Lecture Notes:
Numerical Methods for Large Scale Eigenvalue Problems

D. C. Sorensen
Draft: Spring 1999

1 Introduction

Large scale eigenvalue problems arise in a variety of settings. Two important areas are vibrational analysis of structures and linear stability analysis of fluid flow. The former analysis usually leads to symmetric eigenproblems where the goal typically is to determine the lowest modes. The latter analysis leads to non-symmetric eigenproblems and the interest is in determining if the eigenvalues lie in a particular half of the complex plane.

In both of these settings the discrete problem can become extremely large but only a few eigenvalues are needed to answer the question of interest.

At a later point detailed descriptions of applications will be given. The source of these large scale problems is often from the discretization of a partial differential equation. For example

\begin{equation}
\mathcal{L}u = u \lambda \text{ for } u \in \Omega,
\end{equation}

\begin{equation}
u = 0 \text{ for } u \in \partial \Omega,
\end{equation}

where \( \mathcal{L} \) is some linear differential operator. A number of techniques may be used to discretize \( \mathcal{L} \). The finite element method provides an elegant discretization. If \( \mathcal{W} \) is a space of functions in which the solution to (1.1) may be found and \( \mathcal{W}_n \subset \mathcal{W} \) is an \( n \) dimensional subspace with basis functions \( \{ \phi_j \} \) then an approximate solution \( u_n \) is expanded in the form

\begin{equation}
u_n = \sum_{j=1}^{n} \phi_j \xi_j.
\end{equation}

A variational or a Galerkin principle is used depending on whether \( \mathcal{L} \) is self-adjoint, to obtain

\begin{equation}
< \phi_i, \mathcal{L}(\sum_{j=1}^{n} \phi_j \xi_j) > = \lambda < \phi_i, \sum_{j=1}^{n} \phi_j \xi_j >,
\end{equation}
1 Introduction

where $\langle \cdot, \cdot \rangle$ is an inner product in $W_n$. This leads to the following systems of equations

\[ \sum_{j=1}^{n} < \phi_i, \mathcal{L} \phi_j > \xi_j = \lambda \sum_{j=1}^{n} < \phi_i, \phi_j > \xi_j, \]

for $1 \leq i \leq n$. We may rewrite (1.2) and obtain the matrix equation

\[ Ax = \lambda Mx, \]

where

\[ A_{i,j} = < \phi_i, \mathcal{L} \phi_j >, \]
\[ M_{i,j} = < \phi_i, \phi_j >, \]
\[ x^T = [\xi_1, \ldots, \xi_n]^T, \]

for $1 \leq i, j \leq n$. Typically the basis functions are chosen so that few entries in a row of $A$ or $M$ are nonzero. In structures problems $A$ is called the “stiffness” matrix and $M$ is called the “mass” matrix. In Chemistry and Physics $M$ is often referred to as the “overlap” matrix.

In particular, methods for solving the eigen-problem which avoid the use of matrix factorizations and similarity transformations are of interest. We are going to develop methods that require only matrix-vector products. Typically, only a few eigenpairs are required and the methods we shall develop require storage proportional to $nk$ where $k$ is the number of eigenpairs desired. Advantages of such methods are obvious and we list few.

- Sparsity of the matrices is exploited
- Matrices need not be stored, we only need a formula (subroutine) for computing the necessary matrix-vector product
- Parallelism is easy

1.1 Notation

The real and complex number fields are denoted by $\mathbb{R}$ and $\mathbb{C}$ respectively. The standard $n$-dimensional real and complex vectors are denoted by $\mathbb{R}^n$ and $\mathbb{C}^n$ and the symbols $\mathbb{R}^{m \times n}$ and $\mathbb{C}^{m \times n}$ will denote the real and complex matrices $m$ rows and $n$ columns. Householder notation will be used unless otherwise indicated. Scalars are denoted by lower case Greek letters, vectors are denoted by lower case Latin letters and matrices by capitol Latin letters. For example, a vector $x \in \mathbb{R}^n$ has the form

\[ x = \begin{pmatrix} \xi_1 \\ \vdots \\ \xi_n \end{pmatrix}, \]
while a matrix in $\mathbb{C}^{m \times n}$ is an array of the form
\[
A = \begin{bmatrix}
\alpha_{1,1} & \alpha_{1,2} & \cdots & \alpha_{1,n-1} & \alpha_{1,n} \\
\alpha_{2,1} & \ddots & \cdots & \ddots & \ddots \\
\vdots & \ddots & \ddots & \ddots & \ddots \\
\alpha_{m-1,1} & \cdots & \cdots & \cdots & \alpha_{m-1,n} \\
\alpha_{m,1} & \alpha_{m,2} & \cdots & \cdots & \alpha_{m,n}
\end{bmatrix},
\]
The transpose of an $m \times n$ matrix is defined to be
\[
A^T = \begin{bmatrix}
\alpha_{1,1} & \alpha_{1,2} & \cdots & \alpha_{1,n-1} & \alpha_{1,n} \\
\alpha_{2,1} & \ddots & \cdots & \ddots & \ddots \\
\vdots & \ddots & \ddots & \ddots & \ddots \\
\alpha_{n,1} & \cdots & \cdots & \cdots & \alpha_{n,m}
\end{bmatrix},
\]
or $A^T = (\alpha_{j,i})$. If $A \in \mathbb{C}^{m \times n}$ we define $A^H = (A^T)$, where we denote the complex conjugate of $\alpha_{i,j}$ by $\bar{\alpha}_{i,j}$. Often, the dimension of matrices and vectors will be implied through context and not mentioned explicitly.

For vectors in $\mathbb{R}^n$ ($\mathbb{C}^n$) the scalar or inner product is denoted by
\[
x^Ty = \sum_{j=1}^n \xi_j \eta_j \quad (x^H y = \sum_{j=1}^n \bar{\xi}_j \eta_j),
\]
and matrix multiplication $C = AB$ is defined in the usual way. The symbol, $\| \cdot \|$ will denote the Euclidean or 2-norm of a vector:
\[
\|x\| = (x^T x)^{\frac{1}{2}}, \quad (\|x\| = (x^H x)^{\frac{1}{2}})
\]
if $x \in \mathbb{R}^n$ ($x \in \mathbb{C}^n$). Recall that if $\| \cdot \|$ is a norm on $\mathbb{R}^m$ and $B \in \mathbb{R}^{m \times n}$ is of rank $n$ then
\[
\phi(x) = \|Bx\|
\]
is a norm on $\mathbb{R}^n$. We denote the natural or standard basis of $\mathbb{C}^n$ by the set
\[
\{e_j\}_{j=1}^n
\]
where $e_j \in \mathbb{C}^n$ has the value one in the $j$-th component and all remaining components equal to zero. It follows that the $j$-th component of $x$ is $e_i^T x = \xi_j$.

In similar fashion $e_i^T A e_j = \alpha_{i,j}$, picks out the $i, j$-th element of $A$, while the $i$-th row is $e_i^T A \in \mathbb{C}^{1 \times n}$ and the $j$-th column is $A e_j \in \mathbb{C}^m$.

The determinant of a square matrix $A$ is defined recursively to be
\[
det(A) = \sum_{j=1}^n (-1)^{j+1} \alpha_{1,j} \det(A_{1,j}),
\]
where $A_{1,j}$ is the submatrix obtained by deleting the $j$-th row and the first column of $A$ with
\[
det(\alpha) = \alpha
\]
specified for a $1 \times 1$ matrix with the scalar $\alpha$ as its entry.
1 INTRODUCTION

1.2 Special Matrices

The \( n \times n \) identity matrix is denoted by \( I \) and is defined by \( Ie_j = e_j, \ 1 \leq j \leq n. \) Sometimes \( I_n \) is used to emphasize dimension. A matrix \( R \) is upper triangular if \( \rho_{i,j} = 0 \) when \( i > j. \) A matrix \( H = (\gamma_{i,j}) \) is upper Hessenberg if \( \gamma_{i,j} = 0 \) when \( i > j + 1. \) A matrix is symmetric (Hermitian) if \( A = A^T, \ (A = A^H). \) \( A \in \mathbb{C}^{n \times n} \) is nonsingular if there exits a matrix \( B \) such that

\[ AB = BA = I; \]

\( B = A^{-1} \) denotes the inverse of \( A. \) \( A \) is normal if \( AA^H = A^HA. \) \( Q \) is said to be an unitary (orthogonal) matrix if \( Q^HQ = I = QQ^H(Q^TQ = I = QQ^T). \) The outer product of \( x \in \mathbb{R}^m \) and \( y \in \mathbb{R}^n \) is a matrix in \( \mathbb{R}^{m \times n} \)

\[ xy^T = \begin{bmatrix} x_1y_1 & x_1y_2 & \cdots & x_1y_m \\ x_2y_1 & x_2y_2 & \cdots & x_2y_m \\ \vdots & \vdots & \ddots & \vdots \\ x_my_1 & x_my_2 & \cdots & x_my_m \end{bmatrix}. \]

1.3 Eigenvalues

Let \( A \in \mathbb{C}^{n \times n}. \) The set of numbers \( \sigma(A) \equiv \{ \lambda \in \mathbb{C} : \text{rank}(\lambda I - A) < n \} \) is called the spectrum of \( A. \) The elements of this discrete set are the eigenvalues of \( A. \) It is easily shown that

\[ p(\tau) \equiv \det(\tau I - A). \]

is a polynomial of degree \( n \) and that the spectrum of \( A \) is the set of roots of this characteristic polynomial,

\[ \sigma(A) \equiv \{ \lambda : p(\lambda) = 0 \}. \]

Thus, \( A \) has precisely \( n \) eigenvalues counting multiplicities. From the definition, an eigenvalue is a number \( \lambda \) such that the matrix \( \lambda I - A \) is singular and hence, for each \( \lambda \in \sigma(A) \), there is a nonzero vector \( x \) such that

\[ (\lambda I - A)x = 0, \]

which implies

\[ Ax = x\lambda. \]

We call \( x \) a (right) eigenvector of \( A \) corresponding to the eigenvalue \( \lambda, \) and refer to \( (x, \lambda) \) as an eigenpair. The multiplicity \( n_a(\lambda) \) considered as a root of the characteristic polynomial is called the algebraic multiplicity of \( \lambda. \) Define

\[ S_\lambda = \{ x : (\lambda I - A)x = 0 \} = \text{null}(\lambda I - A). \]
1 INTRODUCTION

$S_\lambda$ is called the (right) eigenspace of $A$ corresponding to $\lambda$. We denote the dimension of $S_\lambda$ by $n_g(\lambda) \equiv \dim(S_\lambda)$, and this number is called the geometric multiplicity of $\lambda$.

Definition: If $n_g(\lambda) < n_a(\lambda)$ then $A$ is called a defective matrix. If $n_g(\lambda) > 1$ then $A$ is called derogatory. Thus $I$ is derogatory but non-defective. If $n_a(\lambda) = 1$ then $\lambda$ is simple.

1.4 Exercises

1. Show that $\text{det}(A) = \lambda_1\lambda_2 \cdots \lambda_n$ where $\{\lambda_1, \lambda_2, \cdots, \lambda_n\} = \sigma(A)$.

2. Show that $trace(A) = \lambda_1 + \lambda_2 + \cdots + \lambda_n$ where $trace(A) \equiv \alpha_{1,1} + \alpha_{2,2} + \cdots + \alpha_{n,n}$.

3. Consider the matrix

$$A = \begin{bmatrix} \mu & 1 & 0 \\ 0 & \mu & 1 \\ 0 & 0 & \mu \end{bmatrix}.$$ 

Show that $n_a(\mu) = 3$ and $n_g(\mu) = 1$. Give a basis for $S_\mu$.

4. Generalize (3) to a matrix with this structure of arbitrary order.

5. Prove for $A \in \mathbb{C}^{n \times n}$ that $n_g(\lambda) \leq n_a(\lambda)$ for all $\lambda \in \sigma(A)$.

6. Show that the diagonal elements of a triangular matrix are its eigenvalues.

7. Let the matrix

$$T = \begin{bmatrix} T_{1,1} & T_{1,2} & \cdots & T_{1,m-1} & T_{1,m} \\ 0 & T_{2,2} & \cdots & \cdots & \cdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & & T_{m-1,m-1} & T_{m-1,m} \\ 0 & 0 & \cdots & 0 & T_{m,m} \end{bmatrix}$$

be block upper triangular with diagonal blocks $T_{j,j}$. Prove that

$$\sigma(T) = \bigcup_{j=1}^m \sigma(T_{j,j}).$$

8. Prove that for every $\lambda \in \sigma(A)$ there is a vector $y \in \mathbb{C}^{n \times n}$ such that $y^H(\lambda I - A) = 0$. Such a vector is called a left eigenvector of $A$ corresponding to $\lambda$. Construct a left eigenvector corresponding to $\mu$ for the matrix in exercise (3). What is the relationship of this vector to the right eigenvector. Develop all of the notions concerning multiplicity, eigenspaces etc. for left eigenvectors. Is the geometric multiplicity with respect to left eigenvectors the same as for right eigenvectors corresponding to the same eigenvalue?
1 INTRODUCTION

1.5 Schur Decomposition

The central theorem in the study of numerical algorithms for eigenproblems is the Schur decomposition. It states that there is a unitary basis in which the representation of a matrix $A \in \mathbb{C}^{n \times n}$ is upper triangular.

**Theorem 1.1 (Schur Decomposition).** Let $A \in \mathbb{C}^{n \times n}$. Then there is a unitary matrix $Q$ and an upper triangular matrix $R$ such that

\[
AQ = QR.
\]

The diagonal elements of $R$ are the eigenvalues of $A$.

**Proof:** We shall prove (1.3) by induction on $n$. The remainder of the proof is left as an exercise. The result is clearly true for matrices of order 1. Assume that (1.3) is true for matrices of order less than $n$. Let $\lambda$ be any eigenvalue of $A$ and $q$ a corresponding normalized eigenvector so that $Aq = q\lambda$ with $\|q\| = 1$. Construct a unitary matrix $Q_1$ such that $Q_1e_1 = q$ (e.g., via Householder transformation). Then $Q_1 = (q, \hat{Q}_1)$ with $\hat{Q}_1^Hq = 0$. Now

\[
AQ_1 = (Aq, A\hat{Q}_1) = (q\lambda, A\hat{Q}_1)
\]

and thus

\[
Q_1^H AQ_1 = \begin{bmatrix}
\lambda & q^HA\hat{Q}_1 \\
0 & \hat{Q}_1^H \hat{A}_1
\end{bmatrix} = \begin{bmatrix}
\lambda & r_1^H \\
0 & \hat{A}_1
\end{bmatrix},
\]

with $\hat{A} \in \mathbb{C}^{(n-1) \times (n-1)}$. By the induction hypothesis there exist an unitary matrix $\hat{Q}$ and an upper triangular matrix $\hat{R}$ such that $\hat{A}\hat{Q} = \hat{Q}\hat{R}$. Thus

\[
Q_1^H AQ_1 = \begin{bmatrix}
1 & 0 \\
0 & \hat{Q}
\end{bmatrix} \begin{bmatrix}
\lambda & r_1^H \\
0 & \hat{R}
\end{bmatrix} \begin{bmatrix}
1 & 0 \\
0 & \hat{Q}^H
\end{bmatrix} = Q_2 R Q_2^H,
\]

where $r_1^H = \hat{Q}^H\hat{Q}$ and

\[
Q_2 = \begin{bmatrix}
1 & 0 \\
0 & \hat{Q}
\end{bmatrix}.
\]

Hence $AQ = QR$ where $Q = Q_1Q_2$ and the induction is complete.

The Schur decomposition provides a means to develop additional fundamental decompositions. The fundamental structure of Hermitian and Normal matrices follows readily from the Schur decomposition and is expressed in the following two Lemmas.

**Lemma 1.2** A matrix $A \in \mathbb{C}^{n \times n}$ is normal ( $AA^H = A^H A$ ) if and only if $A = QAQ^H$ with $Q \in \mathbb{C}^{n \times n}$ unitary and $\Lambda \in \mathbb{C}^{n \times n}$ diagonal. The diagonal entries of $\Lambda$ are the eigenvalues of $A$ and the columns of $Q$ are the corresponding eigenvectors.
1 INTRODUCTION

Proof: From the Schur decomposition \( AQ = QR \) it follows that \( A \) is normal if and only if
\[
RR^H = R^HR,
\]
for if this property holds it is easily shown that \( A \) is normal and if \( A \) is normal then
\[
RR^H = Q^H AQ^H A^H Q = Q^H AA^H Q = Q^HA^H QA = Q^HA^H QQ^H AQ = R^HR.
\]
However, an upper triangular matrix is normal if and only if it is diagonal (see exercise (2) below).

Lemma 1.3 A matrix \( A \in \mathbb{C}^{n \times n} \) is Hermitian (\( A = A^H \)) if and only if \( A = Q\Lambda Q^H \) with \( Q \in \mathbb{C}^{n \times n} \) unitary and \( \Lambda \in \mathbb{R}^{n \times n} \) diagonal.

Proof: If \( A = Q\Lambda Q^H \) as in the hypothesis then it is clearly Hermitian. On the other hand, a Hermitian matrix is necessarily normal and is therefore unitarily similar to a diagonal matrix \( \Lambda \) from the Lemma (1.2). Therefore, this diagonal matrix must also be normal and hence real. (see exercise (3) below)

This lemma indicates that the eigenvalues of a Hermitian matrix are real and the eigenvectors are mutually orthogonal. The eigensystems of Hermitian matrices possess a number of important properties that greatly simplify computational techniques for approximating them. These will be developed later.

1.6 Exercises

1. Complete the proof of the Schur Decomposition. Use determinants to show that the diagonal elements of \( R \) are the eigenvalues of \( A \).

2. Prove that an upper triangular matrix \( R \) is normal if and only if it is diagonal.

3. Show directly from the Schur decomposition that if \( A \) is Hermitian then \( R \) is also Hermitian and thus a real diagonal matrix. If \( A \) is real and symmetric show that the eigenvectors are real and orthogonal.

4. Is a Schur decomposition unique in any sense? Show that for a given matrix \( A \) there is a Schur decomposition \( A = QRQ^H \) with the eigenvalues of \( A \) appearing on the diagonal of \( R \) in any specified order.

5. Suppose \( A, B, X \in \mathbb{C}^{n \times n} \) satisfy \( AX = XB \) with \( X \) nonsingular. Prove (using eigenvectors) that \( \sigma(A) = \sigma(B) \).

6. If there is a nonsingular \( X \) such that \( AX = XB \) then \( A \) is said to be similar to \( B \), i.e. \( A \sim B \). Show that
\[
A \sim A, \quad A \sim B \Rightarrow B \sim A, \quad A \sim B \quad \text{and} \quad B \sim C \Rightarrow A \sim C.
\]
1 INTRODUCTION

7. Give an example where \( \sigma(A) = \sigma(B) \) but \( A \) is not similar to \( B \). What is the weakest additional hypothesis that will imply \( A \sim B \)?

8. Give an algorithm for computing the eigenvectors of an upper triangular matrix \( R \) which has distinct eigenvalues. Discuss the sensitivity of these eigenvectors to perturbations when two or more eigenvalues of \( R \) coalesce. Hint: Partition

\[
R = \begin{bmatrix}
\hat{R} & r \\
0 & \rho
\end{bmatrix},
\]

and \( x^T = [\hat{x}, 1]^T \) and solve \( (R - \rho I)\hat{x} = -r \).

9. Prove that if the eigenvalues of \( A \in \mathbb{C}^{n \times n} \) are distinct then \( A \) is diagonalizable, i.e. show that there exist a nonsingular matrix

\[
X = \begin{bmatrix} x_1, \cdots, x_n \end{bmatrix},
\]

such that \( AX = X\Lambda \) where \( \Lambda \) is a diagonal matrix of the eigenvalues of \( A \) and \( Xe_j \) is the corresponding eigenvector for eigenvalue \( \lambda_j \) the \( j \)-th diagonal element of \( \Lambda \).

1.7 Eigen-decompositions of \( \mathbb{C}^n \)

We have shown that a normal matrix \( A \) has a decomposition

\[ A = Q\Lambda Q^H, \ Q^H Q = I, \ \Lambda \text{ is a diagonal matrix.} \]

Normality of \( A \) is a necessary and sufficient condition for such a decomposition. Now suppose that a normal matrix \( A \) has \( k \) distinct eigenvalues \( \lambda_1, \lambda_2, \cdots, \lambda_k \) where \( \lambda_j \) has multiplicity \( n_j \). By reordering eigenvalues and vectors we can arrange this decomposition in the form

\[ A = \sum_{j=1}^{k} \lambda_j Q_j Q_j^H, \]

where

\[ Q = (Q_1, Q_2, \cdots, Q_k), \ Q_i^H Q_j = 0, \ i \neq j, \ Q_j^H Q_j = I_{n_j}, \ \sum_{j=1}^{k} n_j = n. \]

1.8 Exercise

1. Recall that \( \mathcal{S}_{\lambda_j} = \{ x : (\lambda_j I - A)x = 0 \} \). Show that \( P_j \equiv Q_j Q_j^H \) is an orthogonal projector onto \( \mathcal{S}_{\lambda_j} \) and that \( \mathcal{S}_{\lambda_j} = \text{range}(Q_j) \).

2. Show that \( \sum_{j=1}^{k} P_j = \sum_{j=1}^{k} Q_j Q_j^H = QQ^H = I_n \) so that any \( b \in \mathbb{C}^n \) may be written uniquely as \( b = \sum_{j=1}^{k} Q_j (Q_j^H b) = \sum_{j=1}^{k} q_j \) where \( q_j = Q_j (Q_j^H b) \in \mathcal{S}_{\lambda_j} \). Conclude that \( \mathbb{C}^n = \mathcal{S}_{\lambda_1} \oplus \mathcal{S}_{\lambda_2} \oplus \cdots \oplus \mathcal{S}_{\lambda_k} \) is the direct sum of the subspaces \( \mathcal{S}_{\lambda_j} \).
1 INTRODUCTION

3. Show that if $s \in S_{\lambda_j}$ then $As = s\lambda_j$. Conclude that $AS_{\lambda_j} \subset S_{\lambda_j}$.

1.9 Invariant Subspaces

We have seen in the case of normal matrices that the space $C^n$ is decomposed into the direct sum of the eigen-spaces $S_{\lambda_j}$ of the matrix $A$, and that these subspaces are mutually orthogonal to one another. As we shall see, this property has important consequences numerically since it limits the effects of perturbations (in particular perturbations due to round off).

It is natural therefore to try to understand as best we can the structure of the subspaces $S_{\lambda}$ of general matrices. An understanding of this structure is important to the development of the algorithms for finding a few selected eigenvalues and vectors of a general matrix. As we shall see, the general case is far more complicated than the normal case. Eigenspaces are special cases of a more general class of subspaces associated with a matrix.

Definition: A subspace $\mathcal{S}$ of $C^{n \times n}$ is called an invariant subspace of $A \in C^{n \times n}$ if $AS \subset \mathcal{S}$.

It is straightforward to show

**Lemma 1.4** If $A \in C^{n \times n}$, $X \in C^{n \times k}$ and $B \in C^{k \times k}$ satisfy

\begin{equation}
AX = XB,
\end{equation}

then $\mathcal{S} = \text{Range}(X)$ is an invariant subspace of $A$. Moreover, if $X$ has column rank $k$ then the columns of $X$ form a basis for this subspace and $\sigma(B) \subset \sigma(A)$.

**Proof:** The fact that $\mathcal{S}$ is invariant follows immediately from (1.4) and the assumption of full rank assures $X$ forms a basis. Suppose $By = y\lambda$ is an eigenpair for $B$. Then

$$AXy = XB = Xy\lambda$$

i.e. $Ax = x\lambda$ with $x = Xy$ so $\sigma(B) \subset \sigma(A)$.

1.10 Exercises

1. Why is $x \neq 0$ in the proof of Lemma (1.4) ?

2. Show that $AX = XB$ implies that $\psi(A)X = X\psi(B)$ for any polynomial $\psi$. Recall that if $\psi(\lambda) = \gamma_0 + \gamma_1 \lambda + \cdots + \gamma_k \lambda^k$ then $\psi(A) = \gamma_0 I + \gamma_1 A + \cdots + \gamma_k A^k$.

In many cases it is desirable to have a unitary basis rather than an arbitrary one. It is straightforward to obtain a unitary basis from any given basis by obtaining a $QR$ factorization. We develop this formally in the following lemma.
Lemma 1.5 Assume the hypothesis of Lemma (1.4). If $X = QU$ with $Q^H Q = I_k$ and $U$ an upper triangular matrix then the columns of $Q$ from an orthonormal basis for $S$.

Proof: If

$$AX = XB, \quad X = QU, \quad Q^H Q = I_k$$

then $X$ having rank $k$ implies that $U$ is non-singular. Hence

$$AQU = QUB,$$

so

$$AQ = Q(UBU^{-1}) = Q̂B.$$  

Note that $σ(̂B) = σ(B) ⊂ σ(A)$ and $z \in S$ implies $z = Xw$ for some vector $w \in C^k$ and $Q(Rw) = Q̂w$. So the columns of $Q$ are an orthonormal basis for $S$.

Examination of the proof of the Schur decomposition indicates without loss of generality that a matrix $A ∈ C^{n×n}$ may be decomposed into

$$AQ = QR, \quad Q^H Q = I, \quad R \text{ is upper triangular},$$  

with eigenvalues of $A$ appearing in any order as diagonal elements of $R$ (see exercise 1.6.4). Thus it is possible to obtain $R$ in the form

$$R = \begin{bmatrix}
R_{1,1} & R_{1,2} & \cdots & R_{1,k-1} & R_{1,k} \\
R_{2,1} & R_{2,2} & \cdots & R_{2,k-1} & R_{2,k} \\
\vdots & \vdots & \ddots & \vdots & \vdots \\
0 & R_{k-1,1} & \cdots & R_{k-1,k-1} & R_{k-1,k} \\
0 & 0 & \cdots & R_{k,k-1} & R_{k,k}
\end{bmatrix},$$

where $σ(R_{j,j}) = \{λ_j\}$ where $λ_j \in σ(A)$ for $1 ≤ j ≤ k$ are the distinct eigenvalues of $A$.

The following lemma shows that it is possible to completely decouple the diagonal blocks of $R$ at the expense of losing a unitary basis.

Lemma 1.6 Suppose

$$R = \begin{bmatrix}
R_1 & S \\
0 & R_2
\end{bmatrix},$$

with $σ(R_1) \cap σ(R_2) = \emptyset$. Then

$$R \sim \begin{bmatrix}
R_1 & 0 \\
0 & R_2
\end{bmatrix}.$$
Proof: We shall construct a similarity transformation. Let $R_1 \in \mathbb{C}^{k \times k}$, $R_2 \in \mathbb{C}^{n-k \times n-k}$, and $S$ and $Z \in \mathbb{C}^{k \times n-k}$ where $1 \leq k < n$. Note that

$$\begin{bmatrix} I & Z \\ 0 & I \end{bmatrix} \begin{bmatrix} R_1 & S \\ 0 & R_2 \end{bmatrix} = \begin{bmatrix} R_1 & 0 \\ 0 & R_2 \end{bmatrix} \begin{bmatrix} I & Z \\ 0 & I \end{bmatrix}$$

if and only if $S + ZR_2 = R_1X$ has a unique solution $Z$, i.e.

$$R_1Z - ZR_2 = S.$$ 

This system of equations may be solved by noting that

$$R_1Ze_j - ZR_2e_j = Se_j,$$

i.e.

$$R_1z_j - \rho_{j,j}z_j - \left( \sum_{i=1}^{j-1} z_i \rho_{i,j} \right) = s_j,$$

where

$$R_2 = (\rho_{i,j}), z_j = Ze_j, s_j = Se_j.$$ 

Hence the columns $z_j$ may be determined in the order $j = 1, 2, \ldots, k$ by successively solving the triangular systems

$$(1.7) \quad (R_1 - \rho_{j,j}I)z_j = s_j + \sum_{i=1}^{j-1} z_i \rho_{i,j}.$$ 

Each of these equations has a unique solution since $\sigma(R_1) \cap \sigma(R_2) = \emptyset$ implies that $R_1 - \rho_{j,j}I$ is nonsingular for $1 \leq j \leq k$.

The following figure shows this solution method schematically:

![Diagram](image)

In Figure 1.10, the vertical lines denote the $j$-th column of the respective matrices. The dotted line in $R_2$ and the shaded region of the second $Z$ matrix indicate that the previously computed columns of $Z$ must participate in the calculation of the $j$-th column.
1.11 Exercise

1. Apply Lemma (1.6) repeatedly to \( R \) in (1.6) to obtain a nonsingular \( \hat{X} \) such that \( R \hat{X} = \hat{X} \hat{R} \) where

\[
\hat{R} = \begin{bmatrix}
R_{1,1} & 0 & \cdots & 0 & 0 \\
0 & R_{2,2} & 0 & & \\
\vdots & \vdots & \ddots & \ddots & \vdots \\
0 & 0 & \cdots & R_{k-1,k-1} & 0 \\
0 & 0 & \cdots & 0 & R_{k,k}
\end{bmatrix},
\]

is block diagonal.

Now we have from (1.5) that

\[A Q \hat{X} = Q R \hat{X} = Q \hat{X} \hat{R}\]

and we may express this as \( AX = X \hat{R} \) where \( X = Q \hat{X} \). In partition form

\[A(X_1, X_2, \ldots, X_k) = (X_1, X_2, \ldots, X_k) \begin{bmatrix}
R_{1,1} & 0 & \cdots & 0 \\
0 & R_{2,2} & 0 & \cdots \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & R_{k,k}
\end{bmatrix}.
\]

Setting \( Y = X^{-1} \) gives \( A = X \hat{R} Y \). Partitioning the rows of

\[Y = \begin{bmatrix}
Y_1^H \\
Y_2^H \\
\vdots \\
Y_k^H
\end{bmatrix},
\]

gives

\[A = X_1 R_{1,1} Y_1^H + X_2 R_{2,2} Y_2^H + \cdots + X_k R_{k,k} Y_k^H.
\]

Since \( YX = I \) we have

\[Y_i^H X_j = \begin{cases}
I_{n_j} & i = j \\
0 & i \neq j,
\end{cases}
\]

where \( n_j \) is the order of \( R_{j,j} \). Moreover,

\[I = \sum_{j=1}^{k} X_j Y_j^H \text{ since } XY = I.
\]

Thus any vector \( b \in \mathbb{C}^n \) may be written uniquely as

\[(1.8) \quad b = \sum_{j=1}^{k} X_j (Y_j^H b) = \sum_{j=1}^{k} x_j,
\]
1 INTRODUCTION

where \( x_j \in \mathcal{S}_j \equiv \text{Range}(X_j) \). Hence

\[
\mathbb{C}^n = \mathcal{S}_1 \cup \mathcal{S}_2 \cup \cdots \mathcal{S}_k \quad \text{with} \quad \mathcal{S}_i \cap \mathcal{S}_j = \{0\} \quad \text{if} \quad i \neq j.
\]

There are two main differences between this result and the result for normal matrices. The spaces \( \mathcal{S}_j \) are not mutually orthogonal and a polynomial of degree greater than 1 in \( A \) may be needed to annihilate a space \( \mathcal{S}_j \).

1.12 Exercises

1. Show that \( P_j \equiv X_j Y_j^H \) is a projector of \( \mathbb{C}^n \) onto \( \mathcal{S}_j \equiv \text{Range}(X_j) \).

2. Prove that a projector \( P_j \) is orthogonal if and only if \( P_j \) is Hermitian.

3. Assume the columns of \( X \) each have norm 1 and suppose that \( \dim(\mathcal{S}_j) = 1 \). Show that \( \|P_j\| = \frac{1}{\cos \theta_j} \) where \( \theta_j \) is the angle between \( X_j \) and \( Y_j \).

4. Suppose \( X_j R_j Y_j^H = \lambda_j P_j \) where \( P_j \) is a projector onto \( \mathcal{S}_j \) for all \( j \). Show that \( A \) is a normal matrix if and only if \( P_j \) is an orthogonal projector for every \( j \).

5. Define \( \mathcal{U}_j \equiv \text{Range}(Y_j) \). Show that \( \mathcal{U}_i \perp \mathcal{S}_j \) when \( i \neq j \).

6. Show that although \( \mathcal{S}_i \cap \mathcal{S}_j = \{0\} \) and \( \mathcal{U}_i \cap \mathcal{U}_j = \emptyset \) for \( i \neq j \) that \( \mathcal{U}_i \perp \mathcal{U}_j \) and \( \mathcal{S}_i \perp \mathcal{S}_j \) does not necessarily hold for every \( i \neq j \). Show that if this property does hold for every \( i \neq j \) then \( A \) must be normal.

7. From exercise (1.10.2) we know that \( \psi(A)X_j = X_j \psi(R_{j,j}) \) for any polynomial \( \psi \). Let \( \psi_j(\lambda) = (\lambda_j - \lambda)^{n_j} \). Show that \( \psi(R_{j,j}) = 0 \).

1.13 The Cayley-Hamilton Theorem

With the structure of \( \mathbb{C}^n \) with respect to the eigen-spaces of a given matrix \( A \) it is relatively easy to prove the following

**Theorem 1.7** (Cayley-Hamilton) If \( p(\lambda) \) is the characteristic polynomial of \( A \in \mathbb{C}^{n \times n} \) then \( p(A) = 0 \).

**Proof:** Since \( A \) commutes with itself and the identity matrix, it is possible to write

\[
p(A) = p_j(A)(A - \lambda_j I)^{n_j}
\]

for each distinct eigenvalue \( \lambda_j \in \sigma(A) \) where \( p_j \) is a polynomial of degree \( n - n_j \). Since

\[
(A - \lambda_j I)^{n_j} X_j = X_j (R_{j,j} - \lambda_j I)^{n_j} = 0
\]

we have \( p(A)X_j = 0 \) for \( 1 \leq j \leq k \). From (1.8) it follows that \( p(A)b = 0 \) for every \( b \in \mathbb{C}^n \) and hence that \( p(A) = 0 \).
1 INTRODUCTION

This intriguing theoretical result has profound implications for algorithms to compute eigenvalues and solutions of linear systems. With respect to linear systems $Ax = b$ we know that $p(A)b = 0$ and $p(0) = (-1)^n \det(A) \neq 0$. Thus we may write the polynomial $p(\lambda)/p(0) = 1 - \lambda \phi(\lambda)$. Hence

\[(1.9) \quad (I - A\phi(A))b = \frac{1}{p(0)} p(A)b = 0\]

to give

\[Ax = b \quad \text{with} \quad x \equiv \phi(A)b.\]

Therefore, we are highly motivated to devise algorithms that construct approximate solutions of the form $x = \psi(A)b$ where $\psi$ is a polynomial. Moreover, this result hints at the possibility of an algorithm which finds a solution of this type by computing at most $n$ matrix vector products $w = Av$ since $\psi(A)b$ may be constructed one term at a time from such products. Since the expense of such an algorithm would be directly proportional to the degree of the polynomial $\psi$, the question of minimizing the degree of the polynomial that is satisfied by $A$ or that annihilates a given vector $b$.

This discussion motivates the following two definitions

**Definition:** A monic polynomial $\phi$ of degree $k$ is called the *minimal polynomial* of $A$ if $\phi(A) = 0$ and no monic polynomial of lower degree has this property.

**Definition:** A monic polynomial $\psi$ of degree $k$ is called the *minimal polynomial of A with respect to a vector b* if $\psi(A)b = 0$ and no monic polynomial of lower degree has this property.

1.4 Exercises

1. Use a degree argument to prove that each of the two definitions defines a unique polynomial.

2. Give a non-trivial ($b \neq 0$) example to show that these two concepts are distinct.

3. Prove that when $A$ is normal its minimal polynomial consists of the product of $k$ linear factors $(A - \lambda_j I)$. One for each distinct eigenvalue $\lambda_j$ of $A$.

4. Prove that $A$ is diagonalizable if and only if its minimal polynomial $\psi(\tau) = \Pi_{j=1}^k (\tau - \lambda_j)$, where $\{\lambda_j\}$ are the $k$ distinct eigenvalues of $A$.

5. Prove that every eigenvector of $A$ may be written in the form $\psi(A)b$ where $\psi$ is a polynomial of degree at most $n-1$ and $b$ (almost) any nonzero vector. Which vectors $b$ will not work here?
6. Prove that the polynomial \( \phi(\lambda) \) defined as in (1.9) is the (unique) polynomial of degree \( n - 1 \) that interpolates the function \( \frac{1}{\lambda} \) at the eigenvalues of \( A \). Also show that \( A^{-1} = \phi(A) \). How do you make this well defined when there are multiple eigenvalues?

7. Suppose \( A \) is nonsingular and \( \psi(\tau) \) is the minimal polynomial of \( A \) with respect to \( b \). Prove that \( \psi(0) \neq 0 \). How can this be used to construct a solution to the equation \( Ax = b \)?

8. Let \( \phi(\tau) \) be the minimal polynomial of \( A \). Prove that \( \phi \) divides any polynomial \( \psi \) such that \( \psi(A) = 0 \). (Hint: Express \( \psi(\tau) = \alpha(\tau)\phi(\tau) + \rho(\tau) \) where \( \alpha \) is some polynomial and \( \rho \) is a polynomial of degree less than degree of \( \phi \).

9. Suppose \( \phi(\tau) \) is the minimal polynomial of \( A \) and that \( \phi(\tau) = (\tau - \lambda)^{\ell}\hat{\phi}(\tau) \) with \( \hat{\phi}(\lambda) \neq 0 \). Given a nonzero vector \( v \), define \( x_j = (A - \lambda I)^{\ell-j}\hat{\phi}(A)v \) for \( j = 1, 2, \ldots, \ell \). Prove that there is a vector \( v \) such that \( \{x_j\} \) are linearly independent and

\[
AX_\lambda = X_\lambda J_\ell(\lambda)
\]

where \( X_\lambda = [x_1, x_2, \ldots, x_\ell] \) and \( J_\ell(\lambda) \) is a Jordan block of order \( \ell \) (see next section for definition of \( J \)).

1.15 The Jordan Canonical Form

We see that the structure of the decomposion of \( \mathbb{C}^n \) with respect to eigen-spaces of a matrix \( A \) is closely related to matrix polynomials and their degree. The next question we should ask is what is the smallest degree \( m \) of a polynomial in \( A \) that will annihilate the space \( S_j \). Let’s consider a special case. Define

\[
J_m(\mu) = \begin{bmatrix}
\mu & 1 & 0 & \cdots & 0 \\
\mu & 1 & \cdots & & \\
& \ddots & \ddots & \ddots & 0 \\
& & \ddots & 1 & \\
& & & \mu
\end{bmatrix},
\]

(1.10)

to be an upper triangular matrix of order \( m \) with \( \mu \in \mathbb{C} \).

1.16 Exercise

1. Show that

\[
(\mu I - J_m(\mu))^l \begin{cases} = 0 & l \geq m, \\
\neq 0 & 0 < l < m. \end{cases}
\]

These observations indicate that there may be a further decomposition of the eigen-structure of \( A \) into a more fundamental form. This is indeed true as we show in the following classic result.
Theorem 1.8 (Jordan Canonical Form) Any matrix $A \in \mathbb{C}^{n \times n}$ has a decomposition of the form

$$AX = XJ,$$

where $X \in \mathbb{C}^{n \times n}$ is nonsingular and

$$J = \begin{bmatrix}
J_{n_1}(\lambda_1) & & \\
& J_{n_2}(\lambda_2) & \\
& & \ddots \\
& & & J_{n_k}(\lambda_k)
\end{bmatrix},$$

with diagonal blocks as in (1.10).

Proof: The proof of this theorem begins with the Schur decomposition and through a sequence of similarity transformations, reduces $A$ to the desired form. The steps are:

1. Obtain the Schur decomposition $Q^H A Q = R$ with

$$R = \begin{bmatrix}
R_{1,1} & R_{1,2} & \cdots & R_{1,m} \\
0 & R_{2,2} & \vdots \\
\vdots & \ddots & \ddots \\
0 & 0 & \cdots & R_{m,m}
\end{bmatrix},$$

where $\sigma(A) = \{\lambda_j\}$ are the distinct eigenvalues of $A$ and where $R_{j,j} = \lambda_j I + U_j$ with $U_j$ a strictly upper triangular matrix.

2. Construct a nonsingular matrix $Z$ such that

$$Z^{-1} R Z = \begin{bmatrix}
R_{1,1} & 0 & \cdots & 0 \\
0 & R_{2,2} & \ddots \\
\vdots & \ddots & \ddots \\
0 & \cdots & \cdots & R_{m,m}
\end{bmatrix}.$$

3. Determine nonsingular matrices $Y_j$ such that

$$Y_j^{-1} R_{j,j} Y_j = \lambda_j I + J_j,$$

where

$$J_j = \begin{bmatrix}
J_{k_1}(0) & & \\
& J_{k_2}(0) & \\
& & \ddots \\
& & & J_{k_l}(0)
\end{bmatrix},$$

such that $k_1 \geq k_2 \geq \cdots \geq k_l.$
1 INTRODUCTION

4. This sequence results in

\[(QZY)^{-1}A(QZY) = \begin{bmatrix}
    \lambda_1 I + J_1 \\
    \lambda_2 I + J_2 \\
    \vdots \\
    \lambda_m I + J_m
\end{bmatrix},\]

where

\[Y = \begin{bmatrix}
    Y_1 & 0 & \cdots & 0 \\
    0 & Y_2 & \vdots \\
    \vdots & \ddots & \vdots \\
    0 & \cdots & \cdots & Y_m
\end{bmatrix}.

5. The stated result (JCF) follows with \(X = QZY\) and \(J\) equaling the matrix on the right hand side of 4 since \(J_k(\mu) = \mu I + J_k(0)\).

Steps (1) and (2) have been established with the previous discussion. To accomplish step (3) we need to prove the following.

**Lemma 1.9** If \(U \in \mathbb{C}^{n \times n}\) is strictly upper triangular, then there is a nonsingular \(Y\) such that

\[Y^{-1}UY = \begin{bmatrix}
    J_{k_1}(0) \\
    \vdots \\
    J_{k_m}(0)
\end{bmatrix},\]

such that \(k_1 \geq k_2 \geq \cdots \geq k_m\).

**Proof:** The proof is by induction on \(n\). The lemma clearly holds for \(n = 1\). Assume it is true for strictly upper triangular matrices of order less than \(n\).

Since \(U\) is strictly upper triangular, there is a least positive integer \(k_1\) such that \(U^{k_1} = 0\). Let \(y\) be a vector such that \(U^{k_1-1}y \neq 0\). Put

\[Y_1 = [U^{k_1-1}y, U^{k_1-2}y, \ldots, Uy, y],\]

and observe that

\[UY_1 = Y_1 J_{k_1},\]

and that the columns of \(Y_1\) are linearly independent (due to the strict upper triangular form of \(U\)). Now, complete the basis, i.e., construct \(Q_2\) such that \([Y_1, Q_2]\) is nonsingular and

\[U[Y_1, Q_2] = [Y_1, Q_2] \begin{bmatrix}
    J_{k_1} & U_{12} \\
    0 & U_{22}
\end{bmatrix},\]

with \(U_{22}\) strictly upper triangular. By the induction hypothesis, there is a nonsingular \(\hat{Y}_2\) such that

\[U_{22}\hat{Y}_2 = \hat{Y}_2 \hat{J}\]
where $\hat{J}$ is the direct sum of Jordan matrices of decreasing order. Now, we have

$$U[Y_1, Y_2] = [Y_1, Y_2] \begin{bmatrix} J_{k_1} & \hat{U} \\ 0 & \hat{j} \end{bmatrix},$$

where $Y_2 = Q_2 \hat{Y}_2$ and $\hat{U} = U_{12} \hat{Y}_2$. It is straightforward to verify that

$$U^{\ell+1}[Y_1, Y_2] = [Y_1, Y_2] \begin{bmatrix} J_{k_1} \hat{U} \hat{j} & \sum_{j=0}^{\ell} J_{k_1}^{\ell-j} \hat{U} \hat{j}^j \\ 0 & \hat{j}^{\ell+1} \end{bmatrix},$$

for any positive integer $\ell$ and hence that

$$\sum_{j=0}^{\ell} J_{k_1}^{\ell-j} \hat{U} \hat{j}^j = 0 \text{ and } \hat{j}^{\ell+1} = 0$$

when $\ell = k_1 - 1$. We intend to show that the Sylvester’s equation

$$J_{k_1} W = U + W \hat{j}$$

has a solution $W$ even though it is singular. In fact, we claim that

$$W \equiv \sum_{j=0}^{\ell} S^{\ell-j} \hat{U} \hat{j}^j$$

is a solution where $S \equiv J_{k_1} + e_k e_1^T$ is the right circular shift operator. We shall verify this essentially by direct substitution.

First, we observe that $e_1^T W = 0$, since

$$e_1^T W = e_1^T \sum_{j=0}^{\ell} S^{\ell-j} \hat{U} \hat{j}^j = e_1^T \sum_{j=0}^{\ell} J_{k_1}^{\ell-j} \hat{U} \hat{j}^j = 0,$$

when $\ell = k_1 - 1$. Thus

$$J_{k_1} W = (J_{k_1} + e_k e_1^T) W = SW.$$

Then we compute (assuming $\ell = k_1 - 1$)

$$U + W \hat{j} = U + \sum_{j=0}^{\ell} S^{\ell-j} \hat{U} \hat{j}^{j+1}$$

$$= S^{\ell+1} U + S \sum_{j=0}^{\ell} S^{\ell-j-1} \hat{U} \hat{j}^{j+1}$$

$$= S[S^{\ell} U + \sum_{j=0}^{\ell} S^{\ell-j-1} \hat{U} \hat{j}^{j+1}]$$
\[ S[S'^TU + \sum_{j=1}^{l+1} S'^{l-j+1} U\hat{J}^j] \]
\[ = S[S'^TU + \sum_{j=1}^{l} S'^{l-j} U\hat{J}^j] \]
\[ = SW, \]

where the second to last equality follows from the fact \( \hat{J}^{l+1} = 0 \) and we have used \( S'^{l+1} = I \).

This solution to the Sylvester equation implies
\[
\begin{bmatrix} J_{k_1} & 0 \\ 0 & \hat{J} \end{bmatrix} \begin{bmatrix} I & W \\ 0 & I \end{bmatrix} = \begin{bmatrix} I & W \\ 0 & I \end{bmatrix} \begin{bmatrix} J_{k_1} & \hat{U} \\ 0 & \hat{J} \end{bmatrix},
\]

and hence that
\[ UY = YJ \]

where \( J = \text{diag}(J_{k_1}, \hat{J}) \) and \( Y = [Y_1, Y_2 - Y_1W] \). This completes the induction and concludes the proof.

1.7 Exercise

1. Explain why the matrix factors in the Jordan Canonical Form are not continuous functions of the elements of the original matrix \( A \). What are the implications of this with respect to the development of a numerical algorithm to compute the Jordan Canonical Form?

2. The proof of the Jordan Canonical Form presented here is constructive. At what point is the discontinuity mentioned in the previous exercise apparent in the proof.

Lemma 1.6 uses a special case of the following more general result.

**Theorem 1.10** (Sylvester’s Equation) Suppose \( A \in \mathbb{C}^{n \times n}, B \in \mathbb{C}^{k \times k}, \) and \( C \in \mathbb{C}^{n \times k} \). Then the equation

\[ AX - XB = C, \]

has a unique solution \( X \in \mathbb{C}^{n \times k} \) if and only if

\[ \sigma(A) \cap \sigma(B) = \emptyset. \]

**Proof:** Let

\[ A = QR_AQ^H \text{ and } B = UR_BU^H, \]

KTH Lecture Notes, Draft, Spring 1999, D. C. Sorensen
be the Schur decompositions of $A$ and $B$ respectively. Note that $Q^H Q = I_n$ and $U^H U = I_k$. Set $X = Q S U^H$, with $S \in \mathbb{C}^{n \times k}$ to be determined. Then

$$AX - XB = Q R A S U^H - Q S R_B U^H,$$

and (1.11) is satisfied if and only if

$$R_A S - S R_B = \hat{C},$$

where $\hat{C} = Q^H C U$. The last row, $s_n^H = e_n^T S$ may be determined from

$$e_n^T R_A S - e_n^T S R_B = e_n^T \hat{C},$$

giving the equation

$$\rho_{n,n} e_n^T S - s_n^H R_B = \hat{c}_n^H,$$

i.e.

$$s_n^H (\rho_{n,n} I - R_B) = \hat{c}_n^H.$$

This equation has a unique solution if and only if $\rho_{n,n}$ is not an element of $\sigma(B)$. Moreover, if we partition

$$R_A = \begin{bmatrix} \hat{R}_A & r \\ 0 & \rho_{n,n} \end{bmatrix}, \quad S = \begin{bmatrix} \hat{S} \\ s_n^H \end{bmatrix}, \quad \hat{C} = \begin{bmatrix} \tilde{S} \\ \hat{c}_n^H \end{bmatrix},$$

we obtain the equation

$$\hat{R}_A \hat{S} - \hat{S} R_B = \hat{C} - r s_n^H.$$

The last row of $\hat{S}$ may now be determined using the same mechanism. Repeating this process all the rows of $S$ may be uniquely determined if and only if $\{\rho_{j,j}\}_{j=1}^n \cap \sigma(B) = \emptyset$, i.e. $\sigma(A) \cap \sigma(B) = \emptyset$.

1.18 Exercise

1. In the case $\mu = 1$ in the proof of lemma 1.9 we constructed a solution $M$ to

$$e_1 u_2^H + M J_2 = \hat{J}_1 M,$$

even though $\sigma(J_2) \cap \sigma(\hat{J}_1) = \{0\} \neq \emptyset$. Does this contradict Sylvester’s theorem? Why not?

2. Let $A \in \mathbb{C}^{m \times k}$ and $B \in \mathbb{C}^{n \times l}$. Define the tensor product $A \otimes B$ to be

$$A \otimes B = (\alpha_{i,j} B),$$

where $A \otimes B \in \mathbb{C}^{mn \times kl}$, i.e. a block matrix with the $i,j$-th block equal to $\alpha_{i,j} B$. If $X \in \mathbb{C}^{n \times k}$ and we define

$$\bar{X} = \begin{bmatrix} X e_1 \\ X e_2 \\ \vdots \\ X e_k \end{bmatrix} \quad \text{and} \quad \hat{C} = \begin{bmatrix} C e_1 \\ C e_2 \\ \vdots \\ C e_k \end{bmatrix},$$

KTH Lecture Notes, Draft, Spring 1999, D. C. Sorensen
show that
\[ DX - XE = C \]
if and only if
\[ (I_k \otimes D)\hat{X} - (E^T \otimes I_n)\hat{X} = \hat{C}, \]
that is,
\[ (I_k \otimes D - E^T \otimes I_n)\hat{X} = \hat{C}, \]
where \( D \in \mathbb{C}^{n \times n} \) and \( E \in \mathbb{C}^{k \times k} \).

3. If \( AC \) and \( BC \) are defined then \( (A \otimes B)(C \otimes D) = AC \otimes BD \).

4. If \( AB \) and \( AC \) are defined then \( A \otimes (B + C) = A \otimes B + A \otimes C \).

5. If \( AB \) is defined and \( A \) and \( B \) are nonsingular then \( (A \otimes B)^{-1} = A^{-1} \otimes B^{-1} \).

6. Use the Schur decompositions \( AQ = QR_A \) and \( BU = QR_B \) and the tensor product theory developed in the previous exercises to prove Theorem (1.10).

7. Suppose that
\[ AU = U\Lambda \text{ and } BV = V\Omega \]
where \( V^H V = I_n = U^H U \) and both \( \Lambda \) and \( \Omega \) are diagonal matrices containing the eigenvalues of \( A \) and \( B \) respectively. Show that
\[ (A \otimes B)(U \otimes V) = (U \otimes V)(\Lambda \otimes \Omega). \]

Conclude that \( (A \otimes B)q_{ij} = q_{ij}\lambda_j \omega_j \) where \( q_{ij} = U e_i \otimes V e_j \).

8. Consider the eigenvalue problem,
\[ -\frac{d^2}{d\xi^2}u(\xi) = \nu u(\xi), \]
with boundary conditions \( u(0) = 0 = u(1) \). The eigenvalue problem is to determine the non-zero functions \( u(\xi) \) and the corresponding \( \nu \). Using finite difference approximations on the uniform partition,
\[ 0 = \xi_0 < \xi_1 < \cdots < \xi_n < \xi_{n+1} = 1, \]
we obtain the following finite dimensional system of order \( n \),
\[ -\frac{1}{h^2}T_n u_k = \lambda_k u_k. \]

The eigenvector \( u_k = [\mu_{1,k}, \cdots, \mu_{n,k}]^T \) represents the approximations \( \mu_{i,k} \) to the eigenfunction \( u_k(\xi_i) \) where \( \xi_{i+1} = \xi_i + h \) for \( i, k = 1 : n \) and \( h = \frac{1}{n+1} \).
The coefficient matrix $T_n$ is the 1-D Laplacian which has the following tridiagonal structure:

$$
T_n = \begin{bmatrix}
2 & -1 & 0 & \cdots & 0 \\
-1 & 2 & -1 & \ddots & \vdots \\
0 & \ddots & \ddots & \ddots & 0 \\
\vdots & & -1 & 2 & -1 \\
0 & \cdots & 0 & -1 & 2
\end{bmatrix}.
$$

Use matlab to graph the eigenvectors. Why do they look like $\sin(k\pi\theta)$ for $0 \leq \theta \leq 1$? The 2-D Laplacian is

$$-\Delta u = -\left(\frac{\partial}{\partial \xi} + \frac{\partial}{\partial \eta}\right)u,$$

with boundary condition $u = 0$ on the unit square. If as above we take equally spaced points $\xi_i$ and $\eta_j$ on the unit square show that the discrete 2-D Laplacian is of the form

$$L_n = \begin{bmatrix}
T_n + 2I_n & -I_n & 0 & \cdots & 0 \\
-I_n & T_n + 2I_n & -I_n & \ddots & \vdots \\
0 & \ddots & \ddots & \ddots & 0 \\
\vdots & & -I_n & T_n + 2I_n & -I_n \\
0 & \cdots & 0 & -I_n & T_n + 2I_n
\end{bmatrix}.$$

Also show that $L_n = T_n \otimes I + I \otimes T_n$. What is eigensystem of $L_n$? Use matlab to graph the eigenvectors as 2-D surfaces over the unit square.

## 2 Basic Algorithms

Our main goal is to develop a class of rather sophisticated methods based upon projecting the large matrix onto a low dimensional subspace and extracting approximate eigenpairs from the eigensystem of the relatively small projected matrix. To understand and develop these methods it will be necessary to have more than a passing familiarity with the basic methods for approximating eigenpairs. We need to review the methods available for small dense problems as well as the simple iterations for large sparse problems.

### 2.1 Single Vector Power Iterations

Probably the oldest algorithm for approximating eigenvalues and corresponding eigenvectors of a matrix is the power method. This method is an important tool in its own right when conditions are appropriate. However, it is of fundamental
importance to the development, understanding, and convergence analysis of all of the iterative methods we shall derive here. The basic idea of the power method is extremely simple as indicated with the following illustrative example.

**Example:** Let \( A = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix} \). Then the vector \( e_2 = (0, 1)^T \) is an eigenvector for \( A \) corresponding to the eigenvalue of largest modulus. Suppose we begin with a vector \( v = (\nu_1, 1)^T \) where \( 0 < \nu_1 < 1 \). Repeatedly the following steps are executed

\[
(i) \quad w \leftarrow Av, \quad (ii) \quad v \leftarrow w / (e_2^T w).
\]

It is readily seen that if \( v_k \) denotes the vector \( v \) after \( k \) of these steps then

\[
v_k = A^k v / (e_2^T A^k v) = (2^{-k} \nu_1, 1)^T
\]

so that

\[
\|e_2 - v_k\|_\infty = 2^{-k} \nu_1 < 2^{-k} \|e_2\|_\infty.
\]

Thus the iterates \( \{v_j\} \) converge *linearly* to the eigenvector corresponding to the eigenvalue of largest magnitude with a convergence factor

\[
\frac{1}{2} = \frac{|\lambda_2|}{|\lambda_1|}
\]

The following figure illustrates this discussion. The starting vector is \( v_1^T = (1, 1/2)^T \). The lines labled 1, 2, \ldots, 6 show the directions of the successive vector iterates after scaling to be on the unit ball in the infinity-norm.

This simple example generalizes to arbitrary matrices and for Hermitian matrices this analysis remains virtually unaltered. We express this as

**Algorithm 2.1 (Power Method)**

1. Put \( v = v_0 / \|v_0\|_\infty \) (with \( v_0 \) arbitrary);
2. while ( ),
   (a) \( w \leftarrow Av \);
   (b) \( \lambda = \frac{w^T w}{v^T v} \);
   (c) \( j = \text{imax} (w) \);
   (d) \( v \leftarrow w / (e_j^T w) \);
In order to analyze the convergence of this method, note that the contents of $v$ after $k$-steps of this iteration will be the vector

$$v_k = \left( \frac{1}{e_j^T A^k v_0} \right) A^k v_0 = \left( \frac{\rho_k}{e_j^T A^k v_0} \right) \left( \frac{1}{\rho_k} A^k v_0 \right)$$

for any nonzero scalar $\rho_k$. In particular, this iteration may be analyzed as if the vectors had been scaled by $\rho_k = \lambda_1^k$ at each step, with $\lambda_1$ an eigenvalue of $A$ with largest magnitude. If $A$ is diagonalizable with eigenpairs $\{(x_j, \lambda_j), 1 \leq j \leq n\}$ and $v_0$ has the expansion $v_0 = \sum_{j=1}^n x_j \gamma_j$ in this basis then

$$\frac{1}{\lambda_1^k} A^k v_0 = \frac{1}{\lambda_1^k} \sum_{j=1}^n A^k x_j \gamma_j = \sum_{j=1}^n x_j (\lambda_j/\lambda_1)^k \gamma_j.$$  

If $\lambda_1$ is a simple eigenvalue then

$$\left( \frac{\lambda_j}{\lambda_1} \right)^k \rightarrow 0, \hspace{1cm} 2 \leq j \leq n.$$  

It follows that $v_k \rightarrow x_1/(e_1^T x_1)$ where $j = \arg \max (x_1)$ with a linear rate of $|\lambda_1|$.  

While the power method is useful, it has two obvious drawbacks. Convergence may be arbitrarily slow or may not happen at all, and only one eigenvalue and corresponding vector can be found. The basic iteration may be modified to overcome these difficulties. The most fundamental modification is to employ a spectral transformation.
2.2 Spectral Transformations and the Inverse Power Method

Spectral transformations are matrix valued functions of the given matrix \( A \) which have a more favorable distribution of the spectrum with respect to the power method. These transformations are generally based upon the following:

Let \( A \in \mathbb{C}^{n \times n} \) have an eigenvalue \( \lambda \) with corresponding eigenvector \( x \).

1. Let \( p(\tau) = \gamma_0 + \gamma_1 \tau + \gamma_2 \tau^2 + \ldots + \gamma_k \tau^k \). Then \( p(\lambda) \) is an eigenvalue of the matrix \( p(A) = \gamma_0 I + \gamma_1 A + \gamma_2 A^2 + \ldots + \gamma_k A^k \) with corresponding eigenvector \( x \) (i.e. \( p(A)x = xp(\lambda) \)).

2. If \( r(\tau) = \frac{p(\tau)}{q(\tau)} \) where \( p \) and \( q \) are polynomials with \( q(A) \) nonsingular, define \( r(A) = [q(A)]^{-1}p(A) \). Then \( r(\lambda) \) is an eigenvalue of \( r(A) \) with corresponding eigenvector \( x \).

It is often possible to construct a polynomial or rational function \( \phi(\tau) \) such that

\[
|\phi(\lambda_i)| \gg |\phi(\lambda_j)| \quad \text{for} \quad 1 \leq j \leq n, \quad j \neq i,
\]

where \( \lambda_i \) is an eigenvalue of particular interest. This is called a spectral transformation since the eigenvectors of the transformed matrix \( \phi(A) \) remain the same, but the corresponding eigenvalues \( \lambda_j \) are transformed to \( \phi(\lambda_j) \). Applying the power method with \( \phi(A) \) in place of \( A \) will then produce the eigenvector \( q \equiv x_i \) corresponding to \( \lambda_i \) at a convergence rate of \( |\frac{\phi(\lambda)}{\phi(\lambda_i)}| < 1 \). Once the eigenvector has been found, the eigenvalue \( \lambda \equiv \lambda_i \) may be calculated directly from a Raleigh quotient \( \lambda = q^H Aq / q^H q \).

This simple observation leads to dramatic enhancement of the convergence of the power method. Polynomial transformations may be applied only using matrix-vector products. Rational transformations require the solution of linear systems with the transformed matrix as the coefficient matrix. The simplest rational transformation turns out to be very powerful and is almost exclusively used for this purpose. If \( \mu \notin \sigma(A) \) then \( A - \mu I \) is invertible and \( \sigma([A - \mu I]^{-1}) = \{1/(\lambda - \mu) : \lambda \in \sigma(A) \} \). Using this transformed matrix in place of \( A \) leads to the following algorithm.

**Algorithm 2.2 (Inverse Power Method)**

1. Put \( v = v_0 / \|v_0\|_\infty \) (with \( v_0 \) arbitrary);
2. Select a shift \( \mu \);
3. while ( ),
   (a) Solve \( (A - \mu I)w = v \);
   (b) \( \lambda = \mu + \frac{w^H v}{w^H w} \);
2  BASIC ALGORITHMS

(c)  \( j = \text{max} (w); \)

(d)  \( v \leftarrow w / (e^T w) ; \)

Observe that the formula for \( \lambda \) at step (3,b) is equivalent to forming \( \lambda = (w^H A w) / (w^H w) \) so an additional matrix vector product is not necessary to obtain the Raleigh quotient estimate. The analysis of convergence remains entirely in tact. This iteration converges linearly with the rate

\[
\frac{|\lambda_1 - \mu|}{|\lambda_2 - \mu|}
\]

where the eigenvalues of \( A \) have been re-indexed so that \( |\lambda_1 - \mu| < |\lambda_2 - \mu| \leq |\lambda_3 - \mu| \leq \ldots \leq |\lambda_n - \mu| \).

The following figure illustrates the effect of the shift-invert spectral transformation on a real symmetric matrix. The original eigenvalues are shown on the x-axis and the transformed eigenvalues are shown on the y-axis.

Although not nearly as powerful, polynomial transformations are also useful and can be effective. These are particularly effective when there is no option of factoring the shifted matrix with a sparse direct method. This may happen for instance when no ordering will provide a reasonable limit on fill-in, or when the matrix is not available explicitly.

The following figure illustrates a Chebychev polynomial transformation of the spectrum.
This result is encouraging but still leaves us wondering how to select the shift \( \mu \). In many applications the choice is apparent from the requirements of the problem.

The normalization of the eigenvectors at Steps (3.c, 3.d) is for the most part irrelevant so long as it keeps the vectors uniformly bounded. The scaling here keeps the vector iterates on the unit ball in the \( \infty \) - norm. The Raleigh quotient approximate eigenvalues are independent of the scaling used. However, the vectors are not. Moreover, the convergence of the vectors may be destroyed (in a mild way) when a different scaling is used as the following exercise illustrates.

### 2.3 Exercises

Consider the following Matlab implementation of the Power Method:

```matlab
function [v, lam] = power(A);
    [n, n] = size(A);
    v = rand(n, 1) + sqrt(-1)*rand(n, 1);
    v = v/norm(v);
    lam1 = 0; lam = 1; v1 = zeros(n, 1); k = 1;

    while(k < 1000 & (STOPPING RULE ) ),
        v1 = v;
        lam1 = lam;
        w = A*v;
```

KTH Lecture Notes, Draft, Spring 1999, D. C. Sorensen
lam = v'*w;
  v = w/norm(w);
  k = k+1;
end

Consider two possibilities for STOPPING RULE:

a) abs(lam1-lam) > .000001*abs(lam)

b) norm(v1 - v) > .000001

These Stopping Rules are “satisfied” when the opposite inequality holds (hence the while-loop is exited).

Suppose $A$ is an $n \times n$ matrix such that $A = QAQ^H$ where $Q$ is unitary and $\Lambda$ is a diagonal matrix with complex diagonal elements $\lambda_j$ which satisfy $|\lambda_1| = 10|\lambda_2| > 0$ and $|\lambda_j| \geq |\lambda_{j+1}|$, $j = 2, 3, ..., n - 1$.

1. What special properties does $A$ have? Is $A$ defective or non-defective?

2. If $\lambda_1$ is a positive real number how many iterations will the above code take (in general) to satisfy stopping rule (b)?

3. If $\lambda_1 = \rho e^{i\theta}$ is complex will Stopping Rule (b) ever be satisfied? Explain. (Assume $\rho > 0$ and $\theta$ is not $0$ or $\pi$.)

4. Show that Stopping Rule (a) will (almost) always be satisfied for such a matrix $A$. Hint: Let $v_0$ be the initial vector in the iteration. Express $v_0 = q_1\gamma_1 + \hat{Q}g$ where $Q = (q_1, \hat{Q})$. What does $v_0^HA^mv_0$ look like?

5. How many iterations will it take to satisfy Stopping Rule (a)? What condition on $v_0$ could possibly make the iteration fail to to satisfy Stopping Rule (a)? Do you expect this to happen? Why?

Close examination of the inverse power method suggests an obvious improvement that will assure very impressive convergence rates. Since an estimate of an eigenvalue is available at each iteration why not use this to update the shift selection instead of keeping the shift fixed? The resulting iteration is:

**Algorithm 2.3 (Raleigh Quotient Iteration)**

1. Put $v = v_0/\|v_0\|$ (with $v_0$ arbitrary);

2. Put $\mu = v^HAv$;
3. while ( ),

(a) Solve \((A - \mu I)w = v;\)
(b) \(\mu \leftarrow \mu + \frac{w^H v}{w^H w};\)
(c) \(v \leftarrow w/(w^H v);\)

Again the normalization at step (c) does not effect the convergence of the eigenvalue estimates. However, it does effect the convergence properties of the vectors. As in the other iterations this choice will allow the vectors to converge to a particular eigenvector while normalizing by the 2-norm may prevent convergence of the vectors.

Generally, this method converges at a quadratic rate. However, it will converge at a cubic rate when the matrix is Hermitian. We shall first illustrate the general rate of convergence under the assumption that convergence occurs. Then we shall discuss the beautiful special case of Hermitian matrices in detail.

We denote the updated iterates by the subscript + so

\[
\mu = v^H Av, \quad (A - \mu I)v_+ = v\tau, \quad \mu_+ = \mu + \frac{w^H v}{w^H w},
\]

where \(\tau = 1/(w^H v)\) is the scale factor that causes \(v^H v_+ = 1\). Note that \(\mu_+ = \frac{v_+^H Av_+}{v_+^H v_+}\).

**Theorem 2.4 (RQI Convergence)** Define

\[
\rho \equiv \left\|(A - \mu I)v\right\|/\|v\|,
\]

\[
\rho_+ \equiv \left\|(A - \mu_+ I)v_+\right\|/\|v_+\|.
\]

If \(\mu \notin \sigma(A)\) then there is a positive number \(\gamma(\mu)\) that is uniformly bounded in the neighborhood of a simple eigenvalue of \(A\) such that

\[
\rho_+ \leq \gamma(\mu)\rho^m.
\]

with \(m = 3\) if \(A\) is Hermitian and \(m = 2\) otherwise.

**Proof:** Assume (w.l.o.g.) that \(\|v\| = 1\) and construct \(V = (v, \hat{V})\) to be unitary. Then \(V^H v = e_1\) and the scaling \(1 = v^H v_+ = v^H V V^H v_+ = e_1^T V^H v_+\) implies that \(V^H v_+\) is of the form

\[
\begin{pmatrix}
1 \\
p
\end{pmatrix} = V^H v_+
\]

Let

\[
y = V^H v, \quad y_+ = V^H v_+, \quad \begin{bmatrix}
\mu \\
C
\end{bmatrix} = V^H AV.
\]
Since \((A - \mu I)v_+ = v\tau\) implies \((V^H AV - \mu I)y_+ = y\tau\) we have
\[
\begin{bmatrix}
0 & h^H \\
g & C - \mu I
\end{bmatrix}
\begin{bmatrix}
1 \\
p
\end{bmatrix}
= \begin{bmatrix}
1 \\
0
\end{bmatrix} \tau.
\]
Hence
\[(2.5) \quad h^H p = \tau \text{ and } (C - \mu I)p = -g.\]
Moreover,
\[
(\mu_+ - \mu)v_+^H v_+ = v_+^H (A - \mu I)v_+ = v_+^H v\tau
\]
implies
\[(2.6) \quad (\mu_+ - \mu)(1 + \|p\|^2) = \tau.\]
Since \(V\) is unitary, we have
\[(2.7) \quad \rho = \|(V^H AV - \mu I)y\|/\|y\| = \|(\mu - 1) - \begin{bmatrix}
1 \\
0
\end{bmatrix}\mu\| = \|g\| = \|\rho\|.
\]
and
\[(2.8) \quad (\mu_+ - \mu)(1 + \|p\|^2) = \tau.
\]
Similarly, for \(\rho_+\) we have
\[
\rho_+ = \|(V^H AV - \mu_+ I)y_+\|/\|y_+\| = \|(V^H AV - \mu I)y_+ - (\mu_+ - \mu)y_+\|/\|y_+\| = \|y\tau - (\mu_+ - \mu)y_+\|/\|y_+\|.
\]
Substituting for \(y, y_+\) and \(\tau\) using (2.4, 2.5) will give
\[
\rho_+ = \left\| \begin{bmatrix}
1 \\
0
\end{bmatrix} \right\| (\mu_+ - \mu)(1 + \|p\|^2) - \left\| \begin{bmatrix}
1 \\
p
\end{bmatrix} \right\| (\mu_+ - \mu)(1 + \|p\|^2)\right\|_{1/2}
= |\mu_+ - \mu| \left(\frac{\|p\|^2}{1 + \|p\|^2}\right) \right\|/\|p\|^2\right\|_{1/2}
= |\mu_+ - \mu| \left(\frac{\|p\|^2}{1 + \|p\|^2}\right) \right\|/\|p\|^2\right\|_{1/2}
= |\mu_+ - \mu| \left(\frac{\|p\|^2}{1 + \|p\|^2}\right)
\leq \frac{\|h\|\|p\|^2}{(1 + \|p\|^2)}.
\]
Substituting \(p = -(C - \mu I)^{-1}g\) and \(\|g\| = \rho\) gives
\[
\rho_+ \leq \|h\|\|(C - \mu I)^{-1}g\|^2/(1 + \|(C - \mu I)^{-1}g\|^2)
\leq \left\| \left\| \frac{h}{\rho} \right\| \left\| (C - \mu I) \right\| ^2 (1 + \left\| (C - \mu I)^{-1} g \right\|^2) \right\|^2
\leq \gamma_1(\mu)\rho^2.
\]
Finally note that $A = A^H$ would imply
\[ |h^H p| = |g^H (C - \mu I)^{-1} g| \leq \|(C - \mu I)^{-1}\| \|g\|^2 \]
so that
\[
\begin{align*}
\rho_+ &\leq \frac{|h^H p| \|p\|}{\|(C - \mu I)\|^2/(1 + \|(C - \mu I)^{-1} g\|^2)} \rho^3 \\
(2.9) &\leq \gamma_2(\mu) \rho^3.
\end{align*}
\]

The desired result is established with $\gamma(\mu)$ the maximum of $\gamma_1$ and $\gamma_2$.

This result does not fully analyze the convergence properties of RQI. However, it does establish that if an iteration obtained RQI is convergent to a simple eigenvalue of $A$ then the residuals decrease at a cubic rate. The assumption that the limit of $\{\mu_j\}$ is a simple eigenvalue of $A$ is required to assert the uniform boundedness of the $\{\gamma(\mu_j)\}$

### 2.4 RQI on Hermitian Matrices - Gaps and Convergence

A great deal more about convergence may be said in the Hermitian case. The series of facts leading to these results will prove to be of extreme importance in the design of stopping rules for our algorithms. Therefore, we shall develop them here in some detail.

An important tool in the analysis is the mini-max Theorem

**Theorem 2.5 (Courant-Fischer)** Let $A \in \mathbb{C}^{n \times n}$ be Hermitian with eigenvalues $\lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_n$. Then

\[
\lambda_j = \min_{\dim \mathcal{S} = j} \left\{ \max_{x \in \mathcal{S}} \left\{ \frac{x^H A x}{x^H x} \right\} \right\}
\]

It will be useful to define certain quantities that will be used throughout the remainder of these notes. We define the Raleigh Quotient for a nonzero vector $v$ to be

\[
\mu = \mu(v) = \frac{v^H A v}{v^H v}
\]

We define the (Raleigh Quotient) RQ-residual to be

\[
r(v) = (I - \frac{vv^H}{v^H v}) A v = A v - v \mu.
\]

Observe that by definition $r(v)$ is the orthogonal projection of $A v$ into the orthogonal compliment of $\text{Span}(v)$ and thus $r(v)$ satisfies

\[
\|r(v)\| \leq \|A v - v \theta\| \quad \text{for all } \theta \in \mathbb{C}
\]
The Minimax Theorem implies that every Raleigh quotient $\mu$ is in the interval $[\lambda_1, \lambda_n]$. For a given Raleigh quotient $\mu$ we define

$$\lambda(\mu) \equiv \arg\min \{ |\lambda - \mu| : \lambda \in \sigma(A) \}$$

to be the eigenvalue of $A$ that is closest to $\mu$ and we define the gap between $\mu$ and the remainder of the spectrum to be

$$G(\mu) \equiv \min \{ |\theta - \mu| : \theta \in \sigma(A)/\{\lambda(\mu)\} \}.$$ 

These definitions may be made without the assumption that $A$ is Hermitian. However, the RQ-residual possesses an extremely important properties with respect to estimating the distance of a Raleigh Quotient to an eigenvalue of the matrix.

**Lemma 2.6** Let $A \in \mathbb{C}^{n \times n}$ be Hermitian, $v \in \mathbb{C}^n$ have unit length, $\mu = \mu(v)$. Let $\lambda = \lambda(\mu)$ and let

$$v = q\gamma + z\sigma$$

be the unique decomposition of $v$ into orthogonal components with $q \in S_\lambda, z \in S_\lambda^\perp$, and $\| z \| = \| q \| = 1$. Then

$$|\sigma| \leq \frac{\| r(v) \|}{G(\mu)} \quad \text{and} \quad |\lambda - \mu| \leq \frac{\| r(v) \|^2}{G(\mu)}$$

**Proof:** Note that

$$r(v) = (A - \mu I)v = q(\lambda - \mu)\gamma + (A - \mu I)z\sigma.$$ 

Since $A$ is Hermitian, $q^H(A - \mu I)z = (\lambda - \mu)q^Hz = 0$ and thus

$$(2.12) \quad \| r(v) \|^2 = (\lambda - \mu)^2\gamma^2 + \| (A - \mu I)z \|^2\sigma^2 \geq \| (A - \mu I)z \|^2\sigma^2.$$ 

The Mini-max Theorem implies $z^H(A - \mu I)^2z \geq G(\mu)^2$ so that

$$|\sigma| \leq \frac{\| r(v) \|}{G(\mu)}.$$ 

If $\lambda = \mu$ there is nothing more to prove. Otherwise, $\lambda \neq \mu$ and it follows from

$$0 = v^H r(v) = (\lambda - \mu)\gamma^2 + z^H(A - \mu I)z\sigma^2$$

and the fact $\gamma^2 + \sigma^2 = 1$ that

$$0 = (\lambda - \mu) + [z^H(A - \mu I)z - (\lambda - \mu)]\sigma^2 = (\lambda - \mu) + z^H(A - \lambda I)z\sigma^2.$$ 

Hence

$$\sigma^2 = \frac{\mu - \lambda}{z^H(A - \lambda I)z} \quad \text{and} \quad \gamma^2 = \frac{z^H(A - \lambda I)z}{z^H(A - \lambda I)z}.$$ 

---

KTH Lecture Notes, Draft, Spring 1999, D. C. Sorensen
Substituting this into (2.12) gives

\[(2.13) \quad \|r(v)\|^2 = \frac{(\lambda - \mu)^2 z^H (A - \mu I) z + z^H (A - \mu I)^2 (\mu - \lambda) z}{z^H (A - \lambda I) z} \]

\[= |\lambda - \mu| \left| \frac{z^H (A - \lambda I) (A - \mu I) z}{z^H (A - \lambda I) z} \right| \]

The definition of \( \lambda \) as the closest eigenvalue to \( \mu \) may be used to conclude that the matrix \( (A - \lambda I)(A - \mu I) \) is positive semidefinite and that

\[z^H (A - \lambda I)(A - \mu I) z = \sum_{j=1}^{n} \xi_j^2 |\lambda_j - \lambda| |\lambda_j - \mu| \]

\[\geq \sum_{j=1}^{n} \xi_j^2 |\lambda_j - \lambda| G(\mu) \]

\[\geq |z^H (A - \lambda I) z| G(\mu), \]

where \( \xi_j \) is the component of \( z \) corresponding to \( \lambda_j \in \sigma(A) \) in the expansion of \( z \) in the eigenvector basis. Using this in (2.13) gives

\[|\lambda - \mu| \leq \frac{\|r(v)\|^2}{G(\mu)} \]

as claimed.

This result has important consequences for accessing the accuracy of approximate eigenvalues of Hermitian matrices. It will be used to construct stopping rules for the iterations we shall develop.

One final lemma is needed to establish the rate of convergence of RQI. It will establish the surprising fact that the sequence of residuals are monotone decreasing in norm.

**Lemma 2.7** Let \( A \in \mathbb{C}^{n \times n} \) be Hermitian, \( v \in \mathbb{C}^n \) have unit length, \( \mu = \mu(v) \). If

\[(A - \mu I)w = v \]

then

\[\frac{\|r(w)\|}{\|w\|} \leq \|r(v)\| \]

**Proof:** From the minimization property of \( r(w) \) and the fact that \( \|(A - \mu I)w\| = \|v\| = 1 \) it follows that

\[\|r(w)\| \leq \|Aw - w\mu\| \]

\[= \|(A - \mu I)w\| \cdot \|v\| \]

\[= \|v^H (A - \mu I)w\| \]

\[\leq \|(A - \mu I)v\| \cdot \|w\|. \]
2. BASIC ALGORITHMS

The second equality and the last inequality are both consequences of Cauchy-Schwarz and the assumption that $A$ is Hermitian and the result has been established.

If the RQI sequence of vectors $\{v_k\}$ have been normalized to have unit length then this lemma implies that

\[
\|r(v_{k+1})\| \leq \|r(v_k)\| \quad \text{for} \quad k = 1, 2, 3, \ldots
\]

regardless of the starting vector $v_1$. Hence the sequence $\{\|r(v_k)\|\}$ must converge as it is bounded below by zero. We shall first examine the usual case when this limit is actually equal to zero.

**Theorem 2.8** Let $A \in \mathbb{C}^{n \times n}$ be Hermitian, and suppose $\{v_k\}$ is an RQI sequence of vectors each normalized to have unit length. Let $\mu_k = \mu(v_k)$ be the associated Raleigh Quotients. If $r(v_k) \to 0$ then $\mu_k \to \lambda \in \sigma(A)$ and for any constant $\eta > 1$

\[
|\lambda - \mu_{k+1}| \leq |\lambda - \mu_k|^3 \frac{\eta}{G(\lambda)^2}
\]

for all $k$ sufficiently large.

**Proof:** Let $\lambda^{(k)} \equiv \lambda(\mu_k)$ and let

\[
v_k = q^{(k)} \gamma_k + z^{(k)} \sigma_k
\]

be the orthogonal decomposition of $v_k$ as in Lemma (2.6). That lemma implies

\[
|\sigma_k| \leq \frac{\|r(v_k)\|}{G(\mu_k)} \quad \text{and} \quad |\lambda^{(k)} - \mu_k| \leq \frac{\|r(v_k)\|^2}{G(\mu_k)}.
\]

Since $\{\mu_k\} \subset [\lambda_1, \lambda_n]$ this sequence has limit points and these two inequalities imply that $\sigma_{k_j} \to 0$ and $|\lambda^{(k_j)} - \mu_{k_j}| \to 0$ for each convergent subsequence. Thus each limit point of the sequence $\{\mu_k\}$ is an eigenvalue of $A$ and the positive numbers $G(\mu_k)$ are uniformly bounded away from zero for all $k$ sufficiently large.

Dropping subscripts for the moment, let $\mu = \mu(v), \lambda = \lambda(\mu), (A - \mu I)w = v$ and

\[
\mu_+ = \mu + \frac{v^H w}{w^H w}
\]

be successive RQI iterates. If $v = q\gamma + z\sigma$ then

\[
w = \frac{1}{\lambda - \mu} q \gamma + (A - \mu I)^{-1} z \sigma
\]

and

\[
|\mu_+ - \lambda| = |\mu - \lambda + \frac{1}{(\lambda - \mu)^3} \gamma^2 z^H (A - \mu I)^{-1} z \sigma^2|
\]

\[
= \frac{1}{|\lambda - \mu|^2} \gamma^2 z^H (A - \mu I)^{-1} z + \frac{(\mu - \lambda) z^H (A - \mu I)^{-2} z}{1 + (\frac{z^H}{\gamma^2})^2 (\lambda - \mu) z^H (A - \mu I)^{-2} z}
\]

\[
\leq (\lambda - \mu)^2 (\frac{z}{\gamma})^2 [1/G(\mu) + |\lambda - \mu|/G(\mu)^2]
\]
where the last inequality follows from the Minimax Theorem and the fact that $z$ is orthogonal to all vectors in $S_{\lambda}$.

Re-introducing subscripts, we have established that

$$|\mu_{k+1} - \lambda^{(k)}| \leq (\lambda^{(k)} - \mu_k)^2 \frac{\sigma_k^2}{\gamma_k^2} [1/G(\mu_k) + |\lambda^{(k)} - \mu_k|/G(\mu_k)^2]$$

From our initial discussion it follows that

$$(\lambda^{(k)} - \mu_k)^2 \frac{\sigma_k^2}{\gamma_k^2} [1/G(\mu_k) + |\lambda^{(k)} - \mu_k|/G(\mu_k)^2] \to 0 \text{ and } |\mu_{k+1} - \lambda^{(k+1)}| \to 0.$$ 

Thus

$$|\lambda^{(k+1)} - \lambda^{(k)}| \leq |\lambda^{(k+1)} - \mu_{k+1}| + |\mu_{k+1} - \lambda^{(k)}| \to 0.$$ 

Since each $\lambda^{(k)}$ is in the finite set $\sigma(A)$ we conclude that $\lambda^{(k)} = \lambda \in \sigma(A)$ for all $k$ sufficiently large. This establishes $Q$ super-quadric convergence of the sequence $\{\mu_k\}$ to some $\lambda \in \sigma(A)$.

Again from our proof of Lemma (2.6) we have

$$\sigma_k^2 = \frac{\mu_k - \lambda}{z^{(k)^H} (A - \lambda I) z^{(k)}}$$

but the Min-Max Theorem implies that

$$|z^{(k)^H} (A - \lambda I) z^{(k)}| \geq G(\lambda).$$

The final result then follows.

2.5 Generalized Power Methods

All of the methods discuss so far suffer the disadvantage of only finding one eigenvalue at a time. Even the analysis of the rapidly convergent Rayleigh Quotient Iteration exposes a dependency on the distance of the target eigenvalue to the remainder of the spectrum. This motivates the search for algorithms capable of computing more than one eigenvalue and associated vectors at a time.

An issue that has been avoided to this point is the problem associated with computing the complex eigenvalues of a real square matrix. The eigenvalues and associated eigenvectors occur in complex conjugate pairs. If sufficient care is taken in the organization of the algorithms then it should be possible to compute these without having to resort to complex arithmetic. It is possible to implement the algorithms below in real arithmetic as they stand so long as conjugate-transpose operations are replaced with transpose operations.

The first algorithm to consider is a straightforward generalization of the power method. The algorithm attempts to compute an invariant subspace rather than a single vector.

**Algorithm 2.9 (Subspace Iteration)**
2 BASIC ALGORITHMS

1. Factor $VR = W$ (with $W \in \mathbb{R}^{n \times k}$ arbitrary);

2. $H = 0$;

3. while $\|W - VH\| > tol\|H\|$,
   
   (a) $W \leftarrow AV$;
   
   (b) $H = V^H W$;
   
   (c) Factor $VR = qr(W)$;

When it converges, this iteration will have computed an invariant subspace of the real matrix $A$. Note that

$$H = V^H AV \text{ with } V^H V = I$$

so $H$ is a generalized Raleigh Quotient. Approximate eigenpairs for $A$ are of the form $(x, \theta)$ where $Hy = y\theta$ and $x = Vy$.

The analogy to the shifted inverse power method is a straightforward modification of subspace iteration. However, there is no direct analogy to Raleigh Quotient iteration when $k < n$ is the dimension of the subspace. The iteration becomes quite interesting when $k = n$.

**Hessenberg Form and QR Iteration:** We shall discuss reduction to Hessenberg form and the implicit Q-Theorem in the next chapter. For the moment suppose we have a method for constructing an orthogonal similarity transformation of $A$ to upper Hessenberg form. That is

$$AV = VH \text{ with } V^T V = I, \ H \text{ upper Hessenberg.}$$

If $H = QR$ is the Q-R factorization of $H$ then $W = (VQ)R$ is the Q-R factorization of $W = AV$. Moreover

$$A(VQ) = (VQ)(RQ)$$

$$AV_+ = V_+ H_+ \text{ with } V_+^T V_+ = I, \ H_+ = RQ.$$  

The amazing point here is that $H_+$ remains upper Hessenberg if $H$ is originally upper Hessenberg. Moreover, the Q-R factorization of $H$ by Givens’ method and the associated updating $V_+ = VQ$ requires $O(n^2)$ flops rather than $O(n^3)$ for a dense Q-R factorization. This leads to the following algorithm.

**Algorithm 2.10 (Basic QR Iteration)**

1. Reduce $A$ to Hessenberg form $AV = VH$;

2. while $\,$
   
   (a) $QR = qr(H)$;
2 BASIC ALGORITHMS

(b) \( V \leftarrow VQ; \)
(c) \( H \leftarrow RQ; \)

Let’s examine the behavior of the basic QR Iteration with respect to columns of the matrix \( V \). First note that

\[
AV = (VQ)R = V_+R \quad \text{implies} \quad Av_1 = v_1^+ \rho_{1,1}
\]

where \( \rho_{1,1} \) is the \((1,1)\) element of the upper triangular matrix \( R \) and \( v_1, v_1^+ \)
are the first columns of \( V \) and \( V_+ \) respectively. Thus the basic QR iteration is performing the power method on the first column of the matrix \( V \) as the iteration proceeds. Secondly, multiplying the relation \( AV = V_+R \) on the left by \( V_+^H \) and on the right by \( V^H \) and then taking the conjugate transpose gives the relation

\[
A^HV_+ = VR^H
\]

Equating the last column on both sides will give

\[
A^Hv_n^+ = v_n \rho_{n,n}
\]

So inverse iteration is being performed on the last column of \( V \) with respect to the matrix \( A^H \). The final improvement that might be made is to introduce a shift to enhance convergence. Note that

\[
v_n^HAv_n = \epsilon_n^THe_n
\]

is the Raleigh quotient associated with the last column of \( V \)

\textbf{Algorithm 2.11 (Explicitly Shifted QR Iteration)}

1. Reduce \( A \) to Hessenberg form \( AV = VH; \)
2. \( \mu = H(n, n) \)
3. \textbf{while} ( ),
   (a) \( QR = qr(H - \mu I); \)
   (b) \( V \leftarrow VQ; \)
   (c) \( H \leftarrow RQ + \mu I; \)

The analysis of basic QR iteration is essentially the same for this iteration. One can show as before that

\[
(A - \mu I)v_1 = v_1^+ \rho_{1,1} \quad \text{and} \quad (A - \mu I)^Hv_n^+ = v_n \rho_{n,n}
\]
Where $\rho_{1,1}, \rho_{n,n}$ are the first and last diagonal elements of the upper triangular matrix $R$. Thus at the $k$-th iteration the vector $v_1$ is of the form

$$v_1 = \frac{p_k(A)v_1^{(0)}}{\|p_k(A)v_1^{(0)}\|}$$

where $p_k(\tau) = \prod_{j=1}^{k}(\tau - \mu_j)$, $v_1^{(0)}$ is the initial first column of $V$ and $\{\mu_j\}$ are the $k$ shifts that have been applied. Observe that as $\mu_k \to \lambda \in \sigma(A)$ the polynomial $p_k(A)$ will annihilate all components of $v_1$ in the eigenspace $S_\lambda$. The consequences are quite interesting and will be developed in the next chapter.

It is also apparent that the iteration as defined above is precisely carrying out RQI on the last column of $V$. Our analysis of RQI applies directly to show that the $(n,n)$ element of $H$ should converge to $\lambda$ quadratically when $A$ is non-Hermitian and cubically when $A$ is Hermitian. This choice of shift is rarely used in practice because another selection due to Wilkinson has been found to perform better. This is to choose the eigenvalue of the bottom $2 \times 2$ submatrix of $H$ that is closest to the $(n,n)$ element of $H$.

There are a number of practical issues remaining. Three important issues are: computing in real arithmetic when $A$ is real, deflation of subdiagonal elements of $H$, and the implicit shift mechanism.

The issue of computing in real arithmetic is quite important in practice since storage will double and multiplication will be four times as expensive if complex arithmetic has to be used. Fortunately it can be avoided. Suppose $\mu_1$ and $\mu_2$ are two successive shifts. These successive iterations are as follows

$$Q_1R_1 = qr(H - \mu_1 I)$$
$$H_1 = R_1Q_1 + \mu_1 I$$
$$Q_2R_2 = qr(H_1 - \mu_2 I)$$
$$H_2 = R_2Q_2 + \mu_2 I$$

Note that

$$HQ_1 = (H - \mu_1 I)Q_1 + \mu_1 Q_1$$
$$= Q_1(R_1Q_1 + \mu_1 I)$$
$$= Q_1H_1$$

so $H_1$ is unitarily similar $H$ and a repeat of this argument shows that every $H_j$ throughout the iteration is similar to the original matrix $A$. In particular

$$H_2 = (Q_1Q_2)^H H(Q_1Q_2)$$
Now, observe that
\[
(H - \mu_2 I)(H - \mu_1 I) = (H - \mu_2 I)Q_1 R_1 \\
= Q_1 (H_1 - \mu_1 I)R_1 \\
= (Q_1 Q_2)(R_2 R_1)
\]
so the Q-R factorization of this quadratic polynomial of \(H\) will automatically give the matrix \((Q_1 Q_2)\) needed to perform the similarity transformation that would have been obtained through two successive steps of the QR iteration. Suppose \(\mu = \xi + i\eta\) is complex and that \(\mu_1 = \mu\) and \(\mu_2 = \bar{\mu}\). Then
\[
(H - \mu_2 I)(H - \mu_1 I) = (H - \xi I)^2 + \eta^2 I
\]
so two successive steps of the shifted QR iteration associated with the application of a complex shift followed by it’s conjugate may be performed in real arithmetic as follows.
\[
QR = qr((H - \xi I)^2 + \eta^2 I) \\
H \leftarrow Q^T HQ.
\]

Since this is mathematically equivalent to having done the two steps in succession it follows that the updated \(H\) remains upper Hessenberg as before. However, this would not necessarily hold true in finite precision. There is a very clever way to force the updated \(H\) to be upper Hessenberg and to apply these shifts without forming or factoring the matrices that appear above. The key result that allows this algorithm is the Implicit Q-Theorem. This deceptively simple result has a fundamental deep role in the development of eigenvalue algorithms. It shall be stated here without proof. A more general form will be developed in the following chapter.

**Theorem 2.12** (Implicit Q-Theorem) Let \(A \in \mathbb{C}^{n \times n}\) and suppose that there is a unitary similarity between \(A\) and a Hessenberg matrix \(H\)
\[
AQ = QH, \quad Q^H Q = I.
\]
If every subdiagonal element of \(H\) is positive (\(H\) is unreduced) then the elements of \(H\) and columns \(2 - n\) of \(Q\) are uniquely determined by the choice of the first column of \(Q\).

This theorem implies that any mechanism whatsoever that will produce a similarity transformation of an initial \(H\) to \(H_2\)
\[
HQ = QH_2
\]
with \(Q e_1 = (Q_1 Q_2) e_1\) will be mathematically equivalent to performing the two successive QR steps above. The key idea is to note that if \(A = QR\), the first
column of \( Q \) is just a scaling of the first column of \( A \). In the case of a single real shift \( \mu \) we will need to determine the first column of \( H - \mu I \)

\[
(H - \mu I)e_1 = 
\begin{pmatrix}
\gamma_{1,1} - \mu \\
\gamma_{2,1} \\
0 \\
\vdots \\
0
\end{pmatrix}.
\]

A Givens transformation \( G_{1,2} \) may be constructed to introduce a zero in the \((2,1)\) position of \((H - \mu I)e_1\) and this transformation has the same first column as the orthogonal matrix \( Q \) of the Q-R factorization \((H - \mu I) = QR\). Now the matrix \( H^{(1)} = G_{1,2}^T H G_{1,2} \) will have a nonzero introduced into the \((3,1)\) position. This may be zeroed out by a second Givens transformation introducing a nonzero in the \((4,2)\) position and so on until the “bulge” is chased off the end.

**Algorithm 2.13 (Bulge Chase - Single Shift)**

1. Construct a Givens transformation \( G_{1,2} \) such that \( G_{1,2}^T (H - \mu I)e_1 = e_1 \rho \);
2. \( H \leftarrow G_{1,2}^T H G_{1,2} \);
3. for \( j = 1, 2, \ldots, n - 2 \)
   (a) Construct a Givens transformation \( G_{j+1,j+2} \) to introduce a zero in the \((j+2, j)\) position of \( H \);
   (b) \( H \leftarrow G_{j+1,j+2}^T H G_{j+1,j+2} \);

The double shift case has a similar structure but a \(3 \times 3\) bulge instead of a \(2 \times 2\) bulge is chased in this case. In the case of a double complex shift \( \mu = \xi + i\eta \) followed by \( \bar{\mu} \) we will need to determine the first column of \((H - \xi I)^2 + \eta^2 I \)

\[
((H - \xi I)^2 + \eta^2 I)e_1 = 
\begin{pmatrix}
(\gamma_{1,1} - \xi)^2 + \gamma_{1,2}\gamma_{2,1} + \eta^2 \\
\gamma_{2,1}(\gamma_{1,1} - \xi) + (\gamma_{2,2} - \xi)\gamma_{2,1} \\
0 \\
\vdots \\
0
\end{pmatrix}.
\]

Two Givens transformations or a Householder transformation \( Q_1 \) may be constructed to introduce zeroes below the \((1,1)\) position of \((H - \mu I)(H - \bar{\mu} I)e_1\) and this transformation has the same first column as the orthogonal matrix \( Q \) of the Q-R factorization \((H - \mu I)(H - \bar{\mu} I) = QR\). Now the matrix \( H^{(1)} = Q_1^T H Q_1 \) will have a nonzero elements introduced into the \((3,1), (4,1)\) and \((4,2)\) positions. The entries in positions \((3,1), (4,1)\) may be zeroed out by a second Householder transformation introducing a nonzeros in the \((5,2)\) and \((5,3)\) positions and so on until the “bulge” is chased off the end as in the single shift case.
Algorithm 2.14 (Bulge Chase - Double Complex Shift)

1. Construct a Householder transformation $Q_1$ such that $Q_1^T(H - \mu I)(H - \tilde{\mu}I)e_1 = e_1p_1$;
2. $H \leftarrow Q_1^THQ_1$;
3. for $j = 1, 2, \ldots, n - 3$
   (a) Construct a Householder transformation $Q_{j+1}$ to introduce zeroes in the $(j + 2, j)$ and $(j + 3, j)$ positions of $H$;
   (b) $H \leftarrow Q_{j+1}^THQ_{j+1}$;
4. Construct a final Givens transformation $G_{n-1,n}$ to introduce a zero in the $(n, n-2)$ position of $H$;
5. $H \leftarrow G_{n-1,n}^THG_{n-1,n}$;

The final detail that must be added to make this work is the introduction of deflation. Usually during the course of this iteration, subdiagonal elements of $H$ become small early during the iteration. The hypothesis of the Implicit Q-Theorem is violated when a subdiagonal actually becomes zero. Numerical difficulties may occur when a subdiagonal is small. In this case, a deflation procedure must be enforced. Often the procedure amounts to declaring the $j$-th subdiagonal $\gamma_{j+1,j}$ to be zero when

$$|\gamma_{j+1,j}| < \epsilon \max(|\gamma_{j,j}|, |\gamma_{j+1,j+1}|)$$

is satisfied. When this is done, the problem has been split or deflated into two smaller problems. With the appropriate book-keeping, this provides the final direct enhancement to the process.

3 Krylov Subspaces and Projection Methods

We have seen that although the rate of convergence can be improved to an acceptable level through spectral transformations, power iterations are only able to find one eigenvector at a time. If more vectors are sought, then various deflation techniques (such as orthogonalizing against previously converged eigenvectors) and shift strategies must be introduced. The Implicitly Shifted QR Method is not appropriate in the large scale setting but we might well ask if there is something in between a single vector power iteration and the QR Method. An alternative that is related to both subspace iteration and the QR method may be found in a class of methods called Krylov subspace projection methods.

A natural motivation for this class of methods may be found upon examination of the behavior of the power sequence. Equation (2.1) hints that the
successive vectors produced by a power iteration may contain considerable information along eigenvector directions corresponding to eigenvalues near the one with largest magnitude. A single vector power iteration simply ignores this information. Projections methods offer a means to extract this additional information. Rather than discard the vectors produced during the power iteration, additional eigen-information is obtained by looking at various linear combinations of the sequence of vectors produced by the power iteration instead of discarding them. This immediately leads to a formal consideration of the Krylov subspace

\[ \mathcal{K}_k(A, v_1) = \text{Span} \{ v_1, Av_1, A^2v_1, \ldots, A^{k-1}v_1 \}. \]

It is reasonable to construct approximate eigenpairs from this subspace by imposing a Galerkin condition: A vector \( x \in \mathcal{K}_k(A, v_1) \) is called a Ritz vector with corresponding Ritz value \( \theta \) if the Galerkin condition

\[ \langle w, Ax - x\theta \rangle = 0, \text{ for all } w \in \mathcal{K}_k(A, v_1) \]

is satisfied. There are some immediate consequences of this definition: Let \( W \) be a matrix whose columns form an orthonormal basis for \( \mathcal{K}_k \equiv \mathcal{K}_k(A, v_1) \). Let \( \mathcal{P} = WW^H \) denote the related orthogonal projector onto \( \mathcal{K}_k \) and define \( \hat{A} \equiv \mathcal{P}A\mathcal{P} = BWB^H \) where \( B \equiv W^H BW \). It can be shown that

- \((x, \theta)\) is a Ritz-pair if and only if \( x = Wy \) with \( By = y\theta \).
- \( \|(I - \mathcal{P})AW\| = \|(A - \hat{A})W\| \leq \|(A - C)W\| \) for all \( C \) such that \( C\mathcal{K}_k \subset \mathcal{K}_k \).
- The Ritz-pairs \((x, \theta)\) and the minimum value \( \|(I - \mathcal{P})AW\| \) are independent of the choice of orthonormal basis \( W \).

Item(1) follows immediately from the Galerkin condition since it implies that

\[ 0 = W^H(AWy -Wy\theta) = By - y\theta. \]

Item(2) is easily shown using invariance of \( \| \cdot \| \) under unitary transformations. Item(3) follows from the fact that \( V \) is an orthonormal basis for \( \mathcal{K}_k \) if and only if \( V = WQ \) for some \( k \times k \) unitary matrix \( Q \). With this change of basis \( \hat{A} = VH^H \) where \( H = V^HAV = Q^HBQ \). Since \( H \) is unitarily similar to \( B \) so the Ritz-values remain the same and the Ritz-vectors are of the form \( x = Wy = V\hat{y} \) where \( \hat{y} = Q^Hy \).

These facts are actually valid for any \( k \) dimensional subspace \( S \) in place of \( \mathcal{K}_k \). The following properties are consequences of the fact that every \( w \in \mathcal{K}_k \) is of the form \( w = \phi(A)v_1 \) for some polynomial \( \phi \) of degree less than \( k \).

- If \( q \) is a polynomial of degree less than \( k \) then

\[ q(A)v_1 = q(\hat{A})v_1 = Wq(B)z_1 \]

where \( v_1 = Wz_1 \), and if degree of \( q \) is \( k \) then

\[ \mathcal{P}q(A)v_1 = q(\hat{A})v_1 \]
3 KRYLOV SUBSPACES AND PROJECTION METHODS

- If \( \hat{p}(\lambda) \equiv \text{det}(\lambda I - B) \) is the characteristic polynomial of \( B \) then \( \hat{p}(A) = 0 \)
and \( \|\hat{p}(A)v_1\| \leq \|q(A)v_1\| \) for all monic polynomials of degree \( k \).

- If \( y \) is any vector in \( \mathbb{C}^k \) then \( AWy - WB\gamma y = \gamma\hat{p}(A)v_1 \) for some scalar \( \gamma \).

- If \((x, \theta)\) is any Ritz-pair for \( A \) with respect to \( \mathcal{K}_k \) then

\[
Ax - x\theta = \gamma\hat{p}(A)v_1
\]

for some scalar \( \gamma \).

While these facts may seem esoteric, they have important algorithmic consequences. First, it should be noted that \( \mathcal{K}_k \) is an invariant subspace for \( A \) if and only if \( v_1 = V y \) where \( AV = VR \) with \( V^HV = I_k \) and \( R \) is \( k \times k \) upper triangular. Also, \( \mathcal{K}_k \) is an invariant subspace for \( A \) if \( v_1 = Xy \) where \( X \in \mathbb{C}^{n\times k} \) and \( AX = X\Lambda \) with \( \Lambda \) diagonal. This follows from items 2 and 3 since there is a \( k \) degree monic polynomial \( q \) such that \( q(R) = 0 \) and hence \( \|\hat{p}(A)v_1\| \leq \|q(A)v_1\| = \|Vq(R)y\| = 0 \) (A similar argument holds when \( v_1 = Xy \)).

Secondly, there is some algorithmic motivation to seek a convenient orthonormal basis \( V = WQ \) that will provide a means to successively construct these basis vectors. It is possible to construct a \( k \times k \) unitary \( Q \) using standard Householder transformations such that \( v_1 = V e_1 \) and \( H = Q^H BQ \) is upper Hessenberg with non-negative subdiagonal elements. It is possible to show using item (3) that in this basis,

\[
AV = VH + fe_k^T, \quad \text{where} \quad f = \gamma\hat{p}(A)v_1
\]

and \( V^H f = 0 \) follows from the Galerkin condition.

The first observation shows that if it is possible to obtain a \( v_1 \) as a linear combination of \( k \) eigenvectors of \( A \) then \( f = 0 \) and \( V \) is an orthonormal basis for an invariant subspace of \( A \) and that the Ritz values \( \sigma(H) \subset \sigma(A) \) and corresponding vectors are eigenpairs for \( A \). The second observation leads to the Lanczos/Arnoldi process. Our next task is to explore these observations in an algorithmic setting.

3.1 Reduction to Condensed Form

We have already discussed the advantages of reducing the matrix \( A \) to condensed form before performing QR iterations. There are a number of ways to accomplish these reductions. The most popular is the Householder reduction. The basic tool used is a Householder transformation. These transformations were initially developed to construct a Q-R factorization and to solve least squares problems. They are designed by analogy to elementary eliminators used in constructing an LU decomposition.

These Householder transformations or elementary reflectors as they are often called are usually developed in real arithmetic. The complex form is almost as straightforward but there are a few points that require care. It is easily verified
that a matrix of the form \( U = I - 2vv^H \) with \( \|v\| = 1 \) is unitary. Moreover, if \( a \) is a given nonzero vector, and \( Ua = e_1\beta \) then necessarily \( |\beta| = \|a\| \) and

\[
(I - 2vv^H)a = e_1\beta
\]

implies \((2v^Ha)v = a - e_1\beta\). Hence \( v \) must be of the form

\[
(a - e_1\beta)\tau \quad \text{with} \quad \tau = \frac{1}{\|(a - e_1\beta)\|}
\]

Now observe that

\[
\tau^2 = a^H a - a^H e_1\beta - \beta e_1^T a + |\beta|^2
\]

\[
= 2|\beta|(|\beta| - Re(a^H e_1\beta / |\beta|))
\]

If \( a^H e_1 = \sigma e^{i\phi} \) than a numerically stable choice for \( \beta \) is \( \beta = -\|a\|e^{-i\phi} \) and then

\[
\tau^2 = 2|\beta|(|\beta| + \sigma)
\]

\[
= 2\|a\|(|\|a\| + |e_1^T a||).
\]

Thus the final form of \( U \) is

\[
I - \theta(a - \beta e_1)(a - \beta e_1)^H
\]

with \( \beta = -\|a\|(e_1^T a / |e_1^T a|) \) and \( \theta = \frac{1}{\|a\| + |e_1^T a|} \). It is sometimes desirable to normalize so that the elementary unitary transformation will send the vector \( a \) to the vector \( e_1\|a\| \). To do this just replace \( U \) with \(-e^{i\phi}U\).

The reduction of a matrix \( A \) to upper Hessenberg form is accomplished through the application of a sequence of Householder transformations

\[
H = (U_1 U_2 \ldots U_{n-2})^H A (U_1 U_2 \ldots U_{n-2})
\]

where each \( U_j \) is of the form

\[
U_j = \begin{bmatrix} I_j & 0 \\ 0 & \hat{U}_j \end{bmatrix}
\]

where \( \hat{U}_j \) is a Householder transformation of order \( n - j \). If

\[
A_j = (U_1 U_2 \ldots U_j)^H A (U_1 U_2 \ldots U_j)
\]

\[
= \begin{bmatrix} H_j & M_j \\ a_j e_j^T & A_j \end{bmatrix}
\]

then \( \hat{U}_j \) is constructed so that \( \hat{U}_j a_j = e_1\beta_j \) and as described above it may be arranged so that \( \beta_j \geq 0 \) and we have

\[
A_{j+1} = U_j^H A U_j
\]

\[
= \begin{bmatrix} H_j & M_j \hat{U}_j \\ \beta_j e_1 e_j^T & \hat{U}_j^H \hat{A}_j \hat{U}_j \end{bmatrix}
\]
After the \( n - 2 \) transformations have been applied we have

\[ AV = VH \]

with \( V = U_1 U_2 \ldots U_{n-2} \) and \( H = A_{n-1} \). Finally, we observe in passing that \( Ve_1 = e_1 \) with this construction. If \( v_1 \) is any vector of unit length then we may construct a Householder transformation \( U_o \) such that \( U_o e_1 = v_1 \) and if we apply the procedure just described to \( \hat{A} = U_o^H A U_o \) instead of \( A \) we will construct

\[ AV = VH \quad \text{with} \quad V^H V = I, \quad Ve_1 = v_1, \]

where \( V = U_o \hat{U}_1 \hat{U}_2 \ldots \hat{U}_{n-2} \) and \( H \) is upper Hessenberg with non-negative subdiagonal elements.

Let us now examine the first \( k \) steps of this reduction process by considering the relationships on the left and right of (3.1) between the first \( k \) columns. We have

\[ AV_k = V_k H_k + v_{k+1} \beta_k e_k^T \]

where \( V_k \) is the matrix consisting of the first \( k \) columns of \( V \) and \( v_{k+1} \) is the \( k + 1 \)-st column of \( V \). Since \( V \) is unitary, we have

\[ V_k^H V_k = I_k, \quad V_k^H v_{k+1} = 0, \quad H_k = V_k^H A V_k^H. \]

If \( A = A^H \) then \( H = H^H \) and hence is tridiagonal. Moreover, if we have arranged for the subdiagonals \( \beta_i \geq 0 \) then \( H \) is real symmetric and tridiagonal with non-negative off-diagonal elements.

The amazing fact about these relationships is that if the subdiagonal elements of \( H_k \) are all positive this \( k \) \(-\) step factorization is uniquely determined by the starting vector \( v_1 \). This is the content of the Implicit Q-Theorem which shall be stated and proved now.

**Theorem 3.1 (Implicit Q-Theorem)** Suppose \( A \in \mathbb{C}^{n \times n} \) and that

\[ AV = VH + f e_k^T \]
\[ AQ = QG + g e_k^T \]

where \( Q, V \) have orthonormal columns and \( G, H \in \mathbb{C}^{k \times k} \) are both upper Hessenberg with positive subdiagonal elements.

If \( Q e_1 = V e_1 \) and \( Q^T g = V^T f = 0 \), then \( Q = V, \ G = H, \) and \( g = f \). Moreover, \( f = p(A) v_1 \) where \( p \) is a polynomial of degree at most \( k \).

**Proof:** There is a straightforward inductive proof (or see [16,p367]).
3.2 The Arnoldi Factorization

The proof of the Implicit Q-Theorem actually shows how to reduce a matrix to Hessenberg form in a columnwise fashion. This process is known as the Arnoldi process if $A$ is non-Hermitian and it is known as the Lanczos process if $A$ is Hermitian. The Lanczos process precedes Arnoldi which is a generalization to the non-Hermitian case.

Definition: If $A \in \mathbb{C}^{n \times n}$ then a relation of the form

$$AV = VH + f e_k^T$$

where $V \in \mathbb{C}^{n \times k}$ has orthonormal columns, $V^H f = 0$ and $H \in \mathbb{C}^{k \times k}$ is upper Hessenberg with non-negative subdiagonal elements is called a $k$-step Arnoldi Factorization of $A$. If $A$ is Hermitian then the relation is called a $k$-step Lanczos Factorization of $A$. The columns of $V$ are referred to as the Arnoldi vectors or Lanczos vectors respectively.

We have just seen that a k-step Arnoldi factorization may be viewed as a truncated reduction of an $n \times n$ matrix $A$ to upper Hessenberg form. An alternative way to write this factorization is

$$AV = (V, v) \begin{pmatrix} H \\ \beta e_k^T \end{pmatrix} \quad \text{where} \quad \beta = \|f\| \quad \text{and} \quad v = \frac{1}{\beta} f. \quad (3.1)$$

At this point the factorization may be used to obtain approximate solutions to a linear system $Ax = b$ if $b = v_1 \beta_1$ and this will be discussed in detail later. Our purpose here is to investigate the use of this factorization to obtain approximate eigenvalues and eigenvectors. These are readily available through the factorization. It should be apparent that when $\|f\|$ is small then the columns of $V$ nearly form a basis for an invariant subspace of $A$ and we may use essentially the same ideas to get approximate eigenpairs from the small matrix $H$.

If $Hy = y\theta$ then the vector $x = Vy$ satisfies

$$\|Ax - x\theta\| = \|(AV - VH)y\| = |\beta e_k^T y|.$$

Definition: If an approximate eigenpair $(x, \theta)$ has been obtained from an Arnoldi (Lanczos) factorization then the vector $x$ a is called a Ritz vector and the approximate eigenvalue $\theta$ a is called a Ritz value the number $|\beta e_k^T y|$ is called the Ritz estimate for this approximate eigenpair.

Observe that if $(x, \theta)$ is a Ritz pair then

$$\theta = y^H H y = (Vy)^H A (Vy) = x^H A x$$

is a Raleigh Quotient (assuming $\|y\| = 1$) and the associated RQ-residual is $r(x) = Ax - x\theta$ satisfies

$$\|r(x)\| = |\beta e_k^T y|.$$
This provides rigorous bounds on the quality of the eigenvalues of $H$ as approximations to eigenvalues of $A$ when $A$ is Hermitian, but only on the RQ residual otherwise. However, if $f = 0$ these the Ritz pairs become exact eigenpairs of $A$.

This factorization may be advanced one step at the cost of a (sparse) matrix-vector product involving $A$ and two dense matrix vector products involving $V^T$ and $V$. The explicit steps are:

1. $\beta = \| f \|; \; v \leftarrow f / \beta$;
2. $V \leftarrow (V, v); \; H \leftarrow \begin{pmatrix} H \\ \beta e_k^T \end{pmatrix}$.
3. $z \leftarrow Av$;
4. $h \leftarrow V^T z; \; f \leftarrow z - V h$;
5. $H \leftarrow (H, h)$;

The dense matrix-vector products may be accomplished using level 2 BLAS. In exact arithmetic, the columns of $V$ form an orthonormal basis for the Krylov subspace and $H$ is the orthogonal projection of $A$ onto this space. In finite precision arithmetic, care must be taken to assure that the computed vectors are orthogonal to working precision. The method developed in [?] may be used to accomplish this.

The following figure illustrates the Arnoldi factorization.

The information obtained through this process is completely determined by the choice of the starting vector. Eigen-information of interest may not appear until $k$ gets very large. In this case it becomes intractable to maintain numerical orthogonality of the basis vectors $V$ and it also will require extensive storage. Failure to maintain orthogonality leads to a number of numerical difficulties.

In a certain sense, computation (or approximation) of the projection indicated at Step (3.3.5) has been the main source of research activity in this topic.
The computational difficulty stems from the fact that \( \| f \| = 0 \) if and only if the columns of \( V \) span an invariant subspace of \( A \). When \( V \) "nearly" spans such a subspace \( \| r \| \) will be small. Typically, in this situation, a loss of significant digits will take place at Step (3.3.5) through numerical cancellation unless special care is taken. On the one hand, it is a delightful situation when \( \| f \| \) becomes small because this indicates that the eigenvalues of \( H \) are accurate approximations to the eigenvalues of \( A \). On the other hand, this "convergence" will indicate a probable loss of numerical orthogonality in \( V \). The identification of this phenomenon in the symmetric case and the first rigorous numerical treatment is due to Paige\[22,23]\. There have been several approaches to overcome this problem in the symmetric case. They include: (1) complete re-orthogonalization which may be accomplished through maintaining \( V \) in product Householder form \([15, 34]\) or through the Modified Gram-Schmidt processes with re-orthogonalization \([9, 26]\). (2) Selective re-orthogonalization which has been proposed by Parlett and has been heavily researched by him and his students. Most notably, the thesis and subsequent papers and computer codes of Scott have developed this idea \([24, 25, 31]\). (3) No re-orthogonalization which has been developed by Cullum and her colleagues. This last option introduces the almost certain possibility of introducing spurious eigenvalues. Various techniques have been developed to detect and deal with the presence of spurious eigenvalues \([7, 8]\).

The appearance of spurious eigenvalues may be avoided through complete re-orthogonalization of the Arnoldi (or Lanczos) vectors. Computational cost has been cited as the reason for not employing this option. However, the cost will be reasonable if one is able to fix \( k \) at a modest size and then update the starting vector \( v_1 = Ve_1 \) while repeatedly doing \( k \)-Arnoldi steps. This approach has been explored to some extent in \([2, 28]\). In Cullum \([6]\) has explored the symmetric case.

### 3.3 Structure of the Arnoldi Factorization

In the next chapter we develop an analogy to the Implicitly Shifted QR-Iteration. The algorithm we develop here may be regarded as a truncation of that iteration. We take the point of view here that \( f = f(v_1) \) is a function of the starting vector and then attempt to iteratively update \( v_1 \) in order to force the residual vector \( f(v_1) \) to zero. In order to make sense of this it will be necessary to understand when \( f \) is indeed a function of \( v_1 \) and also to determine its functional form and structure and characterize the zeros of this function.

Since each column of \( V \) is obtained through a normalization of a vector of the form \( p(A)v_1 \), all of these vectors are members of the Krylov subspace

\[
\mathcal{K}_k(A, v_1) = \text{Span} \{ v_1, Av_1, A^2v_1, \ldots, A^{k-1}v_1 \}.
\]

This space plays a central role along with the Krylov matrix

\[
K = (v_1, Av_1, \ldots, A^{k-1}v_1).
\]
3 KRYLOV SUBSPACES AND PROJECTION METHODS

We wish to study the structure of these vectors. An alternate derivation of the Arnoldi process is to consider the companion (or Frobenius) matrix

\[
F \equiv \begin{pmatrix}
0 & \gamma_0 \\
I & \hat{g}
\end{pmatrix} = \begin{pmatrix}
0 & \gamma_0 \\
1 & \gamma_1 \\
& \ddots & \ddots \\
& & 1 & \gamma_{k-1}
\end{pmatrix}
\]

and to observe that

(3.3) \[ AK - KF = \hat{r} e_k^T. \]

where \( \hat{r} = A^k v_1 - Kg \) with \( g^T = (\gamma_0, \hat{g}^T) \). Note that \( \hat{r} = \hat{p}(A)v_1 \) where \( \hat{p}(\lambda) = \lambda^k + \sum_{j=0}^{k-1} \gamma_j \lambda^j \), and also that \( \hat{p}(\lambda) \) is the characteristic polynomial of \( F \). If \( g \) is chosen to solve \( \min_{p \in \mathcal{P}_k(A)} \| A^k v_1 - Kg \|_2 \) then \( \hat{r} \) is orthogonal to all vectors in \( \mathcal{K}_k(A, v_1) \). Moreover, \( \hat{p} \) solves \( \min_{p \in \mathcal{P}_k(A)} \{ \| p(A) v_1 \| \} \) where \( \mathcal{P}_k \) is the set of all monic polynomials of degree \( k \).

To solve the minimization problem in (3.5), one would factor \( K = QR \) where \( Q \) is orthogonal, \( R \) is upper triangular. Note that \( R \) is nonsingular if and only if \( K \) has linearly independent columns and that \( Q \) may be constructed so that \( \rho_{jj} = e_j^T Re_j > 0 \). One then solves

\[ g = R^{-1} Q^T A^k v_1. \]

This choice of \( g \) will minimize the residual and also will assure that \( 0 = Q^T \hat{r} \). Multiplying (3.5) on the right by \( R^{-1} \) gives

\[ A(QR^{-1}) - (QR^{-1})RF R^{-1} = \hat{r} e_k^T R^{-1}, \]

i.e.

(3.4) \[ AQ - QG = \hat{f} e_k^T \]

where \( Q = KR^{-1}, G = RF R^{-1} \) is upper Hessenberg with the same characteristic polynomial as \( F \), and \( \hat{f} = \frac{1}{\rho_{kk}} \hat{r} \). It is easily verified that \( v_1 = Qe_1 = V e_1 \), and \( 0 = Q^T \hat{f} \). Thus, the Implicit Q-Theorem will imply that \( Q = V, G = H, \) and \( \hat{f} = f \). Putting \( H = G \) yields

\[ \beta_j = e_{j+1}^T H e_j = e_{j+1}^T RF R^{-1} e_j = \frac{\rho_{j+1,j+1}}{\rho_{jj}}. \]

Moreover,

\[ \frac{1}{\rho_{jj}} \| \hat{p}_j(A) v_1 \| = \beta_j = \frac{\rho_{j+1,j+1}}{\rho_{jj}}. \]

gives

\[ \rho_{j+1,j+1} = \| \hat{r}_j \| = \| \hat{p}_j(A) v_1 \|. \]

This discussion establishes the following.
3 KRYLOV SUBSPACES AND PROJECTION METHODS

Theorem 3.4 Let $AV_j = V_j H_j + f_j e_j^T$ be a sequence of successive Arnoldi steps $1 \leq j \leq k$ and suppose that $\dim(K_k(A, v_1)) = k$. Then

$$f_j = \frac{1}{\|\hat{p}_{j-1}(A)v_1\|} \hat{p}_j(A)v_1,$$

$$\beta_j = \frac{\|\hat{p}_j(A)v_1\|}{\|\hat{p}_{j-1}(A)v_1\|},$$

where $\hat{p}_j(\lambda)$ is the characteristic polynomial of $H_j$. Moreover,

$$\hat{p}_j \text{ solves } \min_{p \in \mathbb{P}_m} \{ \|p(A)v_1\| \}$$

for $1 \leq j \leq k$.

The development leading to Theorem (3.7) follows and builds upon the development by Ruhe in [27]. The fact that $\|\hat{p}_k(A)v_1\|$ (the characteristic polynomial of $H_k$ acting on $v_1$) will minimize $\|p(A)v_1\|$ over all monic polynomials of degree $k$ was proved by Saad in [29]. Theorem (3.7) points out that that the structure of $\beta_j$ is a somewhat complicated function of $A$ and $v_1$.

The final result of this section will develop necessary and sufficient conditions for a particular starting vector to generate a $k$-dimensional invariant subspace.

Theorem 3.5 Let $AV_k - V_k H_k = f_k e_k^T$ be a $k$-step Arnoldi factorization of $A$, with $H$ unreduced (i.e. $f_j \neq 0$, $1 \leq j \leq k - 1$). Then $f_k = 0$ if and only if $v_1 = Xy$ where $AX = XJ$ with $\text{rank}(X) = k$ and $J$ a Jordan matrix of order $k$ (i.e. the direct sum of Jordan blocks).

Proof: If $r_k = 0$, let $H \hat{X} = \hat{X}J$ be the Jordan canonical form of $H$ and put $X = V_k \hat{X}$. Then $AX = XJ$, $\text{rank}(X) = k$ and

$$v_1 = V_k e_1 = V_k \hat{X} \hat{X}^{-1} e_1 = Xy, \text{ with } y = \hat{X}^{-1} e_1.$$

Suppose now that $AX = XJ$, $\text{rank}(X) = k$, and $v_1 = Xy$. Then $A^m X = X J^m$ for any non-negative integer $m$ and it follows that

$$A^m v_1 = A^m X y = X J^m y \in \text{Range}(X)$$

for all $m$. Hence, $\dim K_{k+1}(A, v_1) = \text{rank}(X) = k$. Now, $H$ unreduced implies $\dim K_j(A, v_1) = j$ for $1 \leq j \leq k - 1$ and it follows from Theorem (3.7) that $f_k = 0$.

A similar result may be formulated in terms of Schur vectors instead of generalized eigenvectors. This result will be stated without its proof which is very similar to the proof of the previous result.

Theorem 3.6 Let $AV_k - V_k H_k = f_k e_k^T$ be a $k$-step Arnoldi factorization of $A$, with $H$ unreduced. Then $f_k = 0$ if and only if $v_1 = Qy$ where $AQ = QR$ with $Q^H Q = I_k$ and $R$ upper triangular of order $k$. 

KTH Lecture Notes, Draft, Spring 1999, D. C. Sorensen
Theorem (3.8) provides the motivation for the algorithms we shall develop. It suggests that one might find an invariant subspace by iteratively replacing the starting vector with a linear combination of approximate eigenvectors corresponding to eigenvalues of interest. Such approximations are readily available through the Arnoldi factorization. This theorem also indicates that it will be impossible to force the residual to zero if \( v_1 \) has a component of a generator of a cyclic subspace of dimension greater than \( k \). Theorem (3.9) indicates that our computations can be carried out within the framework of a truncated Schur decomposition and this leads to the development of an implicit restart method that is analogous to the implicitly shifted QR iteration.

3.4 Restarting the Arnoldi Process

A clear difficulty with the Lanczos/Arnoldi process is that the number of steps required to calculate eigenvalues of interest within a specified accuracy cannot be pre-determined. This depends completely on the starting vector \( v_1 \), and generally eigen-information of interest does not appear until \( k \) gets very large. In Figure 1 the distribution in the complex plane of the Ritz values (shown in blue dots) is compared with the spectrum (shown as black \( + \)'s). The original matrix is a normally distributed random matrix of order 200 and the Ritz values are from a \( k = 50 \) step Arnoldi factorization. Note that hardly any Ritz values appear in the interior and also that very few are well approximated. Eigenvalues at the extremes of the spectrum of \( A \) are clearly better approximated than the others.

![Figure 1: Typical Distribution of Ritz Values](image)

Unfortunately, one cannot know in advance how many steps will be required before eigenvalues of interest are well approximated by Ritz values. This is particu-
ularly true when the problem has a wide range of eigenvalues but the eigenvalues of interest are clustered. In computational chemistry, problems are usually symmetric and positive definite. There is a wide range of eigenvalues varying over many orders of magnitude (8 - 10 orders are common). Only the smallest ones are physically interesting and they are clustered at the low end of the spectrum. Shift and invert is not an option. Without this spectral transformation, many Lanczos steps are required to obtain the smallest eigenvalues. In order to recover eigenvectors, one is obliged to store all of the Lanczos basis vectors (usually on a peripheral device) and to solve very large tridiagonal eigenvalue subproblems at each step. In the Arnoldi process that is used in the non-Hermitian case, not only do the basis vectors have to be stored and orthogonalized, the Hessenberg eigenvalue subproblem must also be solved at the \( k \)-th step.

For large problems, it is clearly intractable to compute and store a numerically orthogonal basis set \( V_k \) for large \( k \). Storage requirements are \( \mathcal{O}(n \cdot k) \) and arithmetic costs are \( \mathcal{O}(n \cdot k^2) \) flops to compute the basis vectors plus \( \mathcal{O}(k^3) \) flops to compute the eigensystem of \( H_k \).

To control this cost, restarting schemes have been developed that iteratively replace the starting vector \( v_1 \) with an “improved” starting vector \( v_1^+ \) and then compute new Arnoldi factorization of fixed length \( k \). Beyond the obvious motivation to control computational cost and storage overhead, there is a clear interest in forcing \( f_k = 0 \). However, this is useful only if the spectrum \( \sigma(H_k) \) has the desired properties. The structure of \( f_k \) guides the strategy. The goal is to iteratively force \( v_1 \) to be a linear combination of eigenvectors of interest.

Theorems 3.5 and 3.6 give the most general conditions for \( f_k \) to vanish, but the following Lemma is perhaps more intuitive. Since \( v_1 \) determines the subspace \( \mathcal{K}_k \), this vector must be constructed to select the eigenvalues of interest. The following lemmas serve as a guide.

**Lemma 3.7** If \( v = \sum_{j=1}^{k} q_j \gamma_j \) where \( Aq_j = q_j \lambda_j \) and

\[
AV = VH + f e_k^T
\]

is a \( k \)-step Arnoldi factorization with unreduced \( H \) then \( f = 0 \) and \( \sigma(H) = \{ \lambda_1, \lambda_2, \ldots, \lambda_k \} \).

This Lemma follows easily from the observation \( \phi(A)v_1 = 0 \) with \( \phi(\tau) = \prod_{i=1}^{k} (\tau - \lambda_j) \) together with the minimization property ?? which implies \( f_k = 0 \).

Restarting was initially proposed by Karush [?] soon after the Lanczos algorithm appeared [?]. Subsequently, there were developments by Paige [?], Cullum and Donath [?], and Golub and Underwood [?]. Then, Saad [?] developed a polynomial restarting scheme for eigenvalue computation based upon the acceleration scheme of Manteuffel [?] for the iterative solution of linear systems.
3.5 Polynomial Restarting

Polynomial restarting strategies replace $v_1$ by

$$v_1 \leftarrow \psi(A)v_1,$$

where $\psi$ is a polynomial constructed to damp unwanted components from the starting vector. The idea is to force the starting vector to be closer and closer to an invariant subspace. This is motivated by Lemmas 3.7 and 3.8. An iteration is defined by a repeatedly restarting until the updated Arnoldi factorization eventually contains the desired information. For more information on the selection of effective restarting vectors, see [?]. One of the more successful approaches is to use Chebyshev polynomials in order to damp unwanted eigenvector components in the available subspace. Explicit restarting techniques are easily parallelized, in contrast to the overhead involved in implicit restarting (next section). The reason is that major part of the work is in matrix vector products. In the situation that one has to solve the eigenproblem on a massively parallel computer for a matrix that involves cheap matrix vector products, this may be an attractive property.

Two possibilities for constructing $\psi$ suggest themselves immediately. One is to construct the polynomial to be “small” in magnitude on the unwanted set of eigenvalues and large on the wanted set. This criteria can be met by constructing a polynomial that best approximates 0 on a specified set that encloses the unwanted set and excludes the wanted set of eigenvalues. The other possibility is to use the best approximation to eigenvectors in hand. Select the current Ritz vectors corresponding to Ritz values that best approximate the wanted eigenvalues and form

$$v_+ = \sum_{j=1}^{k} \hat{q}_j \gamma_j,$$

Since each Ritz vector is of the form $\hat{q}_j = \phi_j(A)v$ where $\phi_j$ is a polynomial of degree $m$, this mechanism is also a polynomial restart. Saad gives some heuristics for choosing the weights $\gamma_j$.

A third way is to specify the polynomial $\psi$ by its roots. A fairly obvious choice is to find the eigenvalues $\theta_j$ of the projected matrix $H$ and sort these into two sets according to a given criteria: The wanted set $\Omega_w = \{\theta_j : j = 1, 2, \ldots, k\}$ and the unwanted set $\Omega_u = \{\theta_j : j = k + 1, k + 2, \ldots, k + p\}$. Then one specifies the polynomial $\psi$ as the polynomial with these unwanted Ritz values as its roots. This choice of roots is called Exact Shifts was suggested in [?].

Morgan [?] found a remarkable property of this strategy. If exact shifts are used to define $\psi(\tau) = \prod_{j=k+1}^{k+p}(\tau - \theta_j)$, then the Krylov space generated by $v_+^1 = \psi(A)v_1$ satisfies

$$K_m(A, v_1^+), = \text{Span}\{\hat{q}_1, \hat{q}_2, \ldots, \hat{q}_k, A\hat{q}_j, A^2\hat{q}_j, \ldots, A^p\hat{q}_j\},$$

for any $j = 1, 2, \ldots, k$. Thus polynomial restarting with exact shifts will generate a new subspace that contains all of the possible choices in 3.5.
This property follows from the fact that $\mathcal{K}_m(A, v_1^+ \psi(A)) = \psi(A)\mathcal{K}_m(A, v_1)$ together with the fact that a Ritz vector $\hat{q}_j$ has the form

$$\hat{q}_j = \prod_{i=1, i \neq j}^k (A - \theta_i I)\psi(A)v_1$$

and thus

$$A^\ell \hat{q}_j = A^\ell \prod_{i=1, i \neq j}^k (A - \theta_i I)v_1^+ \in \mathcal{K}_m(A, v_1^+), \quad \text{for } \ell = 1, 2, \ldots, p.$$ 

Hence

$$\text{Span}\{\hat{q}_1, \hat{q}_2, \ldots, \hat{q}_k, A\hat{q}_j, A^2\hat{q}_j, \ldots, A^p\hat{q}_j\} \subset \mathcal{K}_m(A, v_1^+).$$

A minimal polynomial argument may then be used to establish the linear independence of $\{\hat{q}_1, \hat{q}_2, \ldots, \hat{q}_k, A\hat{q}_j, A^2\hat{q}_j, \ldots, A^p\hat{q}_j\}$. When wanted Ritz values are not distinct, generalized eigenvectors enter in.

Exact shifts have proven to perform remarkably well in practice and have been adopted as the shift selection of choice when no other information is available.

### 3.6 Explicit Restarting

Polynomial restarting was proposed by Manteuffel in the case of linear systems and this idea was modified for eigenvalues by Saad. Again, the idea is to apply a polynomial to the starting vector that is constructed to damp unwanted components from the eigenvector expansion. The resulting algorithm takes the form:

**Repeat while()**

1. Compute an $m$-step Arnoldi factorization

$$AV_m = V_m H_m + f_m e_m^T,$$

with $V_m e_1 = v_1$;

2. Compute $\sigma(H)$ and corresponding Ritz estimates and halt if desired eigenvalues are well approximated.

3. Construct a polynomial $\psi$ based upon $\sigma(H)$ to damp unwanted components.

4. $v_1 \leftarrow \psi(A)v_1; \quad v_1 \leftarrow v_1/\text{norm}(v_1)$;
This is called explicit restarting when the polynomial is applied directly to the starting vector resulting in a new starting vector used to initialize a new Arnoldi factorization. The construction of the polynomial at Step 4 may be guided by a-priori information about the spectrum of $A$ or solely by information gleaned from $\sigma(H)$. A typical scheme is to sort the spectrum of $H$ into two disjoint sets $\Omega_w$ and $\Omega_u$ with $\sigma(H) = \Omega_w \cup \Omega_u$. The Ritz values in the set $\Omega_w$ are to be regarded as approximations to the “wanted” eigenvalues of $A$ and an open convex set $\mathcal{C}_u$ containing $\Omega_u$ with $\Omega_w \cap \mathcal{C}_u = \emptyset$ is to be regarded as a region that approximately encloses the “unwanted” portion of the spectrum of $A$. The polynomial $\psi$ is then constructed to be as small as possible in $\mathcal{C}_u$ when normalized, for example, to take the value 1 at an element of $\Omega_w$ closest to $\partial \mathcal{C}_u$. Chebyshev polynomials are appropriate when $\mathcal{C}_u$ is taken to be an ellipse and this was the original proposal of Saad when he adapted the Manteuffel idea to eigenvalue calculations. Another possibility explored by Saad has been to take $\mathcal{C}_u$ to be the convex hull of $\Omega_u$ and to construct the polynomial $\psi$ that best approximates 0 on this set in the least squares sense. Both of these are based upon well known theory of polynomial approximation. The problem of constructing an optimal ellipse for this problem has been studied by Chatelin and Ho. The reader is referred to \[ for details of constructing these polynomials.

The reasoning behind this type of algorithm is based upon the fact that if $v_1$ is a linear combination of precisely $k$ eigenvectors of $A$ then Arnoldi factorizations terminates in $k$ steps with an orthonormal basis for the invariant subspace spanned by those vectors. The Ritz values $\sigma(H_k)$ will be the corresponding eigenvalues of $A$. The update of the starting vector $v_1$ is designed to enhance the components of this vector in the directions of the wanted eigenvectors and damp its components in the unwanted directions. This effect is achieved at Step 4 since

$$v_1 = \sum_{j=1}^{n} x_j \gamma_j \rightarrow \psi(A)v_1 = \sum_{j=1}^{n} x_j \psi(\lambda_j) \gamma_j.$$ 

If the same polynomial were applied time, then after $M$ iterations, the $j$-th original expansion coefficient would be essentially attenuated by a factor

$$\left(\frac{\psi(\lambda_j)}{\psi(\lambda_1)}\right)^M$$

where the eigenvalues have been ordered according decreasing values $|\psi(\lambda_j)|$. The eigenvalues inside the region $\mathcal{C}_u$ become less and less significant as the iteration proceeds. Hence, the wanted eigenvalues are approximated increasingly well as the iteration proceeds.

The technique just described is referred to as explicit (polynomial) restarting or as the Arnoldi-Chebyshev method. The cost in terms of matrix-vector products $w \leftarrow Av$ is $M \ast (m + \deg(\psi))$ for $M$ major iterations. The cost of the arithmetic in the Arnoldi factorization is $M \ast (2n \ast m^2 + O(m^3))$ Flops (floating point operations). Tradeoffs must be made in terms of cost of the Arnoldi factorization vs. cost of the matrix-vector products $Av$ and also in terms of storage $(nm + O(m^2))$. 

---

KTH Lecture Notes, Draft, Spring 1999, D. C. Sorensen
3 KRYLOV SUBSPACES AND PROJECTION METHODS

3.7 Implicit Restarting

There is another approach to restarting that offers a more efficient and numerically stable formulation. This approach called implicit restarting is a technique for combining the implicitly shifted QR mechanism with a k-step Arnoldi or Lanczos factorization to obtain a truncated form of the implicitly shifted QR-iteration. The numerical difficulties and storage problems normally associated with Arnoldi and Lanczos processes are avoided. The algorithm is capable of computing a few (k) eigenvalues with user specified features such as largest real part or largest magnitude using $2nk + O(k^2)$ storage. No auxiliary storage is required. The computed Schur basis vectors for the desired k-dimensional eigen-space are numerically orthogonal to working precision. This method is well suited to the development of mathematical software and this will be discussed in Section ??.

Implicit restarting provides a means to extract interesting information from very large Krylov subspaces while avoiding the storage and numerical difficulties associated with the standard approach. It does this by continually compressing the interesting information into a fixed size k dimensional subspace. This is accomplished through the implicitly shifted QR mechanism. The basic mechanism is most easily understood in the case of applying a single shift $mu$. Assume that we have an $m$-step Arnoldi factorization

$$AV = VH + f_m e_m^T$$

where we have suppressed the subscripts on the $n \times m$ orthogonal matrix $V$ and on the $m \times m$ upper Hessenberg $H$. An analogy of the explicitly shifted QR algorithm may be applied to this truncated factorization of $A$. It consists of the following four steps. Let $\mu$ be a shift and let $(H - \mu I) = QR$ with $Q$ orthogonal and $R$ upper triangular. Then

\begin{align*}
(4.1.1) & \quad (A - \mu I)V - V(H - \mu I) = f_m e_m^T \\
(4.1.2) & \quad (A - \mu I)V - VQR = f_m e_m^T \\
(4.1.3) & \quad (A - \mu I)(VQ) - (VQ)(RQ) = f_{k+p} c_{k+p}^T Q \\
(4.1.4) & \quad A(VQ) - (VQ)(RQ + \mu I) = f_m e_m^T Q
\end{align*}

Let $V_+ = VQ$ and $H_+ = RQ + \mu I$. Then $H_+$ is upper Hessenberg and applying the matrices in (4.1.2) to the vector $e_1$ to expose the relationship of their first columns gives

$$(A - \mu I)v_1 = v_1^+ \rho_{11}$$

where $\rho_{11} = e_1^T Re_1, v_1^+ = V_+ e_1$.

This idea may be extended for up to $p$ shifts being applied successively. The development will continue using the implicit shift strategy.

An Arnoldi factorization of length $m = k + p$ is compressed to a factorization of length $k$ by applying $p$ implicit shifts resulting in

$$AV_m^+ = V_m^+ H_m^+ + f_m e_m^T Q,$$
where \( V_m^+ = V_m Q \), \( H_m^+ = Q^T H_m Q \), and \( Q = Q_1 Q_2 \cdots Q_p \), with \( Q_j \) the orthogonal matrix associated with the shift \( \mu_j \). It may be shown that the first \( k - 1 \) entries of the vector \( e_m^T Q \) are zero. Equating the first \( k \) columns on both sides yields an updated \( k \)-step Arnoldi factorization

\[
(3.7) \quad AV_k^+ = V_k^+ H_k^+ + f_k^+ e_k^T,
\]

with an updated residual of the form \( f_k^+ = V_{k+p}^+ e_k^{+1} \beta_k + f_{k+p} \sigma \). Using this as a starting point it is possible to use \( p \) additional steps of the Arnoldi process to return to the original \( m \)-step form. Each of these applications implicitly applies a polynomial in \( A \) of degree \( p \) to the starting vector. The roots of this polynomial are the shifts used in the \( QR \) process and these may be selected to filter unwanted information from the starting vector and hence from the Arnoldi factorization.

The basic iteration becomes

Given an \( m \)-step Arnoldi factorization

\[
AV = VH + f e_m^T,
\]

**Repeat while()**

1. Select a set of \( p < m \) shifts \( \mu_1, \mu_2, \ldots, \mu_p \) based upon \( \sigma(H) \) or perhaps other information.
2. \( q^T \leftarrow e_m^T \);
3. For \( j = 1, 2, \ldots, p, \)
   
   (a) Factor \( QR = H - \mu_j I \);
   
   (b) \( H \leftarrow Q^H H Q \); \( V \leftarrow V Q \); \( q \leftarrow q^H Q \);
4. Let \( k = m - p \) and put
   
   \( f_k \leftarrow V(:, k+1) \ast H(k+1, k) + f \ast q(k) \);
   
   \( V_k \leftarrow V(:, 1 : k) \); \( H_k \leftarrow H(1 : k, 1 : k) \);
5. Beginning with the \( k \)-step Arnoldi factorization

\[
AV_k = V_k H_k + f_k e_k^T,
\]

apply \( p \) additional steps of the Arnoldi process to obtain a new \( m \)-step Arnoldi factorization

\[
AV = VH + f e_m^T,
\]

Observe that if \( m = n \) then \( f = 0 \) and this iteration is precisely the same as the Implicitly Shifted QR iteration. Even for \( m < n \), the first \( k \) columns of \( V \) and the Hessenberg submatrix \( H(1 : k, 1 : k) \) are mathematically equivalent to the matrices that would appear in the full Implicitly Shifted QR iteration using the same shifts \( \mu_j \). In this sense, the Implicitly Restarted Arnoldi method may be viewed as a truncation of the Implicitly Shifted QR iteration. The fundamental
difference is that the standard Implicitly Shifted QR iteration selects shifts to drive subdiagonal elements of $H$ to zero from the bottom up while the shift selection in the Implicitly Restarted Arnoldi method is made to drive subdiagonal elements of $H$ to zero from the top down. Important implementation details concerning the deflation (setting to zero) of subdiagonal elements of $H$ and the purging of unwanted but converged Ritz values are beyond the scope of this discussion. However, these details are extremely important to the success of this iteration in difficult cases. Nevertheless, the simple basic iteration described above is often successful.

The above iteration can be used to apply any known polynomial restart. If the roots of the polynomial are no known there is an alternative implementation that only requires one to compute $q_1 = \psi(H)e_1$ where $\psi$ is the desired degree $p$ polynomial. A sequence of Householder transformations may developed to form a unitary matrix $Q$ such that $Qe_1 = q_1$ and $H \leftarrow Q^H Q$ is upper Hessenberg. The details which follow standard developments for the Implicitly Shifted QR iteration will be ommitted here.

A shift selection strategy that has proved successful in practice is called the “Exact Shift Strategy”. In this strategy, one computest $\sigma(H)$ and sorts this into two disjoint sets $\Omega_w$ and $\Omega_u$. The $k$ Ritz values in the set $\Omega_w$ are regarded as approximations to the “wanted” eigenvalues of $A$, and the $p$ Ritz values in the set $\Omega_u$ are taken as the shifts $\mu_j$. An interesting consequence (in exact arithmetic) is that after step (???) above, the spectrum of $H_k$ is $\sigma(H_k) = \Omega_w$ and the updated starting vector $v_1$ is a particular linear combination of the $k$ Ritz vectors associated with these Ritz values. In this sense, the iteration is attempting to achieve the desired state of a starting vector that is a linear combination of wanted eigenvectors.

The following diagram indicates how this iteration proceeds
From the standpoint of numerical stability this updating scheme has several advantages:

(1) Orthogonality can be maintained since the value of $k$ is modest.

(2) There is no question of spurious eigenvalues.

(3) There is a fixed storage requirement.

(4) Deflation techniques similar to those associated with the QR-iteration for dealing with numerically small subdiagonal elements of $H_k$ may be taken advantage of directly.

For the sake of clarity, the Arnoldi iteration and the updating procedure will be completely defined:

**Algorithm 3.7**

**function** $[V, H, f] = \text{Arnoldi} (A, V, H, f, k, p)$

**Input:** $AV - VH = f e_k^T$ with $V^T V = I_k$, $V^T f = 0$.

**Output:** $AV - VH = f e_{k+p}^T$ with $V^T V = I_{k+p}$, $V^T f = 0$.

(1) For $j = 1, 2, \ldots, p$

   (1) $\beta \leftarrow \| f \|_2$; if $\beta < \text{tol}$ then stop;

   (2) $H \leftarrow \begin{pmatrix} H & \frac{1}{\beta} f \\ \beta e_{k+j-1}^T \\ \end{pmatrix}$; $v \leftarrow \frac{1}{\beta} f$; $V \leftarrow (V, v)$;

   (3) $w \leftarrow Av$;

   (4) $h \leftarrow V^T w$; $H \leftarrow (H, h)$;

   (5) $f \leftarrow w - Vh$;

   (6) **while** $\| s \| > \epsilon \| f \|$;

      (1) $s = V^T f$;

      (2) $f \leftarrow f - Vs$;

      (3) $h \leftarrow h + s$;

**Remark 1:** Step (1.6) is Gram Schmidt with iterative refinement to assure orthogonality [9]. For details of implementation see Reichel and Gragg [26]. Computational experience with this device indicates that it is sufficient to do just one step of iterative refinement.

With the basic Arnoldi factorization defined, it is possible to describe the complete iteration:

**Algorithm 3.8**

**function** $[V, H, f] = \text{Arnupd} (A, k, p, \text{tol})$.

(1) initialize $V(:, 1) = v_1$; $H \leftarrow (v_1^T Av_1)$; $f \leftarrow Av_1 - v_1 H$ ;
(2) $[V, H, f] \leftarrow \text{Arnoldi} \ (A, V, H, f, 1, k)$

(3) For $m = 1, 2, \ldots$

(1) if ($\|f\| < \text{tol}$) then stop;
(2) $[V, H, f] \leftarrow \text{Arnoldi} \ (A, H, V, f, k, p)$;
(3) $u = \text{Shifts} \ (H, p)$; \hspace{1em} (defined below)
(4) $Q \leftarrow I_{k+p}$;
(5) for $j = 1, 2, \ldots, p$

(1) $H \leftarrow Q_j^T H Q_j$; \hspace{1em} (Bulge-Chase corresponding to shift $\mu_j = u(j)$)
(2) $Q \leftarrow QQ_j$;
(6) $v \leftarrow (VQ)e_{k+1}$; $V \leftarrow V * Q(:, 1 : k)$;
(7) $f \leftarrow (v \beta_k + f \sigma_k)$; where $\beta_k = e_{k+1}^T H e_k$, $\sigma_k = e_{k+p}^T Q e_k$;

Remark 2: The Bulge Chase at step (4.5.1) is defined implicitly as usual so that $H - \mu_j I = Q_j R_j$; if the shifts are in complex conjugate pairs then the implicit double shift can be implemented to avoid complex arithmetic.

Remark 3: During a Bulge Chase sweep at step (4.5.1), it may happen that a sub-diagonal element $\beta_j$ becomes small. The deflation strategies associated with the QR algorithm are then employed. In this case, the matrix $H$ is split giving

$$H = \begin{pmatrix} H_j & M \\ \beta_j e_1 e_j^T & \hat{H}_j \end{pmatrix} \approx \begin{pmatrix} H_j & M \\ 0 & \hat{H}_j \end{pmatrix}, \hspace{1em} VQ = (V_j, \hat{V}_j).$$

Thus, an invariant subspace of dimension $j$ has been found. If $j \geq k$ and all the shifts have been applied then the iteration is halted. Otherwise $H_j, V_j$ are retained and the iteration proceeds with $\hat{V}_j, \hat{H}_j$ filling the role of $V, H$ respectively. However, $H_j$ continues to participate in the shift selection strategy on subsequent iterations. That is, all of the eigenvalues of $H$ are considered in the selection process. If some of the eigenvalues of $H_j$ are selected as shifts then these are applied implicitly to $H_j$ to split this matrix and the unwanted portion is discarded to form a submatrix of smaller size. If the matrix is non-symmetric the factorization must be explicitly restarted at the $j + 1$ position with a vector that is orthogonal to the first $j$ basis vectors. If the matrix $A$ is symmetric then the corresponding columns of the (updated) matrix $\hat{V}_j$ are discarded and then $\hat{V}_j$ and $\hat{H}_j$ are moved (concatenated) to left. The remaining shifts are applied implicitly to $\hat{H}_j$ and then the Arnoldi factorization is completed to fill out the remainder of the $k + p$ columns of $V$. In this way the iteration is not terminated by deflation until the appropriate approximation to the wanted spectrum has appeared.

As discussed at the beginning of this section, each application of an implicit shift $\mu_j$ will replace the starting vector $v_1$ with $(A - \mu_j I)v_1$. Thus after completion of each cycle of the loop at Step 3 in Algorithm (4.8):

$$V e_1 = v_1 \leftarrow \psi(A)v_1;$$
where $\psi(\lambda) = \frac{1}{\tau} \prod_{j=1}^{p} (\lambda - \mu_j)$ with $\tau$ a normalization factor. Numerous choices are possible for the selection of these $p$ shifts. Some possibilities will be discussed in Section 5. However, there is one immediate possibility to discuss and that is the case of choosing $p$ “exact” shifts with respect to $H$. Thus the selection process might be

**Algorithm 3.9**

function $[u] = \text{Shifts} (H, p)$

(1) Compute $\sigma(H)$ (by QR for example)

(2) Select $p$ unwanted eigenvalues \( \{u(j) \leftarrow \mu_j : 1 \leq j \leq p\} \subset \sigma(H) \)

Some obvious criterion for this selection might be

(i) Sort $\sigma(H)$ according to algebraically largest real part and select the $p$ eigenvalues with smallest real part as shifts;

(ii) Sort $\sigma(H)$ according to largest modulus the $p$ eigenvalues with smallest modulus as shifts;

Selecting these exact shifts has interesting consequences in the iteration.

**Lemma 3.10** Let $\sigma(H) = \{\theta_1, \ldots, \theta_k\} \cup \{\mu_1, \ldots, \mu_p\}$ be a disjoint partition of the spectrum of $H$ and let

$$H_+ = Q^T HQ$$

where $Q = Q_1 Q_2 \cdots Q_p$ with $Q_j$ implicitly determined by the shift $\mu_j$. If $\beta_j \neq 0 \quad 1 \leq j \leq k - 1$ then $\beta_k = 0$ and

$$H_+ = \begin{pmatrix} H^+_k & M^+ \\ 0 & R_p \end{pmatrix}$$

where $\sigma(H^+_k) = \{\theta_1, \ldots, \theta_k\}$, $\sigma(R_p) = \{\mu_1, \mu_2, \ldots, \mu_p\}$. Moreover,

$$v_1^+ = V Q e_1 = \sum x_j$$

where each $x_j$ is a Ritz vector corresponding to the Ritz value $\theta_j$ i.e. $x_j = V y_j$ where $H y_j = y_j \theta_j \quad 1 \leq j \leq k$.

**Proof:** After applying the $p$ implicit shifts we have

$$HQ = Q H_+$$

with

$$q_1 = Q e_1 = \psi(H)e_1, \quad \psi(\lambda) = \frac{1}{\tau} \prod_{j=1}^{p} (\lambda - \mu_j).$$

Therefore $q_1 = \sum_{j=1}^{k} y_j \zeta_j$ where $H y_j = y_j \theta_j$ since $q_1 = \psi(H)e_1$ has annihilated any component of $e_1$ along an eigenvector of $H$ associated with $\mu_j$, $1 \leq j \leq p$. 

KTH Lecture Notes, Draft, Spring 1999, D. C. Sorensen
As a consequence of Theorem (2.8), $\beta_k = 0$ must hold. Moreover, $v_i^+ = V Q e_1 = V q_1 = \sum_{j=1}^k V y_j \zeta_j = \sum_{j=1}^k x_j \zeta_j$.

This lemma provides a very nice interpretation of the iteration when exact shifts are chosen. Casting out the unwanted set of eigenvalues using exact shifts is mathematically equivalent to restarting the Arnoldi Factorization from the beginning after updating $v_1 \leftarrow \sum x_j \zeta_j$ a linear combination of Ritz vectors associated with the “wanted” eigenvalues. Thus the updated starting vector has been implicitly replaced by the sum of $k$ approximate eigenvectors.

If $A$ is symmetric and the $p$ algebraically smallest eigenvalues of $H$ are selected for deletion then this method is similar to the single vector $s$-step Lanczos process described by Cullum and Donath in [5] and expanded on in [6, 8]. The particular linear combination is apparently different. This variant has the advantage that a restart of the entire Lanczos sequence is not required. Approximate eigenvectors from a Krylov subspace of dimension $k + p$ are available at each iteration for a cost of $p$ rather than $k + p$ matrix vector products per iteration.

5 Orthogonal Polynomials

There is a beautiful direct connection between the Lanczos polynomials and the classic theory of moments and orthogonal polynomials. Succinctly stated, the moment problem is:

- Given a sequence of moments $\mu_j = \int \lambda^j d\mu$ construct a discrete measure $\hat{\mu}$ such that
  $$\mu_j = \int \lambda^j d\hat{\mu}, \text{ for } j = 1, 2, \ldots, k$$

A related problem is given a finite measure and its associated inner product construct orthogonal polynomials with respect to the inner product and construct the associated Gaussian Quadrature Formula.

5.1 The Lanczos measure

We shall restrict this discussion to real symmetric matrices $A$. The results generalize in a straightforward manner to the Hermitian case but the non-Hermitian case presents difficulties that will not be addressed here.

So far we have developed the Lanczos factorization of a symmetric $A$ to be

$$AV = VH + f e_k^T \quad (5.1)$$

where $V \in \mathbb{R}^{n \times k}$ has orthonormal columns, $V^T f = 0$ and $H \in \mathbb{R}^{k \times k}$ is symmetric and tridiagonal with non-negative subdiagonal elements. We have noted that

$$Range(V) = \text{Span } \{v_1, Av_1, A^2 v_1, \ldots, A^{k-1} v_1\} = K_k(A, v_1)$$
and that \( v_j = p_{j-1}(A)v_1 \) where \( p_{j-1} \) is a polynomial of degree \( j - 1 \). Moreover, by construction \( s^T v_j = 0 \) for all \( s \in \mathcal{K}_j(A, v_1) \). In other words \( v_j \) is orthogonal to every vector of the form \( \psi(A)v_1 \) where \( \psi \) is a polynomial of degree less than \( j - 1 \). Since \( H \) is symmetric and tridiagonal it has the form

\[
H_k = \begin{bmatrix}
\alpha_1 & \beta_1 & 0 & \cdots & 0 \\
\beta_1 & \alpha_2 & \beta_2 & \ddots & \\
0 & \ddots & \ddots & \ddots & 0 \\
\vdots & & \ddots & \ddots & \beta_{k-2} & \alpha_{k-1} & \beta_{k-1} \\
0 & \cdots & 0 & \beta_{k-1} & \alpha_k
\end{bmatrix}.
\]

For the remainder of this section we assume that all \( \beta_j \) are positive (i.e. that the tridiagonal \( H_k \) is unreduced).

As noted previously, equating the \( j-th \) column on both sides of (5.1) will give the relation

\[
Av_j = v_{j-1}\beta_{j-1} + v_j\alpha_j + v_{j+1}\beta_j
\]

which leads to the three term recurrence

\[
v_{j+1}\beta_j = (A - \alpha_j I)v_j - v_{j-1}\beta_{j-1}
\]

and thus yields the relationship

\[
p_j(A)v_1\beta_j = (A - \alpha_j I)p_{j-1}(A)v_1 - p_{j-2}(A)v_1\beta_{j-1}
\]

with \( p_{-1} = 0 \) and \( p_0 = 1 \). The associated polynomials

\[
p_j(\lambda)\beta_j = (\lambda - \alpha_j)p_{j-1}(\lambda) - p_{j-2}(\lambda)\beta_{j-1}
\]

have been called the Lanczos polynomials.

Let \( A = Q\Lambda Q^T \) be the eigen-decomposition of \( A \), with \( Q^T Q = I \) and \( \Lambda = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n) \). Let \( g = Q^T v_1 = (\gamma_1, \gamma_2, \ldots, \gamma_n)^T \). Then for any polynomial \( \psi \) we have

\[
\psi(A)v_1 = Q\psi(\Lambda)Q^Tv_1 = Q\psi(\Lambda)g.
\]

Hence

\[
v_1^T \psi(A)v_1 = g^T \psi(\Lambda)g = \sum_{j=1}^n \psi(\lambda_j)\gamma_j^2
\]

Observe that

\[
v_1^T \psi(A)v_1 = \int_I \psi(\lambda)d\mu(\lambda)
\]

where \( \mu \) is defined by

\[
\mu(\lambda) = \sum_{j=1}^n \gamma_j^2 h(\lambda - \lambda_j)
\]
with $h(\tau) = 1$ if $\tau \geq 0$ and $h(\tau) = 0$ if $\tau < 0$ and $\mathcal{I}$ is any interval containing $\sigma(A)$. Of course this relation extends to the family of real analytic functions on $\mathcal{I}$.

The natural inner product associated with $\mu$ is

$$< \phi, \psi > \equiv \sum_{j=1}^{n} \phi(\lambda_j) \psi(\lambda_j) \gamma_j^2$$

and we might ask: What are the orthogonal polynomials with respect to this inner product?

The answer to this question may be obvious by now. The Lanczos polynomials by construction satisfy

$$< p_i, p_j > = v_i^T p_i(A)^T p_j(A)v_1 = \delta_{ij}$$

the Kroneker delta. Moreover, recall (see Sec 3.3) with the definition

$$\hat{p}_j \equiv \text{det}(\lambda I - H_j)$$

that $p_j(A)v_1 = \hat{p}_j(A)v_1/\|\hat{p}_j(A)v_1\|$ and that

$$\beta_j = \frac{\|\hat{p}_j(A)v_1\|}{\|\hat{p}_{j-1}(A)v_1\|} = \frac{< \hat{p}_j, \hat{p}_j >^{1/2}}{< \hat{p}_{j-1}, \hat{p}_{j-1} >^{1/2}} \quad (5.2)$$

$$\alpha_{j+1} = v_{j+1}^T A v_{j+1} = v_1^T p_j(A)^T A p_j(A)v_1$$

$$= \frac{< \hat{p}_j, \lambda \hat{p}_j >}{< \hat{p}_j, \hat{p}_j >}$$

From (5.2) we discover that

$$1 = \|\hat{p}_0(A)v_1\|$$

$$\beta_1 = \|\hat{p}_1(A)v_1\|$$

$$\beta_2 \beta_1 = \beta_2 \|\hat{p}_1(A)v_1\| = \|\hat{p}_2(A)v_1\|$$

$$\vdots$$

$$\beta_j \beta_{j-1} \ldots \beta_2 \beta_1 = \beta_j \|\hat{p}_{j-1}(A)v_1\| = \|\hat{p}_j(A)v_1\|$$

As a result of Theorem (3.4) we have

$$\beta_j \beta_{j-1} \ldots \beta_2 \beta_1 = \|\hat{p}_j(A)v_1\| = \min_{p \in P_{\lambda_j}} \{\|p(A)v_1\|\}.$$ 

Finally, since $< p_j, p_j > = 1$, it follows that

$$\hat{p}_j(\lambda) = \beta_j \beta_{j-1} \ldots \beta_2 \beta_1 p_j(\lambda).$$
5 ORTHOGONAL POLYNOMIALS

5.2 Exercises

1. Partition the matrix

\[ H_{j+1} = \begin{bmatrix} H_j & e_j \beta_j \\ \beta_j e_j^T & \alpha_{j+1} \end{bmatrix} \]

and expand the determinant by minors about the last column to verify that

\[ \det(\lambda I - H_{j+1}) = (\lambda - \alpha_{j+1})\det(\lambda I - H_j) - \beta_j^2 \det(\lambda I - H_{j-1}) \]

2. The previous exercise implies that

\[ \hat{p}_{j+1}(\lambda) = (\lambda - \alpha_{j+1})\hat{p}_j(\lambda) - \beta_j^2 \hat{p}_{j-1}(\lambda). \]

Use the relation \( \hat{p}_j(\lambda) = \beta_j \hat{p}_{j-1}(\lambda) \) in this three term recurrence to verify that it gives precisely the same formula as the three term recurrence obtained through the Lanczos process.

3. Apply the Gram-Schmidt process to orthogonalize the set of monomials \( \{1, \lambda, \lambda^2, \ldots, \lambda^k\} \) with respect to the inner product \( \langle \cdot, \cdot \rangle \) to verify that this process gives precisely the same \( \{\alpha_j\} \) and \( \{\beta_j\} \) as the Lanczos process.

5.3 Gaussian Quadrature

Once orthogonal polynomials have been constructed for a particular measure with respect to the associated inner product, it is possible to construct Gaussian Quadrature formulae that may be used to approximate integrals of functions (with respect to the measure). A \( k \)-point Gaussian quadrature formula has the form

\[ \int_I \psi(\lambda) d\mu(\lambda) \approx \sum_{j=1}^k \psi(\theta_j) \omega_j^2. \]

The \( \{\theta_j\} \) are called the Gaussian quadrature points (abscissa) and the \( \{\omega_j^2\} \) are called the Gaussian weights. The classic result is that this formula is exact for the integral of every polynomial of degree less than \( 2k \).

It turns out that the Lanczos process generates these weights and abscissa automatically as part of the eigen-system of the tridiagonal matrices \( H_j \). This relation will be developed now. To begin the development it is helpful to establish a technical lemma that will prove to be useful in other contexts later on. The lemma applies to the general Arnoldi process as well as the symmetric case.

**Lemma 5.1** Let \( A \in \mathbb{R}^{n \times n} \) and suppose that

\[ AV = VH + fe_k^T \]

is a \( k \)-step Arnoldi factorization of \( A \). Let \( \psi \) be a polynomial. If the degree of \( \psi \) is less than \( k \) then

\[ \psi(A)v_1 = V\psi(H)e_1 \]
and if the degree of \( \psi \) is \( k \) then

\[
\psi(A)v_1 = V\psi(H)e_1 + f\omega
\]

where \( \omega = \alpha \beta_{k-1} \cdots \beta_2 \beta_1 \), and \( \alpha \) is the leading coefficient of \( \psi \).

**Proof:** First verify that \( e_k^T \hat{\psi}(H)e_1 = 0 \) if the degree of \( \hat{\psi} \) is less than \( k - 1 \) and \( e_k^T \hat{\psi}(H)e_1 = \beta_{k-1} \cdots \beta_2 \beta_1 \) if \( \hat{\psi} \) is monic of degree \( k - 1 \). Next observe as before that

\[
(A - \mu I)V = V (H - \mu I) + fe_k^T
\]

and thus

\[
(A - \mu I)V e_1 = V (H - \mu I)e_1.
\]

Assume the result is true for polynomials of degree less than \( k - 1 \). If \( \psi \) is a polynomial of degree less than \( k \) then \( \psi(\lambda) = (\lambda - \mu)\hat{\psi}(\lambda) \) with the degree of \( \hat{\psi} \) less than \( k - 1 \). Thus

\[
\psi(A)v_1 = (A - \mu I)e_k^T \phi(A)v_1 \\
= (A - \mu I)V e_k^T \phi(H)e_1 \\
= [V (H - \mu I) + fe_k^T]\phi(H)e_1 \\
= V\phi(H)e_1 + fe_k^T \phi(H)e_1 = V\phi(H)e_1.
\]

This establishes the result by induction for any polynomial \( \psi \) of degree less than \( k \). If the degree of \( \psi \) is \( k \), a similar argument leads to the final result since \( e_k^T \phi(H)e_1 = \alpha \beta_{k-1} \cdots \beta_2 \beta_1 \) where \( \alpha \) is the leading coefficient of \( \psi \).

Now we return to the symmetric case \( A = A^T \) and \( H \) symmetric tridiagonal. Let

\[
H = Y \Theta Y^T
\]

be the eigendecomposition of \( H \) with \( \Theta = \text{diag}(\theta_1, \theta_2, \ldots, \theta_k) \) and \( Y^TY = I \). Define \( (\omega_1, \omega_2, \ldots, \omega_k)^T = e_1^T Y \) and note that \( \sum_{j=1}^k \omega_j^2 = 1 \) since this is the first row of an orthogonal matrix. Furthermore, none of the \( \omega_j \) vanish if \( H \) is unreduced. Noting that the \( \theta_j \) are the roots of the Lanczos polynomial \( p_k \), the sets \( \{\theta_j\} \) and \( \{\omega_j^2\} \) shall serve as the Gaussian quadrature points and weights respectively in (5.3).

Now suppose \( \psi \) is a polynomial of degree \( 2k - 1 \). Express \( \psi(\lambda) = \psi_{k-1}(\lambda)\psi_k(\lambda) \) where the subscripts denote the degree of the polynomial factors. With Lemma (5.1) we have established

\[
\int_T \psi(\lambda)d\mu(\lambda) = \langle \psi_{k-1}, \psi_k \rangle \\
= v_1^T \psi_{k-1}(A)^T \psi_k(A)v_1 \\
= e_1^T \psi_{k-1}(H)^TY^TY\psi_k(H)e_1 \\
= e_1^TY\psi(\Theta)Y^Te_1 \\
= \sum_{j=1}^k \psi(\theta_j)\omega_j^2.
\]
5 ORTHOGONAL POLYNOMIALS

The second to last inequality makes use of symmetry of \( H \) and orthogonality of \( V \). This establishes the desired result. Namely, the Gaussian quadrature formula represents the exact integral for all polynomials \( \psi \) of degree less than \( 2k \).

5.4 Exercises

1. The Tchebyshev polynomials \( \{p_j\} \) are specified by the three term recursion

\[
p_{j+1}(\lambda) = 2\lambda p_j(\lambda) - p_{j-1}(\lambda)
\]

with \( p_{-1} = 0 \), and \( p_0 = 1 \). Verify that these polynomials are orthogonal with respect to the inner product

\[
< \phi, \psi > = \int_{-1}^{1} \phi(\lambda)\psi(\lambda)\omega(\lambda)\,d\lambda
\]

and that the roots of the \( k-th \) polynomial are

\[
\theta_{j,k} = \cos\left(\frac{j\pi}{k+1}\right), \quad j = 1, 2, \ldots, k.
\]

2. Use Matlab to verify the formulas we have derived. Compute the the eigenvalues of the order \( k \) tridiagonal matrix

\[
A_k = \begin{bmatrix}
0 & 1 & 0 & \cdots & 0 \\
1 & 0 & 1 & \ddots & \\
0 & \ddots & \ddots & \ddots & 0 \\
\vdots & 1 & 0 & 1 & \\
0 & \cdots & 0 & 1 & 0
\end{bmatrix}
\]

and compare these eigenvalues to the values given by the formula in (1).

3. Plot the eigenvectors corresponding to a few of the extreme eigenvalues. Can you derive a formula for the eigenvalue entries in terms of the polynomials and roots?

4. What are the eigenvalues of the discrete Laplacian

\[
A_k = \begin{bmatrix}
2 & -1 & 0 & \cdots & 0 \\
-1 & 2 & -1 & \ddots & \\
0 & \ddots & \ddots & \ddots & 0 \\
\vdots & -1 & 2 & -1 & \\
0 & \cdots & 0 & -1 & 2
\end{bmatrix}
\]
5. Let $\mu$ be a positive measure on an interval $\mathcal{I}$ and consider the sequence of moments
\[ \mu_j = \int_{\mathcal{I}} \lambda^j d\mu(\lambda) \quad j = 0, 1, 2, \ldots \]
Define the inner product
\[ < f, g > = \int_{\mathcal{I}} f(\lambda)g(\lambda)d\mu(\lambda). \]
Then $\mu_{i+j} = \langle \lambda^i, \lambda^j \rangle$. Explain how to use the Lanczos algorithm to construct orthogonal polynomials and Gaussian quadrature rules from the sequence of moments $\{\mu_j \mid j = 0, 1, 2, \ldots \}$. To do this consider the Hankel moment matrix of order $n$
\[
A_n = \begin{bmatrix}
\mu_0 & \mu_1 & \mu_2 & \cdots & \mu_{n-1} \\
\mu_1 & \mu_2 & \mu_3 & \cdots & \mu_n \\
\mu_2 & \mu_3 & \mu_4 & \cdots & \mu_{n+1} \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
\mu_{n-2} & \mu_{n-1} & \mu_n & \cdots & \mu_{2n-3} \\
\mu_{n-1} & \mu_n & \mu_{n+1} & \cdots & \mu_{2n-2}
\end{bmatrix}.
\]
and note that
\[ \langle \lambda^{i-1}, \lambda^{j-1} \rangle = e_i^T A_n e_j. \]