ON CLOSEST NORMAL MATRICES

BY

ROBERT L. CAUSEY

TECHNICAL REPORT CS10

JUNE 30, 1964

COMPUTER SCIENCE DIVISION
School of Humanities and Sciences
STANFORD UNIVERSITY

:S10

ON CLOSEST NORMAL MATRICES

BY

ROBERT L. CAUSEY

PREPARED UNDER CONTRACT Nonr-225(37)
(NR-044-211)
OFFICE OF NAVAL RESEARCH

COMPUTER SCIENCE DIVISION
School of Humanities and Sciences
STANFORD UNIVERSITY

ACKNOWLEDGMENTS

The author wishes to express his sincere appreciation and thanks to the following persons:

Professor George E. Forsythe, Stanford University, for his encouragement and for his extremely valuable suggestions and comments in directing this dissertation.

Professor Ingram Olkin, Stanford University, for a critical appraisal of an earlier manuscript and for suggestions which led to a number of improvements, especially in Section 4.2.

Professor John G. Herriot, Stanford University, for reading the manuscript and for numerous valuable suggestions.

Dr. Patricia J. Eberlein, University of Rochester, for some valuable private communications concerning her own work on Mirsky's conjecture.

Professor J. Wallace Givens, Northwestern University and Argonne National Laboratory, for calling the author's attention (before research on this dissertation was initiated) to the useful notion of a differentiable curve of matrices and for some general comments on the problem of this dissertation.

Professor Peter Henrici, Eidgenössische Technische Hochschule, Zurich, Switzerland, Dr. J. H. Wilkinson, National Physical Laboratory, Teddington, England, and Professors Menahem M. Schiffer and Charles Loewner of Stanford University for stimulating conversations on the investigations of this thesis.

My wife, Barbara, for constant help and understanding.

Mrs. Jerri Rudnick for a most careful and competent typing of the manuscript and Mrs. Betty Jo Prine for seeing the manuscript safely through the typing and reproduction phases.

This work was supported in part by the Lockheed Missiles and Space Company Independent Research Program.

TABLE OF CONTENTS

BLANK PAGE

CHAPTER 1

INTRODUCTION AND PRELIMINARIES

1.0 Introduction.

The central problem considered in this paper is the following: Given an n by n matrix A of complex elements, find those normal matrices (called ν-minimal matrices) of order n which are closest to A in the sense of a metric defined in terms of a norm ν. A related problem is that of determining the distance $d_\nu(A)$ between A and the subset $\mathcal{N}$ of all normal matrices of order n. The historical background and precise definition of these problems are given in Section 1.2. The distance problem mentioned above was first studied by Mirsky, who offered Conjecture 1.35 as the general solution for all unitarily invariant norms.

After a proof that ν-minimal matrices always exist (Section 1.5), it is shown in Chapter 2 that the property of ν-minimality and certain other quantities are invariant under certain transformations of matric space. An inequality of Mirsky is sharpened in Chapter 3. A number of important necessary conditions for ϵ-minimal matrices (ϵ denotes the familiar Euclidean norm) are derived in Chapters 4 and 7. In Chapter 9 it is shown that all eigenvalues of an ϵ-minimal matrix lie in the field of values $F(A)$ of A; these eigenvalues are shown to be special extreme points of $F(A)$ in the case $n = 2$. In the final Chapter 10 an iterative computational procedure for finding $d_\epsilon(A)$ or an ϵ-minimal matrix is proposed, but its convergence is not proved.

Probably the most important results of the paper are the following:

1) A characterization of any ϵ-minimal matrix in terms of the Maximum Problem 5.11 (Theorem 5.13).

2) A determination of all ϵ-minimal matrices of order 2 (Theorem 6.24).

3) Mirsky's Conjecture 1.35 is shown to be true for $\nu = \epsilon$ and $n = 2$ (Theorem 6.80), false for $\nu = \epsilon$ and $n \geq 3$ (Theorem 8.5) and false for $n \geq 2$ and $\nu = \nu_p$ $(2 < p \leq \quad)$, where ν_p is defined by (1.15) (Theorem 8.9).

[Note: The fact that Mirsky's conjecture is false for $\nu = \epsilon$ and $n \geq 3$ was first proved by P. J. Eberlein.]

All results stated herein which are not specifically labeled as known or for which no reference is given are believed to be new.

1.1 Notation and Preliminary Definitions.

Let $\underline{R}$ and $\underline{C}$ denote respectively the real and complex number fields and let $\mathcal{M}_n$ denote the algebra of all n by n matrices over $\underline{C}$, where n is a positive integer. If $X \in \mathcal{M}_n$ we denote its complex conjugate transpose by X^*. A matrix X is called <u>normal</u> if $X^*X = XX^*$; in particular X is called <u>hermitian</u> if $X^* = X$, <u>skew-hermitian</u> if $X^* = -X$, and <u>unitary</u> if $X^*X = I$, where I denotes the identity matrix of the same order as X. Let $\mathcal{N}_n$, $\mathcal{H}_n$, $\mathcal{U}_n$, and $\mathcal{D}_n$ denote respectively the subsets of all normal, hermitian, unitary and diagonal matrices in $\mathcal{M}_n$. (We shall sometimes omit the subscript n, if the value of n need not be specified.)

For the meaning of terminology or notation not explicitly defined in this paper, the reader is referred to one of the standard textbooks on the theory of matrices (e.g., Perlis [21]†).

A real valued function ν defined on $\mathcal{M}$ is called a <u>norm</u> if the conditions

$$\nu(A) > 0 \qquad \text{if } A \neq 0 \,, \tag{1.1}$$

$$\nu(cA) = |c|\,\nu(A) \,, \tag{1.2}$$

$$\nu(A + B) \leq \nu(A) + \nu(B) \tag{1.3}$$

are satisfied for all $A, B \in \mathcal{M}$ and for all $c \in \underline{C}$. A norm ν is said

---

† Numbers in square brackets refer to references listed in the bibliography at the end of this paper.

to be <u>unitarily invariant</u> if, in addition to (1.1) - (1.3),

$$\nu(UA) = \nu(AU) = \nu(A) \tag{1.4}$$

holds for all $A \in \mathcal{M}$ and for all $U \in \mathcal{U}$. Moreover, a norm ν is said to be <u>multiplicative</u> if, for arbitrary $A, B \in \mathcal{M}$,

$$\nu(AB) \leq \nu(A)\,\nu(B) \quad . \tag{1.5}$$

Unitarily invariant norms were characterized by von Neumann [19] (see also Fan and Hoffman [9]) as follows. A real-valued function $\varphi(u) \equiv \varphi(u_1, u_2, \ldots, u_n)$ defined for all real n-vectors $u = (u_1, u_2, \ldots, u_n)$ is called a <u>symmetric gauge function</u> if it satisfies, for arbitrary real vectors u, v and real scalars α, the following conditions

$$\varphi(u) > 0 \qquad \text{if } u \neq 0 \ , \tag{1.6}$$

$$\varphi(\alpha u) = |\alpha|\,\varphi(u) \ , \tag{1.7}$$

$$\varphi(u + v) \leq \varphi(u) + \varphi(v) \ , \tag{1.8}$$

$$\varphi(u_1, u_2, \ldots, u_n) = \varphi(\sigma_1 u_{\pi_1}, \sigma_2 u_{\pi_2}, \ldots, \sigma_n u_{\pi_n}) \ . \tag{1.9}$$

where σ_i can be either of the signs ± 1 $(i = 1,2, \ldots, n)$ and where $(\pi_1, \pi_2, \ldots, \pi_n)$ is any permutation of $(1,2, \ldots, n)$.

<u>1.10 Definition</u>. Let $A \in \mathcal{M}$. The nonnegative square roots of the eigenvalues of A^*A are called the <u>singular values</u> of A.

<u>Remark</u>. Since the spectrum of AA^* coincides with that of A^*A, the singular values of A^* are the same as those of A.

1.11 Theorem (von Neumann [19]). A norm ν on $\mathcal{M}_n$ is unitarily invariant if and only if there exists a symmetric gauge function φ_ν of n real variables such that

$$\nu(A) = \varphi_\nu(\alpha_1,\alpha_2, \ldots , \alpha_n) \qquad \text{for all } A \in \mathcal{M}_n \tag{1.12}$$

where $\alpha_1,\alpha_2, \ldots , \alpha_n$ are the singular values of A.

As examples of symmetric gauge functions we may cite

$$\varphi_p(u_1, \ldots , u_n) = \Big(\sum_{i=1}^{n} |u_i|^p\Big)^{1/p} \qquad \text{for } 1 \le p < \infty \ . \tag{1.13}$$

As $p \to \infty$, the function in (1.13) converges to

$$\varphi_\infty(u_1,u_2, \ldots , u_n) = \max_{i=1,\ldots,n} (|u_i|) \ , \tag{1.14}$$

which is also a symmetric gauge function. Thus, for $1 \le p \le \infty$,

$$\nu_p(A) = \varphi_p(\alpha_1, \ldots , \alpha_n) \tag{1.15}$$

is a unitarily invariant norm, where $\alpha_1, \ldots , \alpha_n$ are the singular values of A. The norm ν_∞ is known as the spectral norm of A and is sometimes denoted by the symbol σ, i.e., $\sigma(A) = \nu_\infty(A)$ for all $A \in \mathcal{M}_n$. For $p = 2$ $\nu_p(A)$ in (1.15) coincides with the familiar Euclidean norm $\epsilon(A)$ defined by

$$\epsilon^2(A) = \sum_{i,j=1}^{n} |a_{ij}|^2 \tag{1.16}$$

where $A = (a_{ij}) \in \mathcal{M}_n$ and where by $\epsilon^2(A)$ we mean $[\epsilon(A)]^2$. It can be shown (see e.g., Faddeev and Faddeeva [7] pp. 105-111) that both σ and ϵ are multiplicative norms.

It is well known that $\mathcal{M}_n$ is a Banach algebra with respect to the norm (1.16), that is, $\mathcal{M}_n$ is a Banach space when considered as a linear space with norm ϵ, and the multiplication operation (transformation) $(A,B) \to AB$ is a continuous mapping from the product space $\mathcal{M}_n \times \mathcal{M}_n$ onto $\mathcal{M}_n$. (Cf. Hille and Phillips [13], p. 22.)

1.17 Definition. Let the eigenvalues of $M \in \mathcal{M}_n$ be denoted by $\lambda_1(M), \lambda_2(M), \ldots, \lambda_n(M)$ in some order. Then $\Omega(M)$ is defined by

$$\Omega(M) = \operatorname{diag}(\lambda_1(M), \ldots, \lambda_n(M)) \quad .$$

For any $M = (m_{ij}) \in \mathcal{M}_n$ the trace of M is defined by

$$\operatorname{tr}(M) = \sum_{i=1}^{n} m_{ii} \quad . \tag{1.18}$$

We note the following well-known properties of the trace functional:

$$\operatorname{tr}(AB) = \operatorname{tr}(BA) \quad ; \tag{1.19}$$

$$\operatorname{tr}(\alpha A + \beta B) = \alpha \operatorname{tr}(A) + \beta \operatorname{tr}(B) \quad ; \tag{1.20}$$

$$\operatorname{tr}(M^*) = \overline{\operatorname{tr}(M)} \quad ; \tag{1.21}$$

$$\epsilon^2(M) = \operatorname{tr}(M^*M) = \operatorname{tr}(MM^*) \quad . \tag{1.22}$$

Here A, B, M are any matrices in $\mathcal{M}$ and α, β are any complex numbers.

1.23 Definition. Let $M = (m_{ij}) \in \mathcal{M}_n$. The diagonal of M is defined by

$$dg(M) = \text{diag}(m_{11}, \ldots, m_{nn}) \quad . \tag{1.24}$$

Likewise the off-diagonal of M is defined by

$$\text{offdg}(M) = M - dg(M) \quad . \tag{1.25}$$

1.2 Distance Problems; Mirsky's Conjecture and Bound.

Let $A \in \mathcal{M}$, let ν be any norm on $\mathcal{M}$, and let $\mathcal{S}$ be any subset of $\mathcal{M}$. By a distance problem we mean the problem of determining the "distance"

$$\inf_{X \in \mathcal{S}} \nu(A - X) \tag{1.26}$$

between A and $\mathcal{S}$ with respect to the norm ν. In any case in which the infimum (1.26) is attained by a matrix in $\mathcal{S}$ we may consider the related minimum problem of finding (at least one matrix and preferably all) matrices $X_o \in \mathcal{S}$ such that

$$\nu(A - X_o) = \min_{X \in \mathcal{S}} \nu(A - X) \quad .$$

Several such problems have been considered and solved in the past. We now describe some of these results.

1.27 Theorem (Fan and Hoffman [9]). Let $A \in \mathcal{M}$, let $\mathcal{H}$ denote the subset of all hermitian matrices in $\mathcal{M}$, and let ν denote any unitarily invariant norm on $\mathcal{M}$. Then

$$\min_{X \in \mathcal{H}} \nu(A - X) = \nu(A - \tfrac{1}{2}(A + A^*)) = \tfrac{1}{2}\nu(A - A^*) \quad . \tag{1.28}$$

1.29 Theorem (Fan and Hoffman [9]). Let $\mathcal{U}$ denote the subset of all unitary matrices in $\mathcal{M}$ and let ν denote any unitarily invariant norm on $\mathcal{M}$. Let $A \in \mathcal{M}$ and suppose $A = UH$ where $H \in \mathcal{H}$ is positive semidefinite and $U \in \mathcal{U}$. Then

$$\min_{X \in \mathcal{U}} \nu(A - X) = \nu(\mathrm{diag}(\alpha_1, \ldots, \alpha_n) - I) = \nu(A - U) \tag{1.30}$$

where $\alpha_1, \ldots, \alpha_n$ are the singular values of A.

The next result is apparently new, although its interpretation (see Amir-Moez and Horn [2] and [9] regarding a well-known analogy between matrices and complex numbers) and the method of proof are strictly analogous to those associated with Theorem 1.27.

1.31 Theorem. Let $A \in \mathcal{M}$, let $\mathcal{S}$ denote the subset of all skew-hermitian matrices in $\mathcal{M}$, and let ν denote any unitarily invariant norm on $\mathcal{M}$. Then

$$\min_{X \in \mathcal{S}} \nu(A - X) = \nu(A - \tfrac{1}{2}(A - A^*)) = \tfrac{1}{2}\nu(A + A^*) \quad . \tag{1.32}$$

Proof. Let S be any skew-hermitian matrix. We have

$$A - \frac{A-A^*}{2} = \frac{A-S}{2} + \frac{A^*+S}{2} = \frac{A-S}{2} + \frac{(A-S)^*}{2}$$

whence

$$\nu(A - \tfrac{1}{2}(A - A^*)) \leq \tfrac{1}{2}\,\nu(A - S) + \tfrac{1}{2}\,\nu((A - S)^*) \quad .$$

By Theorem 1.11 and the remark following Definition 1.10, $\nu((A - S)^*) = \nu(A - S)$; consequently

$$\nu(A - \tfrac{1}{2}(A - A^*)) \leq \nu(A - S) \tag{1.33}$$

holds for all $S \in \mathcal{S}$. This proves (1.32).

Let k denote an integer such that $1 \leq k \leq n$. In [15] Mirsky solved distance problems for the subsets $\{X\ ;\ X \in \mathcal{M}_n$ and $\operatorname{rank}(X) \leq k\}$ and $\{X\ ;\ X \in \mathcal{M}_n$ and $\operatorname{rank}(X) = k\}$. As in the above results, the formula for the distance can be put in the same form for all unitarily invariant norms.

In this paper we shall be primarily interested in distance and extremum problems associated with the subset $\mathcal{N}$ of all normal matrices in $\mathcal{M}$. The distance problem was apparently first studied by Mirsky [15]. Let A be any fixed element of $\mathcal{M}$, let ν be any norm on $\mathcal{M}$ and define

$$d_\nu(A) = \inf_{X \in \mathcal{N}} \nu(A - X) \quad . \tag{1.34}$$

Mirsky was unable to determine $d_\nu(A)$, even for special choices of ν, but he obtained an upper bound for $d_\epsilon(A)$ (see Theorem 1.37 below) and

offered the following conjecture for the general solution when ν is unitarily invariant.

1.35 Conjecture (Mirsky). Let ν denote any unitarily invariant norm on $\mathcal{M}_n$. Then

$$d_\nu^2(A) = \tfrac{1}{2}(\nu^2(A) - \nu^2(\Omega(A))) \tag{1.36}$$

holds for all $A \in \mathcal{M}_n$, where $d_\nu^2(A) = [d_\nu(A)]^2$, $\nu^2(A) = [\nu(A)]^2$ and where $\Omega(A)$ is defined in Definition 1.17.

Note. By (1.9) the right side of (1.36) is independent of the order of the λ's in $\Omega(A)$. The singular values of A are the eigenvalues of the positive semidefinite square root (denoted by $(A^*A)^{\frac{1}{2}}$) of A^*A; consequently $\nu(A) = \nu(\Omega((A^*A)^{\frac{1}{2}}))$. A further interpretation of Mirsky's conjecture is contained in Chapter 2 where the nonnegativity of the right side of (1.36) is proved (Lemma 2.4).

1.37 Theorem (Mirsky). Let $A \in \mathcal{M}$. We have

$$d_\epsilon^2(A) = \inf_{X \in \mathcal{N}} \epsilon^2(A - X) \leq \tfrac{1}{2}(\epsilon^2(A) - |\mathrm{tr}(A^2)|) \quad . \tag{1.38}$$

1.39 Definition. Let $A \in \mathcal{M}_n$ and let ν be any norm on $\mathcal{M}_n$. A matrix $N \in \mathcal{N}_n$ such that

$$\nu(A - N) = d_\nu(A) = \inf_{X \in \mathcal{N}_n} \nu(A - X) \tag{1.40}$$

is called a ν-minimizing normal matrix (for A) or N is said to be ν-minimal (for A).

1.3 Differentiable Curves in Matric Space.

Let $A(t) = (a_{ij}(t))$ be a matrix function of the real variable t which is defined for $-\infty \leq a < t < b \leq \infty$. In the sequel we shall assume that each of the scalar functions $a_{ij}(t)$ is sufficiently differentiable throughout its domain of definition. We define the derivative of $A(t)$ by

$$\frac{dA}{dt} = \frac{d}{dt} A(t) = \left(\frac{da_{ij}}{dt}\right) \quad ; \tag{1.41}$$

higher order derivatives are defined in a similar fashion. The exponential function $\exp(A)$ is defined by the power series

$$e^A = I + \sum_{k=1}^{\infty} \frac{1}{k!} A^k \tag{1.42}$$

which converges for all $A \in \mathcal{M}$.

1.43 Lemma. Let $A(t)$, $B(t)$ be any differentiable matrices in $\mathcal{M}$ and let C be any constant matrix in $\mathcal{M}$. We have

$$\frac{d}{dt} [A(t)\, B(t)] = \frac{dA(t)}{dt} B(t) + A(t) \frac{dB(t)}{dt} \quad , \tag{1.44}$$

$$\frac{d}{dt} e^{tC} = e^{tC} \cdot C = C \cdot e^{tC} \quad , \tag{1.45}$$

$$\frac{d}{dt} \operatorname{tr}[A(t)] = \operatorname{tr}\left(\frac{dA(t)}{dt}\right) \quad , \tag{1.46}$$

$$\frac{d}{dt} \operatorname{dg}(A(t)) = \operatorname{dg}\left(\frac{dA(t)}{dt}\right) \quad , \tag{1.47}$$

(1.48) $$\frac{d}{dt} A^*(t) = \left[\frac{dA(t)}{dt}\right]^*$$

where $A^*(t) = [A(t)]^*$ and where dg is defined by (1.24).

The proof of Lemma 1.43 will be omitted since each equation is either very elementary or well known.

1.49 Definition. Let $\mathcal{S}$ be any subset of $\mathcal{M}$. A matrix function $A(t)$ whose range is in $\mathcal{S}$ and which is differentiable in some interval $a < t < b$ $(a < b)$ is called a differentiable curve in $\mathcal{S}$.

For any differentiable curve $A(t)$, the tangent vector to $A(t)$ at any $t_o \in (a,b)$ is defined to be the matrix $[dA(t)/dt]_{t=t_o}$.

Let $U(t)$ be an arbitrary differentiable curve in the subset $\mathcal{U}$ of all unitary matrices in $\mathcal{M}$. Then $U^*(t)\,U(t) = I$ and $U(t)\,U^*(t) = I$ are identities in t. Differentiating these identities and using the rule (1.44) we obtain

(1.50) $$\frac{dU^*(t)}{dt} U(t) + U^*(t) \frac{dU(t)}{dt} = 0 \quad ,$$

(1.51) $$\frac{dU(t)}{dt} U^*(t) + U(t) \frac{dU^*(t)}{dt} = 0 \quad .$$

From either of the last two equations we obtain

(1.52) $$\frac{dU(t)}{dt} = - U(t) \frac{dU^*(t)}{dt} U(t) \quad .$$

1.53 Lemma. Let $U(t)$ be any differentiable curve in $\mathcal{U}$ and let $S_1(t)$, $S_2(t)$ be given by the equations

(1.54) $$S_1(t) = - \frac{dU^*(t)}{dt} U(t) \quad ,$$

(1.55) $$S_2(t) = - U(t) \frac{dU^*(t)}{dt} \quad .$$

Then $S_1(t)$ and $S_2(t)$ are skew-hermitian for every $t \in (a,b)$ and the tangent vector to $U(t)$ at t_o is given by

(1.56) $$\left.\frac{dU(t)}{dt}\right|_{t=t_o} = U(t_o)\, S_1(t_o) = S_2(t_o)\, U(t_o) \quad .$$

Proof. Using (1.48) and the definitions of $S_1(t)$ and $S_2(t)$ we see from (1.50) and (1.51) that $S_1(t) + S_1^*(t) = 0$, $S_2^*(t) + S_2(t) = 0$; i.e., S_1 and S_2 are skew-hermitian for each value of t. The expression (1.56) follows immediately from (1.52).

Let S be any skew-hermitian matrix. It is easy to see that $\exp(tS)$ is unitary for all finite values of the real variable t. In fact, from the definition of the exponential function, $[\exp(tS)]^* = \exp(-tS)$ and, since tS commutes with $-tS$, we have

$$(e^{tS})^*\, e^{tS} = e^{-tS}\, e^{tS} = e^{tS-tS} = e^0 = I \quad .$$

Thus, letting $U(t) = \exp(tS)$, we see from (1.45) that

$$\frac{dU(t)}{dt} = U(t)\, S = S\, U(t) \quad ;$$

consequently any skew-hermitian matrix S can occur in place of $S_1(t_o)$ and $S_2(t_o)$ in (1.56) (for some differentiable curve $U(t)$ in $\mathcal{U}$) and for any value of t_o.

1.4 Normal Matrices.

In the next theorem we list several known characterizations of normal matrices, already defined in Section 1.1.

1.57 Theorem. Let the eigenvalues of $N \in \mathcal{M}_n$ be denoted by $\lambda_1, \lambda_2, \ldots, \lambda_n$. Then N is normal if and only if any one of the following propositions is true:

(a) $N = H_1 + iH_2$ where H_1 and H_2 are hermitian and $H_1H_2 = H_2H_1$.

(b) N has a complete orthonormal set of eigenvectors.

(c) (Toeplitz [23]) N is unitarily similar to a diagonal matrix:

$$N = U^*DU \qquad (U \in \mathcal{U},\ D \in \mathcal{D}) .$$

(d) (Wintner and Murnaghan [26] and Williamson [25], see also Halmos [11], pp. 169-170) There is a positive semidefinite hermitian matrix H and a unitary matrix U such that

$$N = UH = HU . \tag{1.58}$$

(e) (Parker [20], p. 522) There exists a unitary matrix U such that $U(N + N^*)U^* \in \mathcal{D}$ and $U(N - N^*)U^* \in \mathcal{D}$.

(f) (Parker [20], Theorem 1) The eigenvalues of NN^* are $|\lambda_1|^2, |\lambda_2|^2, \ldots, |\lambda_n|^2$.

(g) (Parker [20], Theorem 2) The eigenvalues of $N + N^*$ are $\lambda_1 + \bar{\lambda}_1, \lambda_2 + \bar{\lambda}_2, \ldots, \lambda_n + \bar{\lambda}_n$.

1.59 Theorem (Toeplitz [23] and Parker [20]). A triangular matrix in $\mathcal{M}$ is normal if and only if it is diagonal.

We shall be interested later in utilizing differentiable curves in $\mathcal{N}$. One way of constructing such curves is to use Theorem 1.57 (c) and differentiable curves $U(t)$ and $D(t)$ in $\mathcal{U}$ and $\mathcal{D}$ respectively. Then $N(t) = U^*(t)\, D(t)\, U(t)$ is a differentiable curve in $\mathcal{N}$. Furthermore we can construct differentiable curves in $\mathcal{N}$ which pass through a given normal matrix $N_o = U_o^* D_o U_o$ ($U_o \in \mathcal{U}$, $D_o \in \mathcal{D}$) for some value of t (say $t = 0$) by merely requiring that $U(0) = U_o$ and $D(0) = D_o$. For curves $U(t)$ in $\mathcal{U}$ we shall use the formula

$$U(t) = U_o e^{itH} \tag{1.60}$$

where H is hermitian. There is no loss of generality in restricting ourselves to the formula (1.60) since we shall be concerned with evaluations of the derivative of $U(t)$ at $t = 0$. Note that, as H runs through $\mathcal{H}$, iH runs through the set of all skew-hermitian matrices, so, by Lemma 1.53, all possible tangent vectors to a differentiable curve $U(t)$ in $\mathcal{U}$ at $t = 0$ can occur for curves of the type given by (1.60)

We shall have a need later on for the following result concerning differentiable curves in $\mathcal{D}$.

1.61 Lemma. If $D(t)$ is any differentiable curve in $\mathcal{D}$, then $dD(t)/dt\big|_{t=0}$ is in $\mathcal{D}$. Furthermore every $\Lambda \in \mathcal{D}$ can occur as $dD(t)/dt\big|_{t=0}$ for some differentiable curve $D(t)$ in $\mathcal{D}$.

Proof. Let $D(t) = \operatorname{diag}(\alpha_1(t), \ldots, \alpha_n(t))$. Each of the scalar functions $\alpha_i(t)$ has a scalar derivative so $dD(t)/dt$ is in $\mathcal{D}$ for every value of t. Let z_i $(i = 1, \ldots, n)$ be any complex numbers. Then the derivative of $D(t) = t \operatorname{diag}(z_1, z_2, \ldots, z_n)$ equals $\operatorname{diag}(z_1, \ldots, z_n)$, an arbitrary matrix in $\mathcal{D}$, for all t.

1.62 Theorem. Let $X, Y \in \mathcal{N}$. Then $N = X + Y \in \mathcal{N}$ if and only if

$$XY^* - Y^*X + YX^* - X^*Y = 0 \quad . \tag{1.63}$$

Proof. Since X and Y are normal and have

$$\begin{aligned} NN^* - N^*N &= (X + Y)(X^* + Y^*) - (X^* + Y^*)(X + Y) \\ &= XX^* + XY^* + YX^* + YY^* - (X^*X + X^*Y + Y^*X + Y^*Y) \\ &= XY^* - Y^*X + YX^* - X^*Y \end{aligned}$$

whence N is normal if and only if (1.63) holds.

1.64 Corollary. Let z be any fixed complex number. Then $N \in \mathcal{N}$ if and only if $N + zI \in \mathcal{N}$.

Proof. Clearly $zI \in \mathcal{N}$ for all $z \in \underline{C}$. Setting $X = N$, $Y = zI$ we find that (1.63) is satisfied for all z:

$$XY^* - Y^*X + YX^* - X^*Y = \bar{z}N - \bar{z}N + zN^* - zN^* = 0 \quad .$$

Thus, by Theorem 1.62, $N \in \mathcal{N}$ implies $N + zI \in \mathcal{N}$. The converse implication is clear since $(N + zI) - zI = N$.

1.65 Theorem. Let $X \in \mathcal{N}$. Then $X + tY \in \mathcal{N}$ for all values of t in a real interval of positive length if and only if $Y \in \mathcal{N}$ and (1.63) holds.

Proof. Let $N = X + tY$. A short computation, using the fact that t is real and $X \in \mathcal{N}$, yields

$$(1.66) \quad NN^* - N^*N = t^2(YY^* - Y^*Y) + t(XY^* - Y^*X + YX^* - X^*Y) \ .$$

A polynomial of second degree can have at most two zeros unless all of its coefficients vanish. Thus the assumption that $N \in \mathcal{N}$ for more than two values of t implies immediately (1.63) and $YY^* - Y^*Y = 0$ i.e., $Y \in \mathcal{N}$. Conversely $Y \in \mathcal{N}$ and (1.63) imply via (1.66) that $X + tY \in \mathcal{N}$ for all real values of t.

1.67 Theorem. Let $P_k(X) = z_0 I + \sum_{i=1}^{k} z_i X^i$ denote a polynomial of degree $k(\geq 1)$ in a matrix $X \in \mathcal{M}$ with arbitrary complex coefficients z_i. Then $N \in \mathcal{N}$ implies $P_k(N) \in \mathcal{N}$.

Proof. If $N \in \mathcal{N}$ then by Theorem 1.57 (c) it has a decomposition $N = U^*DU$ $(U \in \mathcal{U}, D \in \mathcal{D})$ and clearly $N^i = U^*D^iU$ for all positive integers i. Thus $P_k(N) = U^*P_k(D)U$, $P_k(D) \in \mathcal{D}$ so by Theorem 1.57 (c) $P_k(N)$ is normal.

1.68 Corollary. Suppose $N \in \mathcal{N}$. Then $N^k \in \mathcal{N}$ for $k = 2, 3, 4, \ldots$. Furthermore if N^{-1} exists then $N^{-k} \in \mathcal{N}$ for $k = 1, 2, 3, \ldots$.

Proof. That $N^k \in \mathcal{N}$ for $k \geq 2$ is obvious from Theorem 1.67. If N^{-1} exists then, by the Cayley-Hamilton Theorem, N^{-1} is a polynomial in N whence N^{-1} is normal. Applying Theorem 1.67 again we see that

$N^{-k} = (N^{-1})^k$ is normal for $k \geq 2$.

1.69 Theorem. Let $A \in \mathcal{M}$ and let α and β be complex numbers. If A is normal then $\alpha A + \beta A^* \in \mathcal{N}$ for all α,β; if A is not normal, $\alpha A + \beta A^* \in \mathcal{N}$ if and only if $|\alpha| = |\beta|$

Proof. Setting $N = \alpha A + \beta A^*$ we find that

$$NN^* - N^*N = (|\alpha|^2 - |\beta|^2)(AA^* - A^*A) \quad . \tag{1.70}$$

The conclusions of the theorem follow immediately from the relation (1.70).

1.71 Corollary. If $A \in \mathcal{M}$ and $\alpha,\beta \in \underline{C}$ with $|\alpha| = |\beta|$, then $P_k(\alpha A + \beta A^*) \in \mathcal{N}$ where P_k is an arbitrary polynomial of degree k with coefficients in $\underline{C}$.

Proof. This is an obvious consequence of Theorems 1.67 and 1.69.

1.72 Theorem. Suppose $\alpha \in \underline{C}$ $(\alpha \neq 0)$, $z \in \underline{C}$, and $U \in \mathcal{U}$ are fixed. Then each of the transformations

$$T_\alpha(N) = \alpha N \quad , \tag{1.73}$$

$$T_z(N) = N + zI \quad , \tag{1.74}$$

$$T_U(N) = U^*NU \tag{1.75}$$

defines a one-to-one mapping of $\mathcal{N}$ onto itself.

Proof. Let N be any normal matrix. Then from Theorem 1.67, Corollary 1.64, and Theorem 1.57 (c) we see that $\alpha^{-1}N$, $N - zI$, and UNU^* respectively are in $\mathcal{N}$. Thus $N = T_\alpha(\alpha^{-1}N) = T_z(N-zI) = T_U(UNU^*)$ which proves that each of the transformations (1.73), (1.74), and (1.75)

is onto. Letting N_1, N_2 denote any pair of normal matrices, one sees easily that any one of the equations $T_\alpha(N_1) = T_\alpha(N_2)$, $T_z(N_1) = T_z(N_2)$, $T_U(N_1) = T_U(N_2)$ implies $N_1 = N_2$, whence each of the transformations is one-to-one.

1.5 Existence of ν-Minimal Matrices.

The definition of a ν-minimal (or ν-minimizing normal) matrix has already been given in Section 1.2 (Definition 1.39). Let ν denote any norm on $\mathcal{M}$. Suppose first that $A \in \mathcal{N}$. Then there is a unique ν-minimal N_o, namely $N_o = A$; for if $A \in \mathcal{N}$ then $d_\nu(A) = 0$ and the infimum in (1.34) is assumed if and only if $X = A$. Our main purpose in the present section is to show that, for any $A \in \mathcal{M}$ and any norm ν, there exists a ν-minimal matrix.

Note. Any two normed linear spaces (over $\underline{C}$) of the same finite dimension are topologically isomorphic (see e.g., [13], p. 13). This implies that the norm topologies induced in $\mathcal{M}_n$ by any two norms are the same; consequently there is only one norm topology for $\mathcal{M}_n$ and we refer to it as the norm topology of $\mathcal{M}_n$.

1.76 Lemma. The set $\mathcal{N}_n$ of all normal matrices (in $\mathcal{M}_n$) is closed in the norm topology of $\mathcal{M}_n$.

Proof. Matrix multiplication is continuous in the norm topology since it is continuous with respect to the ϵ-norm topology (cf. Section 1.1). Let N be any matrix in the closure $\overline{\mathcal{N}}$ of $\mathcal{N}$. Then there is a sequence $\{N_i\}$ ($N_i \in \mathcal{N}$ for $i = 1, 2, 3, \ldots$) such that $N_i \to N$. Since each N_i is normal we have

(1.77) $$N_i N_i^* = N_i^* N_i \qquad (i = 1, 2, \ldots) \ .$$

By virtue of the continuity of multiplication we may pass to the limit in (1.77) and obtain $NN^* = N^*N$, i.e., N is normal. This implies $\overline{\mathcal{N}} \subset \mathcal{N}$ which proves that $\mathcal{N}$ is closed.

1.78 Theorem. Let ν be any norm on $\mathcal{M}_n$ and let A be any fixed matrix in $\mathcal{M}_n$. Then there is a ν-minimal matrix N_o (for A).

Proof. We assume $d_\nu(A) > 0$ (see (1.34)) since otherwise the theorem is trivial. By the definition of $d_\nu(A)$, there is a sequence $\{N_i\}$ of normal matrices such that $\nu(A - N_i) \to d_\nu(A)$. The subset $\mathcal{R} = \{X \ ; \ X \in \mathcal{M}_n$ and $d_\nu(A) \leq \nu(A - X) \leq 2d_\nu(A)\}$ is closed and bounded, hence compact, in $\mathcal{M}_n$. Clearly there is an index k_o so that $N_i \in \mathcal{R}$ for $i \geq k_o$. Thus, since $\mathcal{R}$ is compact (and therefore countably compact), there is a subsequence $\{N_{i_k}\}$ $(i_k \geq k_o$ for $k = 1, 2, 3, \ldots)$ which converges to a matrix $N_o \in \mathcal{R}$. This N_o is a point of closure of $\mathcal{N}$ so, by Lemma 1.76, it is normal. Finally

(1.79) $$d_\nu(A) \leq (A - N_o) \leq \nu(A - N_{i_k}) + \nu(N_{i_k} - N_o) \ ;$$

and, since $\nu(A - N_{i_k}) \to d_\nu(A)$, we have

$$\nu(A - N_{i_k}) = d_\nu(A) + \delta_k \geq d_\nu(A)$$

where $\delta_k \to 0$ as $k \to \infty$. Thus, given $\epsilon > 0$, there is an index k_1 such that the right side of (1.79) is less than $d_\nu(A) + \epsilon$ for $k \geq k_1$. Since

ϵ is arbitrary we have $d_\nu(A) \leq \nu(A - N_o) \leq d_\nu(A)$ which completes the proof of Theorem 1.78.

Theorem 1.78 shows that the distance problem of finding $d_\nu(A)$ is actually a minimum problem for all ν and for all A. This suggests the possibility of finding $d_\nu(A)$ by determining a ν-minimal matrix. We shall investigate this aspect of the distance problem in subsequent sections of this paper.

Since $\mathcal{N}_n$ is not a convex set (the sum of two normal matrices is not necessarily normal), we naturally expect that there might exist matrices $A \in \mathcal{N}_n$ for which there is no unique ν-minimal matrix. We shall show, at least for $\nu = \epsilon$ and $n = 2$, that this is the case.

CHAPTER 2

EQUIVALENCE THEOREMS CONCERNING MIRSKY'S CONJECTURE, ν-MINIMAL MATRICES, AND DISTANCE FORMULAS.

Now that the existence of ν-minimal matrices has been established, it is of interest to determine what transformations of $\mathcal{M}_n$ leave the property of minimality invariant. In this section we shall give three results of this type and we shall also prove three closely parallel results concerning distance formulas resembling Mirsky's formula (1.36). We begin with an examination of the meaning of Mirsky's conjecture.

2.1 Mirsky's Conjecture.

2.1 Definition. A norm ν on $\mathcal{M}_n$ such that

$$\nu(A) = \nu(\Omega(A)) \qquad \text{for all } A \in \mathcal{M}_n \tag{2.2}$$

and

$$\nu(A) > \nu(\Omega(A)) \qquad \text{for all } A \in \mathcal{M}_n,\ A \notin \mathcal{N}_n\ , \tag{2.3}$$

where $\Omega(A)$ is defined in Definition 1.17, is said to have property S.

2.4 Lemma. The Euclidean norm ϵ has property S. Furthermore, for any unitarily invariant norm ν, we have

$$\nu(\Omega((A^*A)^{\frac{1}{2}})) = \nu(A) \geq \nu(\Omega(A)) \qquad \text{for all } A \in \mathcal{M}_n \tag{2.5}$$

with equality holding in (2.5) for all $A \in \mathcal{N}_n$.

Proof. By a well-known theorem of Schur [22] (see e.g., [16], p. 307) every $A \in \mathcal{M}_n$ is unitarily similar to a triangular matrix:

$$VAV^* = \Omega(A) + M \qquad (V \in \mathcal{U}_n) \tag{2.6}$$

where in M only elements above the principal diagonal may be different from zero. Thus, by (1.16), we have

$$\epsilon^2(A) = \epsilon^2(VAV^*) = \epsilon^2(\Omega(A)) + \epsilon^2(M)$$

so

$$\epsilon^2(A) - \epsilon^2(\Omega(A)) = \epsilon^2(M) \tag{2.7}$$

By Theorem 1.59 $M = 0$ if and only if $A \in \mathcal{N}_n$; hence from (2.7) we see that ϵ has property S.

In order to prove (2.5) we shall need the following two results.

2.8 Theorem (Fan [8], Theorem 4). Let $a_1 \geq a_2 \geq \cdots \geq a_n \geq 0$, $b_1 \geq b_2 \geq \cdots \geq b_n \geq 0$. Then

$$\varphi(a_1, a_2, \ldots, a_n) \leq \varphi(b_1, b_2, \ldots, b_n) \tag{2.9}$$

holds for all symmetric gauge functions φ of n real variables if and only if

$$\sum_{i=1}^{k} a_i \leq \sum_{i=1}^{k} b_i \qquad (1 \leq k \leq n) \ . \tag{2.10}$$

2.11 Theorem (Weyl [24], p. 409). Let the eigenvalues λ_i and singular values α_i of $A \in \mathcal{M}_n$ be arranged so that $|\lambda_1| \geq |\lambda_2| \geq \cdots \geq |\lambda_n|$, $\alpha_1 \geq \alpha_2 \geq \cdots \geq \alpha_n$. Then, for any real exponent $s > 0$,

$$(2.12) \qquad \sum_{i=1}^{k} |\lambda_i|^s \leq \sum_{i=1}^{k} \alpha_i^s \qquad (1 \leq k \leq n) \ .$$

Since $\Omega(A)$ is diagonal (and therefore normal) its singular values are $|\lambda_1|, \ldots, |\lambda_n|$ (see Theorem 1.57 (f)). The inequality (2.5) now follows immediately by combining Theorem 1.11, Theorem 2.8 (here we set $a_i = |\lambda_i|$, $b_i = \alpha_i$), and Theorem 2.11 (use $s = 1$). If A is normal it is unitarily similar to $\Omega(A)$ whence equality holds in (2.5) for all $A \in \mathcal{N}_n$.

The question of which unitarily invariant norms $\nu \neq \epsilon$ have property S (i.e., which ones satisfy (2.3)) is apparently open. We shall presently show that the spectral norm σ does not have property S.

2.13 Definition. Let ν be any norm on $\mathcal{M}_n$ which has property S. Then we define

$$(2.14) \qquad p_{\nu,n}(A) = \frac{d_\nu^2(A)}{\nu^2(A) - \nu^2(\Omega(A))} \qquad \text{for } A \notin \mathcal{N}_n \ .$$

For any ν with property S we see from the definition of $d_\nu(A)$ and (2.3) that $p_{\nu,n}(A)$ is a well-defined positive quantity so that $d_\nu(A)$ can be expressed in the form

$$(2.15) \quad d_\nu^2(A) = p_{\nu,n}(A)(\nu^2(A) - \nu^2(\Omega(A))) \text{ for all } A \in \mathcal{M}_n, A \notin \mathcal{N}_n \ .$$

If we assigned to $p_{\nu,n}(A)$ some convenient finite value for $A \in \mathcal{N}_n$, then (2.15) would be valid for all $A \in \mathcal{M}_n$. Note that (2.15) has the same general form as Mirsky's formula (1.36). We now prove the following characterization of Mirsky's conjecture for a particular norm ν.

2.16 Theorem. Let ν be a norm on $\mathcal{M}_n$. Mirsky's conjectured formula (1.36) for $d_\nu(A)$ holds for all $A \in \mathcal{M}_n$ if and only if ν has property S and

$$p_{\nu,n}(A) = \frac{1}{2} \tag{2.17}$$

for all $A \in \mathcal{M}_n$, $A \notin \mathcal{N}_n$ and for $n = 2, 3, \ldots$.

Proof. Since $\mathcal{N}_n$ is closed as a subset of $\mathcal{M}_n$, we see from (1.34) that $d_\nu(A)$ is zero when $A \in \mathcal{N}_n$ and strictly positive when $A \notin \mathcal{N}_n$. Thus, if (1.36) holds for all $A \in \mathcal{M}_n$, ν has property S and (2.17) holds. The converse is obvious.

Let the n by n matrix A $(n \geq 3)$ be given by

$$A = \begin{pmatrix} B & 0 \\ 0 & 0 \end{pmatrix} \quad \text{where} \quad B = \begin{pmatrix} 2 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 0 \end{pmatrix}. \tag{2.18}$$

The nonzero singular values of A are $2, \sqrt{2}$ and the nonzero singular values of $\Omega(A)$ are $2, 1$. From (1.14) and (1.15) we have $\sigma(A) = \sigma(\Omega(A)) = 2$ while, by Theorem 1.59, A is not normal. Thus (2.3) does not hold for $\nu = \sigma$; consequently σ fails to have property S for $n \geq 3$. We have proved

2.19 Theorem. The spectral norm σ does not have property S for $n \geq 3$. Mirsky's conjecture 1.35 is false for $\nu = \sigma$ and for $n \geq 3$.

In Chapter 8 where we discuss some other counterexamples we shall prove that Mirsky's conjecture also fails for $\nu = \sigma$ and $n = 2$. (See Theorem 8.9).

2.2 Invariance of ν-Minimal Matrices.

The content of the next three theorems is, roughly speaking, the following: ν-minimality is generally invariant under the transformations (1.73) - (1.75). The accompanying corollaries show that uniqueness of a ν-minimal matrix is also invariant under the same transformations.

2.20 Theorem. Let ν be any norm on $\mathcal{M}_n$ and suppose $\alpha \in \underline{C}$, $\alpha \neq 0$. Then N_o is ν-minimal for $A \in \mathcal{M}_n$ if and only if αN_o is ν-minimal for αA. Furthermore

$$d_\nu(\alpha A) = |\alpha| d_\nu(A) \quad . \tag{2.21}$$

Proof. By Theorem 1.72 αN runs through $\mathcal{N}$ as N runs through $\mathcal{N}$. Consequently $\nu(A - N)$ assumes its minimum for $N = N_o$ if and only if $|\alpha| \cdot \nu(A - N) = \nu(\alpha A - \alpha N)$ assumes its minimum for $N = N_o$. If αN_o is ν-minimal for αA then

$$d_\nu(\alpha A) = \nu(\alpha A - \alpha N_o) = |\alpha| \nu(A - N_o) = |\alpha| d_\nu(A) \quad .$$

2.22 Corollary. Let ν be any norm on $\mathcal{M}_n$ and suppose $\alpha \in \underline{C}$, $\alpha \neq 0$. $A \in \mathcal{M}_n$ has a unique ν-minimal matrix if and only if αA has a unique ν-minimal matrix.

Proof. If N_o is the unique ν-minimal matrix for A and if N_1, N_2 are ν-minimal for αA, then by Theorem 2.20 $N_o = \alpha^{-1}N_1 = \alpha^{-1}N_2$ whence $N_1 = N_2$. The converse is proved in a similar manner.

2.23 Theorem. Let ν denote any norm on $\mathcal{M}$. If N_o is ν-minimal for $A \in \mathcal{M}$ then $N_o + zI$ is ν-minimal for $A + zI$ for all $z \in \underline{C}$. Conversely, if $N_o + zI$ is ν-minimal for $A + zI$ for one value of $z \in \underline{C}$ then N_o is ν-minimal for A. Furthermore, for all $z \in \underline{C}$, we have

$$d_\nu(A + zI) = d_\nu(A) \quad . \tag{2.24}$$

Proof. Obviously

$$\nu(A - N) = \nu[(A + zI) - (N + zI)] \tag{2.25}$$

holds for any norm ν, for all $A, N \in \mathcal{M}$, and for all $z \in \underline{C}$. By Theorem 1.72 $T_z(N) = N + zI$ runs through $\mathcal{M}$ in a one-to-one manner as N runs through $\mathcal{M}$. Thus, as N runs through $\mathcal{M}$, the left and right sides of (2.25) assume their minima simultaneously. This proves the first two statements of Theorem 2.23. The relation (2.24) follows immediately from (2.25) if one assumes that N is ν-minimal for A.

2.26 Corollary. Let ν be any norm on $\mathcal{M}$ and suppose $z \in \underline{C}$. Then $A \in \mathcal{M}$ has a unique ν-minimal matrix if and only if $A + zI$ has a unique ν-minimal matrix.

Proof. The proof is strictly analogous to that for Corollary 2.22.

2.27 Theorem. Let ν denote any unitarily invariant norm on $\mathcal{M}$ and let $U \in \mathcal{U}$ by fixed. Then N_o is ν-minimal for $A \in \mathcal{M}$ if and only if

U^*N_oU is ν-minimal for U^*AU. Furthermore

$$d_\nu(U^*AU) = d_\nu(A) \quad . \tag{2.28}$$

Proof. Since ν is unitarily invariant

$$\nu(A - N) = \nu(U^*AU - U^*NU) \tag{2.29}$$

holds for all $A, N \in \mathcal{M}$ and for all $U \in \mathcal{U}$. By Theorem 1.72 $T_U(N) = U^*NU$ runs through $\mathcal{M}$ in a one-to-one manner as N runs through $\mathcal{M}$. Therefore, as N runs through $\mathcal{M}$, the left and right sides of (2.29) assume their minima simultaneously. This proves the first assertion in Theorem 2.27. Equation (2.28) follows from (2.29) if it is assumed that N is ν-minimal for A.

2.30 Corollary. Let ν be any unitarily invariant norm on $\mathcal{M}$ and let $U \in \mathcal{U}$. Then $A \in \mathcal{M}$ has a unique ν-minimal matrix if and only if U^*AU has a unique ν-minimal matrix.

The proof is analogous to that for Corollary 2.22 and is therefore omitted.

2.3 Invariance of Distance Formulas.

2.31 Definition. A transformation T whose domain space is $\mathcal{M}_n$ and whose range space is contained in $\mathcal{M}_n$ is said to be discriminating if

$$T(M) \notin \mathcal{N}_n \qquad \text{for } M \notin \mathcal{N}_n$$

and

$$T(M) \in \mathcal{N}_n \qquad \text{for } M \in \mathcal{N}_n \ .$$

We have already seen that, for any norm ν with property S, $d_\nu(A)$ can be represented by the formula

$$(2.32) \qquad d_\nu^2(A) = p_{\nu,n}(A)(\nu^2(A) - \nu^2(\Omega(A))), \ (A \in \mathcal{M}_n, \ A \notin \mathcal{N}_n)$$

where $p_{\nu,n}(A)$ is defined by (2.14). The question naturally arises as to what happens to $p_{\nu,n}(A)$ when A is subjected to various discriminating transformations. The next three theorems provide some answers to this question for the discriminating transformations (1.73) - (1.75).

2.33 Theorem. Let ν be any norm on $\mathcal{M}_n$ which has property S. Let $\alpha \in \underline{C}$, $\alpha \neq 0$ and let $A \in \mathcal{M}_n$, $A \notin \mathcal{N}_n$. Then

$$(2.34) \qquad p_{\nu,n}(\alpha A) = p_{\nu,n}(A) \ .$$

Proof. If $\lambda_1, \lambda_2, \ldots, \lambda_n$ are the eigenvalues of A then $\alpha\lambda_1, \alpha\lambda_2, \ldots, \alpha\lambda_n$ are the eigenvalues of αA; consequently (cf. Definition 1.17) $\Omega(\alpha A) = \alpha\Omega(A)$. Using that fact, (2.32), and (2.21) we obtain

$$\begin{aligned} d_\nu^2(\alpha A) &= |\alpha|^2 \, d_\nu^2(A) \\ &= p_{\nu,n}(A)(\nu^2(\alpha A) - \nu^2(\alpha\Omega(A))) \\ &= p_{\nu,n}(A)(\nu^2(\alpha A) - \nu^2(\Omega(\alpha A))) \ . \end{aligned}$$

The first factor in the last line must be $p_{\nu,n}(\alpha A)$ which proves (2.34).

2.35 Theorem. Let ϵ denote the Euclidean norm (1.16) on $\mathcal{M}_n$. Let A be any nonnormal matrix in $\mathcal{M}_n$ and let $z \in \underline{C}$. Then

$$p_{\epsilon,n}(A + zI) = p_{\epsilon,n}(A) \quad . \tag{2.36}$$

Proof. Let λ_i $(i = 1,2, \ldots, n)$ denote the eigenvalues of $A = (a_{ij})$. From (1.16) we have

$$\epsilon^2(A + zI) = \sum_{i \neq j} |a_{ij}|^2 + \sum_{i=1}^{n} |a_{ii} + z|^2$$

$$= \epsilon^2(A) + \bar{z} \sum_{i=1}^{n} a_{ii} + z \sum_{i=1}^{n} \bar{a}_{ii} + n|z|^2 \quad ;$$

and, since $\sum_1^n \lambda_i = \sum_1^n a_{ii}$, we obtain

$$\epsilon^2(A + zI) = \epsilon^2(A) + \bar{z} \sum_{i=1}^{n} \lambda_i + z \sum_{i=1}^{n} \bar{\lambda}_i + n|z|^2 \quad . \tag{2.37}$$

Furthermore, since $(\lambda_i + z)$ $(i = 1, \ldots, n)$ are the eigenvalues of $A + zI$,

$$\epsilon^2(\Omega(A + zI)) = \sum_{i=1}^{n} (\lambda_i + z)(\bar{\lambda}_i + \bar{z})$$

$$= \epsilon^2(\Omega(A)) + \bar{z} \sum_{i=1}^{n} \lambda_i + z \sum_{i=1}^{n} \bar{\lambda}_i + n|z|^2 \quad . \tag{2.38}$$

Combining (2.37) and (2.38) we find that

$$\epsilon^2(A) - \epsilon^2(\Omega(A)) = \epsilon^2(A + zI) - \epsilon^2(\Omega(A + zI)) \quad . \tag{2.39}$$

Applying (2.24) in the case $\nu = \epsilon$ and using (2.39) we obtain

$$d_\epsilon^2(A + zI) = d_\epsilon^2(A) = p_{\epsilon,n}(A)(\epsilon^2(A) - \epsilon^2(\Omega(A)))$$

$$= p_{\epsilon,n}(A)(\epsilon^2(A + zI) - \epsilon^2(\Omega(A + zI))) \quad .$$

The relation (2.36) now follows from the last equation.

2.40 Theorem. Let ν be any unitarily invariant norm on $\mathcal{M}_n$ which has property S. Let A be any nonnormal matrix in $\mathcal{M}_n$ and let $U \in \mathcal{U}_n$. Then

$$p_{\nu,n}(U^*AU) = p_{\nu,n}(A) \quad . \tag{2.41}$$

Proof. The eigenvalues of a matrix are invariant under a unitary similarity transformation so $\Omega(U^*AU) = \Omega(A)$. Consequently, since ν is unitarily invariant, we have

$$\nu^2(A) - \nu^2(\Omega(A)) = \nu^2(U^*AU) - \nu^2(\Omega(U^*AU)) \quad . \tag{2.42}$$

From (2.28) we have

$$d_\nu^2(U^*AU) = d_\nu^2(A) = p_{\nu,n}(A)(\nu^2(A) - \nu^2(\Omega(A)))$$

and, combining this with (2.42) we obtain (2.41).

CHAPTER 3

IMPROVEMENTS OF MIRSKY'S BOUND

3.1 A New Bound.

In this section we shall obtain an upper bound for $d_\epsilon^2(A)$ which is sharper than the bound (1.38) obtained by Mirsky. Before doing this we prove a lemma which sheds some light on Mirsky's result (Theorem 1.37). This lemma furnishes at least a partial answer to the question: given $A \in \mathcal{M}_n$, what normal matrices lie at the distance from A which is given by Mirsky's bound?

3.1 Definition. For any $A \in \mathcal{M}_n$ we define

$$\eta(A) = \begin{cases} \{\mathrm{tr}(A^2)/|\mathrm{tr}(A^2)|\} & \text{if } \mathrm{tr}(A^2) \neq 0 \\ \{z \,;\, z \in \underline{C} \text{ and } |z| = 1\} & \text{if } \mathrm{tr}(A^2) = 0 \,; \end{cases} \tag{3.2}$$

$$M_o(A) = \{\tfrac{1}{2}(A + \eta A^*) \,;\, \eta \in \eta(A)\} \,. \tag{3.3}$$

Note. $\eta(A)$ is a set of complex numbers of unit modulus. If $\mathrm{tr}(A^2) \neq 0$ then $\eta(A)$ contains a single well-defined number but, if $\mathrm{tr}(A^2) = 0$, $\eta(A)$ consists of all complex numbers on the boundary of the unit disk. An analogous remark applies to the set $M_o(A)$.

3.4 Lemma. Let A be any matrix in $\mathcal{M}_n$. Then every matrix in the set $M_o(A)$ is normal. Furthermore

(3.5) $\epsilon^2(A - X) = \frac{1}{2}(\epsilon^2(A) - |tr(A^2)|)$ for all $X \in M_o(A)$.

Proof. The fact that $M_o(A) \subset \mathcal{N}_n$ follows immediately from Theorem 1.69. Assuming that $|\eta| = 1$ and using the properties (1.20) and (1.22) of the trace, we obtain

$$\epsilon^2(A - \tfrac{1}{2}(A + \eta A^*)) = \tfrac{1}{4} tr[(A^* - \bar{\eta}A)(A - \eta A^*)]$$

(3.6)

$$= \tfrac{1}{4}[tr(A^*A) - \eta\, tr(A^{*2}) - \bar{\eta}\, tr(A^2) + tr(AA^*)] .$$

Using (1.22) again and (1.21) we see that the last equation becomes

(3.7) $$\epsilon^2(A - \tfrac{1}{2}(A + \eta A^*)) = \tfrac{1}{2}\epsilon^2(A) - \tfrac{1}{4}(\eta\, \overline{tr(A^2)} + \bar{\eta}\, tr(A^2)) .$$

If $tr(A^2) \neq 0$ and $\eta \in \eta(A)$ then (3.2) implies $\eta\, \overline{tr(A^2)} = \bar{\eta}\, tr(A^2) = |tr(A^2)|$, whence (3.7) implies (3.5). If $tr(A^2) = 0$ then the last term on the right side of (3.7) vanishes for all $\eta \in \eta(A)$ and the right sides of both (3.5) and (3.7) reduce to $(1/2)\,\epsilon^2(A)$. This proves Lemma 3.4.

Remark. The author believes that the set $M_o(A)$ contains all matrices of the form $\alpha A + \beta A^*$ (with $|\alpha| = |\beta|$) which satisfy

$$\epsilon^2[A - (\alpha A + \beta A^*)] = \tfrac{1}{2}(\epsilon^2(A) - |tr(A^2)|) .$$

However, no attempt will be made here to prove that assertion.

3.8 Definition. For any $A \in \mathcal{N}_n$ we define

$$(3.9)\qquad \zeta(A) = \begin{cases} \{(\mathrm{tr}(A^2) - (1/n)(\mathrm{tr}(A))^2)/|\mathrm{tr}(A^2) - (1/n)(\mathrm{tr}(A))^2|\} \\ \qquad\qquad \text{if } \mathrm{tr}(A^2) \neq \frac{1}{n}(\mathrm{tr}(A))^2 \\ \{z\ ;\ z \in \underline{C} \text{ and } |z| = 1\} \text{ if } \mathrm{tr}(A^2) = \frac{1}{n}(\mathrm{tr}(A))^2 \end{cases};$$

$$(3.10)\qquad M(A) = \{\tfrac{1}{2}(A + \zeta A^*) + \tfrac{1}{2n}\mathrm{tr}(A - \zeta A^*)I\ ;\ \zeta \in \zeta(A)\}\ .$$

Note. The situation with $\zeta(A)$ and $M(A)$ is exactly the same as that for $\eta(A)$ and $M_0(A)$ (see the note following Definition 3.1). If $\mathrm{tr}(A^2) \neq (1/n)(\mathrm{tr}(A))^2$ we use the symbol $N(A)$ to denote the (single) matrix in $M(A)$. In the ambiguous case $\mathrm{tr}(A^2) = (1/n)(\mathrm{tr}(A))^2$ of (3.9) we shall use the notation $N_\theta(A)$ (or $N_\zeta(A)$) to denote the particular matrix in $M(A)$ which corresponds to the element $\zeta = \exp(i\theta)$ of $\zeta(A)$ (here θ is real).

3.11 Theorem. Let $A \in \mathcal{M}_n$ and let ϵ denote the Euclidean norm (1.16) on $\mathcal{M}_n$. Then every matrix in the set $M(A)$ is normal. Furthermore

$$(3.12)\qquad \epsilon^2(A - X) = \tfrac{1}{2}\left(\epsilon^2(A) - \frac{|\mathrm{tr}(A)|^2}{n} - |\mathrm{tr}(A^2) - \frac{(\mathrm{tr}(A))^2}{n}|\right)$$

for all $X \in M(A)$.

Proof. The fact that $M(A) \subset \mathcal{N}_n$ follows immediately from Corollary 1.7l. Let

$$B = A - \frac{\mathrm{tr}(A)}{n} I\ .$$

Then

$$B^2 = A^2 - \frac{2\,\mathrm{tr}(A)}{n} A + \frac{(\mathrm{tr}(A))^2}{n^2} I$$

whence

$$\mathrm{tr}(B^2) = \mathrm{tr}(A^2) - \frac{1}{n} (\mathrm{tr}(A))^2 \quad . \tag{3.13}$$

A comparison of the definitions of the sets (3.2) and (3.9) reveals that

$$\eta(B) = \zeta(A) \quad . \tag{3.13}$$

Furthermore

$$A - [\tfrac{1}{2} (A + \zeta A^*) + \frac{1}{2n} \mathrm{tr}(A - \zeta A^*) I] = \tfrac{1}{2} (A - \frac{\mathrm{tr}(A)}{n} I) - \tfrac{1}{2} \zeta (A^* - \frac{\mathrm{tr}(A^*)}{n} I)$$

$$= \tfrac{1}{2} (B - \zeta B^*) = B - \tfrac{1}{2} (B + \zeta B^*) \quad ;$$

consequently, from (3.5) and (3.14), we obtain

$$\epsilon^2[A - (\tfrac{1}{2} (A + \zeta A^*) + \frac{1}{2n} \mathrm{tr}(A - \zeta A^*) I)] = \tfrac{1}{2} (\epsilon^2(B) - |\mathrm{tr}(B^2)|) \tag{3.15}$$

$$\text{for all } \zeta \in \zeta(A) \quad .$$

Now

$$\epsilon^2(B) = \mathrm{tr}[(A^* - \frac{\mathrm{tr}(A^*)}{n} I)(A - \frac{\mathrm{tr}(A)}{n} I)]$$

$$= \mathrm{tr}[A^*A - \frac{\mathrm{tr}(A)}{n} A^* - \frac{\mathrm{tr}(A^*)}{n} A + \frac{|\mathrm{tr}(A)|^2}{n^2} I] \tag{3.16}$$

$$= \epsilon^2(A) - \frac{|\mathrm{tr}(A)|^2}{n} \quad .$$

Combining (3.13) - (3.16) we obtain (3.12) and this proves Theorem 3.11.

An obvious consequence of Theorem 3.11 is

3.17 Theorem. We have, for all $A \in \mathcal{M}_n$,

$$(3.18) \qquad d_\epsilon^2(A) \leq \frac{1}{2}\left(\epsilon^2(A) - \frac{|\mathrm{tr}(A)|^2}{n} - \left|\mathrm{tr}(A^2) - \frac{(\mathrm{tr}(A))^2}{n}\right|\right) .$$

We shall show that the bound (3.18) is sharper than Mirsky's bound (1.38). We first prove

3.19 Lemma. If $A \in \mathcal{M}_n$ then

$$(3.20) \qquad \epsilon^2(A) - \frac{|\mathrm{tr}(A)|^2}{n} - \left|\mathrm{tr}(A^2) - \frac{(\mathrm{tr}(A))^2}{n}\right| \leq \epsilon^2(A) - |\mathrm{tr}(A^2)| .$$

If $\mathrm{tr}(A) = 0$ we have equality in (3.20) and, if $\mathrm{tr}(A) \neq 0$, equality holds in (3.20) if and only if

$$(3.21) \qquad \frac{n\,\mathrm{tr}(A^2)}{(\mathrm{tr}(A))^2} \geq 1 .$$

Note. The inequality (3.21) is to be interpreted as follows: it is satisfied if and only if the left hand side of (3.21) is both real and greater than or equal to unity.

Proof. The inequality (3.20) and the conditions of equality in it for the case $\mathrm{tr}(A) \neq 0$ are an immediate consequence of the triangle inequality and its conditions of equality (see e.g., [1] pp. 8-9).

While Lemma 3.19 provides, in a certain sense, a complete answer to the questions of equality and inequality of the upper bounds in (1.38) and (3.18), there remain the more interesting questions of what relationship

these bounds have to the actual distance d_ϵ and to the conjectured distance (1.36) in the case $\nu = \epsilon$. We shall discuss the former question in connection with some counterexamples in Chapter 8. The next three results delineate a partial answer to the latter question.

3.22 Lemma. Let $\lambda_1, \lambda_2, \ldots, \lambda_n$ denote complex numbers. Then $|\sum_{k=1}^{n} \lambda_k^2| \leq \sum_{k=1}^{n} |\lambda_k|^2$ with equality holding if and only if all nonzero λ's lie on a single straight line through the origin in the complex plane.

Proof. By the triangle inequality (see [1], p. 9) the desired inequality holds with equality holding if and only if the ratio of the squares of any pair of nonzero λ's is positive:

$$\frac{\lambda_k^2}{\lambda_j^2} = \left(\frac{\lambda_k}{\lambda_j}\right)^2 > 0 \quad , \qquad (\lambda_k \neq 0, \ \lambda_j \neq 0) \quad . \tag{3.23}$$

Obviously (3.23) holds if and only if λ_k/λ_j is real so that, if $\lambda_k = r_k \exp(i\theta_k)$ $(k = 1,2, \ldots, n)$ where $r_k \geq 0$ and θ_k is the principal value of $\arg(\lambda_k)$, $\exp[i(\theta_k - \theta_j)] = \pm 1$ whence either $\theta_k - \theta_j \equiv 0 \pmod{2\pi}$ or $\theta_k - \theta_j \equiv \pi \pmod{2\pi}$ i.e., either $\theta_k = \theta_j$ or $\theta_k = \theta_j \pm \pi$.

3.24 Corollary. Let $A \in \mathcal{M}_n$. We have

$$|\mathrm{tr}(A^2)| \leq \epsilon^2(\Omega(A)) \tag{3.25}$$

with equality holding if and only if all nonzero eigenvalues of A lie on a single straight line through the origin in the complex plane.

Proof. Let $\lambda_1, \ldots, \lambda_n$ denote the eigenvalues of A. Then by Lemma 3.22

$$(3.26) \qquad |\mathrm{tr}(A^2)| = |\sum_{k=1}^{n} \lambda_k^2| \leq \sum_{k=1}^{n} |\lambda_k|^2 = \epsilon^2(\Omega(A))$$

with equality holding as stated in Corollary 3.24.

3.27 Lemma. Let $A \in \mathcal{M}_n$ be given. If, for this particular matrix A, Mirsky's bound in (1.38) is equal to the conjectured distance (1.36) (with $\nu = \epsilon$), i.e., if

$$(3.28) \qquad |\mathrm{tr}(A^2)| = \epsilon^2(\Omega(A)) \quad ,$$

then the bound in (3.18) is also equal to the conjectured distance, i.e.,

$$(3.29) \qquad \frac{|\mathrm{tr}(A)|^2}{n} + |\mathrm{tr}(A^2) - \frac{(\mathrm{tr}(A))^2}{n}| = \epsilon^2(\Omega(A)) \quad ,$$

and

$$(3.30) \qquad \epsilon^2(A - X) = \tfrac{1}{2}(\epsilon^2(A) - \epsilon^2(\Omega(A))) \quad \text{for all } X \in M(A)$$

where M(A) is defined in Definition 3.8.

Proof. By Corollary 3.24 (3.28) holds if and only if the eigenvalues of A can be written in the form $\lambda_k = r_k \exp[i(\theta + p_k)]$ $(k = 1,2, \ldots, n)$ where $r_k \geq 0$, $\exp[ip_k] = \pm 1$ $(k = 1, 2, \ldots, n)$ and θ is a real constant. Thus

$$(3.31) \qquad \mathrm{tr}(A^2) = \sum_{k=1}^{n} \lambda_k^2 = e^{i2\theta} \sum_{k=1}^{n} r_k^2 e^{i2p_k} = e^{i2\theta} \sum_{k=1}^{n} r_k^2 \quad ,$$

$$(3.32) \qquad (\mathrm{tr}(A))^2 = \Big(\sum_{k=1}^{n} \lambda_k\Big)^2 = \Big(e^{i\theta} \sum_{k=1}^{n} r_k e^{ip_k}\Big)^2 = e^{i2\theta} \Big(\sum_{k=1}^{n} r_k e^{ip_k}\Big)^2 \quad .$$

Clearly

$$(3.33) \qquad \Big(\sum_{k=1}^{n} r_k e^{ip_k}\Big)^2 \le \Big(\sum_{k=1}^{n} r_k\Big)^2 \le n \sum_{k=1}^{n} r_k^2 \quad .$$

By Lemma 3.19 and (3.28), (3.29) holds if $\mathrm{tr}(A) = 0$. If $\mathrm{tr}(A) \neq 0$, then at least one r_k is positive and, from (3.31), (3.32), and (3.33), we have

$$\frac{n\ \mathrm{tr}(A^2)}{(\mathrm{tr}(A))^2} = \frac{n \sum r_k^2}{(\sum r_k \exp(ip_k))^2} \ge \frac{n \sum r_k^2}{n \sum r_k^2} = 1 \quad .$$

Therefore, if $\mathrm{tr}(A) \neq 0$, (3.21) holds so that equality holds in (3.20). This proves (3.29). Equation (3.30) then follows immediately from (3.12). This completes the proof of Lemma 3.27.

Remark. The equality (3.29) can hold also in many cases in which (3.28) is not satisfied; furthermore the bound in (3.18) can lie between the bound in (1.38) and the conjectured distance (1.36) (with $\nu = \epsilon$) as the following examples show. Suppose A is of order 4 and has eigenvalues $-1 + i$, $-1 + i$, $-1 + i$, and $2 + 6i$. Then straightforward computations show that

$$\frac{|\mathrm{tr}(A)|^2}{n} + \Big|\mathrm{tr}(A^2) - \frac{(\mathrm{tr}(A))^2}{n}\Big| = \epsilon^2(\Omega(A)) = 46$$

while

$$36.71 < |\mathrm{tr}(A^2)| < 36.72 \quad .$$

Again, if A is of order 4 and has eigenvalues 1, 1, i, and $-i$, then

$$\epsilon^2(\Omega(A)) = 4, \quad \frac{|\mathrm{tr}(A)|^2}{n} + |\mathrm{tr}(A^2) - \frac{(\mathrm{tr}(A))^2}{n}| = 2, \quad |\mathrm{tr}(A^2)| = 0 \quad .$$

We shall show later (Lemma 6.5) that (3.29) and (3.30) hold for all $A \in \mathcal{M}_2$.

3.2 Methods for Obtaining Other Bounds.

Another upper bound for $d_\epsilon(A)$ will be obtained in Chapter 5 (Theorem 5.24) in connection with a maximum problem which is closely related to ϵ-minimal matrices.

We shall now described some problems whose solutions, if they were known, would give rise to upper bounds for $d_\epsilon(A)$.

Let $A \in \mathcal{M}_n$. For each natural number $k \geq 1$ we define the following subsets of $\mathcal{M}_n$:

$$(3.34) \quad \mathcal{L}_k(A) = \{z_0 I + \sum_{\alpha=1}^{k} z_\alpha (A + \zeta A^*)^\alpha;\ \zeta \in \underline{C},\ |\zeta| = 1,\ z_\alpha \in \underline{C} \quad (\alpha = 0, 1, \ldots, k)\} \quad ,$$

$$(3.35) \quad \mathcal{L}_k(A;\zeta) = \{z_0 I + \sum_{\alpha=1}^{k} z_\alpha (A + \zeta A^*)^\alpha;\ z_\alpha \in \underline{C} \quad (\alpha = 0,1, \ldots, k)\}$$

where in (3.35) we require $\zeta \in \underline{C}$ and $|\zeta| = 1$. From Theorems 1.67 and 1.69 we obtain

$$\mathcal{L}_k(A) \subset \mathcal{N}_n \qquad \text{for } A \in \mathcal{M}_n \ , \tag{3.36}$$

$$\mathcal{L}_k(A;\zeta) \subset \mathcal{L}_k(A) \quad \text{for all } |\zeta| = 1 \ . \tag{3.37}$$

Also, obviously,

$$\mathcal{L}_{k-1}(A) \subset \mathcal{L}_k(A),\ \mathcal{L}_{k-1}(A;\zeta) \subset \mathcal{L}_k(A;\zeta) \qquad (k \geq 2) \ . \tag{3.38}$$

From the Cayley-Hamilton Theorem one may deduce

$$\mathcal{L}_k(A) \subset \mathcal{L}_{n-1}(A),\ \mathcal{L}_k(A;\zeta) \subset \mathcal{L}_{n-1}(A;\zeta) \quad \text{where } A \in \mathcal{M}_n, \ k \geq n \ . \tag{3.39}$$

3.40 Problem. Let $A \in \mathcal{M}_n$. Find

$$\operatorname*{Min}_{X \in \mathcal{L}_k(A)} \epsilon^2(A - X) \qquad (1 \leq k \leq n - 1) \tag{3.41}$$

and find all matrices in $\mathcal{L}_k(A)$ for which the minimum in (3.41) is assumed.

A simpler problem is the following

3.42 Problem. Let $A \in \mathcal{M}_n$ and let ζ be a fixed complex number satisfying $|\zeta| = 1$. Find

$$\operatorname*{Min}_{X \in \mathcal{L}_k(A;\zeta)} \epsilon^2(A - X) \qquad (1 \leq k \leq n - 1) \tag{3.43}$$

and find all matrices in $\mathcal{L}_k(A;\zeta)$ for which the minimum in (3.43) is assumed.

By virtue of (3.36) and (3.37) a solution to either of the above problems would yield an upper bound for $d_\epsilon(A)$.

Remark. The author believes that the set $M(A)$ defined in Definition 3.8 provides a complete solution to Problem 3.40 for $k = 1$. No attempt to prove that assertion will be made here.

CHAPTER 4

NECESSARY CONDITIONS FOR ϵ-MINIMAL MATRICES

4.1 Primary Results.

In this section we shall derive a number of conditions which a matrix N_o must satisfy if it is ϵ-minimal. One method which we employ involves the notion of a differentiable curve $N(t)$ in $\mathcal{N}$. If N_o is ν-minimal for A then, by elementary calculus, we know that

$$(4.1) \qquad \frac{d}{dt}\,\nu(A - N(t))\Big|_{t=0} = 0 \quad \text{or} \quad \frac{d}{dt}\,\nu^2(A - N(t))\Big|_{t=0} = 0$$

must hold for <u>every</u> differentiable curve $N(t)$ such that $N(0) = N_o$, provided the derivatives indicated in (4.1) exist in an interval containing $t = 0$ in its interior. This brings up the (possibly difficult) question of how to differentiate an arbitrary (unitarily invariant) norm or some other function whose minima coincide with those of $\nu(A - N(t))$ (such as $\nu^2(A - N(t))$). In the case where ν is the Euclidean norm, the derivatives of all orders can easily be computed. Indeed, from (1.22), we have

$$(4.2) \qquad \begin{aligned} \frac{d}{dt}\,\epsilon^2(A - N(t)) &= \frac{d}{dt}\,\mathrm{tr}[(A - N(t))^*\,(A - N(t))] \\ &= \frac{d}{dt}\,\mathrm{tr}[A^*A + N^*(t)N(t) - A^*N(t) - N^*(t)A] \end{aligned}$$

Using (1.20) and (1.46), we have

$$(4.3)\quad \frac{d}{dt}\,\epsilon^2(A - N(t)) = \frac{d}{dt}\,\mathrm{tr}(N^*(t)N(t)) - \mathrm{tr}\left(A^*\,\frac{dN(t)}{dt} + \frac{dN^*(t)}{dt}\,A\right) .$$

In the proof of the next theorem we shall need the following three elementary lemmas.

4.4 Lemma. Let $N(t) = U^*(t)\,D_0U(t)$ where D_0 is any fixed element of $\mathcal{D}_n$ and $U(t)$ is any differentiable curve in $\mathcal{U}_n$. Then

$$(4.5)\qquad \mathrm{tr}(N^*(t)N(t)) \equiv \epsilon^2(D_0)$$

Proof. This follows immediately from (1.22) and the unitary invariance of ϵ.

4.6 Lemma. (von Neumann [19], p. 290) Let $A \in \mathcal{M}_n$. Then $\mathrm{tr}(AH) = 0$ for all $H \in \mathcal{H}_n$ if and only if $A = 0$.

4.7 Lemma. Let $A = (a_{ij}) \in \mathcal{M}_n$. Then $\mathrm{Re}\ \mathrm{tr}(\Lambda A) = 0$ for all $\Lambda \in \mathcal{D}_n$ if and only if

$$(4.8)\qquad a_{ii} = 0 \qquad (i = 1,2, \ldots , n) .$$

Proof. Letting $\Lambda = \mathrm{diag}(\alpha_1,\alpha_2, \ldots , \alpha_n)$, the sufficiency of (4.8) is obvious from

$$(4.9)\qquad \mathrm{Re}\ \mathrm{tr}(\Lambda A) = \mathrm{Re} \sum_{i=1}^{n} \alpha_i a_{ii} = \sum_{i=1}^{n} \mathrm{Re}(\alpha_i a_{ii}) .$$

For the necessity we set $\Lambda = \mathrm{diag}(\bar{a}_{11},\bar{a}_{22}, \ldots , \bar{a}_{nn})$, obtaining $\mathrm{Re}\ \mathrm{tr}(\Lambda A) = \sum_{i=1}^{n} |a_{ii}|^2 = 0$ which implies (4.8).

4.10 Theorem. Let $A \in \mathcal{M}$ and define, for all $X \in \mathcal{M}$, the operator $L_A(X)$ by

$$L_A(X) = XA^* - A^*X + X^*A - AX^* \quad . \tag{4.11}$$

If $N_o = U_o^* D_o U_o$, where $U_o \in \mathcal{U}$ and $D_o \in \mathcal{D}$, is ϵ-minimal for A then

$$L_A(N_o) = 0 \tag{4.12}$$

and

$$D_o = dg(U_o A U_o^*) \tag{4.13}$$

where the function dg is defined by (1.24).

Proof. We first utilize differentiable curves of the form $N(t) = U^*(t) D_o U(t)$, where $U(t)$ is given by (1.60) and H is any matrix in $\mathcal{H}$. For every $H \in \mathcal{H}$ we have $N(0) = N_o$ and

$$\frac{dU(t)}{dt} = iU_o H e^{itH}, \quad \frac{dU(t)}{dt}\Big|_{t=0} = iU_o H \quad ; \tag{4.14}$$

$$\frac{dN(t)}{dt} = \frac{dU^*(t)}{dt} D_o U(t) + U^*(t) D_o \frac{dU(t)}{dt} \quad ,$$

$$\frac{dN(t)}{dt}\Big|_{t=0} = i(N_o H - H N_o) \quad . \tag{4.15}$$

Using (4.3), Lemma 4.4, (1.48), and (4.15) we obtain

$$\frac{d}{dt}\,\epsilon^2(A - N(t))\Big|_{t=0} = -\,\mathrm{tr}[A^* i(N_o H - H N_o) - i(H N_o^* - N_o^* H)A]$$

$$= i\,\mathrm{tr}(A^* H N_o - A^* N_o H + H N_o^* A - N_o^* H A) \tag{4.16}$$

$$= i\,\mathrm{tr}[(N_o A^* - A^* N_o + N_o^* A - A N_o^*)H]$$

where, in the last step, we used (1.19) twice. According to (4.1) the last expression in (4.16) must vanish for all $H \in \mathcal{H}$. Therefore, by Lemma 4.6, we obtain

$$N_o A^* - A^* N_o + N_o^* A - A N_o^* = 0 \quad , \tag{4.17}$$

i.e., (4.12) holds.

Next let $N(t) = U_o^* D(t) U_o$ where $D(t)$ is a differentiable curve in $\mathcal{D}$ satisfying $D(0) = D_o$. In using (4.1) with this type of curve $N(t)$ we are concerned only with the derivative of $D(t)$ (evaluated at $t = 0$) and not $D(t)$ itself. By Lemma 1.61 we obtain complete generality by setting

$$\frac{dD(t)}{dt}\Big|_{t=0} = \Lambda \tag{4.18}$$

where Λ denotes an arbitrary diagonal matrix. We define N_1 by $N_1 = U_o^* \Lambda U_o$ and note that

$$\frac{dN(t)}{dt}\Big|_{t=0} = U_o^* \frac{dD(t)}{dt} U_o\Big|_{t=0} = U_o^* \Lambda U_o = N_1 \quad . \tag{4.19}$$

Thus, by (1.44),

$$\frac{d}{dt}\,[N^*(t)N(t)]\Big|_{t=0} = N_1^*N_o + N_o^*N_1 \tag{4.20}$$

and, using (4.3), (4.19), and (4.20), we obtain

$$\begin{aligned}\frac{d}{dt}\,\epsilon^2(A - N(t))\Big|_{t=0} &= \mathrm{tr}(N_1^*N_o + N_o^*N_1 - A^*N_1 - N_1^*A)\\ &= \mathrm{tr}(\Lambda^*D_o + D_o^*\Lambda - U_oA^*U_o^*\Lambda - \Lambda^*U_oAU_o^*)\\ &= \mathrm{tr}[\Lambda^*(D_o - U_oAU_o^*) + [\Lambda^*(D_o - U_oAU_o^*)]^*]\\ &= 2\ \mathrm{Re}\ \mathrm{tr}[\Lambda^*(D_o - U_oAU_o^*)]\ .\end{aligned} \tag{4.21}$$

It follows from (4.1) that (4.21) must vanish for all $\Lambda \in \mathcal{D}$; consequently, by Lemma 4.7, all diagonal elements of the matrix $D_o - U_oAU_o^*$ vanish, i.e., (4.13) holds.

Remark. It is instructive to investigate the result of applying (4.1) in the more general situation in which $N(t) = U^*(t)D(t)U(t)$, where $U(t)$ is again given by (1.60) and $D(t)$ is any differentiable curve in $\mathcal{D}$ such that $D(0) = D_o$. The computations can be summarized as follows:

$$\frac{dN(t)}{dt} = U^*(t)\,\frac{dD(t)}{dt}\,U(t) + \frac{dU^*(t)}{dt}\,D(t)U(t) + U^*(t)D(t)\,\frac{dU(t)}{dt}\ , \tag{4.22}$$

$$\frac{dN(t)}{dt}\Big|_{t=0} = N_1 + i(N_oH - HN_o)\ , \tag{4.23}$$

where N_1 is given by (4.19), and

(4.24) $$\frac{d}{dt}\,\epsilon^2(A - N(t))\Big|_{t=0} = 2\,\mathrm{Re}\,\mathrm{tr}[A^*(D_o - U_oAU_o^*)] + i\,\mathrm{tr}[(N_oA^* - A^*N_o + N_o^*A - AN_o^*)H] \quad .$$

Of course the right side of (4.24) must vanish for all $A \in \mathcal{O}$ and for all $H \in \mathcal{H}$; but, from (4.16) and (4.21), we see that this is implied by the separate arguments used in the proof of Theorem 4.10. Therefore nothing new can be obtained by using the more general type of differentiable curves in $\mathcal{M}$.

4.25 Theorem. Let $A \in \mathcal{M}$. If N_o is ϵ-minimal for A then

(4.26) $$\mathrm{tr}(A) = \mathrm{tr}(N_o) \quad ,$$

(4.27) $$\mathrm{tr}[N_o^*(A - N_o)] = \mathrm{tr}[(A - N_o)^*N_o] = 0 \quad .$$

Remark. The reader will note from the following proof that both (4.26) and (4.27) are consequences of (4.13), as are also the following equalities which are equivalent with (4.27):

(4.28) $$\epsilon^2(N_o) = \mathrm{tr}(A^*N_o) = \mathrm{tr}(N_o^*A) \quad ,$$

(4.29) $$d_\epsilon^2(A) = \epsilon^2(A - N_o) = \epsilon^2(A) - \epsilon^2(N_o) \quad .$$

Proof. Since $\mathrm{tr}(D_o) = \mathrm{tr}(N_o)$, (4.26) is an immediate consequence of (4.13). Writing $N_o = U_o^*D_oU_o$ as before, we have

$$\operatorname{tr}[N_o^*(A - N_o)] = \operatorname{tr}[U_o N_o^* U_o^* U_o (A - N_o) U_o^*]$$

(4.30)

$$= \operatorname{tr}[D_o^*(U_o A U_o^* - D_o)] \quad .$$

By (4.13) all diagonal elements of the matrix $U_o A U_o^* - D_o$ are zero, consequently the same holds for the matrix $D_o^*(U_o A U_o^* - D_o)$ hence $\operatorname{tr}[D_o^*(U_o A U_o^* - D_o)] = 0$. Thus, from (4.30), we have

(4.31) $$\operatorname{tr}[N_o^*(A - N_o)] = 0$$

The last equality implies that $\operatorname{tr}[N_o^*(A - N_o)]$ is real; consequently the rest of (4.27) follows from (1.21).

Remark. Let $A, B \in \mathcal{M}_n$. The set $\mathcal{M}_n$ can be considered as a complex Hilbert space with respect to the inner product

(4.32) $$(A,B) = \operatorname{tr}(AB^*) \quad .$$

Comparing (4.32) with (4.27) we see that, if N_o is ϵ-minimal for A,

(4.33) $$(A - N_o, N_o) = 0 \quad ;$$

that is, N_o is orthogonal to $A - N_o$. This geometrical reformulation of the necessary condition (4.27) is not too surprising. Indeed, if N_o is ϵ-minimal, then it must also be ϵ-minimal among all members of the one-dimensional linear subspace (of $\mathcal{M}_n$):

$$\mathcal{L} = \{X \; ; \; X = zN_o \text{ where } z \in \underline{C}\}$$

which is contained in $\mathcal{N}_n$. $\mathcal{L}$ is precisely the "straight line" through N_o and the null matrix. Since no point on $\mathcal{L}$ can be closer to A than N_o, we see that N_o must be the point of intersection of $\mathcal{L}$ and the line through A which is perpendicular to $\mathcal{L}$. This gives immediately the result (4.33) so we have proved (4.27) a second time without using differentiable curves and calculus. One can also prove (4.27) directly (without using (4.13)) by using (4.1) and the particular differentiable curves $N(t) = (1 - t)N_o$ and $N(t) = (1 - it)N_o$. The latter proof, which will not be worked out in detail here, appears to be essentially an analytic reformulation of the geometric considerations just mentioned.

We shall prove later (Theorem 5.13) that (4.13) and another condition on U_o constitute necessary and sufficient conditions for ϵ-minimality. It appears that (4.13) is a much more stringent condition than (4.12); however, the latter is much easier to check than (4.13) and this fact enhances its value. We now describe some consequences of the necessary condition (4.12).

4.34 Theorem. Let $A \in \mathcal{M}$, $N_o \in \mathcal{M}$. Then the following statements are equivalent.

(a) $L_A(N_o) = 0$;

(b) $N_o A^* + N_o^* A \in \mathcal{H}$;

(c) $(N_o - \alpha A)A^* + (N_o - \beta A)^* A \in \mathcal{H}$ for any $\alpha, \beta \in \underline{R}$;

(d) $N_o(A - \alpha N_o)^* + N_o^*(A - \beta N_o) \in \mathcal{H}$ for any $\alpha, \beta \in \underline{R}$;

(e) $N_o(A - zN_o)^* + N_o^*(A - wN_o) = [N_o(A - wN_o)^* + N_o^*(A - zN_o)]^*$ holds for any $z, w \in \underline{C}$ provided $N_o \in \mathcal{N}$.

Proof. Statement (a) is equivalent to (4.17) and we shall show that (b), (c), (d), and (e) are equivalent to (4.17). First note that (4.17) can be rewritten in the form

$$N_0A^* + N_0^*A = AN_0^* + A^*N_0 \quad .$$

The right side of the last equation obviously equals $[N_0A^* + N_0^*A]^*$ which shows that (b) is equivalent to (4.17). For any real numbers α and β we note that

$$N_0A^* - \alpha N_0N_0^* - A^*N_0 + \beta N_0^*N_0 + N_0^*A - \beta N_0^*N_0 - AN_0 + \alpha N_0N_0^* = 0 \quad ,$$

$$N_0A^* - \alpha AA^* - A^*N_0 + \beta A^*A + N_0^*A - \beta A^*A - AN_0 + \alpha AA^* = 0$$

are equivalent to (4.17). Clearly the last two equations are equivalent respectively to

$$N_0(A - \alpha N_0)^* - (A - \beta N_0)^*N_0 + N_0^*(A - \beta N_0) - (A - \alpha N_0)N_0^* = 0 \quad ,$$

$$(N_0 - \alpha A)A^* - A^*(N_0 - \beta A) + (N_0 - \beta A)^*A - A(N_0 - \alpha A)^* = 0$$

which are equivalent respectively to statements (c) and (d). If we assume that N_0 is normal and let z and w be any complex numbers, we obtain from (4.17)

$$N_0A^* - \bar{z}N_0N_0^* - A^*N_0 + \bar{z}N_0^*N_0 + N_0^*A - wN_0^*N_0 - AN_0^* + wN_0N_0^* = 0$$

or

$$N_0(A - zN_0)^* - (A - zN_0)^*N_0 + N_0^*(A - wN_0) - (A - wN_0)N_0^* = 0$$

which is equivalent to (e). This proves Theorem 4.34.

In terms of the operator (4.11) we observe the following rather obvious consequences of Theorem 4.34. If $A, N_o \in \mathcal{M}$, then the following statements are equivalent to (4.12):

(4.35) $\quad L_A(N_o - \alpha A) = 0 \qquad$ for any $\alpha \in \underline{R}$,

(4.36) $\quad L_{A-\alpha N_o}(N_o) = 0 \qquad$ for any $\alpha \in \underline{R}$,

(4.37) $\quad L_{A-zN_o}(N_o) = 0 \qquad$ for any $z \in \underline{C}$ provided $N_o \in \mathcal{M}$.

4.38 Lemma. Let $A \in \mathcal{M}$ and suppose the operator $L_A(X)$ is defined by (4.11). Then, for all $X, Y \in \mathcal{M}$, we have

(4.39) $\quad L_A(\alpha X) = \alpha L_A(X) \qquad$ for all $\alpha \in \underline{R}$,

(4.40) $\quad L_A(X + Y) = L_A(X) + L_A(Y)$.

Proof. For any $\alpha \in \underline{C}$ we have

(4.41) $\quad L_A(\alpha X) = \alpha(XA^* - A^*X) + \bar{\alpha}(X^*A - AX^*)$

whence (4.39) holds if α is real. The verification of (4.40) is equally simple and need not be given here.

Remark. Equations (4.39) and (4.40) show that, given any $A \in \mathcal{M}$,

$$\{X \; ; \; X \in \mathcal{M} \text{ and } L_A(X) = 0\}$$

is a real linear subspace of $\mathcal{M}$, i.e., a linear subspace of $\mathcal{M}$ over the field $\underline{R}$ of real scalars.

Page Intentionally Left Blank

BLANK PAGE

The result (4.39) cannot in general be extended to all complex scalars α as the following simple example shows. Suppose $A \notin \mathcal{N}$. Then from (4.11) and (4.41) we find that $L_A(A) = 0$ but $L_A(iA) = 2i(AA^* - A^*A) \neq 0$. Of course, in certain special situations, (4.39) can hold for all $\alpha \in \underline{C}$. Examples of the latter can be deduced from the following easily verified results which are valid for any $A \in \mathcal{M}_n$:

(4.42) $$L_A(zI) = 0 \qquad \text{for all } z \in \underline{C} \ ;$$

(4.43) $$L_A(\alpha A) = 0 \qquad \text{for all } \alpha \in \underline{R} \ ;$$

(4.44) $$L_A(z(A^*)^k) = 0 \qquad \text{for all } z \in \underline{C} \text{ and for } k = 1, 2, \ldots \ ;$$

(4.45) $$L_A(zA^2 + \bar{z}(AA^* + A^*A)) = 0 \qquad \text{for all } z \in \underline{C} \ ;$$

(4.46) $$L_A(\alpha A + zA^*) = 0 \qquad \text{for all } \alpha \in \underline{R}, \ z \in \underline{C} \ ;$$

(4.47) $$L_A[z(A + \tfrac{\bar{z}}{z} A^*)^2] = 0 \qquad \text{for all } z \in \underline{C} \text{ with } z \neq 0 \ .$$

Note. The content of the next lemma is that every matrix in the set M(A) defined in Definition 3.8 satisfies (for any value of n) all of the necessary conditions given in Theorems 4.10 and 4.25 with the possible exception of (4.13).

Remark. While condition (4.13) is apparently the most crucial test of a candidate for ϵ-minimality of all the necessary conditions derived so far, it cannot be checked unless the normal matrix in question is first diagonalized by a unitary transformation.

4.48 Lemma. Let $A \in \mathcal{M}_n$ and let $M(A)$ be defined as in Definition 3.8. Then the necessary conditions (4.12), (4.26), and (4.27) are satisfied when N_o is replaced by any matrix in $M(A)$.

Proof. Let

$$P_\zeta(A) = \tfrac{1}{2}(A + \zeta A^*) + \tfrac{1}{2n}\operatorname{tr}(A - \zeta A^*)I \quad . \tag{4.49}$$

The fact that $P_\zeta(A)$ satisfies (4.12) for all $\zeta \in \underline{C}$ follows immediately from (4.40), (4.42), and (4.46). This implies that every matrix in $M(A)$ satisfies (4.12). Furthermore

$$\operatorname{tr}[P_\zeta(A)] = \tfrac{1}{2}\operatorname{tr}(A) + \tfrac{1}{2}\zeta\operatorname{tr}(A^*) + \tfrac{1}{2}\operatorname{tr}(A - \zeta A^*) = \operatorname{tr}(A)$$

for all $\zeta \in \underline{C}$ hence every matrix in $M(A)$ satisfies (4.26). Let

$$z = \tfrac{1}{n}\operatorname{tr}(A - \zeta A^*) \quad . \tag{4.50}$$

Then $\bar{z} = (1/n)\operatorname{tr}(A^* - \bar{\zeta}A)$, $P_\zeta(A) = (1/2)(A + \zeta A^* + zI)$, and $A - P_\zeta(A) = (1/2)(A - \zeta A^* - zI)$. Consequently, for $|\zeta| = 1$,

$$P_\zeta^*(A)(A - P_\zeta(A)) = \tfrac{1}{4}(A^* + \bar{\zeta}A + \bar{z}I)(A - \zeta A^* - zI)$$

$$= \tfrac{1}{4}(A^*A - \zeta A^{*2} + \bar{\zeta}A^2 - AA^* - z(A^* + \bar{\zeta}A) + \bar{z}(A - \zeta A^*) - z\bar{z}I)$$

so that

$$(4.51)\qquad 4\,\mathrm{tr}[P_\zeta^*(A)(A - P_\zeta(A))] = \bar{\zeta}\,\mathrm{tr}(A^2) - \zeta\,\mathrm{tr}(A^{*2}) - z\,\mathrm{tr}(A^* + \bar{\zeta}A) + \bar{z}\,\mathrm{tr}(A - \zeta A^*) - nz\bar{z}\,, \quad (|\zeta| = 1)\ .$$

Using (4.50) in the last term of equation (4.51) we see that the last two terms on the right side of (4.51) cancel. Using (4.50) again we find, for $|\zeta| = 1$,

$$\begin{aligned} z\,\mathrm{tr}(A^* + \bar{\zeta}A) &= \tfrac{1}{n}\,[\mathrm{tr}(A) - \zeta\,\mathrm{tr}(A^*)][\mathrm{tr}(A^*) + \bar{\zeta}\,\mathrm{tr}(A)] \\ &= \tfrac{1}{n}\,[|\mathrm{tr}(A)|^2 + \bar{\zeta}(\mathrm{tr}(A))^2 - \zeta(\mathrm{tr}(A^*))^2 - |\mathrm{tr}(A)|^2] \qquad (4.52) \\ &= \tfrac{1}{n}\,[\bar{\zeta}(\mathrm{tr}(A))^2 - \zeta(\mathrm{tr}(A^*))^2]\ . \end{aligned}$$

Combining (4.51) and (4.52) we obtain

$$(4.53)\qquad 4\,\mathrm{tr}[P_\zeta^*(A)(A - P_\zeta(A))] = \bar{\zeta}(\mathrm{tr}(A^2) - \tfrac{1}{n}(\mathrm{tr}(A))^2) - \overline{\zeta(\mathrm{tr}(A^2) - \tfrac{1}{n}(\mathrm{tr}(A))^2)}\ .$$

If $\mathrm{tr}(A^2) = (1/n)(\mathrm{tr}(A))^2$ then the right side of (4.53) vanishes for all $\zeta \in \zeta(A)$. On the other hand, if $\mathrm{tr}(A^2) \neq (1/n)(\mathrm{tr}(A))^2$ then by (3.9) the right side of (4.53) becomes

$$|\mathrm{tr}(A^2) - \tfrac{1}{n}(\mathrm{tr}(A))^2| - |\mathrm{tr}(A^2) - \tfrac{1}{n}(\mathrm{tr}(A))^2| = 0\ .$$

Thus every matrix in $M(A)$ satisfies (4.27) and this proves Lemma 4.48.

In the light of Lemma 4.48, it is natural to ask whether the matrices in the set $M(A)$ can ever be ϵ-minimal. We shall answer this question in the affirmative in Section 6.2 below.

4.2 Use of Lagrange Multipliers in Finding Necessary Conditions.

As was observed in Section 1.5, the problem of finding a ν-minimal matrix can be considered as a problem of minimizing a real valued function of several complex variables. The minimum must be taken over normal matrices only, so certain constraints on the variables are inevitable. This point of view suggests that the method of Lagrange multipliers might be useful in deriving necessary conditions for a ν-minimal matrix. Since the variables and constraints are complex, it first appears that the initial computational labor of deriving such conditions would be prohibitive. However, by using conjugate complex coordinates (see e.g., [18] pp. 16-21) and the complex (matrix) differential calculus, the algebraic manipulations can be made almost inconsequential. In the next two paragraphs we outline briefly the mathematical basis of the technique we shall use in deriving necessary conditions for ϵ-minimal matrices. The equations which will be derived appear to be very difficult to solve in general, primarily because they involve a large number of auxiliary unknown variables (Lagrange multipliers). As a consequence these equations are not of immediate value in finding ϵ-minimal matrices. Nevertheless, the equations themselves and the technique of deriving

them seem interesting. Furthermore as byproducts we obtain alternative derivations of the primary necessary conditions (4.12), (4.13) and also additional necessary conditions for ϵ-minimal matrices involving theoretically interesting interpretations of certain Lagrange multipliers.

Let x and y denote a pair of real variables (coordinates). Following Nehari [18] we define $z = x + iy$ and $\bar{z} = x - iy$ to be the corresponding pair of conjugate complex variables (coordinates). We define formally a pair of "partial differential operators" by the expressions

$$\frac{\partial}{\partial z} = \frac{1}{2}\left(\frac{\partial}{\partial x} - i\frac{\partial}{\partial y}\right) , \tag{4.54}$$

$$\frac{\partial}{\partial \bar{z}} = \frac{1}{2}\left(\frac{\partial}{\partial x} + i\frac{\partial}{\partial y}\right) . \tag{4.55}$$

As shown by Nehari, these operations can be carried out, at least for a wide class of functions of x and y (or equivalently of z and $\bar{z}$), by treating z and $\bar{z}$ as independent variables and differentiating by the usual rules of calculus. As a simple illustration of the method, suppose $g(x,y) = f(z,\bar{z})$ is a real valued function for which the stationary points are desired. The standard method of elementary calculus involves solving the simultaneous pair of equations

$$\frac{\partial g}{\partial x} = 0 , \qquad \frac{\partial g}{\partial y} = 0 ; \tag{4.56}$$

whereas in the conjugate variables method one merely works with the single (complex) equation

(4.57) $$\frac{\partial f}{\partial \bar{z}} = 0$$

which is equivalent to the pair of equations (4.56). It is obvious how one may extend this procedure to the problem of finding stationary points of real valued functions of $2k$ real variables where $k = 2,3, \ldots$.

An alternative technique, which is often easier to apply in matrix problems, is based on the use of differentials. Nehari showed that the operators (4.54) and (4.55) behave like genuine partial derivatives in the sense that the differential of a real function $g(x,y) = f(z,\bar{z})$ is given by

(4.58) $$dg = df = \frac{\partial f}{\partial z}\, dz + \frac{\partial f}{\partial \bar{z}}\, d\bar{z}$$

where $dz = dx + i\, dy$ and $d\bar{z} = dx - i\, dy$. In elementary calculus the equivalent formulation of (4.56) in terms of differentials is the statement that $dg = (\partial g/\partial x)dx + (\partial g/\partial y)dy$ vanishes for all values of the independent differentials dx and dy. This does not correspond exactly to the statement that (4.58) vanishes for all dz and $d\bar{z}$ considered as independent differentials. For, from that statement, one would obtain both $\partial f/\partial z = 0$ and $\partial f/\partial \bar{z} = 0$ which would imply (4.56) twice since $\partial f/\partial z = 0$ if and only if $\partial f/\partial \bar{z} = 0$. It is sufficient to compute the complex "half differential"

$$(dg)_{\bar{z}} = (df)_{\bar{z}} = \frac{\partial f}{\partial \bar{z}}\, d\bar{z}$$

and say that it must vanish for all $d\bar{z}$. In the case of functions of more than two real variables, this rule for deriving necessary conditions can be generalized as follows.

4.59 Theorem. Let $z = (z_1, z_2, \ldots, z_k)$ denote a vector with complex components and suppose $\phi(z,\bar{z}) \equiv \phi(z_1, z_2, \ldots, z_k, \bar{z}_1, \bar{z}_2, \ldots, \bar{z}_k)$ is a real valued differentiable function of the $2k$ real variables x_α, y_α $(\alpha = 1, \ldots, k)$ where $z_\alpha = x_\alpha + i\, y_\alpha$ $(\alpha = 1, \ldots, k)$. Let the half differential $(d\phi)_{\underset{z}{}}$ be defined by

$$(4.60) \qquad (d\phi)_{\underset{z}{}}(z,\bar{z}) \equiv \frac{\partial\phi(z,\bar{z})}{\partial\bar{z}_1}\, d\bar{z}_1 + \frac{\partial\phi(z,\bar{z})}{\partial\bar{z}_2}\, d\bar{z}_2 + \cdots + \frac{\partial\phi(z,\bar{z})}{\partial\bar{z}_k}\, d\bar{z}_k \ .$$

If ϕ has an extreme value at the point $z_o = (z_1^{(o)}, \ldots, z_k^{(k)})$, then $(d\phi)_{\underset{z}{}}(z_o,\bar{z}_o)$ must vanish for all values of the independent differentials $d\bar{z}_1, d\bar{z}_2, \ldots, d\bar{z}_k$.

We refrain from giving a proof of Theorem 4.59.

Let $A = (a_{ij}) \in \mathcal{M}_n$ and $N = (n_{ij}) \in \mathcal{M}_n$. We first consider the straightforward problem of minimizing $\epsilon^2(N - A)$ subject to the constraint

$$(4.61) \qquad N^*N - NN^* = 0 \ .$$

Since the matrix $N^*N - NN^*$ is hermitian, the matrix equation (4.61) amounts to: (1) at most $(1/2)n(n-1)$ independent complex constraints on the n_{ij} (these correspond to all elements of (4.61) which lie on one side of the diagonal), and (2) at most n independent real constraints (these correspond to the n (real) diagonal elements of (4.61)). Thus

(4.61) involves at most n^2 real constraints on the n^2 complex quantities n_{ij}. Clearly then, we need at most n^2 real Lagrange multipliers. Nevertheless, it is convenient in deriving the necessary conditions to introduce at the beginning $2n^2$ real Lagrange multipliers, one for each of the real constraints

$$(4.62) \qquad \mathrm{Re}[(N^*N - NN^*)_{ij}] = 0 \ , \qquad \mathrm{Im}[(N^*N - NN^*)_{ij}] = 0$$

where $1 \leq i \leq n$, $1 \leq j \leq n$. The introduction of superfluous Lagrange multipliers causes no difficulty since the superfluous ones either drop out or combine automatically with others to form the maximal number of independent multipliers. We let λ_{ij}, μ_{ij} be real Lagrange multipliers which correspond (in that order) to the constraints (4.62). We now define the scalar quantities

$$(4.63) \qquad \tau_{ij} = \lambda_{ij} + i\,\mu_{ij} \qquad (1 \leq i, j \leq n)$$

and call the τ_{ij} <u>complex Lagrange multipliers</u>. For later use we also define the matrix

$$(4.64) \qquad T = (\tau_{ij}) \ .$$

In the classical way of applying the method of Lagrange multipliers one would consider the function

$$(4.65) \quad \epsilon^2(N - A) + \sum_{i,j=1}^{n} [\lambda_{ij}\,\mathrm{Re}[(N^*N-NN^*)_{ij}] + \mu_{ij}\,\mathrm{Im}[(N^*N-NN^*)_{ij}]]$$

and differentiate it partially with respect to the real variables $\mathrm{Re}(n_{rc})$, $\mathrm{Im}(n_{rc})$ $(1 \leq r, c \leq n)$. As was suggested above, this procedure is unnecessarily complicated and can be greatly simplified and shortened by using complex analysis.

Note that

$$\lambda_{ij}\,\mathrm{Re}[(N^*N - NN^*)_{ij}] + \mu_{ij}\,\mathrm{Im}[(N^*N - NN^*)_{ij}] = \mathrm{Re}[\bar{\tau}_{ij}(N^*N - NN^*)_{ij}]$$

so that the sum on the right side of (4.65) can be written in the form

$$\mathrm{Re}\Big\{\sum_{i,j=1}^{n} \bar{\tau}_{ij}(N^*N - NN^*)_{ij}\Big\} = \mathrm{Re}\ \mathrm{tr}[(N^*N - NN^*)T^*]$$

$$= \frac{1}{2}\left(\mathrm{tr}[(N^*N - NN^*)T^*] + \overline{\mathrm{tr}[(N^*N - NN^*)T^*]}\right) \tag{4.66}$$

$$= \frac{1}{2}\,\mathrm{tr}[(N^*N - NN^*)(T + T^*)]$$

where T is given by (4.64). In view of the form of the constraint equations (4.61) the factor (1/2) in (4.66) can be omitted upon substitution of (4.66) into (4.65). Thus, writing $\epsilon^2(N - A) = \mathrm{tr}[(N - A)^*(N - A)]$, we consider the problem of minimizing the function

$$\Phi(N,N^*) = \mathrm{tr}[(N^* - A^*)(N - A)] + \mathrm{tr}[(N^*N - NN^*)H] \tag{4.67}$$

where H is given by

$$H = T + T^* \ . \tag{4.68}$$

Equation (4.67) shows that, as far as the Lagrange multipliers (4.63) are concerned, it is the matrix H which has a definite meaning (i.e., which has to be determined using (4.61)) and not T. Clearly the Lagrange multipliers μ_{ii} $(i = 1, \ldots, n)$ which correspond to nonexistent constraints in (4.61) have dropped out and the others have combined to form the n^2 "actual Lagrange multipliers":

$$2\lambda_{ii} \qquad (i = 1, \ldots, n) \ ,$$

$$\lambda_{ij} + \lambda_{ji} \ , \ \mu_{ij} - \mu_{ji} \qquad (1 \leq i < j \leq n) \ .$$

Differentiating (4.67) we obtain

$$\begin{aligned} (d\Phi)_{N^*} &= \mathrm{tr}[dN^*(N - A) + (dN^*N - NdN^*)H] \\ &= \mathrm{tr}[(N - A + NH - HN)dN^*] \ . \end{aligned} \tag{4.69}$$

By Theorem 4.59 Φ has a minimum at the point (matrix) N only if $(d\Phi)_{N^*}$ vanishes for all matrices dN^*. It is easy to show from von Neumann's Lemma 4.6 that (4.69) vanishes for all $dN^* \in \mathcal{M}_n$ if and only if

$$N - A + NH - HN = 0 \ .$$

If we rewrite the last equation and the constraint (4.61) and note that (4.68) implies $H \in \mathcal{H}_n$, we obtain the following set of three simultaneous matrix equations for the unknown matrices N and H.

$$N + NH - HN = A \ , \tag{4.70}$$

(4.71) $$H = H^* \ ,$$

(4.72) $$N^*N - NN^* = 0 \ .$$

4.73 Theorem. Let $A \in \mathcal{M}_n$. If N_o is ϵ-minimal for A then there exists a matrix $H_o \in \mathcal{M}_n$ such that (4.70), (4.71), and (4.72) are satisfied when N and H are replaced respectively by N_o and H_o.

Proof. If N_o is ϵ-minimal it minimizes the functional $\epsilon^2(N - A)$ and satisfies (4.72). Consequently N_o is a stationary point of $\epsilon^2(N - A)$ which implies the existence of complex Lagrange multipliers τ_{ij} such that N_o is also a stationary point of (4.67). By the above derivation of equation (4.70) there exists a hermitian matrix H_o such that $N_o + N_oH_o - H_oN_o = A$. This proves Theorem 4.73.

Remark. It should be noted that the hermitian matrix H of Lagrange multipliers need not necessarily be of the form of the most general hermitian matrix. For example in the case $n = 2$ the diagonal elements of the left side of (4.72) differ only in sign (cf. Lemma 6.1 below) so one additional real Lagrange multiplier can be eliminated, allowing H to assume the form

$$H = \begin{pmatrix} \lambda_1 & \lambda + i\mu \\ \lambda - i\mu & -\lambda_1 \end{pmatrix}$$

where λ_1, λ, and μ are real. It should also be noted, however, that H may be assumed to be a general hermitian matrix. There is no harm in repeating constraints in the method of Lagrange multipliers.

Since at least one solution of the necessary condition described in Theorem 4.73 is ϵ-minimal, it is natural to investigate the problem of solving the system (4.70) - (4.72). In theory one could carry out the following procedure. Assuming that H is hermitian, solve (4.70) for N as a function of A and H, i.e., find

$$N = F(A,H) \quad . \tag{4.74}$$

Then substitute (4.74) into (4.72) and solve the latter equation for the hermitian matrix H. Once H was determined it would be eliminated by substitution into (4.74) which would yield a normal matrix.

Actually, it is possible to solve the system (4.70) - (4.72) formally, although not explicitly, by a method which runs somewhat along the lines of the procedure outlined in the previous paragraph. Since the matrix H of Lagrange multipliers is hermitian it can be diagonalized by a unitary transformation:

$$UHU^* = D \qquad (U \in \mathcal{U}_n, \ D \in \mathcal{D}_n) \tag{4.75}$$

where D is real. By means of (4.75) we replace the unknown hermitian matrix H by two other matrices, namely an unknown unitary matrix U and an unknown diagonal matrix D. Now let

$$UNU^* = M = (m_{ij}) \ , \qquad UAU^* = B = (b_{ij}) \tag{4.76}$$

where U is the same matrix which occurs in (4.75). By premultiplication by U and postmultiplication by U^* we see that the equations (4.70) and (4.72) are equivalent respectively to

(4.77) $$M + MD - DM = B$$

and

(4.78) $$M^*M - MM^* = 0 \quad .$$

Note that N is normal if and only if M is normal.

4.79 Definition. Let the matrix (4.75) be given by $D = \mathrm{diag}(d_1, d_2, \ldots, d_n)$, let $\Delta_{ij} = d_j - d_i$ $(1 \leq i, j \leq n)$, and let $\Delta = \{\Delta_{ij};\ 1 \leq i < j \leq n\}$. Then $M(A,U,\Delta)$ is defined to be the matrix

(4.80) $$M(A,U,\Delta) = \left(\frac{b_{ij}}{1 + \Delta_{ij}}\right)$$

where $B = (b_{ij})$ is defined by (4.76).

Using the notation of (4.76) and the preceding definition we see that (4.77) (which can be rewritten in the form $M(I + D) - DM = B$) amounts to a set of scalar equations:

$$m_{ij}(1 + d_j) - d_i\, m_{ij} = b_{ij}$$

or

(4.81) $$m_{ij}(1 + d_j - d_i) = m_{ij}(1 + \Delta_{ij}) = b_{ij} \qquad (1 \leq i, j \leq n) \quad .$$

If we choose the d_i's in such a way that $1 + d_j - d_i = 1 + \Delta_{ij} \neq 0$ for all pairs (i,j) then we can solve (4.81) for the m_{ij}. From (4.80) we obtain $M = M(A,U,\Delta)$, a known function of A, U, and Δ. The set Δ is determined by D so the matrix (4.90) is actually a function of A, U, and D. Thus, if we choose U and D so as to make the matrix (4.80) normal then

BLANK PAGE

(4.88) $$T_{U_o,\Delta_o}(A) \in \mathcal{N}_n$$

and

(4.89) $$\epsilon^2(U_oAU_o^* - T_{U_o,\Delta_o}(A)) = \text{Min } \epsilon^2(UAU^* - T_{U,\Delta}(A))$$

where the minimum is taken over all U, D such that $T_{U,\Delta}(A) \in \mathcal{N}_n$ and where Δ and Δ_o are determined respectively by D and D_o as in Definition 4.79.

In order to examine this problem we note that the transformation (4.86) is the composition $T_\Delta \cdot T_{U^*}$ where the transformation on the right is to be carried out first, where T_U is defined by (1.75), and where

(4.90) $$T_\Delta(M) = \left(\frac{m_{ij}}{1 + \Delta_{ij}}\right) \quad , \quad (M = (m_{ij}) \in \mathcal{N}_n \, , \, \Delta_{ij} \neq -1) \quad .$$

Thus $T_{U,\Delta}(A) = T_\Delta(T_{U^*}(A))$. If A is not normal, then neither is $B = T_{U^*}(A)$; hence the second transformation T_Δ plays an indispensable role in achieving the normality of $T_{U,\Delta}(A)$. On the other hand, it can be shown by examples that $T_{I,\Delta}(A) \notin \mathcal{N}_n$ can hold for all Δ's which arise from real diagonal matrices D. Thus we have gained some insight into the meaning and role of the matrix H of Lagrange multipliers: The <u>possibility</u> that

(4.91) $$T_{U,\Delta}(A) \in \mathcal{N}_n$$

for some Δ depends on the choice of U; the actual realization of (4.91) depends on the choice of Δ and not on U. If we translate back into the coordinate system of the set of equations (4.70) - (4.72), we can summarize those observations in the following manner.

4.92 Theorem. Let $A \in \mathcal{M}_n$, $A \notin \mathcal{N}_n$ and let $H \in \mathcal{H}_n$. Then the possibility that (4.70) has a normal solution N (for some choice of the eigenvalues of H) depends on the eigenvectors of H. The choice of the eigenvalues (and not the eigenvectors) of H determines whether or not (4.70) has a normal solution N.

Proof. We need only observe from (4.75) that the eigenvectors of H are the columns of U^* and that Δ is determined by the eigenvalues d_i of H.

Instead of Problem 4.87 we may consider the simpler

4.93 Problem. Let $A \in \mathcal{M}_n$, $A \notin \mathcal{N}_n$ and define

$$\mathcal{S}(A) = \{U ; U \in \mathcal{U}_n \text{ and } T_{U,\Delta}(A) \in \mathcal{N}_n \text{ for some admissible } \Delta\} .$$

Find a $U \in \mathcal{S}(A)$. Then, for this U, find a $D_o \in \mathcal{D}_n$ such that

$$T_{U,\Delta_o}(A) \in \mathcal{N}_n$$

and

$$\epsilon^2(UAU^* - T_{U,\Delta_o}(A)) = \text{Min } \epsilon^2(UAU^* - T_{U,\Delta}(A))$$

where the minimum is taken over all D such that $T_{U,\Delta}(A) \in \mathcal{N}_n$ and where Δ and Δ_o are determined respectively by D and D_o as in Definition 4.79.

The explicit determination of solutions to either Problem 4.87 or Problem 4.93 appears to be difficult to carry out for $n \geq 3$. The author has found solutions to Problem 4.93 for all $A \in \mathcal{M}_2$ and, on the basis of the results in Chapter 6 below, these have all been ϵ-minimal. No further investigation of Problems 4.87 or 4.93 will be made here except for the following simple observation.

4.94 Lemma. Let $U \in \mathcal{U}_n$ and let $\Delta = [\Delta_{ij}]$ $(1 \leq i, j \leq n)$ where $\Delta_{ij} \neq -1$ for all i and j. Then the transformation $T_{U,\Delta}(A)$ defined by (4.86) is a linear transformation on $\mathcal{M}_n$.

Proof. The transformations T_U of (1.75) and T_Δ of (4.90) are both obviously linear. Thus $T_{U,\Delta}$, being a composition of linear transformations, is linear too.

Note. We shall find another application for equation (4.70) in Chapter 7 below.

We consider next another approach to the use of Lagrange multipliers. By Theorem 1.57 (c) $N = U^*DU$ runs through $\mathcal{M}_n$ as U and D run independently through $\mathcal{U}_n$ and $\mathcal{D}_n$ respectively. Thus we can find an ϵ-minimal matrix for A by minimizing

$$\epsilon^2(U^*DU - A) = \mathrm{tr}[(U^*D^*U - A^*)(U^*DU - A)] \tag{4.95}$$

subject to the constraints

$$U^*U = I \ , \tag{4.96}$$

$$D \in \mathcal{D}_n \ . \tag{4.97}$$

As in the case of the constraint (4.61) we introduce complex Lagrange multipliers τ_{ij} corresponding to the (i,j) elements $(U^*U - I)_{ij}$ of the constraint equation (4.96) and write $T = (\tau_{ij})$ as before. If we merely keep in mind that D is diagonal, we need not introduce any Lagrange multipliers for the constraint (4.97). We have

$$2 \operatorname{Re} \operatorname{tr}(U^*UT^*) = \operatorname{tr}[U^*U(T + T^*)] \quad .$$

Here, as in the previous constrained minimization problem, it is not T which has to be determined but the hermitian matrix $H = T + T^*$. Therefore the function we wish to minimize in the method of Lagrange multipliers is

$$(4.98) \qquad \Phi(U,U^*,D,D^*) = \operatorname{tr}[(U^*D^*U - A^*)(U^*DU - A) + U^*UH] \quad .$$

Differentiating this we obtain

$$\begin{aligned}(d\Phi)_{U^*,D^*} &= \operatorname{tr}[dU^*D^*U(U^*DU - A) + (U^*D^*U - A)dU^*DU + dU^*UH] \\ &\quad + \operatorname{tr}[U^*dD^*U(U^*DU - A)] \\ &= \operatorname{tr}[[D^*U(U^*DU - A) + DU(U^*D^*U - A^*) + UH]dU^*] \\ &\quad + \operatorname{tr}[dD^*U(U^*DU - A)U^*] \quad .\end{aligned}$$

By Theorem 4.59 Φ has a minimum at the point U^*DU only if $(d\Phi)_{U^*,D^*}$ vanishes for all $dU^* \in \mathcal{M}_n$ and $dD^* \in \mathcal{D}_n$. Setting $dU^* = 0$, varying dD^* in $\mathcal{D}_n$, and using Lemma 4.7 we obtain (4.99) below.

Setting $dD^* = 0$ and varying dU^* we obtain (4.100) below. If we write down all the relevant constraints with these equations we obtain the following complete set of five simultaneous matrix equations for the unknown matrices U, D, and H.

$$dg[U(U^*DU - A)U^*] = 0 \quad , \tag{4.99}$$

$$D^*U(U^*DU - A) + DU(U^*DU - A)^* + UH = 0 \quad , \tag{4.100}$$

$$U^*U = I \quad , \tag{4.101}$$

$$H = H^* \quad , \tag{4.102}$$

$$\text{offdg}(D) = 0 \quad . \tag{4.103}$$

4.104 Theorem. Let $A \in \mathcal{M}_n$. If $N_o = U_o^* D_o U_o$ where $U_o \in \mathcal{U}_n$ and $D_o \in \mathcal{D}_n$ is ϵ-minimal for A, then there exists a matrix $H_o \in \mathcal{M}_n$ such that (4.99) through (4.103) are satisfied when U,D, and H are replaced by U_o, D_o, and H_o respectively.

Proof. The proof of Theorem 4.104 follows along the same lines as the proof of Theorem 4.73.

Remark. We can use Theorem 4.104 to derive again the necessary conditions (4.12) and (4.13). Let $N_o = U_o^* D_o U_o$ be ϵ-minimal as in Theorem 4.104 and let H_o be a hermitian matrix of Lagrange multipliers which corresponds to it. Then using the conclusion of Theorem 4.104 we can immediately combine (4.99) and (4.101) to obtain $dg(D_o - U_o A U_o^*) = 0$ which is the same as (4.13). Premultiplying (4.100) by U_o^* and again

using (4.101) we find, upon setting $N_o = U_o^* D_o U_o$, that

$$N_o^* N_o + N_o N_o^* + H = N_o^* A + N_o A^* \quad .$$

The last equation shows clearly that $N_o^* A + N_o A^*$ is hermitian so, by Theorem 4.34, we have again derived the necessary condition (4.12).

Still another approach to the use of Lagrange multipliers is suggested by the representation of normal matrices given by Theorem 1.57 (a). Here we set

$$N = H + iK \quad . \tag{4.105}$$

By Theorem 1.57, (4.105) runs through $\mathcal{N}$ as H and K run through the set of all pairs of commuting hermitian matrices. Thus we try to minimize $\epsilon^2(H + iK - A)$ subject to the constraints

$$H = H^* \quad , \quad K = K^* \quad , \quad \text{and} \quad HK = KH \quad . \tag{4.106}$$

We introduce matrices Z, Π, and T of complex Lagrange multipliers which correspond respectively to the matrix constraints (4.106). Without going through the computational details we simply give here the results of the differentiations, which again turn out to be matrix equations:

$$H + iK - A + TK^* - K^*T + (Z - Z^*) = 0 \quad , \tag{4.107}$$

$$-i(H + iK - A) + H^*T - TH^* + (\Pi - \Pi^*) = 0 \quad . \tag{4.108}$$

Here, by analogy with what happened in the previous derivations, it is not the matrices Z and Π which have to be determined but the

skew-hermitian matrices $S = Z - Z^*$ and $P = \Pi - \Pi^*$ instead. Note, however, that no combinations of elements of T occurs so that every element of T is an "actual Lagrange multiplier". This was to be expected, since the equation $HK = KH$ involves no duplications.

If we rewrite (4.107) and (4.108), replacing $Z - Z^*$ and $\Pi - \Pi^*$ by S and P respectively, and write down all the relevant auxiliary equations we obtain the following set of simultaneous matrix equations for the five unknown matrices H, K, T, S, and P.

$$H + iK - A + TK^* - K^*T + S = 0 \quad , \tag{4.109}$$

$$-iH + K + iA + H^*T - TH^* + P = 0 \quad , \tag{4.110}$$

$$H = H^* \quad , \tag{4.111}$$

$$K = K^* \quad , \tag{4.112}$$

$$HK = KH \quad , \tag{4.113}$$

$$S + S^* = 0 \quad , \tag{4.114}$$

$$P + P^* = 0 \quad . \tag{4.115}$$

As in the previous constrained minimization problems (cf. Theorems 4.73 and 4.104) we can prove

4.116 Theorem. Let $A \in \mathcal{M}_n$. If $N_o = H_o + iK_o$ where $H_o, K_o \in \mathcal{H}_n$ and $H_oK_o = K_oH_o$ is ϵ-minimal for A then there exist matrices $T_o, S_o, P_o \in \mathcal{M}_n$ such that (4.109) through (4.115) are satisfied when $H, K, T, S,$ and P are replaced by $H_o, K_o, T_o, S_o,$ and P_o respectively.

CHAPTER 5

CHARACTERIZATION OF ϵ-MINIMAL MATRICES

We begin the present chapter by proving the following simple

5.1 Lemma. Let ν denote any norm on $\mathcal{M}$. We have

$$d_\nu(A) = \inf_{\substack{U \in \mathcal{U} \\ D \in \mathcal{D}}} \nu(A - U^*DU) = \inf_{U \in \mathcal{U}} \left(\inf_{D \in \mathcal{D}} \nu(A - U^*DU) \right) \tag{5.2}$$

$$= \inf_{D \in \mathcal{D}} \left(\inf_{U \in \mathcal{U}} \nu(A - U^*DU) \right) .$$

Proof. Let

$$d' = \inf_{U \in \mathcal{U}} \left(\inf_{D \in \mathcal{D}} \nu(A - U^*DU) \right) ,$$

$$\alpha(U) = \inf_{D \in \mathcal{D}} \nu(A - U^*DU) .$$

We want to show that $d' = d_\nu(A)$. Clearly $d' \leq \alpha(U)$ and

$$d_\nu(A) \leq \alpha(U) \tag{5.3}$$

hold for all $U \in \mathcal{U}$. Taking the infimum over $\mathcal{U}$ on the right side of (5.3) we obtain $d_\nu(A) \leq d'$. Suppose $d_\nu(A) < d'$; then there is a pair of matrices $U_o \in \mathcal{U}$, $D_o \in \mathcal{D}$ such that $\nu(A - U_o^* D_o U_o) < d'$. But this implies $\alpha(U_o) < d'$ which contradicts the inequality $d' \leq \alpha(U_o)$. Thus $d_\nu(A) = d'$ which proves the second equality in (5.2). A similar argument will establish the last equality in (5.2).

By very much the same arguments used in proving Theorem 1.7, it can be shown that any of the infima over $\mathcal{D}$ in (5.2) are assumed for some $D \in \mathcal{D}$. Furthermore, since ν is continuous and $\mathcal{U}$ is compact in the norm topology of $\mathcal{M}$, all infima over $\mathcal{U}$ are likewise assumed. Thus, in (5.2), we can replace "inf" by "Min" in every instance. We shall use this fact in proving

5.4 Theorem. Let $A \in \mathcal{M}$ and let ϵ denote the Euclidean norm (1.16) on $\mathcal{M}$. Then

$$d_\epsilon^2(A) = \epsilon^2(A) - \operatorname*{Max}_{U \in \mathcal{U}} \epsilon^2(dg(UAU^*)) \quad . \tag{5.5}$$

Proof. Since ϵ is unitarily invariant we have

$$\epsilon^2(A - U^*DU) = \epsilon^2(UAU^* - D) \quad . \tag{5.6}$$

From Definition 1.23 and (1.16) we can easily deduce the following equalities which hold for any $M \in \mathcal{M}$:

$$\operatorname*{Min}_{D \in \mathcal{D}} \epsilon^2(M - D) = \epsilon^2(offdg(M)) \quad , \tag{5.7}$$

$$\epsilon^2(M) = \epsilon^2(offdg(M) + dg(M)) = \epsilon^2(offdg(M)) + \epsilon^2(dg(M)) \tag{5.8}$$

where, in (5.7), the minimum is assumed if and only if $D = dg(M)$. Thus, for every $U \in \mathcal{U}$, we have

$$\operatorname*{Min}_{D \in \mathcal{D}} \epsilon^2(UAU^* - D) = \epsilon^2(UAU^* - dg(UAU^*)) \tag{5.9}$$

whence, using (5.2),

$$d_\epsilon^2(A) = \min_{U \in \mathcal{U}} \epsilon^2(UAU^* - dg(UAU^*)) \quad . \tag{5.10}$$

Noting that $\epsilon^2(UAU^*) = \epsilon^2(A)$ and combining (5.8) and (5.10) we obtain

$$\begin{aligned} d_\epsilon^2(A) &= \min_{U \in \mathcal{U}} [\epsilon^2(UAU^*) - \epsilon^2(dg(UAU^*))] \\ &= \min_{U \in \mathcal{U}} [\epsilon^2(A) - \epsilon^2(dg(UAU^*))] \\ &= \epsilon^2(A) - \max_{U \in \mathcal{U}} \epsilon^2(dg(UAU^*)) \end{aligned}$$

which proves (5.5).

5.11 Maximum Problem. Let $A \in \mathcal{M}_n$. Find a $U_o \in \mathcal{U}_n$ such that

$$\epsilon^2(dg(U_o A U_o^*)) = \max_{U \in \mathcal{U}_n} \epsilon^2(dg(UAU^*)) \quad . \tag{5.12}$$

5.13 Theorem (Characterization Theorem). Let $A \in \mathcal{M}_n$ and let $N_o = U_o^* D_o U_o$ where $U_o \in \mathcal{U}_n$ and $D_o \in \mathcal{D}_n$. Then N_o is ϵ-minimal for A if and only if U_o satisfies (5.12) (i.e., U_o solves the Maximum Problem 5.11) and

$$D_o = dg(U_o A U_o^*) \quad . \tag{5.14}$$

Proof. Let U_o be any unitary matrix which satisfies (5.12) and let D_o be given by (5.14). Then, from (5.5), we have

$$d_\epsilon^2(A) = \epsilon^2(A) - \epsilon^2(dg(U_o A U_o^*))$$

$$= \epsilon^2(U_o A U_o^*) - \epsilon^2(dg(U_o A U_o^*))$$

$$= \epsilon^2(offdg(U_o A U_o^*))$$

$$= \epsilon^2(U_o A U_o^* - dg(U_o A U_o^*)) = \epsilon^2(U_o A U_o^* - D_o)$$

$$= \epsilon^2(A - U_o^* D_o U_o) = \epsilon^2(A - N_o)$$

so that N_o is ϵ-minimal for A. This proves the sufficiency of the conditions (5.12) and (5.14). For their necessity we first note from Theorem 4.10 that the necessity of (5.14) has already been established. Thus, if N_o is ϵ-minimal for A, we have

$$d_\epsilon^2(A) = \epsilon^2(A - N_o) = \epsilon^2(U_o A U_o^* - D_o)$$

$$= \epsilon^2(U_o A U_o^* - dg(U_o A U_o^*))$$

$$= \epsilon^2(offdg(U_o A U_o^*))$$

$$= \epsilon^2(U_o A U_o^*) - \epsilon^2(dg(U_o A U_o^*))$$

$$= \epsilon^2(A) - \epsilon^2(dg(U_o A U_o^*)) \quad .$$

Comparing the last equation with (5.5) we see that U_o satisfies (5.12). This completes the proof of Theorem 5.13.

An immediate corollary of Theorem 5.13 is the following result which indicates the importance of the Maximum Problem 5.11 in so far as ϵ-minimal matrices are concerned.

5.15 Theorem. Let U_o be a solution of the Maximum Problem 5.11. Then

$$U_o^* \, dg(U_o A U_o^*) U_o \tag{5.16}$$

is ϵ-minimal for A.

Let u and v denote complex column n-vectors. The (Euclidean) inner product of u and v is defined by

$$(u,v) = v^* u$$

where v^* denotes the conjugate transpose of v. We call u and v orthogonal if $(u,v) = 0$. A set of vectors $u_1, u_2, \ldots, u_k$ $(k \leq n)$ is called orthonormal if

$$(u_i,u_j) = \delta_{ij} = \begin{cases} 1 \ , & i = j \\ 0 \ , & i \neq j \end{cases} .$$

5.17 Maximum Problem. Let $A \in \mathcal{M}_n$. Find an orthonormal set of column n-vectors $u_1, u_2, \ldots, u_n$ such that

$$\sum_{i=1}^{n} |(Au_i,u_i)|^2 = \text{Max} \sum_{i=1}^{n} |(Av_i,v_i)|^2$$

where the maximum is taken over all orthonormal sets $v_1, v_2, \ldots, v_n$ of column n-vectors.

Suppose $U \in \mathcal{M}_n$ and let $v_1, \ldots, v_n$ denote the columns of the matrix U^*. Clearly U is unitary if and only if $v_1, \ldots, v_n$ is orthonormal. Since $v^* A u = (Au,v)$ for all vectors u,v, we see from

(1.24) that

$$dg(UAU^*) = diag((Av_1,v_1), \ldots , (Av_n,v_n))$$

whence, by (1.16),

$$\epsilon^2(dg(UAU^*)) = \sum_{i=1}^{n} |(Av_i,v_i)|^2 \quad . \tag{5.19}$$

From (5.12), (5.18), and (5.19) one can immediately deduce

5.20 Theorem. The Maximum Problems 5.11 and 5.17 are equivalent in the sense that a solution of one yields immediately a solution of the other.

From the decomposition $N_o = U_o^* D_o U_o$ one finds that the columns of U_o^* are eigenvectors of N_o. Therefore, by Theorem 5.13, we have

5.21 Theorem. Let $A \in \mathcal{M}_n$ and let $N_o = U_o^* D_o U_o$ where $U_o \in \mathcal{U}_n$ and $D_o \in \mathcal{D}_n$. Then N_o is ϵ-minimal for A if and only if it has an orthonormal set of eigenvectors which solves the Maximum Problem 5.17.

The next result exhibits the relationship between the maximum (5.12) and distance formulas of the type (2.32).

5.22 Lemma. Let $A \in \mathcal{M}_n$, $A \notin \mathcal{M}_n$ and let $p_{\epsilon,n}(A)$ be defined by Definition 2.13 in the case $\nu = \epsilon$. Then

$$\max_{U \in \mathcal{U}_n} \epsilon^2(dg(UAU^*)) = (1 - p_{\epsilon,n}(A))\epsilon^2(A) + p_{\epsilon,n}(A)\, \epsilon^2(\Omega(A)) \quad . \tag{5.23}$$

Proof. Since ϵ has property S (see Definition 2.1), (2.32) holds for $\nu = \epsilon$. That equation and (5.5) yield immediately (5.23).

Remark. Theorem 5.4 opens up a new avenue for obtaining estimates of $d_\epsilon(A)$. Thus, a lower bound for the maximum (5.12) will yield an

upper bound for $d_\epsilon(A)$ and an upper bound for (5.12) will furnish a lower bound for $d_\epsilon(A)$. The next theorem is a result of this type.

5.24 Theorem. Let $A \in \mathcal{M}_n$ and let $A = WH$ where H is positive semidefinite hermitian and $W \in \mathcal{U}_n$. Furthermore let $y_1, y_2, \ldots, y_n$ be an orthonormal set of eigenvectors of H, i.e., $Hy_i = \alpha_i y_i$ $(1 \leq i \leq n)$ where α_i are the singular values of A and $(y_i, y_j) = \delta_{ij}$ $(1 \leq i, j \leq n)$. Then

$$d_\epsilon^2(A) \leq \sum_{i=1}^{n} \alpha_i^2 (1 - |(Wy_i, y_i)|^2) \quad . \tag{5.25}$$

Proof. We have

$$(Ay_i, y_i) = (WHy_i, y_i) = \alpha_i (Wy_i, y_i) \quad \text{for } i = 1, 2, \ldots, n \quad .$$

Since $y_1, \ldots, y_n$ is an orthonormal set, we see from (5.18) and (5.19) that

$$\sum_{i=1}^{n} |(Ay_i, y_i)|^2 = \sum_{i=1}^{n} \alpha_i^2 |(Wy_i, y_i)|^2 \tag{5.26}$$

is a lower bound for the maximum (5.12). As was observed in Section 1.1, $\epsilon(A)$ coincides with (1.15) for $p = 2$; consequently

$$\epsilon^2(A) = \sum_{i=1}^{n} \alpha_i^2 \quad . \tag{5.27}$$

Therefore, from (5.5), (5.26), and (5.27) we obtain

$$d_\epsilon^2(A) \le \sum_{i=1}^{n} \alpha_i^2 - \sum_{i=1}^{n} \alpha_i^2 \, |(Wy_i, y_i)|^2$$

which is the same as (5.25).

The next result provides a slight simplification of the problem of finding a solution to the Maximum Problem 5.11.

5.28 Lemma. Let $\mathcal{U}_n^+$ denote the set of all unitary matrices of order n which have nonnegative diagonal elements. Then, for any $A \in \mathcal{M}_n$, we have

$$\max_{U \in \mathcal{U}_n} \epsilon^2(dg(UAU^*)) = \max_{U \in \mathcal{U}_n^+} \epsilon^2(dg(UAU^*)) \quad . \tag{5.29}$$

Proof. Let U be any unitary matrix of order n and let $u_{kk} = r_k \exp(i\theta_k)$ be its k-th diagonal element where $r_k \ge 0$ and θ_k is real. By factoring out $\exp(i\theta_k)$ from the k-th row of U ($k = 1,2, \ldots, n$) we can write $U = \Lambda V$ where $V \in \mathcal{U}_n^+$ and $\Lambda = \mathrm{diag}(\exp(i\theta_1), \ldots, \exp(i\theta_n))$ is a diagonal unitary matrix, i.e., an element of $\mathcal{D}_n \cap \mathcal{U}_n$. Noting that $dg(\Lambda M \Lambda^*) = dg(M)$ holds for all $M \in \mathcal{M}_n$ and for all $\Lambda \in \mathcal{D}_n \cap \mathcal{U}_n$ we find that

$$\epsilon^2(dg(UAU^*)) = \epsilon^2(dg(\Lambda VAV^*\Lambda^*)) = \epsilon^2(dg(VAV^*)) \quad .$$

The last equation shows that, in computing the maximum (5.12), it suffices to consider only unitary matrices in $\mathcal{U}_n^+$. The proof of Lemma 5.28 is now complete.

CHAPTER 6

ϵ-MINIMAL MATRICES OF ORDER 2

6.1 Preliminaries Concerning 2 by 2 Matrices.

6.1 Lemma. Let $N \in \mathcal{M}_2$ be given by

$$N = \begin{pmatrix} n_1 & n_2 \\ n_3 & n_4 \end{pmatrix} . \tag{6.2}$$

Then N is normal if and only if

$$|n_2|^2 = |n_3|^2 \tag{6.3}$$

and

$$(n_1 - n_4)\bar{n}_3 = n_2(\bar{n}_1 - \bar{n}_4) . \tag{6.4}$$

Proof. We have

$$NN^* = \begin{pmatrix} |n_1|^2 + |n_2|^2 & n_1\bar{n}_3 + n_2\bar{n}_4 \\ \bar{n}_1 n_3 + \bar{n}_2 n_4 & |n_3|^2 + |n_4|^2 \end{pmatrix} ,$$

$$N^*N = \begin{pmatrix} |n_1|^2 + |n_3|^2 & \bar{n}_1 n_2 + \bar{n}_3 n_4 \\ n_1\bar{n}_2 + n_3\bar{n}_4 & |n_2|^2 + |n_4|^2 \end{pmatrix}$$

whence $NN^* - N^*N = 0$ if and only if (6.3) and (6.4) hold.

6.5 Lemma. Let $A \in \mathcal{M}_2$. Let $\Omega(A)$ and $M(A)$ be defined as in Definitions 1.17 and 3.8 respectively. Then

$$\epsilon^2(A - X) = \frac{1}{2}\,(\epsilon^2(A) - \epsilon^2(\Omega(A))) \quad \text{for all} \quad X \in M(A) \quad . \tag{6.6}$$

Proof. Let the eigenvalues of A be denoted by λ_1, λ_2. Then the eigenvalues of A^2 are λ_1^2 and λ_2^2 so we have $\mathrm{tr}(A) = \lambda_1 + \lambda_2$ and $\mathrm{tr}(A^2) = \lambda_1^2 + \lambda_2^2$. Thus

$$\begin{aligned} \frac{1}{2}\,|\mathrm{tr}(A)|^2 &= \frac{1}{2}\,(\lambda_1 + \lambda_2)(\bar{\lambda}_1 + \bar{\lambda}_2) \\ &= \frac{1}{2}\,(|\lambda_1|^2 + |\lambda_2|^2 + \lambda_1\bar{\lambda}_2 + \bar{\lambda}_1\lambda_2) \end{aligned}$$

or

$$\frac{1}{2}\,|\mathrm{tr}(A)|^2 = \frac{1}{2}\,\epsilon^2(\Omega(A)) + \mathrm{Re}(\lambda_1\bar{\lambda}_2) \quad . \tag{6.7}$$

Furthermore

$$\begin{aligned} \mathrm{tr}(A^2) - \frac{1}{2}\,(\mathrm{tr}(A))^2 &= \lambda_1^2 + \lambda_2^2 - \frac{1}{2}\,(\lambda_1 + \lambda_2)^2 \\ &= \lambda_1^2 + \lambda_2^2 - \frac{1}{2}\,(\lambda_1^2 + \lambda_2^2 + 2\lambda_1\lambda_2) \end{aligned}$$

or

$$\mathrm{tr}(A^2) - \frac{1}{2}\,(\mathrm{tr}(A))^2 = \frac{1}{2}\,(\lambda_1 - \lambda_2)^2 \quad . \tag{6.8}$$

From the last equation we obtain

$$|\mathrm{tr}(A^2) - \tfrac{1}{2}(\mathrm{tr}(A))^2| = \tfrac{1}{2}|\lambda_1 - \lambda_2|^2$$

$$= \tfrac{1}{2}(\lambda_1 - \lambda_2)(\bar{\lambda}_1 - \bar{\lambda}_2)$$

$$= \tfrac{1}{2}(|\lambda_1|^2 + |\lambda_2|^2 - \lambda_1\bar{\lambda}_2 - \bar{\lambda}_1\lambda_2)$$

or

$$|\mathrm{tr}(A^2) - \tfrac{1}{2}(\mathrm{tr}(A))^2| = \tfrac{1}{2}\epsilon^2(\Omega(A)) - \mathrm{Re}(\lambda_1\bar{\lambda}_2) \quad . \tag{6.9}$$

Combining (6.7) and (6.9) we find that

$$\frac{|\mathrm{tr}(A)|^2}{2} + |\mathrm{tr}(A^2) - \frac{(\mathrm{tr}(A))^2}{2}| = \epsilon^2(\Omega(A)) \quad . \tag{6.10}$$

The equation (6.6) now follows immediately from (6.10) and Theorem 3.11.

The next lemma shows how the simplification provided by Lemma 5.28 works out in the case $n = 2$.

<u>6.11 Lemma</u>. Let $A \in \mathcal{M}_2$. Let $\mu \in \underline{C}$ and let

$$W(\mu) = \frac{1}{\sqrt{1 + \mu\bar{\mu}}} \begin{pmatrix} 1 & -\bar{\mu} \\ \mu & 1 \end{pmatrix} . \tag{6.12}$$

Then $W(\mu) \in \mathcal{U}_2$ for all $\mu \in \underline{C}$ and

$$\max_{U \in \mathcal{U}_2} \epsilon^2(\mathrm{dg}(UAU^*)) = \max_{\mu \in \underline{C}} \epsilon^2(\mathrm{dg}(W(\mu)\, AW^*(\mu))) \quad . \tag{6.13}$$

<u>Proof</u>. It was shown in [3] (see also [17] for another parametrization of 2 by 2 unitary matrices) that every $U \in \mathcal{U}_2$ can be obtained from the formula

$$(6.14) \qquad U = \begin{pmatrix} e^{i\alpha} \cos\theta & -e^{i\beta} \sin\theta \\ e^{i\gamma} \sin\theta & e^{i\delta} \cos\theta \end{pmatrix}$$

where $\theta \in \underline{R}$ and where α, β, γ, δ are real numbers satisfying

$$(6.15) \qquad \alpha - \beta - \gamma + \delta \equiv 0 \pmod{2\pi} \ .$$

By factoring out $\exp(i\alpha)$ and $\exp(i\delta)$ respectively from the first and second rows of (6.14) and using (6.15) we see that (6.14) can be written in the form

$$(6.16) \qquad U = \begin{pmatrix} e^{i\alpha} & 0 \\ 0 & e^{i\delta} \end{pmatrix} \begin{pmatrix} \cos\theta & -e^{-i(\gamma-\delta)} \sin\theta \\ e^{i(\gamma-\delta)} \sin\theta & \cos\theta \end{pmatrix} = \Lambda W(\mu)$$

where $\Lambda = \mathrm{diag}(\exp(i\alpha), \exp(i\delta))$ and where $W(\mu)$ is given by (6.12) with $\mu = \exp[i(\gamma-\delta)]\tan\theta$. Since the real quantities θ, α, β, γ, δ in (6.14) are arbitrary except for the constraint (6.15) which was used in obtaining (and which is automatically satisfied by) the product $\Lambda W(\mu)$ in (6.16), one sees easily that (6.16) furnishes a parametrization of all $U \in \mathcal{U}_2$ in the following manner. If we consider θ, α, δ, and $\psi = \gamma - \delta$ as independent real variables, or, alternatively, if we consider α and δ as independent real variables and $\mu = \exp(i\psi)\tan\theta$ as an independent complex variable, then, as μ runs through $\underline{C}$ and as α, δ run through $\underline{R}$, $U = \Lambda W(\mu)$ runs through $\mathcal{U}_2$. As was shown in the proof of Lemma 5.28, the factor Λ in (6.16) can be disregarded

in computing $\epsilon^2(dg(UAU^*))$. Therefore only the complex parameter μ matters and we have (6.13).

6.17 Lemma. Let $A \in \mathcal{N}_2$, $A \notin \mathcal{N}_2$. Then A is unitarily similar either to a matrix of the form

$$A_1 = \begin{pmatrix} a_1 & a_2 \\ 0 & a_1 \end{pmatrix} \qquad \text{with } a_2 \neq 0 \tag{6.18}$$

or to a matrix of the form

$$A_2 = \begin{pmatrix} a_1 & a_2 \\ a_3 & a_1 \end{pmatrix} \qquad \text{with } a_2 \neq 0 \,, \; a_3 \neq 0 \,. \tag{6.19}$$

Proof. By Schur's theorem [22] (or see [16] p. 307) every A is unitarily similar to a triangular matrix:

$$VAV^* = B = \begin{pmatrix} b_1 & b_2 \\ 0 & b_4 \end{pmatrix} \qquad (V \in \mathcal{U}_2) \;; \tag{6.20}$$

and, if $A \notin \mathcal{N}_2$, we have $b_2 \neq 0$. If $b_1 = b_4$, i.e., if the two eigenvalues of A are equal, then (6.20) is already of the form (6.18). It remains only to show that, if $b_1 \neq b_4$, A is unitarily similar to a matrix of the form (6.19). To prove the latter statement we consider

$$W(\mu)\, BW^*(\mu) = \begin{pmatrix} a_1 & a_2 \\ a_3 & a_4 \end{pmatrix} \tag{6.21}$$

where $W(\mu)$ is given by (6.12) and where $\mu \in \underline{C}$. We have

$$a_1 - a_4 = (1 + \mu\bar{\mu})^{-1}[(b_1 - b_4)(1 - \mu\bar{\mu}) - 2\mu b_2]$$

hence $a_1 = a_4$ if and only if $2\mu b_2 = (b_1 - b_4)(1 - \mu\bar{\mu})$ or

$$2|\mu|e^{i \arg \mu} |b_2|e^{i \arg b_2} = |b_1 - b_4|e^{i \arg(b_1 - b_4)}(1 - |\mu|^2) \quad .$$

The last equation will be satisfied if

$$\arg \mu \equiv \arg(b_1 - b_4) - \arg b_2 \pmod{2\pi} \quad \text{and} \tag{6.22}$$

$$|\mu| = \frac{-|b_2| + \sqrt{|b_2|^2 + |b_1 - b_4|^2}}{|b_1 - b_4|} \quad . \tag{6.23}$$

Thus, if μ is chosen to satisfy (6.22) and (6.23), $a_1 = a_4$ holds in (6.21). Furthermore, for this choice of μ, neither a_2 nor a_3 in (6.21) can vanish since otherwise the eigenvalues of A, namely b_1 and b_4, would have to be equal. This completes the proof of Lemma 6.17.

6.2 Determination of all 2 by 2 ϵ-Minimal Matrices.

6.24 Theorem. Let $A \in \mathcal{M}_2$ and let M(A) be defined as in Definition 3.8. Then the set M(A) represents the totality of all ϵ-minimal matrices for A.

Proof. Suppose first that $A \in \mathcal{M}_2$. Then by Lemma 2.4 $\epsilon^2(A) = \epsilon^2(\Omega(A))$ whence Lemma 6.5 implies

$$\epsilon^2(A - X) = 0 \quad \text{for all } X \in M(A) \text{ where } A \in \mathcal{M}_2 \quad . \tag{6.25}$$

The meaning of the last equation is that every matrix in the set (3.10) is equal to A:

$$M(A) = \{A\} \qquad (A \in \mathcal{N}_2) \tag{6.26}$$

that is, if $A \in \mathcal{N}_2$, then the set (3.10) contains exactly one matrix, namely A itself! This proves Theorem 6.24 in the case $A \in \mathcal{N}_2$.

We assume henceforth in this proof that $A \notin \mathcal{N}_2$. The principal tools we shall employ in the case $A \notin \mathcal{N}_2$ are Theorem 5.13 and Lemma 6.11. In order to keep the computations manageable we shall employ a change of coordinates defined by a unitary matrix which transforms A into one of the forms (6.18) or (6.19). This procedure is justified in the next paragraph. The general outline of our proof is as follows. We shall first show that every matrix in the set M(A) is ϵ-minimal. Following this we show that there are no other ϵ-minimal matrices.

From (3.9) and the invariance of the trace under unitary similarity we have

$$\zeta(A) = \zeta(UAU^*) \qquad (A \in \mathcal{M}_n, \ U \in \mathcal{U}_n) \ . \tag{6.27}$$

In like manner we obtain the identity

$$P_\zeta(UAU^*) = UP_\zeta(A)U^* \qquad (A \in \mathcal{M}_n, \ U \in \mathcal{U}_n) \tag{6.28}$$

for all $\zeta \in \underline{C}$ where P_ζ is defined by (4.49). In view of (3.10), (6.28), and Theorem 2.27 it will suffice to prove the conclusion of Theorem 6.24 for some unitary transform of A. As mentioned in the

preceding paragraph we shall use the forms (6.18) and (6.19) for this purpose.

We first observe from (6.8) that the case in which A is unitarily similar to (6.18) corresponds to the ambiguous case of (3.9). Similarly the case in which A is similar to (6.19) corresponds to the non-ambiguous case of (3.9) where $\zeta(A)$ contains exactly one number. As stated in the note following Definition 3.8, we use the notation $N_\theta(A_1)$ to denote one of the matrices in $M(A_1)$ and $N(A_2)$ to denote the (single) matrix in $M(A_2)$. Straightforward calculations based on Defintion 3.8 yield

$$(6.29) \qquad N_\theta(A_1) = \begin{pmatrix} a_1 & \frac{1}{2}a_2 \\ \frac{1}{2}\bar{a}_2 e^{i\theta} & a_1 \end{pmatrix} \qquad (\theta \in \underline{R}) \ ,$$

$$(6.30) \quad N(A_2) = \begin{pmatrix} a_1 & \frac{1}{2}(a_2 + \zeta\bar{a}_3) \\ \frac{1}{2}(a_3 + \zeta\bar{a}_2) & a_1 \end{pmatrix} \quad \text{where} \quad \zeta = \frac{a_2 a_3}{|a_2 a_3|} \ .$$

In order to prove that $N_\theta(A_1)$ is ϵ-minimal for A_1 for all $\theta \in \underline{R}$ and that $N(A_2)$ is ϵ-minimal for A_2, we shall diagonalize (6.29) and (6.30) by unitary matrices of the form (6.12) and verify the sufficient conditions of Theorem 5.13.

Let

$$M = \begin{pmatrix} m_1 & m_2 \\ m_3 & m_4 \end{pmatrix}$$

and let $W(\mu)$ be given by (6.12); then

$$(6.31)\qquad W(\mu)MW^*(\mu) = \frac{1}{1+\mu\bar{\mu}}\begin{pmatrix} m_1-\mu m_2-\bar{\mu}m_3+\mu\bar{\mu}m_4 & \bar{\mu}m_1+m_2-\bar{\mu}^2m_3-\bar{\mu}m_4 \\ \mu m_1-\mu^2 m_2+m_3-\mu m_4 & \mu\bar{\mu}m_1+\mu m_2+\bar{\mu}m_3+m_4 \end{pmatrix} .$$

In order to diagonalize $N_\theta(A_1)$ and $N(A_2)$ it will suffice, by Theorem 1.59, to triangularize them. Thus, from (6.29) and (6.31) we see that $W(\mu_{1\theta})N_\theta(A_1)W^*(\mu_{1\theta})$ is diagonal if and only if

$$(6.32)\qquad \mu_{1\theta}^2 = \frac{\bar{a}_2 e^{i\theta}}{a_2} .$$

Similarly, $W(\mu_2)N(A_2)W^*(\mu_2)$ is diagonal if and only if

$$(6.33)\qquad \mu_2^2 = \frac{a_3+\zeta\bar{a}_2}{a_2+\zeta\bar{a}_3} = \frac{a_3}{a_2}\cdot\frac{|a_2a_3|+|a_2|^2}{|a_2a_3|+|a_3|^2} = \frac{a_3|a_2|}{a_2|a_3|} .$$

For notational simplicity we shall usually omit the subscript θ from $\mu_{1\theta}$ and write only μ_1. However, it should always be understood that μ_1 depends on the parameter θ. Clearly, each of equations (6.32) and (6.33) has two solutions and we are at liberty to choose either solution in diagonalizing $N_\theta(A_1)$ or $N(A_2)$. In what follows it will not matter which solution is chosen, so we use the symbols μ_1 and μ_2 to denote <u>any</u> solutions of (6.32) and (6.33). We note from (6.32) that

$$(6.34)\qquad |\mu_{1\theta}| = 1 ,$$

$$(6.35)\qquad \mu_{1\theta}^2 a_2^2 = |a_2|^2 e^{i\theta} .$$

From (6.33) we have

$$\mu_2^2 a_2 \bar{a}_3 = \bar{\mu}_2^2 \bar{a}_2 a_3 = |a_2 a_3| > 0 \quad ; \tag{6.36}$$

and, from (6.30), (6.33), and Lemma 6.1 (equation (6.3)), we obtain

$$|\mu_2| = 1 \quad . \tag{6.37}$$

By straightforward calculations we find that

$$W(\mu_1)N_\theta(A_1)W^*(\mu_1) = \frac{1}{2}\begin{pmatrix} 2a_1 - \frac{1}{2}(\mu_1 a_2 + \bar{\mu}_1 \bar{a}_2 e^{i\theta}) & 0 \\ 0 & 2a_1 + \frac{1}{2}(\mu_1 a_2 + \bar{\mu}_1 \bar{a}_2 e^{i\theta}) \end{pmatrix} , \tag{6.38}$$

$$W(\mu_2)N(A_2)W^*(\mu_2) = \frac{1}{2}\begin{pmatrix} 2a_1 - \frac{1}{2}[\mu_2(a_2+\zeta\bar{a}_3) + \bar{\mu}_2(a_3+\zeta\bar{a}_2)] & 0 \\ 0 & 2a_1 + \frac{1}{2}[\mu_2(a_2+\zeta\bar{a}_3) + \bar{\mu}_2(a_3+\zeta\bar{a}_2)] \end{pmatrix} , \tag{6.39}$$

$$W(\mu_1)A_1W^*(\mu_1) = \frac{1}{2}\begin{pmatrix} 2a_1 - \mu_1 a_2 & a_2 \\ -\mu_1^2 a_2 & 2a_1 + \mu_1 a_2 \end{pmatrix} , \tag{6.40}$$

$$W(\mu_2)A_2W^*(\mu_2) = \frac{1}{2}\begin{pmatrix} 2a_1 - (\mu_2 a_2 + \bar{\mu}_2 a_3) & a_2 - \bar{\mu}_2^2 a_3 \\ a_3 - \mu_2^2 a_2 & 2a_1 + (\mu_2 a_2 + \bar{\mu}_2 a_3) \end{pmatrix} . \tag{6.41}$$

From (6.32) and (6.34) we obtain $\mu_1 a_2 = \bar{\mu}_1 \bar{a}_2 \exp(i\theta)$ whence, using (6.38) and (6.40),

$$(6.42) \quad W(\mu_1)N_\theta(A_1)W^*(\mu_1) = \tfrac{1}{2}\,\mathrm{diag}(2a_1 - \mu_1 a_2, 2a_1 + \mu_1 a_2) = \mathrm{dg}(W(\mu_1)A_1 W^*(\mu_1)) \ .$$

Using (6.37) we can write

$$\mu_2(a_2 + \zeta\bar{a}_3) + \bar{\mu}_2(a_3 + \zeta\bar{a}_2) = \bar{\mu}_2[\mu_2^2 a_2 + \mu_2^2\zeta\bar{a}_3 + a_3 + \zeta\bar{a}_2]$$

and, by (6.30) and (6.36), we have

$$\mu_2^2\zeta\bar{a}_3 = \mu_2^2 a_2\bar{a}_3 \frac{1}{|a_2 a_3|}\, a_3 = a_3 \ ,$$

$$\zeta\bar{a}_2 = a_2 \frac{\bar{a}_2 a_3}{|a_2 a_3|} = a_2 \frac{1}{\bar{\mu}_2^2} = \mu_2^2 a_2$$

so that

$$(6.43) \quad \mu_2(a_2+\zeta\bar{a}_3) + \bar{\mu}_2(a_3+\zeta\bar{a}_2) = 2\bar{\mu}_2(\mu_2^2 a_2 + a_3) = 2(\mu_2 a_2 + \bar{\mu}_2 a_3) \ .$$

Combining (6.39), (6.41) and (6.43) we obtain

$$W(\mu_2)N(A_2)W^*(\mu_2) = \tfrac{1}{2}\,\mathrm{diag}[2a_1 - (\mu_2 a_2 + \bar{\mu}_2 a_3),\ 2a_1 + (\mu_2 a_2 + \bar{\mu}_2 a_3)]$$

$$(6.44) \qquad = \mathrm{dg}(W(\mu_2)A_2 W^*(\mu_2)) \ .$$

Equations (6.42) and (6.44) merely express the fact that the decompositions of $N_\theta(A_1)$ and $N(A_2)$ implied in (6.38) and (6.39) satisfy the condition (5.14) of Theorem 5.13.

In order to show that every matrix in the set $M(A)$ is ϵ-minimal for A, it remains only to prove that $W(\mu_i)$ solves the Maximum Problem 5.11 for A_i $(i = 1,2)$. By Lemma 6.11 it will suffice to prove that

$$\sigma_1(\lambda) \leq \epsilon^2[dg(W(\mu_i)A_iW^*(\mu_i))] \qquad (i = 1,2) \tag{6.45}$$

hold for all $\lambda \in \underline{C}$ where

$$\sigma_1(\lambda) = \epsilon^2[dg(W(\lambda)W(\mu_i)A_iW^*(\mu_i)W^*(\lambda))] \qquad (i = 1,2) \ . \tag{6.46}$$

Straightforward calculations using (6.12), (6.32) through (6.37), (6.40), and (6.41) yield

$$\epsilon^2[dg(W(\mu_1)A_1W^*(\mu_1))] = 2\ |a_1|^2 + \tfrac{1}{2}\ |a_2|^2 \ , \tag{6.47}$$

$$\epsilon^2[dg(W(\mu_2)A_2W^*(\mu_2))] = 2\ |a_1|^2 + \tfrac{1}{2}\ (|a_2| + |a_3|)^2 \ , \tag{6.48}$$

$$\sigma_1(\lambda) = \frac{1}{\alpha^2}\ (|\alpha a_1 + \beta_1 a_2|^2 + |\alpha a_1 - \beta_1 a_2|^2) \ , \tag{6.49}$$

$$\sigma_2(\lambda) = \frac{1}{\alpha^2}\ (|\alpha a_1 + \beta_2 a_2 + \bar{\beta}_2 a_3|^2 + |\alpha a_1 - \beta_2 a_2 - \bar{\beta}_2 a_3|^2) \tag{6.50}$$

where

$$\alpha = 2(1 + \lambda\bar{\lambda}) \tag{6.51}$$

and

(6.52) $$\beta_i = \mu_i^2\bar{\lambda} + \mu_i\lambda\bar{\lambda} - \mu_i - \lambda \qquad (i = 1,2) \ .$$

After further calculations we find that

(6.53) $$\sigma_1(\lambda) = 2|a_1|^2 + \frac{2|\beta_1|^2}{\alpha^2}\,|a_2|^2 \ ,$$

(6.54) $$\sigma_2(\lambda) = 2|a_1|^2 + \frac{2}{\alpha^2}\,|\beta_2 a_2 + \bar{\beta}_2 a_3|^2 \ ,$$

(6.55) $$\beta_i = -\mu_i(\mu_i + \lambda)(\bar{\mu}_i - \bar{\lambda}) \qquad (i = 1,2) \ ,$$

(6.56) $$|\beta_i| = |\mu_i^2 - \lambda^2| \qquad (i = 1,2) \ .$$

Rewriting (6.45) using (6.47) and (6.48) we obtain

(6.57) $$\sigma_1(\lambda) \le 2|a_1|^2 + \tfrac{1}{2}\,|a_2|^2 \ ,$$

(6.58) $$\sigma_2(\lambda) \le 2|a_1|^2 + \tfrac{1}{2}\,(|a_2| + |a_3|)^2 \ .$$

We wish to show that (6.57) and (6.58) hold for all $\lambda \in \underline{C}$. One sees from (6.53) that (6.57) holds if and only if $2|\beta_1|^2/\alpha^2 \le 1/2$ that is

(6.59) $$\frac{|\beta_1|^2}{(1 + \lambda\bar{\lambda})^2} \le 1 \qquad \text{for all } \lambda \in \underline{C} \ .$$

Furthermore equality holds in (6.57) if and only if equality holds in (6.59). From (6.34) and (6.37) we have $|\mu_i| = 1$ $(i = 1,2)$;

consequently

$$|\mu_i^2 - \lambda^2| \le |\mu_i|^2 + |\lambda|^2 = 1 + \lambda\bar{\lambda} \qquad (i = 1,2)$$

so, from (6.56) and the last inequality, we see that

$$(6.60) \qquad \frac{|\beta_i|^2}{(1 + \lambda\bar{\lambda})^2} = \frac{|\mu_i^2 - \lambda^2|^2}{(1 + \lambda\bar{\lambda})^2} \le \frac{(1 + \lambda\bar{\lambda})^2}{(1 + \lambda\bar{\lambda})^2} = 1 \qquad (i = 1,2)$$

hold for all $\lambda \in \underline{C}$. Equality holds in (6.60) if and only if $-\bar{\mu}_i^2\lambda^2$ is real and nonnegative, i.e., $-\bar{\mu}_i^2\lambda^2 \ge 0$. By the triangle inequality

$$(6.61) \qquad |\beta_2 a_2 + \bar{\beta}_2 a_3|^2 \le |\beta_2|^2 (|a_2| + |a_3|)^2$$

with equality holding if and only if $\beta_2^2 a_2 \bar{a}_3 \ge 0$. From (6.36) and (6.55) we have

$$(6.62) \quad \beta_2^2 a_2 \bar{a}_3 = \mu_2^2 a_2 \bar{a}_3 (\mu_2 + \lambda)^2 (\bar{\mu}_2 - \bar{\lambda})^2 = |a_2 a_3| (\mu_2 + \lambda)^2 (\bar{\mu}_2 - \bar{\lambda})^2 \ .$$

From (6.51), (6.54), and (6.61) we obtain

$$\sigma_2(\lambda) \le 2|a_1|^2 + \frac{|\beta_2|^2}{(1 + \lambda\bar{\lambda})^2} \cdot \frac{1}{2} (|a_2| + |a_3|)^2$$

consequently, by (6.60) and (6.62), (6.58) is valid for all $\lambda \in \underline{C}$ with equality if and only if

$$(6.63) \qquad (\mu_2 + \lambda)^2 (\bar{\mu}_2 - \bar{\lambda})^2 \ge 0$$

and

$$(6.64) \qquad -\bar{\mu}_2^{-2}\lambda^2 \geq 0 \quad .$$

Furthermore, from (6.60) and (6.59), we see that (6.57) holds for all $\lambda \in \underline{C}$ with equality if and only if

$$(6.65) \qquad -\bar{\mu}_{1\theta}^{-2}\lambda^2 \geq 0 \quad .$$

This completes the proof of the fact that every matrix in the set $M(A)$ is ϵ-minimal for A.

In order to determine all ϵ-minimal matrices for A it will suffice, by Theorem 2.27 and Lemma 6.17, to determine all ϵ-minimal matrices for $B_i = W(\mu_i)A_iW^*(\mu_i)$ $(i = 1,2)$. The following lemma will provide help in making that determination.

6.66 Lemma. Let $B \in \mathcal{M}_2$ and let $N_o = U_o^* D_o U_o$, where $U_o \in \mathcal{U}_2$ and $D_o \in \mathcal{D}_2$, be ϵ-minimal for B. Then there exists a $\lambda_o \in \underline{C}$ such that

$$N_o = W^*(\lambda_o)\, dg(W(\lambda_o)BW^*(\lambda_o))W(\lambda_o) \quad .$$

Proof. By the parametric representation of $\mathcal{U}_2$ developed in the proof of Lemma 6.11 there exist $\Lambda_o \in \mathcal{D}_2 \cap \mathcal{U}_2$ and $\lambda_o \in \underline{C}$ such that $U_o = \Lambda_o W(\lambda_o)$. Thus, recalling that $dg(\Lambda_o M \Lambda_o^*) = dg(M)$ for any $M \in \mathcal{M}_2$, we have

$$dg(U_o B U_o^*) = dg(W(\lambda_o)BW^*(\lambda_o)) \quad ;$$

consequently by Theorem 5.13 and the fact that diagonal matrices commute

$$N_o = U_o^* \ dg(W(\lambda_o)BW^*(\lambda_o))U_o$$

$$= W^*(\lambda_o)\Lambda_o^* \ dg(W(\lambda_o)BW^*(\lambda_o))\Lambda_o W(\lambda_o)$$

$$= W^*(\lambda_o) \ dg(W(\lambda_o)BW^*(\lambda_o))W(\lambda_o)$$

as desired.

Since $W(\mu_i)$ solves the Maximum Problem 5.11 for A_i $(i = 1,2)$ we see from Lemma 6.66 that, if we determine all $\lambda_o \in \underline{C}$ such that

$$\epsilon^2(dg(B_i)) = \epsilon^2(dg(W(\lambda_o)B_iW^*(\lambda_o))) \ ,$$

then we will have determined all ϵ-minimal matrices for B_i $(i = 1,2)$. By the definition (6.46) of $\sigma_i(\lambda)$ this amounts to determining all cases of equality in (6.57) and (6.58). We shall do this by using the next two lemmas.

6.67 Lemma. The inequalities

$$(6.68) \qquad -\overline{\mu}_k^2\lambda^2 \geq 0 \qquad (k = 1,2)$$

hold if and only if

$$(6.69) \qquad \lambda = \rho i \ \mu_k \qquad (k = 1,2)$$

where ρ is real.

Proof. Let $\lambda = r \exp(i\varphi)$ where $r \geq 0$ and φ is real. Since $|\mu_k| = 1$ $(k = 1,2)$ we can set $\mu_k = \exp(i\varphi_k)$. Thus

(6.70) $$-\bar{\mu}_k^2\lambda^2 = -r^2 e^{i\,2(\varphi-\varphi_k)} .$$

Obviously (6.68) holds if $r = 0$. If $r > 0$ then, from (6.70), we see that (6.68) holds if and only if $\exp[i\,2(\varphi - \varphi_k)] = -1$ i.e., if and only if $\exp(i\varphi) = \pm i\,\mu_k$ or $\lambda = \pm r\,i\,\mu_k$. Setting $\rho = \pm r$ we see that (6.68) holds if and only if λ has the form (6.69) where ρ is any real number.

6.71 Lemma. The inequality (6.63) holds if and only if

(6.72) $$\lambda = \sigma\mu_2$$

where σ is real.

Proof. Straightforward calculations using (6.37) yield

(6.73) $$(\mu_2+\lambda)^2(\bar{\mu}_2-\bar{\lambda})^2 = [1-4|\lambda|^2+|\lambda|^4 + 2\,\mathrm{Re}(\lambda^2\bar{\mu}_2^2)]+[2(|\lambda|^2-1)(\mu_2\bar{\lambda}-\bar{\mu}_2\lambda)] .$$

The first term in square brackets on the right of (6.73) is real while the second such term is pure imaginary. Therefore in order that (6.63) hold it is necessary and sufficient that

(6.74) $$(|\lambda|^2 - 1)(\mu_2\bar{\lambda} - \bar{\mu}_2\lambda) = 0$$

and

(6.75) $$f(\lambda,\mu_2) = 1 - 4|\lambda|^2 + |\lambda|^4 + 2\,\mathrm{Re}(\lambda^2\bar{\mu}_2^2) \geq 0 .$$

Obviously (6.74) holds if and only if either $|\lambda| = 1$ or $\mu_2\bar{\lambda} = \bar{\mu}_2\lambda$. If $|\lambda| = 1$ then

$$f(\lambda,\mu_2) = 2(\mathrm{Re}(\lambda^2\bar{\mu}_2^2) - 1) \leq 0$$

whence (6.75) holds if and only if equality holds in the last inequality, i.e.,

$$\lambda = \pm \mu_2 \quad . \tag{6.76}$$

If $|\lambda| \neq 1$ then (6.74) holds if and only if $\mu_2\bar{\lambda} = \bar{\mu}_2\lambda$, i.e., $\lambda = \mu_2^2\bar{\lambda}$. Let $\lambda = r\exp(i\varphi)$ where $r \geq 0$ and φ is real. We need consider only the case $r > 0$ since (6.63) holds if $\lambda = 0$. Thus $r\exp(i\varphi) = \mu_2^2 r\exp(-i\varphi)$ or $\exp(i2\varphi) = \mu_2^2$ or $\exp(i\varphi) = \pm\mu_2$. Consequently

$$\lambda = \pm r\mu_2 \quad . \tag{6.77}$$

Substituting this into $f(\lambda,\mu_2)$ defined in (6.75) we have

$$f(\lambda,\mu_2) = 1 + 4r^2 + r^4 + 2r^2 = r^4 - 2r^2 + 1 = (r^2 - 1)^2 \geq 0$$

for all r. Setting $\sigma = \pm r$ and combining (6.76) and (6.77) we see that (6.74) and (6.75) hold simultaneously if and only if λ has the form (6.72) where σ is any real number. This completes the proof of Lemma 6.71.

Consider first the case in which $M(A)$ contains exactly one matrix, i.e., the case in which A is unitarily similar to A_2. Here we wish to determine all cases of equality in (6.58). We have previously

observed that equality holds in (6.58) if and only if (6.63) and (6.64) hold simultaneously. Therefore by Lemmas 6.67 and 6.71 we must have $\lambda = \sigma\mu_2$ and $\lambda = \rho i\mu_2$ where ρ and σ are real. But the last two equations represent straight lines in the complex plane which intersect only for $\rho = \sigma = 0$, i.e., only for $\lambda = 0$. Thus equality holds in (6.58) only for $\lambda = 0$. Since $W(0) = I$ we see from Theorem 5.13 and Lemma 6.66 that

$$N(A_2) = W^*(\mu_2)\, dg(W(\mu_2)A_2W^*(\mu_2))W(\mu_2)$$

is the only ϵ-minimal matrix for A_2. By Corollary 2.30 A has exactly one ϵ-minimal matrix, namely $N(A)$.

Now consider the case in which A is unitarily similar to A_1. Here $M(A)$ contains infinitely many matrices and we wish to determine all cases of equality in (6.57). We observed previously that equality in (6.57) if and only if (6.65) holds. By Lemma 6.67 equality holds in (6.57) if and only if $\lambda = \rho i\mu_{1\theta}$ where ρ is real and $\mu_{1\theta}$ is defined by (6.32). Using (6.34) and (6.12) we obtain

$$W(\lambda)W(\mu_1) = \frac{1}{\sqrt{2}\sqrt{1+\rho^2}} \begin{pmatrix} 1 & \rho i\bar{\mu}_1 \\ \rho i\mu_1 & 1 \end{pmatrix}\begin{pmatrix} 1 & -\bar{\mu}_1 \\ \mu_1 & 1 \end{pmatrix}$$

$$= \frac{1}{\sqrt{2}\sqrt{1+\rho^2}} \begin{pmatrix} 1+\rho i & \bar{\mu}_1(-1+\rho i) \\ \mu_1(1+\rho i) & 1-\rho i \end{pmatrix}$$

$$= \frac{1}{\sqrt{1+\rho^2}} \begin{pmatrix} 1+\rho i & 0 \\ 0 & 1-\rho i \end{pmatrix}\frac{1}{\sqrt{2}}\begin{pmatrix} 1 & -\bar{\mu}_1\overline{\left(\frac{1+\rho i}{1-\rho i}\right)} \\ \mu_1\frac{1+\rho i}{1-\rho i} & 1 \end{pmatrix}$$

$$= \Lambda_\rho\, W(\mu_1 e^{i\theta_\rho/2})$$

where

$$\Lambda_\rho = \begin{pmatrix} \frac{1+\rho i}{\sqrt{1+\rho^2}} & 0 \\ 0 & \frac{1-\rho i}{\sqrt{1+\rho^2}} \end{pmatrix} \in \mathcal{D}_2 \cap \mathcal{U}_2$$

and where

$$e^{i\theta_\rho/2} = \frac{1+\rho i}{1-\rho i} \quad .$$

From (6.32) we have

$$\mu_{1\theta}\, e^{i\theta_\rho/2} = \sqrt{a_2/a_2}\; e^{i\frac{\theta+\theta_\rho}{2}} = \mu_{1,\theta+\theta_\rho} \quad .$$

Therefore by Lemma 6.66 every ϵ-minimal matrix for A_1 is of the form

$$W^*(\mu_1)W^*(\lambda)\, dg(W(\lambda)W(\mu_1)A_1W^*(\mu_1)W^*(\lambda))W(\lambda)W(\mu_1)$$

$$= W^*(\mu_{1,\theta+\theta_\rho})\Lambda_\rho^*\, dg(\Lambda_\rho W(\mu_{1,\theta+\theta_\rho})A_1W^*(\mu_{1,\theta+\theta_\rho})\Lambda_\rho^*)\Lambda_\rho W(\mu_{1,\theta+\theta_\rho})$$

$$= W^*(\mu_{1,\theta+\theta_\rho})\, dg(W(\mu_{1,\theta+\theta_\rho})A_1W^*(\mu_{1,\theta+\theta_\rho}))W(\mu_{1,\theta+\theta_\rho})$$

$$= N_{\theta+\theta_\rho}(A_1)$$

which is just another one of the matrices in the set $M(A_1)$ no matter what the value of ρ is. Thus $M(A_1)$ contains all the ϵ-minimal matrices for A_1. This completes the proof of Theorem 6.24.

Some immediate consequences of Theorem 6.24 are the following.

6.78 Corollary. If $A \in \mathcal{N}_2$ then the set (3.10) contains exactly one matrix, namely A.

6.79 Theorem. Let λ_1 and λ_2 denote the eigenvalues of $A \in \mathcal{N}_2$. If $A \notin \mathcal{N}_2$ then there is a unique ϵ-minimal matrix for A if and only if $\lambda_1 \neq \lambda_2$; if $\lambda_1 = \lambda_2$ there are infinitely many ϵ-minimal matrices for A.

Proof. By Theorem 6.24 A has a unique ϵ-minimal matrix if and only if $\mathrm{tr}(A^2) - (1/2)(\mathrm{tr}(A))^2 \neq 0$ (cf. Definition 3.8). By (6.8) this happens if and only if $\lambda_1 \neq \lambda_2$. Similarly, if $\lambda_1 = \lambda_2$, M(A) contains an infinite number of matrices.

6.80 Theorem. Mirsky's Conjecture 1.35 is true for $\nu = \epsilon$ and $n = 2$.

Proof. This follows immediately from Theorem 6.24, Lemma 6.5, and the definition (1.34) of $d_\epsilon(A)$.

6.81 Theorem. The set M(A) of (3.10) provides a complete solution to Problem 3.40 for $n = 2$ and $k = 1$.

Proof. Clearly $M(A) \subset \mathcal{L}_1(A) \subset \mathcal{N}_2$ so by Theorem 6.24 we have the desired conclusion.

For any subset $\mathcal{S}$ of $\mathcal{M}_n$ we denote the set of all real matrices in $\mathcal{S}$ by $\mathcal{S}/\underline{R}$ (read: $\mathcal{S}$ restricted to $\underline{R}$). Theorem 6.24 provides a complete solution to the distance problem (and associated minimum problem) of finding (1.26) where $\nu = \epsilon$, $A \in \mathcal{N}_2/\underline{R}$ and where $\mathcal{S}$ is replaced by $\mathcal{N}_2/\underline{R}$. For, if $A \in \mathcal{N}_2/\underline{R}$ has eigenvalues λ_1, λ_2, then the set M(A) contains all matrices in $\mathcal{N}_2$ which are ϵ-minimal for A. If $\lambda_1 \neq \lambda_2$ (cf. (6.8)) then by (3.9) $\zeta(A)$ contains exactly one number and that is real; consequently by (3.10) M(A) contains exactly one matrix and that is real. If $\lambda_1 = \lambda_2$ then $\zeta(A)$ contains all

complex numbers of unit modulus and by (3.10) $N_\zeta(A)$ is real if and only if either $\zeta = +1$ or $\zeta = -1$. We have proved

6.82 Theorem. Let $A \in \mathcal{M}_2/\underline{R}$, $A \notin \mathcal{M}_2$ and let the eigenvalues of A be denoted by λ_1, λ_2. If $\lambda_1 \neq \lambda_2$ there is a unique real ϵ-minimal matrix for A given by (3.10). If $\lambda_1 = \lambda_2$ there are exactly two real ϵ-minimal matrices for A, namely

$$\frac{1}{2}(A \pm A^*) + \frac{1}{4}\,\mathrm{tr}(A \mp A^*)I \quad .$$

6.3 The Maximum Problem 5.11 in the Case $n = 2$.

6.83 Theorem. Let $A \in \mathcal{M}_2$. The identity matrix I solves the Maximum Problem 5.11 for A, i.e., A satisfies

$$\epsilon^2(dg(A)) = \max_{U \in \mathcal{U}_2} \epsilon^2(dg(UAU^*)) \quad , \tag{6.84}$$

if and only if

$$A + \zeta A^* \in \mathcal{D}_2 \tag{6.85}$$

for some ζ in the set $\zeta(A)$ defined by (3.9).

Proof. If (6.85) holds for $\zeta \in \zeta(A)$ then by (3.10) $N_\zeta(A) \in \mathcal{D}_2$ whence I diagonalizes $N_\zeta(A)$. By Theorem 6.24 $N_\zeta(A)$ is ϵ-minimal so by Theorem 5.13 I solves the Maximum Problem 5.11. Suppose now that (6.84) holds. Then by Theorem 5.13 $dg(A)$ is ϵ-minimal for A whence $dg(A) \in M(A)$ by Theorem 6.24. Thus there is a $\zeta \in \zeta(A)$ such that

$N_\zeta(A) = dg(A)$, i.e., $N_\zeta(A) \in \mathcal{D}_2$, and this implies (6.85).

6.86 Theorem. Let $A \in \mathcal{M}_2$. U_o solves the Maximum Problem 5.11 for A if and only if $U_o \in \mathcal{U}_2$ and

$$U_o(A + \zeta A^*)U_o^* \in \mathcal{D}_2 \tag{6.87}$$

for some ζ in the set $\zeta(A)$ defined by (3.9).

Proof. Every $V \in \mathcal{U}_2$ can be written uniquely as $V = UU_o$ and V runs through $\mathcal{U}_2$ if and only if U runs through $\mathcal{U}_2$. Thus

$$\epsilon^2(dg(U_oAU_o^*)) = \max_{U \in \mathcal{U}_2} \epsilon^2(dg(U(U_oAU_o^*)U^*)) \tag{6.88}$$

whence $U_oAU_o^*$ satisfies (6.88) if and only if U_o solves the Maximum Problem 5.11. By virtue of (6.27) Theorem 6.86 follows immediately from Theorem 6.83.

Remark. The significance of Theorem 6.86 lies in the fact that it characterizes any solution of the Maximum Problem 5.11 (for $n = 2$) in terms of an algebraic condition which is very easy to check.

Note. A more precise determination of the values of ζ for which (6.85) and (6.87) hold in the ambiguous case of (3.9) will be made in Chapter 7 (Theorems 7.24 and 7.26).

CHAPTER 7

FURTHER NECESSARY CONDITIONS FOR ϵ-MINIMAL MATRICES

According to Theorem 5.13 all ϵ-minimal matrices are determined by solutions to the Maximum Problem 5.11. Therefore a necessary condition on a unitary matrix solving this maximum problem will, indirectly, be a necessary condition on an ϵ-minimal matrix. By working through the Maximum Problem 5.11 we shall be able to derive some additional necessary conditions in the present chapter.

The identity matrix I solves the Maximum Problem 5.11 for $B \in \mathcal{M}_n$ if and only if B satisfies the condition

$$\epsilon^2(dg(B)) = \max_{U \in \mathcal{U}_n} \epsilon^2(dg(UBU^*)) \quad . \tag{7.1}$$

If B satisfies (7.1) then, by Theorem 5.15, $dg(B)$ is ϵ-minimal for B and, by Theorem 4.10, $dg(B)$ satisfies the necessary condition (4.12), i.e.,

$$dg(B)B^* - B^* dg(B) + dg(B^*)B - B\, dg(B^*) = 0 \quad . \tag{7.2}$$

Letting $B = (b_{ij})$ we can express the last equation in terms of the elements of B as follows

$$\bar{b}_{ji}(b_{ii} - b_{jj}) + b_{ij}(\bar{b}_{ii} - \bar{b}_{jj}) = 0 \quad (1 \leq i, j \leq n) \quad . \tag{7.3}$$

In like manner we find from Theorem 4.73 that there is a hermitian matrix H such that

$$dg(B) + dg(B)H - H\,dg(B) = B = dg(B) + offdg(B)$$

or

$$offdg(B) = dg(B)H - H\,dg(B) \quad .$$

Writing $H = (h_{ij})$ we can express the last equation in terms of the elements of B and H as follows

$$b_{ij} = h_{ij}(b_{ii} - b_{jj}) \qquad (i \neq j) \quad .$$

Since H is hermitian we have $h_{ji} = \bar{h}_{ij}$ so $b_{ji} = \bar{h}_{ij}(b_{jj} - b_{ii}) = -\bar{h}_{ij}(b_{ii} - b_{jj})$. Therefore we find that every 2 by 2 principal submatrix B_{ij} of B is of the form

$$(7.4) \qquad B_{ij} = \begin{pmatrix} b_{ii} & b_{ij} \\ b_{ji} & b_{jj} \end{pmatrix} = \begin{pmatrix} b_{ii} & h_{ij}(b_{ii}-b_{jj}) \\ -\bar{h}_{ij}(b_{ii}-b_{jj}) & b_{jj} \end{pmatrix} \quad (1 \leq i < j \leq n).$$

Consider now unitary matrices $U = (u_{rc}) \in \mathcal{U}_n$ of the following special type. For any pair (i,j) of row and column indices satisfying $1 \leq i < j \leq n$ we let the 2 by 2 principal submatrix

$$U_{ij} = \begin{pmatrix} u_{ii} & u_{ij} \\ u_{ji} & u_{jj} \end{pmatrix}$$

be unrestricted (except for the requirement $U \in \mathcal{U}_n$) and specify that all other elements of U satisfy

$$u_{rc} = \delta_{rc} = \begin{cases} 1 \ , & r = c \\ 0 \ , & r \neq c \end{cases} .$$

One may easily verify that U is unitary if and only if U_{ij} is unitary. If B satisfies (7.1) then certainly $\epsilon^2(dg(B)) \geq \epsilon^2(dg(UBU^*))$ holds for all U of the special type just described. The effect of the transformation UBU^* on the submatrix B_{ij} is that of replacing B_{ij} by $U_{ij} B_{ij} U_{ij}^*$; therefore, since U_{ij} can be any matrix in $\mathcal{U}_2$, we see that B_{ij} itself satisfies (7.1) with n replaced by 2:

$$\epsilon^2(dg(B_{ij})) = \max_{U \in \mathcal{U}_2} \epsilon^2(dg(UB_{ij}U^*)) \qquad (1 \leq i < j \leq n) .$$

We now find from Theorem 6.83 that

$$B_{ij} + \zeta_{ij} B_{ij}^* \in \mathcal{D}_2 \qquad (1 \leq i < j \leq n) \tag{7.5}$$

holds for some $\zeta_{ij} \in \zeta(B_{ij})$. Expressed in terms of elements, the condition (7.5) states that

$$b_{ij} + \zeta_{ij}\bar{b}_{ji} = b_{ji} + \zeta_{ij}\bar{b}_{ij} = 0 \qquad (1 \leq i < j \leq n) . \tag{7.6}$$

Since $|\zeta_{ij}| = 1$ the last equation implies $|b_{ij}| = |b_{ji}|$.

We can obtain some information about the complex Lagrange multipliers h_{ij} in (7.4) as follows. From (7.3) and (7.4) we obtain

$$-\bar{h}_{ij}(b_{ii} - b_{jj})^2 + h_{ij} |b_{ii} - b_{jj}|^2 = 0$$

so, unless $b_{ii} = b_{jj}$ (in which case the value of h_{ij} in (7.4) does not matter at all) or $h_{ij} = 0$, we have

$$\frac{h_{ij}^2}{|h_{ij}|^2} = \frac{(b_{ii} - b_{jj})^2}{|b_{ii} - b_{jj}|^2}$$

or

$$\frac{h_{ij}}{|h_{ij}|} = \pm \frac{b_{ii} - b_{jj}}{|b_{ii} - b_{jj}|} .$$

Straightforward computations using (7.4) show that

$$(7.7) \qquad \mathrm{tr}(B_{ij}^2) - \tfrac{1}{2}(\mathrm{tr}(B_{ij}))^2 = (\tfrac{1}{2} - 2|h_{ij}|^2)(b_{ii} - b_{jj})^2 .$$

If (7.7) does not vanish we see from (3.9) that ζ_{ij} in (7.5) must have the value

$$(7.8) \qquad \zeta_{ij} = s_{ij} \frac{b_{ii} - b_{jj}}{\bar{b}_{ii} - \bar{b}_{jj}}$$

where

$$s_{ij} = \mathrm{sgn}(\tfrac{1}{2} - 2|h_{ij}|^2)$$

and where for $a \in \underline{R}$ $\mathrm{sgn}(a)$ denotes the sign of a:

$$\mathrm{sgn}(a) = \begin{cases} +1 & \text{if } a > 0 \\ -1 & \text{if } a < 0 \end{cases} .$$

Using (7.4) and (7.8) we obtain

$$b_{ij} + \zeta_{ij}\bar{b}_{ji} = h_{ij}(1 - s_{ij})(b_{ii} - b_{jj}) \quad ,$$

$$b_{ji} + \zeta_{ij}\bar{b}_{ij} = \bar{h}_{ij}(s_{ij} - 1)(b_{ii} - b_{jj})$$

whence (7.6) holds if and only if either $h_{ij} = 0$ or $s_{ij} = 1$. Clearly $s_{ij} = 1$ holds if and only if $(1/2) - 2|h_{ij}|^2 > 0$, i.e.,

$$|h_{ij}| < \tfrac{1}{2} \quad . \tag{7.9}$$

The other possibility is that $h_{ij} = 0$ in which case (7.9) still holds. In the ambiguous case of (3.9) where (7.7) vanishes we obtain

$$b_{ij} + \zeta_{ij}\bar{b}_{ji} = h_{ij}[(b_{ii} - b_{jj}) - \zeta_{ij}\overline{(b_{ii} - b_{jj})}] \quad , \tag{7.10}$$

$$b_{ji} + \zeta_{ij}\bar{b}_{ij} = -\bar{h}_{ij}[(b_{ii} - b_{jj}) - \zeta_{ij}\overline{(b_{ii} - b_{jj})}] \quad . \tag{7.11}$$

If we disregard the uninteresting case $b_{ii} = b_{jj}$ then obviously (7.7) vanishes if and only if $|h_{ij}| = (1/2)$. Thus, from (7.10) and (7.11), (7.6) holds if and only if

$$\zeta_{ij} = \frac{b_{ii} - b_{jj}}{\bar{b}_{ii} - \bar{b}_{jj}} \quad .$$

We now summarize the above results in the following

7.12 Theorem. Let $B = (b_{ij}) \in \mathcal{M}_n$ satisfy the condition (7.1). For each pair of indices (i,j) satisfying $1 \leq i < j \leq n$ the following statements are true.

(7.13) $$b_{ij}\overline{(b_{ii} - b_{jj})} = \overline{b}_{ji}(b_{ii} - b_{jj}) \ .$$

(7.14) $$|b_{ij}| = |b_{ji}| \ .$$

There exists a complex number (= complex Lagrange multiplier) h_{ij} such that

(7.15) $$B_{ij} = \begin{pmatrix} b_{ii} & b_{ij} \\ b_{ji} & b_{jj} \end{pmatrix} = \begin{pmatrix} b_{ii} & h_{ij}(b_{ii}-b_{jj}) \\ -\overline{h}_{ij}(b_{ii}-b_{jj}) & b_{jj} \end{pmatrix} \ .$$

If $b_{ii} = b_{jj}$ then $b_{ij} = b_{ji} = 0$. If $b_{ii} \neq b_{jj}$ then

(7.16) $$|h_{ij}| \leq \frac{1}{2} \ .$$

More precisely, if

(7.17) $$\mathrm{tr}(B_{ij}^2) - \tfrac{1}{2}(\mathrm{tr}(B_{ij}))^2 \neq 0 \ ,$$

then

(7.18) $$|h_{ij}| < \frac{1}{2}$$

and, if

(7.19) $$\mathrm{tr}(B_{ij}^2) - \tfrac{1}{2}(\mathrm{tr}(B_{ij}))^2 = 0$$

and $b_{ii} \neq b_{jj}$, then

(7.20) $$|h_{ij}| = \frac{1}{2} \quad .$$

If $h_{ij} \neq 0$ and $b_{ii} \neq b_{jj}$ then

(7.21) $$\frac{h_{ij}}{|h_{ij}|} = \pm \frac{b_{ii} - b_{jj}}{|b_{ii} - b_{jj}|} \quad .$$

Furthermore, if $b_{ii} \neq b_{jj}$, there is a uniquely determined complex number ζ_{ij} satisfying $|\zeta_{ij}| = 1$ such that

(7.22) $$b_{ij} + \zeta_{ij}\bar{b}_{ji} = b_{ji} + \zeta_{ij}\bar{b}_{ij} = 0 \quad .$$

If (7.17) holds $\zeta_{ij} \in \zeta(B_{ij})$ where the set $\zeta(B_{ij})$ is defined by (3.9). If (7.19) holds then

(7.23) $$\zeta_{ij} = \frac{b_{ii} - b_{jj}}{\bar{b}_{ii} - \bar{b}_{jj}} \quad .$$

Specializing Theorem 7.12 to the case $n = 2$ we can improve Theorems 6.83 and 6.86 as follows.

<u>7.24 Theorem</u>. Let $A = (a_{ij}) \in \mathcal{M}_2$ satisfy (6.84). If $a_{11} = a_{22}$ then $A = a_{11}I$ and (6.85) holds for all $\zeta \in \zeta(A)$. If $a_{11} \neq a_{22}$ and and if $tr(A^2) - (1/2)(tr(A))^2 = 0$ then (6.85) holds only for

(7.25) $$\zeta = \frac{a_{11} - a_{22}}{\bar{a}_{11} - \bar{a}_{22}} \quad .$$

7.26 Theorem. Let $A \in \mathcal{M}_2$, let U_o solve the Maximum Problem 5.11 for A, and let $B = (b_{ij}) = U_o A U_o^*$. If $b_{11} = b_{22}$ then $B = b_{11} I$ and (6.87) holds for all $\zeta \in \zeta(A)$. If $b_{11} \neq b_{22}$ and if $\mathrm{tr}(A^2) - (1/2)(\mathrm{tr}(A))^2 = 0$ then (6.87) holds only for

$$\zeta = \frac{b_{11} - b_{22}}{\bar{b}_{11} - \bar{b}_{22}} \quad . \tag{7.27}$$

7.28 Theorem. Let $A \in \mathcal{M}_n$, let U_o solve the Maximum Problem 5.11 for A, and let $B = (b_{ij}) = U_o A U_o^*$. Then all the conclusions of Theorem 7.12 hold.

Proof. If U_o solves the Maximum Problem 5.11 for A then, by an argument similar to that used in the proof of Theorem 6.86, B satisfies (7.1).

CHAPTER 8

COUNTEREXAMPLES TO MIRSKY'S CONJECTURE

In the present chapter we shall present some selected examples which shed some light on the distance problem of finding $d_\nu(A)$ and which also show that Mirsky's Conjecture 1.35 is incorrect in a number of instances.

Consider first the following class of matrices of order n. We define $A = (a_{k\ell}) \in \mathcal{M}_n$ as follows: $a_{k\ell} = 0$ $(1 \leq k, \ell \leq n)$ with the following exceptions

$$a_{k,k+1} = e^{i\theta_k} \qquad (k = 1,2, \ldots, n-1) \tag{8.1}$$

where the θ_k are arbitrary real numbers. Now define $N_\varphi = (b_{k\ell}) \in \mathcal{M}_n$ by

$$b_{n1} = e^{i\varphi}, \quad b_{k\ell} = a_{k\ell} \text{ otherwise} \tag{8.2}$$

where φ is any real number. Clearly the rows and columns of N_φ form orthonormal sets so that N_φ is unitary (for all $\theta_1, \ldots, \theta_{n-1}, \varphi \in \underline{R}$) and therefore normal. (Actually N_φ is the product of a diagonal unitary matrix with a permutation matrix.)

For all $\theta_1, \ldots, \theta_{n-1}$ we have $\Omega(A) = 0$ and $\epsilon^2(A) = n - 1$ so

$$\epsilon^2(A) - \epsilon^2(\Omega(A)) = n - 1 \ . \tag{8.3}$$

An elementary calculation shows that $f(c) = \epsilon^2(A - cN_\varphi)$, where c is real, assumes its absolute minimum if and only if $c = (n-1)/n$.

Furthermore, by (8.3)

$$\epsilon^2(A - \tfrac{n-1}{n} N_\varphi) = \tfrac{n-1}{n} = \tfrac{1}{n}(\epsilon^2(A) - \epsilon^2(\Omega(A))) \tag{8.4}$$

holds for all $\theta_1, \ldots, \theta_{n-1}, \varphi \in \underline{R}$. This proves (cf. Theorem 2.16).

8.5 Theorem. Mirsky's Conjecture 1.35 is false for $\nu = \epsilon$ and $n \geq 3$. There exist matrices A in $\mathcal{M}_n$ such that

$$p_{\epsilon,n}(A) \leq \tfrac{1}{n} \qquad (n \geq 3) \tag{8.6}$$

where $p_{\epsilon,n}(A)$ is defined (for $\nu = \epsilon$) by (2.14).

Note. An example of order 3 similar to the pair A, $(n-1)/n\, N_\varphi$ in the case $n = 3$ is due to Eberlein [6]. She also obtains counter-examples to Mirsky's conjecture for $n \geq 4$ by bordering her 3 by 3 example with zeros.

Consider next the class of 2 by 2 matrices of the form

$$A = \begin{pmatrix} \lambda & m \\ 0 & 0 \end{pmatrix} \quad \text{where } \lambda, m \in \underline{C}, \quad m \neq 0 . \tag{8.7}$$

Let N be any ϵ-minimal matrix for A, that is, any one of the matrices

$$\tfrac{1}{2}\begin{pmatrix} 2\lambda & m \\ \zeta\bar{m} & 0 \end{pmatrix} \qquad (\zeta \in \zeta(A)) \tag{8.8}$$

where $\zeta(A)$ is defined by (3.9) (cf. Theorem 6.24). One finds easily

$$AA^* = \operatorname{diag}(|\lambda|^2 + |m|^2, 0), \quad (A - N)(A - N)^* = \tfrac{1}{4}\operatorname{diag}(|m|^2, |m|^2)$$

whence, by Definition 1.10, the singular values of A are

$$\sqrt{|\lambda|^2 + |m|^2} \quad , \quad 0$$

and the singular values of $A - N$ are

$$\tfrac{1}{2}\,|m| \quad , \quad \tfrac{1}{2}\,|m| \quad .$$

Using the defining formulas (1.13) - (1.15) for the unitarily invariant norms ν_p, we find that the following hold for all p $(1 \leq p \leq \infty)$:

$$\nu_p(A) = \left[\left(\sqrt{|\lambda|^2 + |m|^2}\right)^p\right]^{1/p} = \sqrt{|\lambda|^2 + |m|^2} \quad ,$$

$$\nu_p(\Omega(A)) = |\lambda| \quad ,$$

$$\nu_p(A - N) = \left[2\left(\frac{|m|}{2}\right)^p\right]^{1/p} = 2^{1/p-1}\,|m|$$

so

$$\nu_p^2(A) - \nu_p^2(\Omega(A)) = |m|^2 \quad ,$$

$$\nu_p^2(A - N) = 2^{2/p-2}\,|m|^2$$

or

$$\nu_p^2(A - N) = 2^{2/p-2}\left(\nu_p^2(A) - \nu_p^2(\Omega(A))\right) \quad .$$

Now $2^{2/p-2} < (1/2)$ if and only if $2^{2/p} < 2$ and this happens if and only if $2/p < 1$, i.e., $p > 2$. Thus the pair A, N provides a

counterexample to Mirsky's conjecture for $n = 2$ and for all $p > 2$. By bordering the matrices A, N with zeros and by carrying out the relevant computations with (1.13) - (1.15) for higher values of n, we then obtain counterexamples to the conjecture for all $n \geq 3$. We have proved

8.9 Theorem. Mirsky's Conjecture 1.35 is false for $\nu = \nu_p$ $(2 < p \leq \infty)$ and $n \geq 2$ where ν_p is given by (1.15).

Remark. Since $\nu_\infty = \sigma$ we have again proved the second statement of Theorem 2.19.

CHAPTER 9

THE FIELD OF VALUES AND EIGENVALUES OF ϵ-MINIMAL MATRICES

The <u>field of values</u> (or <u>numerical range</u>) of a matrix $A \in \mathcal{M}_n$ is defined to be the following set of complex numbers.

$$F(A) = \{(Ax,x);\ (x,x) = 1 \text{ and } x \text{ is a complex column vector}\} . \tag{9.1}$$

The following are known facts concerning $F(A)$:

<u>9.2 Theorem</u>. (Toeplitz [23] and Hausdorff [12]) For any $A \in \mathcal{M}_n$ $F(A)$ is a closed, bounded, connected, convex subset of $\underline{C}$.

<u>9.3 Theorem</u>. (Toeplitz [23]) Let $A \in \mathcal{M}_n$. All eigenvalues of A are in $F(A)$. If $A \in \mathcal{N}_n$ then $F(A)$ coincides with the convex hull $C(A)$ of the eigenvalues of A.

<u>9.4 Theorem</u>. (Hausdorff [12]) The field of values is invariant under a unitary similarity transformation:

$$F(A) = F(UAU^*) \text{ where } A \in \mathcal{M}_n,\ U \in \mathcal{U}_n . \tag{9.5}$$

A simple consequence of Theorems 9.2 and 9.3 is the following

<u>9.6 Theorem</u>. Let $A \in \mathcal{M}_n$. Then the convex hull $C(A)$ of the eigenvalues of A is contained in $F(A)$:

$$C(A) \subset F(A) . \tag{9.7}$$

An elementary computation using Theorem 9.4 and results of Toeplitz [23] and Donoghue [5] yields the following result whose proof is omitted.

9.8 Theorem. Let $A \in \mathcal{M}_2$ and let

$$VAV^* = \begin{pmatrix} \lambda_1 & m \\ 0 & \lambda_2 \end{pmatrix} \qquad (V \in \mathcal{U}_2) \tag{9.9}$$

be a Schur triangular form for A. If A is not normal, i.e., if $m \neq 0$, and if $\lambda_1 \neq \lambda_2$, then F(A) is the interior and boundary of an ellipse whose foci are λ_1 and λ_2, whose minor axis has length $|m|$, and whose major axis has length $(|m|^2 + |\lambda_1 - \lambda_2|^2)^{1/2}$; if $\lambda_1 = \lambda_2$ then F(A) is the interior and boundary of a circle with center at λ_1 and diameter $|m|$. If A is normal, i.e., if $m = 0$, and if $\lambda_1 \neq \lambda_2$, then F(A) is the straight line segment connnecting λ_1 and λ_2; if $\lambda_1 = \lambda_2$, then F(A) reduces to a single point, namely λ_1.

Our first objective in the present chapter is to prove

9.10 Theorem. Let $A \in \mathcal{M}_2$, $A \notin \mathcal{N}_2$ and let (9.9) be a Schur form for A. If $\lambda_1 \neq \lambda_2$ then the eigenvalues of the (unique) ϵ-minimal matrix for A are the end points of the major axis of the ellipse which is the boundary of F(A). If $\lambda_1 = \lambda_2$ and if $N_\theta(A)$ (see the note following Definition 3.8) is one of the ϵ-minimal matrices for A, then the eigenvalues of $N_\theta(A)$ are

$$\lambda_1 \pm \tfrac{1}{2} |m| e^{i\theta/2} \quad ; \tag{9.11}$$

given any diameter of the circle which is the boundary of F(A), there is a $\theta \in \underline{R}$ such that the eigenvalues of $N_\theta(A)$ are the endpoints of that diameter.

Proof. In view of Theorem 9.4, (6.27), and (6.28) we need only prove the result for some unitary transform of A. We shall use the form (9.9) for this purpose. Consider first the case $\lambda_1 = \lambda_2$. Then by putting $a_1 = \lambda_1$, $a_2 = m$ in (6.29) and (6.40) we see that

$$N_\theta(VAV^*) = \begin{pmatrix} \lambda_1 & \frac{1}{2} m \\ \frac{1}{2} \bar{m} e^{i\theta} & \lambda_1 \end{pmatrix} ;$$

furthermore, by Theorem 5.13, the eigenvalues of N_θ are the diagonal elements of (6.40), namely

$$\lambda_1 \pm \frac{1}{2} \mu_{1\theta} m \quad , \tag{9.12}$$

where, by (6.35), $(\mu_{1\theta} m)^2 = |m|^2 \exp(i\theta)$ or

$$\mu_{1\theta} m = \pm |m| e^{i\theta/2} \quad . \tag{9.13}$$

Whatever sign is chosen in (9.13) we see from (9.12) that the eigenvalues of N_θ are given by (9.11). By Theorem 9.8 the boundary of $F(A)$ is a circle with center λ_1 and radius $(1/2)|m|$. As θ increases from 0 to 2π it is clear from (9.11) that the eigenvalues of N_θ are the endpoints of a diameter of that circle which rotates through an angle of π. That is, every diameter of the circle which is the boundary of $F(A)$ is included as one of those which can occur as the line segment connecting the eigenvalues of some ϵ-minimal matrix for A.

We now consider the case $\lambda_1 \neq \lambda_2$. Straightforward computations

based on Definition 3.8 show that

$$(9.14) \qquad N(VAV^*) = \begin{pmatrix} \lambda_1 & \frac{1}{2}m \\ \frac{1}{2}\zeta\bar{m} & \lambda_2 \end{pmatrix}, \quad \text{where} \quad \zeta = \frac{\lambda_1 - \lambda_2}{\bar{\lambda}_1 - \bar{\lambda}_2}$$

and, according to Theorems 6.24 and 6.79, this is the only ϵ-minimal matrix for (9.9). The eigenvalues μ_1, μ_2 of $N(VAV^*)$ are the roots of the quadratic equation $\mu^2 - (\lambda_1 + \lambda_2)\mu + \lambda_1\lambda_2 - (1/4)|m|^2\zeta = 0$. Using the quadratic formula and the expression for ζ in (9.14) we find

$$(9.15) \qquad \begin{cases} \mu_1 = \frac{1}{2}(\lambda_1 + \lambda_2) + \frac{1}{2}\sqrt{\zeta}\,\sqrt{|m|^2 + |\lambda_1 - \lambda_2|^2}\,, \\ \mu_2 = \frac{1}{2}(\lambda_1 + \lambda_2) - \frac{1}{2}\sqrt{\zeta}\,\sqrt{|m|^2 + |\lambda_1 - \lambda_2|^2}\,. \end{cases}$$

Now

$$\zeta = \frac{\lambda_1 - \lambda_2}{\bar{\lambda}_1 - \bar{\lambda}_2} = \frac{(\lambda_1 - \lambda_2)^2}{|\lambda_1 - \lambda_2|^2}$$

whence

$$(9.16) \qquad \arg(\sqrt{\zeta}) = \arg(\lambda_1 - \lambda_2)\,.$$

According to Theorem 9.8 the boundary of $F(A)$ is an ellipse with foci λ_1, λ_2, center $(1/2)(\lambda_1 + \lambda_2)$ and major axis of length $(|m|^2 + |\lambda_1 - \lambda_2|^2)^{1/2}$. From (9.15) and (9.16) we find

$$(9.17) \quad \arg(\lambda_1 - \tfrac{1}{2}(\lambda_1 + \lambda_2)) = \arg(\lambda_1 - \lambda_2) = \arg(\mu_1 - \tfrac{1}{2}(\lambda_1 + \lambda_2))\,,$$

BLANK PAGE

(9.18) $\arg(\lambda_2 - \frac{1}{2}(\lambda_1 + \lambda_2)) = \arg[-(\lambda_1 - \lambda_2)] = \arg(\mu_2 - \frac{1}{2}(\lambda_1 + \lambda_2))$.

Relations (9.17) and (9.18) show that λ_1, λ_2, μ_1, and μ_2 all lie on the same straight line through the center $(1/2)(\lambda_1 + \lambda_2)$ of the ellipse. Since λ_1 and λ_2 lie on the major axis, since $|\zeta| = 1$, and since the length of the major axis is $(|m|^2 + |\lambda_1 - \lambda_2|^2)^{1/2}$, we see from (9.15) that μ_1 and μ_2 are the endpoints of the major axis. This completes the proof of Theorem 9.10.

9.19 Corollary. Let $A \in \mathcal{N}_2$, $A \notin \mathcal{N}_2$. If $F(A)$ is a circular disk, then given any diameter of that disk there is a $\theta \in \underline{R}$ such that $F(N_\theta(A))$ coincides with that diameter. If $F(A)$ is the interior and boundary of an ellipse (not a circle) then $F(N(A))$ coincides with the major axis of that ellipse.

Proof. This follows immediately from Theorems 9.8 and 9.10, since every matrix in $M(A)$ is normal.

9.20 Corollary. Let $A \in \mathcal{N}_2$, $A \notin \mathcal{N}_2$ and let N_o be ϵ-minimal for A. Then the eigenvalues μ_1, μ_2 of N_o are extreme points of $F(A)$. Furthermore $|\mu_1 - \mu_2| = \operatorname{diam}(F(A)) = \sup_{z,w \in F(A)} |z - w|$.

9.21 Theorem. Let $A \in \mathcal{N}_n$ and let N_o be ϵ-minimal for A. Then every eigenvalue of N_o belongs to $F(A)$ and

(9.22) $$F(N_o) \subset F(A) \quad .$$

Proof. Let $N_o = U_o^* D_o U_o$ where $U_o \in \mathcal{U}_n$ and $D_o \in \mathcal{D}_n$. According to Theorem 5.13 the eigenvalues of N_o are the diagonal elements of the matrix $U_o A U_o^*$. If u_k is the k-th column of U_o^* then the k-th

diagonal element of $U_o A U_o^*$ is given by (Au_k, u_k) and, by (9.1), this is in $F(A)$. Since $F(A)$ is convex and since $F(N_o)$ is the convex hull of the eigenvalues of N_o, we obtain immediately (9.22).

CHAPTER 10

A GENERALIZATION OF THE JACOBI AND GOLDSTINE-HORWITZ METHODS

Long ago Jacobi [14] devised a method for diagonalizing a real symmetric matrix. The method utilized coordinate-plane rotations and was essentially dependent only on an elementary technique for diagonalizing a real symmetric matrix of order 2 using an orthogonal similarity transformation. Since 1950 Jacobi's method has been extensively studied and generalized (see e.g., [3], [4]). In the present chapter we shall describe still another generalization of Jacobi's method which amounts to a computational technique for solving the Maximum Problem 5.11. This new technique also generalizes and simplifies a method devised by Goldstine and Horwitz [10].

Consider the following computational algorithm. Let $A \in \mathcal{M}_n$ and set $A_0 = A$. One calculates a sequence of matrices $A_1, A_2, \ldots,$ $A_k = (a_{rc}^{(k)}), \ldots$ which are unitarily similar to A by the recurrence relation

$$A_{k+1} = U_k A_k U_k^* \qquad (k = 0,1,2, \ldots) \quad . \tag{10.1}$$

The $U_k = (u_{rc}^{(k)})$ are special unitary matrices of order n. For every value of k there is specified a pair $\varkappa_k = (i_k, j_k) = (i,j)$ of indices (we omit the subscript k in the sequel for notational simplicity) satisfying $1 \leq i < j \leq n$ such that the 2 by 2 matrix

(10.2) $$V_k = \begin{pmatrix} u_{ii}^{(k)} & u_{ij}^{(k)} \\ u_{ji}^{(k)} & u_{jj}^{(k)} \end{pmatrix} ,$$

which is a principal submatrix of U_k, is unitary. All other elements of U_k satisfy

$$u_{rc}^{(k)} = \delta_{rc} = \begin{cases} 1 , & r = c \\ 0 , & r \neq c \end{cases} .$$

The matrices U_k are completely determined by the pairs π_k and the 2 by 2 unitary matrices V_k. We shall always take V_k to be a matrix of the form (6.12).

Any set of rules for choosing the sequence $\{U_k\}$ will be called a method of Jacobi type. The following example which is defined only for $A \in \mathcal{H}_n$ and which is a straightforward extension of Jacobi's original method to hermitian matrices will be referred to as the classical Jacobi method (cf. [4]). Here one chooses π_k such that

(10.3) $$|a_{ij}^{(k)}| = \max_{r \neq c} |a_{rc}^{(k)}|$$

and chooses V_k to be a matrix of the form (6.12) such that

(10.4) $$a_{ij}^{(k+1)} = 0 .$$

The new generalization of Jacobi's method which was announced in the first paragraph of this chapter is a method of Jacobi type with the

following set of rules for determining the U_k of (10.1). Let

$$A_{rc}^{(k)} = \begin{pmatrix} a_{rr}^{(k)} & a_{rc}^{(k)} \\ a_{cr}^{(k)} & a_{cc}^{(k)} \end{pmatrix} \qquad (1 \le r < c \le n) \tag{10.5}$$

and define

$$\Delta_k(r,c) = \max_{U \in \mathcal{U}_2} \epsilon^2(\mathrm{dg}(UA_{rc}^{(k)}U^*)) - \epsilon^2(\mathrm{dg}(A_{rc}^{(k)})) \ . \tag{10.6}$$

Choose s_k so that

$$\Delta_k(i,j) = \max_{1 \le r < c \le n} \Delta_k(r,c) \tag{10.7}$$

and choose V_k to be a matrix of the form (6.12) which solves the Maximum Problem 5.11 for $A_{ij}^{(k)}$. We show in the next paragraph how to calculate easily $\Delta_k(r,c)$ and V_k.

By Theorems 5.13 and 6.24 a unitary matrix U_o solves the Maximum Problem 5.11 for $A \in \mathcal{M}_2$ if and only if $U_o N_\zeta(A) U_o^*$ is diagonal and, by (3.10), that happens if and only if $U_o(A + \zeta A^*)U_o^*$ is diagonal where $\zeta \in \zeta(A)$. Therefore all we need to do in order to solve the Maximum Problem 5.11 for any of the submatrices $A_{rc}^{(k)}$ is to find a $\mu_{rc}^{(k)} \in \underline{C}$ such that

$$W(\mu_{rc}^{(k)})(A_{rc}^{(k)} + \zeta\, A_{rc}^{(k)})W^*(\mu_{rc}^{(k)}) \in \mathcal{D}_2 \tag{10.8}$$

where ζ is any number in $\zeta(A_{rc}^{(k)})$. Thus from (10.6) we have

$$(10.9) \qquad \Delta_k(r,c) = \epsilon^2(dg(W(\mu_{rc}^{(k)})A_{rc}^{(k)}W^*(\mu_{rc}^{(k)}))) - \epsilon^2(dg(A_{rc}^{(k)}))$$

and we set

$$(10.10) \qquad V_k = W(\mu_{ij}^{(k)}) \quad .$$

We now show that if $A \in \mathcal{H}_n$ the set of rules (10.7) and (10.10) coincide respectively with (10.3) and (10.4). This will show that our new method is a generalization of the classical Jacobi method. If $A \in \mathcal{H}_n$ then, from (10.1), $A_k \in \mathcal{H}_n$ $(k = 0,1,2, \ldots)$ which implies $A_{rc}^{(k)} \in \mathcal{H}_2$ for $1 \leq r < c \leq n$ and for all k. Since $A_{rc}^{(k)}$ is hermitian it can be diagonalized by a unitary transformation and the value of the maximum in (10.6) is given by $\epsilon^2(A_{rc}^{(k)})$. Thus we find that

$$\Delta_k(r,c) = 2|a_{rc}^{(k)}| \qquad (1 \leq r < c \leq n)$$

whence the rule (10.7) reduces to (10.3). Furthermore, by Corollary 6.78, $M(A_{rc}^{(k)}) = |A_{rc}^{(k)}|$ whence the V_k given by (10.10) diagonalizes $A_{ij}^{(k)}$ i.e., (10.4) holds.

Since the classical Jacobi method diagonalizes any $A \in \mathcal{H}_n$ we have proved that the method of Jacobi type (10.7), (10.10) will diagonalize any hermitian matrix. By a similar argument we could show that the method will also diagonalize any skew-hermitian matrix.

In [10] Goldstine and Horwitz devised a method of Jacobi type which was applicable to any normal matrix $A \in \mathcal{M}_n$. At each stage they sought, by a very complicated procedure, to determine a V_k of the form (6.12)

such that

$$\tau^2(A_{ij}^{(k+1)}) = \epsilon^2(\mathrm{offdg}(V_k A_{ij}^{(k)} V_k^*))$$

was minimized. It is easy to see that this is the same as determining V_k so that $\epsilon^2(\mathrm{dg}(V_k A_{ij}^{(k)} V_k^*))$ is maximized i.e., solving the Maximum Problem 5.11 for $A_{ij}^{(k)}$. Thus our technique provides a simple solution to the problem studied by Goldstine and Horwitz and at the same time makes its application to arbitrary $A \in \mathcal{M}_n$ meaningful in the context of solving the Maximum Problem 5.11 for A.

Of course the main question here is whether or not the method of Jacobi type (10.7), (10.10) will actually solve the Maximum Problem 5.11 for any $A \in \mathcal{M}_n$, i.e., whether or not $\epsilon^2(\mathrm{dg}(A_k))$ converges to the maximum (5.12). A related question is whether or not the infinite product $\cdots U_k U_{k-1} \cdots U_1 U_0$ of unitary matrices converges to a unitary matrix which solves the Maximum Problem 5.11 (cf. [4]). If the answers to these questions were affirmative then we would have a constructive method of computing ϵ-minimal matrices for any $A \in \mathcal{M}_n$. If the answer to the first question were affirmative then, by Theorem 5.4, we would have a constructive method of computing $d_\epsilon(A)$. These convergence questions appear to be rather difficult and we content ourselves here by proving only the following

10.11 Lemma. Let $A \in \mathcal{M}_n$ and let the matrices U_k of (10.1) be determined for each k by the rules (10.7) and (10.10). Then

$$\Delta_k(i,j) \to 0 \quad \text{as } k \to \infty \ . \tag{10.12}$$

Proof. Let $\sigma_k = \epsilon^2(dg(A_k))$ for $k = 0,1,2, \ldots$. Since each A_k is unitarily similar to A, we have

$$\sigma_k \leq \epsilon^2(A) \qquad (k = 0,1,2, \ldots) \quad . \tag{10.13}$$

The only diagonal elements of A_k affected by the transformation (10.1) are the ones in the i-th and j-th rows. Moreover, since the transformation $A_{ij}^{(k+1)} = V_k A_{ij}^{(k)} V_k^*$ increases the function $\epsilon^2(dg(A_{ij}))$ by the amount $\Delta_k(i,j)$, we have

$$\sigma_{k+1} = \sigma_k + \Delta_k(i,j) \geq \sigma_k \qquad \text{for } k = 0,1,2, \ldots \quad . \tag{10.14}$$

The relations (10.13) and (10.14) show that (σ_k) is a monotonically increasing sequence of positive numbers which is bounded from above. Consequently $\sigma_k \to \sigma \leq \epsilon^2(A)$ as $k \to \infty$ and from (10.14) we obtain the desired conclusion (10.12).

BIBLIOGRAPHY

[1] L. V. Ahlfors, Complex Analysis, McGraw-Hill, New York, 1953.

[2] A. R. Amir-Moéz and A. Horn, "Singular values of a matrix," Amer. Math. Monthly, vol. 65, (1958), pp. 742-748.

[3] R. L. Causey, "Computing eigenvalues of non-hermitian matrices by methods of Jacobi type," J. Soc. Indust. Appl. Math., vol. 6, (1958), pp. 172-181.

[4] R. L. Causey and P. Henrici, "Convergence of approximate eigenvectors in Jacobi methods," Numerische Math., vol. 2, (1960), pp. 67-78.

[5] W. F. Donoghue, Jr., "On the numerical range of a bounded operator," Mich. Math. J., vol. 4, (1957), pp. 261-263.

[6] P. J. Eberlein, "On a conjecture of Mirsky," submitted for publication.

[7] D. K. Faddeev and V. N. Faddeeva, Computational Methods of Linear Algebra (translated from Russian by Robert C. Williams), W. H. Freeman and Co., San Francisco, 1963.

[8] Ky Fan, "Maximum properties and inequalities for eigenvalues of completely continuous operators," Proc. Nat. Acad. Sci. U.S.A., vol. 37, (1951), pp. 760-766.

[9] Ky Fan and A. J. Hoffman, "Some metric inequalities in the space of matrices," Proc. Amer. Math. Soc., vol. 6, (1951), pp. 760-766.

[10] H. H. Goldstine and L. P. Horwitz, "A procedure for the diagonalization of normal matrices," J. Assoc. Comput. Mach., vol. 6, (1959), pp. 176-195.

[11] P. R. Halmos, Finite-Dimensional Vector Spaces, 2nd Ed., Van Nostrand, New York, 1958.

[12] F. Hausdorff, "Der Wertvorrat einer Bilinearform," Math. Zeitschrift, vol. 3, (1919), pp. 314-316.

[13] E. Hille and R. Phillips, Functional Analysis and Semi-groups, Amer. Math. Soc. Colloq. Publ., vol. XXXI, New York, 1957.

[14] C. G. J. Jacobi, "Über ein leichtes Verfahren, die in der Theorie der Säkularstörungen vorkommenden Gleichungen numerisch aufzulösen," J. reine angew. Math., vol. 30, (1846), pp. 51-95.

[15] L. Mirsky, "Symmetric gauge functions and unitarily invariant norms," Quart. J. Math. Oxford (2), vol. 11, (1960), pp. 50-59.

[16] L. Mirsky, An Introduction to Linear Algebra, Oxford Univ. Press, London, 1955.

[17] F. D. Murnaghan, "On a convenient system of parameters for the unitary group," Proc. Nat. Acad. Sci. U.S.A., vol. 38, (1952), pp. 127-129.

[18] Z. Nehari, Introduction to Complex Analysis, Allyn and Bacon, Boston, 1961.

[19] J. von Neumann, "Some matrix-inequalities and metrization of matric space," Tomsk Univ. Review, vol. 1, (1937), pp. 286-300; Collected Works, vol. IV, p. 205.

[20] W. V. Parker, "The characteristic roots of matrices," Duke Math. J., vol. 12, (1945), pp. 519-526.

[21] S. Perlis, Theory of Matrices, Addison-Wesley Publishing Co., Inc., Cambridge, Mass., 1952.

[22] I. Schur, "Über die charakteristischen Wurzeln einer linearen Substitution mit einer Anwendung auf die Theorie der Integralgleichungen," Math. Annalen, vol. 66, (1904), pp. 488-510.

[23] O. Toeplitz, "Das algebraische Analogon zu einem Satze von Fejér," Math. Zeitschrift, vol. 2, (1918), pp. 187-197.

[24] H. Weyl, "Inequalities between the two kinds of eigenvalues of a linear transformation," Proc. Nat. Acad. Sci. U.S.A., vol. 35, (1949), pp. 408-411.

[25] J. Williamson, "A polar representation of singular matrices," Bull. Amer. Math. Soc., vol. 41, (1935), pp. 118-123.

[26] A. Wintner and F. D. Murnaghan, "On a polar representation of non-singular square matrices," Proc. Nat. Acad. Sci. U.S.A., vol. 17, (1931), pp. 676-678.