CONVEXITY OF THE FIELD
OF A LINEAR TRANSFORMATION

A.J. Goldman and M. Marcus

(received June 16, 1958)

Let $U_n$ be an $n$-dimensional unitary space with inner product $(x, y) = \overline{(y, x)}$. In $U_n$ let $S_{n-1}$ denote the unit sphere:

$$S_{n-1} = \{ x \mid (x, x) = 1 \}.$$

Let $A$ be an arbitrary linear transformation of $U_n$. The subset

$$F(A) = \{ \zeta \mid \zeta = (Ax, x), x \in S_{n-1} \}$$

of the $\zeta$-plane ($\zeta = \xi + i\eta$) is called the field of $A$.

As the image of $S_{n-1}$ under the continuous mapping $x \to (Ax, x)$, $F(A)$ must be compact and connected. Toeplitz proved in [4] that the boundary of $F(A)$ is a convex curve. Hausdorff then showed [2] that $F(A)$ actually fills the interior of this curve (i.e., that $F(A)$ is convex). Proofs of the convexity of $F(A)$ also appear in [3] and [5].

The purpose of this note is to provide a simple inductive proof for the convexity of $F(A)$ which reduces the essential computation to the single case $n = 2$. We then dispose of this case by verifying directly that $F(A)$ satisfies the definition of a convex set.

**THEOREM.** $F(A)$ is convex.

**Proof.** (a) If $n = 1$, then $F(A)$ is a single point.

(b) Deferring the case $n = 2$, we suppose $n \geq 3$ and consider the inductive step from $n - 1$ to $n$. Let $x$ and $y$ be any two vectors of $S_{n-1}$; we must show that $F(A)$ contains the segment joining the points $(Ax, x)$ and $(Ay, y)$ in the $\zeta$-plane. Since $n \geq 3$, we can find a vector $u$ in $U_n$ such that $(u, x) = (u, y) = 0$. The unitary-orthogonal complement in $U_n$ of the line $L$ spanned by $u$

is a subspace $U_{n-1}$ of $U_n$ whose unit sphere $S_{n-2}$ is contained in $S_{n-1}$; furthermore, $x$ and $y$ lie in $S_{n-2}$. Any vector $w$ in $U_n$ admits a unique decomposition $w = v + z$, with $v$ in $L$ and $z$ in $U_n$; the unitary-orthogonal projection $P$ of $U_n$ onto $U_{n-1}$ is defined by $Pw = z$. Obviously $A_0 = PAP = P(AP)$ is a linear transformation of $U_{n-1}$ into itself. For any $z$ in $S_{n-2}$ (and thus in $S_{n-1}$) we have $Pz = z$ and thus, decomposing $Az = v_1 + z_1$, \[(Az,z) = (v_1 + z_1, z) = (z_1, z) = (P(Az), z) = (PAPz, z) = (A_0z, z);\] since $(A_0z, z) = (Az, z)$, $F(A_0)$ is a subset of $F(A)$. Also, taking $z = x$ and $z = y$, we see that $(Ax, x)$ and $(Ay, y)$ are in $F(A_0)$; $F(A_0)$ is convex by hypothesis, and so the segment joining $(Ax, x)$ and $(Ay, y)$ lies in $F(A_0)$ and thus in $F(A)$, as desired.

(c) We turn now to the case $n=2$. It is well known (see [1], for example) that there exists a coordinate system (or equivalently, a basis) in $U_2$ with respect to which the matrix of $A$ takes a "superdiagonal" form

\[A = \begin{pmatrix} a & c \\ 0 & b \end{pmatrix}\]

so that for any vector $x$ in the "unit circle" $S_1$ of $U_2$, with coordinates $x_1, x_2$ relative to the system, we have

\[(Ax, x) = a|x_1|^2 + b|x_2|^2 + c\bar{x_1}x_2 \quad (|x_1|^2 + |x_2|^2 = 1)\]

\[= b + (a-b)|x_1|^2 + c\bar{x_1}x_2.\]

If, using the convention $\arg (0) = 0$, we let

\[\alpha = |a-b| \quad (\alpha \geq 0)\]

\[t = \arg(a-b)\]

\[s = |x_1|^2 \quad (0 \leq s \leq 1)\]

\[\theta = \arg x_2 - \arg x_1 - t,\]

and consider the set $S = [F(A) - b]\exp(-it)$, we find that
\[ S = \{ \zeta \mid \zeta = \alpha s + c(s(1-s))^{1/2} \exp(i\theta); \, 0 \leq s \leq 1, \, 0 \leq \theta \leq 2\pi \} \].

Since \( S \) is congruent to \( F(A) \), it suffices to prove that \( S \) is convex.

If \( c = 0 \), then \( S \) is a line segment and therefore convex.

If \( c \neq 0 \) then we can assume \( c = 1 \), since \( F(A) \) is convex if and only if \( c^{-1}F(A) = F(c^{-1}A) \) is convex. Thus we can take \( S \) to be the union of the circles

\[ C(s): \mid \zeta - \alpha s \mid = (s(1-s))^{1/2} = f(s) \quad (0 \leq s \leq 1). \]

Let \( \zeta_1 \) and \( \zeta_2 \) be any points of \( S \) and let \( \zeta_o \) be any point on the line joining them: we must show that \( \zeta_o \) lies in \( S \). Let \( C(s_1) \) and \( C(s_2) \) be circles on which \( \zeta_1, \zeta_2 \) lie, and use the fact that \( \zeta_o \) can be written in the form

\[ \zeta_o = r \zeta_1 + (1-r)\zeta_2 \quad (0 \leq r \leq 1) \]

to define \( s_o = rs_1 + (1-r)s_2 \).

Consider \( G(s) = \mid \zeta_o - \alpha s \mid - f(s) \). Obviously \( G(0) = \mid \zeta_o \mid \geq 0 \) (i.e., \( \zeta_o \) lies outside or on \( C(0) \)). We will show that \( G(s_o) \leq 0 \) (i.e., that \( \zeta_o \) lies inside or on \( C(s_o) \)). It follows that \( G(s^*) = 0 \) (i.e., that \( \zeta_o \) lies on \( C(s^*) \)) for some \( s^* \) with \( 0 \leq s^* \leq s_o \leq 1 \), so that \( \zeta_o \) lies in \( S \) and the convexity of \( S \) will be proved.

To show that \( G(s_o) \leq 0 \), we apply the triangle inequality:

\[ \mid \zeta_o - \alpha s_o \mid \leq \mid r \zeta_o - \alpha s_1 \mid + (1-r) \mid \zeta_o - \alpha s_2 \mid = rf(s_1) + (1-r)f(s_2). \]

Since \( f''(s) \leq 0 \) for \( 0 < s < 1 \), we have

\[ rf(s_1) + (1-r)f(s_2) \leq f(s_o) \]

and so \( \mid \zeta_o - \alpha s_o \mid \leq f(s_o) \) (i.e., \( G(s_o) \leq 0 \)). This completes the proof.
REFERENCES


National Bureau of Standards
and
University of British Columbia