APPROXIMATION OF THE EIGENVALUES OF NON SELF-ADJOINT OPERATORS

By John E. Osborn

1. Introduction and Setting. The general nature of the problem considered in this paper is that of approximate calculation of the eigenvalues of a non self-adjoint operator. These eigenvalues will in general be complex numbers so it is desired to find regions of the complex plane which contain some or all of these eigenvalues. The class of operators considered and the exact nature of the regions found will be described in later paragraphs.

This same problem for self-adjoint operators has been extensively explored. The eigenvalues are in this case real and many methods have been developed for obtaining upper and lower bounds for the eigenvalues. For a description of some of these methods see [4].

A basic theorem on the location of the eigenvalues of a matrix operator is the Gershgorin theorem which locates the eigenvalues of a matrix in circles centered at the diagonal elements. For a discussion of this theorem and extensions of it see [5].

Mysovskih [6] has developed a method for approximating the eigenvalues of a Fredholm integral operator with a continuous, not necessarily Hermitian, kernel. In this method the kernel is approximated by a second kernel, usually degenerate, for which the eigenvalues are known. If a particular eigenvalue of the second kernel is chosen then under certain conditions it is possible to describe a circle about it which contains an eigenvalue of the original operator.

Fichera [2] obtained a convergence result for Picone's method. This method applies to unbounded operators whose inverses are compact. Let E be such an operator and $\{u_i\}$ a complete system of vectors. Consider the functional

$$\mu_n(\lambda) = \inf \frac{\| Eu - \lambda u \|^2}{\| u \|^2}, \qquad u = \sum_{i=1}^{n} c_i u_i .$$

The method of Picone uses the values of λ which minimize this functional as approximations to the eigenvalues of E. The result is as follows. Let C_ρ be a circle of radius ρ centered at the origin. For each n let $L_\rho(n)$ be the set of those minimizing points of $\mu_n(\lambda)$ which satisfy an additional condition described in [2]. Let Λ_ρ be the set of eigenvalues of E in C_ρ. Then $\lim_{n\to\infty} L_\rho(n) = \Lambda_\rho$ where the convergence is in the Hausdorff metric topology on compact subsets of the plane. No estimate of the rate of convergence is given.

For differential operators Petrov [8] estimates the difference between an arbitrary approximate eigenvalue defined by Galerkin's method and any exact eigenvalue which satisfies certain special conditions.

In this paper the class of operators considered is as follows. Let L be the inverse of a positive definite Hilbert-Schmidt operator on a complex, separable, infinite dimensional Hilbert space H. Let V be the domain of L. Since H is a complex space and L^{-1} is positive definite L will be self-adjoint in the sense that $(Lx, y) =$

(x, Ly) for all $x, y \in V$. The spectrum of L consists of a countable set of positive eigenvalues whose only limit point is $+\infty$. Let λ_n be the n^{th} eigenvalue of L where the eigenvalues are counted in increasing order taking account of geometric multiplicities, i.e. the dimensions of the eigenspaces of the eigenvalues. The eigenvectors x_n of L can be chosen to be orthonormal; the set $\{x_n\}$ is complete. Now let $A: V \to H$ be a bounded, not necessarily self-adjoint, operator and define $\tilde{L} = L + A$. Assuming that the eigenvalues and eigenvectors of L are known, the problem considered is the approximation of the eigenvalues of $\tilde{L}$.

The main results of Section 2, Theorems 2.1 and 2.2, locate the eigenvalues of $\tilde{L}$ in circles centered at the eigenvalues of L. More precisely, in a slightly restricted form, these theorems assert the following. If $\| A \| < \lambda_1$, and if the circles

$$C_m \equiv \{\lambda \,|\, |\lambda - \lambda_m| \leqq \| A \|\}, \qquad m = 1, 2, \cdots,$$

are mutually disjoint, then the eigenvalues of $\tilde{L}$ are contained in the circles C_m, one eigenvalue to each circle. It should be mentioned here that the result of Theorem 2.1 has been obtained by J. T. Schwartz [9] and similar results have been obtained by Gavurin [3].

In Section 3 the Rayleigh-Ritz method is applied to $\tilde{L}$, using the eigenvectors of L as a basis, to obtain improvable approximations for the eigenvalues of L. Estimates are obtained for the errors which arise. We assume that $\| A \| < \lambda_1$ and that the circles C_m are disjoint. Let μ_p be the eigenvalue of $\tilde{L}$ in C_p; for $n \geqq p$ let $\eta_1(n), \cdots, \eta_n(n)$ be the eigenvalues of $\tilde{L}_n \equiv (P_n \tilde{L})|_{V_n}$, where V_n is the subspace spanned by the first n eigenvectors of L and P_n is the projection onto that subspace. Theorem 3.4 states that if φ is any one of the quantities $\| \tilde{L}_n^* - \tilde{L}_n \|$, $|||\tilde{L}_n^* - \tilde{L}_n|||$, $\| A^* - A \|$, $|||A^* - A|||$, $2\| A \|$, or $2|||A|||$, where $|||\cdot|||$ denotes the Hilbert-Schmidt norm, and if

$$\varphi \max_{1 \leqq k \leqq n} \sum_{j=1, j \neq k}^{n} |\eta_k(n) - \bar{\eta}_j(n)|^{-1} < 1,$$

then for some i, say i_0, where $1 \leqq i_0 \leqq n$,

$$|\mu_p - \eta_{i_0}(n)| \leqq \left(\frac{1 + \varphi \max_k \sum_{j \neq k} |\eta_k(n) - \bar{\eta}_j(n)|^{-1}}{1 - \varphi \max_k \sum_{j \neq k} |\eta_k(n) - \bar{\eta}_j(n)|^{-1}} \right)^{\frac{1}{2}} \frac{\| A \|^2}{((\lambda_{n+1} - \lambda_p)^2 - 2(\lambda_{n+1} - \lambda_p)\| A \|)^{\frac{1}{2}}}$$

$$\equiv \epsilon_p(n).$$

$\epsilon_p(n)$ is readily computable. Theorem 3.5 is a convergence result. It asserts that, for fixed p, $\lim_{n \to \infty} \epsilon_p(n) = 0$ if $\varphi \sum_{i=1}^{\infty} (\lambda_{i+1} - \lambda_i - 2\| A \|)^{-1} \leqq \alpha$, where φ is any one of the quantities $\| A^* - A \|$, $|||A^* - A|||$, $2\| A \|$, or $2|||A|||$ and $\alpha < 1$.

Although the assumption that $\| A \| < \lambda_1$ is used in the proof of all the main theorems of the paper it does not restrict the applications of the method. If $\| A \| \geqq \lambda_1$ we can consider the operator $L' = L + cI$ where $c > \| A \| - \lambda_1$. $(L')^{-1}$ is a positive definite Hilbert-Schmidt operator and has eigenvalues $\lambda_i' = \lambda_i + c$. Since $\| A \| < \lambda_1'$ we can apply the results of the paper to $L' + A$ and hence obtain approximations to the eigenvalues of $L + A$ by subtracting c from the corresponding approximations to the eigenvalues of $L' + A$.

In Section 4 the results of Sections 2 and 3 are illustrated by examples. The computations for the first example were done on the facilities of the Numerical Analysis Center at the University of Minnesota. The computer program used in these computations is based on a program by Ehrlich [1] for finding the eigenvalues of a non self-adjoint matrix and this latter program is based on a program due to E. E. Osborne [7]. No account of round-off error has been taken in these examples.

The author wishes to thank Professor Hans F. Weinberger for helpful guidance during the preparation of this paper.

This research was partially supported by the National Science Foundation Grant No. GP-3904 with the University of Minnesota.

2. Basic Location Theorems. *Lemma* 2.1. Suppose μ is an eigenvalue of $\tilde{L}$ and y a corresponding unit eigenvector. Then $\sum_{i=1}^{\infty} |(\mu - \lambda_i)(y, x_i)|^2 \leqq \| A \|^2$.

Proof. For each i we have

$$(Ay, x_i) = ((\tilde{L} - L)y, x_i) = (\mu - \lambda_i)(y, x_i).$$

Thus by Parseval's relation we get

$$\sum_{i=1}^{\infty} |(\mu - \lambda_i)(y, x_i)|^2 = \| Ay \|^2 \leqq \| A \|^2.$$

Theorem 2.1. Suppose μ is an eigenvalue of $\tilde{L}$. Then for some i, say i_0, $| \mu - \lambda_{i_0} | \leqq \| A \|$.

Proof. Suppose $| \mu - \lambda_i | > \| A \|$ for all i. Then

$$\sum_{i=1}^{\infty} |(\mu - \lambda_i)(y, x_i)|^2 > \| A \|^2 \sum_{i=1}^{\infty} |(y, x_i)|^2 = \| A \|^2.$$

This contradicts Lemma 2.1; hence we get the result.

Theorem 2.1 asserts that the eigenvalues of $\tilde{L}$ are contained in $\bigcup_{m=1}^{\infty} C_m$.

Lemma 2.2. If $\| A \| < \lambda_1$ then $\mathfrak{L}_t \equiv L + tA$, $0 \leqq t \leqq 1$, has for each fixed t, an inverse which is a Hilbert-Schmidt operator on H.

Proof. For each t we have $L + tA = (I_H + tAL^{-1})L$. Since $\| A \| < \lambda_1$ we see that $\| tAL^{-1} \| < 1$; this implies that $(I_H + tAL^{-1})^{-1}$ exists as a bounded operator on H. Hence $L + tA$ is invertible and

$$(L + tA)^{-1} = L^{-1}(I_H + tAL^{-1})^{-1}.$$

$(L + tA)^{-1}$ is a Hilbert-Schmidt operator since it is the product of a bounded operator and a Hilbert-Schmidt operator.

Definition. The algebraic multiplicity of an eigenvalue μ of $\mathfrak{L}_t$ is the dimension of the null space of $(\mathfrak{L}_t^{-1} - (1/\mu)I)^{\nu}$ where ν is the smallest integer such that $(\mathfrak{L}_t^{-1} - (1/\mu)I)^{\nu}$ and $(\mathfrak{L}_t^{-1} - (1/\mu)I)^{\nu+1}$ have the same null space.

The algebraic multiplicity of an eigenvalue of $\mathfrak{L}_t$ is greater than or equal to its geometric multiplicity and the two multiplicities are equal if $\mathfrak{L}_t$ is normal.

Since, if $\| A \| < \lambda_1$, $\mathfrak{L}_t^{-1}$ is a Hilbert-Schmidt operator, it has a Fredholm determinant; denote this determinant by $\delta_t(\lambda)$. We assume the following properties of $\delta_t(\lambda)$ are known. $\delta_t(\lambda)$ is entire in λ, the zeros of $\delta_t(\lambda)$ are the eigenvalues of $\mathfrak{L}_t$, and the algebraic multiplicity of an eigenvalue of $\mathfrak{L}_t$ is equal to its order

as a zero of $\delta_t(\lambda)$. Furthermore $\delta_t(\lambda)$ is given by the power series $\delta_t(\lambda) = \sum_{i=0}^{\infty} \delta_{i,t}\lambda^i$, where

$$\delta_{0,t} = 1, \qquad \delta_{i,t} = \frac{(-1)^i}{i!} \det \begin{bmatrix} 0 & i-1 & \cdot & \cdot & \cdot & \cdot 0 \\ \sigma_{2,t} & 0 & \ddots & \ddots & & \vdots \\ \sigma_{3,t} & \sigma_{2,t} & \ddots & \ddots & \ddots & \\ \vdots & \vdots & & \ddots & \ddots & 1 \\ \sigma_{i,t} & \sigma_{i-1,t} & \cdot & \cdot & \cdot & \cdot 0 \end{bmatrix}, \qquad i = 1, 2, \cdots,$$

$$\sigma_{j,t} = \operatorname{tr}\left((\mathfrak{L}_t^{-1})^j\right) = \operatorname{trace}\left((\mathfrak{L}_t^{-1})^j\right), \qquad j = 2, 3, \cdots.$$

Also

$$|\,\delta_{i,t}\,| \leqq \frac{e^{i/2}\,|||\,\mathfrak{L}_t^{-1}\,|||^i}{i^{i/2}}.$$

(A complete discussion of these ideas can be found in Zaanen [10].) One further property is needed and is given in the following lemma.

Lemma 2.3. Let S be a compact set not containing any of the zeros of $\delta_t(\lambda)$ for any t. Then on $[0, 1] \times S$, $[(d/d\lambda)(\delta_t(\lambda))]/\delta_t(\lambda)$ is continuous in t uniformly in λ.

Proof. As a first step we show that $\mathfrak{L}_t^{-1}$ is a continuous function of t in the topology of the Hilbert-Schmidt norm. Since $\|\,tAL^{-1}\,\| < 1$ we see that $\|(I_H + tAL^{-1})^{-1}\,\| \leqq (1 - \|\,tAL^{-1}\,\|)^{-1}$. Thus

$$|||\,\mathfrak{L}_t^{-1}\,||| = |||\,L^{-1}(I_H + tAL^{-1})\,||| \leqq |||\,L^{-1}\,|||(1 - \|\,tAL^{-1}\,\|)^{-1}$$
$$\leqq |||\,L^{-1}\,|||(1 - \|\,AL^{-1}\,\|)^{-1} \equiv K.$$

Now $\mathfrak{L}_t^{-1} - \mathfrak{L}_{t_0}^{-1} = \mathfrak{L}_t^{-1}(\mathfrak{L}_{t_0} - \mathfrak{L}_t)\mathfrak{L}_{t_0}^{-1} = (t_0 - t)\mathfrak{L}_t^{-1}A\mathfrak{L}_{t_0}^{-1}$ and hence $|||\,\mathfrak{L}_t^{-1} - \mathfrak{L}_{t_0}^{-1}\,||| \leqq |\,t_0 - t\,|\,\|\,A\,\|\,K^2$. This implies the continuity of $\mathfrak{L}_t^{-1}$. Next we show that $\delta_t(\lambda)$ is a continuous function of $(t, \lambda) \in [0, 1] \times S$. Let $R = \max\{|\,\lambda\,|\,|\,\lambda \in S\}$. Since

$$|\,\delta_{i,t}\lambda^i\,| \leqq \frac{e^{i/2}\,|||\,\mathfrak{L}_t^{-1}\,|||^i}{i^{i/2}}\,|\,\lambda\,|^i \leqq \frac{e^{i/2}}{i^{i/2}}\,(KR)^i \equiv a_i$$

and $\sum_{i=0}^{\infty} a_i$ converges we see that $\sum_{i=0}^{\infty} \delta_{i,t}\lambda^i$ converges uniformly in t and λ. $\delta_{i,t}\lambda^i$ depends continuously on λ and $\delta_{i,t}$ depends continuously on $\sigma_{j,t}$, $j = 2, 3, \cdots, i$. Hence it is sufficient to show the continuity of $\sigma_{j,t}$, $j = 2, 3, \cdots$, as a function of t. The following inequality shows this.

$$\begin{aligned} |\,\sigma_{j,t} - \sigma_{j,t_0}\,| &= |\,\operatorname{tr}[(\mathfrak{L}_t^{-1})^j - (\mathfrak{L}_{t_0}^{-1})^j]\,| \\ &= |\,\textstyle\sum_{i=1}^{j} \operatorname{tr}[(\mathfrak{L}_t^{-1})^{i-1}(\mathfrak{L}_t^{-1} - \mathfrak{L}_{t_0}^{-1})(\mathfrak{L}_{t_0}^{-1})^{j-i}]\,| \\ &\leqq \textstyle\sum_{i=1}^{j} |||(\mathfrak{L}_t^{-1})^{i-1}\,|||\,|||\,\mathfrak{L}_t^{-1} - \mathfrak{L}_{t_0}^{-1}\,|||\,|||(\mathfrak{L}_{t_0}^{-1})^{j-i}\,||| \\ &\leqq |\,t_0 - t\,|\,jK^{j+1}\,\|\,A\,\|. \end{aligned}$$

Essentially the same argument shows that $(d/d\lambda)(\delta_t(\lambda))$ is also continuous on $[0, 1] \times S$. Finally, since both $\delta_t(\lambda)$ and $(d/d\lambda)(\delta_t(\lambda))$ are continuous on $[0, 1] \times S$ and $\delta_t(\lambda)$ is never zero, $[(d/d\lambda)(\delta_t(\lambda))]/\delta_t(\lambda)$ is continuous and hence

uniformly continuous on $[0, 1] \times S$. In particular $[(d/d\lambda)(\delta_t(\lambda))]/\delta_t(\lambda)$ is continuous in t, uniformly in λ.

Theorem 2.2. Suppose that $\| A \| < \lambda_1$ and that k of the circles C_m form a connected set C which does not intersect any of the other circles: $C = \bigcup_{i=1}^{k} C_{m_i}$. Then counting according to algebraic multiplicities, there are k eigenvalues of $\tilde{L}$ in C.

Proof. For $t \in [0, 1]$, $\| \mathfrak{L}_t - L \| = t \| A \|$. If we apply Theorem 2.1 to $\mathfrak{L}_t$ instead on $\tilde{L}$ we get that the eigenvalues of $\mathfrak{L}_t$ lie in the circles

$$C_{m_i}^t \equiv \{\lambda \mid | \lambda - \lambda_m | \leqq t \| A \|\}.$$

Let $C_{m_i}^*$ be obtained from C_{m_i} by increasing the radius of the circles by a number r which is small enough so that $C^* \equiv \bigcup_{i=1}^{k} C_{m_i}^*$ does not intersect C_j, $j \neq m_1, \cdots, m_k$. ∂C^* is a compact set not containing any zeros of $\delta_t(\lambda)$ for any t. By the argument principle the number of zeros of $\delta_t(\lambda)$ in the interior of C^* is given by

$$N(t) \equiv \frac{1}{2\pi i} \int_{\partial C^*} \frac{(d/d\lambda)(\delta_t(\lambda))}{\delta_t(\lambda)} \, d\lambda.$$

$N(t)$ is integer valued and $N(0) = k$ since $\mathfrak{L}_0 = L$. The fact that $[(d/d\lambda)(\delta_t(\lambda))]/\delta_t(\lambda)$ is continuous in t, uniformly for $\lambda \in \partial C^*$, implies that $N(t)$ is continuous on $[0, 1]$. Thus $N(t)$ is identically equal to k. In particular $N(1) = k$ and therefore the number of eigenvalues of $\tilde{L}$ in the interior of C^* is k. Since r is arbitrary we have shown that there are k eigenvalues of $\tilde{L}$ in C.

Corollary 2.1. If $\| A \| < \lambda_1$ and $\bigcup_{m=1}^{\infty} C_m$ is not connected then $\tilde{L}$ has at least one eigenvalue.

Proof. In this situation there is an integer p such that $C_1 \cup \cdots \cup C_p$ is a connected set not intersecting any of the other circles. Theorem 2.2 applied to this set gives the result.

Corollary 2.2. If $\| A \| < \lambda_1$, the eigenvalues of $\tilde{L}$ are symmetric with respect to the real axis, and the circle C_{m_0} does not intersect any of the other circles, then the eigenvalue of $\tilde{L}$ in C_{m_0} is real.

Proof. By Theorem 2.2 there is exactly one eigenvalue of $\tilde{L}$ in C_{m_0}. If this eigenvalue is not real then its complex conjugate will also be an eigenvalue, will be in C_{m_0}, and will be different from it. This contradicts the fact that there is only one eigenvalue in C_{m_0}.

3. Improvable Approximations for the Initial Eigenvalues of $\tilde{L}$. Let $V_n = \text{span}\,(x_1, \cdots, x_n)$ where $x_1, \cdots, x_n$ are the first n eigenvectors of L. Let $L_n = (P_n L)|_{V_n}$. The n eigenvalues of $\tilde{L}_n$ will be called the n^{th} stage Rayleigh-Ritz eigenvalues of $\tilde{L}$ relative to $\{x_i\}$.

Lemma 3.1. Suppose η is an eigenvalue of $\tilde{L}_n$ and z a corresponding unit eigenvector. Then $\sum_{i=1}^{n} |(\eta - \lambda_i)(z, x_i)|^2 \leqq \| \tilde{L}_n - L_n \|^2$.

Proof. The proof of Lemma 2.1 with L and $\tilde{L}$ replaced by L_n and $\tilde{L}_n$ respectively gives this result.

Theorem 3.1. Suppose η is an eigenvalue of $\tilde{L}_n$. Then for some i, say i_0, where $1 \leqq i_0 \leqq n$, $|\eta - \lambda_{i_0}| \leqq \| \tilde{L}_n - L_n \| \leqq \| A \|$.

Proof. The first inequality follows from Lemma 3.1 just as Theorem 2.1 follows from Lemma 2.1, except that here the summation is from 1 to n. The second inequality is easily verified.

Theorem 3.1 asserts that the n^{th} stage Rayleigh-Ritz eigenvalues of $\tilde{L}$ are contained in $\bigcup_{m=1}^{n} C_m$.

Theorem 3.2. Suppose k of the circles $C_1, \cdots, C_n$ form a connected set C not intersecting the rest of the circles: $C = \bigcup_{i=1}^{k} C_{m_i}$. Then, counting according to algebraic multiplicities, there are k eigenvalues of $\tilde{L}_n$ in C.

Proof. The proof of this theorem is similar to the proof of Theorem 2.2. The essential idea in that proof was the application of the argument principle to the Ferdholm determinant. Here we can use the ordinary determinant.

For the remainder of this section we assume that the circles C_m are mutually disjoint and that $\| A \| < \lambda_1$. Theorems 2.1 and 2.2 apply in this situation and assert that the eigenvalues of $\tilde{L}$ are contained in the circles C_m, one eigenvalue to each circle; denote the one in C_j by μ_j. Also, for n fixed, Theorem 3.2 asserts that there is in each of the circles $C_1, \cdots, C_n$ exactly one of the n^{th} stage Rayleigh-Ritz eigenvalues of $\tilde{L}$; denote the one in C_j by $\eta_j(n)$.

Now we derive estimates on the size of the error which arises when $\eta_p(n)$, the p^{th} of the n^{th} stage Rayleigh-Ritz eigenvalues of $\tilde{L}$, is used as an approximation for μ_p, the p^{th} eigenvalue of $\tilde{L}$, $p = 1, 2, \cdots, n$. Let y_p be a unit eigenvector corresponding to μ_p. Let y_p be orthogonally decomposed as follows, $y_p = y_p(n) + w_p(n)$, $y_p(n) \in V_n$, $w_p(n) \in V_n^{\perp}$. Let $z_j(n)$ be a unit eigenvector of $\tilde{L}_n$ corresponding to $\eta_j(n)$. The vectors $z_1(n), \cdots, z_n(n)$ span V_n.

Lemma 3.2. Let $\alpha_{p,1}(n), \cdots, \alpha_{p,n}(n)$ be the coefficients in the expansion of $y_p(n)$ in the basis $z_1(n), \cdots, z_n(n)$. Then we have

$$\| \textstyle\sum_{i=1}^{n} \alpha_{p,i}(n) z_i(n) \| = (1 - \| w_p(n) \|^2)^{\frac{1}{2}}, \quad \text{and}$$

$$\| \textstyle\sum_{i=1}^{n} \alpha_{p,i}(n)(\mu_p - \eta_i(n)) z_i(n) \| \leqq \| A \| \, \| w_p(n) \|.$$

Proof. The first statement is established by noting that

$$y_p(n) = \textstyle\sum_{i=1}^{n} \alpha_{p,i}(n) z_i(n) \quad \text{and} \quad 1 = \| y_p \|^2 = \| y_p(n) \|^2 + \| w_p(n) \|^2.$$

Now

$$\textstyle\sum_{i=1}^{n} \alpha_{p,i}(n)(\mu_p - \eta_i(n)) z_i(n) = \mu_p y_p(n) - \tilde{L}_n y_p(n)$$

$$= \mu_p P_n y_p - P_n(L + A)(y_p - w_p(n)) = P_n A w_p(n).$$

The second statement follows clearly from this.

Lemma 3.3. Let $\xi_1, \cdots, \xi_n$ be a basis of unit vectors for an n-dimensional inner-product space. Let $\zeta = \sum_{i=1}^{n} \tau_i \xi_i$. Then

$$m(\xi_1, \cdots, \xi_n) \textstyle\sum_{i=1}^{n} |\tau_i|^2 \leqq \| \xi \|^2 \leqq M(\xi_1, \cdots, \xi_n) \sum_{i=1}^{n} |\tau_i|^2,$$

where $M(\xi_1, \cdots, \xi_n)$ and $m(\xi_1, \cdots, \xi_n)$ are respectively the greatest and least eigenvalues of the matrix $((\xi_i, \xi_j))$.

Proof. We have $\| \zeta \|^2 = (\zeta, \zeta) = \sum_{i,j=1}^{n} \tau_i \bar{\tau}_j (\xi_i, \xi_j)$. Now $((\xi_i, \xi_j))$ is a Hermitian matrix and thus $\| \zeta \|^2$ is a Hermitian form in $(\bar{\tau}_1, \cdots, \bar{\tau}_n)$. The result follows from the fact that the value of a Hermitian form lies between the least eigenvalue of its matrix times the Euclidean length of the vector and the greatest eigenvalue of its matrix times the Euclidean length of the vector.

Lemma 3.4. $1 - \| w_p(n) \|^2 \leqq M(z_1(n), \cdots, z_n(n)) \sum_{i=1}^{n} | \alpha_{p,i}(n) |^2$, and

$$m(z_1(n), \cdots, z_n(n)) \sum_{i=1}^{n} | \alpha_{p,i}(n)(\mu_p - \eta_i(n)) |^2 \leqq \| A \|^2 \| w_p(n) \|^2.$$

Proof. This follows directly from Lemmas 3.2 and 3.3.

Theorem 3.3. For some i, say i_0, where $1 \leqq i_0 \leqq n$, we have

$$| \mu_p - \eta_{i_0}(n) | \leqq \left(\frac{M(z_1(n), \cdots, z_n(n))}{m(z_1(n), \cdots, z_n(n))} \right)^{\frac{1}{2}} \frac{\| A \| \| w_p(n) \|}{(1 - \| w_p(n) \|^2)^{\frac{1}{2}}}.$$

Proof. We write M and m for $M(z_1(n), \cdots, z_n(n))$ and $m(z_1(n), \cdots, z_n(n))$ for the remainder of this section. Suppose

$$| \mu_p - \eta_i(n) | > (M/m)^{\frac{1}{2}} \frac{\| A \| \| w_p(n) \|}{(1 - \| w_p(n) \|^2)^{\frac{1}{2}}}$$

for $i = 1, 2, \cdots, n$. Then we have, using the first inequality of Lemma 3.4,

$$m \sum_{i=1}^{n} | \alpha_{p,i}(n)(\mu_p - \eta_i(n)) |^2 > M \frac{\| A \|^2 \| w_p(n) \|^2}{1 - \| w_p(n) \|^2} \sum_{i=1}^{n} | \alpha_{p,i}(n) |^2 \geqq \| A \|^2 \| w_p(n) \|^2.$$

This contradicts the second inequality of Lemma 3.4; thus we get the result.

Theorem 3.3 can be used to approximate the eigenvalues of $\tilde{L}$ but the quantities M/m and $\| w_p(n) \|$ are difficult to compute. The following Lemmas 3.5 and 3.6 provide more computable bounds for these quantities. Theorem 3.4 combines Theorem 3.3 and Lemmas 3.5 and 3.6.

Lemma 3.5. $\| w_p(n) \| \leqq \| A \| \, | \mu_p - \lambda_{n+1} |^{-1}$ for $n \geqq p$.

Proof. For $n \geqq p$ we have

$$\begin{aligned} \| w_p(n) \|^2 &= \sum_{i=n+1}^{\infty} |(y_p, x_i)|^2 \\ &\leqq | \mu_p - \lambda_{n+1} |^{-2} \sum_{i=n+1}^{\infty} |(y_p, x_i)(\mu_p - \lambda_i)|^2 \\ &\leqq | \mu_p - \lambda_{n+1} |^{-2} \sum_{i=1}^{\infty} |(y_p, x_i)(\mu_p - \lambda_i)|^2. \end{aligned}$$

The result now follows from Lemma 2.1.

Lemma 3.6. If φ is any one of the quantities $\| \tilde{L}_n^* - \tilde{L}_n \|$, $\| \| \tilde{L}_n^* - \tilde{L}_n \| \|$, $\| A^* - A \|$, $\| \| A^* - A \| \|$, $2 \| A \|$, or $2 \| \| A \| \|$ and if

$$\varphi \max_{1 \leqq k \leqq n} \sum_{j=1, j \neq k}^{n} | \eta_k(n) - \bar{\eta}_j(n) |^{-1} < 1,$$

then

$$(M/m)^{\frac{1}{2}} \leqq \left(\frac{1 + \varphi \max_k \sum_{j \neq k} | \eta_k(n) - \bar{\eta}_j(n) |^{-1}}{1 - \varphi \max_k \sum_{j \neq k} | \eta_k(n) - \bar{\eta}_j(n) |^{-1}} \right)^{\frac{1}{2}}.$$

Proof. The expression $\sum_{j \neq k} | \eta_k(n) - \bar{\eta}_j(n) |^{-1}$ is defined to be zero if $n = 1$.

With this convention the theorem is trivial if $n = 1$. M and m are the greatest and least eigenvalues of the matrix $(z_i(n), z_j(n))$. The diagonal elements of this matrix are all equal to 1 and hence we can use Gershgorin's theorem to locate the eigenvalues in circles centered at 1. The radii of the n Gershgorin circles are given by $R_k \equiv \sum_{j=1, j\neq k}^{n} |(z_k(n), z_j(n))|, k = 1, 2, \cdots, n$. If we let $R = \max_{1\leqq k\leqq n} R_k$ we have by Gershgorin's theorem that $M \leqq 1 + R$ and $1 - R \leqq m$. Now we estimate R. A simple calculation shows that for each $i \neq j$ we have

$$(z_i(n), z_j(n)) = \frac{(z_i(n), (\tilde{L}_n^* - \tilde{L}_n)z_j(n))}{\eta_i(n) - \bar{\eta}_j(n)}.$$

Thus

$$R_k \leqq \sum_{j\neq k} \frac{\| z_k(n) \| \, \| \tilde{L}_n^* - \tilde{L}_n \| \, \| z_j(n) \|}{| \eta_k(n) - \bar{\eta}_j(n) |} = \| \tilde{L}_n^* - \tilde{L}_n \| \sum_{j\neq k} |\eta_k(n) - \bar{\eta}_j(n)|^{-1}.$$

Hence for M and m we have the estimates

$$M \leqq 1 + \| \tilde{L}_n^* - \tilde{L}_n \| \max_k \sum_{j\neq k} | \eta_k(n) - \bar{\eta}_j(n)|^{-1}, \quad \text{and}$$
$$1 - \| \tilde{L}_n^* - \tilde{L}_n \| \max_k \sum_{j\neq k} | \eta_k(n) - \bar{\eta}_j(n)|^{-1} \leqq m.$$

Since

$$\| \tilde{L}_n^* - \tilde{L}_n \| \leqq ||| \tilde{L}_n^* - \tilde{L}_n |||,$$
$$\| \tilde{L}_n^* - \tilde{L}_n \| \leqq \| A^* - A \| \leqq ||| A^* - A ||| \leqq 2 ||| A |||, \quad \text{and}$$
$$\| A^* - A \| \leqq 2 \| A \|,$$

$\| \tilde{L}_n^* - \tilde{L}_n \|$ may be replaced by $||| \tilde{L}_n^* - \tilde{L}_n |||, \| A^* - A \|, ||| A^* - A |||, 2 \| A \|$, or $2 ||| A |||$ in these estimates. The desired result now follows easily.

Theorem 3.4. If φ is any one of the quantities $\| \tilde{L}_n^* - \tilde{L}_n \|, ||| \tilde{L}_n^* - \tilde{L}_n |||$, $\| A^* - A \|, ||| A^* - A |||, 2 \| A \|$, or $2 ||| A |||$, if

$$(\dagger) \qquad \varphi \max_k \sum_{j\neq k} | \eta_k(n) - \bar{\eta}_j(n)|^{-1} < 1,$$

and if $n \geqq p$, then for some i, say i_0, where $1 \leqq i_0 \leqq n$, we have

$$| \mu_p - \eta_{i_0}(n) | \leqq \left(\frac{1 + \varphi \max_k \sum_{j\neq k} | \eta_k(n) - \bar{\eta}_j(n) |^{-1}}{1 - \varphi \max_k \sum_{j\neq k} | \eta_k(n) - \bar{\eta}_j(n) |^{-1}} \right)^{\frac{1}{2}} \cdot \frac{\| A \|^2}{((\lambda_{n+1} - \lambda_p)^2 - 2(\lambda_{n+1} - \lambda_p) \| A \|)^{\frac{1}{2}}}$$
$$\equiv \epsilon_p(n).$$

Proof. A direct application of Theorem 3.3 and Lemmas 3.5 and 3.6 together with the inequality $| \mu_p - \lambda_{n+1} | \geqq \lambda_{n+1} - \lambda_p - \| A \|$ yields the result.

For fixed p we can use Theorem 3.4 to approximate μ_p, taking n to be any integer greater than or equal to p which satisfies $(\dagger)$. In many cases it is possible to conclude that $i_0 = p$. Compare the last paragraph of this section and Section 4. To gain the most precise information on the location of μ_p we can use the value of

n which minimizes $\epsilon_p(n)$, or we can use different values of n and consider the intersection of the corresponding circles; this necessitates the explicit calculation of $\epsilon_p(n)$ for the various values of n. Compare Section 4. The next theorem gives conditions under which we can use any $n \geqq p$ in the application of Theorem 3.4 and under which $\lim_{n\to\infty} \epsilon_p(n) = 0$.

Theorem 3.5. If $\varphi \sum_{i=1}^{\infty} (\lambda_{i+1} - \lambda_i - 2 \| A \|)^{-1} \leqq \alpha$ where φ is any one of the quantities $\| A^* - A \|$, $|||A^* - A|||$, $2\| A \|$, or $2|||A|||$ and $\alpha < 1$, then for p fixed the hypothesis (†) of Theorem 3.4 holds for all $n \geqq p$ and $\lim_{n\to\infty} \epsilon_p(n) = 0$.

Proof. $\epsilon_p(n)$ is the product of

$$\left(\frac{1 + \varphi \max_k \sum_{j\neq k} |\eta_k(n) - \bar{\eta}_j(n)|^{-1}}{1 - \varphi \max_k \sum_{j\neq k} |\eta_k(n) - \bar{\eta}_j(n)|^{-1}}\right)^{\frac{1}{2}}$$

and

$$\frac{\| A \|^2}{((\lambda_{n+1} - \lambda_p) - 2(\lambda_{n+1} - \lambda_p) \| A \|)^{\frac{1}{2}}}.$$

The second factor converges to zero since $\lim_{n\to\infty} \lambda_n = \infty$. Now we show that the first factor is bounded in n. For any distinct k and j between 1 and n we have

$$|\eta_k(n) - \bar{\eta}_j(n)| \geqq |\lambda_k - \lambda_j| - 2\| A \| > 0.$$

Thus

$$\begin{aligned}
&\sum_{j=1, j\neq k}^{n} |\eta_k(n) - \bar{\eta}_j(n)|^{-1} \\
&\quad \leqq \sum_{j=1, j\neq k}^{n} (|\lambda_k - \lambda_j| - 2\| A \|)^{-1} \\
&\quad \leqq \sum_{j=1}^{k-1} (|\lambda_k - \lambda_j| - 2\| A \|)^{-1} + \sum_{j=k+1}^{n} (|\lambda_k - \lambda_j| - 2\| A \|)^{-1} \\
&\quad \leqq \sum_{j=1}^{k-1} (|\lambda_{j+1} - \lambda_j| - 2\| A \|)^{-1} + \sum_{j=k}^{n-1} (|\lambda_j - \lambda_{j+1}| - 2\| A \|)^{-1} \\
&\quad \leqq \sum_{j=1}^{\infty} (\lambda_{j+1} - \lambda_j - 2\| A \|)^{-1}.
\end{aligned}$$

Hence $\varphi \max_k \sum_{j\neq k} |\eta_k(n) - \bar{\eta}_j(n)|^{-1} \leqq \alpha$ for all n and this shows that (†) is satisfied. We also have that

$$\left(\frac{1 + \varphi \max_k \sum_{j\neq k} |\eta_k(n) - \bar{\eta}_j(n)|^{-1}}{1 - \varphi \max_k \sum_{j\neq k} |\eta_k(n) - \bar{\eta}_j(n)|^{-1}}\right)^{\frac{1}{2}} \leqq \left(\frac{1 + \alpha}{1 - \alpha}\right)^{\frac{1}{2}}$$

for all n. This completes the proof.

From the above proof we see that under the conditions of Theorem 3.5 we can use

$$\frac{[(1 + \alpha)/(1 - \alpha)]^{\frac{1}{2}} \| A \|^2}{((\lambda_{n+1} - \lambda_p)^2 - 2(\lambda_{n+1} - \lambda_p) \| A \|)^{\frac{1}{2}}} \equiv \epsilon_p'(n)$$

instead of $\epsilon_p(n)$ as an error estimate. Also if $\epsilon > 0$ we can solve the inequality $\epsilon_p'(n) < \epsilon$ for λ_{n+1}. We get

$$\lambda_{n+1} > \lambda_p + \| A \| \left[1 + \left(1 + \frac{1 + \alpha}{1 - \alpha} \frac{\| A \|^2}{\epsilon^2}\right)^{\frac{1}{2}}\right].$$

This inequality enables one to choose a value of n which will guarantee that the error is less then ϵ.

In the paragraph following the proof of Theorem 3.4 we mentioned the possibility of concluding that $i_0 = p$ where i_0 was defined in Theorem 3.4. Under the hypothesis of Theorem 3.5 this can be done as follows. If $|\mu_p - \eta_j(n)| > \epsilon_p(n)$ for $j = 1, \cdots, p-1, p+1, \cdots, n$ then we clearly have that $i_0 = p$. Since $|\mu_p - \eta_j(n)| \geqq |\lambda_p - \lambda_j| - 2\|A\|$ and $\epsilon_p(n) \leqq \epsilon_p'(n)$ it is sufficient that $|\lambda_p - \lambda_j| - 2\|A\| > \epsilon_p'(n)$ for $j = 1, \cdots, p-1, p+1, \cdots, n$. This condition can be verified without calculating the Rayleigh-Ritz eigenvalues. Compare example (b) of Section 4.

4. Examples. (a) Let H be the sequence space l_2. Let V be the subset of sequences $z = (z_1, z_2, \cdots)$ such that $\sum_{i=1}^{\infty} i^4 |z_i|^2 < \infty$. Let $L : V \to H$ be defined by $Lz = (z_1, 4z_2, \cdots, m^2 z_m, \cdots)$. L^{-1} is a positive definite, Hilbert-Schmidt operator on H. The eigenvalues of L are $\lambda_i = i^2$, $i = 1, 2, \cdots$ and the eigenvectors are the natural basis vectors in l_2. Let $Az \equiv \frac{1}{2}(z_1 + z_2, \frac{1}{2}(z_1 + z_3), \frac{1}{3}(z_1 + z_4), \cdots)$. A is bounded; in fact if it is regarded as an operator on H it is a Hilbert-Schmidt operator and $\|A\| \leqq |||A||| = \pi/2\sqrt{3} \simeq .907$. Let $\tilde{L} = L + A$.

Since $\|A\| < \lambda_1$ and the circles C_m are disjoint we can apply Theorem 2.2 and conclude that the eigenvalues of $\tilde{L}$ are contained in the circles C_m, one eigenvalue in each circle. Since the eigenvalues of $\tilde{L}$ occur in conjugate pairs we see by Corollary 2.2 that the eigenvalues of $\tilde{L}$ are real.

The hypothesis (†) of Theorem 3.4, with $\varphi = |||\tilde{L}_n^* - \tilde{L}_n|||$, is satisfied for $n = 1, 2, \cdots, 20$. This was checked by explicit calculation. Hence if $p \leqq n$, then for some i_0, $1 \leqq i_0 \leqq n$, we have $|\mu_p - \lambda_{i_0}| \leqq \epsilon_p(n)$. By comparing the circles C_m and the computed values of $\epsilon_p(n)$ we can conclude that $i_0 = p$. The most precise information found on the location of μ_1 is that

$$1.44785 \leqq \mu_1 \leqq 1.45618.$$

Here we used $n = 20$, for an examination of the values of $\epsilon_1(n)$ showed that $\epsilon_1(20)$ was the smallest of the numbers $\epsilon_1(1), \cdots, \epsilon_1(20)$. The most precise information on μ_{20} is that

$$399.95441 \leqq \mu_{20} \leqq 400.04559.$$

(b) Let $H = l_2$ and $V = \{z \mid \sum_{i=1}^{\infty} i^8 |z_i|^2 < \infty\}$. Let $Lz = (z_1, 16z_2, \cdots, m^4 z_4, \cdots)$; the eigenvalues of L are $\lambda_i = i^4$, $i = 1, 2, \cdots$. Let

$$Az = i/2(z_1 + z_2 + \tfrac{1}{2}z_3 + \tfrac{1}{3}z_4 + \cdots, \tfrac{1}{2}z_1, \tfrac{1}{3}z_1, \cdots).$$

Clearly $\|A\| \leqq |||A||| \simeq .907$. Theorem 3.4 can be applied with $\varphi = |||\tilde{L}_n^* - \tilde{L}_n|||$. The results of the computations for $n = 1, 2$ are

$$\eta_1(1) = 1 + i/2, \qquad \epsilon_1(1) \leqq .059, \qquad \eta_1(2) = 1.008 + .501i,$$

$$\eta_2(2) = 15.992 - .001i, \qquad \epsilon_1(2) \leqq .011, \qquad \epsilon_2(2) \leqq .014.$$

Thus μ_1 is in the circle centered at $1.008 + .501i$ with radius .011 and μ_2 is in the circle centered at $15.992 - .001i$ with radius .014.

Since

$$\sum_{i=1}^{\infty} (\lambda_{i+1} - \lambda_i - 2 \| A \|)^{-1} < \sum_{i=1}^{\infty} ((i+4)^4 - i^4 - 2)^{-1}$$

$$< \frac{1}{13} + \frac{1}{63} + \sum_{i=3}^{\infty} (4i^3 - 2)^{-1} < \frac{76}{819} + \frac{1}{4} \sum_{i=2}^{\infty} i^{-3}$$

$$< \frac{76}{819} + \frac{1}{4} \int_1^{\infty} x^{-3}\, dx = \frac{1427}{6552},$$

we have

$$\varphi \textstyle\sum_{i=1}^{\infty} (\lambda_{i+1} - \lambda_i - 2 \| A \|)^{-1} < 1427/3276 < 1$$

if we take $\varphi = 2 \| A \|$. Thus the hypothesis of Theorem 3.5 is satisfied with $\alpha = 1427/3276$. Hence, if we take $\varphi = 2 \| A \|$, Theorem 3.4 can be applied for any value of $n \geqq p$ and $\lim_{n \to \infty} \epsilon_p(n) = 0$. Furthermore we always have $i_0 = p$ since $| \lambda_p - \lambda_j | - 2 \| A \| > \epsilon_p'(n)$ for all appropriate values of n, p, and j.

As was mentioned in the paragraph following the proof of Theorem 3.5 we can choose the value of n which will insure a stated level of accuracy. If, for instance, in this example we wish to approximate μ_1 to within .001 it is sufficient to choose $n = 10$.

BIBLIOGRAPHY

1. EHRLICH, L. W. "F4 UTEX MATSUB, Eigenvalues and eigenvectors of complex non-Hermitian matrices using the direct and inverse power methods and matrix deflation", *Systems Programs for CO-OP*, Control Data Corporation.
2. FICHERA, G. "Su un metodo del Picone per il calcolo degli autovalori e delli autosoluzioni", *Ann. Mat. Pura. Appl.*, Ser. 4, **40** (1955), pp. 239–259.
3. GAVURIN, M. K. "On estimates for eigen-values and vectors of a perturbed operator", *Doklady Akad. Nauk. SSSR (N.S.)* **96** (1954), pp. 1093–1095.
4. GOULD, S. H. *Variational methods for eigenvalue problems. An introduction to Weinstein's theory of intermediate problems, 2nd edition.* Toronto: University of Toronto Press, 1966.
5. MARCUS, M. AND MINC, H. *A survey of matrix theory and matrix inequalities.* Boston: Allyn and Bacon, Inc., 1964.
6. MYSOVSKIH, I. P. "On an estimate of the error in eigenvalues calculated by replacing the kernel by another close to it", *Mat. Sb. (N.S.)* **49** (91) (1959), pp. 331–340.
7. OSBORNE, E. E. "On acceleration and matrix deflation processes used with the power method", *J. Soc. Indust. Appl. Math.* **6** (1958), pp. 279–287.
8. PETROV, G. I. "Estimation of accuracy in the approximate calculation of an eigenvalue by the method of Galerkin", *Prikl. Mat. Meh.* **21** (1957), pp. 184–188.
9. SCHWARTZ, J. T. "Perturbations of spectral operators, and applications. I. Bounded perturbations", *Pacific J. Math.* **4** (1954), pp. 415–458.
10. ZAANEN, A. C. *Linear analysis.* New York: Interscience Publishers Inc. (1953).

UNIVERSITY OF MINNESOTA
UNIVERSITY OF MARYLAND

(Received October 22, 1965)