PROCEEDINGS OF THE
AMERICAN MATHEMATICAL SOCIETY
Volume 80, Number 2, October 1980

BEST APPROXIMATION OF A NORMAL OPERATOR IN THE SCHATTEN p-NORM

RICHARD BOULDIN

ABSTRACT. Let A be a fixed normal operator and let $\mathfrak{N}(\Lambda)$ denote the normal operators with spectrum contained in Λ. Provided there is some N in $\mathfrak{N}(\Lambda)$ such that $A - N$ belongs to the Schatten class c_p, $p > 2$, the main result of this paper obtains a best approximation for A from $\mathfrak{N}(\Lambda)$ with respect to the Schatten p-norm. A necessary and sufficient condition is given for A to have a unique best approximation in that case.

1. Introduction. If Λ is a closed nonempty set in the complex plane then $\mathfrak{N}(\Lambda)$ denotes the normal (bounded linear) operators on the fixed separable Hilbert space H with spectrum contained in Λ. For any compact operator T let $|T| = (T^*T)^{1/2}$ and let $s_1(T), s_2(T), \ldots$ be the eigenvalues of $|T|$ in nonincreasing order repeated according to multiplicity. If, for some $p > 1$, one has

$$\sum_{j=1}^{\infty} s_j(T)^p < \infty$$

then one says that T belongs to the *Schatten class* c_p which is normed with

$$\|T\|_p = \left(\sum_{j=1}^{\infty} s_j(T)^p \right)^{1/p}.$$

A good reference for the general theory of Schatten classes is **[8]**. The problem considered in this paper is to find a best approximation for a fixed normal operator A from $\mathfrak{N}(\Lambda)$ using the norm $\|\cdot\|_p$. The problem of determining when A has a unique best approximation is also considered.

2. Main results. In [12] P. R. Halmos constructed a best approximation of the fixed normal operator A from $\mathfrak{N}(\Lambda)$ using the usual operator norm. In order to state his result it is necessary to discuss the class of complex valued functions of a complex variable which are called retracts. One says that $F(z)$ is a *distance minimizing retract* onto Λ provided each $F(z)$ belongs to Λ and

$$|z - F(z)| \leq |z - \lambda| \quad \text{for all } \lambda \text{ in } \Lambda.$$

Provided Λ is closed and nonempty there is a Borel measurable distance minimizing retract onto Λ; see [12] for a nice proof. If Λ is convex and nonempty then there is a unique distance minimizing retract; see [13, Theorem 7.8, p. 94]. For A and Λ as above, the theorem of Halmos in [12] asserts that

$$\|A - F(A)\| \leq \|A - N\| \quad \text{for every } N \in \mathfrak{N}(\Lambda),$$

Received by the editors May 9, 1979 and, in revised form, August 24, 1979 and October 9, 1979.
AMS (MOS) subject classifications (1970). Primary 47B20.

where $F(z)$ is a Borel measurable distance minimizing retract onto Λ. Note $F(A)$ belongs to $\mathfrak{N}(\Lambda)$.

The main results are now stated; the proofs are given in the next section.

THEOREM 1. *Let A be a fixed normal operator with spectrum $\sigma(A)$. In order for there to exist some $N \in \mathfrak{N}(\Lambda)$ such that $A - N$ belongs to $c_p, p > 2$, it is necessary and sufficient that $\sigma(A) \setminus \Lambda$ is a (possibly empty or possibly infinite) countable set of finite dimensional isolated eigenvalues $\{\alpha_1, \alpha_2, \ldots, \alpha_l\}$, repeated according to multiplicity, such that $\Sigma_j(\operatorname{dist}(\alpha_j, \Lambda))^p$ is finite.*

THEOREM 2. *Let A be a fixed normal operator and let $F(z)$ be a Borel measurable distance minimizing retract of the complex plane onto Λ. If there exists some $N \in \mathfrak{N}(\Lambda)$ such that $A - N$ belongs to $c_p, p > 2$, then $A - F(A)$ belongs to c_p and*

$$\|A - F(A)\|_p \leq \|A - N\|_p. \qquad (*)$$

Furthermore, $F(A)$ is the unique choice of N producing equality in $(*)$ if and only if every point of $\sigma(A)$ has a unique closest point in Λ. In particular, if Λ is convex then equality in $(*)$ implies $N = F(A)$.

In the case that A is an invertible nonnegative operator and Λ is the unit circle, then the theorem was proved in [2] by means of Fréchet derivatives. It should be noted that if $F(z)$ is a distance minimizing retract onto the unit circle and A is an invertible nonnegative operator then $F(A)$ is the identity operator. The reformulation of the result given in [2] shows that it extends theorems in [1], [6], [7] which are relevant to quantum chemistry. Also, [10, Lemma 3.1, p. 323] is a special case of the theorem.

The assertion in the theorem that $A - F(A)$ belongs to c_p provides a remarkable contrast to previously known results about closure properties of c_2. Since $F(z) = z$ for every z in Λ, $F(N)$ equals N and the statement that $A - F(A)$ belongs to c_2 is equivalent to the statement that $F(A) - F(N)$ belongs to c_2. In [4] the best result of this type asserts that $f(V) - f(U)$ belongs to c_2 when $V - U$ belongs to c_2, V and U are unitary and $f(z)$ is a function on the unit circle with its derivative satisfying a Lipshitz condition.

Let $\Lambda = \{0, 1\}$ and $A = (1/2)P$ where P is the orthogonal projection onto some finite dimensional subspace of H. Then any orthogonal projection R onto a subspace of the range of P has the property that

$$\|A - F(A)\|_p \geq \|A - R\|_p$$

for $p > 1$ and any retract $F(z)$ onto Λ. Thus, the uniqueness statement of the theorem is false without some additional hypothesis.

3. Proof of the main results. For the reader's convenience a proof to the following well-known lemma is included.

LEMMA 1. *Let A be a fixed normal operator. If there exists some $N \in \mathfrak{N}(\Lambda)$ such that $(A - N) \in c_p$ then the only points in the spectrum of A, denoted $\sigma(A)$, not contained in Λ are isolated eigenvalues with finite multiplicity.*

Proof. Note that A is a compact perturbation of N. According to Weyl's theorem for normal operators, A and N have the same Weyl spectrum. The reader can find a contemporary discussion of Weyl's theorem in [3]. For any normal operator T the Weyl spectrum coincides with the points of $\sigma(T)$ which are not isolated eigenvalues with finite multiplicity. (See [5, Theorem 3] or [3, Theorem 5.1].) The operators for which the above set coincides with the Weyl spectrum are characterized in [11]. Since the Weyl spectrum of N–and, hence, the Weyl spectrum of A–is contained in Λ, the conclusion of the lemma follows.

Lemma 2. *If N is a normal operator, α is some scalar and e is some unit vector then*

$$\|(\alpha - N)e\| > \operatorname{dist}(\alpha, \sigma(N)). \qquad (*)$$

If there is a unique point β in $\sigma(N)$ which is closest to α and equality holds in $(*)$ then e is an eigenvector for N and β is the corresponding eigenvalue.

Proof. The proof of $(*)$ given in [12] is incorporated in the following. Let $E(\cdot)$ be the spectral measure of N and note that

$$\begin{aligned}\|(\alpha - N)e\|^2 &= \int_{\sigma(N)} |\alpha - z|^2 d\langle E(z)e, e\rangle \\ &> \int_{\sigma(N)} \operatorname{dist}(\alpha, \sigma(N))^2 d\langle E(z)e, e\rangle \\ &= \operatorname{dist}(\alpha, \sigma(N))^2.\end{aligned}$$

Thus, $(*)$ above holds.

Assume that equality holds in $(*)$ and β is the unique point of $\sigma(N)$ closest to α. It follows that

$$|\alpha - z| = \operatorname{dist}(\alpha, \sigma(N))$$

or

$$z = \beta$$

almost everywhere with respect to the measure $\langle E(\cdot)e, e\rangle$. Thus, one has

$$\|(N - \beta)e\|^2 = \int_{\sigma(N)} |z - \beta|^2 d\langle E(z)e, e\rangle = 0$$

and the lemma is proved.

Lemma 3. *Let T be in c_p and let $\{e_1, \ldots, e_l\}$ be a (possibly infinite) orthonormal set. Then one has the inequality*

$$\|T\|_p^p > \sum_{j=1}^{l} \langle |T|e_j, e_j\rangle^p$$

for $p > 1$.

Proof. See [10, Item 5, p. 94].

Lemma 4. *Let T be in c_p, $p > 2$. If $\{e_1, e_2, \ldots\}$ is an orthonormal sequence then $\|T\|_p^p > \Sigma_j \|Te_j\|^p$.*

PROOF.

$$\begin{aligned}\|T\|_p^p &= \|\,|T|\,\|_p^p = \sum_j s_j(|T|)^p \\ &= \sum_j s_j(|T|^2)^{p/2} = \|\,|T|^2\,\|_{p/2}^{p/2} \\ &\geq \sum_j \langle |T|^2 e_j, e_j\rangle^{p/2} \quad \text{by Lemma 3} \\ &= \sum_j \|Te_j\|^p.\end{aligned}$$

It is worth noting that if $\{e_j\}$ is an orthonormal basis then $\|T\|_2^2 = \Sigma_j \|Te_j\|^2$, while if $p = 1$, the reverse inequality holds and may be strict: $\|T\|_1 \leq \Sigma_j \|Te_j\|$.

PROOF OF THEOREM 1. Note that Lemma 1 applies to A and let $\{e_1, \ldots, e_l\}$ be a maximal orthonormal set of eigenvectors for A corresponding to the isolated eigenvalues $\{\alpha_1, \ldots, \alpha_l\}$ of A not contained in Λ. In order to show the inequality $(*)$ one observes the following

$$\begin{aligned}\|A - N\|_p^p &\geq \sum_j \|(A - N)e_j\|^p \quad \text{by Lemma 4} \\ &\geq \sum_j \operatorname{dist}(\alpha_j, \sigma(N))^p \quad \text{by Lemma 2} \\ &\geq \sum_j \operatorname{dist}(\alpha_j, \Lambda)^p.\end{aligned}$$

In order to prove the converse, write A as $A_1 \oplus A_2$ relative to the decomposition $H = E(\Lambda)H \oplus E(\Lambda^c)H$, where $E(\cdot)$ is the spectral measure of A and Λ^c means the complement of Λ. Note that $A_1 \in \mathfrak{N}(\Lambda)$ and $A_2 = \Sigma_{j=1}^l \langle \cdot, e_j\rangle \alpha_j e_j$ where $\{e_1, \ldots, e_l\}$ is a maximal orthonormal set of eigenvectors for A corresponding to $\{\alpha_1, \ldots, \alpha_l\}$. Note that

$$F(A) = A_1 \oplus F(A_2) = A_1 \oplus \sum_{j=1}^{l} \langle \cdot, e_j\rangle F(\alpha_j) e_j \in \mathfrak{N}(\Lambda).$$

Also observe that

$$\begin{aligned}\|A - F(A)\|_p^p &= \left\| 0 \oplus \sum_{j=1}^{l} \langle \cdot, e_j\rangle(\alpha_j - F(\alpha_j))e_j \right\|_p^p \\ &= \sum_{j=1}^{l} |\alpha_j - F(\alpha_j)|^p = \sum_{j=1}^{l} \operatorname{dist}(\alpha_j, \Lambda)^p < \infty.\end{aligned}$$

PROOF OF THEOREM 2. By Lemma 4 and Lemma 2, with the notation of the preceding proof, one obtains

$$\begin{aligned}\|A - N\|_p^p &\geq \sum_{j=1}^{l} \|(A - N)e_j\|^p \\ &\geq \sum_{j=1}^{l} \|(\alpha_j - N)e_j\|^p \geq \sum_{j=1}^{l} \operatorname{dist}(\alpha_j, \sigma(N))^p \\ &= \sum_{j=1}^{l} \operatorname{dist}(\alpha_j, \Lambda)^p = \sum_{j=1}^{l} |\alpha_j - F(\alpha_j)|^p.\end{aligned}$$

It will now be shown that the last sum is $\|A - F(A)\|_p^p$. Write A as $A_1 \oplus A_2$ relative to the decomposition $H = E(\Lambda)H \oplus E(\Lambda^c)H$, where $E(\cdot)$ is the spectral measure of A. Since $F(z) = z$ for all z in Λ one has

$$F(A) = F(A_1) \oplus F(A_2) = A_1 \oplus F(A_2).$$

Thus, if $\{f_1, f_2, \ldots\}$ is any orthogonal basis for $E(\Lambda)H$ then $\{e_1, \ldots, e_l, f_1, f_2, \ldots\}$ diagonalizes $A - F(A)$ and the corresponding eigenvalues are $\{\alpha_1 - F(\alpha_1), \ldots, \alpha_l - F(\alpha_l), 0, 0, \ldots\}$, respectively. It is now elementary that

$$\|A - F(A)\|_p^p = \sum_{j=1}^{l} |\alpha_j - F(\alpha_j)|^p$$

and, hence,

$$\|A - N\|_p^p \geqslant \|A - F(A)\|_p^p.$$

Assume that each point of $\sigma(A)$ has a unique closest point in Λ and let N be some operator from $\mathfrak{N}(\Lambda)$ for which equality holds in (*). Thus, equality holds throughout the inequalities of the first paragraph of this proof. In particular, using Lemma 2, for $j = 1, \ldots, l$, one has

$$\|(\alpha_j - N)e_j\| = \operatorname{dist}(\alpha_j, \Lambda) = \operatorname{dist}(\alpha_j, \sigma(N)).$$

Lemma 2 shows that e_j is an eigenvector for N with corresponding eigenvalue $F(\alpha_j)$. Choosing $\{f_1, f_2, \ldots\}$ as in the second paragraph of this proof, one notes that Lemma 4 implies

$$\|A - N\|_p^p \geqslant \sum_{j=1}^{l} |(A - N)e_j|^p + \sum_j \|(A - N)f_j\|^p.$$

Since equality holds throughout the inequalities of the first paragraph of this proof, it must be that

$$\|(A - N)f_j\| = 0, \qquad j = 1, 2, \ldots.$$

Thus, the restriction of A and N to $E(\Lambda)H$ coincide. Consequently the restrictions of A, N and $F(A)$ coincide. Since N and $F(A)$ coincide on closed span $\{e_1, \ldots, e_l\} = E(\Lambda^c)H$, it is proved that $N = F(A)$.

In the event that Λ is convex every point in the complex plane has a unique nearest point in Λ and so the preceding proof shows that $N = F(A)$.

If there exists some $\lambda \in \sigma(A)$ such that $|\lambda - \mu| = |\lambda - F(\lambda)|$ and $F(\lambda) \neq \mu \in \Lambda$ then the definition of F can be altered by setting $F(\lambda) = \mu$. Thus, there are two Borel measurable distance minimizing retracts onto Λ which are different on $\sigma(A)$. This proves $F(A)$ is not the unique best approximation.

REFERENCES

1. J. G. Aiken, H. B. Jonassen and H. S. Aldrich, *Löwdin orthonormalization as a minimum energy perturbation*, J. Chem. Phys. **62** (1975), 2745–2746.

2. J. G. Aiken, J. A. Erdös and J. A. Goldstein, *Unitary approximation of positive operators* (preprint).

3. S. K. Berberian, *The Weyl spectrum of an operator*, Indiana Univ. Math. J. **20** (1970), 529–544.

4. M. S. Birman and M. Z. Solomyak, *Stieltjes double-integral operators*, Spectral Theory and Wave Processes (Topics in Mathematical Physics, vol. 1), Consultants Bureau, New York, 1976.

5. R. Bouldin, *Essential spectrum for a Hilbert space operator*, Trans. Amer. Math. Soc. **163** (1972), 437–445.

6. B. C. Carlson and J. M. Keller, *Orthogonalization procedures and localization of Wannier functions*, Phys. Rev. **105** (1957), 102–103.

7. A. J. Coleman, *Structure of fermion density matrices*, Rev. Modern Phys. **35** (1963), 668–689.

8. N. Dunford and J. T. Schwartz, *Linear operators*, vol. II, Interscience, New York, 1963.

9. K. Fan, *Maximum properties and inequalities for eigenvalues of completely continuous operators*, Proc. Nat. Acad. Sci. U.S.A. **37** (1951), 760–766.

10. I. C. Gohberg and M. G. Krein, *Introduction to the theory of linear nonselfadjoint operators*, Transl. Math. Monographs, vol. 18, Amer. Math. Soc., Providence, R.I., 1969.

11. K. Gustafson, *Necessary and sufficient conditions for Weyl's theorem*, Michigan Math. J. **19** (1972), 71–81.

12. P. R. Halmos, *Spectral approximants of normal operator*, Proc. Edinburgh Math. Soc. **19** (1974), 51–58.

13. F. A. Valentine, *Convex sets*, McGraw-Hill, New York, 1964.

14. H. Weyl, *Inequalities between the two kinds of eigenvalues of a linear transformation*, Proc. Nat. Acad. Sci. U.S.A. **35** (1949), 408–411.

Department of Mathematics, University of Georgia, Athens, Georgia 30602