Acta Sci. Math., 43 (1981), 109—115

C_p-minimal positive approximants

DONALD D. ROGERS and JOSEPH D. WARD

§ 1. Introduction

In [8], P. R. HALMOS initiated the study of positive operator approximation. Among other things he established the proximinality of the convex set of positive operators on Hilbert space by producing a canonical best positive approximant. This approximant, hereafter referred to as the Halmos approximant was later shown by R. H. BOULDIN [2] to be maximal, in the sense of order, among all positive approximants to a given operator.

This paper originated in the attempt to find a canonical minimal approximant since canonical approximants shed much light on the structure of the set of best approximants [3], [4], [5]. As will be shown in 4, there need not be a positive approximant minimal in the sense of order. Nevertheless, we construct a positive approximant P_m that is minimal in a sense given by the following theorem, in which $\|\cdot\|_p$ denotes the usual C_p norm on finite matrices.

Theorem 1.1. *Each operator $A=B+iC$ on a finite dimensional complex Hilbert space $\mathfrak{H}$ has a positive approximant P_m such that $A-P_m$ is a normal operator and such that for each positive operator $Q\neq P_m$ it follows that $\|A-Q\|_p>\|A-P_m\|_p$ for all finite p sufficiently large. This operator P_m will be referred to as the C_p-minimal positive approximant of A.*

In section 2, relevant background information is given along with needed notation. Section 3 contains the proof of the main theorem, the heart of which involves an inductive construction. There are many open questions related to our result, and these questions along with some examples comprise section 4.

Received February 28, 1980.

§ 2. Preliminaries

The term operator shall mean a bounded linear operator on a complex Hilbert space, and the operator norm of an operator X is denoted by $\|X\| = \sup \{\|Xf\| : f \in \mathfrak{H}, \|f\| = 1\}$. If $\mathfrak{M}$ is a set of operators, then an operator $Y_0 \in \mathfrak{M}$ is an $\mathfrak{M}$-approximant of X if $\|X - Y_0\| = \inf \{\|X - Y\| : Y \in \mathfrak{M}\}$; approximants using other norms are defined similarly. We shall follow Halmos's convention of using "positive operator" as synonymous with "nonnegative operator" and "approximant" in place of "best approximant". For the reader's convenience we restate the following results proved by Halmos in [8].

Theorem 2.1. *If $B+iC$ is the usual Cartesian representation for the operator A, then*

$$\inf \{A - P : P \geqq 0\} = \inf \{r : r \geqq \|C\|, B + (r^2 - C^2)^{1/2} \geqq 0\}.$$

The first infimum shall henceforth be denoted $\delta(A)$.

Theorem 2.2. *If $B+iC$ is the usual Cartesian representation for the operator A and if $P_H = B + ((\delta(A))^2 - C^2)^{1/2}$, then P_H is a positive approximant of A.*

The operator P_H is the Halmos approximant referred to in the introduction.

Theorem 2.3. *Any operator A has a representation of the form $P + U\delta(A)$ where $P \geqq 0$ and U is unitary with negative real part. If A is not a positive operator, then the above representation is unique.*

In another direction, the notion of a *strict approximant* was introduced by J. R. Rice [10] in the course of his investigations into l_∞ approximation as a method of selecting one approximant among many. A full discussion of strict approximants would lead us too far astray but, roughly speaking, to find a strict approximant one minimizes as much as one can. The following example will serve to illustrate.

Example. Consider the vector $v \equiv (2i, i, 0)$ viewed as an element of $l_\infty(3)$. The distance of v to the set of positive functions is 2, and there are clearly an infinite number of positive approximants. The vector (0, 0, 0), however, is the unique strict approximant since 0 is the nearest nonnegative number to $2i$, i and 0.

It was later shown by B. Mitiagin [9] and J. Descloux [6] that the strict approximants have an additional approximation property.

Theorem 2.4. *Let $l_p(n)$ denote n-dimensional complex Cartesian space endowed with the l_p norm and M a subspace of $l_p(n)$. If $x \in l_p(n) \setminus M$, let y_p denote an approximant from M. Then $y = \lim_{p \to \infty} y_p$ exists, and y is the strict approximant of x in $l_\infty(n)$.*

The construction in the next section is modelled after the construction of the strict approximant, although the fact that the space of $n \times n$ matrices is not a commutative algebra introduces some new twists into the construction.

§ 3. The Main Result

In this section the proof of Theorem 1.1 is given. The first lemma is stated in more generality than is needed, but it seems of interest in its own right.

Lemma 3.1. *Let $\mathfrak{C}$ be a norm-closed convex set of compact operators on a uniformly convex Banach space $\mathfrak{B}\neq 0$. Define $d=\inf\{\|X\|: X\in\mathfrak{C}\}$ and*

$$\mathfrak{D} = \{X\in\mathfrak{C}: \|X\| = d\}.$$

If $\mathfrak{D}$ is separable, then there exist unit vectors $y, z\in\mathfrak{B}$ such that for every $X\in\mathfrak{D}$ it follows that $Xy=dz$. In particular, if D is in $\mathfrak{D}$, then

$$\bigcap_{x \text{ in } \mathfrak{D}} \ker(X-D) \neq \{0\}.$$

Proof. Let $\{X_1, X_2, \dots\}$ be dense in $\mathfrak{D}$; define operators $Y_n\in\mathfrak{D}$ to be the corresponding Cesaro means, i.e. $Y_n=(X_1+\dots+X_n)/n$. Because each Y_n is a compact operator, there exists a unit vector $y_n\in\mathfrak{B}$ such that $\|Y_n y_n\|=d$; define the unit vector $z_n=Y_n y_n/d$. Since $\mathfrak{B}$ is reflexive, the sequences $\{y_n\}$ and $\{z_n\}$ have weak cluster points in the unit ball of $\mathfrak{B}$. Thus it is possible to find vectors y, z and subsequences $\{y_{n,j}\}$ and $\{z_{n,j}\}$ that converge weakly to y and to z. Fix $k\geqq 1$. Because X_k is compact it follows that $X_k(y_{n,j})$ converges to $X_k(y)$ in norm, as $j\to\infty$. But $X_k(y_{n,j})=dz_{n,j}$ for all j sufficiently large, by the definition of $Y_{n,j}$ and the fact that $\mathfrak{B}$ is uniformly convex. Thus $dz_{n,j}$ converges to $X_k(y)$ in norm. Hence $\|X_k(y)\|=d$, which implies $\|y\|=1$ since $\|X_k\|=d$. Also, $X_k y=dz$ since $dz_{n,j}$ converges weakly to dz; thus $\|z\|=1$. Since $\{X_k\}$ is dense in $\mathfrak{D}$, it follows that $Xy=dz$ for each $X\in\mathfrak{D}$.

The next lemma is crucial in what follows. It is a slight generalization of a lemma appearing in [2].

Lemma 3.2. *If $X=X^*$, $Y=Y^*$, $P=P^*$, and $d=\|X+iY-P\|$, then $P\leqq X+\sqrt{d^2I-Y^2}$.*

Proof. As in [1], [8] it follows that $(P-X)^2+Y^2\leqq d^2I$. Because the square root function is order-preserving, it follows that $P-X\leqq\sqrt{(P-X)^2}\leqq\sqrt{d^2I-Y^2}$.

Proof of Theorem 1.1. We proceed with constructing the operator P_m by defining numbers $\{\delta_k\}$ and subspaces $\{M_k\}$ that reduce C. If $C(k)$ denotes the part of C on M_k and $I(k)$ denotes the orthogonal projection from H onto M_k, then $P_m=B+\Sigma\sqrt{\delta_k^2 I(k)-C^2(k)}$. The construction of the sequences $\{\delta_k\}$ and $\{M_k\}$ is by induction.

Define $\delta_1=\delta(A)$ (recall the definition immediately following Theorem 2.1) and $M_1=\bigcap_Q \ker(B+\sqrt{\delta_1^2-C^2}-Q)$ where this intersection is taken over all positive approximants Q. Lemma 3.1 can be applied to the convex sets $\mathfrak{C}_1=\{A-P\colon P\geqq 0\}$ and $\mathfrak{D}_1=\{A-Q\colon Q$ is a positive approximant of $A\}$ using $d=\delta_1$ and $D=A-(B+\sqrt{\delta_1^2-C^2})$ to show $M_1\neq\{0\}$.

The fact that M_1 reduces C is shown in [1, proof of Lemma 4.1]; a different proof is given here. Let f be a unit vector in M_1 and let Q be a positive approximant of A. Then $(B-Q)f=-\sqrt{\delta_1^2-C^2}f$ by the definition of M_1; thus both $A-Q$ and $(A-Q)^*$ attain their norm at f. Hence $|A-Q|^2f=|(A-Q)^*|^2f$, and this implies that $(B-Q)Cf=C(B-Q)f$. Thus $(B-Q)Cf=C(B-Q)f=C(-\sqrt{\delta_1^2-C^2})f=-\sqrt{\delta_1^2-C^2}(Cf)$. Hence $(B+\sqrt{\delta_1^2-C^2}-Q)(Cf)=0$, so that $Cf\in M_1$.

Thus M_1 reduces C, and it also reduces $A-Q$ for each approximant Q. Clearly A has a unique approximant if and only if $M_1=H$. Define the subspace $H_1=H$ and the projection $E_1=I$.

Let H_1, E_1, $\mathfrak{C}_1$, δ_1, $\mathfrak{D}_1$, M_1 be as defined above. Define $H_2=H\ominus M_1$ with orthogonal projection $E_2\colon H\to H_2$. Put $\mathfrak{C}_2=\{(A-Q)E_2\colon Q\geqq 0$ and $(A-Q)E_1\in\mathfrak{D}_1\}$; this set $\mathfrak{C}_2$ is convex because $\mathfrak{D}_1$ is convex. Define $\delta_2=\min\{\|X\|\colon X\in\mathfrak{C}_2\}$ and $\mathfrak{D}_2=\{X\in\mathfrak{C}_2\colon \|X\|=\delta_2\}$; this set $\mathfrak{D}_2$ is convex because $\mathfrak{C}_2$ is convex.

The construction of M_2 is as follows. For an arbitrary operator X on H let $X_2=E_2XE_2$; clearly $M_1\subseteqq\ker X_2$ and M_1 reduces X_2. Choose $Q\geqq 0$ such that $(A-Q)E_2\in\mathfrak{D}_2$. Then $0\leqq Q_2\leqq B_2+\sqrt{\delta_2^2E_2-C_2^2}$ because M_1 reduces $A-Q$; this inequality follows from Lemma 3.2 with $X=B_2$, $Y=C_2$, $P=Q_2$ and $d=\delta_2$. Notice that for each such Q it follows that $Q|M_1=(B+\sqrt{\delta_1^2I(1)-C(1)^2})|M_1$ by the definition of M_1. Hence the operator $Z=B+\sqrt{\delta_1^2I(1)-C(1)^2}+\sqrt{\delta_2^2E_2-C_2^2}$ satisfies $Z\geqq Q\geqq 0$ for each such Q. Thus the operator $D_2=iC_2-\sqrt{\delta_2^2E_2-C_2^2}$ is $\in\mathfrak{D}_2$ because the operator $Z=B+\sqrt{\delta_1^2I(1)-C^2(1)}+\sqrt{\delta_2^2E_2-C_2^2}$ is a positive operator such that $(A-Z)E_2\in\mathfrak{C}_2$. Define $M_2=\bigcap_X\ker(X-D_2)\cap H_2$ where the intersection is over all $X\in\mathfrak{D}_2$. From Lemma 3.1 with $\mathfrak{C}=\mathfrak{C}_2$, $d=\delta_2$, $\mathfrak{D}=\mathfrak{D}_2$ and $D=D_2$, considered as operators from H_2 to itself, it follows that $M_2\neq\{0\}$ if $H_2\neq\{0\}$. If $H_2=\{0\}$, then $M_2=\{0\}$.

The fact that M_2 reduces the operator $C_2=C|H_2$ is shown by a proof similar to that used for M_1. Let f be a unit vector in M_2 and let $Q\geqq 0$ be such that $(A-Q)E_2\in\mathfrak{D}_2$. Then H_2 reduces $(A-Q)E_2$ and $(B-Q)f=-\sqrt{\delta_2^2E_2-C_2^2}f$ by the definition of M_2; thus both $(A-Q)E_2$ and $(A-Q)^*E_2$ attain their norm at f. Hence $|A_2-Q_2|^2f=|(A_2-Q_2)^*|^2f=\delta_2^2f$; this implies $(B_2+\sqrt{\delta_2^2E_2-C_2^2}-Q)(C_2f)=0$ as before. In other words, $C_2f\in M_2$, and thus M_2 reduces C_2.

In general, once H_k, E_k, $\mathfrak{C}_k$, δ_k, $\mathfrak{D}_k$, M_k have been defined, put $H_{k+1}=H\ominus(M_1\oplus\ldots\oplus M_k)$ with orthogonal projection $E_{k+1}\colon H\to H_{k+1}$. Let $\mathfrak{C}_{k+1}=\{(A-Q)E_{k+1}\colon Q\geqq 0$ and $(A-Q)E_k\in\mathfrak{D}_k\}$; this set $\mathfrak{C}_{k+1}$ is convex because $\mathfrak{D}_k$ is convex. Define $\delta_{k+1}=\min\{\|X\|\colon X\in\mathfrak{C}_{k+1}\}$ and $\mathfrak{D}_{k+1}=\{X\in\mathfrak{C}_{k+1}\colon \|X\|=\delta_{k+1}\}$; this set $\mathfrak{D}_{k+1}$ is convex because $\mathfrak{C}_{k+1}$ is convex.

To define M_{k+1}, write $X_{k+1}=E_{k+1}XE_{k+1}$ for each X; clearly $M_j\subset\ker X_{k+1}$ for $1\leqq j\leqq k$. The operator $D_{k+1}=iC_{k+1}-\sqrt{\delta_{k+1}^2E_{k+1}-C_{k+1}^2}$ is in $\mathfrak{D}_{k+1}$ because the operator $Z=B+\sqrt{\delta_1^2I(1)-C(1)^2}+\ldots+\sqrt{\delta_k^2I(k)-C(k)^2}+\sqrt{\delta_{k+1}^2E_{k+1}-C_{k+1}^2}$ is a positive operator such that $(A-Z)E_{k+1}$ is in $\mathfrak{C}_{k+1}$. Define the subspace M_{k+1} by $M_{k+1}=\bigcap_X \ker(X-D_{k+1})\cap H_{k+1}$; this intersection is taken over all $X\in\mathfrak{D}_{k+1}$. Lemma 3.1 shows that $M_{k+1}\neq\{0\}$ if $H_{k+1}\neq\{0\}$, and the operator D_{k+1} can be used to show M_{k+1} reduces C_{k+1}. This completes the inductive definition.

Thus for each integer k it is possible to define M_k and δ_k. Because H is finite-dimensional, the subspaces H_{k+1} will be $\{0\}$ for all k sufficiently large. Thus it is possible to define the positive operator P_m by

$$P_m = B+\Sigma\sqrt{\delta_k^2I(k)-C(k)^2}.$$

Clearly $A-P_m$ is a normal operator.

It remains to establish the minimality of P_m. If Q is a positive operator different from P_m, then there exists a least integer $k\geqq 1$ such that $(A-Q)E_k\notin\mathfrak{D}_k$. If $k=1$, then $\|A-Q\|>\delta_1$. Hence if h denotes the dimension of H, then for all p sufficiently large it follows that $\|A-Q\|_p^p\geqq\|A-Q\|^p>h\delta_1^p\geqq\|A-P_m\|_p^p$. If $k>1$, then let $\Delta_k=\|(A-Q)E_k\|$. Then $\Delta_k>\delta_k$ because $(A-Q)E_j$ is in $\mathfrak{D}_j$ for each $j\leqq k-1$. For each $j\leqq k-1$ the subspace M_j reduces $A-Q$, and the part of $A-Q$ on M_j is equal to the part of $A-P_m$ on M_j, which is $iC(j)-\sqrt{\delta_j^2I(j)-C(j)^2}$ and is δ_j times a unitary operator. Thus for all p sufficiently large and m_j=dimension of M_j, it follows that

$$\|A-Q\|_p^p\geqq m_1\delta_1^p+\ldots+m_{k-1}\delta_{k-1}^p+\Delta_k^p> m_1\delta_1^p+\ldots+m_{k-1}\delta_{k-1}^p+$$
$$+(h-m_1-\ldots-m_{k-1})\delta_k^p\geqq\|A-P_m\|_p^p.$$

This proves Theorem 1.1. ∎

§ 4. Examples and Open Questions

Example 4.1. There does not always exist a positive approximant that is minimal in the sense of order.

Let A be the self-adjoint 3×3 matrix given by $A=\begin{pmatrix}-1&0&0\\0&0&0\\0&0&2\end{pmatrix}$. It is easily seen that $\delta(A)=1$, and that no positive approximant is smaller than $P_0=\begin{pmatrix}0&0&0\\0&0&0\\0&0&1\end{pmatrix}$. For if there were such an approximant P_1, then $P_0-P_1\geqq 0$ and P_1

necessarily would have the form $\begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & \alpha \end{pmatrix}$, $\alpha<1$. But then P_1 would no longer be an approximant of A, so P_0 is the only candidate to be minimal. On the other hand, it is easily checked that $P_2=\begin{pmatrix} 0 & 0 & 0 \\ 0 & 1/8 & 1/2 \\ 0 & 1/2 & 5/2 \end{pmatrix}$ is an approximant of A and clearly $P_2-P_0\geqq 0$.

For a given matrix $A=B+iC$ let $\|A\|_p$ denote the C_p norm of A. It is well known (and follows easily from [7, p. 94]) that B is a self-adjoint approximant of A in the C_p norm for all p, and it is unique in case $1<p<\infty$. Thus if S_p denotes the self-adjoint C_p approximant to A, then $S_p=B$ so $\lim_{p\to\infty}\|S_p-B\|=0$. Let R_p denote a positive approximant to A in the C_p norm which again is unique if $1<p<\infty$.

Q1. For a given matrix A and corresponding C_p minimal positive approximant P_m, does $\lim_{p\to\infty}\|R_p-P_m\|=0$?

A weaker question is:

Q2. For a given A, does the corresponding net $\{R_p\}$ have a limit in the uniform norm as $p\to\infty$?

Note that the C_p-minimal positive approximant P_m seems to be the operator analogue of the strict approximant mentioned in section 2. Since the strict approximant of Rice is a limit of l_p approximants by Theorem 2.4, the answer to Q1 could likewise be yes. Moreover Q1 and Q2 both have affirmative answers in the case A is a 2×2 matrix. This follows from the fact that for a given 2×2 matrix A and any positive approximant P, $A-P$ is normal; each convergent subnet of $\{R_p\}$ must converge to a uniform positive approximant, which can in this case be shown to be P_m by using the minimality condition defining P_m. To establish that $A-P$ is normal, note that one of two cases occurs:

i) P_H is the unique approximant so that $A-P_H$ is a multiple of a unitary by Theorem 2.3.

ii) The subspace M_1 mentioned in the proof of Theorem 1.1 is 1-dimensional. In this case for any approximant P the errors $A-P$ and $A-P_H$ can differ only in the (2, 2) entry (when viewed as matrices with respect to the subspaces M_1 and $M_1^{\perp}$). Thus $A-P$ is normal.

Questions analogous to Q1 and Q2 may be asked for $p\to 1$:

Q3. Does $\lim_{p\to 1} R_p$ exist?

If the answer to Q3 is yes, then

Q4. Can the limit in Q3 be identified by any characteristics?

An affirmative answer to Q4 would yield a canonical approximant for positive approximation in the trace norm.

Finally it seems as if Theorem 1.1 must have some extension at least to the compact operator case. Relevant to this problem is the following

Example 4.2. There exists a compact operator with no compact positive operator approximant.

Indeed, let $\{e_1, e_2, \dots\}$ denote an orthonormal basis and let f be the vector $f \equiv \Sigma e_k/k$. Define Q to be the rank one orthogonal projection onto sp $\{f\}$, C the compact operator given by $C(e_k)=e_k/k$, $B \equiv (1-Q)-\sqrt{1-C^2}$, and finally set $A=B+iC$. Then A is a compact operator and has a unique positive approximant P_H [1, p. 282]. Now P_H is not compact since $A-P_H$ is a multiple of a unitary.

Q5. Which compact operators admit compact positive approximants; is there a "minimal" approximant in this case?

References

[1] T. Ando, T. Sekiguchi and T. Suzuki, Approximation by positive operators, *Math. Z.*, **131** (1973), 273—282.

[2] R. H. Bouldin, Positive approximants, *Trans. Amer. Math. Soc.*, **177** (1973), 391—403.

[3] R. H. Bouldin and D. D. Rogers, Normal dilations and operator approximations, *Acta Sci. Math.*, **39** (1977), 233—243.

[4] C. K. Chui, P. W. Smith and J. D. Ward, Approximation with restricted spectra, *Math. Z.*, **144** (1975), 289—297.

[5] C. K. Chui, P. W. Smith and J. D. Ward, Favard's solution is the limit of $W^{k,p}$-splines, *Trans. Amer. Math. Soc.*, **220** (1976), 299—305.

[6] J. Descloux, Approximations in L^p and Chebyshev approximations, *J. Soc. Indust. Appl. Math.*, **11** (1963), 1017—1026.

[7] I. C. Gohberg and M. G. Krein, *Introduction to the theory of linear nonselfadjoint operators*, AMS Translations of Math. Monographs, **18** (1969).

[8] P. R. Halmos, Positive approximants of operators, *Indiana Univ. Math. J.*, **21** (1971), 951—960.

[9] B. Mitjagin, The extremal points of a certain family of convex functions, *Sibirsk. Mat. Ž.*, **6** (1965), 556—563 (Russian).

[10] J. R. Rice, Tchebycheff approximation in a compact metric space, *Bull. Amer. Math. Soc.*, **68** (1962), 405—410.

DEPARTMENT OF MATHEMATICS
TEXAS A&M UNIVERSITY
COLLEGE STATION, TEXAS 77843