Estimates for Some Computational Techniques in Linear Algebra

By Shmuel Kaniel*

1. In this paper we shall be concerned with two methods. The first one is that of conjugate gradients or least squares used for approximate computation of $Ax = f$ where A is a positive definite (symmetric) $n \times n$ matrix [2, 3, 6, 11]. It amounts to minimizing (with respect to the α_i) the expression:

$$\left\| A \sum_{i=0}^{k-1} \alpha_i A^i f - f \right\| \tag{1.1}$$

where k is a predetermined integer (usually much smaller than n) and where $\| \ \|$ denotes the L^2 norm. Then $g = \sum_{i=0}^{k-1} \alpha_i A^i f$ is taken as an approximate solution.

The second method is the generalized gradient or minimal iteration method used for an approximate computation of eigenvalues and eigenfunctions [3, 5].

It can be described as follows: Denote by H_k the subspace spanned by the vectors $f, Af, \cdots, A^k f$; denote by P the orthogonal projection on H_k and by B the restriction of PA to H_k (it is a well-defined $k+1 \times k+1$ matrix). Then $\mu_{\max}$, the largest eigenvalue of B is an approximation to $\lambda_{\max}$, the largest eigenvalue of A, likewise $\mu_{\min}$ approximates $\lambda_{\min}$ (we exclude here certain singular cases). There are, though, quite a few computational techniques for a numerical solution of this problem.

In Section 2 we shall discuss the matrix B. In Section 3 we shall establish a minimax property of general positive measure that will be needed in Sections 4 and 5.

In Section 4 we shall derive estimates for the conjugate gradients method. In Section 5 we shall discuss the rate of convergence of the generalized gradient method in terms of λ_i (the eigenvalues of A). It turns out that we get fast convergence for the leading eigenvalues of A (positive or negative); the convergence process for the positive eigenvalues is not perturbed so much by the existence of negative eigenvalues having large modulus and the effect of having two close eigenvalues is limited.

Let us note that there are some global estimates of the error that do not depend on the initial function f nor on the distribution of the eigenvalues of A. These estimates depend on k and $\| A \|$ only [9]. However the "guaranteed" rate of convergence is very slow and can not serve as a basis for computational techniques. In these estimates the convergence is not necessarily towards the largest or smallest eigenvalue.

The reader will notice that the order of the matrix A does not play any role. Therefore the estimates presented in this paper can be carried to the more general case where A is a Hilbert space operator. The theorems and proofs are exactly the same.

2. The Auxiliary Matrix B.

PROPOSITION 2.1. *Let $p(x)$ be a polynomial satisfying $| p(\lambda_i) | \leqq \epsilon$ where λ_i are the eigenvalues of A then for any vector $f \in H$: $\| p(A)f \| \leqq \epsilon \| f \|$.*

Received June 11, 1964. Revised August 16, 1965.

* This work was sponsored in part by the Army Research Office DA-ARO(B)-3-124-G 157

PROPOSITION 2.2. *If for some* μ *and* f: $\| Af - \mu f \| \leqq \epsilon \| f \|$ *then there exists* λ *eigenvalue of* A *so that* $| \lambda - \mu | \leqq \epsilon$.

These two propositions are a direct consequence of the spectral decompositio theorem.

Let us prove now:

THEOREM 2.1. *Let* $p(x)$ *be a polynomial of degree at most* k *satisfying* $| p(\mu_i) | \leqq$ *where* μ_i *are the eigenvalues of* B *then*:

$$\| p(A)f \| \leqq \epsilon \| f \|. \tag{2.1}$$

Proof. B operates in $H_k \subset H$ so for any $q(x)$: $q(B)f$ is well defined and belong to H_k.

We shall prove inductively that for $i = 0, 1, \cdots, k$, $A^i f = B^i f$. Indeed:

$$A^i f = PA^i f = PAA^{i-1} f = PA(PA)^{i-1} f = BB^{i-1} f = B^i f.$$

Therefore if $p(x)$ has degree at most k then $p(A)f = p(B)f$. Now by Proposition 1 $\| p(B)f \| \leqq \epsilon f$ from which (2.1) follows.

PROPOSITION 2.3. *If* λ_i^+ *is the ith largest eigenvalue of* A; μ_i^+ *the ith largest eigen value of* B; λ_i^- *the ith smallest eigenvalue of* A; μ_i^- *the ith smallest eigenvalue of* B *then*

$$\| A \| \geqq \lambda_i^+ \geqq \mu_i^+. \tag{2.2}$$

$$-\| A \| \leqq \lambda_i^- \leqq \mu_i^-. \tag{2.3}$$

These inequalities are direct consequences of the minimax principle. For example

$$\begin{aligned} \lambda_i^+ &= \min_{\phi_1, \phi_2, \cdots, \phi_{i-1}} \max_{f \perp \phi_1, \phi_2, \cdots, \phi_{i-1}} \frac{(Af, f)}{\|f\|^2} \\ &\geqq \min_{\phi_1, \phi_2, \cdots, \phi_{i-1}} \max_{f \perp \phi_1, \phi_2, \cdots, \phi_{i-1}} \frac{(Af, f)}{\|f\|^2}, \qquad f \in H_k \\ &= \min_{\psi_1, \psi_2, \cdots, \psi_{i-1}} \max_{f \perp \psi_1, \cdots, \psi_{i-1}} \frac{(Af, f)}{\|f\|^2}, \qquad f \in H_k \end{aligned}$$

where $\psi_j = P\phi_j$. If $\langle \phi_1, \cdots, \phi_{i-1} \rangle$ get all possible values in H then $\langle \psi_1, \cdots, \psi_{i-1} \rangle$ get all possible values in H_k. Moreover for $f \in H_k$ we have by the definition of B: $(Af, f) = (Bf, f)$ so by the minimax principle the last expression is μ_i^+.

THEOREM 2.2. *If* B *has a multiple eigenvalue then* A *is invariant in* H_k *and* B *is the restriction of* A *to* H_k. *In this case if* $A^{-1}f$ *exists it belongs to* H_k *and the eigenvalues of* B *coincide with some eigenvalues of* A.

Proof. It is sufficient to prove that A is invariant in H_k because then for any $g \in H_k$, $Ag = PAg = Bg$ and the other properties are well-known properties of invariant subspaces. This can be restated as proving that for any j:

$$A^j f = \sum_{i=0}^{l} \alpha_{ij} A^i f, \qquad l \leqq k, \tag{2.4}$$

which will be proved by induction.

Let g_1 and g_2 be two different eigenfunctions that correspond to the same eigenvalue μ. $g_1 = \sum_{i=0}^{k} \beta_i A^i f$; $g_2 = \sum_{i=0}^{k} \gamma_i A^i f$. There exists a linear combination of g_1 and g_2 that satisfies $g = \sum_{i=0}^{k-1} \alpha_i A^i f$. We have by the definition of P:

$$0 = PAg - \mu g = PA\sum_{i=0}^{k-1} \alpha_i A^i f - \mu \sum_{i=0}^{k-1} \alpha_i A^i f = \sum_{i=0}^{k-1} \alpha_i A^{i+1} f - \mu \sum_{i=0}^{k-1} \alpha_i A^i f.$$

So for some $l \leqq k$ we have:

$$A^{l+1} f = \sum_{i=0}^{l} \alpha_{il+1} A^i f. \tag{2.5}$$

Suppose now that (2.4) holds for j:

$$A^{j+1} f = AA^j f = A\sum_{i=0}^{l} \alpha_{ij} A^i f = \sum_{i=0}^{l-1} \alpha_{ij} A^{i+1} f + \alpha_{lj} A^{l+1} f$$

$$= \sum_{i=0}^{l-1} \alpha_{ij} A^{i+1} f + \alpha_{lj} \sum_{i=0}^{l} \alpha_{il+1} A^i f.$$

The last expression can be regrouped as in (2.4) to complete the induction and the proof of the theorem.

THEOREM 2.3. *The conclusions of Theorem* 2.2 *remain valid if some eigenfunction of* A *belongs to* H_k.

Proof. In view of the proof of Theorem 2.2 it is sufficient to establish (2.5). Indeed, since we can express the eigenfunction g as $g = \sum_{i=0}^{l} \alpha_i A^i f$ where $l \leqq k$ it follows that the equation $Ag - \lambda g = 0$ can be written in the desired form.

3. A Minimax Property for Measures.

THEOREM 3.1. *Let* V *be a positive measure supported by the interval* $[\alpha, \beta]$ *and satisfying:* $|V| = \int_\alpha^\beta dV = C$. *Let* γ *be outside* $[\alpha, \beta]$. *Let* $q(x)$ *be polynomials of degree not more than* k *that satisfy* $q(\gamma) = 1$. *Then:*

$$\sup_{|V|=C} \min_{q(x)} \left(\int_\alpha^\beta |q(x)|^2 \, dv \right)^{1/2} = (C)^{1/2} \left[T_k \left(\frac{2\gamma - (\beta + \alpha)}{\beta - \alpha} \right) \right]^{-1} \tag{3.1}$$

where $T_k(x)$ *is the Tchebisheff polynomial of order* k *i.e.*

$$T_k(x) = \tfrac{1}{2}\{(x + (x^2 - 1)^{1/2})^k + (x - (x^2 - 1)^{1/2})^k\}.$$

Proof. Without loss of generality we can transform α to -1 and β to 1. This way γ will be transformed to $\nu = (2\gamma - (\beta + \alpha))/(\beta - \alpha)$.

Let us allow the function under the integral sign to be any polynomial $p(x)$ of degree $2k$ that is nonnegative on $[-1, 1]$ and satisfies $p(\nu) = 1$. This set is convex and compact. The set of all positive measures V on $[-1, 1]$ that satisfy $|V| = C$ is convex and compact in the weak* topology of $C(-1, 1)$ (uniform convergence). Moreover:

$$\int_{-1}^{1} \{tp_1(x) + (1 - t)p_2(x)\} \, dV = t \int_{-1}^{1} p_1(x) \, dV + (1 - t) \int_{-1}^{1} p_2(x) \, dV.$$

$$\int_{-1}^{1} p(x) \, d[tV_1 + (1 - t)V_2] = t \int_{-1}^{1} p(x) \, dV_1 + (1 - t) \int_{-1}^{1} p(x) \, dV_2.$$

These are the standard conditions [10] that insure:

$$\sup_{|V|=C} \min_{p(x)} \int p(x) \, dV = \min_{p(x)} \max_{|V|=C} \int p(x) \, dV, \qquad p(\nu) = 1.$$

For any fixed polynomial the measure that maximized the integral is $C\delta(x_0)$ where x_0 is a maximum for $p(x)$. Hence:

$$\min_{p(x)} \max_{|V|=C} \int p(x)\, dV = \min_{p(x)} \max_{-1 \leqq x \leqq 1} p(x) \cdot C, \qquad p(\nu) = 1.$$

It is well known that this minimum is attained for the polynomial $[T_k(x)/T_k(\nu)]^2$. Thus the proof is complete.

Remark. The theorem can be inferred from [13]. We believe that our proof, though not elementary, is shorter.

4. The Conjugate Gradients Method. In this section we shall use the theory of polynomials of best approximation.

Definition. Let $f(z)$ be a continuous function on a closed set of complex numbers D. Let k be a fixed integer.

A polynomial $t(z)$ of degree at most k is said to be *the polynomial of best approximation to $f(x)$ on D if:*

$$\max_{z \in D} | t(z) - f(z) | = \alpha$$

while for any polynomial $r(z) \neq t(z)$ of degree at most k:

$$\max_{z \in D} | r(z) - f(z) | > \alpha.$$

If D contains at least $k + 1$ points then $t(z)$ exists and is unique.

We shall use the following particular case:

THEOREM 4.1. *Let x_i, $i = 1, \cdots, k + 2$, be real numbers satisfying: $x_1 < x_2 \cdots < x_{k+2}$ and let $\{y_i\}$, $i = 1, \cdots, k + 2$ be any sequence of real numbers then the polynomial $t(x)$ of degree k that comes closest to y_i on x_i satisfies:*

$$t(x_i) - y_i = (-1)^i \eta \alpha \tag{4.1}$$

where $\eta = \pm 1$.

Conversely if $t(x)$ satisfies (4.1) for some α then it is the polynomial of best approximation to y_i on x_i.

This is a well-known theorem. For the proof cf. [15].

Now we can restate the minimization problem as follows: Find the polynomial $p(x)$ that has degree at most k and satisfies $p(0) = -1$ so that $\| p(A)f \|$ is minimal. We can estimate the error $\| p(A)f \|$ by estimating $\| q(A)f \|$ where $q(x)$ is any polynomial satisfying: $q(0) = -1$. In view of Theorem 2.1 a good choice of $q(x)$ will be the polynomial that has least deviation from zero on μ_i. In order to find it let us consider the polynomial of best approximation to -1 on 0 and to 0 on μ_i, $i = 1, 2, \cdots, k + 1$ ($k + 2$ points in total). Define a polynomial $r(x)$:

$$r(x) = \sum_{j=1}^{k+1} (-1)^j \frac{\prod_{i \neq j} (x - \mu_i)}{\prod_{i \neq j} (\mu_j - \mu_i)}.$$

It is easy to see that $r(\mu_j) = (-1)^j$ and that $q(x) = -r(x)/r(0)$ is the desired polynomial so we have:

THEOREM 4.2. *Let g be the approximate solution of $Ax = f$ which is constructed by*

the conjugate gradients method. Then:

$$\| Ag - f \| \leqq \left(\sum_{j=1}^{k+1} \left| \frac{\prod_{i \neq j} \mu_i}{\prod_{i \neq j} (\mu_j - \mu_i)} \right| \right)^{-1} \| f \|, \tag{4.2}$$

where μ_i *are the eigenvalues of* B.

If we define $T(x)$ to be the polynomial that has the least deviation from 0 on the interval $[\mu_{\min}, \mu_{\max}]$ and satisfies $T(0) = -1$ we can use it to get a simpler estimate than (4.2). If the deviation from zero is ϵ, then $| T(\mu_i) | \leqq \epsilon$ and we may apply Theorem 2.2. It is well known that:

$$-T(x) = T_k\left(\frac{\mu_{\max} + \mu_{\min}}{\mu_{\max} - \mu_{\min}} - \frac{2}{\mu_{\max} - \mu_{\min}} x\right)\left[T_k\left(\frac{\mu_{\max} + \mu_{\min}}{\mu_{\max} - \mu_{\min}}\right)\right]^{-1}$$

and that the deviation ϵ is the last factor. So we get:

THEOREM 4.3. *Under the hypotheses of Theorem* 3.2 *the following is true:*

$$\| Ag - f \| \leqq \left[T_k\left(\frac{\mu_{\max} + \mu_{\min}}{\mu_{\max} - \mu_{\min}}\right)\right]^{-1} \| f \|. \tag{4.3}$$

It is obvious that (4.3) is worse than (4.2). If we use inequality (2.2) we can reduce (4.3) to:

$$\| Ag - f \| \leqq \left[T_k\left(\frac{\lambda_{\max} + \lambda_{\min}}{\lambda_{\max} - \lambda_{\min}}\right)\right]^{-1} \| f \|. \tag{4.4}$$

This result is known [13]. It can be established without using the auxiliary matrix B. It follows directly from Proposition 2.1. Certainly (4.4) is worse than (4.3). In the general case one can hope that (4.2) is considerably better than (4.4) as the following example shows:

Example. If $\lambda_{\min}/\lambda_{\max}$ *is small,* $\mu_{\min} = \mu_1 < \mu_2 < \cdots < \mu_{2k+1} = \mu_{\max}$ *and* $\mu_{i+1} - \mu_i = d$, *then estimate* (4.2) *yields*:

$$\| Ag - f \| < \left[\frac{\mu_{\min}}{\mu_{\max}} \binom{2k}{k}\right]^{-1} \| f \| \sim \frac{\mu_{\max}}{\mu_{\min}} \cdot k^{1/2} \cdot 2^{-2k} \| f \|.$$

Estimate (4.4) results in:

$$\| Ag - f \| \leqq \left[T_k\left(\frac{\lambda_{\max} + \lambda_{\min}}{\lambda_{\max} - \lambda_{\min}}\right)\right]^{-1} \| f \| \sim 2\left(1 + 2\left(\frac{\lambda_{\min}}{\lambda_{\max}}\right)^{1/2}\right)^{-2k} \| f \|.$$

The first estimate is obtained by taking into account only the middle term of (4.2), while the second estimate is obtained by approximating $T_k(1 + \delta)$ by $\frac{1}{2}(1 + (2\delta)^{1/2})^k$.

We did not assume a particularly favorable distribution. A look at (4.2) reveals that if two eigenvalues are close one to the other, then the right-hand side of (4.2) is very small. In the limiting case where the two eigenvalues coincide we get $\| Ag - f \| = 0$. This result can be inferred from Theorem (2.2).

Nevertheless, estimate (4.4) cannot be improved. This fact is known [13]. Let us use Theorem 3.1 in order to prove it. By the spectral decomposition theorem we can find for any operator A and any function $f \in H$ a positive measure V so that:

$$| V | = \int_{\lambda_{\min}}^{\lambda_{\max}} dV = \| f \|^2; \qquad \| q(A)f \|^2 = \int_{\lambda_{\min}}^{\lambda_{\max}} | q(x) |^2 \, dV$$

for any polynomial $q(x)$. So if we substitute in Theorem 3.1 $\lambda_{\min}$ for α, $\lambda_{\max}$ for β and 0 for γ we get that the right-hand side is equal to

$$T_k((\lambda_{\max} + \lambda_{\min})(\lambda_{\max} - \lambda_{\min}))$$

proving (4.4) and showing that it is the best possible.

Still we have:

THEOREM 4.4. *If in* (4.4) *the equality sign holds for* $k = m$ *then we get an exact solution for* $k = m + 1$.

Proof. If the equality sign holds it follows that $\mu_{\max} = \lambda_{\max}$. Let $Bh = \mu_{\max} h$ hence:

$$\lambda_{\max} = \mu_{\max} = \frac{(Ah, h)}{\| h \|^2}.$$

Therefore by the minimax principle h is an eigenfunction of A. By Theorem 2.3 it follows that A is invariant in H_k therefore there exists $g \in H_k$ so that $Ag = f$, (Observe that in (1.1) we minimize the expression $\| Ah - f \|$ where $h \in H_{k-1}$.)

5. The Generalized Gradient Method. This method uses moments so like any other method of its kind [1] it detects only the eigenfunctions that are not orthogonal to the initial vector f. This means that we compute the eigenvalues of the restriction of A to some invariant subspace of H. In order to simplify the notations let us keep the same letter A for the reduced matrix. Therefore, we may assume that the eigenvalues of A are simple and that f is not orthogonal to any eigenfunction.

Let us denote the eigenfunctions of A corresponding to λ_i^+ and λ_i^- by ϕ_i^+ and ϕ_i^-, respectively. We shall also denote the eigenfunctions of $B = PA$ corresponding to μ_i^+ and μ_i^- by ψ_i^+ and ψ_i^-. First of all let us establish an estimate for the approximation of the largest (smallest) eigenvalue.

THEOREM 5.1. *Let* d *denote the distance between* $\lambda_{\max}$ *and the rest of the eigenvalues. Let* f *be normalized:* $\| f \| = 1$ *and let* b *denote* $(f, \phi_{\max})$. *Let* k *be chosen and* B *be constructed. Then:*

$$\lambda_{\max} \geqq \mu_{\max} \geqq \lambda_{\max} - \frac{\lambda_{\max} - \lambda_{\min}}{b^2} \left[T_k\left(\frac{\lambda_{\max} - \lambda_{\min} + d}{\lambda_{\max} - \lambda_{\min} - d} \right) \right]^{-2}. \tag{5.1}$$

Proof. By the spectral decomposition theorem there exists a positive measure V that is supported by the interval $[\lambda_{\min}, \lambda_{\max} - d]$ and the point $\lambda_{\max}$ so that:

$$1 = \| f \|^2 = | V | = b^2 + \int_{\lambda_{\min}}^{\lambda_{\max} - d} dV.$$

$$\| q(A)f \|^2 = q^2(\lambda_{\max}) b^2 + \int_{\lambda_{\min}}^{\lambda_{\max}} | q(x) |^2 \, dV.$$

Let us try to estimate $\lambda_{\max}$ by constructing a polynomial $p(x)$ that has the degree at most k, satisfies: $p(\lambda_{\max}) = 1$ and minimizes the expression $\int_{\lambda_{\min}}^{\lambda_{\max} - d} | p(x) |^2 \, dV$.

Hence we have to use Theorem 3.1 for $\alpha = \lambda_{\min}$, $\beta = \lambda_{\max} - d$, $\gamma = \lambda_{\max}$ and $C = \| f \|^2 - b^2$ to get:

$$(5.2)\qquad \int_{\lambda_{\min}}^{\lambda_{\max}-d} |p(x)|^2\,dV \leqq (\|f\|^2 - b^2)\left[T_k\left(\frac{\lambda_{\max}-\lambda_{\min}+d}{\lambda_{\max}-\lambda_{\min}-d}\right)\right]^{-2}.$$

Define now $h = p(B)f$. We have:

$$\lambda_{\max} = \max_{g\in H_k}\frac{(Ag,g)}{\|g\|^2} \geqq \frac{(Ah,h)}{\|h\|^2} = \frac{\lambda_{\max}b^2 + \int_{\lambda_{\min}}^{\lambda_{\max}-d} x\,|p(x)|^2\,dV}{b^2 + \int_{\lambda_{\min}}^{\lambda_{\max}-d}|p(x)|^2\,dV}$$

$$\geqq \lambda_{\max} - \frac{(\lambda_{\max}-\lambda_{\min})\int_{\lambda_{\min}}^{\lambda_{\max}-d}|p(x)|^2\,dV}{b^2 + \int_{\lambda_{\min}}^{\lambda_{\max}-d}|p(x)|^2\,dV}.$$

Taking (5.2) into account we get the right-hand side of (5.1). The left-hand side is obvious.

Remark. Observe, as in Section 4, that for small d the rate of convergence is about $(1 + (d/|\lambda_{\max}|)^{1/2})^{-2}$.

We need now a lemma that will enable us to take advantage of a particular distribution of the eigenvalues.

LEMMA 5.1. *Let the eigenvalues of A be divided into three sets: L having l elements; M and the set consisting of λ_j^+ only. Let any eigenvalue in M satisfy*:

$$\lambda_i < \lambda_j^+, \qquad \lambda_i \in M.$$

Let us denote by d_j the distance between λ_j^+ and M and by $\gamma_{\min}$ the smallest eigenvalue belonging to M. Let b_j denote (f, ϕ_j). Let k be chosen and H_k be constructed. Under these conditions there exists $g \in H_k : \|g\| = 1$ so that:

$$(5.3)\qquad \frac{(Ag,g)}{\|g\|^2} \geqq \lambda_j^+ - (\lambda_j^+ - \gamma_{\min})\left[b_j\prod_{\lambda_i\in L}(\lambda_j^+ - \lambda_i)\right]^{-2} \cdot\left[T_{k-l}\left(\frac{\lambda_j^+ - \gamma_{\min}+d_j}{\lambda_j^+ - \gamma_{\min}-d_j}\right)\right]^{-2}.$$

Proof. Consider $f_L = \prod_{\lambda_i\in L}(A - \lambda_i I)f$ and consider the iterates $A^j f_L$. For any j and any i: $\lambda_i \in L$, $A^j f_L$ is orthogonal to ϕ_i. Therefore we may consider instead of H the subspace $\bar{H}$ which is the orthogonal complement of all $\phi_i : \lambda_i \in L$ and $\bar{A}$ the restriction of A to $\bar{H}$. Denote by $H_k(L)$ the span of $\{f_L, Af_L, \cdots, A^{k-l}f_L\}$; by P_L the projection on $H_k(L)$ and by B_L the restriction of $P_L\bar{A}$ to $H_k(L)$.

We substitute now in Theorem 5.1 $\bar{A}$ for A; f_L for f; $k - l$ for k; $H_k(L)$ for H_k and B_L for B. By the proof of Theorem 5.1 we get that there exists $g \in H_k(L)$ satisfying (5.3). Since $H_k(L)$ is composed of linear combinations of $A^i f$, $0 \leqq i \leqq k$, it follows that $H_k(L) \subset H_k$. Thus the proof is complete.

Let us illustrate the use of Lemma 5.1 in the following:

THEOREM 5.2. *Let ϵ denote the distance between $\lambda_{\max}$ and λ_2^+; let d denote the distance between $\lambda_{\max}$ and the rest of the eigenvalues then*

$$(5.4)\qquad \lambda_{\max} \geqq \mu_{\max} \geqq \lambda_{\max} - \frac{\lambda_{\max}-\lambda_{\min}}{\epsilon^2 b^2}\cdot\left[T_{k-1}\left(\frac{\lambda_{\max}-\lambda_{\min}+d}{\lambda_{\max}-\lambda_{\min}-d}\right)\right]^{-2}.$$

Proof. Take the set L to be λ_2^+ and substitute in Lemma 5.1.

We see in this example that if λ_2^+ is close to $\lambda_{\max}$ then the rate of convergence is measured in terms of the distance to λ_3^+. We start, of course, with a large initial error.

As another application of Lemma 5.1 let us consider the approximation to the few largest eigenvalues. For that we need the following.

LEMMA 5.2. *Let* $\phi_i^+ = \psi_i^+ + \eta_i$ *where* $\| \eta_i \| = \epsilon_i$, *let* $g \in H_k : \| g \| = 1$ *be orthogonal to* $\phi_1^+, \phi_2^+, \cdots, \phi_{j-1}^+$. *Then*:

$$\lambda_j^+ \geqq \mu_j^+ \geqq (Ag, g) - \sum_{i=1}^{j-1} \mu_i^+ \epsilon_i^2 \geqq (Ag, g) - \sum_{i=1}^{j-1} \lambda_i^+ \epsilon_i^2. \tag{5.5}$$

Proof. Express g as: $g = g_1 + \sum_{i=1}^{j-1} \alpha_i \psi_i^+$ where $g_1 \perp \psi_i^+$, $i = 1, \cdots, j-1$. Taking scalar product with ψ_i^+ we have:

$$| \alpha_i | = | (g, \phi_i^+) | = | (g, \psi_i^+ - \eta_i) | = | (g, \eta_i) | \leqq \epsilon_i .$$

Since ψ_i^+ satisfy $(A\psi_i^+, h) = (\mu_i^+ \psi_i^+, h)$ for any $h \in H_k$ we have:

$$\frac{(Ag, g)}{\| g \|^2} = \frac{(Ag_1, g_1)}{\| g \|^2} + \frac{\sum_{i=1}^{j-1} \mu_i^+ \alpha_i^2}{\| g \|^2} \leqq \frac{(Ag_1, g_1)}{\| g_1 \|^2} + \sum_{i=1}^{j-1} \mu_i^+ \alpha_i^2.$$

Since $g_1 \in H_k$ and is orthogonal to ψ_i^+ we have by the minimax principle: $((Ag_1, g_1) \cdot \| g_1 \|^2)^{-1} \leqq \mu_j^+$ from which the right-hand side of (5.5) follows. The left-hand side of (5.5) was proved in Section 2.

We see that (5.5) involves errors in the eigenfunctions so we need a lemma that relates the errors in the eigenfunctions to the errors in the eigenvalues.

LEMMA 5.3. *Let* $\phi_i^+ = \psi_i^+ + \eta_i$ *where* $\| \eta_i \| = \epsilon_i$, *denote by* d_j *the distance between* λ_j^+ *and the smaller eigenvalues, and let* δ_j *denote the error in the* j*th largest eigenvalues i.e.* $\delta_j = \lambda_j^+ - \mu_j^+$. *Then*

$$\epsilon_j^2 \leqq \frac{\delta_j + \sum_{i=1}^{j-1} (\lambda_i^+ - \lambda_j^+) \epsilon_i^2}{d_j} + \sum_{i=1}^{j-1} \epsilon_i^2. \tag{5.6}$$

Proof. Express ψ_j^+ by: $\psi_j^+ = \sum_i \alpha_i \phi_i^+$. In a similar way to the proof of Lemma 5.2 we have for $i \leqq j - 1$

$$| \alpha_i | = | (\psi_j^+, \phi_i^+) | = | (\psi_j^+, \psi_i^+ + \eta_i) | = | (\psi_i^+, \eta_i) | \leqq \epsilon_i .$$

So:

$$\begin{aligned} \delta_j &= \lambda_j^+ - \mu_j^+ = \lambda_j^+ (\psi_j^+, \psi_j^+) - (A\psi_j^+, \psi_j^+) \\ &= \sum_{i=1}^{j-1} (\lambda_j^+ - \lambda_i^+) \alpha_i^2 + \sum_{i>j} (\lambda_j^+ - \lambda_i^+) \alpha_i^2. \end{aligned}$$

Therefore:

$$\sum_{i>j} \alpha_i^2 \leqq \frac{\delta_j + \sum_{i=1}^{j-1} (\lambda_i^+ - \lambda_j^+) \alpha_i^2}{d_j} .$$

Noting that $\epsilon_j^2 = \sum_{i \neq j} \alpha_i^2$ we get the desired result.

Remark. Observe that (5.5) and (5.6) relate errors in the eigenvalues to squares of errors in the eigenfunctions.

Combining Lemmas 5.1, 5.2 and 5.3 we get:

THEOREM 5.3. *Under the conditions of Lemma* 5.1 *the following estimate holds*:

$$\lambda_j^+ \geqq \mu_j^+ \geqq \lambda_j^+ - (\lambda_j^+ - \lambda_{\min}) \left[b_j \cdot \prod_{i=1}^{j-1} (\lambda_i^+ - \lambda_j^+) \right]^{-2} \cdot \left[T_{k-j+1} \left(\frac{\lambda_j^+ - \lambda_{\min} + d_j}{\lambda_j^+ - \lambda_{\min} - d_j} \right) \right]^{-2} - \sum_{i=1}^{j-1} \lambda_i^+ \epsilon_i^2. \tag{5.7}$$

Proof. Take in Lemma 5.1 the set L to consist of the eigenvalues λ_i^+ for $i < j$. The function g defined in (5.3) satisfies the requirements of Lemma 5.2. Therefore combining (5.3) and (5.5) we get the desired result.

Let us compute these estimates in the following example: $\lambda_{\max} = 1.00$, $\lambda_2^+ = 0.99$, $\lambda_3^+ = 0.96$, $\lambda_i \leqq 0.9$ for $i > 3$ and $\lambda_{\min} = 0$. Suppose, furthermore, that the components of f on the first three eigenfunctions have the absolute value 0.01. Then for $k = 52$ we have:

$$\mu_{\max} \geqq 1 - (0.01 \cdot 0.01 \cdot 0.04)^{-2} \cdot T_{50} \left(\frac{1 + 0 \cdot 1}{1 - 0 \cdot 1} \right)^{-2}$$

$$\geqq 1 - 16^{-1} \cdot 10^{12} \cdot 2 \cdot (1.9)^{-100} \geqq 1 - 10^{-16}.$$

Hence $\delta_1 \leqq 10^{-16}$; $\epsilon_1^2 \leqq 10^{-14}$.

$$\mu_2^+ \geqq 0.99 - (0.01 \cdot 0.01 \cdot 0.03)^{-2} \cdot T_{50} \left(\frac{1 + 0.09}{1 - 0.09} \right)^{-2} - 10^{-14}$$

$$\geqq 0.99 - 9^{-1} \cdot 10^{12} \cdot (1.78)^{-100} \geqq 0.99 - 10^{-13}.$$

Hence $\delta_2 \leqq 10^{-13}$; $\epsilon_2^2 \leqq 3.10^{-12}$.

$$\mu_3^+ \geqq 0.96 - (0.01 \cdot 0.03 \cdot 0.04)^{-2} \cdot T_{50} \left(\frac{1 + 0.06}{1 - 0.06} \right)^{-2} - 10^{-14} - 3.10^{-12}$$

$$\geqq 0.96 - 12^{-2} \cdot 10^{12} \cdot (1.6)^{-100} \geqq 0.96 - 10^{-10}.$$

Hence $\delta_3 \leqq 10^{-10}$; $\epsilon_3^2 \leqq 2.10^{-8}$.

If we use the power method we get $\delta_1 \leqq 10^{-2}$. Therefore the method described in this paper may be considered wherever the matrices are large and the computation of the few largest (or smallest) eigenvalues is needed.

The University of Chicago
Chicago, Illinois

Stanford University
Stanford, California

1. F. L. BAUER & A. S. HOUSEHOLDER, "Moments and characteristic roots," *Numer. Math.*, v. 2, 1960, pp. 42–53. MR **22** #1070.

2. E. BODEWIG, *Matrix Calculus*, 2nd rev. ed., North-Holland, Amsterdam; Interscience, New York, 1959. MR **23** #B563.

3. D. K. FADDEEV & V. N. FADDEEVA, *Computational Methods of Linear Algebra*, W. H. Freeman, San Francisco, Calif., 1963. MR **28** #1742.

4. M. K. GAVURIN, "The use of polynomials of best approximation for improving the convergence of iterative processes," *Uspehi Mat. Nauk*, v. 5, 1950, no. 3(37), pp. 156–160. (Russian) MR **12**, 209.

5. M. R. HESTENES & W. KARUSH, "A method of gradients for the calculation of the characteristic roots and vectors of a real symmetric matrix," *J. Res. Nat. Bur. Standards*, v. 47, 1951, pp. 45–61. MR **13,** 283.

6. M. R. HESTENES & E. STIEFEL, "Methods of conjugate gradients for solving linear systems," *J. Res. Nat. Bur. Standards*, v. 49, 1952, pp. 409–436. MR **15,** 651.

7. A. S. HOUSEHOLDER, *Principles of Numerical Analysis*, McGraw-Hill, New York, 1953. MR **15,** 470.

8. A. S. HOUSEHOLDER, *The Theory of Matrices in Numerical Analysis*, Blaisdell, New York, 1964. MR **30** #5475.

9. S. KANIEL, "On the approximation of symmetric operators by operators of finite rank," *Israel J. Math.*, v. 3, 1965, pp. 1–5.

10. S. KARLIN, *Mathematical Methods and Theory in Games Programming and Economics*, Vol 2: *The Theory of Infinite Games*, Addison-Wesley, London, 1959. MR **22** #2496.

11. C. LANCZOS, "Solution of systems of linear equations by minimized iterations," *J. Res. Nat. Bur. Standards*, v. 49, 1952, pp. 33–53. MR **14,** 501.

12. C. LANCZOS, *Applied Analysis*, Prentice-Hall, Englewood Cliffs, N. J., 1956. MR **18,** 823.

13. G. MEINARDUS, "Über eine Verallgemeinerung einer Ungleichung von L. V. Kantorowitsch," *Numer. Math.*, v. 5, 1963, pp. 14–23. MR **28** #3525.

14. E. L. STIEFEL, "Kernel polynomials in linear algebra and their numerical applications," *Nat. Bur. Standards Appl. Math. Ser.*, No. 49, 1958, pp. 1–22. MR **19,** 1080.

15. E. STIEFEL, "Über diskrete und lineare Tschebycheff-Approximationen," *Numer. Math.*, v. 1, 1959, pp. 1–28.