ON THE UNITARY COMPLETION OF A MATRIX

BY

M. Marcus and P. Greiner

I. Introduction

Suppose a certain set of entries in an n-square array are prescribed. We consider the following question: to determine some nontrivial necessary conditions under which the rest of the entries may be constructed so that the resulting matrix is unitary. For example, a trivial necessary condition is that the sum of the squares of the absolute values of the prescribed entries in any particular row or column is at most 1. Define a *diagonal* of an n-square array to be a set of positions $(i, \sigma(i))$, $i = 1, \cdots, n$, where σ is a permutation of $1, \cdots, n$. Two diagonals will overlap in k places if the corresponding permutations agree on k integers in $1, \cdots, n$, and we shall refer to two such diagonals as being *k-overlapping*. Our results will show for example that if A is a 9-square matrix whose entries in a fixed pair of nonoverlapping diagonals are all $\frac{2}{3} + \varepsilon$, $\varepsilon > 0$, then A cannot be completed to a unitary matrix. On the other hand if the absolute value of the sum of the elements in two nonoverlapping diagonals is to be no greater than 10, there exists a 9-square unitary matrix for which this value of the sum is taken on. This last statement becomes false however if 10 is replaced by $10 + \varepsilon$, $\varepsilon > 0$.

For the group of n-square unitary matrices we obtain the maximum and minimum over all pairs of k-overlapping diagonals d_1 and d_2 of the maximum absolute value of the sum of the entries in d_1 and d_2. Our main results are contained in the

Theorem. *Consider a fixed pair of k-overlapping diagonals of an n-square array, $1 \leqq k \leqq n$. Let s_k be the maximum taken over all n-square unitary matrices of the absolute value of the sum of the elements in the given pair of diagonals. Then*

(i) $s_k \leqq n$ *if* $n = k + 2$,

(ii) $s_k \leqq n - 4 + 2 \cot \pi/8$ *if* $n = k + 4$,

(iii) $s_k \leqq n + \alpha$ *if* $n = k + 3\alpha$,

$s_k \leqq n + \alpha - 5 + 2 \csc \pi/10$ *if* $n = k + 3\alpha + 5$,

$s_k \leqq n + \alpha - 7 + 2 \csc \pi/14$ *if* $n = k + 3\alpha + 7$,

(iv) $s_k \geqq n$ *if* $n = k + 2\alpha$,

$s_k \geqq n + 1$ *if* $n = k + 2\alpha + 3$.

Moreover for each of the bounds in (i)–(iv) *there exist a pair of k-overlapping diagonals and a unitary matrix for which the absolute value of the sum of the elements in these diagonals is the appropriate bound.*

In addition if $0 \leqq \mu \leqq s_k$, then there exists a unitary matrix such that the absolute value of the sum of the elements in the fixed pair of diagonals is μ.

Received December 21, 1959.

We first introduce some preliminary definitions and results. Then in Section II we shall consider the case of the two k-overlapping diagonals. In Sections III and IV we state and prove the necessary trigonometric inequalities to complete the proof of the above result. In Section V we indicate a further application of our methods.

Let O_n be the group of n-square unitary matrices over the complex numbers. Let S be any set of pairs (i, j), $1 \leqq i, j \leqq n$, and define the real valued function on O_n

$$g_S(A) = \left| \sum_{(i,j)\epsilon S} a_{ij} \right| . \tag{1.1}$$

The general problem is to determine the maximum value of $g_S(A)$ as A varies over O_n. Let H_S be the n-square matrix whose (i, j) element is the number of integers t such that (i, t) and (j, t) both belong to S. We have

LEMMA 1. *The matrix H_S is positive semidefinite, and*

$$0 \leqq g_S(A) \leqq \operatorname{tr}\left(\sqrt{H_S}\right) \qquad \textit{for} \quad A \epsilon O_n \tag{1.2}$$

where $\sqrt{\ }$ indicates the positive semidefinite determination of the square root. Every value between the two bounds is achievable by $g_S(A)$ for an appropriate unitary A.

Proof. Let P_S be the n-square matrix whose (i, j) element is 1 if $(j, i) \epsilon S$ and 0 otherwise. Then

$$g_S(A) = \left| \sum_{(i,j)\epsilon S} a_{ij} \right| = \left| \operatorname{tr}(P_S A) \right| . \tag{1.3}$$

This equality (1.3) follows immediately upon noting that the (t, t) element of $P_S A$ is the sum $\sum_s a_{st}$ for $(s, t) \epsilon S$. We next apply a result of von Neumann [2, 3] to conclude that

$$g_S(A) = \left| \operatorname{tr}(P_S A) \right| \leqq \operatorname{tr}\left(\sqrt{P_S P_S'}\right) = \operatorname{tr}\left(\sqrt{H_S}\right) \tag{1.4}$$

where A is any unitary matrix. For the (i, j) element of $P_S P_S'$ is precisely the number of columns in which rows i and j of P_S have a 1 in common, and thus $P_S P_S' = H_S$. We include a short proof of (1.4) for completeness. By the polar factorization theorem it is clear that we may assume

$$g_S(A) = \left| \operatorname{tr}(K_S A) \right|$$

where $K_S = \sqrt{H_S}$. Letting $x_1, \cdots, x_n$ be an orthonormal basis and setting $y_i = Ax_i$, $i = 1, \cdots, n$, we have

$$\begin{aligned} g_S(A) &= \left| \operatorname{tr}(K_S A) \right| = \left| \sum_{i=1}^n (K_S A x_i, x_i) \right| \\ &= \left| \sum_{i=1}^n (y_i, K_S x_i) \right| \leqq \sum_{i=1}^n \left| (y_i, K_S x_i) \right| \\ &\leqq \sum_{i=1}^n (K_S x_i, x_i)^{1/2} (K_S y_i, y_i)^{1/2} \\ &\leqq \left\{ \sum_{i=1}^n (K_S x_i, x_i) \right\}^{1/2} \left\{ \sum_{i=1}^n (K_S y_i, y_i) \right\}^{1/2} \\ &= \{\operatorname{tr}(K_S)\}^{1/2} \{\operatorname{tr}(K_S)\}^{1/2} = \operatorname{tr}\left(\sqrt{H_S}\right). \end{aligned}$$

To see that $g_S(O_n)$ is precisely the closed interval $[0, \operatorname{tr}(\sqrt{H_S})]$ we note the following facts. First, $g_S(A)$ is clearly continuous with respect to the usual distance function in O_n, $d(U, V) = (\sum_{i,j=1}^n |u_{ij} - v_{ij}|^2)^{1/2}$. Second, if U and V are in O_n, there exists a continuous one-parameter family $A(t) \in O_n$, $0 \leqq t \leqq 1$, such that $A(0) = U$ and $A(1) = V$; for let $S \in O_n$ be chosen so that $S^{-1}(U^{-1}V)S = \operatorname{diag}(e^{i\theta_1}, \cdots, e^{i\theta_n})$, and set

$$A(t) = US \operatorname{diag}(e^{i\theta_1 t}, \cdots, e^{i\theta_n t})S^{-1}.$$

Finally, $g_S(A) = 0$ for an appropriate $A \in O_n$; for $g_S(A) = |\operatorname{tr}(K_S A)|$, and by selecting $x_1, \cdots, x_n$ to be an orthonormal set of eigenvectors of K_S and choosing $A \in O_n$ such that $Ax_i = x_{i+1} \pmod{n}$, we have

$$g_S(A) = |\textstyle\sum_{i=1}^n (K_S A x_i, x_i)| = |\sum_{i=1}^n \lambda_i(x_{i+1}, x_i)| = 0.$$

II. Sums down pairs of diagonals

Let S be a set of pairs (i, j), $1 \leqq i, j \leqq n$, determined by two k-overlapping diagonals, $0 \leqq k \leqq n$. In this case P_S has precisely one entry 1 in each of exactly k rows and columns and precisely two entries 1 in each of the remaining $n - k$ rows and columns. Then

$$H_S = P_S P_S' \simeq (I_k \dotplus (P + Q))(I_k \dotplus (P' + Q')), \tag{2.1}$$

where in (2.1) $A \simeq B$ means $RAR' = B$ for some permutation matrix R, $\dotplus$ indicates direct sum, P and Q are both $(n - k)$-square permutation matrices, and I_k is the k-square identity matrix. Hence from (2.1) we see that

$$H_S \simeq I_k \dotplus (2I_{n-k} + PQ' + (PQ')'). \tag{2.2}$$

Thus from (2.2) we conclude that the eigenvalues of $\sqrt{H_S}$ are 1 with multiplicity k and $(2 + \lambda_i + \lambda_i^{-1})^{1/2}$, $i = 1, \cdots, n - k$, where the λ_i are the eigenvalues of the permutation matrix PQ'. Let σ and γ be the permutations on $n - k$ symbols corresponding to P and Q respectively. Then $\sigma\gamma^{-1}$ holds no symbol fixed; otherwise the two diagonals would be at least $(k + 1)$-overlapping. Thus $\sigma\gamma^{-1}$ has a decomposition into the product of disjoint cycles each of which has length at least 2. Let $m_1, \cdots, m_p$ be the cycle lengths in this decomposition,

$$\textstyle\sum_{j=1}^p m_j = n - k, \qquad m_i \geqq 2.$$

Then it is well known that the characteristic polynomial of PQ' is $\sum_{i=1}^p (x^{m_i} - 1)$. Hence $PQ' + (PQ')'$ has as eigenvalues the numbers

$$\begin{gathered} e^{2\pi ik/m_t} + e^{-2\pi ik/m_t} = 2\cos(2\pi k/m_t), \\ k = 0, \cdots, m_t - 1, \quad t = 1, \cdots, p. \end{gathered} \tag{2.3}$$

Since $(2 + 2\cos 2\theta)^{1/2} = 2|\cos\theta|$ and

$$\frac{1}{2} + \sum_{t=1}^n \cos t\theta = \frac{\sin(n + \frac{1}{2})\theta}{2\sin(\theta/2)},$$

we conclude from (2.3) that

$$\operatorname{tr}\left(\sqrt{H_S}\right) = k + 2 \textstyle\sum_{t=1}^{p} f_{m_t} \tag{2.4}$$

where

$$f_\alpha = \textstyle\sum_{k=0}^{\alpha-1} \mid \cos(\pi k/\alpha) \mid = \begin{cases} \cot(\pi/2\alpha) & \text{for } \alpha \text{ even} \\ \csc(\pi/2\alpha) & \text{for } \alpha \text{ odd.} \end{cases} \tag{2.5}$$

From (2.4) we see that the proof of the theorem depends upon finding the maximum and minimum values of

$$\psi(\gamma) = \textstyle\sum_{t=1}^{p} f_{m_t} \tag{2.6}$$

where γ varies over all partitions of the form

$$\begin{gathered} \gamma : m_1 + \cdots + m_p = n - k, \\ m_i \geqq 2 \text{ for each } i = 1, \cdots, p. \end{gathered} \tag{2.7}$$

In the next section we state and prove the inequalities necessary to evaluate the extreme values of $\psi(\gamma)$.

III. Some inequalities

The necessary inequalities are contained in the following lemmas.

LEMMA 2. *If α and β are even, then*

$$f_\alpha + f_\beta \leqq f_{\alpha+\beta}\,. \tag{3.1}$$

LEMMA 3. *If α and β are odd, then*

$$f_\alpha + f_\beta \geqq f_{\alpha+\beta}\,. \tag{3.2}$$

LEMMA 4. *If $\alpha \geqq 13$, then*

$$f_3 + f_{\alpha-3} \geqq f_\alpha\,. \tag{3.3}$$

LEMMA 5. *If α is odd, then*

$$f_2 + f_\alpha \leqq f_{\alpha+2}\,. \tag{3.4}$$

We prove Lemma 2 first. Let $g(x) = \cot(\pi/x)$, and note that

$$g''(x) = \frac{2\pi}{x^3} \csc^2 \frac{\pi}{x} \left(\frac{\pi}{x} \cot \frac{\pi}{x} - 1 \right),$$

and since $(\pi/x) \cot(\pi/x) < 1$ for $x > 2$, we conclude that $g(x)$ is a concave function for $x \geqq 2$. Thus

$$2g(\alpha + \beta) \geqq g(2\alpha) + g(2\beta),$$

$$2 \cot \frac{\pi}{\alpha + \beta} \geqq \cot \frac{\pi}{2\alpha} + \cot \frac{\pi}{2\beta},$$

and finally

$$\cot \frac{\pi}{2\alpha + 2\beta} \geqq 2 \cot \frac{\pi}{\alpha + \beta} \geqq \cot \frac{\pi}{2\alpha} + \cot \frac{\pi}{2\beta},$$

$$f_{\alpha+\beta} \geqq f_\alpha + f_\beta\,.$$

This completes the proof of Lemma 2.

To prove Lemma 3 let $g(x) = \csc(\pi/x)$, and note that

$$g''(x) = \frac{\pi}{x^3}\csc\frac{\pi}{x}\left(\frac{\pi}{x}\csc^2\frac{\pi}{x} + \frac{\pi}{x}\cot^2\frac{\pi}{x} - 2\cot\frac{\pi}{x}\right).$$

Let $\theta = \pi/x > 0$, and observe that

$$\theta\,(\csc^2\theta + \cot^2\theta) - 2\cot\theta > 0$$

if and only if

$$h(\theta) = \theta\,(1 + \cos^2\theta) - \sin 2\theta > 0.$$

Now $h(0) = 0$, and $h'(\theta) > 0$ if and only if $\tan\theta > \frac{2}{3}\theta$. This shows that $g(x) = \csc(\pi/x)$ is convex for $x \geqq 2$, and hence

$$f_{\alpha+\beta} = \cot\frac{\pi}{2\alpha + 2\beta} \leqq 2\csc\frac{\pi}{\alpha+\beta} \leqq \csc\frac{\pi}{2\alpha} + \csc\frac{\pi}{2\beta} = f_\alpha + f_\beta\,.$$

A somewhat less direct argument is necessary for Lemma 4. First note that for α even Lemma 4 follows from Lemma 3; thus our problem is to show that for α odd

$$2 + \cot\frac{\pi}{2(\alpha - 3)} \geqq \csc\frac{\pi}{2\alpha}, \qquad \alpha \geqq 13. \tag{3.5}$$

We write a sequence of inequalities each of which implies its predecessor;

$$2 + \frac{2(\alpha-3)}{\pi} \geqq \frac{1}{\sin(\pi/2\alpha)\cos(\pi/2(\alpha-3))}\,,$$

$$\sin\frac{\pi(2\alpha-3)}{2\alpha(\alpha-3)} - \sin\frac{3\pi}{2\alpha(\alpha-3)} \geqq \frac{\pi}{\pi - 3 + \alpha}\,,$$

and from the series expansion for the sine function,

$$\frac{\pi(2\alpha-3)}{2\alpha(\alpha-3)} - \frac{\pi^3(2\alpha-3)^3}{48\alpha^3(\alpha-3)^3} \geqq \frac{\pi}{\pi-3+\alpha} + \frac{3\pi}{2\alpha(\alpha-3)}\,,$$

$$\frac{\pi - 3}{\alpha(\pi - 3 + \alpha)} \geqq \frac{\pi^2}{6(\alpha-3)^3}\,. \tag{3.6}$$

Now (3.6) holds if

$$C(\alpha) = 6(\alpha-3)^3(\pi-3) - \pi^2\alpha(\pi - 3 + \alpha) \geqq 0.$$

Now $C(23) > 0$ may be checked, and it may also be directly verified that the largest root of $C'(\alpha)$ is less than 23. Hence $C(\alpha) > 0$ for $\alpha \geqq 23$, and we check separately the values $\alpha = 13, 15, 17, 19, 21$ to complete the proof. We remark that to check these values requires a table containing hundredths of a degree. We omit the similar proof of Lemma 5.

IV. The proof of the theorem

Now (i) is clear, and (ii) follows from Lemma 2 since the only partition of 4 is $4 = 2 + 2$. Next, every integer greater than or equal to 4 is either even and greater than or equal to 6, or odd and at least 9 except for the integers 3, 4, 5, 7.

Therefore since $f_9 \leqq 3f_3$ and $f_{11} \leqq 2f_3 + f_5$, we see by repeated applications of (3.1), (3.2), (3.3), and (3.4) that $\psi(\gamma)$ is dominated by the value of ψ on a partition of $n - k$ which involves only the integers 3, 4, 5, 7 with appropriate multiplicities. Checking separately that

$$\begin{aligned} f_3 + f_4 &\leqq f_7\,, & f_5 + f_5 &\leqq f_3 + f_7\,, \\ f_4 + f_5 &\leqq 3f_3\,, & f_5 + f_7 &\leqq 4f_3\,, \\ f_4 + f_7 &\leqq 2f_3 + f_5\,, & f_7 + f_7 &\leqq 3f_3 + f_5\,, \end{aligned}$$

we conclude that the value of ψ on any partition of $n - k$ is dominated by its value on a partition consisting of all 3's, or all 3's and a 5, or all 3's and 7. This representation is of course unique since $n - k \equiv 0$, 5, or 7 (mod 3). This completes the proof of (iii). The proof of (iv) proceeds in an analogous way. The last statement in the theorem is precisely the content of Lemma 1.

V. Another application

In problem 4845 of the advanced problem section of the American Mathematical Monthly [1] the following question is posed: Find the maximum of $g_S(A)$ for $A \in O_n$ where S is the set of (i, j) satisfying $i \geqq j$. We answer a generalization of this question in which we assume $i \geqq j + p$, p a fixed nonnegative integer. In this case P_S becomes the n-square matrix whose i^{th} row is $(0, \cdots, 0, 1, 1, \cdots, 1)$ where the first 1 appears in the $i + p$ position; if $i + p > n$, then the i^{th} row of P_S is the zero vector. Then it is easy to check that

$$H_S = \begin{bmatrix} 0 \cdots 0 & \cdots & \cdots & \cdots & \cdots & 0 \\ \vdots \quad \vdots & & & & & \vdots \\ 0 \cdots 0 & \cdots & \cdots & \cdots & \cdots & 0 \\ \cdot \quad \cdot & 1 & \cdots & \cdots & \cdots & 1 \\ \cdot \quad \cdot & \cdot & 2 & \cdots & \cdots & 2 \\ \cdot \quad \cdot & \cdot & \cdot & 3 & \cdots & 3 \\ \vdots \quad \vdots & \vdots & \vdots & \vdots & & \vdots \\ 0 \cdots 0 & 1 & 2 & 3 & \cdots & n-p \end{bmatrix}.$$

Now we observe that if

$$C_S = \begin{bmatrix} 0 \cdots\cdots 0 & \cdots & \cdots & \cdots & 0 \\ \vdots \qquad \vdots & & & & \vdots \\ \cdot \qquad \cdot & & & & 0 \\ 0 \cdots\cdots 0 & \cdots & \cdots & 0 & 1 \\ \cdot \qquad \cdot & & & 1 & 1 \\ \vdots \qquad \vdots & & \cdot^{\cdot^{\cdot}} & & \vdots \\ \cdot \qquad 0 & 1 & & & \cdot \\ 0 \cdots 0 \quad 1 & 1 & \cdots & \cdots & 1 \end{bmatrix},$$

then $C_S^2 = H_S$.

Thus by Lemma 1 we know that the maximum of the function $g_S(A)$ for $A \in O_n$ is $\sum_{j=1}^{n-p} |\lambda_j|$ where λ_j, $j = 1, \cdots n - p$, are the $n - p$ nonzero eigenvalues of C_S. In a private communication Professor A. C. Aitken proved the following result:

$$\lambda_j = \frac{(-1)^{j-1}}{2} \csc \frac{(2j - 1)\pi}{4(n - p) + 2}, \qquad j = 1, \cdots, n - p.$$

These values were obtained by the very elegant observation that the inverse of the lower right nonsingular $(n - p)$-square block of C_S is a differencing matrix whose eigenvectors can be readily computed.

We then can conclude finally that

$$\max_{A \in O_n} \Big| \sum_{i \geqq j+p} a_{ij} \Big| = \frac{1}{2} \sum_{j=1}^{n-p} \left| \csc \frac{(2j - 1)\pi}{4(n - p) + 2} \right|.$$

References

1. F. Koehler, *Problem 4845*, Amer. Math. Monthly, vol. 66 (1959), p. 426.
2. M. Marcus and B. N. Moyls, *On the maximum principle of Ky Fan*, Canadian J. Math., vol. 9 (1957), pp. 313–320.
3. J. von Neumann, *Some matrix-inequalities and metrization of matric-space*, Tomsk. Gos. Univ. Uč. Zap., vol. 1 (1937), pp. 286–299.

THE UNIVERSITY OF BRITISH COLUMBIA
VANCOUVER, CANADA