $n = (m+1)k$*. Let* $U = C_n(k, k, \ldots, k)$ *be the striped companion matrix in lower*

Hessenberg form to $p(x)$ as given below:

$$U = \left[\begin{array}{cccc|c}
\vec{0} & & I_{k-1} & & 0 \\
\hline
-c_m & -c_m a_1 & \dots & -c_m a_{k-1} & \\
0 & 0 & \dots & 0 & \\
\vdots & \vdots & \dots & \vdots & \\
0 & 0 & \dots & 0 & \\
-c_i & -c_i a_1 & \dots & -c_i a_{k-1} & \\
0 & 0 & \dots & 0 & \\
\vdots & \vdots & \dots & \vdots & \\
0 & 0 & \dots & 0 & I_{n-k-1} \\
-c_2 & -c_2 a_1 & \dots & -c_2 a_{k-1} & \\
0 & 0 & \dots & 0 & \\
\vdots & \vdots & \dots & \vdots & \\
0 & 0 & \dots & 0 & \\
-c_1 & -c_1 a_1 & \dots & -c_1 a_{k-1} & \\
0 & 0 & \dots & 0 & \\
\vdots & \vdots & \dots & \vdots & \\
0 & 0 & \dots & 0 & \\
-1 & -a_1 & \dots & -a_{k-1} & \vec{0}^T
\end{array}\right].$$

If $(1 + c_1^2 + \cdots + c_m^2)$ is sufficiently large, then for any Fiedler companion matrix M to $p(x)$, it is the case that

$$\left[\frac{\kappa(M)}{\kappa(U)}\right]^2 = \frac{||M^{-1}||^2}{||U^{-1}||^2} \approx (a_1^2 + \cdots + a_{k-1}^2 + 1).$$

Proof. By Theorem 4.7

$$\left[\frac{\kappa(M)}{\kappa(U)}\right]^2 = \frac{||M^{-1}||^2}{||U^{-1}||^2} = \frac{(1 + c_1^2 + \cdots + c_m^2)(a_1^2 + \cdots + a_{k-1}^2) + (c_1^2 + \cdots + c_m^2) + n}{(a_1^2 + \cdots + a_{k-1}^2) + (c_1^2 + \cdots + c_m^2) + n}.$$

Let $C = (c_1^2 + \cdots + c_m^2)$, and let $A = (a_1^2 + \cdots + a_{k-1}^2)$. Then

$$\lim_{C\to\infty} \left[\frac{\kappa(M)}{\kappa(U)}\right]^2 = \lim_{C\to\infty} \frac{(C+1)A + C + n}{A + C + n}.$$

If we divided all the terms in the denominator and the numerator by $(C+1)$ we are left with

$$\lim_{C\to\infty} \frac{A + \frac{C}{C+1} + \frac{n}{C+1}}{\frac{A}{C+1} + 1 + \frac{n}{C+1}} = \frac{A + 1 + 0}{0 + 1 + 0} = a_1^2 + \cdots + a_{k-1}^2 + 1.$$

□

The previous result shows that the ratio of the condition numbers of the companion matrices can be arbitrarily large when we fix the rank of the striped R-submatrix from Theorem 2.13 to be 1. The rank 1 condition is not a requirement for a striped companion matrix to have a lower condition number in general. In Example 4.5 we did not have an R-submatrix with rank 1, but we were able to create a striped companion matrix that had better condition number than the Fiedler matrices. Example 4.6 demonstrates how large we can make the ratio.

4.2 Striped Companion Matrices with Different Sized Stripes

So far in this chapter, we have considered only striped companion matrices that have all the same sized stripes. We wish to consider also when the stripes are not all the same size. Recall the structure of any companion matrix A in lower Hessenberg form to some characteristic polynomial $p(x) = x^n + a_{n-1}x^{n-1} + \cdots + a_1x + a_0$ from Lemma 2.26

$$A = \left[\begin{array}{c|c|c} \vec{0} & I_m & 0 \\ \hline \vec{u} & H & I_{n-m-1} \\ \hline -a_0 & \vec{y}^T & \vec{0}^T \end{array}\right]. \tag{4.2}$$

Consider now some striped companion matrix $U = C_n(\mathbf{s})$ where $\mathbf{s} = (s_1, s_2, \ldots, s_r)$, $s_r \leq s_i \leq s_1$ for $i \in \{2, \ldots, r-1\}$. Then the bottom stripe will fall completely in the row vector $\vec{y}^T$, except the $-a_0$ entry, followed by $s_1 - s_r$ zeros. The $\vec{u}$ column vector of the striped companion matrix U will have the first entry of each nonzero stripe according to (4.2), and zeros elsewhere corresponding to the zero stripes of U. If $\mathbf{s} = (s_1, s_2, \ldots, s_r)$, then there will be r nonzero entries in the column vector $\vec{u}$ of U which we will call a_{ℓ_j} for $j \in \{1, \ldots, r-1\}$ since $a_{\ell_r} = -a_0$. With this notation, we consider the following theorem.

Theorem 4.9. *Consider the monic polynomial:*

$$p(x) = x^n + a_{n-1}x^{n-1} + a_{n-1}x^{n-2} + \cdots + a_1x + 1,$$

where n is any positive integer, and $a_1, \ldots, a_{n-1} \in \mathbb{R}$. Consider any striped companion matrix $U = C_n(\boldsymbol{s})$ where $\boldsymbol{s} = (s_1, s_2, \ldots, s_r)$, $s_r < s_i \leq s_1$ for all $i \in \{2, \ldots, r-1\}$. Assume that each nonzero stripe has a nonzero entry in $\vec{u}$, and let a_{ℓ_j} for $j \in \{1, \ldots, r-1\}$ be the the nonzero coefficients of the polynomial $p(x)$ that fall in the column vector $\vec{u}$ of the matrix U. If M is any lower Hessenberg Fiedler companion matrix to the polynomial $p(x)$, then $\kappa_F(U) \leq \kappa_F(M)$ if the entries of U satisfy:

$$|a_k a_{\ell_j} - a_{k+\ell_j}| \leq |a_{k+\ell_j}| \text{ for } k \in \{1, \ldots, t\}$$

Proof. By definition, $\kappa_F(U) \leq \kappa_F(M)$ if and only if

$$||U|| \cdot ||U^{-1}|| \leq ||M|| \cdot ||M^{-1}||.$$

Let $s_r = t$. It suffices to show that $||U^{-1}||^2 \leq ||M^{-1}||^2$ if the entries of U satisfy:

$$|a_k a_{\ell_j} - a_{\ell_j+k}| \leq |a_{\ell_j+k}|$$

for $k \in \{1, \dots, t\}$, and $j \in \{1, \dots, r-1\}$. U can be represtend as follows:

$$\left[\begin{array}{c|c} \vec{0} \;\; I_m & 0 \\ \hline R & \begin{array}{c} I_{n-m-1} \\ \vec{0}^T \end{array} \end{array}\right],$$

where

$$R = \begin{bmatrix}
-a_{\ell_1} & -a_{\ell_1+1} & \dots & -a_{\ell_1+t} & \bigstar & \dots & \bigstar & \dots & -a_{n-1} \\
0 & 0 & \dots & 0 & 0 & \dots & 0 & \dots & 0 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
0 & 0 & \dots & 0 & 0 & \dots & 0 & \dots & 0 \\
-a_{\ell_{r-j}} & -a_{\ell_{r-j}+1} & \dots & -a_{\ell_{r-j}+t} & \bigstar & \dots & \bigstar & \dots & \bigstar \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
0 & 0 & \dots & 0 & 0 & \dots & 0 & \dots & 0 \\
-a_{\ell_{r-2}} & -a_{\ell_{r-2}+1} & \dots & -a_{\ell_{r-2}+t} & \bigstar & \dots & \bigstar & \dots & \bigstar \\
0 & 0 & \dots & 0 & 0 & \dots & 0 & \dots & 0 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
0 & 0 & \dots & 0 & 0 & \dots & 0 & \dots & 0 \\
-a_{\ell_{r-1}} & -a_{\ell_{r-1}+1} & \dots & -a_{\ell_{r-1}+t} & \bigstar & \dots & \bigstar & \dots & \bigstar \\
0 & 0 & \dots & 0 & 0 & \dots & 0 & \dots & 0 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
0 & 0 & \dots & 0 & 0 & \dots & 0 & \dots & 0 \\
-1 & -a_1 & \dots & -a_t & 0 & \dots & 0 & \dots & 0
\end{bmatrix},$$

where $\bigstar$s are either coefficients or zeros that fall outside of the $(n-t) \times (t+1)$ submatrix in the bottom left corner of U. Note from Lemma 2.26 that, since $a_0 = 1$,

$$U^{-1} = \left[\begin{array}{cccc|c|c}
-a_1 & -a_2 & \ldots & -a_t & \vec{0}^T & -1 \\ \hline
\multicolumn{4}{c|}{I_t} & 0 & \vec{0} \\ \hline
a_1a_{\ell_1} - a_{\ell_1+1} & a_2a_{\ell_1} - a_{\ell_1+2} & \ldots & a_ta_{\ell_1} - a_{\ell_1+t} & & -a_{\ell_1} \\
0 & 0 & \ldots & 0 & & 0 \\
\vdots & \vdots & \ldots & \vdots & & \vdots \\
0 & 0 & \ldots & 0 & & 0 \\
a_1a_{\ell_{r-j}} - a_{\ell_{r-j}+1} & a_2a_{\ell_{r-j}} - a_{\ell_{r-j}+2} & \ldots & a_ta_{\ell_{r-j}} - a_{\ell_{r-j}+t} & & -a_{\ell_{r-j}} \\
\vdots & \vdots & \ldots & \vdots & & \vdots \\
\vdots & \vdots & \ldots & \vdots & & \vdots \\
0 & 0 & \ldots & 0 & W & 0 \\
a_1a_{\ell_{r-2}} - a_{\ell_{r-2}+1} & a_2a_{\ell_{r-2}} - a_{\ell_{r-2}+2} & \ldots & a_ta_{\ell_{r-2}} - a_{\ell_{r-2}+t} & & -a_{\ell_{r-2}} \\
0 & 0 & \ldots & 0 & & 0 \\
\vdots & \vdots & \ldots & \vdots & & \vdots \\
0 & 0 & \ldots & 0 & & 0 \\
a_1a_{\ell_{r-1}} - a_{\ell_{r-1}+1} & a_2a_{\ell_{r-1}} - a_{\ell_{r-1}+2} & \ldots & a_ta_{\ell_{r-1}} - a_{\ell_{r-1}+t} & & -a_{\ell_{r-1}} \\
0 & 0 & \ldots & 0 & & 0 \\
\vdots & \vdots & \ldots & \vdots & & \vdots \\
0 & 0 & \ldots & 0 & & 0
\end{array}\right].$$

Since the ★s fall outside the first $(t+1)$ columns of U, when we take the inverse of U, all of the nonzero ★s will be monomial elements in U^{-1}, represented by the nonzero elements of W. Note that when we compare condition numbers of the striped and Fiedler companion matrices all of the elements of $*$ in U^{-1} will cancel out with some a_i in M^{-1} for $i \in \{1, \ldots n-1\}$. By Corollary 3.8, when $a_0 = 1$, then

$$||M^{-1}||_F^2 = \sum_{i=1}^{n-1} |a_i|^2 + n$$

for any Fiedler companion matrix M to $p(x)$. From Lemma 2.26, both U^{-1} and M^{-1} will have n ones (as $a_0 = 1$). And from our hypothesis we have that

$$|a_k a_{\ell_j} - a_{\ell_j+k}| \leq |a_{\ell_j+k}|$$

for $k \in \{1, \ldots, t\}$, and $j \in \{1, \ldots, r-1\}$. Thus, $||U^{-1}||^2 \leq ||M^{-1}||^2$. □

Theorem 4.10. *Consider the monic polynomial:*

$$p(x) = x^n + a_{n-1}x^{n-1} + a_{n-1}x^{n-2} + \cdots + a_1x + a_0$$

where $a_0 = 1$, and $a_1, \ldots, a_{n-1} \in \mathbb{R}$. Let U be a striped companion matrix to the polynomial $p(x)$. If

$$U = \left[\begin{array}{c|c}
\vec{0}\ \ I_m & 0 \\ \hline
R & \begin{array}{c} I_{n-m-1} \\ \vec{0}^T \end{array}
\end{array}\right]$$

with R a submatrix, defined in Theorem 2.13, of the form

$$R = \left[\begin{array}{c|c} B & \begin{array}{c} * \\ \hline 0\cdots 0 \end{array} \end{array}\right],$$

and B is given by

$$B = \begin{bmatrix} -a_{\ell_1} & -a_{\ell_1+1} & \dots & -a_{\ell_1+t} \\ 0 & 0 & \dots & 0 \\ \vdots & \vdots & \dots & \vdots \\ 0 & 0 & \dots & 0 \\ -a_{\ell_{r-j}} & -a_{\ell_{r-j}+1} & \dots & -a_{\ell_{r-j}+t} \\ \vdots & \vdots & \dots & \vdots \\ 0 & 0 & \dots & 0 \\ -a_{\ell_{r-2}} & -a_{\ell_{r-2}+1} & \dots & -a_{\ell_{r-2}+t} \\ 0 & 0 & \dots & 0 \\ \vdots & \vdots & \dots & \vdots \\ 0 & 0 & \dots & 0 \\ -a_{\ell_{r-1}} & -a_{\ell_{r-1}+1} & \dots & -a_{\ell_{r-1}+t} \\ 0 & 0 & \dots & 0 \\ \vdots & \vdots & \dots & \vdots \\ 0 & 0 & \dots & 0 \\ -1 & -a_1 & \dots & -a_t \end{bmatrix},$$

then

$$\sum_{j=1}^{r-1}\left(\sum_{i=1}^{t}|a_i a_{\ell_j} - a_{\ell_j+i}|^2)\right) = 0 \quad \Leftrightarrow \quad \text{rank}(B) = 1.$$

Proof. Let

$$B' = \begin{bmatrix} -a_{\ell_1} & -a_{\ell_1+1} & \dots & -a_{\ell_1+t} \\ -a_{\ell_{r-j}} & -a_{\ell_{r-j}+1} & \dots & -a_{\ell_{r-j}+t} \\ -a_{\ell_{r-2}} & -a_{\ell_{r-2}+1} & \dots & -a_{\ell_{r-2}+t} \\ -a_{\ell_{r-1}} & -a_{\ell_{r-1}+1} & \dots & -a_{\ell_{r-1}+t} \\ -1 & -a_1 & \dots & -a_t \end{bmatrix}.$$

be the matrix consisting of the nonzero rows of B. ($\Rightarrow$) Note that for $1 \leq i \leq t$, and $1 \leq j \leq r-1$, $\sum_{j=1}^{r-1}(\sum_{i=1}^{t}|a_i a_{\ell_j} - a_{\ell_j+i}|^2) = 0$ if and only if $|a_i a_{\ell_j} - a_{\ell_j+i}| = 0$ if and only if for all the 2×2 submatrices of B' involving the first column, and the last row of B' have a determinant of zero, i.e.:

$$\det\begin{bmatrix} -a_{\ell_j} & -a_{\ell_j+i} \\ -1 & -a_i \end{bmatrix} = 0, \ \text{ for } i \in \{1,\dots,t\}, \text{ and } j \in \{1,\dots,r-1\}.$$

We choose one column and one row, and consider the corners of that square. First consider just the bottom two rows of B'.

$$B'' = \begin{bmatrix} -a_{\ell_{r-1}} & -a_{\ell_{r-1}+1} & \dots & -a_{\ell_{r-1}+t} \\ -1 & -a_1 & \dots & -a_t \end{bmatrix}.$$

If the 2×2 submatrices of B'' involving the first column have determinant 0, that is equivalent to saying that each column of B'' is a scalar multiple of $[-a_{\ell_{r-1}} \quad -1]^T$. So $\text{rank}(B'') = 1$, which is true if and only if row 2 is a multiple of row 1. We can do the same process for every row of B' in fact. And so every row of B' is a multiple of the bottom row, which implies that $\text{rank}(B') = 1$ which implies $\text{rank}(B) = 1$.

($\Leftarrow$) If the rank of B is q, that is equivalent to saying that the largest square submatrix of B with nonzero determinant is a $q \times q$ matrix [8]. So if $q = 1$, then every 2×2 submatrix of B has a zero determinant, which means that $|a_i a_{\ell_j} - a_{\ell_j+i}| = 0$ for $1 \leq i \leq t$, and $1 \leq j \leq r-1$. This implies that $\sum_{j=1}^{r-1}(\sum_{i=1}^{t} |a_i a_{\ell_j} - a_{\ell_j+i}|^2) = 0$. □

In this chapter we saw that under the right hypotheses striped companion matrices can have a smaller condition number to any Fiedler companion matrix. Striped companion matrices are the main class of companion matrices we studied in this paper, but these results give hope to potentially generalizing some of these results to compare the condition number of any sparse companion matrix to the Fiedler companion matrices.

Chapter 5

Singular Values of Companion Matrices

In this chapter, we change gears to discussing some of the results about singular values and the spectral condition number previously studied by De Téran et. al. in 2012 [1]. The spectral condition number uses the spectral norm, which uses eigenvalues to evaluate the size of a matrix. We will discuss in more detail through this chapter how to find the singular values of a matrix, and how the singular values will help in the calculation of the spectral condition number of a matrix.

Definition 5.1. Let A be an $m \times n$ matrix. The *conjugate transpose* of the matrix A, denoted A^*, is obtained by taking the complex conjugate of each entry of A, and then transposing the matrix. That is,

$$A^* = \bar{A}^T.$$

Note that for this project we work primarily in the real numbers $\mathbb{R}$. This fact implies that the conjugate transpose is merely the transpose, i.e., $A^* = \bar{A}^T = A^T$.

Definition 5.2. A matrix A is called *unitary* if its conjugate transpose is also its inverse. That is,

$$A^T A = AA^T = I.$$

Note that $A^T A$ is a symmetric matrix. Thus the eigenvalues of $A^T A$ are real numbers. Since $\vec{v}^T\vec{v} \geq 0$ for any real vector $\vec{v}$, setting $\vec{v} = A\vec{x}$ gives $\vec{x}^T A^T A\vec{x} \geq 0$. If λ is an eigenvalue of $A^T A$ then $A^T A\vec{x} = \lambda\vec{x}$ for a real nonzero vector $\vec{x}$ and hence $\lambda\vec{v}^T\vec{v} = \vec{x}^T A^T A\vec{x} \geq 0$. Thus the eigenvalues of $A^T A$ are nonnegative real numbers.

Definition 5.3. The *singular values* of a matrix A are the square roots of the eigenvalues of $A^T A$.

Definition 5.4. [5] The *spectral norm* is defined as the square root of the maximum eigenvalue of $A^T A$. That is,

$$||A||_2 = \sqrt{\max\{\lambda \mid \lambda \text{ is an eigenvalue of } A^T A\}}.$$

Thus the spectral norm of A is the largest singular value of A.

We can show that the spectral norm of any matrix A is in fact a matrix norm. 0

- $||A||_2 = 0$ if and only if $A = 0$ is true since a zero matrix can only have 0 eigenvalues.
- Next we wish to show that $||kA||_2 = |k| \cdot ||A||_2$. Suppose that λ_{max} is the largest eigenvalue of A. Note that if $\vec{x}$ is an eigenvector of A then $A\vec{x} = \lambda\vec{x}$ and $kA\vec{x} = k\lambda\vec{x}$, and hence, $(k\lambda)$ is an eigenvector of (kA). So

$$\begin{aligned} ||kA||_2 &= \sqrt{\max\{\lambda \mid \lambda \text{ is an eigenvalue of } (kA)^T(kA)\}} \\ &= \sqrt{k^2 \cdot \max\{\lambda \mid \lambda \text{ is an eigenvalue of } A^TA\}} \\ &= |k| \cdot \sqrt{\max\{\lambda \mid \lambda \text{ is an eigenvalue of } A^TA\}}. \end{aligned}$$

  And so $||kA||_2 = |k| \cdot ||A||_2$.

- $||A + B||_2 \leq ||A||_2 + ||B||_2$ follow from Weyl's Theorem (see [5, Theorem 4.3.1]).

Definition 5.5. The *spectral condition number* of a matrix A, denoted $\kappa_2(A)$, is $\kappa_2(A) = ||A||_2 \cdot ||A^{-1}||_2$.

Example 5.6. Consider the $n \times n$ matrix given by $[h_{ij}] = \dfrac{1}{(i + j - 1)}$, called a *Hilbert* matrix. For example, when $n = 4$

$$H_4 = \begin{bmatrix} 1 & 1/2 & 1/3 & 1/4 \\ 1/2 & 1/3 & 1/4 & 1/5 \\ 1/3 & 1/4 & 1/5 & 1/6 \\ 1/4 & 1/5 & 1/6 & 1/7 \end{bmatrix}.$$

When one calculates the spectral condition number $\kappa_2(H_4)$, we see that even for a small n, the condition number can be quite large since $\kappa_2(H_4) \approx 15,500$.

Lemma 5.7. [6] *Let A be an $n \times n$ invertible matrix. Suppose also that s_1 and s_2 are the maximum and minimum singular values of A respectively. Then*

$$\kappa_2(A) = \frac{s_1}{s_2}.$$

Proof. By the definition of the spectral norm, $||A||_2 = s_1$. For this proof we need the following statement: If A is invertible, then

$$\lambda \text{ is an eigenvalue of } A \Leftrightarrow \lambda^{-1} \text{ is an eigenvalue of } A^{-1}.$$

This is true since for invertible matrices A,

$$\begin{aligned} A\vec{x} &= \lambda\vec{x} \\ \Leftrightarrow A^{-1}(A\vec{x}) &= \lambda A^{-1}\vec{x} \\ \Leftrightarrow \lambda^{-1}(A^{-1}A)\vec{x} &= (\lambda^{-1}\lambda)A^{-1}\vec{x} \\ \Leftrightarrow \lambda^{-1}\vec{x} &= A^{-1}\vec{x}. \end{aligned}$$

Now consider

$$\begin{aligned}||A^{-1}||_2^2 &= \max\{\lambda \mid \lambda \text{ is an eigenvalue of } (A^{-1})^T A^{-1}\} \\ &= \max\{\lambda \mid \lambda \text{ is an eigenvalue of } (AA^T)^{-1}\} = \max\{\frac{1}{\lambda} \mid \frac{1}{\lambda} \text{ is an eigenvalue of } (AA^T)\} \\ &= \frac{1}{\min\{\lambda \mid \lambda \text{ is an eigenvalue of } (AA^T)\}} = \frac{1}{s_2}.\end{aligned}$$

And so we have

$$\kappa_2(A) = ||A||_2 ||A^{-1}||_2 = \frac{s_1}{s_2}.$$

□

The next theorem exhibits a relation between $\kappa_2(A)$ and $\kappa_F(A)$.

Theorem 5.8. [10] *For any $n \times n$ matrix A,*

$$||A||_2 \leq ||A||_F \leq \sqrt{n} \cdot ||A||_2.$$

Consequently

$$\kappa_2(A) \leq \kappa_F(A) \leq n \cdot \kappa_2(A). \tag{5.1}$$

Proof. Recall that the *trace* of a matrix $A = [a_{ij}]$ is given by

$$\text{tr}(A) = \sum_{i=1}^{n} a_{ii},$$

and also that $\text{tr}(A)$ is the sum of the eigenvalues of A. Suppose then, that the eigenvalues of $A^T A$ are

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n \geq 0.$$

Then

$$||A||_2 = \sqrt{\lambda_1} \leq \sqrt{\sum_{i=1}^{n} \lambda_i} = \sqrt{\text{tr}(A^T A)} = ||A||_F,$$

and

$$||A||_F = \sqrt{\sum_{i=1}^{n} \lambda_i} \leq \sqrt{n\lambda_1} = \sqrt{n}\sqrt{\lambda_1} = \sqrt{n}||A||_2.$$

The proof for (5.1) is similar. □

The next goal is to describe some ideas from linear algebra to help us factor companion matrices. Theorem 5.16 will allow us to see how these factored matrices are going to aid us in the calculation of the singular values of a sparse companion matrix.

Lemma 5.9. [8, Theorem 3.29] *Let V be a vector space of dimension n over a $\mathbb{R}$, and let*

$$B = \{\vec{b_1}, \vec{b_2}, \dots, \vec{b_n}\}$$

be an ordered basis for V. Then for every $v \in V$, there is a unique linear combination of the basis vectors that equals v, that is, there exists unique $\alpha_1, \dots, \alpha_n \in \mathbb{R}$, such that

$$v = \alpha_1 \vec{b_1} + \cdots + \alpha_n \vec{b_n}.$$

Definition 5.10. The *coordinate vector* of v relative to an ordered basis B is the ordered tuple of coefficients corresponding to each v, that is,

$$[v]_B = (\alpha_1, \dots, \alpha_n).$$

Lemma 5.11. [7] *Let B be an $m \times n$ matrix with rank r. Then there exists a matrix $L \in \mathbb{R}^{m\times r}$ and a matrix $R \in \mathbb{R}^{r\times n}$ such that $B = LR$.*

Proof. Let

$$B = \begin{bmatrix} \vec{b}_1 \\ \vec{b}_2 \\ \vdots \\ \vec{b}_n \end{bmatrix}$$

where $\{\vec{b}_1, \dots, \vec{b}_n\}$ are the row vectors of B. If $r = \text{rank}(B)$, then there exists r row vectors $\beta = \{\vec{b_{i_1}}, \dots, \vec{b_{i_r}}\}$ that form a basis for the row space of B. We can represent each row of B with a list of coefficients corresponding to the linear combination it represents over the whole field. Applying Lemma 5.10, for each $i = 1, \dots, n$, let $[\vec{b_i}]_\beta = (c_1^i, \dots, c_r^i)$ be the coordinates of $\vec{b_i}$ with respect to the ordered basis β. With this notation we can simply choose

$$L = \begin{bmatrix} [\vec{b}_1]_\beta \\ [\vec{b}_2]_\beta \\ \vdots \\ [\vec{b}_n]_\beta \end{bmatrix} \quad \text{and} \quad R = \begin{bmatrix} \vec{b_{i_1}} \\ \vec{b_{i_2}} \\ \vdots \\ \vec{b_{i_r}} \end{bmatrix}, \text{ then } LR = B.$$

□

This next result is a generalization of [1, Lemma 6.4].

Lemma 5.12. *Let $p(x) = x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0$ be a polynomial over $\mathbb{R}$. Suppose A is an $n \times n$ sparse companion matrix to $p(x)$ in lower Hessenberg form. Let*

$$U = \begin{bmatrix} 0 & 1 & 0 & 0 & \dots & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & \dots & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & \dots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \dots & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & \dots & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & \dots & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & \dots & 0 & 0 & 0 \end{bmatrix}.$$

Then there exists a matrix $L \in \mathbb{R}^{n\times r}$, and a matrix $R \in \mathbb{R}^{r\times n}$ where r is the rank of the matrix $A - U$, such that $A = U + LR$ where

Proof. The result follows by applying Lemma 5.11 to the matrix $A - U$. □

Example 5.13. Consider the 9×9 companion matrix

$$A = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -a_7 & -a_8 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & -a_5 & -a_6 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & -a_3 & -a_4 & 0 & 0 & 0 & 0 & 1 \\ -a_0 & -a_1 & -a_2 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}.$$

Then

$$A - U = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -a_7 & -a_8 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -a_5 & -a_6 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -a_3 & -a_4 & 0 & 0 & 0 & 0 & 0 \\ -a_0 - 1 & -a_1 & -a_2 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}.$$

Our task is to decompose the matrix $A - U$ into a product LR. Using row reductions, we can show $A - U$ has rank at most 3. Without loss of generality, let the fourth row of $A - U$ be dependent upon the rows 6 and 8, and we suppose that $\operatorname{rank}(A - U) = 3$. Then the rows of

$$R = \begin{bmatrix} 0 & 0 & -a_5 & -a_6 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -a_3 & -a_4 & 0 & 0 & 0 & 0 & 0 \\ -a_0 - 1 & -a_1 & -a_2 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

are a basis for the row space of $A - U$. Next, since the fourth row is a linear combination of the sixth and eighth, there exists coefficients s and t such that $a_7 = sa_5 + ta_3$ and $a_8 = sa_6 + ta_4$.

Let

$$L = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ s & t & 0 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

Then

$$\begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ s & t & 0 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 0 & 0 & -a_5 & -a_6 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -a_3 & -a_4 & 0 & 0 & 0 & 0 & 0 \\ -a_0-1 & -a_1 & -a_2 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -a_7 & -a_8 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -a_5 & -a_6 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -a_3 & -a_4 & 0 & 0 & 0 & 0 & 0 \\ -a_0-1 & -a_1 & -a_2 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}.$$

Example 5.14. Consider the 9×9 companion matrix

$$A = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ -a_6 & -a_7 & -a_8 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ -a_3 & -a_4 & -a_5 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ -a_0 & -a_1 & -a_2 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}.$$

The matrix $A - U$ also has rank at most 3. Suppose $A - U$ has rank 3. Let

$$L = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad \text{and} \quad R = \begin{bmatrix} -a_6 & -a_7 & -a_8 & 0 & 0 & 0 & 0 & 0 & 0 \\ -a_3 & -a_4 & -a_5 & 0 & 0 & 0 & 0 & 0 & 0 \\ -a_0-1 & -a_1 & -a_2 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}.$$

Then $LR = A - U$.

Theorem 5.15. [5, Theorem 1.3.20] *Let $A \in \mathbb{R}^{m\times n}$ and $B \in \mathbb{R}^{n\times m}$ with $m \leq n$. Then BA has the same eigenvalues of AB, counting multiplicity, together with and additional $n-m$ eigenvalues equal to 0; that is $p_{BA}(x) = x^{n-m}p_{AB}(x)$. If $m = n$ and at least one of A or B is nonsingular, then AB and BA are similar matrices.*

With this theorem we can state and prove the following result from De Terán et. al [1].

Theorem 5.16. [1, Lemma 6.3] *Let $A = U + LR \in \mathbb{R}^{n\times n}$, where $U \in \mathbb{R}^{n\times n}$ is a unitary matrix, $L \in \mathbb{R}^{n\times r}$, and $R \in \mathbb{R}^{r\times n}$. If $2r < n$, then A has at least $n-2r$ singular values equal to 1, and the other $2r$ singular values are the square roots of the eigenvalues of the matrix*

$$H = I + \left[\frac{R}{L^TU}\right]\left[\begin{array}{c|c} U^TL + R^TL^TL & R^T \end{array}\right].$$

Proof. The singular values of A are the square roots of the eigenvalues of A^TA. So we first compute A^TA. We have

$$\begin{aligned} A^TA =&(U+LR)^T(U+LR) = U^TU + R^TL^TU + U^TLR + R^TL^TLR \\ =&I + [U^TL + R^TL^TL \quad R^T]\begin{bmatrix} R \\ L^TU \end{bmatrix} = \ I + \tilde{L}\tilde{R}. \end{aligned}$$

Since $\tilde{L} \in \mathbb{R}^{n\times 2r}$ and $\tilde{R} \in \mathbb{R}^{2r\times n}$, then $\operatorname{rank}(\tilde{L}\tilde{R}) \leq 2r$. We have that $\tilde{L}\tilde{R} \in \mathbb{R}^{n\times n}$, so by Theorem 5.15, its eigenvalues are the same as the eigenvalues of $\tilde{R}\tilde{L} \in \mathbb{R}^{2r\times 2r}$, plus $n-2r$ eigenvalues equal to 0 . So the eigenvalues of $I+\tilde{R}\tilde{L} \in \mathbb{R}^{2r\times 2r}$ together with $n-2r$ eigenvalues equal to 1 are the eigenvalues of $H = I + \tilde{L}\tilde{R} \in \mathbb{R}^{n\times n}$, which are the squares of the singular values of A. □

Example 5.17. Consider the following companion matrix to the polynomial $x^6 - 6x^5 - 3x^4 - 4x^3 - 2x^2 - 2x - 2$:

$$M = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 3 & 6 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 2 & 4 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 2 & 2 & 0 & 0 & 0 & 0 \end{bmatrix}.$$

Setting

$$L = \begin{bmatrix} 0 \\ 3 \\ 0 \\ 2 \\ 0 \\ 1 \end{bmatrix} \quad \text{and} \quad R = \begin{bmatrix} 1 & 2 & 0 & 0 & 0 & 0 \end{bmatrix},$$

we decompose this matrix into

$$
\begin{aligned}
M &= U + LR \\
&= \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} + \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 3 & 6 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 2 & 4 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 2 & 0 & 0 & 0 & 0 \end{bmatrix} \\
&= \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} + \begin{bmatrix} 0 \\ 3 \\ 0 \\ 2 \\ 0 \\ 1 \end{bmatrix} \begin{bmatrix} 1 & 2 & 0 & 0 & 0 & 0 \end{bmatrix}.
\end{aligned}
$$

Next we compute the matrix H from Lemma 5.16:

$$
\begin{aligned}
H &= I + \begin{bmatrix} R \\ L^T U \end{bmatrix} \begin{bmatrix} U^T L + R^T L^T L & R^T \end{bmatrix} \\
&= I + \begin{bmatrix} 1 & 2 & 0 & 0 & 0 & 0 \\ 1 & 0 & 3 & 0 & 2 & 0 \end{bmatrix} \begin{bmatrix} 15 & 1 \\ 28 & 2 \\ 3 & 0 \\ 0 & 0 \\ 2 & 0 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} 72 & 5 \\ 28 & 2 \end{bmatrix}.
\end{aligned}
$$

The eigenvalues of H are

$$\lambda_1 = 37 + \sqrt{1365} \quad , \quad \lambda_2 = 37 - \sqrt{1365},$$

which implies that the singular values of M are

$$s_1 = \sqrt{37 + \sqrt{1365}} \quad , \quad s_2 = \sqrt{37 - \sqrt{1365}},$$

and the other four singular values of the matrix are 1's according to Theorem 5.16.

We want to find the condition number of M, which requires the maximum singular values of M and M^{-1}. Now

$$M^{-1} = \tilde{U} + \tilde{L}\tilde{R}$$

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

The right most matrix, $M^{-1} - \tilde{U}$ has rank 1, as all the rows are scalar multiples of each other. Thus

$$M^{-1} - \tilde{U} = \begin{bmatrix} -1 \\ 0 \\ -3 \\ 0 \\ -2 \\ 0 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & \frac{1}{2} \end{bmatrix}.$$

We then compute $\tilde{H}$:

$$\tilde{H} = I + \begin{bmatrix} \tilde{R} \\ \tilde{L}^T\tilde{U} \end{bmatrix} \begin{bmatrix} \tilde{U}^T\tilde{L} + \tilde{R}^T\tilde{L}^T\tilde{L} & \tilde{R}^T \end{bmatrix}$$

$$\tilde{H} = I + \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & \frac{1}{2} \\ 0 & -3 & 0 & -2 & 0 & -1 \end{bmatrix} \begin{bmatrix} 14 & 1 \\ -3 & 0 \\ 0 & 0 \\ -2 & 0 \\ 0 & 0 \\ 6 & \frac{1}{2} \end{bmatrix} = I + \begin{bmatrix} 17 & \frac{5}{4} \\ 7 & -\frac{1}{2} \end{bmatrix} = \begin{bmatrix} 18 & \frac{5}{4} \\ 7 & \frac{1}{2} \end{bmatrix}.$$

When we find the eigenvalues of $\tilde{H}$ we get

$$\tilde{\lambda_1} = \frac{1}{4}(37 + \sqrt{1365}) \quad , \quad \tilde{\lambda_2} = \frac{1}{4}(37 - \sqrt{1365}),$$

which implies that the singular values of M^{-1} are

$$\tilde{s_1} = \sqrt{\frac{1}{4}(37 + \sqrt{1365})} \quad , \quad \tilde{s_2} = \sqrt{\frac{1}{4}(37 - \sqrt{1365})}.$$

Again, the other four singular values of the matrix are 1's according to Theorem 5.16. Thus we get that

$$\begin{aligned} \kappa_2(M) = ||M||_2||M^{-1}||_2 =& (\sqrt{37 + \sqrt{1365}})(\sqrt{\frac{1}{4}(37 + \sqrt{1365})}) \\ =& \frac{1}{2}(37 + \sqrt{1365}) \\ \approx& \, 36.972953. \end{aligned}$$

While we used Theorem 5.16 for illustration purposes to calculate $||M^{-1}||_2$, we could also use Lemma 5.7. In particular Lemma 5.7 implies

$$||M^{-1}||_2 = \frac{1}{s_2} \text{ and } \kappa_2(M) = \frac{s_1}{s_2}.$$

If we can find all the eigenvalues of an $n \times n$ matrix, we can find the spectral condition number of a sparse companion matrix according to the definition. Theorem 5.16 tells us that we can find the condition number of a sparse companion matrix by finding the eigenvalues of a smaller $2r \times 2r$ matrix.

Chapter 6

Conclusion/Open Questions

Throughout the majority of this project we explored companion matrices, and the conditions under which a sparse companion matrix can have a better condition number than Fiedler companion matrices. We focused on the striped companion matrices and discovered that under the right circumstances we could successfully do this. Given the formula of the condition number, we consider the following question:

Question 6.1. *For some striped companion matrix $U = C_n(\boldsymbol{s})$, does there exist a general formula for $||U||_F$ and $||U^{-1}||_F$?*

If we could find explicit formulas for these, we could easily compare the condition number $\kappa_F(U)$ with any Fiedler companion matrix. And with an explicit formula it might be possible to construct an ordering for the condition number of the striped companion matrices, as De Téran et. al. did for the Fiedler companion matrices [1]. In chapter 4 we saw that we could also make the ratio of condition numbers for striped and Fiedler companion matrices as big as we wanted for sufficiently large constants when each striped was linearly dependant on one another. From this result, another natural question would be:

Question 6.2. *What other striped companion matrices $C_n(\boldsymbol{s})$ can we have the property that $\kappa(C) < \kappa(M)$ for every Fiedler matrix M, other than when the striped matrix has all linearly dependent stripes.*

As previously noted, this project compared the condition numbers of the Fiedler companion matrices with striped companion matrices. In future work, the next step would be to explore other classes of sparse companion matrices that we can compare to Fiedler companion matrices. The ultimate goal is to answer the following question:

Question 6.3. *Given a polynomial $p(x)$, can we determine which sparse companion matrices to $p(x)$ yield the smallest condition numbers*

Chapter 7

Appendix

Algorithm 7.1. [3, Algorithm 2.10] This algorithm finds a matrix P to transform M into a lower Hessenberg form. In particular, the matrix P has the property that PMP^T is in lower Hessenberg form if M is a sparse companion matrix.

```
P = nxn matrix
    j = 0;
    for i in range(n):
        for k in range(n):
            if A[k,i]==-an:
                j=i;
                break
    P[j,0] = 1;

    for r in range(1,n):
        for s in range(n):
            if A[j,s]==1:
                j=s;
                break
        P[s,r] = 1;
    return P;
```

Bibliography

[1] F. de Terán, F.M. Dopico, J. Pérez, Condition numbers for inversion of Fiedler companion matrices, *Linear Algebra and its Applications* 439 (2013) 944–981.

[2] B. Eastman, I.J. Kim, B.L Shader, K.N. Vander Meulen, Companion Matrix Patterns, *Linear Algebra and its Applications* 463 (2014) 255–272.

[3] B. Eastman, K.N. Vander Meulen, Pentadiagonal Companion Matrices, *Special Matrices* 4 (2016) 13–30.

[4] M. Fiedler, A Note on Companion Matrices, *Linear Algebra and its Applications* 372 (2003) 325–331.

[5] R.A. Horn, C.R. Johnson, *Matrix Analysis*, Cambridge University Press, Cambridge, (1985).

[6] C. Kenney, A.J. Laub, Controllability and Stability Radii for companion Form Systems, *Mathematics of Control, Signals and Systems* (1988) 239–256

[7] R. Piziak, P.L. Odell, Full Rank Factorization of Matrices, *Mathematics Magazine* 72 (1999) 193–201

[8] D. Poole, *Linear Algebra: A Modern Introduction*, Cengage Learning (2015).

[9] K.N. Vander Meulen, T. Vanderwoerd, Bounds on Polynomial Roots using Intercyclic Companion Matrices, *Linear Algebra and its Applications* 539 (2018) 94–116.

[10] D. Watkins, *Fundamentals of Matrix Computation*, Wiley (2010) 122–145