Numer. Math. 28, 307 – 327 (1977)

Numerische Mathematik

The Tchebychev Iteration for Nonsymmetric Linear Systems

Thomas A. Manteuffel
Division 8325, Sandia Laboratories, Livermore, CA 94550, USA

Summary. In this paper an iterative method for solving nonsymmetric linear systems based on the Tchebychev polynomials in the complex plane is discussed. The iteration is shown to converge whenever the eigenvalues of the linear system lie in the open right half complex plane. An algorithm is developed for finding optimal iteration parameters as a function of the convex hull of the spectrum.

Subject Classifications. AMS (Mos): 65 F 10; CR: 5.14.

1. Introduction

Mathematical models of physical phenomena often yield large, sparse linear systems

$$Ax=b \tag{1.1}$$

where A is an $n\times n$ real valued matrix and x and b are $n\times 1$ vectors. Frequently, these systems are nonsymmetric. A standard technique for solving nonsymmetric systems is to consider the equivalent system

$$A^TAx=A^Tb,$$

and apply techniques available for solving positive definite systems. However, the condition of the matrix A^TA may be much worse than the condition of A.

This paper will discuss an iterative method for solving large, sparse, nonsymmetric systems whose eigenvalues lie in the right (left) half plane. This includes the important special case when the symmetric part of A,

$$M=1/2(A+A^T)$$

is positive (negative) definite. The method is based upon the Tchebychev polynomials in the complex plane.

The Tchebychev algorithm has the following properties:

1. The method does not depend upon the nonzero structure of the matrix A.

2. The method is sensitive to the condition of A rather than of A^TA.

3. The method requires only one matrix vector multiplication per step.

4. The method can be used in conjunction with factorization and splitting methods.

Few iterative methods have been developed to treat nonsymmetric systems. None share these properties with the Tchebychev algorithm.

It will be shown in Section 2 that the method will converge whenever the spectrum of A can be enclosed in an ellipse that does not contain the origin. It will also be shown that the Tchebychev iteration is optimal, in a certain sense, over all other polynomial based methods. Section 3 will show how optimal iteration parameters may be found as the solution to a mini-max problem of two real variables, while Section 4 will provide an algorithm for solving the min-max problem. An adaptive procedure for dynamically estimating optimal iteration parameters as well as numerical results will be presented in a subsequent paper.

Work has been done by Kjellberg [15] and Young and Edison [25, 26] on choosing the optimal SOR parameter, ω, when the eigenvalues of the Jacobi matrix are complex. The methods used here for finding the optimal Tchebychev parameters are similar. In this case, however, the mini-max problem involves different functions and two parameters. Wachspress [21], Wrigley [24], Kincaid [13, 14], and Hageman [9] have considered the Tchebychev iteration in the complex plane. This paper finds the optimal iteration parameters in terms of the eigenvalues of the matrix A. Many of the ideas for this work came from Diamond [3], Hestenes [10], Rutishauser [5], Stiefel [5, 10, 18], Golub [8], and Varga [8, 19].

2. Tchebychev Iteration

2.1. General Polynomial Based Iteration

In the remainder of this paper we will be concerned with solving the system (1.1). It will be assumed that the eigenvalues of A, λ_i, lie in the open right half plane. Later it will be assumed that the matrix A is real. That assumption is not necessary to the preliminary discussion which is presented as a motivation for the use of polynomial based methods (Rutishauser [5]).

If x_0 is the initial guess at the solution x, then we can define an iteration with general step

$$x_n = x_{n-1} + \sum_{i=1}^{n-1} \gamma_{ni} r_i, \tag{2.1}$$

where the γ_{ij}'s are constants and

$$r_i = b - Ax_i$$

is the residual at step i. Let $e_i = x - x_i$ be the error at the i^{th} step. An inductive

argument yields

$$e_n = P_n(A)\, e_0 \tag{2.2}$$

where $P_n(z)$ is a polynomial of degree n such that $P_n(0)=1$.

Any sequence of polynomials, $\{P_i(z)\}_{i=1}^{n}$, with $P_i(0)=1$, can be generated by choosing the constants, $\{\gamma_{ij}\}$. We would like to choose the sequence of polynomials so that

$$\|e_n\| \leqq \|P_n(A)\|\, \|e_0\| \tag{2.3}$$

is small [1]. Let us examine $P_n(A)$. If A is diagonalizable, then the Jordan form of A is a diagonal matrix (Birkhoff and Maclane [2]). There exists a nonsingular S such that $A=S^{-1}JS$ and

$$J=\begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_k \end{pmatrix}.$$

Now

$$P_n(A)=P_n(S^{-1}JS)=S^{-1}P_n(J)S,$$

and since J is a diagonal matrix we have

$$P_n(J)=\begin{pmatrix} P_n(\lambda_1) & & \\ & \ddots & \\ & & P_n(\lambda_k) \end{pmatrix}.$$

This yields the following result:

Theorem 2.1. If A is diagonalizable, then $\|P_n(A)\| \to 0$ as $n \to \infty$ if and only if $P_n(\lambda_i) \to 0$ as $n \to \infty$ for every eigenvalue, λ_i, of A.

Suppose A is not diagonalizable; that is, suppose A has nonlinear elementary divisors. The Jordan form of A has nontrivial Jordan blocks. We have $A=S^{-1}JS$ where

$$J=\begin{pmatrix} J_1 & & \\ & \ddots & \\ & & J_k \end{pmatrix},$$

and

$$J_i=\begin{pmatrix} \lambda_i & 1 & & \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ & & & \lambda_i \end{pmatrix}$$

[1] Here $\|\cdot\|$ represents the l_2-norm of a vector and the corresponding *lub* norm of the operator (Householder [12])

is the Jordan block associated with the eigenvalue λ_i with invariant subspace of dimension d_i. In this case

$$P_n(A)=P_n(S^{-1}JS)=S^{-1}P_n(J)S,$$

and

$$P_n(A)=\begin{pmatrix} P_n(J_1) & & \\ & \ddots & \\ & & P_n(J_k) \end{pmatrix},$$

where

$$P_n(J_i)=\begin{pmatrix} P_n(\lambda_i) & P_n'(\lambda_i) & \frac{1}{2!}P_n''(\lambda_i) & \cdots & \frac{1}{d_i!}P_n^{(d_i-1)}(\lambda_i) \\ & \ddots & \ddots & \ddots & \vdots \\ & & \ddots & \ddots & \frac{1}{2!}P_n''(\lambda_i) \\ & & & \ddots & P_n'(\lambda_i) \\ & & & & P_n(\lambda_i) \end{pmatrix}$$

This yields the following theorem:

Theorem 2.2. If λ_i is an eigenvalue of A with invariant subspace of dimension d_i, then $\|P_n(A)\|\to 0$ as $n\to\infty$ if and only if $P_n^{(j)}(\lambda_i)\to 0$ as $n\to\infty$ for every $j<d_i$, for each eigenvalue λ_i.

When looking for a sequence of polynomials to suppress the eigenvalues of A, we must find one whose derivatives also suppress the eigenvalues of A.

In light of the previous discussion we can establish three criteria upon which to choose a sequence of polynomials:

1. We must choose $P_n(z)$, among polynomials of like degree such that $P_n(0)=1$, to be "as small as possible" on the spectrum of A.
2. If A has nonlinear elementary divisors, then we must choose $P_n(z)$ so that its derivatives are small on the spectrum of A.
3. We must choose $P_n(z)$ to have some recursive properties so that all of the previous residuals need not be stored.

In the following sections we will see that the scaled and translated Tchebychev polynomials fit these criteria.

2.2. *The Tchebychev Polynomials*

The Tchebychev polynomials were discovered a century ago by the Russian mathematician Tchebychev (the spelling of which has many variations). Their importance for practical computation, however, was rediscovered about forty years ago by C. Lanczos. Since then they have found many uses in numerical analysis (Fox and Parker [7]).

The Tchebychev polynomials are given by:

$$\begin{aligned} T_0(z)&=1, \\ T_1(z)&=z, \\ T_{n+1}(z)&=2zT_n(z)-T_{n-1}(z), \quad n>1. \end{aligned} \tag{2.4}$$

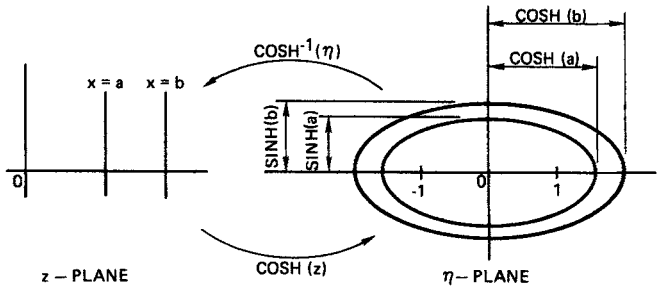

Fig. 1

They may also be written:

$$T_n(z) = \cosh(n \cosh^{-1}(z)). \tag{2.5}$$

Consider the map $\eta = \cosh(z)$. Let $z = x + iy$, $\eta = u + iv$ $(x, y, u, v \in \mathscr{R})$. Then, $\cosh(z) = \cosh(x+iy) = u + iv = \eta$. Using the expansion formula for the cosh, we have

$$\cosh(x+iy) = \cosh(x)\cos(y) + i\sinh(x)\sin(y),$$

or

$$u = \cosh(x)\cos(y),$$

$$v = \sinh(x)\sin(y).$$

Suppose we fix $x > 0$ and allow y to vary. Then u and v satisfy

$$\frac{u^2}{\cosh^2(x)} + \frac{v^2}{\sinh^2(x)} = 1.$$

That is, the line $x = \text{constant}$ is mapped onto an ellipse with semi-major axis $|\cosh(x)|$, semi-minor axis $|\sinh(x)|$, and foci at 1 and -1 (see Fig. 1). This map has period $2\pi i$. Since $\cosh(x)$ and $\sinh(x)$ are increasing for $x \geqq 0$, if $0 < a < b$, then the line $x = a$ is mapped onto an ellipse inside and confocal to the ellipse that the line $x = b$ is mapped onto (see Fig. 1). If $x = 0$ we have

$$u = \cos(y),$$

$$v = 0,$$

and the ellipse has collapsed onto the real line segment $[-1, 1]$.

Because of periodicity, the map $\eta = \cosh(z)$ takes the region

$$0 \leqq \mathrm{Re}(z)$$

$$0 \leqq \mathrm{Im}(z) < 2\pi \tag{2.6}$$

onto the entire η-plane. Each vertical line in this region is mapped onto an ellipse in the η-plane. This region is the branch of $\cosh^{-1}$ used in the definition of the Tchebychev polynomials.

The function $\cosh^{-1}$ may also be written in log form:

$$\cosh^{-1}(w) = \ln(w + (w^2 - 1)^{1/2}). \tag{2.7}$$

Care must be taken when choosing the branch of the square root. The branch chosen depends on the argument w and should be chosen so that $(w^2)^{1/2} = w$.

The n^{th} Tchebychev polynomial, $T_n(z) = \cosh(n \cosh^{-1}(z))$, maps an ellipse onto a vertical line segment in the region (2.6) above, multiplies this line segment by n, and maps the new line segment back onto another ellipse (see Fig. 1). Since the new line segment cuts through n branches of $\cosh^{-1}$, it is wrapped around the new ellipse n times. The degenerate ellipse, the line segment $[-1, 1]$, is mapped onto the line segment $[-1, 1]$. Since it is wrapped around n times, $T_n(z)$ has n zeros on the line segment $[-1, 1]$.

To establish some notation, let $\mathfrak{F}(d, c)$ be the family of ellipses in the complex plane centered at d with foci at $d+c$ and $d-c$. Let $F(d, c) \in \mathfrak{F}(d, c)$ be a member of this family. Let $F_i(d, c) \subset F_j(d, c)$ mean that the ellipse $F_i(d, c)$ is inside the ellipse $F_j(d, c)$. Let $z \in F_i(d, c)$ mean that the point z is on the ellipse $F_i(d, c)$. The Tchebychev polynomials the map members of $\mathfrak{F}(0, 1)$ onto other members of $\mathfrak{F}(0, 1)$. In this notation we have the following result:

Lemma 2.3. Suppose $z_i \in F_i(0, 1)$, $z_j \in F_j(0, 1)$; then

$$\mathrm{Re}(\cosh^{-1}(z_i)) < \mathrm{Re}(\cosh^{-1}(z_j)) \Leftrightarrow F_i(0, 1) \subset F_j(0, 1), \tag{2.8a}$$

$$\mathrm{Re}(\cosh^{-1}(z_i)) = \mathrm{Re}(\cosh^{-1}(z_j)) \Leftrightarrow F_i(0, 1) = F_j(0, 1). \tag{2.8b}$$

Now consider the scaled and translated Tchebychev polynomials,

$$P_n(\lambda) = \frac{T_n\left(\frac{d-\lambda}{c}\right)}{T_n\left(\frac{d}{c}\right)}, \tag{2.9}$$

where d and c are complex numbers. Notice that $P_n(0) = 1$ as is required for use in a polynomial based iterative method. Using the definition of the cosh, we see that if $\frac{d}{c} \notin [-1, 1]$, $\left(\frac{d-\lambda}{c}\right) \notin [-1, 1]$, then

$$P_n(\lambda) = \frac{e^{n \cosh^{-1}\left(\frac{d-\lambda}{c}\right)} + e^{-n \cosh^{-1}\left(\frac{d-\lambda}{c}\right)}}{e^{n \cosh^{-1}\left(\frac{d}{c}\right)} + e^{-n \cosh^{-1}\left(\frac{d}{c}\right)}}$$

$$\doteq e^{n\left(\cosh^{-1}\left(\frac{d-\lambda}{c}\right) - \cosh^{-1}\left(\frac{d}{c}\right)\right)}, \tag{2.10}$$

for large n. This motivates the following definition:

Definition. Let

$$r(\lambda) = \lim_{n \to \infty} |P_n(\lambda)^{1/n}| \tag{2.11}$$

be the *asymptotic convergence factor* of $P_n(\lambda)$ at the point λ.

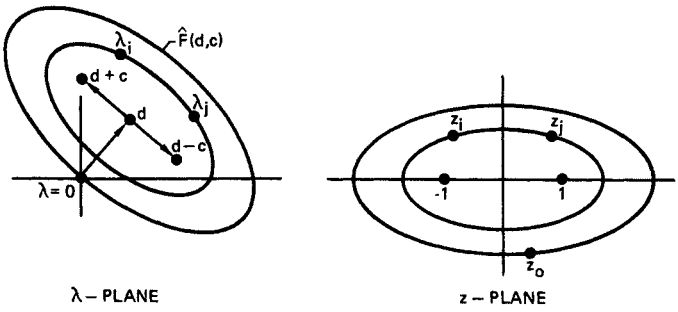

Fig. 2

The asymptotic convergence factor, which will be referred to as the convergence factor, is related to the rate of convergence as defined by Young [26] and Varga [20] in that the rate of convergence equals $-\ln(r(\lambda))$. From (2.10) we have

$$r(\lambda) = |e^{\cosh^{-1}\left(\frac{d-\lambda}{c}\right) - \cosh^{-1}\left(\frac{d}{c}\right)}|$$
$$= e^{\mathrm{Re}\left(\cosh^{-1}\left(\frac{d-\lambda}{c}\right) - \cosh^{-1}\left(\frac{d}{c}\right)\right)}. \tag{2.12}$$

The choice of d and c determines a family of ellipses, $\mathfrak{F}(d, c)$, with foci at $d+c$ and $d-c$ (see Fig. 2).

Let $\hat{F}(d, c) \in \mathfrak{F}(d, c)$ be the member of the family passing through the origin. The translation $z = \dfrac{d-\lambda}{c}$ maps members of $\mathfrak{F}(d, c)$ in the λ-plane onto members of $\mathfrak{F}(0, 1)$ in the z-plane. From (2.12) and Lemma 2.3 we have that if $\lambda_i \in F_i(d, c)$, $\lambda_j \in F_j(d, c)$, then

$$r(\lambda_i) < r(\lambda_j) \Leftrightarrow F_i(d, c) \subset F_j(d, c), \tag{2.13a}$$
$$r(\lambda_i) = r(\lambda_j) \Leftrightarrow F_i(d, c) = F_j(d, c), \tag{2.13b}$$
$$r(\lambda) = 1 \Leftrightarrow \lambda \in \hat{F}(d, c). \tag{2.13c}$$

Asymptotically, the polynomials $P_n(\lambda)$ in (2.9) map the ellipse $F_i(d, c)$ onto the circle of radius $r(\lambda_i)^n$ where λ_i is any point on $F_i(d, c)$. Thus, we have from (2.13)

$$\lim_{n\to\infty} P_n(\lambda) = \begin{cases} 0 & \text{if } \lambda \text{ is inside } \hat{F}(d, c) \\ \infty & \text{if } \lambda \text{ is outside } \hat{F}(d, c). \end{cases} \tag{2.14}$$

2.3. *Optimal Properties of the Tchebychev Polynomials*

The first criterion mentioned in Section 2.1 suggests that when choosing a sequence of polynomials upon which to base an iterative method, it is desirable to choose polynomials that are small on the spectrum of the matrix A. Since the spectrum of A is seldom exactly known, it is more practical to choose polynomials that are small on a region containing the spectrum of A. If the region is bounded by a circle or an

ellipse, the scaled and translated Tchebychev polynomials have certain optimal properties. Much is known of the optimal properties of the monic scaled and translated Tchebychev polynomials over all monic polynomials of like degree on regions bounded by ellipses (Hille [11]; Walsh [22]). Similar results, but not as strong, can be shown for polynomials *normalized at the origin.*

Definition. Let $S_n = \{\text{all polynomials, } s_n(\lambda), \text{ of degree } n \text{ such that } s_n(0) = 1\}$. The elements of S_n are said to be *normalized at the origin.*

Complex function theory yields the following useful result (Hille [11]).

Theorem 2.4. Let E be a closed and bounded infinite set in the complex plane. There exists a unique $t_n \in S_n$ such that

$$\max_{z \in E} |t_n(z)| = \min_{s_n \in S_n} \max_{z \in E} |s_n(z)|. \tag{2.15}$$

If the region E is bounded by an ellipse, a circle being a special case of an ellipse, then the maximum modulus of an analytic function will occur on the boundary. Using the notation of Section 2.2, let $F(d, c, a)$ be the member of the family $\mathfrak{F}(d, c)$ with semi-major axis $a > 0$. Instead of taking the maximum over the entire region we may take the maximum over the boundary, $F(d, c, a)$. The circle with center d and radius a is denoted $F(d, o, a)$. If the spectrum of A is contained in a region that is bounded by a circle that does not include the origin in its interior, we have the following well known result (Varga [19]; Manteuffel [16]).

Theorem 2.5. Suppose $F(d, o, a)$ does not include the origin in its interior; that is, $a \leqq |d|$. If $t_n \in S_n$ satisfies (2.15) with $E = F(d, o, a)$ then

$$t_n = \left(\frac{d - \lambda}{d}\right)^n. \tag{2.16}$$

If the region is bounded by an ellipse with real foci that does not contain the origin in its interior, we have the following result due to Clayton (Wrigley [24]).

Theorem 2.6. Let $o < c \leqq a \leqq d$. If $t_n \in S_n$ satisfies (2.15) with $E = F(d, c, a)$ then

$$t_n(\lambda) = P_n(\lambda) = \frac{T_n\left(\frac{d - \lambda}{c}\right)}{T_n\left(\frac{d}{c}\right)}, \tag{2.17}$$

the associated scaled and translated Tchebychev polynomial.

This result cannot be extended to d and c with complex values. For example, if $d > 0$, $c = i \cdot \frac{d}{10}$, then for $0 < a < d$

$$\max_{\lambda \in F(d, c, a)} |P_2(\lambda)| < \max_{\lambda \in F(d, c, a)} |P_3(\lambda)|.$$

However, it can be shown to be asymptotically true. Let

$$m(s_n) = \min_{\lambda \in F(d,c,a)} |s_n(\lambda)|, \tag{2.18a}$$

$$M(s_n) = \max_{\lambda \in F(d,c,a)} |s_n(\lambda)|. \tag{2.18b}$$

We have the following.

Lemma 2.7. Suppose $F(d, c, a)$ does not contain the origin in its interior. If $t_n \in S_n$ satisfies (2.15) with $E = F(d, c, a)$ and $P_n(\lambda)$ is igiven by (2.9). then,

$$m(P_n) \leqq M(t_n) \leqq M(P_n). \tag{2.19}$$

Proof. The second inequality is true by hypothesis. Suppose that

$$M(t_n) < m(P_n);$$

then

$$t_n(\lambda) < P_n(\lambda)$$

for every $\lambda \in F(d, c, a)$. By Rouché's theorem, the polynomial $P_n(\lambda) - t_n(\lambda)$ has the same number of zeros in the interior of $F(d, c, a)$ as $P_n(\lambda)$ does. $P_n(\lambda)$ has n zeros on the line segment joining the foci, $d+c$ and $d-c$. Notice that $P_n(0) - t_n(0) = 0$. Since $\lambda = 0$ is not in the interior of $F(d, c, a)$, $P_n(\lambda) - t_n(\lambda)$ is a polynomial of degree n with $n+1$ zeros. We can conclude that $P_n(\lambda) = t_n(\lambda)$, and the lemma is proved.

Theorem 2.8. Suppose $F(d, c, a)$ does not include the origin in its interior. If $t_n \in S_n$ satisfies (2.15) with $E = F(d, c, a)$ and $P_n(\lambda)$ is given by (2.9), then

$$\lim_{n \to \infty} [M(t_n)^{1/n}] = \lim_{n \to \infty} [M(P_n)^{1/n}]. \tag{2.20}$$

Proof. From Lemmma 2.7 we know that $m(P_n) \leqq M(t_n) \leqq M(P_n)$. It is sufficient to show that

$$\lim_{n \to \infty} [m(P_n)^{1/n}] = \lim_{n \to \infty} [M(P_n)^{1/n}].$$

By (2.13) all points on the ellipse $F(d, c, a)$ have the same convergence factor; thus, we have

$$r(\lambda) = \lim_{n \to \infty} [m(P_n)^{1/n}] = \lim_{n \to \infty} [M(P_n)^{1/n}]$$

for every $\lambda \in F(d, c, a)$. This proves the theorem.

Because of the nature of the cosh function, the asymptotic convergence factor is achieved very quickly; thus, the scaled and translated Tchebychev polynomials tend very quickly to the optimal polynomial in S_n.

As the focal length c approaches 0, the ellipse $F(d, c, a)$ is deformed into the circle $F(d, o, a)$. It is easy to show that the result for circles is compatible with the result for ellipses; that is,

$$\lim_{c \to 0} \frac{T_n\left(\dfrac{d-\lambda}{c}\right)}{T_n\left(\dfrac{d}{c}\right)} = \left(\frac{d-\lambda}{d}\right)^n. \tag{2.21}$$

2.4. Convergence of $P_n^{(j)}(\lambda)$

Recall from Section 2.1 that if the matrix A has nonlinear elementary divisors, the derivatives of the sequence of polynomials must also converge to zero on the eigenvalues of A. Since $\{P_n(\lambda)\}$ is a sequence of analytic functions which converge to zero uniformly on the closure of the interior of any member of the family $\mathfrak{F}(d, c)$ that does not include the origin, the drivatives also converge to zero. It can be shown that

$$|P_n^{(j)}(\lambda)| \leqq K(\lambda, j)\, n^j\, r(\lambda)^n \tag{2.22}$$

for large n, where $K(\lambda, j)$ is a constant depending only on λ and j (Manteuffel [16]). We can conclude that when the ellipse $\hat{F}(d, c)$ contains the spectrum of A in its interior, an iteration based upon the associated scaled and translated Tchebychev polynomials will converge, although more slowly, in spite of the presence of nonlinear elementary divisors.

2.5. The Tchebychev Iteration

In light of the recursive formulae (2.4) for the Tchebychev polynomials, a three term recorsion can be carried out that satisfies (2.2) where $P_n(\lambda)$ satisfy (2.9) (Stiefel [18]). Given x_0 and parameters d and c, let

$$x_1 = x_0 + \Delta_0,$$

$$\Delta_0 = \frac{1}{d} r_0, \qquad r_0 = b - A x_0,$$

$$x_{n+1} = x_n + \Delta_n,$$

$$\Delta_n = \alpha_n r_n + \beta_n \Delta_{n-1}, \qquad r_n = b - A x_n,$$

$$\alpha_n = \frac{2}{c} \frac{T_n\left(\frac{d}{c}\right)}{T_{n-1}\left(\frac{d}{c}\right)}, \qquad \beta_n = \frac{T_{n-1}\left(\frac{d}{c}\right)}{T_{n+1}\left(\frac{d}{c}\right)}. \tag{2.23}$$

Again appealing to the formulae (2.4) we can generate α_n and β_n recursively as follows:

$$\alpha_1 = \frac{2d}{2d^2 - c^2}, \qquad \beta_1 = d\alpha_1 - 1,$$

$$\alpha_n = \left[d - \left(\frac{c}{2}\right)^2 \alpha_{n-1}\right]^{-1}, \qquad \beta_n = d\alpha_n - 1. \tag{2.24}$$

If the spectrum of the matrix A lies in the right half plane, then it can be enclosed in an ellipse that does not contain the origin in its interior. The associated scaled and translated Tchebychev polynomials meet the criteria established in Section 2.1.

They have minimal maximum moculus properties on ellipses, their derivatives sequences also converge, and the iteration can be carried out by a three term recursion. The remainder of this paper will be devoted to implementing an iteration based upon the Tchebychev polynomials.

3. Optimal Parameters

The spectrum of the matrix A can be enclosed in many different ellipses. In fact, given any family of ellipses $\mathfrak{F}(d, c)$, there is some member of the family that contains the spectrum of A in its interior. If the spectrum of A lies in the interior of $\hat{F}(d, c)$, the member of the family $\mathfrak{F}(d, c)$ passing through the origin, then the iteration based on the associated scaled and translated Tchebychev polynomials will converge. We would like to choose $\mathfrak{F}(d, c)$ so that this convergence is optimal in some sense.

3.1. The Mini-Max Problem

Suppose d and c have been chosen. We have seen in Section 2.2 that each point in the λ-plane is associated with a convergence factor,

$$r(\lambda) = |e^{[\cosh^{-1}(\frac{d-\lambda}{c}) - \cosh^{-1}(\frac{d}{c})]}|.$$

If we use (2.7), the log form of the $\cosh^{-1}$, then

$$r(\lambda) = \left| \frac{\left(\frac{d-\lambda}{c}\right) + \left(\left(\frac{d-\lambda}{c}\right)^2 - 1\right)^{1/2}}{\left(\frac{d}{c}\right) + \left(\left(\frac{d}{c}\right)^2 - 1\right)^{1/2}} \right|$$

$$= \left| \frac{(d-\lambda) + ((d-\lambda)^2 - c^2)^{1/2}}{d + (d^2 - c^2)^{1/2}} \right|. \tag{3.1}$$

In particular, each eigenvalue, λ_i, is associated with the convergence factor $r(\lambda_i)$. One way to optimize the choice of d and c is to make the maximum $r(\lambda_i)$ as small as possible. The parameters d and c will then satisfy

$$\min_{d,c} \max_{\lambda_i} r(\lambda_i) = \min_{d,c} \max_{\lambda_i} \left| \frac{(d-\lambda_i) + ((d-\lambda_i)^2 - c^2)^{1/2}}{d + (d^2 - c^2)^{1/2}} \right|. \tag{3.2}$$

A more rigorous argument which yields this same mini-max problem is as follows. With a polynomial based iterative method the error is suppressed in accordance with the equation $e_n = P_n(A)\, e_0$. The following definition of rate of convergence is used by Young [26].

Definition. The *rate of convergence* of a polynomial based iterative method applied to the system $Ax = b$ is

$$R(A) = -\log(\lim_{n \to \infty} (\|P_n(A)\|^{1/n})). \tag{3.3}$$

We would like to choose d and c to make $R(A)$ as large as possible or, equivalently, to make $\lim_{n\to\infty}(\|P_n(A)\|^{1/n})$ as small as possible. Let

$$S(\lambda)=\frac{(d-\lambda)+((d-\lambda)^2-c^2)^{1/2}}{d+(d^2-c^2)^{1/2}}.$$

From (2.10) and (3.1) we have for $\left(\frac{d-\lambda}{c}\right)\notin[-1,1]$

$$P_n(\lambda)\doteq(S(\lambda))^n \tag{3.4}$$

for large n. Since $M(\lambda)$ is analytic (with trivial exceptions) in an open set containing the spectrum of A, there exists an operator $S(A)$. The eigenvalues of $S(A)$ are $S(\lambda_i)$ where λ_i is an eigenvalue of A (Dunford and Schwartz [4]). Since for λ such that $\left(\frac{d-\lambda}{c}\right)\in[-1,1]$ we have

$$P_n(\lambda)\doteq(S(\lambda))^n\doteq 0,$$

we can write

$$P_n(A)\doteq(S(A))^n$$

for large n; moreover

$$\lim_{n\to\infty}(\|P_n(A)\|^{1/n})=\lim_{n\to\infty}(\|S(A)^n\|^{1/n})$$
$$=\text{spectral radius of } S(A).$$

The spectral radius of $S(A)$ is

$$\max_{\lambda_i}|S(\lambda_i)|=\max_{\lambda_i} r(\lambda_i).$$

The choice of d and c which yields the optimal rate of convergence is the solution to the mini-max problem (3.2) above. Since $r(\lambda)$ is a function of d and c as well as λ, let us write the mini-max problem as

$$\min_{d,c}\max_{\lambda_i} r(\lambda_i,d,c). \tag{3.5}$$

3.2. Restrictions

In the remainder of the paper it will be assumed that A is a real valued matrix. In this section we will show that if A is a real valued matrix, the mini-max problem (3.5) can be restricted so that the maximum is taken over a subset of the eigenvalues and the minimum is taken over d and c such that d and c^2 are real, and (d, c^2) lies in a certain region of the real plane, $\mathscr{R}\times\mathscr{R}$.

If d and c are fixed, then the convergence factor associated with each eigenvalue will be determined by the member of the family $\mathfrak{F}(d,c)$ that it is on. The eigenvalue

that has the maximum convergence factor will be on a member of $\mathfrak{F}(d, c)$ that includes the rest of the spectrum in the closure of its interior. This motivates the following.

Definition. Let $H = \{\lambda_i | \lambda_i$ is a vertex of the smallest convex polygon enclosing the spectrum of $A\}$. We will refer to H as the *hull* of the spectrum.

In light of (2.13) and the fact that ellipses are convex it can be seen that the elements of H completely determine the mini-max problem (3.5); that is, for any d and c we have

$$\max_{\lambda_i} r(\lambda_i, d, c) = \max_{\lambda_i \in H} r(\lambda_i, d, c). \tag{3.6}$$

If A is a real valued matrix, then the eigenvalues of A are real or appear in complex conjugate pairs. The hull, H, of the spectrum is symmetric with respect to the real axis. It can be shown that if A is a real valued matrix and d and c satisfy (3.5) then the family $\mathfrak{F}(d, c)$ is symmetric with respect to the real axis. (Although this result seems geometrically obvious, the proof is very tedious because of the nature of the functions $r(\lambda_i, d, c)$ (Manteuffel [17]).) Such a family has foci that are either both real or are a complex conjugate pair. Since the foci are $d+c$ and $d-c$, then d is real and c is either real or pure imaginary. In either case c^2 is real. Notice in (3.1) that c appears only as c^2. Let

$$r(\lambda, d, c^2) = \left| \frac{(d-\lambda) + ((d-\lambda)^2 - c^2)^{1/2}}{d + (d^2 - c^2)^{1/2}} \right|. \tag{3.7}$$

Since the families $\mathfrak{F}(d, c)$ and $\mathfrak{F}(d, -c)$ have the same foci, $d+c$ and $d-c$, the parameters d and c^2 uniquely determine the family $\mathfrak{F}(d, c)$.

For A real we may restrict d and c^2 to real values. In addition we may ignore those values of d and c^2 for which convergence clearly does not occur.

Definition. Let $R = \{(d, c^2) | 0 < d$ and $c^2 < d^2\}$.

Lemma 3.1. If A is a real valued matrix with eigenvalues in the open right half plane the mini-max problem (3.5) can be written

$$\min_{(d, c^2) \in R} \max_{\lambda_i \in H} r(\lambda_i, d, c^2). \tag{3.8}$$

Proof. It is clear from the discussion above that d and c^2 may be restricted to the real numbers. By hypothesis, A has eigenvalues in the open right half plane. If $d \leqq 0$, then every eigenvalue of A would be outside $\hat{F}(d, c)$, the ellipse passing through the origin. Convergence could not occur for this choice of d. If $d > 0$ and $c^2 \geqq d^2$, then $c \geqq d$ and $d - c \leqq 0 \leqq d + c$. The family $\mathfrak{F}(d, c)$ has one foci on each side of the origin. The ellipse $\hat{F}(d, c)$ is the degenerate ellipse. Since $\hat{F}(d, c)$ has no interior, there is no region of convergence.

Since A has its eigenvalues in the open right half plane, there is some d and c^2 for which convergence will occur. The solution of the mini-max problem is in the set R, and the lemma is proved.

Notice that for $(d, c^2) \in R$ we have from (3.7)

$$r(\lambda, d, c^2) = r(\bar{\lambda}, d, c^2). \tag{3.9}$$

The maximum is completely determined by the eigenvalues with nonnegative imaginary part.

Definition. Let $H^+ = \{\lambda_i \in H \mid \mathrm{Im}(\lambda_i) \geqq 0\}$. We will refer to H^+ as the *positive hull* of the spectrum.

Lemma 3.2. If A is a real valued matrix with eigenvalues in the right half plane, then the mini-max problem can be written

$$\min_{(d,c^2)\in R} \max_{\lambda_i \in H^+} r(\lambda_i, d, c^2). \tag{3.10}$$

A further reduction of the set H^+ is possible. We would like to find the smallest set of eigenvalues that completely determines $\max_{\lambda_i} r(\lambda_i, d, c^2)$ when $(d, c^2) \in R$.

Definition. Let $K = \{\lambda_k \in H^+ \mid$ there exists $(d, c^2) \in R$ such that $r(\lambda_k, d, c^2) = \max_{\lambda_i} r(\lambda_i, d, c^2)\}$. The elements of K will be referred to as *key elements.*

Clearly, if $(d, c^2) \in R$, then

$$\max_{\lambda_i \in K} r(\lambda_i, d, c^2) = \max_{\lambda_i \in H^+} r(\lambda_i, d, c^2).$$

Criteria to determine when an engenvalue is in the set K are needed.

Lemma 3.3. If $\lambda_k \in K$, then one of the following is true:

1. $\mathrm{Re}(\lambda_k) \leqq \mathrm{Re}(\lambda_i)$ for every $\lambda_i \in \mathrm{H}^+$.
2. $\mathrm{Re}(\lambda_k) \geqq \mathrm{Re}(\lambda_i)$ for every $\lambda_i \in H^+$.
3. There exist $\lambda_l, \lambda_m \in H^+$ such that there is an ellipse, $F_k(d, c)$, with $(d, c^2) \in R$, passing through λ_k, λ_l, and λ_m, containing the spectrum of A in the closure of its interior.

Proof. Every point $(d, c^2) \in R$ is associated with a family of ellipses, $\mathfrak{F}(d, c)$. Let $F_k(d, c)$ be the member of $\mathfrak{F}(d, c)$ passing through λ_k. As (d, c^2) is moved through the region R, $F_k(d, c)$ is continuously deformed.

If $\lambda_k \in K$ then there is some point $(d_1, c_1^2) \in R$ such that

$$r(\lambda_k, d_1, c_1^2) = \max_{\lambda_i} r(\lambda_i, d_1, c_1^2).$$

Since $r(\lambda_k, d_1, c_1^2)$ is maximal, we know from Equations (2.13) that $F_k(d_1, c_1)$ contains the spectrum in the closure of its interior.

Suppose λ_k is the only eigenvalue on the ellipse $F_k(d_1, c_1)$. If λ_k does not satisfy 1. or 2. of the hypothesis, then there are at least two other eigenvalues in H^+, say λ_j and λ_n, such that

$$\mathrm{Re}(\lambda_j) < \mathrm{Re}(\lambda_k) < \mathrm{Re}(\lambda_n).$$

Consider deforming the ellipse $F_k(d_1, c_1)$ into the degenerate ellipse, the line segment connecting λ_k and $\bar{\lambda}_k$, by moving (d, c^2) through R (see Fig. 3a). The eigenvalues λ_j and λ_n are inside $F_k(d_1, c_1)$ but outside the degenerate ellipse. One of the intermediate ellipses must have passed through another eigenvalue. As (d, c^2)

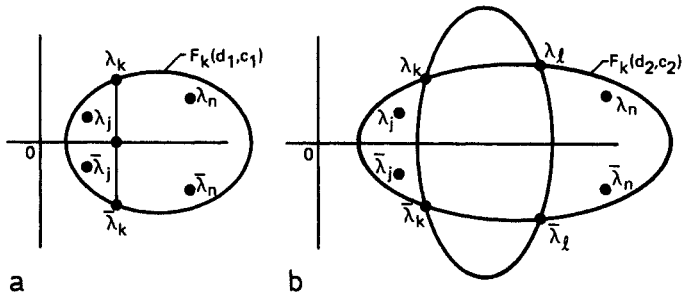

Fig. 3

moves from (d_1, c_1^2), let (d_2, c_2^2) be the first point such that the ellipse $F_k(d_2, c_2)$ passes through another eigenvalue, say λ_l. Since it was the first, $F_k(d_2, c_2)$ still encloses the spectrum, and, from Lemma 2.3, $\lambda_l \in H^+$.

Suppose λ_k and λ_l are the only eigenvalues on the ellipse $F_k(d_2, c_2)$. We can move (d, c^2) through R in such a way that $F_k(d, c)$ passes through λ_k and λ_l. As c^2 gets negatively large the foci of the ellipse $F_k(d, c)$ have large imaginary part and the ellipse is deformed into the infinite column between $\mathrm{Re}(\lambda_k)$ and $\mathrm{Re}(\lambda_l)$ (see Fig. 3 b). If $\mathrm{Re}(\lambda_k) < \mathrm{Re}(\lambda_l)$ then one of the intermediate ellipses must have passed through λ_j. If $\mathrm{Re}(\lambda_k) > \mathrm{Re}(\lambda_l)$, then one of the intermediate ellipses must have passed through λ_n. In either case, as (d, c^2) moves from (d_2, c_2^2) let $F_k(d_3, c_3)$ be the first ellipse to pass through a third eigenvalue, say λ_m. Since it is the first, $F_k(d_3, c_3)$ still encloses the spectrum, and, from Lemma 3.2, $\lambda_m \in H^+$. This proves the lemma.

The lemma provides criteria by which certain elements of the set H^+ can be ignored. The implementation of these criteria will arise naturally from the algorithm to be presented at the end of Section 4.

The results of this section are summed up in the following theorem.

Theorem 3.4. If A is a real valued matrix with eigenvalues in the right half plane, then the parameters d and c which yield the optimal rate of convergence can be found in terms of d and c^2 as the solution of the mini-max problem

$$\min_{(d, c^2) \in R} \max_{\lambda_i \in K} r(\lambda_i, d, c^2). \tag{3.11}$$

Notice that in the formulae (2.24) of the iteration c appears only as c^2. The iteration can be carried out in real arithmetic even when the scaling parameter c is pure imaginary.

4. The Mini-Max Solution

The optimal iteration parameters can be found as the point, $(d, c^2) \in R$, that minimizes the maximum of a finite number of real valued functions of two real variables. Consider each function, $r(\lambda_i, d, c^2)$, to be a surface over the d, c^2-plane. Let

$$m(d, c^2) = \max_{\lambda_i \in H^+} r(\lambda_i, d, c^2)$$

be the maximum surface. The optimal point is the minimum for this surface in the region R. In this section we will find the mini-max solution explicitly in terms of the eigenvalues of A. For details see Manteuffel [16].

4.1. *The Alternative Theorem*

We borrow the following useful result from functional analysis (Bartle [1]).

Theorem 4.1 (Alternative Theorem). If $\{f_i(x, y)\}$ is a finite set of real valued functions of two real variables, each of which is continuous on a closed and bounded region S and

$$m(x, y) = \max_i f_i(x, y),$$

then $m(x, y)$ takes on a minimum at some point (x_0, y_0) in the region S. If (x_0, y_0) is in the interior of S, then one of the following holds:

1. The point (x_0, y_0) is a local minimum of $f_i(x, y)$ for some i such that $m(x_0, y_0) = f_i(x_0, y_0)$.
2. The point (x_0, y_0) is a local minimum among the locus $\{(x, y) \in S \mid f_i(x, y) = f_j(x, y)\}$ for some i and j such that $m(x_0, y_0) = f_i(x_0, y_0) = f_j(x_0, y_0)$.
3. The point (x_0, y_0) is such that for some i, j, and k, $m(x_0, y_0) = f_i(x_0, y_0) = f_j(x_0, y_0) = f_k(x_0, y_0)$.

The solution to the mini-max problem lies in the open region R. It can be easily seen that there is some compact $S \subset R$ which contains the solution in its interior. Thus, we can apply the Alternative Theorem.

4.2. *The Mini-Max Solution*

Each point $(d, c^2) \in R$ is associated with a family of ellipses, $\mathfrak{F}(d, c)$, in the λ-plane whose members are the level lines of $r(\lambda, d, c^2)$. Let $F_i(d, c)$ be the ellipse in this family passing through λ_i. From (2.13) we know that $r(\lambda, d, c^2)$ takes on the same value at each $\lambda \in F_i(d, c)$. In particular, consider λ_0 in Figure 4. Since $(d - \lambda_0)^2 \geqq c^2$ we have

$$r(\lambda_i, d, c^2) = r(\lambda_0, d, c^2) = \frac{(d - \lambda_0) + ((d - \lambda_0)^2 - c^2)^{1/2}}{d + (d^2 - c^2)^{1/2}}.$$

If we let $(d - \lambda_0)^2 = a^2$, then every point $\lambda = x + iy$ on the ellipse $F_i(d, c)$ will satisfy

$$\frac{(d - x)^2}{a^2} + \frac{y^2}{a^2 - c^2} = 1.$$

Letting $\lambda_i = \mathrm{x}_i + iy_i$, we can write

$$r(\lambda_i, d, c^2) = \frac{a + (a^2 - c^2)^{1/2}}{d + (d^2 - c^2)^{1/2}}, \tag{4.1}$$

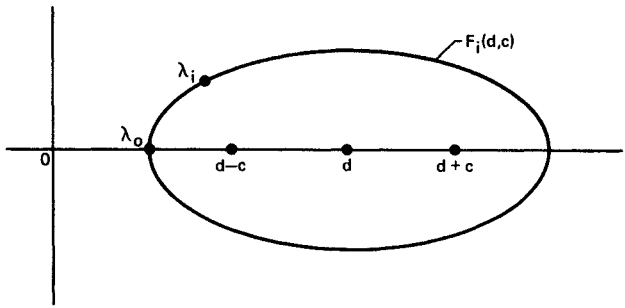

Fig. 4

subject to the constraint

$$\frac{(d-x_i)^2}{a^2}+\frac{y_i^2}{a^2-c^2}=1. \tag{4.2}$$

(If the ellipse $F_i(d, c)$ is degenerate, the above constraint does not hold. If $y_i \neq 0$, the degenerate case gives $a^2=0$. If $y_i=0$, the degenerate case gives $a^2=c^2$. In any case, a^2 varies continuously with d and c^2.)

We have the following results (Manteuffel [16]).

a) *One Eigenvalue.* Suppose the positive hull, H^+, contains only one eigenvalue, $\lambda_1=x_1+iy_1$. The *only* local minimum of the function $r(\lambda_1, d, c^2)$ occurs at

$$d=x_1,$$
$$c^2=-y_1^2.$$

That is, the family of ellipses $\mathfrak{F}(d, c)$ has foci at λ_1 and $\bar{\lambda}_1$ and the member of the family passing through λ_1 is the degenerate ellipse. We have

$$a^2=0,$$

which yields from (4.1)

$$r(\lambda_1, x_1, -y_1^2)=\frac{y_1}{x_1+(x_1^2+y_1^2)^{1/2}}. \tag{4.3}$$

b) *Two Eigenvalues.* Suppose the positive hull contains two eigenvalues, $\lambda_1=x_1+iy_1$, $\lambda_2=x_2+iy_2$. The mini-max problem becomes

$$\min_{(d,c^2)\in R} \max\{r(\lambda_1, d, c^2),\ r(\lambda_2, d, c^2)\}.$$

It is easily shown that

$$r(\lambda_1, x_1, -y_1^2)<r(\lambda_2, x_1, -y_1^2),$$
$$r(\lambda_1, x_2, -y_2^2)>r(\lambda_2, x_2, -y_2^2).$$

Since there is only one local minimum on each surface, the Alternative Theorem 4.1 yields that the solution must occur along the intersection of the two surfaces,

$$r(\lambda_1, d, c^2)=r(\lambda_2, d, c^2). \tag{4.4}$$

Consider a point (d, c^2) along this intersection. The eigenvalues λ_1 and λ_2 lie on the same member of the family $\mathfrak{F}(d, c)$ in the λ-plane. Moreover, they both satisfy the equation of that ellipse,

$$\begin{aligned} &\frac{(d-x_1)^2}{a^2}+\frac{(y_1)^2}{a^2-c^2}=1,\\ &\frac{(d-x_2)^2}{a^2}+\frac{(y_2)^2}{a^2-c^2}=1. \end{aligned} \tag{4.5}$$

Let

$$A=\frac{x_2-x_1}{2}, \quad B=\frac{x_2+x_1}{2},$$

$$S=\frac{y_2-y_1}{2}, \quad T=\frac{y_2+y_1}{2}.$$

We may assume that $x_2>x_1$; then, we have $A>0$, $B>0$, and $T\geqq 0$. With this notation, Equation (4.5) yield the following relationships. If $S=0$, then

$$\begin{aligned} &d=B,\\ &c^2=\frac{a^2(a^2-(A^2+T^2))}{(a^2-A^2)}. \end{aligned} \tag{4.6}$$

If $S\neq 0$, then

$$\begin{aligned} c^2&=\frac{\left(d-\left(B+\frac{ST}{A}\right)\right)\left(d-\left(B-A\frac{T}{S}\right)\right)\left(d-\left(BA\frac{S}{T}\right)\right.}{(d-B)}\\ a^2&=\left(d-\left(B-A\frac{T}{S}\right)\right)\left(d-\left(B-A\frac{S}{T}\right)\right). \end{aligned} \tag{4.7}$$

If $S=0$, the *only* local minimum, an absolute minimum, along the intersection of the surfaces can be found in terms of $y=a^2$, as the *only* real root of the cubic polynomial

$$q_1 y^3+q_2 y^2+q_3 y+q_4=0$$

in the interval (A^2, B^2). The coefficients are:

$$\begin{aligned} q_1&=(B^2+T^2),\\ q_2&=-3A^2B^2,\\ q_3&=3A^4B^2,\\ q_4&=-A^4B^2(A^2+T^2). \end{aligned} \tag{4.8}$$

(If $T=0$, then $c^2=a^2=A^2$ which corresponds to the symmetric matrix problem (Golub and Varga [8]).)

If $S\neq 0$, the *only* local minimum along the intersection of the surfaces can be found in terms of $z=d-B$, as the root of the polynomial

$$p_1 z^5 + p_2 z^4 + p_3 z^3 + p_4 z^2 + p_5 z + p_6 = 0$$

in the interval

$(0, A)$ for $S > 0$,

$(-A, 0)$ for $S < 0$.

The coefficients are:

$$\begin{aligned}
p_1 &= \left(2B - A\left(\frac{T}{S} + \frac{S}{T}\right)\right)\left(2B + \frac{ST}{A} - A\left(\frac{T}{S} + \frac{S}{T}\right)\right),\\
p_2 &= \left(2B + \frac{ST}{A} - A\left(\frac{T}{S} + \frac{S}{T}\right)\right)\left((2AB + ST)\left(\frac{T}{S} + \frac{S}{T}\right) + 4A^2\right)\\
&\quad + B^2\left(2B - A\left(\frac{T}{S} + \frac{S}{T}\right)\right) + B(B^2 - A^2),\\
p_3 &= 4A^4 - 4A^3 B\left(\frac{T}{S} + \frac{S}{T}\right) + A^2 ST\left(\left(\frac{T^3}{S^3} + \frac{S^3}{T^3}\right) - 3\left(\frac{T}{S} + \frac{S}{T}\right)\right)\\
&\quad + A^2 B^2\left(\frac{T^2}{S^2} + \frac{S^2}{T^2} + 3\right),\\
p_4 &= AST\left(\left(B - A\frac{T}{S}\right)\left(B - 3A\frac{T}{S}\right) + \left(B - A\frac{S}{T}\right)\left(B - 3A\frac{S}{T}\right)\right),\\
p_5 &= -3A^3 ST\left(2B - A\left(\frac{T}{S} + \frac{S}{T}\right)\right),\\
p_6 &= -3A^3 ST(B^2 - A^2).
\end{aligned} \tag{4.9}$$

The point thus found represents the best parameters when the positive hull contains two eigenvalues. Such a point will be referred to as a *pair-wise best point*, and the associated ellipse passing through the two eigenvalues will be referred to as the *pair-wise best ellipse*. The convergence factor associated with the pair-wise best ellipse is as in (4.1).

c) *Three or More Eigenvalues.* Suppose the positive hull contains three or more eigenvalues. From the Alternative Theorem, the mini-max solution must be a pair-wise best point or a point of intersection of three surfaces. Let λ_1, λ_2 be two eigenvalues in the positive hull with pair-wise best point $(\underline{d}, \underline{c}^2)$. Then $(\underline{d}, \underline{c}^2)$ is the mini-max solution if and only if the pair-wise best ellipse contains the other eigenvalues in the closure of its interior. If this is the case, we have

$$\begin{aligned}
\min_{(d, c^2) \in R} \max_{\lambda_i \in H^+} \{r(\lambda_i, d, c^2)\} &\geqq \min_{(d, c^2) \in R} \max \{r(\lambda_1, d, c^2), r(\lambda_2, d, c^2)\}\\
&= r(\lambda_1, \underline{d}, \underline{c}^2)\\
&= \max_{\lambda_i \in H^+} \{r(\lambda_i, \underline{d}, \underline{c}^2)\}.
\end{aligned} \tag{4.10}$$

The last equality holds from (2.13) since the other eigenvalues are insider or on the pair-wise best ellipse.

If no pair-wise best point is the mini-max solution, then the solution must be a point of intersection of three surfaces, referred to as a *three way point*. Let $\lambda_1 = x_1 + i y_1$, $\lambda_2 = x_2 + i y_2$, and $\lambda_3 = x_3 + i y_3$, where $x_1 < x_2 < x_3$, be three eigenvalues in the positive hull. There will be a point (d, c^2) such that

$$r(\lambda_1, d, c^2) = r(\lambda_2, d, c^2) = r(\lambda_3, d, c^2)$$

only if

$$(x_2 - x_1)(y_3^2 - y_1^2) < (x_3 - x_1)(y_2^2 - y_1^2), \tag{4.11}$$

and if it exists it will be unique. Equivalently, there will be a unique ellipse, referred to as a *three-way ellipse*, symmetric with respect to the real axis, passing through λ_1, λ_2 and λ_3. We have

$$\begin{aligned} d &= \frac{1}{2} \frac{(y_1^2(x_2^2 - x_3^2) + y_2^2(x_3^2 - x_1^2) + y_3^2(x_1^2 - x_2^2))}{(y_1^2(x_2 - x_3) + y_2^2(x_3 - x_1) + y_3^2(x_1 - x_2))}, \\ a^2 &= d^2 - \frac{(y_1^2 x_2 x_3(x_2 - x_3) + y_2^2 x_1 x_3(x_3 - x_1) + y_3^2 x_1 x_2(x_1 - x_2))}{(y_1^2(x_2 - x_3) + y_2^2(x_3 - x_1) + y_3^2(x_1 - x_2))}, \\ c^2 &= a^2 \left(1 - \frac{(y_1^2(x_2 - x_3) + y_2^2(x_3 - x_1) + y_3^2(x_1 - x_2))}{(x_1 - x_2)(x_2 - x_3)(x_3 - x_1)}\right). \end{aligned} \tag{4.12}$$

Such a three-way point can be the mini-max solution only if the associated three-way ellipse passing through λ_1, λ_2, and λ_3 contains the spectrum in the closure of its interior. The convergence factor associated with this three-way ellipse is as in (4.1).

4.3. The Algorithm

If the eigenvalues in the positive hull are known, the optimal iteration parameters can be found by the following algorithm.

1. For each pair of eigenvalues in the positive hull, find the pair-wise best point. If the pair-wise best ellipse contains the other members of the positive hull in the closure of its interior, then the mini-max solution is found.

2. If no pair-wise best point is the solution, find the three-way point, if it exists, for each set of three eigenvalues in the positive hull. If the associated three-way ellipse contains the other members of the positive hull in the closure of its interior, then this point is a candidate.

3. The three-way candidate with the smallest convergence factor will be the mini-max solution.

Notice that those eigenvalues in H^+ which are involved in some combination of three eigenvalues that produced a three-way candidate are exactly the key elements described in Section 3.2. In the course of finding the mini-max solution, the key elements are determined. Thus, if a sequence of eigenvalue estimates were made available during execution of Tchebychev iteration, only the key elements need be retained and used in subsequent searches for optimal parameters.

An adaptive procedure that dynamically yields eigenvalue estimates, along with numerical results will appear in a subsequent paper.

References

1. Bartle, R.G.: Elements of real analysis. New York: Wiley 1964
2. Birkhoff, G., Maclane, S.: A survey of modern algebra. New York: MacMillan 1953
3. Diamond, M.A.: An economical algorithm for the solution of elliptic difference equations independent of user-supplied parameters. Ph.D. Dissertation, Department of Computer Science, University of Ill., 1972
4. Dunford, N., Schwartz, J.L.: Linear operators. New York: Interscience 1958
5. Engeli, M., Ginsburg, T.H., Rutishauser, H., Stiefel, E.L.: Refined iterative methods for computation of the solution and Eigenvalues of self-adjoint boundary value problems. Mitteilungen aus dem Institut für Angewandte Mathematik, No. 8, pp. 1–78, 1959
6. Faddeev, D.K., Faddeeva, U.N.: Computational methods of linear algebra. San Francisco: Freeman 1963
7. Fox, L., Parker, I.B.: Chebyshev polynomials in numerical analysis. London: Oxford University Press 1968
8. Golub, G.H., Varga, R.S.: Chebyshev semi-iterative methods, successive over relaxation iterative methods and second order Richardson iterative methods. Numerische Math. **3**, 147 (1961)
9. Hageman, L.A.: The estimation of acceleration parameters for the Chebyshev polynomial and the successive over relaxation iteration methods. AEC Research and Development Report WAPD-TM-1038, June, 1972
10. Hestenes, M.R., Stiefel, E.L.: Methods of conjugate gradients for solving linear systems. N.B.S.J. of Res. **49**, 409–436 (1952)
11. Hille, E.: Analytic function theory, Vol. II, Ch. 16, pp. 264–274. Boston: Ginn 1962
12. Householder, A.S.: The theory of matricies in numerical analysis, pp. 37–57. New York: Blaisdell 1964
13. Kincaid, D.R.: On complex second-degree iterative methods. SIAM J. numer. Analysis **II**, No. 2, 211–218 (1974)
14. Kincaid, D.R.: Numerical results of the application of complex second-degree and semi-iterative methods. Center for Numerical Analysis Report, CNA-90, Oct., 1974
15. Kjellberg, G.: On the convergence of successive over relaxation applied to a class of linear systems of equations with complex Eigenvalues. Ericsson Technics, No. 2, pp. 245–258, 1958
16. Manteuffel, T.A.: An iterative method for solving nonsymmetric linear systems with dynamic estimation of parameters. Digital Computer Laboratory Reports, Rep. UIUCDS-R-75-758, University of Ill., Oct. 1975
17. Manteuffel, T.A.: Unpublished notes, 1974
18. Stiefel, E.L.: Kernel polynomials in linear algebra and their applications. U.S. N.B.S. Applied Math Series **49**, 1–22 (1958)
19. Varga, R.S.: A comparison of successive over relaxation and semi-iterative methods using Chebyshev polynomials. SIAM J. numer. Analysis **5**, 39–46 (1957)
20. Varga, R.S.: Matrix iterative analysis. Englewood Cliffs, N.J.: Prentice-Hall 1962
21. Wachspress, E.L.: Iterative solution of elliptic systems, pp. 157–158. Engelwood Cliffs, N.J.: Prentice-Hall 1962
22. Walsh, J.L.: Interpolation and approximation by rational functions in the complex domain, revised ed., Colloquium Publications, Vol. 20. Providence, R.I.: A.M.S. 1956
23. Wilkinson, J.H.: The algebraic Eigenvalue problem. Oxford: Clarendon Press 1965
24. Wrigley, H.E.: Accelerating the Jacobi method for solving simultaneous equations by Chebyshev extrapolation when the Eigenvalues of the iteration matrix are complex. Computer J. **6**, 169–176 (1963)
25. Young, D.M., Edison, H.D.: On the determination of the optimum relaxation factor for the SOR method when the Eigenvalues of the Jacobi method are complex. Center for Numerical Analysis Report, CNA-1, September 1970
26. Young, D.: Iterative solution of large linear systems, pp. 191–200. New York-London: Academic Press 1971

Received June 9, 1976