# *Projection methods for solving large sparse eigenvalue problems*

***Youcef Saad***

Technical Report # 224

April 2, 1982

This work was supported in part by the U.S. Office of Naval Research under grant N000014-76-C-0277 and in part by NSF Grant.

## Table of Contents

## *Abstract*

We present a unified approach to several methods for computing eigenvalues and eigenvectors of large sparse matrices. The methods considered are projection methods, i.e. Galerkin type methods, and include the most commonly used algorithms for solving large sparse eigenproblems like the Lanczos algorithm, Arnoldi's method, the subspace iteration, etc.. We first derive some a priori error bounds for general projection methods, in terms of the distance of the exact eigenvector from the subspace of approximation. Then this distance is estimated for some typical methods, particularly those for unsymmetric problems.

## 1. Introduction

In the previous few years a fairly important effort has been devoted to solving large sparse eigenvalue problems. Although more attention has been directed towards symmetric eigenvalue problems, many applications are now encountered where one requires the eigenvalues of a large unsymmetric matrix.

The purpose of this paper is to attempt to present a unified view of the most commonly used algorithms for solving large sparse eigenproblems. We will start by reviewing the general framework of projection methods and describe orthogonal as well as oblique projection methods. A projection method consists in approximating the exact eigenvector u, by a vector $\tilde{u}$ belonging to some subspace K, referred to as the right subspace, by requiring that the residual vector of $\tilde{u}$ satisfies the Petrov-Galerkin condition that it is orthogonal to some subspace L, possibly different from K, often rederred to as the left subspace. When L=K we have an orthogonal projection method otherwise we say that the method is an oblique projection method. As it turns out most methods for solving large sparse eigensystems can be formulated in terms of projection methods . As will be seen, the common feature which makes these methods work, is that the exact eigenvector is well approximated by some vector of the subspace K. It becomes then important to analyse the distance between the exact eigenvector and the subspace of approximation. This will be done for several methods with emphasis on those for solving unsymmetric problems, including the subspace iteration, the method of Arnoldi, etc..

Concerning the subspace iteration method we will see that Chebyshev acceleration can also be efficiently used and we will derive some estimates of the convergence factor.

Throughout the paper, the norm $\| \; \|$ represents the Euclidean norm. The spectrum of a matrix is denoted by $\sigma(A)$. The matrices treated may be complex and the transpose conjugate of A is denoted by $A^H$.

## 2. General projection methods for matrix eigenvalue problems.

In this section we present the general projection methods which provide a unified approach to many methods for computing eigenvalues and eigenvectors of large matrices.

### 2.1. Orthogonal projection methods.

The material of this subsection summarizes part of the previous paper [11]. Let A be an NxN complex matrix and **K** be an m-dimensional subspace of $\mathbf{C}^N$. We will make the notational convention of representing by the same symbol A the matrix and the linear operator represented by the matrix A.

Consider the eigenvalue problem

$$A\,u = \lambda\,u \tag{1}$$

An orthogonal projection method on the subspace **K** seeks an approximate eigenpair $\tilde{\lambda}$, $\tilde{u}$ to problem (1), which belongs to the subspace **K** and such that the following Galerkin condition is satisfied.

$$A\,\tilde{u} - \tilde{\lambda}\,\tilde{u} \perp \mathbf{K} \tag{2}$$

In terms of projection operators, if $\Pi$ represents the orthogonal projector onto the subsapce **K**, then the above Galerkin condition (2) can be rewritten as

$$\Pi(\,A\,\tilde{u} - \tilde{\lambda}\,\tilde{u}\,) = 0 \tag{3}$$

We will refer to (2) or (3) as the approximate problem. Assuming that we have an orthonormal basis $V = [v_1,v_2,..v_m]$ of **K** we can solve the approximate problem (2) by expressing the approximation $\tilde{u}$ in the basis V as

$$\tilde{u} = V\,y \tag{4}$$

in which case $\tilde{\lambda}$ and y constitute an eigenpair of the m dimensional eigenproblem derived from (2):

$$B_m\,y = \tilde{\lambda}\,y \tag{5}$$

with

$$B_m = V^H\,A\,V \tag{6}$$

We will denote by $A_m$ the linear application of rank m, defined by $A_m = \Pi A \Pi$ . Note that the restriction of $A_m$ to **K** is represented by the matrix $B_m$ with respect to the basis V. An important quantity for the convergence properties of the method is the distance $||(I-\Pi)u||$ of the exact eigenvector u, which is supposed of norm 1, from the subspace **K**. First it is clear that the eigenvector u cannot be well approximated from **K** if $||(I-\Pi)u\ ||$ is not small , because

$$||u-\tilde{u}|| \geq ||(I-\Pi)u||$$

The fundamental distance $||(I-\Pi)u||$ can also be interpreted as the sine of the acute angle between the eigenvector u and the subspace of approximation. In [11] it was shown that if we consider the exact eigenvector u as an approximate eigenvector of $A_m$ then the corresponding residual vector satisfies the inequality:

$$||\ (A_m - \lambda I)\ u\ || \leq (\ |\lambda|^2 + \ \gamma^2\ )^{1/2}\ ||\ (I-\Pi)\ u\ || \tag{7}$$

where

$$\gamma = ||\ \Pi A(I-\Pi)||$$

Note that $\gamma$ can be bounded by $||A||$, and this indicates that we can obtain a good approximation provided that the distance $||(I-\Pi)u||$ is small. Among the methods which are of the type described above let us mention the symmetric Lanczos method (see e.g. [7]) , some of the Subspace iteration methods [7, 16], the method of Arnoldi [1, 11].

## 2.2. Oblique projection methods.

Several methods for large matrices can be interpreted in terms of oblique projection methods, or Petrov-Galerkin methods. In these methods we are given a second subspace **L** which may be different from **K**, and we seek an approximation $\tilde{u}$ belonging to **K** and satisfting the Petrov Galerkin condition:

$$A\ \tilde{u} - \tilde{\lambda}\ \tilde{u}\ \perp\ \mathbf{L} \tag{8}$$

The subsapce **K** is often referred to as the right subspace and **L** as the left subspace. In order to interpret the above condition in terms of operators we require the oblique projector Q onto **K** and orthogonal to **L** which is defined by:

$$Qx \in \mathbf{K}$$

$$x - Qx \perp \mathbf{L}$$

Note that the vector Qx is uniquely defined only under the assumption that *no vector of* **L** *is orthogonal to* **K**. It is easy to see that this fundamental assumption is equivalent to:

Assumption: For any two bases V and W of **K** and **L** respectively we have

$$\mathrm{Det}(\ W^H V\ ) \neq 0 \tag{9}$$

The projector Q is illustrated in figure (2-1).

We can rewrite equation (8) as

$$Q(A\ \tilde{u} - \tilde{\lambda}\ \tilde{u}) = 0 \tag{10}$$

which again can be translated matricially by expressing the approximate eigenvector $\tilde{u}$ in an appropriate basis of **K**. Assume that a basis V of **K** and a basis W of **L** can be found such that V and W form a biorthogonal pair i.e. such that

$$V^H W = I$$

where I is the identity matrix. Then if we write $\tilde{u}=Vy$ , the Petrov-Galerkin condition (8) gives the same approximate problem as (5) except that the matrix $B_m$ is this time defined by:

$$B_m = W^H\ A\ V$$

We should however emphasize that in order for a biorthogonal pair V, W to exist we must make the above assumption (9).

We can establish the following theorem which generalizes the result (7) of [11] .

**Theorem 1:** *Let $\gamma=||Q(A-\lambda I)(I-\Pi)||$. Then the following two inequalities hold:*

1) $$|| (A_m - \lambda I)\ \Pi u\ || \leq\ \gamma\ ||\ (I-\Pi)\ u\ || \tag{11}$$

2) $$|| (A_m - \lambda I)\ u\ || \leq \sqrt{|\lambda|^2 + \gamma^2}\ \ \ ||\ (I-\Pi)\ u\ || \tag{12}$$

**Proof:**

1) Since $\Pi y$ belongs to **K** for any vector y, we have $Q\Pi=\Pi$ and

$$\begin{aligned}(A_m - \lambda I)\ \Pi u &= Q\ (A - \lambda I)\ \Pi u \\ &= Q(A-\lambda I)(\Pi u - u) \\ &= -\ Q(A-\lambda I)(I-\Pi)u\end{aligned} \tag{13}$$

and since (I-Π) is a projector

$$(A_m - \lambda I)\ \Pi u = -\ Q(A-\lambda I)(I-\Pi).(I-\Pi)u$$

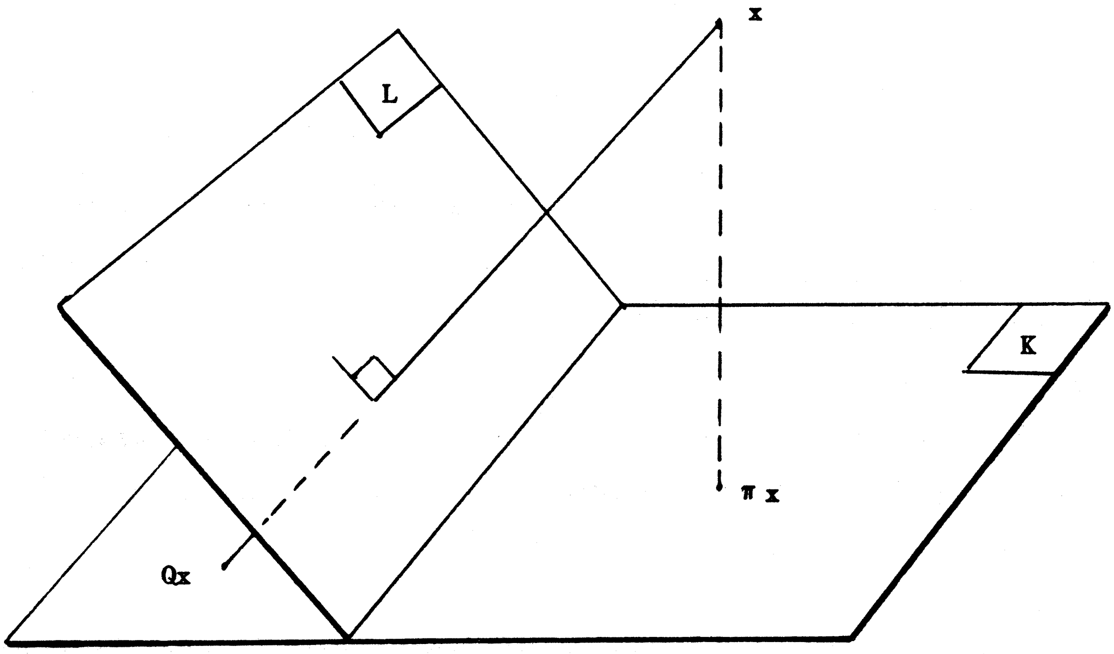

**Figure 2-1:** The oblique projector Q

Taking the Euclidean norms of both sides we immediately obtain the result (11)

2) We have

$$(A_m - \lambda I)u = (A_m - \lambda I)[\Pi u + (I-\Pi)u]$$

$$= (A_m - \lambda I)\Pi u + (A_m - \lambda I)(I-\Pi)u$$

Noticing that $A_m(I-\Pi)=0$, this becomes:

$$(A_m - \lambda I)u = (A_m - \lambda I)\Pi u - \lambda(I-\Pi)u$$

Since the two terms of the right hand side are orthogonal we obtain

$$||(A_m - \lambda I)u||^2 = ||(A_m - \lambda I)\Pi u||^2 + ||\lambda(I-\Pi)u||^2$$

Using inequality (11) this immediatly gives the result (12). □

Note that (12) is a simple generalization of (7). In the case of orthogonal projection methods we have $||Q||=1$ and $\gamma$ may be bounded by $||A||$. It may seem that since we obtain very similar error bounds for both the orthogonal projection method and the oblique projection method we are likely to get similar errors when we use the same subspace **K**. This is unfortunately not the case because unlike in the orthogonal projection method, the scalar $\gamma$ can no longer be bounded by $||A||$, since we have $||Q|| \geq 1$ and $||Q||$ is unknown in general. In fact the constant $\gamma$ might very well be a large number even if A is a matrix with a moderate norm. Besides $\gamma$ we also have the difficulty that we do not have any information about the conditioning of the approximate problem.

## 2.3. Application: orthogonal projection methods for symmetric problems

We will now restrict ourself to the case where an orthogonal projection method is applied to a symmetric problem and will establish a few consequences of theorem 1. First notice the important fact that the first inequality can be matricially translated in the subspace **K** by expressing the vector $\Pi u$ and the operator $A_m$ in a certain basis $V:=\{v_1,v_2,..v_m\}$ of **K**. Assuming V is orthonormal let us set $\Pi u/|| \Pi u ||=Vy$ and define $B_m$ as in (6). Then inequality (11) translates itself into:

$$|| (B_m - \lambda I)y || \leq \gamma || (I-\Pi) u ||/|| \Pi u || \qquad (14)$$

where $B_m$ is the representation of the restriction of $A_m$ into **K** as defined by (6).

As mentioned above when A is unsymmetric and Q is oblique then we may encounter some difficulties because we have little information on the conditioning of the approximate eigenvalue problem and because the constant $\gamma$ may be large. However, in the symmetric case and when **L=K**, the situation is somehow simpler, and we are able to give more precise bounds. We would like to apply the following well-known result , in conjunction with our result (11).

**Theorem 2:** *Let $B$ be any symmetric matrix, and $y$ a vector of norm 1. Consider the Rayleigh quotient $\mu=(By,y)$ and let $\rho$ be the norm of the residual vector associated with $y$, i.e. $\rho=\|(B-\mu I)y\|$ Then there exists an eigenvalue $\tilde{\lambda}$ of $B$ with associated eigenvector $\tilde{u}$ such that:*

$$|\mu - \tilde{\lambda}| \leq \rho^2 / \delta \tag{15}$$

$$\sin[\theta(u,\tilde{u})] \leq \rho / \delta \tag{16}$$

*where $\delta$ is the distance between $\mu$ and $\sigma(B)$-$\{\tilde{\lambda}\}$ i.e.*

$$\delta = \min\{ |\mu - \tilde{\lambda}'|, \tilde{\lambda}' \in \sigma(B), \tilde{\lambda}' \neq \tilde{\lambda} \}.$$

*and $\theta(u,\tilde{u})$ denotes the acute angle between $u$ and $\tilde{u}$.*

For the proof see e.g. [7] or [2]. The above fundamental result shows in particular the well known fact that the error for the Rayleigh quotient as an eigenvalue is of order the square of the residual norm. Our next objective is naturally to show that the Rayleigh quotient of $\Pi u$ is not too different from the exact eigenvalue $\lambda$ when $\|(I-\Pi)u\|$ is small.

**Lemma 3:** *The Rayleigh quotient $\tilde{\mu}$ of the vector $\Pi u$, where $u$ is the exact eigenvector of $A$ associated with $\lambda$, is such that:*

$$|\lambda - \tilde{\mu}| \leq \|A - \lambda I\| \, \|(I-\Pi)u\|^2 / \|\Pi u\|^2 \tag{17}$$

**Proof:** We have

$$|\lambda - \tilde{\mu}| = \frac{((A-\lambda I)\Pi u, \Pi u)}{(\Pi u, \Pi u)} = \frac{((A-\lambda I)(I-\Pi)u , (I-\Pi)u)}{(\Pi u, \Pi u)}$$

Hence the result by use of the Cauchy Schwartz inequality. □

With the help of the above two results we can prove:

**Theorem 4:** *Let $\lambda$ be an eigenvalue of A, with associated eigenvector u of norm 1 and let*

$$\epsilon = || (I-\Pi)u ||/|| \Pi u ||$$

*Then there exists an eigenvalue $\tilde{\lambda}$ of the approximate problem satisfying:*

$$| \lambda - \tilde{\lambda} | \leq \tau \epsilon^2 \tag{18}$$

*in which $\tau = ||A-\lambda I|| + (\gamma/\delta)^2$, with $\gamma$ defined in theorem 1 and where $\delta$ is the distance from $\mu$, the Rayleigh quotient of $\Pi u$, to the approximate eigenvalues $\tilde{\lambda}$' different from $\tilde{\lambda}$.*

**Proof:** The result is a consequence of the triangle inequality

$$|\lambda - \tilde{\lambda}| \leq | \lambda - \mu| + | \mu - \tilde{\lambda}|$$

and lemma (3), theoerm (2) , and inequality (14). □

The above result expresses the error in the eigenvalue directly in terms of he distance between the exact eigenvector u and the subspace of approximants and shows that this error is of order the square of this distance. A somehow weaker but more general inequality can be derived from the well known result [7]:

**Lemma 5:** *Let B be any symmetric matrix, y a vector of norm 1 and $\lambda$ any scalar. Let $\rho$ be the norm of the residual vector associated with y and $\lambda$ i.e.*

$$\rho = || (B-\lambda I)y ||$$

*Then there exists an eigenvalue $\tilde{\lambda}$ of B such that:*

$$| \lambda - \tilde{\lambda} | \leq \rho \tag{19}$$

This with inequality (11) immediately proves the theorem:

**Theorem 6:** *Let $\lambda$ be an eigenvalue of A. Then there exists an eigenvalue $\tilde{\lambda}$ of the approximate problem such that :*

$$| \lambda - \tilde{\lambda} | \leq \gamma \epsilon \tag{20}$$

*where $\gamma$ is defined in theorem 1, and $\epsilon$ is as in theorem 4.*

Clearly inequality (20) is weaker than the previous result (18). It has, however, its own importance. Most projection methods use a sequence a sequence of subspaces $\mathbf{K}_m$ for which it is known that $|| (I-\Pi_m)u||$ converges to zero when m tends to infinity. Thus (20) proves the global convergence of an approximate eigenvalue towards the exact eigenvalue $\lambda$ in this situation. As an example for the symmetric Lanczos algorithm it is known that $\| (I-\Pi_m)u \|$ converges to zero under the assumption that the initial vector in the Lanczos algorithm is not deficient in the direction u , see [12]. Therefore there will be at least one sequence of eigenvalues converging towards the exact eigenvalue. Note however that in this particular case there are alternative a priori error bounds which are more accurate, see [12]

We now turn to the problem of estimating the error in the eigenvector.

**Theorem 7:** *Let u be a unit eigenvector of A and let $\theta(u,\mathbf{K})$ be the acute angle between u and the subspace* **K**, *defined by*

$$\sin[\theta(u,\mathbf{K})] = \| (I-\Pi)u \| \tag{21}$$

*Then there exists an eigenvector $\tilde{u}$ of the approximate problem such that*

$$\sin[\,\theta(u,\tilde{u})] \leq \sqrt{1 + \gamma^2/d^2}\ \sin[\theta(u,\mathbf{K})] \tag{22}$$

*where $\gamma$ defined in theorem 1 and where* d *is the distance between $\lambda$ and the set of approximate eigenvalues other than $\tilde{\lambda}$.*

**Proof**: This result was shown in [12] in the context of the Lanczos algorithm , i.e. K is a Krylov subspace, see also [7] for an elegant presentation. It is transparent from the proofs presented there that we do not make use at any time of the fact that the subspace of approximation is a Krylov subspace. The result is therefore true for any orthogonal projection method. □

These results do not extend immediately to unsymmetric problems or to oblique projection methods. The main reason is that we do not have at our disposal the powerful theorem 2. In a recent article, Kahan, Parlett and Jiang have derived alternative error bounds generalizing (2) but using a residual norm of the right and the left approximate eigenvectors [4]. Their idea may be used in our context and this remains to be done. Without the knowledge of the left eigenvector , the best we can hope for is some partial information whereby the bound contains parameters from the approximate problem which are not known a priori. As an example the inequality (20) becomes in the unsymmetric case, assuming $B_m$ diagonalizable [18]:

$$| \lambda - \tilde{\lambda} | \leq \gamma \| X \| \, \| X^{-1} \| \, \| (I-\Pi)u \|$$

where X is a matrix which diagonalizes $B_m$. Clearly the global condition number $\| X \| \, \| X^{-1} \|$ is not known a priori. Note that even in the symmetric case many a priori error bounds use some a posteriori knowledge on the eigenvalues like for example the distance $\delta$ in theorem (2). The result (20) is an exception. Finally we mention that some analysis of the norm of the projector Q is proposed in [14].

## 3. The subspace iteration for nonsymmetric eigenvalue problems.

In this class of methods the space of approximants is a subspace $\mathbf{S}_m$ spanned by a system $S_m$ given by $S_m = A^m S_o$ where $S_o$ is some initial system of r linearly independent vectors. There are two main versions of the method, both originally due to Bauer. The first uses two subspaces and is an oblique projection method [3] known originally under the name of bi-iteration. The second uses one subspace and is an orthogonal projection method, originally named treppeniteration. We will restrict ourself to this second class of methods which require less work in general. An efficient way of implementing the method has been presented by Stewart [16] and an analysis of the convergence was made by Parlett and Poole [9], and by Stewart [16]. We would like to show how our results of section 1 can be applied here.

We will denote by $\Pi_m$ the orthogonal projector onto the subspace $\mathbf{S}_m$ and will assume that the eigenvalues of A are labelled in decreasing order of magnitude and that $|\lambda_{r+1}|<|\lambda_r|$ i.e.

$$|\lambda_1| \geq |\lambda_2| \geq |\lambda_3| \ldots\ldots \geq |\lambda_r| > |\lambda_{r+1}| \geq \ldots \geq \lambda_N$$

Again $u_i$ denotes an eigenvector of A of norm unity associated with $\lambda_i$. The spectral projector associated with the invariant subspace corresponding to $\lambda_1, .. \lambda_r$ will be denoted by P. We can then establish the following theorem concerning the distance $||(I-\Pi_m)u_i||$

**Theorem 8:** *Let $S_o = \{x_1, x_2, .. x_r\}$ and assume that $S_o$ is such that the system of vectors $\{P x_i\}_{i=1,..r}$ is linearly independent . Then for each $u_i$ $i=1,2..r$ there exists at least one vector $s_i$ in the subspace $\mathbf{S}_m = span\{A^m S_o\}$ such that $Ps_i = u_i$. Moreover the following inequality is satisfied:*

$$|| (I-\Pi_m)u_i || \leq || u_i - s_i || ( | \lambda_{r+1} / \lambda_i| + \epsilon_m )^m \qquad (23)$$

*where $\epsilon_m$ tends to zero as m tends to infinity.*

**Proof:** Let us write any vector s of $S_o$ as

$$s = \sum_{j=1}^{r} \xi_j x_j$$

Projecting this onto the invariant subspace associated with $\lambda_1, .. \lambda_r$ we get

$$Ps = \sum_{j=1}^{r} \xi_j \, Px_j$$

Since the $Px_j$'s are assumed linearly independent and since $u_i$ belongs to the invariant subspace associated with $\lambda_1,..\lambda_r$, there exists at least one vector $s_i$ such that $Ps_i=u_i$. Then the vector $s_i$ is such that

$$s_i=u_i+w \tag{24}$$

where $w=(I-P)s_i$. Note that $s_i$ is not unique, since adding to $s_i$ any vector of the intersection of $\mathbf{S}_o$ and $(I-P)\mathbf{C}^N$ would still give a vector satifying the requirement $Ps_i=u_i$. Next consider the vector y of $\mathbf{S}_i$ defined by $y=(1/\lambda_i)^m A^m s_i$. We have from (24) that

$$u_i - y = (1/\lambda_i)^m A^m w \tag{25}$$

Denoting by W the invariant subspace corresponding to the remaining eigenvalues $\lambda_{r+1},..\lambda_N$ , and noticing that w belongs to W, we clearly have

$$u_i - y = (1/\lambda_i)^m \, [A_{|W}]^m \, w$$

Hence

$$|| \, u_i - y \, || \leq \; || \, [\lambda_i^{-1} \, A_{|W}]^m|| \; || \, w \, || \tag{26}$$

Since the eigenvalues of $A_{|W}$ are $\lambda_j$ with $j>r$, the spectral radius of $\lambda_i^{-1} \, A_{|W}$ is simply $|\lambda_{r+1}/\lambda_i|$ and from a well known result [17] we have

$$|| \, [\lambda_i^{-1} \, A_{|W}]^m|| \; = \; [ \; |\lambda_{r+1}/\lambda_i| + \epsilon_m \, ]^m \tag{27}$$

where $\epsilon_m$ tends to zero as $m \rightarrow \infty$. Using the fact that

$$|| \, (I-\Pi_m)u_i \, || = \min \{ \; || \, s - u_i \, ||, \, s \in \mathbf{S}_m \; \} \; ,$$

the inequality (26), and equality (27) yield the desired result (23). □

A few remarks are in order. First notice that we can take advantage of the nonuniqueness of the vector $s_i$ to improve the bound (23) by replacing $s_i$ in the theorem by the vector $\bar{s}_i$ satisfying $Ps_i=u_i$ for which the norm $|| \, s_i - u_i \, ||$ is minimum.

A second remark concerns the sequence $\epsilon_m$ for which we can be more specific by using the more precise bound for $|| \, B^m \, ||$ given in [17]

$$|| \, B^m \, || \leq \alpha \, \rho^m \, m^{\eta-1} \tag{28}$$

where B is any matrix, $\rho$ represents the spectral radius of B, $\eta$ is the dimension of its largest Jordan block, and $\alpha$ is some constant ***independent of m***. We may assume without loss of generality that $\alpha \geq 1$ (otherwise we can consider the weaker bound obtained by replacing $\alpha$ by $\alpha'=1$). Thus in the case where A is diagonalizable[1] we have $\eta=1$ , and after taking the m-th root of both sides of (28) the above inequality (23) becomes simply

$$\| (I-\Pi_m)u_i \| \leq \alpha \| u_i - s_i \| ( | \lambda_{r+1} / \lambda_i | )^m \tag{29}$$

In the diagonalizable case $s_i$ is a vector having the eigenexpansion

$$s_i = u_i + \sum_{j=r+1}^{N} \xi_j u_j$$

and letting $\beta = \sum_{j=r+1}^{N} | \xi_j |$ the proof of theorem 8 can be repeated to yield the inequality

$$\| (I-\Pi_m)u_i \| \leq \beta ( | \lambda_{r+1} / \lambda_i | )^m$$

In case $A_{|W}$ is not diagonalizable, then fron the result (28) we can majorize $\epsilon_m$ as follows:

$$\epsilon_m \leq |\lambda_i/\lambda_{r+1}| \, [\alpha^{1/m} \, m^{(\eta-1)/m} - 1 \, ]$$

which tends to zero as m tends to infinity.

Finally we would like to interpret the assumption of the theorem. It can be easily shown that the assumption that $\{Px_i\}$ is a linearly independent system, is equivalent to the condition

$$\det(U^H S_o) \neq 0$$

in which U is any basis of the invariant subspace. Clearly this is a generalization of a similar condition required for the convergence of the power method.

An experiment described in [11] indicated that even in the unsymmetric case the Chebyshev polynomials can often be efficiently used to accelerate the convergence of the subspace iteration. Let us assume that we can find an ellipse of center d and eccentrencity e which contains all the eigenvalues of A except the r domiminant ones $\lambda_1,...\lambda_r$, see figure 3-1

Then Rutishauser's symmetric subspace iteration can be generalized to unsymmetric problems by

[1] In fact we only need that the restriction of A to W defined in the above proof to be diagonalizable.

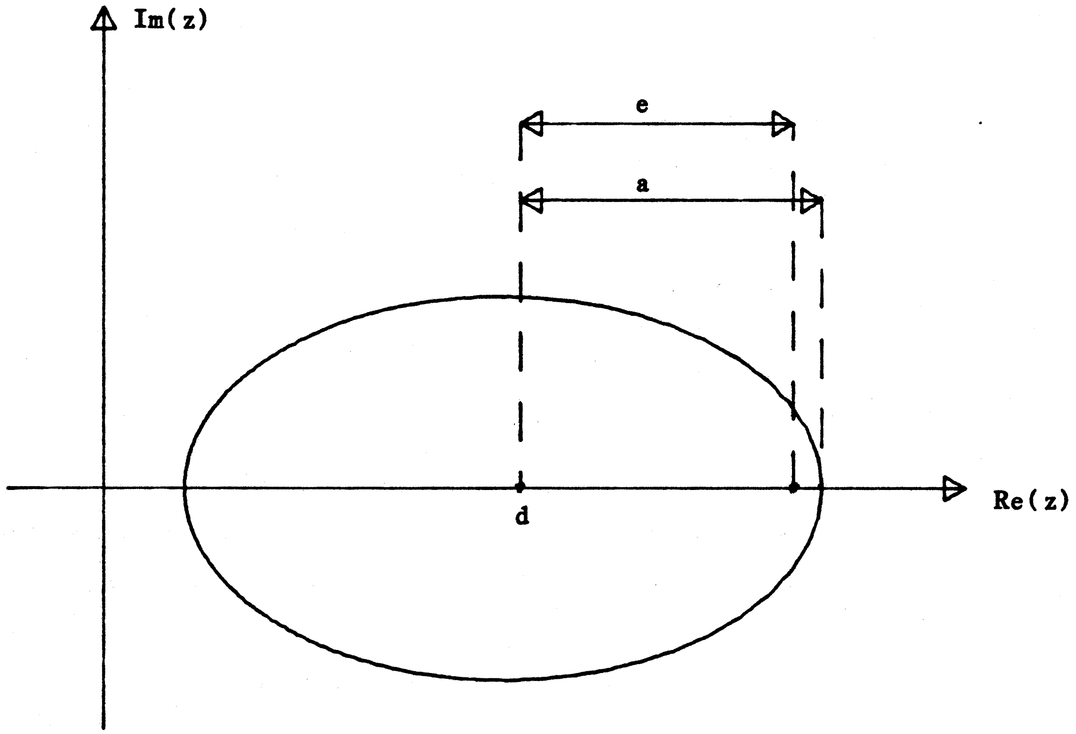

**Figure 3-1:** Optimal ellipse for the unsymmetric subspace iteration

simply replacing the subspace $\mathbf{S}_m$ by the richer subspace $\mathbf{R}_m$ defined by

$$\mathbf{R}_m = \text{span}\{ C_m(A)S_0 \}$$

in which $C_m$ is the shifted Chebyshev polynomial

$$C_m(z) = T_m[(z-d)/e]$$

Assuming that A is diagonalizable then theorem 8 can be extended as follows.

**Theorem 9:** *Let $S_o = \{x_1, x_2,.. x_r\}$ be such that the vectors $\{Px_j, j=1,..r\}$ are linearly independent . Assume that the eigenvalues $\lambda_{r+1}, \lambda_{r+2}..\lambda_N$ are contained in an ellipse E of center d, eccentricity e and major semi axis a. Then for each $u_i$ $i=1,2..r$ there exists at least one vector $s_i$ in the subspace $\mathbf{R}_m$=span$\{C_m(A)S_o\}$ such that*

$$s_i = u_i + \sum_{j=r+1}^{N} \xi_j u_j \tag{30}$$

*Moreover letting $\beta = \sum_{j=r+1}^{N} | \xi_j |$, the following inequality is satisfied:*

$$|| (I-\Pi_m)u_i || \leq \beta \frac{T_m(a/e)}{|T_m[(\lambda_i-d)/e)|} \tag{31}$$

**Proof:** The existence of $s_i$ defined by (30) can be proved in the same way as for theorem 8. Proceeding as in the proof of theorem 8, consider the vector y of $\mathbf{R}_m$ defined by

$$y=(C_m(A)/C_m(\lambda_i)\ )\ s_i$$

Then it is clear that

$$y-u_i = (1/C_m(\lambda_i)) \sum_{j=r+1}^{N} C_m(\lambda_j)\ )\xi_j\ u_j$$

Taking the norm of both sides we obtain the bound

$$|| y-u_i || \leq\ |1/C_m(\lambda_i)| \sum_{j=r+1}^{N} |\ C_m(\lambda_j)\ |\ |\xi_j| \tag{32}$$

$$\leq \max_{z \in E}\ |C_m(z)/C_m(\lambda_i)| \sum_{j=r+1}^{N} |\xi_j|$$

where E is the ellipse containing the eigenvalues $\lambda_{r+1},...\lambda_N$. It is easy to show that the above maximum is achieved for m different points on the boundary of the ellipse including the point on the major semi axis d+a. Replacing this in (32) the proof can be completed in the same way as for theorem 8 □

Again as in theorem 8, because of the nonuniqueness of $s_i$, the constant $\beta$ can be replaced by the smallest possible $\beta$:

$$\overline{\beta} = \min\{ \sum_{j=r+1}^{N} \mid \xi_j \mid ; \text{ all } \xi_j \text{ such that } u_i+\sum_{j=r+1}^{N} \xi_j u_j \in \mathbf{R}_m \}$$

The above bound is a generalization of Rutishauser's result. In the case where all eigenvalues of A are real we can take for E the degenerate ellipse which has eccentricity e=d, i.e. E is the interval [d-e,d+e] and a=e. In this case, assuming the eigenvalues are labelled in increasing order, the numerator of (31) becomes one and the denominator can be written as:

$$T_m(1+2\gamma_r)$$

with

$$\gamma_r=(\lambda_i-\lambda_{r+1})/(\lambda_{r+1}-\lambda_N)$$

This is precisely the result obtained by Rutishauser in the symmetric case, see [7]. Note that generally the result is more difficult to interpret in the presence of complex eigenvalues . It can be shown that the right hand side of (31) always tends to zero, see [19, 6]. When the ellipse is flat, i.e. when the eigenvalues have small imaginary parts , then the convergence will be faster because a/e is closer to one. The ideal case is when al! eigenvalues are real.

## 4. Methods using Krylov subspaces.

There are several techniques which realize a projection method on Krylov subspaces of the form $\mathbf{K}_m = \text{span}\{v_1,Av_1,..A^{m-1}v_1\}$. Unlike in the subspace iteration method where the dimension of the subspace of approximation is fixed during the iteration, here the dimension of $\mathbf{K}_m$ increases by one at every step, i.e. dim( $\mathbf{K}_m$ )=m .

Among the methods which use Krylov subspaces, we mention the following:

- The symmetric Lanczos algorithm, see e.g. [7].
- The method of Arnoldi for unsymmetric systems, [1, 11].
- The unsymmetric Lanczos method [5, 8]
- The method of incomplete orthogonalization [11, 14].

The first two methods are orthogonal projection methods while the last two are oblique projection methods.

We now show a characteristic property for all techniques which realize an orthogonal projection method onto the Krylov subspace $\mathbf{K}_m$.

**Theorem 10:** *Assume that an orthogonal projection technique is applied to A using the subspace* $\mathbf{K}_m$ *and let* $\overline{p}_m(t)$ *be the characteristic polynomial of the approximate problem. Then* $\overline{p}_m$ *minimzes the norm* $\| p(A)\, v_1 \|$ *over all monic polynomials p of degree m.*

**Proof:** By Cayley Hamilton's theorem, we have $\overline{p}_m(A_m) = 0$ so that clearly

$$(\overline{p}_m(A_m)v_1, v) = 0 \quad \text{for any v in } \mathbf{K}_m \tag{33}$$

It can easily shown by induction that for $k \leq m$ we have the property

$$(A_m)^k\, v_1 = \Pi_m\, A^k\, v_1 \tag{34}$$

Therefore (33) becomes

$$(\Pi_m\, \overline{p}_m(A)v_1\, , v\,) = 0, \quad \forall\ v \text{ in } \mathbf{K}_m$$

or

$$(\overline{p}_m(A)v_1\, ,\Pi_m\ \ v\,) = 0, \quad \forall\ v \text{ in } \mathbf{K}_m$$

which is equivalent to

$$(\overline{p}_m(A)v_1\, , v\,) = 0 \quad \forall\ v \text{ in } \mathbf{K}_m$$

Now writing $\overline{p}_m(t)$ as $\overline{p}_m(t) = t^m - q(t)$, where q is a polynomial of degree less than m, we obtain the equality

$$(A^m v_1 - q(A)v_1\, , v\,) = 0 \quad \forall\ v \text{ in } \mathbf{K}_m \tag{35}$$

which is equivalent to

$$(A^m v_1 - q(A)v_1\, , A^j v_1\,) = 0 \quad j{=}0,1,2....m\text{-}1$$

In the above system of equations we recognize the normal equations for minimizing the Euclidean norm of $A^m v_1\text{-}s(A)v_1$ over all polynomials s of degree $\leq$m-1 and the result is proved. □

The above characteristic property was shown in the particular context of the Lanczos algorithm for symmetric problems in [15]. What we have just shown is that it holds for any orthogonal projection method onto a Krylov subspace $\mathbf{K}_m$ and that it is independent of the particular algorithm applied. It indicates that the method can be regarded as an optimization process whereby we attempt to mininize some norm of the minimal polynomial of $v_1$. Indeed under the assumption that the minimal polynomial of $v_1$ is of degree at least m the $\| p(A)\ v_1 \|$ can be regarded as a discrete norm on the set of polynomials of degree not exceeding m-1.

Let us now consider the distance of a particular exact eigenvector $u_i$ from the subspace of approximation $\mathbf{K}_m$. It is simplifying to assume that A is diagonalizable and to denote by $\epsilon_i^{(m)}$, the quantity

$$\epsilon_i^{(m)} = \min_{p \in P_{m-1}} \max_{\lambda \in \sigma(A)-\{\lambda_i\}} |p(\lambda)|$$

where $\mathbf{P}_{m-1}$ represents the set of all polyomials p of degree not exceeding m-1 such that $p(\lambda_i)=1$.

It can be easily shown that $\| (I-\Pi_m)u_i \|$ is related to $\epsilon_i^{(m)}$ by the following inequality, see [11]

$$\| (I-\Pi_m)u_i \| \leq \| v_1 \|_1 \, \epsilon_i^{(m)}$$

where $\| x \|_1$ is the norm defined as the sum of the absolute values of the components of x in the eigenbasis, assuming the eigenvectors are all of norm unity. This means that we will obtain an estimate for $\| (I-\Pi_m)u_i \|$ from an estimate of $\epsilon_i^{(m)}$.

Without loss of generality we assume that i=1. In [13] a result similar to the following one was stated without proof:

**Theorem 11:** *Let m<N. Then there exists m eigenvalues which can be numbered $\lambda_{2}, \ldots \lambda_{m+1}$ such that*

$$\epsilon_1^{(\blacksquare)} = [ \sum_{j=2}^{m+1} \prod_{\substack{k=2 \\ k \neq j}}^{m+1} \frac{|\lambda_k - \lambda_1|}{|\lambda_k - \lambda_j|} ]^{-1}$$

A proof of this equality is proposed in the appendix.

Let us suppose that all of the eigenvalues except $\lambda_1$, lie inside a certain cercle see figure (4-1). Then as m increases $\epsilon_1^{(m)}$ becomes rapidely smaller. This can be made clear by only considering the product term in (11) associated with the eigenvalue nearest to $\lambda_1$, called $\lambda_j$ in figure (4-1). As is seen in fig. (4-1), we will have in general $|\lambda_1 - \lambda_k| > |\lambda_j - \lambda_k|$ and therefore the product

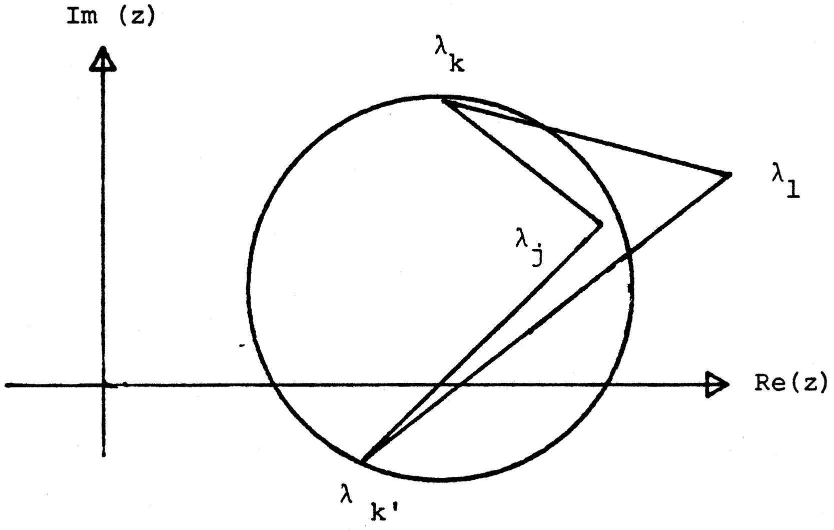

**Figure 4-1:** Illustration of theorem 11

$$\prod_{k=2}^{m+1} \frac{|\lambda_k - \lambda_1|}{|\lambda_k - \lambda_j|}$$

in (11) will be large in general, thus showing that $\epsilon_1^{(m)}$ will be a small quantity. Unfortunately we do not known in general what the eigenvalues $\lambda_2,.. \lambda_{m+1}$ are. The interesting indication provided by the theorem is that the convergence of $\epsilon_1^{(m)}$ towards zero is fastest when $\lambda_1$ is the outest part of the spectrum. We now propose a few illustrations.

**Example 1.** Assume that $\lambda_k = (k-1)/(N-1)$, $k=1,2,..N$ (Uniform distribution), and take $m=N-1$. Then $\epsilon_1^{(m)} = 1/(2^m - 1)$

**Proof:** Since m=N-1 there is no choice for the $\lambda_j$'s but the remaining eigenvalues $\lambda_2,...\lambda_N$. We have from (11)

$$\begin{aligned}
[\epsilon_1^{(m)}]^{-1} &= \sum_{j=2}^{m+1} \prod_{k=2:m+1,\, k\neq j} |k-1|/|k-j| \\
&= \sum_{j=2}^{m+1} \frac{m!}{(j-1)!(m+1-j)!} \\
&= \sum_{l=1}^{m} \binom{l}{m} = 2^m - 1
\end{aligned}$$

**Example 2.** Let again $m=N-1$ and assume the following distribution of the eigenvalues: $\lambda_k = e^{i\,2(k-1)\pi/N}$, $k=1,2..N$. Then $\epsilon_1^{(m)} = 1/m$

**Proof:** Here we have

$$\prod_{k=2:m+1,\, k\neq j} \frac{|\,w^{k-1}-1\,|}{|w^{k-1}-w^{j-1}|} = \prod_{k=1:m,\, k\neq j} \frac{|\,w^k-1|}{|w^k-w^j|}$$

$$\begin{aligned}
&= [|w^j-1| \prod_{k=1:m,\, k\neq j} |w^k-1|] \;/\; [\;|w^j-1| \prod_{k=1:m,\, k\neq j} |w^k-w^j|\;] \\
&= [\prod_{k=1}^{m} |w^k-1|\,] \;/\; [\,|w^j-1| \,.\, \prod_{k=1:m,\, k\neq j} |w^{k-j}-1|\;]
\end{aligned}$$

Recalling that the $w^k$'s are the powers of the N-th root of unity, a simple renumbering of the products in the denominator of the above expression shows that this expression has modulus one. Thus taking the sum over the m different eigenvalues yields the result □

In [13] we have shown a number of ways of bounding a quantity similar to $\epsilon_1^{(m)}$ which occurs when solving a system of linear equations by projection methods onto Krylov subspaces. Similar results hold for the dominant eigenvalue $\lambda_1$, in the context of the eigenvalue problem. For example we have

**Proposition 12:** *Assume that all the eigenvalues of A except $\lambda_1$ lie inside the ellipse having center d, focii d+e,d-e and major semi axis a. Then*

$$\epsilon_1^{(m)} \leq \frac{T_{m-1}(a/e)}{|T_{m-1}[(\lambda_1-d)/e]\,|}$$

*where $T_{m-1}$ is the Chebyshev polynomial of degree m-1 of the first kind.*

Note that a/e is a real positive number although for generality e may be complex , in case the main axis of the ellipse is not on the real line.

## 5. APPENDIX: Proof of theorem 11.

In this appendix we propose a proof of theorem 11. We need the following lemma from approximation theory, see [10].

**Lemma 13:** *Let $\tilde{q}$ be the best uniform approximation of a function f by a set of n polynomials satisfying the Haar conditions, on a compact set $\sigma$ consisting of at least n+1 points. Then there exists at least n+1 critical points $\lambda_{0},..\lambda_n$ of $\sigma$ such that if we set $\tilde{e}(z)=f(z)-\tilde{q}(z)$ we have*

$$|\tilde{e}(\lambda_j)| = \max_{s \in \sigma} \quad |\tilde{e}(z)|$$

Consider

$$\epsilon_i^{(n)} = \{ \min_{p \in P_{n-1}} \max_{\lambda \in \sigma(A)-\{\lambda_i\}} |p(\lambda)|$$

Clearly $\epsilon_i^{(n)}$ represents the smallest possible uniform norm on the set $\sigma(A)$ of polynomials of the form $1-(z-\lambda_1)s(z)$ with s of degree not excceeding n-2. Otherwise stated this means that we seek for the best approximation of the constant function unity over the set $\sigma(A)$ by polynomials of degree $\leq$n-1 which are linear combinations of the polynomials $\omega_1,\omega_2,..\omega_n$ where

$$\omega_j(z)=(z-\lambda_1)\, z^{j-1} \tag{36}$$

Since the set of polynomials (36) verifies the Haar conditions, from the above lemma there exists at least n+1 critical points, i.e. points where the maximum error is reached. We will denote by $\tilde{p}(z)$ the optimal polynomial $1-\tilde{q}$.

We can easily prove

**Lemma 14:** *Let $\lambda_{2}$, $\lambda_{3}$...$\lambda_{n+2}$ be the n+1 critical points. Then there exists a nontrivial solution to the system of equations:*

$$\sum_{i=2}^{n+2} \omega_j(\lambda_i) z_i = 0 \ , \ \ j=1,2...n \tag{37}$$

**Proof.**This is a system of n equations with n+1 unknowns. Because of the Haar conditions when we :solate one unknown, e.g. the last one, then the n by n resulting system is nonsingular □

**Lemma 15:** *Let $z_j$, j=2,...n+2 a certain solution of the system (37) and let us write $z_k=\delta_k e^{-i\theta_k}$, where $\delta_k$ is real positive. Then the best approximation polynomial $\tilde{p}$ is :*

$$\tilde{p}(z) = \sum_{k=2}^{n+2} e^{i\theta_k}\, l_k(z) \,/ \sum_{k=2}^{n+2} e^{i\theta_k}\, l_k(\lambda_1) \tag{38}$$

*where $l_k(z)$ is the Lagrange polynomial of degree $n$ at the points $\lambda_2, \lambda_3..\lambda_{n+2}$, taking the value one at $\lambda_k$:*

$$l_k(z) = \prod_{j=2:n+2,\, j\neq k} \frac{z - \lambda_j}{\lambda_k - \lambda_j}$$

**Proof:** Because of (37) for any v belonging to the space of polynomials $Q_n = \text{span}\{\omega_1, .. \omega_n\}$ we have:

$$\sum_{k=2}^{n+2} \delta_k\, e^{-i\theta_k}\, v(\lambda_k) = 0 \tag{39}$$

Let $\tilde{p}$ be defined by (38). We have to show that

$$\| \tilde{p} + v \|_\infty \geq \| \tilde{p} \|_\infty \quad \text{for any } v \text{ in } Q_n \tag{40}$$

where $\|.\|_\infty$ represents the uniform norm on the set $\sigma$. Let us set

$$\rho = [\sum_{k=2}^{n+2} e^{i\theta_k}\, l_k(\lambda_1)]^{-1} \tag{41}$$

Notice that $|\rho|$ is precisely the uniform norm of $\tilde{p}$ in $\sigma$. From (39) it is clear that for some k' we have

$$\text{Re}\, [\rho\, e^{-i\theta_{k'}}\, v(\lambda_{k'})\,] \geq 0$$

Therefore

$$\begin{aligned} \| \tilde{p} + v \|_\infty^2 &= \max_{j=2,n+2} \{ \quad (\tilde{p} + v)(\lambda_j)\}^2 \geq |\tilde{p}(\lambda_{k'}) + v(\lambda_{k'})|^2 \\ &= |\tilde{p}(\lambda_{k'}) + v(\lambda_{k'})|^2 = |\rho\, e^{i\theta_{k'}} + v(\lambda_{k'})|^2 \\ &= |\rho|^2 + |v(\lambda_{k'})|^2 + 2\,\text{Re}\,\{\rho e^{-i\theta_{k'}}\, v(\lambda_{k'})\} \quad \geq |\rho|^2 \end{aligned}$$

which shows that (40) is true and completes the proof of the lemma □

**Proof of theorem 11.** The system (37) can be solved by using Cramer's rule and some Vandermonde determinant equalities. Doing this it is possible to show that one particular solution of (37) is $z_k = l_k(\lambda_1)$, k=2,..n+2. Hence

$$e^{i\theta_k} = \overline{l_k}(\lambda_1) \,/\, |\, l_k(\lambda_1)\, |$$

replacing this in the expression (41) gives the desired result □

# *References*

[1] W.E. Arnoldi.
The principle of minimized iteration in the solution of the matrix eigenvalue problem.
*Quart. Appl. Math.* 9:17-29, 1951.

[2] F. Chatelin.
*Spectral approximation of linear operators.*
Academic Press, New york, 1983.

[3] M. Clint and A. Jennings.
The evaluation of eigenvalues and eigenvectors of real symmetric matrices by simultaneous iteration method.
*J. Inst. Math. Appl.* 8:111-121, 1971.

[4] W. Kahan, B.N. Parlett and E. Jiang.
Residual bounds on approximate eigensystems of nonormal matrices.
*SIAM J. on Numer. Anal.* : , .
(to appear).

[5] C. Lanczos.
An iteration method for the solution of the eigenvalue problem of linear differential and integral operator.
*J. Res. Nat. Bur. of Standards* 45:255-282, 1950.

[6] T.A. Manteuffel.
*An iterative method for solving nonsymmetric linear systems with dynamic estimation of parameters.*
Technical Report UIUCDCS-75-758, University of illinois at Urbana-Champaign, 1975.
Ph.D. dissertation.

[7] B.N. Parlett.
*The Symmetric Eigenvalue Problem.*
Prentice Hall, Englewood Cliffs, 1980.

[8] B.N. Parlett and D. Taylor.
*A look ahead Lanczos algorithm for unsymmetric matrices.*
Technical Report PAM-43, Center for Pure and Applied Mathematics, 1981.

[9] B.N. Parlett and W.G. Poole.
A geometric theory of the QR, LU and Power iterations.
*SIAM j. of Num. Anal.* 10 #2,:389-412, 1973.

[10] Rivlin T.J.
*The Chebyshev Polynomials.*
J.Wiley and sons Inc., New York, 1976.

[11] Y. Saad.
Variations on Arnoldi's method for computing eigenelements of large unsymmetric matrices.
*Linear Algebra and its Applications* 34:269-295, 1980.

[12] Y. Saad.
On the rates of convergence of the Lanczos and the block Lanczos methods.
*SIAM J. Numer. Anal.* =17:687-706, 1980.

[13] Y. Saad.
Krylov subspace methods for solving large unsymmetric linear systems.
*Mathematics of Computation* 37:105-126, 1981.

[14] Y. Saad.
The Lanczos biorthogonalization algorithm and other oblique projection methods for solving large unsymmetric systems.
*SIAM J. on Numer. Anal.* 19 # 3: , 1982.

[15] Y. Saad.
*Computation of eigenvalues of large Hermitian matrices by Partitioning techniques.*
Technical Report, INPG- University of Grenoble, 1974.
Dissertation.

[16] G.W. Stewart.
Simultaneous iteration for Computing invariant subspaces of non-Hermitian matrices.
*Numer. Mat.* 25:123-136, 1976.

[17] R.S. Varga.
*Matrix Iterative Analysis.*
Prentice Hall, Englewood Cliffs, New Jersey, 1962.

[18] J. H. Wilkinson.
*The Algebraic Eigenvalue Problem.*
Clarendon Press, Oxford, 1965.

[19] H.E. Wrigley.
Accelerating the Jacobi method for solving Simultaneous equations by Chebyshev extrapolation when the eigenvalues of the Iteration Matrix are complex.
*Computer Journal* 6:169-176, 1963.