DILATIONS PRESERVING A BOUND ON THE NORM OF AN OPERATOR

Chandler Davis

ABSTRACT

Given Hilbert space operators A, B, C, one seeks an operator D such that $\left\|\begin{pmatrix} A & C \\ B & D \end{pmatrix}\right\| \leqq \mu$ for a pre-assigned bound μ. In some contexts only the existence of D is called for; in others, one may want to know something about those D which work. This paper gives an exposition of the main theorems and of two of the many applications.

INTRODUCTION

The problem is to choose the missing entry in the operator-matrix $\begin{pmatrix} A & C \\ B & ? \end{pmatrix}$ so as to satisfy a bound on the norm. Now for $\left\|\begin{pmatrix} A & C \\ B & D \end{pmatrix}\right\| \leqq \mu$ it is obviously necessary that $\left\|\begin{pmatrix} A \\ B \end{pmatrix}\right\| \leqq \mu$ and $\|(A \quad C)\| \leqq \mu$. Is this pair of necessary conditions also sufficient? One is led to doubt it if one hastily presumes that the choice $D = 0$ is norm-minimizing. Although $\left\|\begin{pmatrix} 1 \\ 1 \end{pmatrix}\right\| = \|(1 \quad 1)\| = \sqrt{2}$, still we cannot complete the matrix $\begin{pmatrix} 1 & 1 \\ 1 & ? \end{pmatrix}$ to $\begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix}$ while preserving the norm, indeed $\left\|\begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix}\right\| = \frac{1+\sqrt{5}}{2}$. Optimism is revived by the observation that a different choice of D does preserve the norm: $\left\|\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}\right\| = \sqrt{2}$. This simple relationship prevails in general.

<u>EXISTENCE THEOREM</u> *Given* A, B, C *such that* $\left\|\begin{pmatrix} A \\ B \end{pmatrix}\right\| \leqq \mu$ *and* $\|(A \quad C)\| \leqq \mu$, *there exists* D *such that* $\left\|\begin{pmatrix} A & C \\ B & D \end{pmatrix}\right\| \leqq \mu$.

The example considered so far is so simple it gives no basis for guessing to what extent the solution of the problem is unique. Sticking to very simple cases for the moment, we can shed a little light on this issue.

If A, B, C (hence also the desired D) are 1-dimensional, and if $\left\|\binom{A}{B}\right\| = \mu$ while $\|(A\ C)\| \leqq \mu$, then we write

$$\begin{pmatrix} A & C \\ B & ? \end{pmatrix} = \mu \begin{pmatrix} \cos\theta & t\sin\theta \\ \sin\theta & ? \end{pmatrix}$$

for some t with $|t| \leqq 1$ (the loss of generality in assuming A and B real is of no consequence). A little reflection reveals that the solution if still unique. On the other hand, the 1-dimensional case will not give uniqueness if both $\left\|\binom{A}{B}\right\|$ and $\|(A\ C)\|$ are strictly less than the imposed bound μ. Thus consider using the bound 2 in the original example $\begin{pmatrix} 1 & 1 \\ 1 & ? \end{pmatrix}$. The hypotheses $\left\|\binom{A}{B}\right\| \leqq 2$ and $\|(A\ C)\| \leqq 2$ are satisfied and some to spare. In order for the completed matrix $\begin{pmatrix} 1 & 1 \\ 1 & t \end{pmatrix}$ to have norm $\leqq 2$ for real t, it is necessary and sufficient (one can verify directly) that $-\frac{5}{3} \leqq t \leqq 1$. This should prepare the reader for the situation which is encountered in general: if the given part of the matrix contains (perhaps after changing the coordinate system) a column of norm exactly equal to the bound imposed, some uniqueness will result; similarly if the given part of the matrix contains a row of norm exactly equal to the bound; otherwise, there will be some leeway in completing the matrix. I will describe two sorts of results concerning this. The first solves the following.

<u>CHARACTERIZATION PROBLEM</u> *Given* A, B, C *such that* $\|(A\ C)\| \leqq \mu$ *and* $\left\|\binom{A}{B}\right\| \leqq \mu$, *find all* D *such that* $\left\|\begin{pmatrix} A & C \\ B & D \end{pmatrix}\right\| \leqq \mu$.

The second selects one particular D which is optimal in a certain sense. The discussion will assume all operators given are compact; the reader who thinks of them as finite-dimensional will lose nothing essential.

I will then proceed, however, to describe in Sections 2 and 3 two of the applications of the theory, and these do involve infinite-dimensional spaces.

The Existence Theorem was first discovered by W.M. Kahan and H.F. Weinbergerin 1973; the Characterization Problem was solved soon afterward in collaboration with me. The optimal solution was presented in [3], but the other parts of the result were not published for several years, and in the meantime independent discoveries of the ideas were made by S. Parrott [7], Gr. Arsene & A. Ghondea [1], and Yu.L. Shmul'yan and R.N. Yanovskaya [8]. Several different approaches to the proof may be found in these papers and in my survey [4]. The treatment in the following Section 1 is based on [5], [3].

1. THE MAIN IDEAS

The hypothesis $\left\|\binom{A}{B}\right\| \leqq \mu$ may equivalently be written $A^*A+B^*B \leqq \mu^2$; from this it is not hard to see that it is also equivalent to the condition that N be expressible in the form $B = KW$, where Q denotes (now and henceforth) the operator $(\mu^2 - A^*A)^{1/2}$ and where K may be any contraction. Similarly, the hypothesis $\|(A\ \ C)\| \leqq \mu$ is equivalent to C being expressible in the form $C = Q_*L$, where Q_* means $(\mu^2 - AA^*)^{1/2}$ and L is any contraction. Note the dependence of Q and Q_* upon μ.

In these terms it is easy to prove the Existence Theorem. Indeed, I will show that $D = -KA^*L$ is one solution to the problem.

This is done by the factorization

$$\begin{pmatrix} A & C \\ B & -KA^*L \end{pmatrix} = \begin{pmatrix} A & Q_*L \\ KQ & -KA^*L \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & K \end{pmatrix} \begin{pmatrix} A & Q_* \\ Q & -A^* \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & L \end{pmatrix} .$$

On the right, the first and last factors are contractions. The middle factor is μ times a unitary. To verify this last assertion, left-multiply $\begin{pmatrix} A & Q_* \\ Q & -A^* \end{pmatrix}$ by its adjoint; the result must be shown to equal $\mu^2 \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$. The identity which will do it is $Q_*A = AQ$, for which see [9] or [6].

This discussion has provided all the ideas which go into the Characterization Problem. Here is the result.

Given A, $B = KQ$, *and* $C = Q_*L$ *with the definitions above, we will have* $\left\| \begin{pmatrix} A & C \\ B & D \end{pmatrix} \right\| \leq \mu$ *if and only if* D *is of the form*

$$D = -KA^*L + \mu(1-KK^*)^{1/2} Z (1-L^*L)^{1/2}$$

for some contraction Z.

This may be proved by considering the factorization, valid when D is given by the stated expression,

$$\begin{pmatrix} A & C \\ B & D \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & K & (1-KK^*)^{\frac{1}{2}} \end{pmatrix} \begin{pmatrix} A & Q_* & 0 \\ Q & -A^* & 0 \\ 0 & 0 & \mu Z \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & L \\ 0 & (1-L^*L)^{\frac{1}{2}} \end{pmatrix} .$$

Let me give another numerical example, complex enough to show what sort of behaviour is found in general. Let

$$A = \begin{pmatrix} 3 & 0 \\ 0 & 1 \end{pmatrix}, \quad B = \begin{pmatrix} 2 & 0 \\ 0 & 3 \end{pmatrix}, \quad C = \begin{pmatrix} 1 \\ 1 \end{pmatrix} .$$

First, using $\mu = \sqrt{13}$, the tightest bound we compute

$$Q = Q_* = \begin{pmatrix} 2 & 0 \\ 0 & \sqrt{12} \end{pmatrix}, \quad K = \begin{pmatrix} 1 & 0 \\ 0 & \sqrt{3}/2 \end{pmatrix}, \quad L = \frac{1}{6}\begin{pmatrix} 3 \\ \sqrt{3} \end{pmatrix}, \quad \text{whence}$$

$$D = \begin{pmatrix} -\frac{3}{2} \\ -\frac{1}{4} + \frac{\sqrt{78}}{6} t \end{pmatrix} \quad \text{for } |t| \leqq 1.$$

The reason the first component of D is fixed can be seen by looking at

$$\begin{pmatrix} A & C \\ B & D \end{pmatrix} = \begin{pmatrix} 3 & 0 & 1 \\ 0 & 1 & 1 \\ 2 & 0 & d_1 \\ 0 & 3 & d_2 \end{pmatrix} .$$

The first column has norm exactly $\sqrt{13} = \mu$. If the norm of the whole is to be no more than μ, then the third column will have to be orthogonal to this first, extremal, one. Indeed, in this case $\begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$ is certainly a right singular vector belonging to the singular value $\sqrt{13}$; plainly the associated left singular vector is (in its row form) $(3 \;\; 0 \;\; 2 \;\; 0)$ normalized; thus we need

$$(3 \quad 0 \quad 2 \quad 0)\begin{pmatrix} 3 & 0 & 1 \\ 0 & 1 & 1 \\ 2 & 0 & d_1 \\ 0 & 3 & d_2 \end{pmatrix} = (13 \quad 0 \quad 0) ,$$

forcing $d_1 = -\frac{3}{2}$.

This does not happen if our bound is a little looser, say $\mu = 4$. Then computing $-KA^*L$ by the same procedure we obtain $(-\frac{6}{7} \quad -\frac{1}{5})^T$; there is leeway in both components of D. (There has better be, else the present solution would not include those D found before, which satisfy a strictly stronger condition!)

One piece of wisdom to be drawn from this comparison is that there is no very far-reaching significance to the "central" solution $-KA^*L$; for we saw it affected a good deal by a little change in our bound. (As μ gets large, $-KA^*L$ goes to 0 as μ^{-2}.)

Returning again to the tight bound $\mu = \sqrt{13}$, once we have perceived that a definite value for d_1 is required, we may assign it that value and then ask, not for all admissible values of d_2 , but for one which will be most economical. What should we mean by this? A natural interpretation is the following. The first singular value of $\begin{pmatrix} A & C \\ B & D \end{pmatrix}$ has now been fixed, and right and left singular vectors belonging to it. The remaining right singular vectors are constrained to be orthogonal to the one that was fixed, and likewise for left singular vectors. The effect is that the problem of choosing d_2 is a problem $\begin{pmatrix} A' & C' \\ B' & ? \end{pmatrix}$ where A' is 2×1 . So let us solve this problem, not with the bound $\mu = \max\{\|\binom{A}{B}\|, \|(A \;\; C)\|\}$ inherited from the original problem, but with the bound $\mu' = \max\{\|\binom{A'}{B'}\|, \|(A' \;\; C')\|\}$ appropriate to the new, smaller one. The result is $\mu' = \sqrt{10}$, leading to the unique entry $d_2 = -\frac{1}{3}$.

The result of [3] is that this process always succeeds, giving $\begin{pmatrix} A & C \\ B & D \end{pmatrix}$ with singular values which are lexicographically optimal; they may be estimated in terms of those of $\binom{A}{B}$ and $(A \;\; C)$.

Let us extract one final piece of wisdom. It is a little exaggerated to describe the problems treated here as the finding of dilations which minimize the norm. Whatever we compute depends upon μ; if (as is often the case in applications) we have only an imprecise bound, this will lead to different D from what we would find given the best bound. As for the sequential optimizing procedure, it doesn't work at all except with exact bounds! It seems a reasonable project for future research to find a good analogue to it which does not suffer from this limitation.

2. APPLICATION TO HANKEL MATRICES

Z. Nehari proved that every Hankel operator can be obtained as a compression (in the operator theorist's terminology) of a Laurent operator having the same norm. His proof, and most proofs, rely on the realization of the space where the operator is defined as a space of analytic functions. S. Parrott [7] noted that the theorem can be stated entirely in terms of infinite matrices considered as operators on ℓ_2 spaces, and he was led to the following ingenious "clean" proof.

First the matrix formulation. Let H denote the Hilbert space of sequences $(\xi_0, \xi_1, \ldots)$ with norm $(\Sigma_{n=0}^{\infty}|\xi_n|^2)^{\frac{1}{2}} < \infty$. As usual, an operator A on H is called a Hankel operator if it is given by a matrix whose entries are constant on all transverse diagonals, so that we can write the i-th entry of $A(\xi_0,\xi_1,\dots)$ as $\Sigma_{j=0}^{\infty} a_{i+j}\xi_j$ for coefficients $a_0,a_1,\dots$. Assuming we know a bound $\|A\| \leqq \mu$ upon the operator so defined, our object is to imbed H in the space $\hat{H}$ of ℓ_2 bilateral sequences in the usual way, and to find an operator $\hat{A}$ on $\hat{H}$ such that

(i) it is a dilation of A, i.e., $A = P_H\hat{A}|_H$;

(ii) it is still given by a matrix whose entries are constant on all transverse diagonals, i.e., for new coefficients $a_{-1},a_{-2},\dots$ along with the old, we can write the i-th entry of $\hat{A}(\dots,\xi_{-1},\xi_0,\xi_1,\dots)$ as $\Sigma_{j=-\infty}^{\infty} a_{i+j}\xi_j$;

(iii) $\|\hat{A}\| \leqq \mu$.

Now in filling in the matrix of $\hat{A}$ there are a good many places where we know what we have to put, by the Hankel condition (ii):

$$\left[\begin{array}{c|cccccc}
 & ? & ? & a_0 & \cdots & & \\ \hline
 & ? & a_0 & a_1 & a_2 & \cdots & \\ \cline{2-7}
? & a_0 & a_1 & a_2 & a_3 & \cdots & \\
a_0 & a_1 & a_2 & a_3 & \cdots & & \\
a_1 & a_2 & a_3 & \cdots & & & \\
 & \vdots & & & & &
\end{array}\right] .$$

Below and to the right of the solid line is the matrix we begin with, all filled in, representing an operator A on H. Now look at the part below and to the right of the dotted line. It represents an operator A on H to $H_1 = \text{span}(H, e_{-1})$, in which one entry remains to be filled in. This is the situation described in Section 1. The given A is playing the role of $(A \ \ C)$ there. The role of $\binom{A}{B}$ of Section 1 is played by the matrix below and to the right of the dotted line except its first column; this is an operator on $H \ominus \{e_0\}$ to H_1, and since it has just the same matrix as A it has the same norm. Now the Existence Theorem assures us that there is a number a_{-1} which can be placed in the spot under the dotted lines so that A_1 will satisfy $\|A_1\| \leqq \mu$. Then naturally we will place the same number a_{-1} above each a_0 in the array.

But this process can be repeated indefinitely! The result is a definition of an operator $\hat{A}$ on all sequences in $\hat{H}$ having only finitely many non-zero components of negative index. But the set of such sequencs is dense; so the proof is finished in a standard way.

3. APPLICATION TO OPTIMAL ERROR BOUNDS

In this Section I will show how the norm-preserving dilation problem expresses a certain sort of optimality problem in numerical analysis. Suppose we are trying to approximate a linear operator T which acts on a Hilbert space B to a Hilbert space S. For instance, the elements of B may be possible data in a boundary-value problem, and the solutions may be elements of S. Suppose we intend, for any w that may be given in B, to represent it by selecting a finite set of numerical information; let us say that this process consists of mapping w to an element Nw of a Euclidean space E_n. Next a linear computation $\hat{T}$ upon this will produce a new tuple of numbers $\hat{T}Nw$ belonging to another Euclidean space E_m. Finally, these must be interpreted as an approximate solution; this may be by some linear injection M of E_m into S. We hope that the result $M\hat{T}Nw$ will not be far from the exact solution Tw.

I want to focus on the optimal selection of the algorithm $\hat{T}$. That is, I assume that the norms on B and S are appropriate to the needs of the problem, and that the procedures of discretization N and interpolation M have already been decided upon as well. In this case we can equivalently identify E_m with ME_m, a subspace of S, and use the norm it thereby gets from S; and similarly we can identify E_n with the subspace $N^*E_n = \text{null}(N)^\perp$ of B, using on it the norm of B. The result is to regard N and M^* as orthoprojectors in spaces B and S respectively. (Notice that if N uses values of functions in B and (dually) if M is really an interpolation process, then the norms in B and S can not be L_2 norms, but must be such that point evaluation will be continuous.)

Now $\|Tw-M\hat{T}Nw\|$ can not be bounded exclusively from the data so far discussed, for there is an infinite-dimensional subspace of B whose elements give zero data, namely null(N). The most that can be hoped is to bound $\|Tw-M\hat{T}Nw\|$ given an a priori bound upon $\|w\|$ —that is, to bound the operator norm $\|T-M\hat{T}N\|$. Let me now interpret this as a problem in norm-preserving dilation. That is, I will exhibit $T-M\hat{T}N$ as an operator of the form $\begin{pmatrix} A & C \\ B & D \end{pmatrix}$ from a direct sum $H_1 \oplus H_2$ to a direct sum $K_1 \oplus K_2$, with the partial operators $A{:}H_1 \to K_1$ and $B{:}\ H_1 \to K_2$ and $C{:}\ H_2 \to K_1$ all being prescribed while only the partial operator $D{:}\ H_2 \to K_2$ is left for us to choose.

Namely , H_1 here will be null(N) and K_2 will be range(M). Our choice is to consist only of specifying $\hat{T}$; this will have no effect on any w which is annihilated by N , and it can not yield any output not in the range of M , but these are the only restrictions. We being by bounding the norms of the partial operators $\binom{A}{B} = T|_{\text{null}(N)}$ and $(A \quad C) = P_{\text{range}(M)^{\perp}}T$, using our knowledge of the operator T under investigation. Then if μ is a bound upon both of these norms, the Existence Theorem tells us that there is some linear $\hat{T}$ on E_n to E_m such that $\|T-M\hat{T}N\| \leqq \mu$, and the Characterization result permits us to write one down.

This has two limitations in practice. It may be hard to find a good value for μ ; and the formula obtained from the abstract theorem may be troublesome to convert into a finite $m \times n$ matrix of numbers. The operator T whose approximation was treated in [6] was a very simple one: the operator of indefinite integration on an interval, regarded as acting between certain Hilbert spaces of differentiable functions. One of the optimizations done in that paper leads to a procedure in which (to oversimplify the situation slightly) the matrix for $\hat{T}$ agrees closely with that for applying the trapezoidal rule, except near the main diagonal, where improvement is gained by departures from that rule. In subsequent work, Weinberger [10] has shown how to make the improvement more attractive by retaining the simple trapezoidal rule exactly far from the main diagonal, so that only a relatively very small number of the $m \times n$ matrix entries need arduous computing.

REFERENCES

[1] Gr. Arsene and A. Ghondea, *Completing matrix contractions*, J. Operator Theory 7 (1982), 179-189.

[2] M. Crandall, *Norm preserving extensions of linear transformations in Hilbert spaces*, Proc. Amer. Math. Soc. 21 (1969), 335-340.

[3] Ch. Davis, *An extremal problem for extensions of a sesquilinear form*, Linear Algebra Appl. 13 (1976), 91-102.

[4] Ch. Davis. *Some dilation and representation theorems*, Proceedings of the Second International Symposium in West Africa on Functional Analysis and its Applications, Kumasi, 1979 [published 1982], pp. 159-182.

[5] Ch. Davis, *A factorization of an arbitrary* $m \times n$ *contractive operator-matrix*, Proceedings of the Toeplitz Centennial Conference, Birkhäuser, 1982, pp. 217-232.

[6] Ch. Davis and W.M. Kahan and H.F. Weinberger, *Norm-preserving dilations and their applications to optimal error bounds*, SIAM J. Numer. Anal. 19 (1982), 445-469.

[7] S. Parrott, *On a quotient norm and the Sz.-Nagy-Foiaş lifting theorem*, J. Functional Anal. 30 (1978), 311-328.

[8] Yu.L. Shmul'yan and R.N. Yanofskaya, *The blocks of contractive operator-matrices*, Izv. Vyssh. Uchebn. Zaved. Mat. 1982.

[9] B. Sz.-Nagy and C. Foiaş, *Harmonic analysis of operators on Hilbert space*, Akadémiai Kiadó, Budapest, 1970.

[10] H.F. Weinberger, *Some remarks on good, simple, and optimal quadrature formulas*, Proceedings of the Symposium on Recent Advances in Numerical Analysis in honor of J. Barkley Rosser, edited by Golub and de Boor, MRC publication 41, University of Wisconsin, 1978.

Department of Mathematics,
University of Toronto,
Toronto, Ontario M5S 1A1,
CANADA.