EIGENVALUES AND EIGENSPACES OF GENERAL LINEAR OPERATORS ON A FINITE DIMENSIONAL INNER-PRODUCT SPACE

BY

PAUL EUGENE HARPER

Bachelor of Science
Oregon State University
Corvallis, Oregon
1949

Master of Science
Oregon State University
Corvallis, Oregon
1950

Submitted to the Faculty of the Graduate School of
the Oklahoma State University
in partial fulfillment of the requirements
for the degree of
DOCTOR OF PHILOSOPHY
May, 1964

EIGENVALUES AND EIGENSPACES OF GENERAL LINEAR OPERATORS ON A FINITE DIMENSIONAL INNER-PRODUCT SPACE

Thesis Approved:

R B Deal
Thesis Adviser

John E Hoffman

Earl E Marshall

O. H. Hamilton

J. H. Boggs
Dean of the Graduate School

PREFACE

With the growth in speed and memory size of modern digital computers has come an increased interest in many mathematicians concerned with efficient solution of large matrix problems. The general subject of this thesis, which belongs to this category, is concerned with the difficult eigenvalue problem arising in the study of general linear operators. For their assistance and encouragement to stick with this problem, I am deeply indebted to Dr. Thomas C. Doyle and Dr. Edmond D. Cashwell both of the Los Alamos Scientific Laboratory, Los Alamos, New Mexico.

When I made the decision to return to graduate school, I was fortunate to be chosen by the Advisory Committee to participate in the LASL's Advanced Study Program. Financial support from the Laboratory, which is operated by the University of California under contract with the United States Atomic Energy Commission, is gratefully acknowledged by my family and myself. I wish to thank the many members of Laboratory Group T-1, of which Bengt G. Carlson is leader, for their assistance and general interest in my progress. A special note of gratitude is accorded to Mrs. Betty J. Simes, who typed the final draft of this thesis; to Mrs. Barbara C. Powell, who did the illustrations; and to Miss Katherine Heinz of the LASL Library for her assistance in locating and obtaining many reference publications.

For their courses and for their assistance while acting as my Graduate Advisory Committee, I am indebted to the following Oklahoma

State University faculty members: Professors Roy B. Deal, Jr., Olan H. Hamilton, John E. Hoffman, Carl E. Marshall, and Jan J. Tuma. Dr. Roy B. Deal, Jr., was my thesis adviser. His guidance and interest are gratefully acknowledged.

Even with all the help given by those listed above and by many others who remain unmentioned but also not forgotten, these last four years would have been even more disheartening without my family. To my boys, Ronny and Bryce and Dana, who never understood why I had to prefer a book in mathematics to almost all their suggestions for a game: I hope that I can make it up. To my lovely wife Patty, who never once complained and was always there to give a word of needed encouragement: I can only restate my deepest love and affection.

TABLE OF CONTENTS

CHAPTER I

INTRODUCTION

In this paper one is concerned with linear operators mapping an n-dimensional complex inner-product space X into itself. The space X is called "unitary" and is composed of the linear space of n-columns over the complex field C with a complex-valued inner product function defined on the cross-product.

If $f : X \to X$ is a linear operator, the problem is to devise a method of calculating all the eigenvalues and eigenspaces of f suitable for use with a high-speed digital computer. This involves finding a scalar $\lambda \in C$ and a maximal linearly independent set $S \subset X$ such that $f \underline{x}^i = \lambda \underline{x}^i$ for all $\underline{x}^i \in S$. Each number λ is called an eigenvalue, and each corresponding vector an eigenvector. The subspace generated by the elements of S is called the eigenspace corresponding to the eigenvalue.

The problem has the geometric interpretation of finding those vectors $\underline{x}$ whose directions remain invariant under f. The eigenvalue λ is then the factor by which the norm of such a vector is changed. If $\underline{x}$ is an eigenvector corresponding to λ, it follows from the defining equation $f \underline{x} = \lambda \underline{x}$ that $a \underline{x}$ for all $0 \neq a \in C$ is also an eigenvector corresponding to λ. Thus no generality is lost in assuming the eigenvectors have unit length.

Since a linear operator is completely determined by specifying

the images under the operator of an orthonormal set of basis vectors for X, one can view any operator as giving rise to a unique n × n matrix A. The problem then corresponds to solving the system of homogeneous linear equations $(A - \lambda I)\underline{x} = \underline{0}$ where I is the n × n unit matrix determined by the identity operator 1 on X into itself. Algebraically, this is equivalent to asking for those values of λ that allow the system of n equations in the n unknown coordinates of $\underline{x}$ to possess non-trivial solution vectors. Such a solution exists if and only if the determinant of the matrix of coefficients is zero. Expansion of this determinant yields an n'th degree polynomial in λ and is called the characteristic polynomial of A and denoted by $p_A(\lambda)$. The n complex zeros of $p_A(\lambda)$ are the eigenvalues of f, and the set of k distinct zeros is called the spectrum of f and denoted by $\lambda(f)$ or $\lambda(A)$. A particular eigenvalue in the spectrum of f is denoted by $\lambda_i(f)$ or $\lambda_i(A)$ and simply by λ_i when there is no confusion concerning the operator involved. When the elements of the spectrum are real, they are assumed ordered as follows: $\lambda_1 > \lambda_2 > \cdots > \lambda_k$. Since the dimension of the space is finite and defined over the complex field, the spectrum contains at least one element and thus a non-empty eigenspace exists.

The most successful numerical methods at present are applied to linear operators that give rise to real, symmetric matrices; and probably the best of these are those that may be termed "indirect methods" in which evaluation of the determinant by way of the characteristic polynomial is avoided. (Givens, 1954 and von Holdt, 1956). In all these methods no attempt is made to determine the eigenspaces of dimension greater than one; and in some, no eigenvectors are calcu-

lated at all. The procedure developed in this paper is a "direct method" which evaluates the determinant for each approximating eigenvalue although the analytic expression for the characteristic polynomial is not known. Although the application of the method is to linear operators, no restriction at all is placed on the type of matrices which they determine; and the eigenspaces are found for each element in the spectrum.

Probably one of the most basic reasons for wanting to solve such a problem stems from the need to solve a system of n linear differential equations which are encountered in a wide variety of fields. One method used in finding a solution of the system employs the eigenvalue technique. If primes denote differentiation with respect to a variable t, then the system can be written in matrix notation as $\underline{x}' = A\,\underline{x}$ where $\underline{x}'$ and $\underline{x}$ are n-column vectors and A is the $n \times n$ matrix of coefficients. One seeks a solution of the form $\underline{x} = \underline{v}\, e^{\lambda t}$ where $\underline{v}$ is an n-column vector. Differentiation of the trial solution vector followed by substitution in the matrix equation leads to an equivalent system $\underline{v}\,\lambda\, e^{\lambda t} = A\,\underline{v}\, e^{\lambda t}$ or $(A - \lambda I)\underline{v}\, e^{\lambda t} = \underline{0}$. Thus an eigenvalue λ, together with a corresponding eigenvector $\underline{v}$, yields a solution vector $\underline{x} = \underline{v}\, e^{\lambda t}$ of the system of differential equations. The same substitution will yield a solution for a second order system of linear differential equations.

CHAPTER II

NOTATION

The underlying linear space of n-columns over the field of complex numbers C will also be denoted by the symbol X. Points or vectors in X are represented by underscored letters such as $\underline{x}$ or $\underline{y}$. In this paper one is restricted to the subset of linear functions which forms a subspace of the space of all functions on X into itself. Use is also made of the fact that this subset forms a linear space with addition, scalar multiplication, and composition of functions defined as follows:

1. $(f + g)\underline{x} = f\,\underline{x} + g\,\underline{x}$
2. $(a\,f)\underline{x} = a\,f\,\underline{x}$
3. $(f\,g)\underline{x} = f(g\,\underline{x})$

where $f : X \to X$, $g : X \to X$, $a \in C$, and $f(\underline{x})$ or $f\,\underline{x}$ is the value of f evaluated at $\underline{x}$. Since f is linear, $f(a\,\underline{x} + b\,\underline{y}) = a\,f\,\underline{x} + b\,f\,\underline{y}$ for all $\underline{x}, \underline{y} \in X$ and for all $a, b \in C$. N(f) and R(f) denote respectively the null space and range space of f, and these subspaces are subsets of X defined as follows:

$$N(f) = \{\underline{x} \in X : f\,\underline{x} = \underline{0}\}$$
$$R(f) = \{\underline{y} \in X : \underline{y} = f\,\underline{x} \text{ for some } \underline{x} \in X\}\ .$$

The point ${}_k\underline{x}$ with a left subscript is a row vector in X and $\underline{x}^k$ with a right superscript a column vector. If $a \in C$, then a is considered as an ordered pair of real numbers (a_1, b_1) where $a = a_1 + i\, b_1$, and $\bar{a} \in C$ is the ordered pair $(a_1, -b_1)$ the complex conjugate of a.

The ordered set $E = (\underline{e}^i)_{i=1}^n$ of orthonormal column vectors is chosen as a fixed basis in X. The word "ordered" refers to vector position in E regarded as a row vector $(\underline{e}^1, \underline{e}^2, \cdots, \underline{e}^n)$ in which the elements are unit column vectors and mutually orthogonal. If $\underline{x} \in X$ is arbitrary, then $\underline{x} = e^\alpha\, {}_\alpha x$ is a linear combination of the elements of E where the repeated index indicates summation. Unless it is stated otherwise, the range of a repeated index α is such that $\alpha = 1, 2, \cdots, n$. Thus the notation $\underline{x} = e^\alpha\, {}_\alpha x$ means $\underline{x} = e^1\, {}_1x + \underline{e}^2\, {}_2x + \cdots + \underline{e}^n\, {}_nx$. The elements of the set $({}_1x)_{i=1}^n$ belong to C and are called the coordinates or components of $\underline{x}$ relative to the basis E. The summation convention is to be invoked only when the repeated index is a lower case Greek letter. The notation $\underline{e}^k\, {}_kx$ is then multiplication of the k'th entry in E by the scalar ${}_kx$. It should be noted that no restriction is placed on the position of the repeated index as a subscript or superscript.

A column vector $\underline{x}^k$, its transpose ${}^T\underline{x}^k$, and its conjugate transpose ${}^*\underline{x}^k$ are written as follows:

$$\underline{x}^k = \begin{bmatrix} {}_1x^k \\ {}_2x^k \\ \vdots \\ {}_nx^k \end{bmatrix}, \quad {}^T\underline{x}^k = [{}_1x^k,\ {}_2x^k,\ \cdots,\ {}_nx^k],$$

$${}^*\underline{x}^k = [{}_1\bar{x}^k,\ {}_2\bar{x}^k,\ \cdots,\ {}_n\bar{x}^k]$$

The n × n matrix A in which the entries in column j are the components in basis E of the image vectors $(f\ \underline{e}^j)_{j=1}^n$ is called "the matrix for the operator f : X → X with respect to the basis E". If $(\lambda_i)_{i=1}^n$ is the set of zeros of $p_A(\lambda)$, it is known that the number $a = \sum_{i=1}^{n} \lambda_i$ is the sum of the principle diagonal elements of A. This number is called the trace of A and denoted by trace (A). Since E is an ordered orthonormal set, A is uniquely determined and will be interpreted as a linear operator on the n-dimensional space X into itself under the assignment $\underline{x} \to A\ \underline{x}$ where $A\ \underline{x}$ is computed using ordinary matrix multiplication. Thus all statements made concerning properties of eigenvalues and eigenvectors of linear operators on an n-dimensional space can be translated into statements concerning n × n matrices and vice versa. An attempt is made in the formulation and proof of statements to use the language that avoids cumbersome notation and offers greater clarity.

In the notation $A = ({}_i a^j)$, where i and j are respectively the row and column indices, the matrix A is considered as a set of n columns or n rows in which ${}_i a^j = {}_i(A)^j$ denotes the element in the i'th row and j'th column of A. The vector $\underline{a}^j$ or $(A)^j$ is the j'th column and ${}_i\underline{a}$ or ${}_i(A)$ the i'th row vector. Thus for the n × n identity matrix $I = ({}_i\delta^j)$ the symbol $\underline{\delta}^j$ is the column with entries 1 if $i = j$ and 0 if $i \neq j$. Similarly ${}_i\underline{\delta}$ is the row with entries 1 when $j = i$ and 0 if $j \neq i$. For $n \geq 2$ the notation $|A|$ is the determinant of A and the modulus of the single entry if $n = 1$. If $|A| \neq 0$, then A is said to be non-singular; and in this case, the inverse matrix denoted by A^{-1} exists such that $AA^{-1} = A^{-1}A = I$. If A is the n × n matrix determined by a linear operator f, then A^{-1} exists if and only

if $N(f) = (\underline{0})$.

If S and T are subsets of X, then the set difference is $S \setminus T = (\underline{s} \in S \ni \underline{s} \notin T)$. The set of all linear combinations of elements of S is called the "linear manifold of S" or the subspace generated by S and denoted by M(S). If $S \subset T$ and $M(S) = T$, then T is the subspace generated or spanned by S and the dimension of T is equal to the number of vectors in a maximal linearly independent set contained in S. This dimension is abbreviated as dim T or dim M(S). If $M(S) = (\underline{0})$, then M(S) is called the trivial subspace.

In this paper one is concerned with square matrices only. If A is an $n \times n$ square matrix, then the symbol r(A) denotes the rank of A. The rank of A is defined to be the maximum number of linearly independent column vectors contained in A. (Amir - Moez and Fass, 1962). Since rank is an invariant function of matrices under elementary row transformations, this number is determined by examining the matrix in its pivotal form which is discussed in Chapter V.

Definition 2.1. A non-empty set $S \subset X$ such that $(\underline{0}) \notin S$ is linearly independent if no element of S is a linear combination of other elements of S. Otherwise S is said to be a linearly dependent set.

CHAPTER III

BOUNDS FOR THE SPECTRUM

For the complex inner-product space one defines a certain function of two vectors which maps the cross-product of the linear space X into the field of complex numbers C. This function is called the inner product; and if the vectors are $\underline{x},\underline{y}$, the inner product of $\underline{x}$ with $\underline{y}$ is denoted by $(\underline{x},\underline{y})$. The inner-product function is formally analogous to the dot product of ordinary vector analysis, and its properties are contained in the following definition.

Definition 3.1. A complex linear space X is called an inner-product space if there is defined on $X \times X$ a complex-valued function $(\underline{x},\underline{y})$ (called the inner product of $\underline{x}$ with $\underline{y}$) with the following properties:

1. $(\underline{x},\underline{y} + \underline{z}) = (\underline{x},\underline{y}) + (\underline{x},\underline{z})$.
2. $(\underline{x}, a\,\underline{y}) = a(\underline{x},\underline{y})$, for all $a \in C$.
3. $(\underline{x},\underline{y}) = \overline{(\underline{y},\underline{x})}$.
4. $(\underline{x},\underline{x}) \geq 0$ and $(\underline{x},\underline{x}) \neq 0$ if $\underline{x} \neq \underline{0}$.

It is seen from this definition that condition (3) implies $(\underline{x},\underline{x})$ is real; and as a consequence of the properties listed, $(\underline{x},\underline{y})$ has the further properties:

5. $(\underline{x} + \underline{y},\underline{z}) = (\underline{x},\underline{z}) + (\underline{y},\underline{z})$.

6. $(a\,\underline{x},\underline{y}) = \bar{a}(\underline{x},\underline{y})$, for all $a \in C$.

The formula given in the following definition for computing the inner product satisfies the properties stated in Definition 3.1. It is this inner product that is used throughout this paper.

Definition 3.2. If $\underline{x},\underline{y} \in X$ have components $({}_i x)_{i=1}^n$ and $({}_i y)_{i=1}^n$ respectively relative to E, the number $(\underline{x},\underline{y}) = {}_\alpha\bar{x}\,{}_\alpha y$ is the inner product of $\underline{x}$ with $\underline{y}$.

It can be shown that if X is an inner-product space, then $(\underline{x},\underline{x})^{1/2}$ has the properties of a norm. (Taylor, 1961). One writes $\|\underline{x}\| = (\underline{x},\underline{x})^{1/2}$ and calls this number the norm or length of the vector $\underline{x}$. If $\|\underline{x}\| = 1$, then $\underline{x}$ is a unit vector; and two non-zero vectors $\underline{x}$ and $\underline{y}$ with $(\underline{x},\underline{y}) = 0$ are said to be orthogonal and written $\underline{x} \perp \underline{y}$. For the basis E it follows that $(\underline{e}^i,\underline{e}^j) = {}_i\delta^j$ the Kronecker delta.

Definition 3.3. The transposed conjugate of a matrix A is called the adjoint matrix of A and denoted by ${}^*A$. If $A = {}^*A$, then A is called self-adjoint.

Definition 3.4. If f and g are linear operators on $X \to X$ and if $(f\,\underline{x},\underline{y}) = (\underline{x},g\,\underline{y})$ for all $\underline{x},\underline{y} \in X$, then g is called the adjoint operator of f and denoted by ${}^*f$. If $f = {}^*f$, then f is called self-adjoint. For the composition of two functions the usual rule holds that ${}^*(f\,g) = {}^*g\,{}^*f$.

Theorem 3.1. If ${}^*f$ is the adjoint operator of $f : X \to X$ and A is the matrix for f relative to E, then ${}^*A$ is the matrix for ${}^*f$ relative to E.

Proof. By assumption $(f\,\underline{x},\underline{y}) = (\underline{x},{}^*f\,\underline{y})$ for all $\underline{x},\underline{y} \in X$ and for $A = ({}_i a^j)$, $f\,\underline{e}^j = \underline{e}^\beta\,{}_\beta a^j$. One lets $B = ({}_i b^j)$ be the matrix for ${}^*f$ and shows that ${}_i\bar{a}^j = (f\,\underline{e}^j,\underline{e}^i) = (\underline{e}^j,{}^*f\,\underline{e}^i) = {}_j b^i$. Using the properties of the inner product and the fact that E is orthonormal it follows that

$$(f\,\underline{e}^j,\underline{e}^i) = (\underline{e}^\beta\,{}_\beta a^j,\underline{e}^i) =$$

$$\overline{(\underline{e}^i,\underline{e}^\beta\,{}_\beta a^j)} = \overline{{}_\beta a^j(\underline{e}^i,\underline{e}^\beta)} = \overline{{}_\beta a^j\,{}_i\delta^\beta} = {}_i\bar{a}^j$$

and similarly

$$(\underline{e}^j,{}^*f\,\underline{e}^i) = (\underline{e}^j,\underline{e}^\alpha\,{}_\alpha b^i) = {}_\alpha b^i(\underline{e}^j,\underline{e}^\alpha) = {}_\alpha b^i\,{}_j\delta^\alpha = {}_j b^i$$

But $(f\,\underline{e}^j,\underline{e}^i) = (\underline{e}^j,{}^*f\,\underline{e}^i)$ by Definition 3.4, and hence ${}_i\bar{a}^j = {}_j b^i$ and thus $B = {}^*A$.

The preceding theorem shows that any linear operator f on the finite dimensional space X has a unique adjoint ${}^*f$, and ${}^*f$ is completely defined by ${}^*A$ where A is the matrix for f relative to E.

Lemma 3.1. An operator $f : X \to X$ is self-adjoint if and only if $(\underline{x},f\,\underline{x})$ is real for all $\underline{x} \in X$.

Proof. If $f = {}^*f$, then $\overline{(\underline{x},f\,\underline{x})} = (f\,\underline{x},\underline{x}) = ({}^*f\,\underline{x},\underline{x}) = (\underline{x},f\,\underline{x})$ and hence $(x,f\,x)$ is real. Conversely, if $(\underline{x},f\,\underline{x})$ is real for all $\underline{x} \in X$, then $(\underline{x},f\,\underline{x}) = (f\,\underline{x},\underline{x}) = (\underline{x},{}^*f\,\underline{x})$ which implies that $f = {}^*f$ and consequently f is self-adjoint.

It follows immediately from the preceding lemma that the operators ${}^*f + f$ and $i({}^*f - f)$ are self-adjoint; and if g is self-adjoint, then $a\,g$ is self-adjoint if and only if $a \in C$ is real. The number i employed above is the ordered pair $(0,1)$. Thus any operator $f : X \to X$ has a unique Cartesian decomposition $f = g + i\,h$ with $g = \frac{1}{2}({}^*f + f)$ and $h = \frac{i}{2}({}^*f - f)$ where g and h are self-adjoint.

Lemma 3.2. If $f : X \to X$ is self-adjoint, then the eigenvalues of f are real; and the eigenvectors corresponding to distinct eigenvalues are orthogonal.

Proof. If λ is an eigenvalue with corresponding eigenvector $\underline{x}$, then $f\,\underline{x} = {}^*f\,\underline{x} = \lambda\,\underline{x}$. Using the properties of the inner product it follows that $\lambda(\underline{x},\underline{x}) = (\underline{x},\lambda\,\underline{x}) = (\underline{x},f\,\underline{x}) = ({}^*f\,\underline{x},\underline{x}) = (\lambda\,\underline{x},\underline{x}) = \overline{\lambda}(\underline{x},\underline{x})$ and hence $\lambda = \overline{\lambda}$ since $(\underline{x},\underline{x})$ is real and non-zero for $\underline{x} \neq \underline{0}$. Thus the eigenvalues are real. Suppose now that $\mu \neq \lambda$ is an eigenvalue with eigenvector $\underline{y}$. Then $f\,\underline{y} = \mu\,\underline{y}$ and it follows that $\mu(\underline{x},\underline{y}) = (\underline{x},\mu\,\underline{y}) = (\underline{x},f\,\underline{y}) = ({}^*f\,\underline{x},\underline{y}) = (f\,\underline{x},\underline{y}) = (\lambda\,\underline{x},\underline{y}) = \lambda(\underline{x},\underline{y})$. Therefore $(\mu - \lambda)(\underline{x},\underline{y}) = 0$; and since $\mu \neq \lambda$, we have $(\underline{x},\underline{y}) = 0$ and consequently $\underline{x} \perp \underline{y}$.

If $f : X \to X$ is self-adjoint with spectrum $\lambda(f)$, it is known

by Fischer's Minimax Principle that $\lambda_i(f) = \sup_{\|\underline{x}\| = 1} (\underline{x}, f\,\underline{x})$ and $\lambda_k(f) = \inf_{\|\underline{x}\| = 1} (\underline{x}, f\,\underline{x})$. (Amir - Moez and Fass, 1962). In what follows $\mu(g)$ and $\nu(h)$ are respectively the spectrums of the self-adjoint operators g and h in the Cartesian decomposition of a given operator. By Lemma 3.2 the spectrums contain only real numbers.

Theorem 3.2. If $f : X \to X$ has eigenvalues $\lambda(f) = \{(a_i, b_i)\}_{i=1}^{k}$ and Cartesian decomposition $g + i\,h$, then

1. $\mu_k(g) \le a_i \le \mu_1(g)$, $i = 1, 2, \cdots, k$ and

2. $\gamma_k(h) \le b_i \le \gamma_1(h)$, $i = 1, 2, \cdots, k$.

Proof. Let $\underline{x}$ with $\|\underline{x}\| = 1$ be an eigenvector of f corresponding to the eigenvalue $\lambda_i(f)$. Then by Fischer's Minimax Principle $\mu_k(g) \le (\underline{x}, g\,\underline{x}) \le \mu_1(g)$ and $\gamma_k(h) \le (\underline{x}, h\,\underline{x}) \le \gamma_1(h)$. Using the definitions of g and h and the properties of the inner product function one has

$$(\underline{x}, g\,\underline{x}) = \tfrac{1}{2}(\underline{x}, ({}^{*}f + f)\underline{x}) = \tfrac{1}{2}[(\underline{x}, {}^{*}f\,\underline{x}) + (\underline{x}, f\,\underline{x})] =$$

$$\tfrac{1}{2}[\overline{(\underline{x}, f\,\underline{x})} + (\underline{x}, f\,\underline{x})] = \tfrac{1}{2}[\overline{\lambda_i(f)} + \lambda_i(f)] = a_i$$

which proves the first assertion. Similarly

$$(\underline{x}, h\,\underline{x}) = \tfrac{i}{2}(\underline{x}, ({}^{*}f - f)\underline{x}) = \tfrac{i}{2}[(\underline{x}, {}^{*}f\,\underline{x}) - (\underline{x}, f\,\underline{x})] =$$

$$\frac{i}{2}[\overline{(\underline{x}, f\,\underline{x})} - (\underline{x}, f\,\underline{x})] = \frac{i}{2}[\overline{\lambda_i(f)} - \lambda_i(f)] = b_i$$

and thus the second assertion is established.

Thus for any operator f with Cartesian decomposition g + i h, the eigenvalues of f must lie in the rectangle in the complex plane determined by the Cartesian product of the closed intervals $[\mu_k(g), \mu_1(g)]$ and $[\gamma_k(h), \gamma_1(h)]$. Since the operators g and h are self adjoint, each commutes with its adjoint and hence belongs to the class of normal operators. The normal operators on an n-dimensional space are known to be diagonal, and an operator is diagonal if and only if it has n linearly independent eigenvectors. It follows that the eigenvalues of g and h are real with each having n linearly independent eigenvectors, and the basic power method is applicable to compute their maximum and minimum eigenvalues. (Bodewig, 1959). Numerical techniques using the power method, or variations of the power method in some cases, have been developed to approximate these extremal values with a high degree of accuracy. However, for the iterative scheme described in Chapter VIII the following rough approximations to the vertices of R are sufficient and are more readily obtained.

Theorem 3.3. Let the set $(\lambda_i)_{i=1}^n$ be the zeros of the characteristic polynomial determined by the self-adjoint operator $f : X \to X$. If $(\underline{s}^j)_{j=1}^n$ is any set of orthonormal column vectors in X, then $(f\,\underline{s}^\alpha, \underline{s}^\alpha) = \sum_{i=1}^{n} \lambda_i$.

Proof. Since f is self-adjoint, there exists a set $(\underline{x}^i)_{i=1}^n$ of orthonormal vectors such that $f\,\underline{x}^i = \lambda_i\,\underline{x}^i$, i = 1, 2, ···, n. Since this set is a basis for X, $\underline{s}^j = \underline{x}^\alpha\,{}_\alpha x^j$, j = 1, 2, ···, n. Now taking the inner product of both sides with $\underline{x}^\alpha$ gives $(\underline{s}^j,\underline{x}^\alpha) = {}_\alpha x^j(\underline{x}^\alpha,\underline{x}^\alpha) = {}_\alpha x^j$ and hence $\underline{s}^j = (\underline{s}^j,\underline{x}^\alpha)\underline{x}^\alpha$. Then $f\,\underline{s}^j = (\underline{s}^j,\underline{x}^\alpha) f\,\underline{x}^\alpha = \lambda_\alpha(\underline{s}^j,\underline{x}^\alpha)\underline{x}^\alpha$ and $(f\,\underline{s}^j,\underline{s}^j) = \lambda_\alpha(\underline{s}^j,\underline{x}^\alpha)(\underline{x}^\alpha,\underline{s}^j) = \lambda_\alpha(\underline{s}^j,\underline{x}^\alpha)\overline{(\underline{s}^j,\underline{x}^\alpha)} = \lambda_\alpha|(\underline{s}^j,\underline{x}^\alpha)|^2$. Now summing on j gives $(f\,\underline{s}^\beta,\underline{s}^\beta) = \lambda_\alpha \sum_{j=1}^n |(\underline{s}^j,\underline{x}^\alpha)|^2 = \lambda_\alpha\|\underline{x}^\alpha\|^2$. But $\|\underline{x}^\alpha\|^2 = 1$ and hence $\lambda_\alpha\|\underline{x}^\alpha\|^2 = \sum_{i=1}^n \lambda_i$.

Corollary 3.3.1. If $A = ({}_i a^j)$ is the matrix for f relative to E, then $(f\,\underline{e}^\alpha,\underline{e}^\alpha) = \text{trace}\,(A) = {}_\alpha a^\alpha$.

Proof. Follows directly from the definition of the trace (A) and the fact that this number is equal to the sum of the eigenvalues of A.

Corollary 3.3.2. The operator $f^2 : X \to X$ has the set $(\lambda_i^2)_{i=1}^n$ as its eigenvalues and $(f^2\,\underline{s}^\alpha,\underline{s}^\alpha) = \sum_{i=1}^n \lambda_i^2$.

Proof. $f^2 = {}^*ff$ since f is self-adjoint, and ${}^*({}^*ff) = {}^*ff$ implies that f^2 is self-adjoint. Since $f\,\underline{x}^i = \lambda_i\,\underline{x}^i$, then $f^2\,\underline{x}^i = \lambda_i(f\,\underline{x}^i) = \lambda_i^2\,\underline{x}^i$ implies that λ_i^2 is an eigenvalue of f^2.

Theorem 3.4. Let g + i h be the Cartesian decomposition of

$f : X \to X$. Let $(\mu_i)_{i=1}^n$ and $(\gamma_i)_{i=1}^n$ be respectively the eigenvalues of the $n \times n$ matrices $G = ({}_ig^j)$ and $H = ({}_ih^j)$ where G and H are determined respectively by the operators g and h relative to E. Then

1. $\mu_1 \leq [{}_\alpha\overline{g}^\beta \; {}_\alpha g^\beta]^{1/2}$

2. $\mu_n \geq {}_\alpha g^\alpha - n[{}_\alpha\overline{g}^\beta \; {}_\alpha g^\beta]^{1/2}$

3. $\gamma_1 \leq [{}_\alpha\overline{h}^\beta \; {}_\alpha h^\beta]^{1/2}$

4. $\gamma_n \geq {}_\alpha h^\alpha - n[{}_\alpha\overline{h}^\beta \; {}_\alpha h^\beta]^{1/2}$

Proof. Since g is self-adjoint it follows from Corollary 3.3.2 that

$$\mu_1^2 \leq \sum_{i=1}^n \mu_i^2 = (g^2 \, \underline{e}^\alpha, \underline{e}^\alpha) = (g \, \underline{e}^\alpha, g \, \underline{e}^\alpha) = (\underline{g}^\alpha, \underline{g}^\alpha) = {}_\alpha\overline{g}^\beta \; {}_\alpha g^\beta$$

which proves the first assertion. The second assertion follows from Theorem 3.3 and Corollary 3.3.1 since

$$\mu_n = {}_\alpha g^\alpha - \sum_{i=1}^{n-1} \mu_i \geq {}_\alpha g^\alpha - n|\mu_1| \geq {}_\alpha g^\alpha - n[{}_\alpha\overline{g}^\beta \; {}_\alpha g^\beta]^{1/2}$$

The last two assertions are proved in exactly the same way using the operator h instead of g.

The bounds found here are probably not the best that one could

obtain. At the n'th iteration, calculations are carried out in a rectangle whose area is approximately the area of the original rectangle R multiplied by a factor of $1/2^{2n}$. For this reason very crude approximations are adequate. Actually the iterative process described in Chapter VIII requires that all the eigenvalues of a given operator must lie in the interior of a square. This can be any square that contains the rectangle R, but one usually chooses the square such that one side coincides with one of the two longest sides of R.

CHAPTER IV

THE ZEROS OF THE CHARACTERISTIC POLYNOMIAL

It has already been pointed out that the most highly developed spectral theory in numerical analysis is that for self-adjoint operators. These are the operators that give rise to self-adjoint matrices relative to the basis E. The dominant fact about such operators is that there exists an orthonormal basis for X such that the matrix determined by the operator relative to this basis is diagonal. (Householder, 1953). The elements on the diagonal are the eigenvalues of the operator. This is the basis for the fact that self-adjoint operators determine their eigenvalues very well, i.e., small changes in the elements of the matrix produce small changes in the eigenvalues. This form of stability does not necessarily hold for operators that give rise to general matrices. Consider for example, the 10 x 10 matrix A of the following form:

$$(A) = \begin{bmatrix} \circ & 1 & & & \\ & \circ & \ddots & & \\ & & \ddots & 1 & \\ & & & \ddots & \\ & & & & \circ \end{bmatrix}$$

All elements in A are zero except for the first superdiagonal, all of whose elements are unity, and the element $_1(A)^{10}$ whose value is ϵ. It is easily seen that $p_A(\lambda) = \lambda^{10} - \epsilon$; and if ϵ is zero, the only eigenvalue of A is zero with algebraic multiplicity ten. If $\epsilon = 10^{-10}$ and ω is the first principal n'th root of unity, then $\{\omega^i \cdot 10^{-1}\}_{i=0}^{9}$ is the set of eigenvalues of A with each having modulus 10^{-1}. Thus a change of 10^{-10} in one element of A produced a change in moduli of the eigenvalues 10^9 as great. This is indicative of what may happen in general, and such a matrix does not well determine its eigenvalues. Probably no numerical method can give good answers to a problem similar to this if the elements of the matrix are measured quantities, and the accuracy of the results is expected to be as good as the data. It is quite possible in the ideal situation that distinct eigenvalues are sufficiently separated in modulus. It is not proclaimed that the method given in this paper will give good results in all problems. It is believed, however, that the method will handle practical problems even those that give rise to repeated eigenvalues.

If $p_A(\lambda)$ is the characteristic polynomial of f with spectrum $\lambda(f)$, then each element in $\lambda(f)$ is a zero of $p_A(\lambda)$ of some multiplicity a_i, $i = 1, 2, \cdots, k$. The number a_i is called the "algebraic multiplicity" of the eigenvalue $\lambda_i(f)$ with $\sum_{i=1}^{k} a_i = n$ and the factorization of $p_A(\lambda)$ given by $p_A(\lambda) = \prod_{i=1}^{k}[\lambda - \lambda_i(f)]^{a_i}$. Each eigenvalue $\lambda_i(f)$ has at least one associated eigenvector. The maximal number of linearly independent eigenvectors associated with the eigenvalue $\lambda_i(f)$ is called the "geometric multiplicity" of $\lambda_i(f)$ and denoted by g_i. The linear manifold of this maximal set is the eigenspace of f associ-

ated with $\lambda_i(f)$. The method developed here for numerical computation of the eigenvalues is restricted perhaps since one of the basic requirements laid down is that the eigenspace be found for each eigenvalue. That this requirement is satisfied is demonstrated in Corollary 5.3.2.

The name "deflation" is given to any technique used after finding one eigenvalue and corresponding eigenspace to reduce the original problem to one in which the known value and vector or vectors are no longer present. It is believed that the present method has greatly reduced the complexity of this problem, and the reduction is found in part in the proof of the following.

Lemma 4.1. If $f : X \to X$ has an eigenvalue $\lambda_i(f)$ whose algebraic and geometric multiplicities are respectively a_i and g_i, then $g_i \leq a_i$.

Proof. Let $\{\underline{x}^k\}_{k=1}^{g_i}$ be the linearly independent set generating the eigenspace associated with $\lambda_i(f)$. Complete this set to a basis B for X. One now constructs the matrix A for f relative to B. Since $f\,\underline{x}^k = \lambda_i(f)\underline{x}^k$ by assumption, the first g_i columns of A will be $\lambda_i(f)\underline{e}^k$, $k = 1, 2, \cdots, g_i$. It follows that the first g_i columns of the matrix $(A - \lambda I)$ will then be $(\lambda_i(f) - \lambda)\underline{e}^k$, $k = 1, 2, \cdots, g_i$ and hence $(\lambda_i(f) - \lambda)^{g_i}$ is a factor of $p_A(\lambda)$ and thus $g_i \leq a_i$.

It will be seen that the method employed is direct in that for a given λ a polynomial $P_n(\lambda)$ is evaluated which is proportional to

$p_A(\lambda)$, and deflation is obtained by considering the expression $P_n(\lambda) \Big/ \prod_{i=1}^{k} (\lambda_i(f) - \lambda)^{g_i}$ where the first k distinct eigenvalues are known.

CHAPTER V

THE PIVOTAL MATRIX AND ITS PROPERTIES

Let A be an $n \times n$ matrix relative to E for the operator $f : X \to X$. The nonsingular $n \times n$ elementary matrices G_{ij}, $G_i(c)$ and $G_{ij}(c)$ are called Types I, II and III respectively. Employing the usual definition of matrix multiplication, these matrices are defined as follows:

> Type I. The matrix G_{ij} A is the matrix A with rows i and j interchanged.
>
> Type II. The matrix $G_i(c)$A is the matrix A with row i multiplied by a non-zero constant $c \in C$.
>
> Type III. The matrix $G_{ij}(c)$A is the matrix A with row j replaced by $c\ \underline{a}_i + \underline{a}_j$, i.e., the row j is replaced by the sum of c times row i and row j.

These three types of matrices as defined here are elementary row transformations which are applied on the left of a given operator in matrix representation. They are used here in reducing $n \times n$ matrices to pivotal form which is defined in Definition 5.2. It can

be shown that the rank function of matrices is an invariant under row equivalence. (Stoll, 1952).

Definition 5.1. Two matrices are row equivalent if one can be obtained from the other by a finite succession of elementary row operations.

Thus matrices A and B are row equivalent if B = G A where G is a nonsingular finite product of elementary matrices.

Definition 5.2. A square matrix $P = ({}_ip^j)$ is called a pivotal matrix if the following conditions hold:

1. ${}_ip^j = 0$ for $i > j$

2. ${}_ip^i \neq 0 \Rightarrow \underline{p}^i = \underline{\delta}^i$

3. ${}_ip^i = 0 \Rightarrow {}_i\underline{p} = \underline{0}$.

Thus it is seen that a pivotal matrix is by definition upper triangular and that the elements on the main diagonal are either zero or one. The name given to the matrix comes from the fact that in the reduction of a matrix to pivotal form one chooses the element of maximum absolute value in a given column and shifts this element to a position on the main diagonal. In the construction process described in the next theorem, it becomes apparent that this is done to give greater accuracy and less round-off error in the numerical

computations that are necessary for such a reduction.

Theorem 5.1. Every square matrix $A = ({}_i a^j)$ is row equivalent to a pivotal matrix.

Proof. The proof is constructive and in the construction the columns of A are taken care of one by one. Let $\underline{a}^{j_1}$ be the non-zero column of A with smallest index j_1 and ${}_k a^{j_1}$ the largest in absolute value of its non-zero entries. Multiply A in succession by $G_k\left(\frac{1}{{}_k a^{j_1}}\right)$, $G_{kj}\left(-{}_j a^{j_1}\right)$, $j = 1, 2, \cdots, k-1, k+1, \cdots, n$ and G_{kj_1}. The result is a row equivalent matrix $P_1 = G_1 A$ with ${}_i(P_1)^j = 0$ for $i \geq j$ and $j < j_1$, $(P_1)^{j_1} = \underline{\delta}^{j_1}$, and G_1 a product of at most $n+1$ elementary matrices. The transformations $G_{kj}\left(-{}_j a^{j_1}\right)$ are applied to reduce the elements ${}_j a^{j_1}$, $j = 1, 2, \cdots, k-1, k+1, \cdots, n$ to zero; and of course, are not applied if the element is already zero. Next locate in P_1, with row j_1 deleted, the non-zero column with smallest index $j_2 > j_1$, and let ${}_k a^{j_2}$ be the largest in absolute value of the non-zero entries in column j_2 excluding the entry appearing in row j_1. Multiply P_1 by $G_k\left(\frac{1}{{}_k a^{j_2}}\right)$, $G_{kj}\left(-{}_j a^{j_2}\right)$, $j = 1, 2, \cdots, k-1, k+1, \cdots, n$ and G_{kj_2} in succession. These multiplications do not affect the columns of P_1 with column index $j < j_2$. This follows since all columns of P_1 with column index less than j_1 are zero columns, $(P_1)^{j_1} = \underline{\delta}^{j_1}$, and

any column with index i such that $j_1 < i < j_2$ must have all zero entries except possibly the entry appearing in ${}_{j_1}(P_1)$. The result is a row equivalent matrix $P_2 = G_2\, G_1\, A$ with ${}_i(P_2)^j = 0$ for $i > j$ and $j \leq j_2$, $(P_2)^{j_2} = \underline{\delta}^{j_2}$ and $G_2\, G_1$ a product of at most $2(n + 1)$ elementary matrices. The process terminates in $m \leq n$ steps and $P = G_m \cdot G_{m-1} \cdots G_1 \cdot A$ involves the multiplication of A by a finite number of elementary matrices. By construction it is seen that P satisfies conditions (1) and (2) in the definition of a pivotal matrix. To show that condition (3) is also satisfied, let ${}_jp^j = 0$ and assume there exists some $k > j$ such that ${}_jp^k \neq 0$. Then ${}_kp^k = 0$ by condition (2), but this is impossible since by condition (1) and by assumption ${}_ip^j = 0$ for $i \geq j$ and the transformation G_{kj} of Type I could be applied.

A simple example is given below (using a 4×4 real matrix A) to help clarify the numerical procedure used in obtaining the pivotal form. The last matrix given in the sequence of reductions is the pivotal form P of A. The transformations used in the reduction (applied from left to right) are as follows:

1. $G_4(1/2)$, $G_{41}(-1)$, $G_{42}(1)$, $G_{43}(-1)$, G_{41}

2. $G_3(2/3)$, $G_{31}(-5/2)$, $G_{32}(-1/2)$, $G_{34}(1/2)$, G_{32}

3. $G_4(3)$, $G_{41}(-1/3)$, $G_{42}(1/3)$, $G_{43}(-2/3)$

The numbers in parenthesis above the arrows below refer to the sequence of transformations given above. These were applied to the matrices on the left to obtain those on the right of the arrow.

$$A = \begin{bmatrix} 1 & 2 & 0 & 0 \\ -1 & -2 & 0 & 1 \\ 1 & 4 & -2 & -1 \\ 2 & 5 & -1 & -1 \end{bmatrix}$$

$$A \xrightarrow{(1)} \begin{bmatrix} 1 & 5/2 & -1/2 & -1/2 \\ 0 & 1/2 & -1/2 & 1/2 \\ 0 & 3/2 & -3/2 & -1/2 \\ 0 & -1/2 & 1/2 & 1/2 \end{bmatrix} = P_1$$

$$P_1 \xrightarrow{(2)} \begin{bmatrix} 1 & 0 & 2 & 1/3 \\ 0 & 1 & -1 & -1/3 \\ 0 & 0 & 0 & 2/3 \\ 0 & 0 & 0 & 1/3 \end{bmatrix} = P_2$$

$$P_2 \xrightarrow{(3)} \begin{bmatrix} 1 & 0 & 2 & 0 \\ 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} = P$$

Before discussing some of the properties of a pivotal matrix, it may be advantageous to examine the matrix P just obtained. First, the r(A) is three since there are three linearly independent columns in P. This implies that the determinant of A is zero, and there exists a one dimensional subspace belonging to N(A). The vector ${}^{T}\underline{x} = (-2, 1, 1, 0)$ spans this subspace. This vector is obtained from P by computing $\underline{x} = \underline{\delta}^3 - \underline{p}^3$. Theorem 5.3 shows that a more general result is true.

Theorem 5.2. If $P = ({}_{i}p^{j}) = (\underline{p}^1, \underline{p}^2, \cdots, \underline{p}^n)$ is a pivotal matrix, then $P^2 = P$.

Proof. The element in the i'th row and j'th column of P^2 is given by ${}_{i}(P^2)^{j} = {}_{i}p^{\alpha}\,{}_{\alpha}p^{j}$. If ${}_{i}p^{i} = 0$, then by definition ${}_{i}\underline{p}$ is a zero row, and hence ${}_{i}(P^2)^{j} = {}_{i}p^{\alpha}\,{}_{\alpha}p^{j} = {}_{i}(P)^{j} = 0$. If ${}_{i}p^{i} = 1$ and ${}_{i}p^{k} = 0$ for $k \neq i$, then ${}_{i}(P^2)^{j} = {}_{i}p^{\alpha}\,{}_{\alpha}p^{j} = {}_{i}p^{j} = {}_{i}(P)^{j}$. If ${}_{i}p^{i} = 1$ and ${}_{i}p^{k} \neq 0$ for $k \neq i$, then ${}_{k}p^{k} = 0$ which implies ${}_{k}\underline{p} = 0$ and hence ${}_{k}p^{j} = 0$ for $j = 1, 2, \cdots, n$. Thus ${}_{i}(P^2)^{j} = {}_{i}p^{\alpha}\,{}_{\alpha}p^{j} = {}_{i}p^{j} = {}_{i}(P)^{j}$.

Corollary 5.2.1. If $\underline{x}^i$ is any column of a pivotal matrix P, then $P\,\underline{x}^i = \underline{x}^i$.

Proof. The proof is immediate since $P^2 = P$, and $\underline{x}^i$ is a column vector of P.

Theorem 5.3. Let $P = ({}_{i}p^{j}) = (\underline{p}^1, \underline{p}^2, \cdots, \underline{p}^n)$ be a pivotal

matrix obtained from an n × n matrix B such that P = G B. Let $\{S_j\}$ be the non-zero vectors contained in the set $(\underline{\delta}^j - \underline{p}^j)$, j = 1, 2, •••, n. Then $\{S_j\}$ is a set of $m \leq n$ linearly independent vectors spanning N(B).

Proof. Let $\underline{x}^i \in \{S_j\}$. Then since $P^2 = P$,

$$B\,\underline{x}^i = B(\underline{\delta}^i - \underline{p}^i) = B\,\underline{\delta}^i - B\,\underline{p}^i =$$

$$G^{-1} P\,\underline{\delta}^i - G^{-1} P\,\underline{p}^i = G^{-1}\,\underline{p}^i - G^{-1}\,\underline{p}^i = \underline{0}$$

and hence $\underline{x}^i \in N(B)$. If there exists scalers $({}_1a)_{i=1}^m \in C$ not all of which are zero such that $\underline{x}^\alpha\,{}_\alpha a = \underline{0}$, then for some ${}_k a \neq 0$, $\underline{x}^k = -\frac{1}{{}_k a} \sum_{j\neq k}^m \underline{x}^j\,{}_j a$. Now $\underline{x}^k \in \{S_j\}$ implies $\underline{\delta}^k \neq \underline{p}^k$ and hence ${}_k p^k = 0$. But this implies that ${}_k\underline{p} = \underline{0}$, and thus it follows for the k'th component of $\underline{x}^k$ that

$${}_k x^k = -\frac{1}{{}_k a}\left(\sum_{j\neq k}^m {}_k x^j\,{}_j a\right) =$$

$$-\frac{1}{{}_k a}\left(\sum_{j\neq k}^m \left({}_k\delta^i - {}_k p^j\right){}_j a\right) = \frac{1}{{}_k a}\left(\sum_{j\neq k}^m {}_k p^j\,{}_j a\right) = 0$$

since ${}_k p^j = 0$, j = 1, 2, •••, k - 1, k + 1, •••, m. But by definition of $\underline{x}^k$ it follows that ${}_k x^k = {}_k\delta^k - {}_k p^k = 1$ since ${}_k p^k = 0$. This is a contradiction and hence $\{S_j\}$ is a linearly independent set. Since P = G B and G is non-

singular, it follows that dim R(P) = dim R(B). Since $\underline{\delta}^{j_i}$, $i = 1, 2, \cdots, (n - m)$ are linearly independent columns of P and $P\,\underline{\delta}^{j_i} = \underline{\delta}^{j_i}$, it follows that dim R(P) = dim R(B) $\geq$ $n - m$, so that dim N(B) = n - dim R(B) $\leq n - (n - m) = m$. Since $\{S_j\}$ contains m linearly independent vectors that belong to N(B), it follows that dim N(B) = m and hence $\{S_j\}$ spans N(B).

Corollary 5.3.1. If $B = (A - \lambda_j(f)\, I)$ where A is the $n \times n$ matrix for an operator f with eigenvalue $\lambda_j(f)$, then $g_j = m = n - r(P)$ where g_j is the geometric multiplicity of the eigenvalue $\lambda_j(f)$ and P the pivotal form of B.

Proof. The proof is immediate when one notes that each $\underline{x}^k \in \{S_j\}$ is an eigenvector corresponding to the eigenvalue $\lambda_j(f)$, and $\{S_j\}$ is a linearly independent set of m vectors spanning N(B).

Corollary 5.3.2. If $B = (A - \lambda_j(f)\, I)$, then $M(\{S_j\})$ is the eigenspace of A associated with the eigenvalue $\lambda_j(f)$.

Proof. The set $\{S_j\}$ is linearly independent and if $\underline{x}^k \in \{S_j\}$, then $\underline{x}^k \in N(B)$. Since $\{S_j\}$ is a maximal linearly independent set, it follows that $M(\{S_j\})$ is the eigenspace.

In the next chapter an "almost triangular" matrix is defined, and a method is given for reducing any matrix A to an almost tri-

angular form B whose eigenvalues are the same as A. The results of the present chapter are applied to the matrix $(B - \lambda I)$ where λ is an approximation to an eigenvalue of B. In Theorem 7.1 it is shown that the determinant of $(B - \lambda I)$ can be calculated by evaluating a polynomial $P_n(\lambda)$. This evaluation depends on reducing the matrix $(B - \lambda I)$ to a pivotal form P. An approximation λ_j is close to an actual eigenvalue when the modulus or norm of $P_n(\lambda_j)$ is small. If λ_j is considered the j'th iterate in a sequence of approximations and the norm of $P_n(\lambda_j)$ approaches zero, then the sequence approaches an eigenvalue of A. Since the pivotal form is obtained at each step in the iteration, one obtains simultaneously the corresponding eigenspace.

CHAPTER VI

REDUCTION OF A MATRIX TO ALMOST TRIANGULAR FORM

Definition 6.1. A square matrix A is almost lower triangular if ${}_ia^j = 0$ for $j > i + 1$.

If $f : X \to X$ and A is the $n \times n$ matrix determined by f relative to E, then A can be reduced to an almost triangular matrix B whose eigenvalues are the same as A. The process is a generalization of an idea of J. W. Givens (1958) in his method of solving the eigenvalue problem for real symmetric matrices.

Definition 6.2. An operator f is unitary if the composition ${}^*ff$ is the identity operator.

If f is unitary, then from Definition 6.2 $(\underline{x},\underline{y}) = ({}^*ff\ \underline{x},\underline{y}) = (f\ \underline{x}, f\ \underline{y})$ for all $\underline{x},\underline{y} \in X$. Thus upon setting $\underline{y} = \underline{x}$, it follows that $\|f\ \underline{x}\|^2 = \|\underline{x}\|^2$. This implies that $N(f) = \{\underline{0}\}$ so that f has an inverse and $f^{-1} = {}^*f$. Since ${}^*ff = ({}^*ff)(ff^{-1}) = f({}^*ff)\ f^{-1} = 1$, it follows that ${}^*f$ is also unitary. The matrix determined by the unitary operator f relative to any orthonormal basis is also called unitary.

Definition 6.3. If $T_{i,j}$ is an $n \times n$ unitary matrix and ${}_ib^j$ is zero in the matrix $B = {}^*T_{i,j}\ A\ T_{i,j}$ then $T_{i,j}$ is a plane

rotation matrix.

If an arrow is read as "is replaced by", then a rotation matrix is obtained from the unit matrix where ${}_{i+1}\delta^{i+1} \rightarrow c$, ${}_{j}\delta^{j} \rightarrow c$, ${}_{i+1}\delta^{j} \rightarrow -s$, and ${}_{j}\delta^{i+1} \rightarrow s$. Thus a rotation matrix has the following form:

$$T_{i,j} = \begin{bmatrix} 1 & & & & & & & \\ & \ddots & & & & & & \\ & & 1 & & & & & \\ & & & {}_{i+1}C^{i+1} & & {}_{i+1}-S^{j} & & \\ & & & & 1 & & & \\ & & & & & \ddots & & \\ & & & {}_{j}S^{i+1} & & {}_{j}C^{j} & & \\ & & & & & & 1 & \\ & & & & & & & \ddots \\ & & & & & & & & 1 \end{bmatrix}$$

A direct calculation shows that ${}^{*}T_{i,j}\, T_{i,j} = T_{i,j}\, {}^{*}T_{i,j} = I$ provided $c^2 + s^2 = 1$; and for the matrix $({}_{i}b^{j}) = {}^{*}T_{i,j}\, A\, T_{i,j}$, the element ${}_{i}b^{j}$ has the value $c\ {}_{i}a^{j} - s\ {}_{i}a^{i+1}$. If now c and s are chosen such that $c = {}_{i}a^{i+1}/[({}_{i}a^{i+1})^2 + ({}_{i}a^{j})^2]^{1/2}$ and $s = {}_{i}a^{j}/[({}_{i}a^{i+1})^2 + ({}_{i}a^{j})^2]^{1/2}$, then $c^2 + s^2 = 1$ and ${}_{i}b^{j} = 0$. It should be noted that $T_{i,j}$ applied on the right and ${}^{*}T_{i,j}$ applied on the left of a matrix

affects only the $(i + 1)$'th and j'th columns and rows. To obtain the almost triangular form, one begins with the rotation matrix $T_{1,3}$ to reduce ${}_1a^3$ to zero and sets $A_1 = {}^*T_{1,3}\, A\, T_{1,3}$. Using the matrix A_1 one computes ${}^*T_{1,4}\, A_1\, T_{1,4}$, and the result is a matrix A_2 such that ${}_1(A_2)^3 = {}_1(A_2)^4 = 0$ since the element ${}_1(A_1)^3$ is not affected. The process is continued across the first row until the rotation $T_{1,n}$ is applied with the result that ${}_1(A_{n-2})^j = 0$ for $j = 3, 4, \cdots, n$. For the matrix A_{n-2} one begins with the rotation T_{24} to annihilate ${}_2(A_{n-2})^4$ and continues across the second row terminating with $T_{2,n}$. In general one annihilates in order the elements in each row i using the rotation matrices $T_{i,j}$, $j = i + 2, i + 3, \cdots, n$ where $1 \le i \le n - 2$. Of course if the element is already zero, the rotation is not applied; and a brief calculation shows that at most $\frac{1}{2}(n - 1)(n - 2)$ rotations are needed to bring the matrix A into the form $({}_ib^j) = B = {}^*T\, A\, T$ where ${}_1b^j = 0$ for $j > i + 1$ and $T = T_{1,3} \cdots T_{1,n}\, T_{2,4} \cdots T_{2,n}\, T_{3,5} \cdots T_{n-2,n}$. Since $(B - \lambda I) = ({}^*T\, A\, T - \lambda I) = ({}^*T\, A\, T - \lambda\, {}^*T\, T) = {}^*T(A - \lambda I)\, T$, it follows that $|B - \lambda I| = |A - \lambda I|$ and hence the eigenvalues of B are the same as those of A. If B has an eigenvalue λ with corresponding eigenvector $\underline{x}$, then $B\,\underline{x} = \lambda\,\underline{x}$ implies that ${}^*T\, A\, T\,\underline{x} = \lambda\,\underline{x}$ or $A\, T\,\underline{x} = \lambda\, T\,\underline{x}$ and hence $T\,\underline{x}$ is the corresponding eigenvector of A. It should be noted that from a computational standpoint T is obtained if the columns of the unit matrix are operated on precisely as are the columns of A in the reduction. This will avoid full matrix multiplication. The preceding establishes the following.

Theorem 6.1. A well-defined sequence of unitary transfor-

mations, affecting at each step only two rows and two columns, will reduce an arbitrary n × n matrix A to an almost lower triangular matrix B whose eigenvalues coincide with those of A; and if $\underline{x}$ is an eigenvector of B, then T $\underline{x}$ is an eigenvector of A where T is the product of the unitary transformations defined above.

CHAPTER VII

THE POLYNOMIAL $P_n(\lambda)$

In this chapter $B = {}^{*}T\,A\,T$ is the matrix obtained by Theorem 6.1 from the $n \times n$ matrix A determined by the operator $f : X \to X$. It has been shown that $|B - \lambda I| = |A - \lambda I| = p_A(\lambda)$ where the matrix for the operator $(B - \lambda I)$ has the following form:

$$(B - \lambda I) = \begin{bmatrix} {}_1b^1 - \lambda & {}_1b^2 & 0 & 0 & \cdots & 0 \\ {}_2b^1 & {}_2b^2 - \lambda & {}_2b^3 & 0 & \cdots & 0 \\ \vdots & & \vdots & & \vdots & \\ {}_{n-1}b^1 & {}_{n-1}b^2 & {}_{n-1}b^3 & {}_{n-1}b^4 & \cdots & {}_{n-1}b^n \\ {}_nb^1 & {}_nb^2 & {}_nb^3 & {}_nb^4 & \cdots & {}_nb^n - \lambda \end{bmatrix}$$

One considers the elements on the first super-diagonal of $(B - \lambda I)$ and defines a constant $k = (-1)^{n+1}\left[\prod_{i=1}^{n-1} {}_ib^{i+1}\right]$. It is assumed through Theorem 7.1 that k is not zero. The case where

one or more of the elements in this product are zero is considered later. It will be shown that k is the proportionality constant which occurs in the special way that the determinant of $(B - \lambda I)$ is evaluated for any given value of λ.

If $\underline{x}^j$ is an n-column vector with components $\{{}_i x^j\}_{i=1}^n$, then the n equations in the n unknown components of $\underline{x}^j$ for the homogeneous system $(B - \lambda_j I)\underline{x}^j = \underline{0}$ can be written as follows:

$$P_1(\lambda_j) = ({}_1b^1 - \lambda_j)\,{}_1x^j + {}_1b^2\,{}_2x^j + 0 \cdot {}_3x^j + \cdots + 0 \cdot {}_nx^j = 0$$

$$P_2(\lambda_j) = {}_2b^1\,{}_1x^j + ({}_2b^2 - \lambda_j)\,{}_2x^j + {}_2b^3\,{}_3x^j + \cdots + 0 \cdot {}_nx^j = 0$$

$$\vdots \qquad\qquad \vdots \qquad\qquad \vdots$$

$$P_{n-1}(\lambda_j) = {}_{n-1}b^1\,{}_1x^j + {}_{n-1}b^2\,{}_2x^j + {}_{n-1}b^3\,{}_3x^j + \cdots + {}_{n-1}b^n\,{}_nx^j = 0$$

$$P_n(\lambda_j) = {}_nb^1\,{}_1x^j + {}_nb^2\,{}_2x^j + {}_nb^3\,{}_3x^j + \cdots + ({}_nb^n - \lambda_j)\,{}_nx^j = 0$$

It should be pointed out that the n'th equation above defines the polynomial $P_n(\lambda) = {}_nb^\alpha\,{}_\alpha x^j - \lambda_n x^j$. Theorem 7.1 of this chapter shows that this polynomial evaluated for a given value of λ is proportional to the determinant of the operator $(B - \lambda I)$. This determinant is an n'th degree polynomial in λ with the same zeros as the characteristic polynomial $p_A(\lambda)$. Under the assumption that $k \neq 0$, one can show that $P_n(\lambda)$ can also be regarded as an n'th degree polynomial in λ. For

consider the set $\{P_i(\lambda) = 0\}_{i=1}^{n-1}$ of linear equations defined above. Since $k \neq 0$, one can solve for ${}_2x^j$ in $P_1(\lambda) = 0$ obtaining a linear expression in λ. Substitution of this value of ${}_2x^j$ in $P_2(\lambda) = 0$ shows that ${}_3x^j$ is a quadratic expression in λ. If one continues these successive substitutions through $P_{n-1}(\lambda) = 0$, the n'th component can be represented as a polynomial in λ of degree $(n - 1)$. Thus $P_n(\lambda) = {}_nb^\alpha {}_\alpha x^j - \lambda {}_nx^j$ can be regarded as an n'th degree polynomial in λ.

Suppose now that λ_j is an approximation to an eigenvalue of B. The iterative process in the next chapter will produce a sequence (λ_j) such that $(|P_n(\lambda_j)|) \to 0$. Thus the sequence (λ_j) approaches an eigenvalue of B which by Theorem 6.1 is also an eigenvalue of A. The components of the vector $\underline{x}^j$ used in the evaluation of the norm of $P_n(\lambda_j)$ are chosen in a special way. The choice of $\underline{x}^j$ depends on the pivotal form of a certain matrix that is closely related to matrix B. This related matrix, which depends on λ and is denoted by M, is constructed in such a way that $N(M - \lambda I)$ always contains a non-zero vector where I is the $(n + 1) \times (n + 1)$ identity matrix.

Consider now the $(n + 1) \times (n + 1)$ matrix M with the following form:

$$(M) = \begin{bmatrix} {}_1b^1 & {}_1b^2 & 0 & 0 \cdots 0 & 0 \\ {}_2b^1 & {}_2b^2 & {}_2b^3 & 0 \cdots 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \\ {}_nb^1 & {}_nb^2 & {}_nb^3 & {}_nb^4 \cdots {}_nb^n & 1 \\ 0 & 0 & 0 & 0 \cdots 0 & \lambda \end{bmatrix}$$

It should be noted that the matrix M contains the matrix B as a submatrix in the upper left hand corner. The entries in the (n + 1)st row and column are all zero except ${}_{n-1}(M)^n$ and ${}_n(M)^n$ which have the values 1 and λ respectively. Thus any value for λ is an eigenvalue of M since the matrix $(M - \lambda I)$ contains a zero row and consequently $|M - \lambda I| = 0$. The matrix M is almost lower triangular by construction, and the product of its first superdiagonal elements is the constant k defined earlier. The matrix M just defined and the assumption that $k \neq 0$ are used in the following

Lemma 7.1. The null space of the operator $(M - \lambda I)$ for a fixed value of λ is non-trivial; and if $\underline{x}^j$ with components $({}_ix^j)_{i=1}^{n+1}$ is any solution, then the component ${}_1x^j$ is different from zero.

Proof. That non-trivial solutions exist follows from the fact that $|M - \lambda I| = 0$ for every value of λ. If λ has the fixed value λ_j, then by Theorem 5.3 the null space of $(M - \lambda_j I)$ is generated by the set $\{S_j\}$ of linearly independent column vectors. Let $\underline{x}^j \in N(M - \lambda_j I)$ and assume ${}_1x^j = 0$. Then the first equation in the system $(M - \lambda_j I)\underline{x}^j = \underline{0}$ implies that ${}_2x^j = 0$ since ${}_1(M)^i = 0$, $i = 3, 4, \cdots, (n + 1)$. But then the second equation implies that ${}_3x^j = 0$ since ${}_2(M)^i = 0$, $i = 4, 5, \cdots, (n + 1)$. The same reasoning carried through the n'th equation in the system implies that $\underline{x}^j$ is the zero vector. This is a contradiction since $\underline{x}^j \in \{S_j\}$ and cannot be the null vector.

Several important observations can be made concerning the solutions of the system $(M - \lambda_j I)\underline{x}^j = \underline{0}$. A basis $\{S_j\}$ for $N(M - \lambda_j I)$ is obtained by reducing the matrix $(M - \lambda_j I)$ to pivotal form P where $P = G(M - \lambda_j I)$. This reduction will never affect the bottom row of all zeros, so $\underline{\delta}^{n+1}$ subtracted from the last column of P will always be in the basis for the null space of $(M - \lambda_j I)$. Now the last column of P is simply the effect of G on the $(n + 1)$-column vector $\underline{\delta}^n$, and the transpose of this last column always has the form $({}_1y^j, {}_2y^j, \cdots, {}_ny^j, 0)$. Since by Lemma 7.1 the first component cannot be zero, one can define the n-column vector $\underline{x}^j$ with components $\{{}_iy^j/{}_1y^j\}_{i=1}^n$ where ${}_1x^j$ is unity. If $\underline{x}^j$ is the vector just defined and $\{P_i(\lambda_j)\}_{i=1}^n$ the set defined earlier for the system $(B - \lambda_j I)\underline{x}^j = \underline{0}$, then it follows directly that $P_i(\lambda_j) = 0$ for $i = 1, 2, \cdots, (n - 1)$. It is this vector $\underline{x}^j$ that is used in the next theorem which shows that $P_n(\lambda_j) = {}_nb^\alpha {}_\alpha x^j - (\lambda_j){}_nx^j$ is proportional to the determinant of $(B - \lambda_j I)$.

Theorem 7.1. Let P be the pivotal form of $(M - \lambda_j I)$ and $({}_1y^j, {}_2y^j, \cdots, {}_ny^j, 0)$ the transpose of the last column of P. If $\underline{x}^j$ is an n-column vector with components $\{{}_iy^j/{}_1y^j\}_{i=1}^n$, then $p_A(\lambda_j) = k \cdot P_n(\lambda_j)$.

Proof. Consider the matrix for $(B - \lambda_j I)$ in which the first column is multiplied by ${}_1x^j$. Now one multiplies column two by ${}_2x^j$ and adds to column one, multiplies column three by ${}_3x^j$ and adds to column one, and so on until column n is multiplied by ${}_nx^j$ and added to column one. The result is a matrix B_1

whose determinant value is the same as $(B - \lambda I)$. Since ${}_1x^j = 1$, B_1 has the following form:

$$(B_1) = \begin{bmatrix} P_1(\lambda_j) & {}_1b^2 & 0 & 0 & \cdots & 0 \\ P_2(\lambda_j) & ({}_2b^2 - \lambda_j) & {}_2b^3 & 0 & \cdots & 0 \\ \vdots & & \vdots & \vdots & & \\ P_{n-1}(\lambda_j) & {}_{n-1}b^2 & {}_{n-1}b^3 & {}_{n-1}b^4 & \cdots & {}_{n-1}b^n \\ P_n(\lambda_j) & {}_nb^2 & {}_nb^3 & {}_nb^4 & \cdots & ({}_nb^n - \lambda_j) \end{bmatrix}$$

Now by assumption $P_i(\lambda_j) = 0$ for $i = 1, 2, \cdots, (n - 1)$. Thus expanding in terms of the first column, one obtains $|B - \lambda_j I| = p_A(\lambda_j) = |B_1| = k \cdot P_n(\lambda_j)$. This result follows since the cofactor of the element $P_n(\lambda_j)$ in $|B_1|$ is $(-1)^{n+1}$ times the determinant of the submatrix of order $(n - 1)$ obtained from B_1 by deleting the first column and last row. This submatrix is lower triangular, and the value of its determinant is the product of its diagonal elements. But this product multiplied by $(-1)^{n+1}$ was previously defined as the constant k.

It was pointed out earlier that no use was made of the last row of zeros in $(M - \lambda_j I)$ when this matrix was reduced to pivotal form. Thus from a computational standpoint this last row may simply be ignored. The first n components of the vector $({}_1y^j, {}_2y^j, \cdots, {}_ny^j, 0)$, which were used to obtain $\underline{x}^j$, are then the effect of G on the n-column vector $\underline{\delta}^n$ where $P = G(B - \lambda_j I)$ is the pivotal from of $(B - \lambda_j I)$.

But since G will be the inverse of $(B - \lambda_j I)$ if λ_j is not an eigenvalue of B, one may as well record G by operating on I as the reduction proceeds. Then the transpose of the last column of G will be the vector $({}_1y^j, {}_2y^j, \cdots, {}_ny^j)$. This observation is used in the next chapter in producing the sequence of approximations $\{\lambda_j\}$ from which the sequence $\{|P_n(\lambda_j)|\}$ is calculated.

In the case k = 0 one or more of the elements ${}_ib^{i+1}$, i = 1, 2, $\cdots$, (n - 1) must be zero. When this occurs the complexity of the problem is actually reduced. Before stating a general procedure, it is believed that the following example corresponding to a 9 × 9 matrix for the almost triangular matrix B will help to clarify the procedure. Consider the matrix B obtained from A with the following form:

$$(B) = \begin{bmatrix} \times & \times & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \times & \times & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \times & \times & \times & \times & 0 & 0 & 0 & 0 & 0 \\ \times & \times & \times & \times & \times & 0 & 0 & 0 & 0 \\ \times & \times & \times & \times & \times & 0 & 0 & 0 & 0 \\ \times & \times & \times & \times & \times & \times & \times & 0 & 0 \\ \times & \times & \times & \times & \times & \times & \times & \times & 0 \\ \times & \times & \times & \times & \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times & \times & \times & \times & \times \end{bmatrix}$$

Let A_1, A_2, and A_3 denote respectively the 2 x 2, 3 x 3, and 4 x 4 matrices indicated in blocked form above. Then using Laplace's expansion theorem it is seen that $|B - \lambda I| = |A_1 - \lambda I| \cdot |A_2 - \lambda I| \cdot |A_3 - \lambda I|$, and hence $p_A(\lambda) = p_{A_1}(\lambda) \cdot p_{A_2}(\lambda) \cdot p_{A_3}(\lambda)$. Thus an eigenvalue of any of the three smaller matrices is also an eigenvalue of A. Since all of these matrices are almost triangular with the product of their first superdiagonal elements different from zero, Lemma 7.1 and Theorem 7.1 can be applied to obtain the eigenvalues and eigenspaces. Once an eigenvalue is found, then one uses Corollary 5.3.2 to obtain the corresponding eigenspace.

In the example above one notes by Theorem 7.1 that $p_{A_i}(\lambda) = k_i P_{n_i}(\lambda)$, $i = 1, 2, 3$ where $k_1 = {}_1b^2$, $k_2 = \prod_{i=3}^{4} {}_ib^{i+1}$, $k_3 = \prod_{i=6}^{8} {}_ib^{i+1}$, and n_i is the size of A_i. If $K = \prod_{i=1}^{3} k_i$, then one can write $p_A(\lambda) = K \prod_{k=2}^{4} P_k(\lambda)$. In general then, using the technique described above, one can write $p_A(\lambda)$ as a product of $j \leq n$ polynomials. For each i such that $1 \leq i \leq j$ the polynomial $p_{A_i}(\lambda)$ is proportional to $P_{n_i}(\lambda)$. Thus if K is the product of the non-zero elements on the first superdiagonal of B, then one can write $p_A(\lambda) = K \prod_{i=1}^{j} P_{n_i}(\lambda)$. The process reduces then to finding the eigenvalues in turn of the smaller matrices.

CHAPTER VIII

THE SEQUENCE $\{\lambda_j\}$

Let A be the n × n matrix determined by f : X → X relative to E and B = ${}^*T$ A T the almost lower triangular matrix obtained from A. For each approximation λ_j to an eigenvalue of B, one determines the vector $\underline{x}^j$ by Theorem 7.1 such that $P_i(\lambda_j) = 0$ for i = 1, 2, •••, (n - 1). If $\epsilon > 0$ is arbitrarily small and $|P_n(\lambda_j)| < \epsilon$, then λ_j is arbitrarily close to an eigenvalue of A with corresponding eigenvector $\underline{x}^j$. The aim then is to construct a sequence $\{\lambda_j\}$ of approximations such that the sequence $\{|P_n(\lambda_j)|\}$ converges to zero.

The first process described here to obtain the sequence $\{\lambda_j\}$ is one for which the author has been unable to determine conditions which would insure that $\{|P_n(\lambda_j)|\}$ is a null sequence. However, the sequence $\{\lambda_j\}$ does converge by the nature in which the next approximation is chosen; and consequently, the sequence $\{|P_n(\lambda_j)|\}$ must approach a limit. It is conjectured that one can always choose a sequence $\{\lambda_j\}$ in a prescribed way such that this limit is zero, and hence an eigenvalue can be found. This conjecture is based on the results obtained in solving numerous experimental problems on a digital computer. Before discussing the construction process of the sequence $\{\lambda_j\}$, it may be instructive to discuss some of these results. Since $p_A(\lambda)$ can be calculated for any value of λ, the process of finding an eigenvalue of A is equivalent to finding a zero of an n'th degree polynomial though

the analytic expression for $p_A(\lambda)$ is not known. Thus the process may be used to find the zeros of any explicit polynomial, and the results given in Tables I and II were obtained using polynomials of degree six and ten respectively. The results in Table I show that the process can find the zeros although all the zeros are very close with almost equal moduli. The zeros of the polynomial in Table II occur in widely separated groups where the zeros in each group are again close together with almost equal moduli. If one compares the columns of calculated zeros with those for the actual zeros, it will be seen that the zeros were determined with remarkable accuracy.

In order to construct the sequence $\{\lambda_j\}$ it is simpler to assume that all the zeros of $p_A(\lambda)$ lie in the interior of a square D which contains the rectangle determined in Chapter III. The boundaries of D are parallel to the rectangular axes in the complex plane with the length of each side denoted by d. To start the iterative process one divides D into quadrants $\{Q_i\}_{i=1}^4$ and computes the $\min_{1\le i\le 4}\{|P_n(\lambda_i)|\}$ where λ_i is the center of Q_i. The first element in the sequence $\{\lambda_j\}$ is then the approximation that corresponds to this minimum. Suppose that λ_1 is one (there may be more than one) of the approximations for which this minimum is attained. One now constructs a square D_1 that contains quadrant Q_1 with the restriction that $D_1 \subset D$. The length of each side of D_1 is $d_1 = (0.5 + a)d$ where $0 \le a < 0.5$. Thus if $a = 0$, then D_1 coincides with the quadrant Q_1. Now one divides the square D_1 into quadrants $\{Q_i^1\}_{i=1}^4$ and again computes $\min_{1\le i\le 4}\{|P_n(\lambda_i)|\}$ where λ_i is now the center of Q_i^1. The second element in the sequence $\{\lambda_j\}$ is then the approximation that corresponds to this minimum. Assuming that this approximation is λ_3, one constructs a square D_2 that con-

TABLE I

$$f(z) = \sum_{i=0}^{6} (a_i, b_i)\, z^i$$

i	(a_i, b_i)
0	$(\ 0.0000000 \times 10^{0}\ ,\ -1.0053292 \times 10^{5})$
1	$(\ 6.2550117 \times 10^{4}\ ,\ 6.2550117 \times 10^{4})$
2	$(-3.2431428 \times 10^{4}\ ,\ 0.0000000 \times 10^{0})$
3	$(\ 4.4840669 \times 10^{3}\ ,\ -4.4840669 \times 10^{3})$
4	$(\ 0.0000000 \times 10^{0}\ ,\ 6.9747711 \times 10^{2})$
5	$(-2.8930499 \times 10^{1}\ ,\ -2.8930499 \times 10^{1})$
6	$(\ 1.0000000 \times 10^{0}\ ,\ 0.0000000 \times 10^{0})$

Actual Zeros	Calculated Zeros
(4.8350 , 4.8350)	(4.8350000 , 4.8350000)
(4.8200 , 4.8200)	(4.8200000 , 4.8200000)
(4.8100 , 4.8100)	(4.8099999 , 4.8099999)
(4.8355 , 4.8355)	(4.8354999 , 4.8354999)
(4.8000 , 4.8000)	(4.8000000 , 4.8000000)
(4.8300 , 4.8300)	(4.8300000 , 4.8300000)

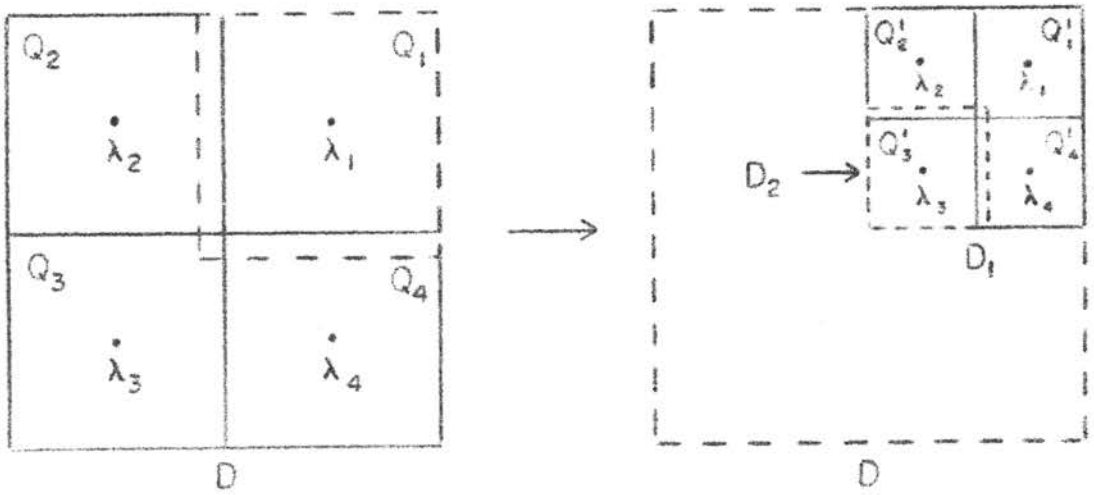
TABLE II

$$f(z) = \sum_{i=0}^{10} (a_i, b_i)\, z^i$$

i	(a_i, b_i)
0	$(5.2564917 \times 10^{16}, -2.9475105 \times 10^{16})$
1	$(-3.3636498 \times 10^{16}, 1.2393803 \times 10^{17})$
2	$(-6.1771960 \times 10^{16}, -1.0493940 \times 10^{17})$
3	$(4.2771960 \times 10^{16}, 9.5221140 \times 10^{15})$
4	$(8.0460358 \times 10^{13}, 9.5298119 \times 10^{14})$
5	$(-8.8699900 \times 10^{12}, 1.4361220 \times 10^{12})$
6	$(-1.4874623 \times 10^{09}, -5.3403376 \times 10^{10})$
7	$(2.1432931 \times 10^{08}, 5.4320499 \times 10^{07})$
8	$(-3.9491300 \times 10^{05}, 5.4320499 \times 10^{05})$
9	$(-6.0100000 \times 10^{02}, -1.1050000 \times 10^{08})$
10	$(1.0000000 \times 10^{00}, 0.0000000 \times 10^{00})$

Actual Zeros	Calculated Zeros
(-100 , 100)	$(-1.0000000 \times 10^{02}, 1.0000000 \times 10^{02})$
(-101 , 100)	$(-1.0100000 \times 10^{02}, 1.0000000 \times 10^{02})$
(-101 , 101)	$(-1.0100000 \times 10^{02}, 1.0100000 \times 10^{02})$
(200 , 300)	$(2.0000000 \times 10^{02}, 3.0000000 \times 10^{02})$
(201 , 300)	$(2.0100000 \times 10^{02}, 3.0000000 \times 10^{02})$
(201 , 301)	$(2.0100000 \times 10^{02}, 3.0100000 \times 10^{02})$
(001 , 000)	$(9.9999999 \times 10^{-1}, -4.6481929 \times 10^{-8})$
(000 , 001)	$(1.9072392 \times 10^{-11}, 1.0000000 \times 10^{00})$
(001 , 001)	$(9.9999999 \times 10^{-1}, 9.9999999 \times 10^{-1})$
(299 , -099)	$(2.9900000 \times 10^{02}, -9.9000000 \times 10^{01})$

tains quadrant Q_3^1 with the restriction that $D_2 \subset D_1$ and the length of each side of D_2 being $d_2 = (0.5 + a)d_1$. Then the entire process is repeated. Before giving the general formula for constructing the square D_{n+1} from D_n, it may be helpful to consider the following diagram showing the construction of the first two squares from the domain D.



The first square shows the quadrants of D with D_1 (indicated in the first square in dotted lines) constructed about Q_1 for a = 0.05. The second square shows the quadrants of D_1 with D_2 shown in dotted lines and constructed about Q_3^1 for the same value of a. In general then if D_n is divided into quadrants $(Q_i^n)_{i=1}^4$, one computes $\min_{1 \le i \le 4}(|P_n(\lambda_i)|)$ where λ_i is the center of Q_i^n. Assuming that λ_k corresponds to this minimum, one then constructs the square D_{n+1} which contains quadrant Q_k^n such that $D_{n+1} \subset D_n$ and $d_{n+1} = (0.5 + a)d_n$. The calculations in Tables I and II were obtained for a = 0.05.

Since the construction process produces a nested sequence of

squares with D_{n+1} properly contained in D_n, the sequence $\{\lambda_j\}$ must converge which implies that the sequence $\{|P_n(\lambda_j)|\}$ must also converge. A simple calculation shows that for small values of a the length of d_n is approximately $(1/2^n)d$ and $|\lambda_{n+1} - \lambda_n| < (1/2^{n+2})d$. For the test problems that have been done, there is evidence that the method is most effective when the eigenvalues are relatively close together. The following shows that d can be made unity by a simple transformation that puts the eigenvalues in the unit circle with trace zero. The transformation is easily recovered and does not destroy the almost triangular character.

Let $a,b \in C$ with $a \neq 0$, and $M = \max_{1\leq i\leq 4}(|V_i|)$ where V_i are the vertices of the rectangle that contains the set $\{\lambda_i\}_{i=1}^n$ of eigenvalues of B. Then $|B - \lambda I| = \prod_{i=1}^n(\lambda_i - \lambda) = p_A(\lambda)$, and

$$|a B + b I - \lambda I| = a^n\left|B - \left(\frac{\lambda - b}{a}\right)I\right| =$$
$$a^n \prod_{i=1}^n\left(\lambda_i - \frac{\lambda - b}{a}\right) = \prod_{i=1}^n(a \lambda_i + b - \lambda)$$

so that the set $\{\mu_i\}_{i=1}^n$ where $\mu_i = a \lambda_i + b$ contains the eigenvalues of a B + b I with multiplicities included. If T denotes the trace of B, then trace $(a B + b I) = \sum_{i=1}^n\mu_i = a \sum_{i=1}^n\lambda_i + n b = a T + n b$; and it follows that trace $(a B + b I) = 0$ for $b = -a T/n$. Furthermore

$$|\mu_i| = |a \lambda_i - a T/n| \leq |a|(|\lambda_i| + |T|/n) < |a|(M + |T|/n) ,$$

and for $a = n/(n M + |T|)$ it follows that $|\mu_i| < 1$, $i = 1, 2, \cdots, n$. Also if $\underline{x}$ is an eigenvector of a B + b I corresponding to μ_i, then $(a B + b I)\underline{x} = (\mu_i)\underline{x} = (a \lambda_i + b)\underline{x}$. Since $a \neq 0$, this implies that $\underline{x}$ is an eigenvector of B corresponding to the eigenvalue λ_i.

The success of many iterative methods will often depend on good

first approximations. Although this does not hold for the method described above, it is true of Newton's method for finding the zeros of a polynomial. (Householder, 1953). For the application of Newton's method, it is believed that the method of nested squares can be used to great advantage in obtaining first guesses using only a few iterations. It may be advantageous to use the transformed matrix with eigenvalues in the unit circle. In any case suppose this has been done to obtain an approximation λ_n to a zero of $p_A(\lambda) = |B - \lambda I|$.

If $P = G(B - \lambda_n I)$ is the pivotal form of $(B - \lambda_n I)$ and λ_n is not an eigenvalue of B, then G is the inverse of $(B - \lambda_n I)$; and

$$\text{trace } (C) = |B - \lambda_n I| \text{ trace } [(B - \lambda_n I)^{-1}] = p_A(\lambda_n) \text{ trace } (G)$$

where ${}_i(C)^j$ is the cofactor of ${}_j(B - \lambda_n I)^i$. Using this result and the definition for the derivative of a determinant, it follows for $p_A(\lambda) = |B - \lambda I|$ that

$$p_A'(\lambda_n) = -\sum_{i=1}^{n} \text{cofactor } ({}_ib^i - \lambda_n) = -\text{ trace } (C) = -p_A(\lambda_n) \text{ trace } (G)$$

where the prime denotes differentiation and $p_A'(\lambda_n)$ is the derivative of $|B - \lambda I|$ with respect to λ evaluated for $\lambda = \lambda_n$. Thus Newton's method can be applied to obtain λ_{n+1} as the next approximation where

$$\lambda_{n+1} = \lambda_n - \frac{p_A(\lambda_n)}{p_A'(\lambda_n)} = \lambda_n + \frac{1}{\text{trace } (B - \lambda_n I)^{-1}} = \lambda_n + \frac{1}{\text{trace } (G)} .$$

Since the trace (G) is known for each approximation λ_n, the iterative process is established to obtain the sequence $\{\lambda_j\}$. In either method one uses $P_n(\lambda_j)/\prod_{i=1}^{n}(\lambda_j - \lambda_i)^{g_i}$ for the test of convergence where eigenvalues $\{\lambda_i\}_{i=1}^{k}$ have already been determined, and g_i is the algebraic multiplicity of the eigenvalue λ_i obtained by Corollary 5.3.1.

CHAPTER IX

SUMMARY

The problem discussed in this thesis is that of devising a method for calculating all the eigenvalues and corresponding eigenspaces of a linear operator suitable for use with a high-speed digital computer. The eigenvalues and eigenspaces are found using the unique matrix A that is determined by the operator relative to a fixed orthonormal basis. Using a sequence of unitary transformations the matrix A is first reduced to an almost lower triangular matrix B with the same eigenvalues as A. It is shown that every square matrix is row equivalent to a pivotal form from which the null space of the original matrix can be determined. For each approximation λ to an eigenvalue of A, the determinant of $(B - \lambda I)$ can be calculated although the analytic expression for the characteristic polynomial is not known. Bounds for the spectrum of the operator are calculated, and two iterative methods are developed producing a sequence of approximations converging to an eigenvalue.

The most successful numerical methods at the present time are applied to linear operators that give rise to real symmetric matrices, and probably the best of these are those that may be termed indirect methods in which evaluation of the determinant is avoided. The iterative procedures developed in this thesis are direct methods which evaluate the characteristic polynomial for each approximating eigen-

value. One of the procedures uses the well known Newton's method for finding the roots of a polynomial. It is also well known that the success of this method in some cases depends on good first approximations. The other method uses a sequence of nested squares to produce a sequence of approximations converging to an eigenvalue. This method can be used to obtain good starting values for Newton's method using only the first few iterations, or it may be employed as a separate routine to calculate the roots of a general polynomial. Once an eigenvalue is found using either method, the corresponding eigenspace is obtained simultaneously from the pivotal form. Both procedures avoid the usually difficult problem of deflation.

A SELECTED BIBLIOGRAPHY

Aitken, A. C. "The Evaluation of the Latent Roots and Latent Vectors of a Matrix." Proceedings of the Royal Society of Edinburgh, Section A, 57 (1937), pp. 268-304.

Amir-Moez, A. R. and A. L. Fass. Elements of Linear Spaces. New York: Pergamon Press, 1962.

Bodewig, E. Matrix Calculus. New York: Amsterdam Interscience Publishers, 1959.

Brauer, Alfred. "Bounds for Characteristic Roots of Matrices." National Bureau of Standards, Applied Mathematics Series, 29 (1953), pp. 101-107.

Browne, E. T. "Limits to the Characteristic Roots of a Matrix." American Mathematical Monthly, 46 (1939), pp. 252-265.

Feller, W. and G. E. Forsythe. "New Matrix Transformations for Obtaining Characteristic Vectors." Quarterly of Applied Mathematics, 8 (1951), pp. 325-331.

Givens, W. "Computation of Plane Unitary Rotations Transforming a General Matrix to Triangular Form." Journal of the Society for Industrial and Applied Mathematics, 6 (1958), pp. 26-51.

_____. "Numerical Computation of the Characteristic Values of a Real Symmetric Matrix." Oak Ridge National Laboratory, ORNL-1574, 1954.

Greenstadt, J. "A Method for Finding Roots of Arbitrary Matrices." Mathematical Tables and Other Aids to Computation, 9 (1955), pp. 47-52.

Hestenes, M. R. "Determination of Eigenvalues and Eigenvectors of Matrices." National Bureau of Standards, Applied Mathematics Series, 29 (1953), pp. 89-95.

Hildebrand, F. B. Introduction to Numerical Analysis. New York: McGraw-Hill Book Company, Inc., 1956.

Householder, Alston S. Principles of Numerical Analysis. New York: McGraw-Hill Book Company, Inc., 1953.

Householder, Alston S. "Unitary Triangularization of a Nonsymmetric Matrix." Journal of the Association for Computing Machinery, 5 (1958), pp. 339-342.

James, G. and R. C. James, Eds. Mathematics Dictionary. Princeton: D. Van Nostrand Company, Inc., 1959.

Kincaid, W. M. "Numerical Methods for Finding Characteristic Roots and Vectors of Matrices." Quarterly of Applied Mathematics, 5 (1947), pp. 320-345.

Lanczos, Cornelius. Applied Analysis. Englewood Cliffs, New Jersey: Prentice Hall, Inc., 1956.

Lotkin, M. "Characteristic Values of Arbitrary Matrices." Quarterly of Applied Mathematics, 14 (1956), pp. 267-275.

Stoll, Robert R. Linear Algebra and Matrix Theory. New York: McGraw-Hill Book Company, Inc., 1952.

Taylor, Angus E. Introduction to Functional Analysis. New York: John Wiley and Sons, Inc., 1961.

Traussky, Olga. "Bounds for Characteristic Roots of Matrices." Duke Mathematical Journal, 15 (1948), pp. 1043-1044.

von Holdt, Richard E. "An Iterative Procedure for the Calculation of the Eigenvalues and Eigenvectors of a Real Symmetric Matrix." Journal of the Association for Computing Machinery, 3 (1956), pp. 223-239.

Wielandt, H. "Inclusion Theorems for Eigenvalues." National Bureau of Standards, Applied Mathematics Series, 29 (1953), pp. 75-79.

Wilkinson, J. H. "Stability of the Reduction of a Matrix to Almost Triangular and Triangular Forms by Elementary Transformations." Journal of the Association for Computing Machinery, 6 (1959), pp. 336-360.

_____. "The Use of Iterative Methods for Finding the Latent Roots and Vectors of Matrices." Mathematical Tables and Other Aids to Computation, 9 (1955), pp. 184-191.

VITA

Paul Eugene Harper

Candidate for the Degree of

Doctor of Philosophy

Thesis: EIGENVALUES AND EIGENSPACES OF GENERAL LINEAR OPERATORS ON A FINITE DIMENSIONAL INNER-PRODUCT SPACE

Major Field: Mathematics

Biographical:

Personal Data: Born near Garvin, Oklahoma, November 6, 1923, the son of Wheeler R. and Bertha Belle Harper.

Education: Attended grade school in Lincoln and Eldorado, Oklahoma; graduated from Lincoln High School in 1942; received the Bachelor of Science degree from the Oregon State University, with a major in Mathematics, in June, 1949; received the Master of Science degree from the Oregon State University, with a major in Mathematics, in August, 1950; attended the Los Alamos Graduate Center of the University of New Mexico; completed requirements for the Doctor of Philosophy degree in May, 1964.

Professional experience: Joined the Los Alamos Scientific Laboratory, Los Alamos, New Mexico, as a Research Assistant in September, 1950; was promoted to Staff Member in January, 1954; served as an applied mathematician on a variety of scientific problems, with constantly growing interest in numerical analysis and computer applications; was chosen in 1960 to complete requirements for the Doctor of Philosophy degree with the support of the Los Alamos Scientific Laboratory's Advanced Study Program.