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

RICHARD BOULDIN

ABSTRACT. Let \( A \) be a fixed normal operator and let \( \mathcal{R}(A) \) denote the normal operators with spectrum contained in \( A \). Provided there is some \( N \) in \( \mathcal{R}(A) \) 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 \( \mathcal{R}(A) \) 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 \( \Lambda \) is a closed nonempty set in the complex plane then \( \mathcal{R}(\Lambda) \) denotes the normal (bounded linear) operators on the fixed separable Hilbert space \( H \) with spectrum contained in \( \Lambda \). 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 \( \mathcal{R}(\Lambda) \) 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 \( \mathcal{R}(\Lambda) \) 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 \( \Lambda \) provided each \( F(z) \) belongs to \( \Lambda \) and

\[
|z - F(z)| < |z - \lambda| \quad \text{for all } \lambda \text{ in } \Lambda.
\]

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

\[
\|A - F(A)\| < \|A - N\| \quad \text{for every } N \in \mathcal{R}(\Lambda),
\]

Received by the editors May 9, 1979 and, in revised form, August 24, 1979 and October 9, 1979.


© 1980 American Mathematical Society 0002-9939/80/0000-0517/$02.50
where $F(z)$ is a Borel measurable distance minimizing retract onto $\Lambda$. Note $F(A)$ belongs to $\mathcal{R}(\Lambda)$.

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 $N \in \mathcal{R}(\Lambda)$ 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 $\{\alpha_1, \alpha_2, \ldots, \alpha_t\}$, repeated according to multiplicity, such that $\sum_j (\text{dist}(\alpha_j, \Lambda))^p$ 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 $\Lambda$. If there exists some $N \in \mathcal{R}(\Lambda)$ such that $A - N$ belongs to $c_p$, $p > 2$, then $A - F(A)$ belongs to $c_p$ and

$$\|A - F(A)\|_p < \|A - N\|_p.$$  \hfill (\ast)

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

In the case that $A$ is an invertible nonnegative operator and $\Lambda$ 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 $\Lambda$, $F(N)$ equals $N$ and the statement that $A - F(A)$ belongs to $c_2$ is equivalent to the statement that $F(A) - N$ 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 $c_p$, $V$ and $U$ are unitary and $f(z)$ is a function on the unit circle with its derivative satisfying a Lipschitz 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

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

for $p > 1$ and any retract $F(z)$ onto $\Lambda$. 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 $N \in \mathcal{R}(\Lambda)$ such that $(A - N) \in c_p$ then the only points in the spectrum of $A$, denoted $\sigma(A)$, not contained in $\Lambda$ 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 $\Lambda$, the conclusion of the lemma follows.

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

$$
\|(\alpha - N)e\| \geq \text{dist}(\alpha, \sigma(N)).
$$

(*)

If there is a unique point $\beta$ in $\sigma(N)$ which is closest to $\alpha$ and equality holds in (*) then $e$ is an eigenvector for $N$ and $\beta$ 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

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

$$
> \int_{\sigma(N)} \text{dist}(\alpha, \sigma(N))^2 d\langle E(z)e, e \rangle
$$

$$
= \text{dist}(\alpha, \sigma(N))^2.
$$

Thus, (*) above holds.

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

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

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 $\{e_1, \ldots , e_i\}$ be a (possibly infinite) orthonormal set. Then one has the inequality

$$
\|T\|^p_p > \sum_{j=1}^{I} \langle |T|_e, e_j \rangle^p
$$

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

$$
\|T\|^p_p > \sum_{j=1}^{I} \|T_{e_j}\|^p.
$$

License or copyright restrictions may apply to redistribution; see https://www.ams.org/journal-terms-of-use
Proof.

\[ \| T \|_p^p = \| |T| \|_p^p = \sum_j s_j(|T|)^p \]
\[ = \sum_j s_j(|T|^2)^{p/2} = \| |T|^2 \|_p^{p/2} \]
\[ > \sum_j \langle |T|^2 e_j, e_j \rangle^{p/2} \quad \text{by Lemma 3} \]
\[ = \sum_j \| T e_j \|^p. \]

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

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 \( \Lambda \). In order to show the inequality (*) one observes the following

\[ \| A - N \|_p^p > \sum_j \| (A - N) e_j \|^p \quad \text{by Lemma 4} \]
\[ > \sum_j \text{dist}(\alpha_j, \sigma(N))^p \quad \text{by Lemma 2} \]
\[ > \sum_j \text{dist}(\alpha_j, \Lambda)^p. \]

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 \( \Lambda^c \) means the complement of \( \Lambda \). Note that \( A_1 \in \mathcal{U}(\Lambda) \) and \( A_2 = \sum_{j=1}^l \langle \cdot, e_j \rangle \alpha_j e_j \) where \( \{ e_1, \ldots, e_l \} \) is a maximal orthonormal set of eigenvectors for \( A \) corresponding to \( \{ \alpha_1, \ldots, \alpha_l \} \). Note that

\[ F(A) = A_1 \oplus F(A_2) = A_1 \oplus \sum_{j=1}^l \langle \cdot, e_j \rangle F(\alpha_j) e_j \in \mathcal{U}(\Lambda). \]

Also observe that

\[ \| A - F(A) \|_p^p = \left\| 0 \oplus \sum_{j=1}^l \langle \cdot, e_j \rangle (\alpha_j - F(\alpha_j)) e_j \right\|_p^p \]
\[ = \sum_{j=1}^l |\alpha_j - F(\alpha_j)|^p = \sum_{j=1}^l \text{dist}(\alpha_j, \Lambda)^p < \infty. \]

Proof of Theorem 2. By Lemma 4 and Lemma 2, with the notation of the preceding proof, one obtains

\[ \| A - N \|_p^p > \sum_{j=1}^l \| (A - N) e_j \|^p \]
\[ > \sum_{j=1}^l \| (\alpha_j - N) e_j \|^p > \sum_{j=1}^l \text{dist}(\alpha_j, \sigma(N))^p \]
\[ = \sum_{j=1}^l \text{dist}(\alpha_j, \Lambda)^p = \sum_{j=1}^l |\alpha_j - F(\alpha_j)|^p. \]
NORMAL OPERATOR IN THE SCHATTEN $p$-NORM

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 $\Lambda$ 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,

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

Assume that each point of $\sigma(A)$ has a unique closest point in $\Lambda$ and let $N$ be some operator from $\mathcal{R}(\Lambda)$ for which equality holds in $(\ast)$. 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

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

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

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

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

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

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 $\Lambda$ is convex every point in the complex plane has a unique nearest point in $\Lambda$ 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 $\Lambda$ which are different on $\sigma(A)$. This proves $F(A)$ is not the unique best approximation.

REFERENCES


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