PRODUCTS OF POSITIVE DEFINITE MATRICES. I

C. S. BALLANTINE

For each positive integer n, this paper gives necessary and sufficient conditions (nasc) on a 2×2 real matrix S (of positive determinant) that S be a product of n positive definite real (symmetric 2×2) matrices. Also, when S is the product of (real 2×2) positive definite matrices $P_1, P_2, \cdots, P_n$, it is shown that $P_1, P_2, \cdots, P_n$, and S must satisfy a condition which roughly speaking measures by how much (depending on S) $P_1, P_2, \cdots, P_n$ must collectively differ from scalar matrices.

For $n = 1$ or 2, the abovementioned nasc are known. For $n = 3$, the nasc on

$$S = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$$

is (besides that $ad - bc > 0$) that

$$(c - b)^2 > 4(ad - bc) \quad \text{whenever} \quad a + d \leq 0 .$$

Thus S is here, in a sense, bounded away from the set of negative definite matrices, but its characteristic values may be both negative, since $(a + d)$ and $(ad - bc)$ may be chosen arbitrarily (except of course that $ad - bc$ must be positive) and independently. For $n = 4$, the nasc on S is (besides that $ad - bc$ be positive) just that S not be a negative scalar matrix. For $n \geq 5$, the nasc is just that $ad - bc$ be positive.

These results are extensions of [1, Th. 1] and are obtained, without much additional effort, from the mechanism used in [1, l. c].

2. Main results. We follow the notation of [1]. Let

$$(1) \qquad T = \Delta^{-(1/2)} S = \begin{bmatrix} a & b \\ c & d \end{bmatrix}, \, ad - bc = 1 ,$$

where $\Delta = \det S$ (note the slight change in notation from the previous section). We call T the *unimodular part* of S. Let

$$(2) \qquad p = [(a + d)^2 + (c - b)^2]^{1/2} = [(a - d)^2 + (c + d)^2 + 4]^{1/2} ,$$

$$(3) \qquad \zeta = \tfrac{1}{2} p e^{i\beta} = \tfrac{1}{2} [(a + d) + i(c - b)] \qquad (|\beta| \leq \pi) ,$$

$$(4) \qquad z = [(a + d) + i(c - b)]^{-1} [(a - d) + i(c + b)] ,$$

$$(5) \qquad \rho = 2 \operatorname{argtanh} |z| = 2 \operatorname{argcosh}\left(\frac{p}{2}\right)$$

(since $p = 2(1 - |z|^2)^{-(1/2)}$).

We shall mean by an "n-gonal arc" a polygonal curve which may or may not be closed and which has at most n sides. For any kind of arc we call the segment from its terminal point to its initial point (so directed, if the arc is directed) the *chord* of the arc and its length the *chordlength* of the arc. For a rectifiable arc we define the *area* of the arc to be the algebraic area of the closed curve consisting of the arc plus its chord. We define functions $F_n(A, \rho)$ and $F(A, \rho)$ geometrically in terms of plane *hyperbolic geometry* as follows: $F_n(A, \rho)$ is, for given A and ρ, the minimum arclength of n-gonal arcs of area A and chordlength ρ; and $F(A, \rho)$ is, for given A and ρ, the minimum arclength of rectifiable arcs of area A and chordlength ρ. It is known that, given A and ρ for which at least one n-gonal arc has area A and chordlength ρ, there is, up to congruence, exactly one n-gonal arc of area A, chordlength ρ, and arclength $F_n(A, \rho)$, namely, the simple *regular* n-gonal arc (if $\rho = 0$ this is just a closed regular polygon of exactly n sides, and if $\rho > 0$ this is a simple polygonal arc of exactly n sides which are all equal and which include equal angles) of area A and chordlength ρ. Similarly, it is known that, given $\rho \geqq 0$ and A, there is, up to congruence, exactly one rectifiable arc of area A, chordlength ρ, and arclength $F(A, \rho)$, namely, the simple circular arc of area A and chordlength ρ. Clearly

$$F_n(A, \rho) > F(A, \rho)$$

for every n, and

$$F_n(A, \rho) \longrightarrow F(A, \rho) \text{ as } n \longrightarrow \infty .$$

Now, $F(A, \rho)$ is defined for all $\rho \geqq 0$ and all real A (and we have $F(-A, \rho) = F(A, \rho)$ and likewise for F_n), but the domain of F_n is somewhat more restricted, as given by the following result from hyperbolic geometry.

LEMMA 1. *Let n be a positive integer, ρ be a nonnegative real number, and A a real number. Then there exists an n-gonal arc of area A and chordlength ρ (i. e., $F_n(A, \rho)$ is defined) if and only if*

$$(6) \qquad\qquad |A| < (n - 2)\pi + 2 \arccos \operatorname{sech}\left(\frac{\rho}{2}\right)$$

when $n > 2$, and when $n = 2$ and $\rho > 0$; and $A = 0$ otherwise (i.e., when $n = 1$, and when $n = 2$ and $\rho = 0$).

The proof of this lemma proceeds by standard arguments of hyperbolic geometry, once we notice that the right side of the ine-

quality (6) is, for $n = 2$, just the area of an infinite isosceles "triangle" of base ρ (here "isosceles" means the base angles are equal).

The analytic formulas for F_n and F are rather complicated and are not given in standard references. We give them for reference at the end of this article.

We are now in position to state and prove our main result.

THEOREM 1. *Let* $P_1, P_2, \cdots, P_n$ *be real (symmetric)* 2×2 *positive definite matrices and let*

$$(7) \qquad\qquad S = P_n \cdots P_2 P_1 .$$

For each $j, 1 \leqq j \leqq n$, *let* λ_j *be the larger eigenvalue of* P_j. *Let* $\Delta = \det S$. *Then* $F_n(2\beta, \rho)$ *is defined, and*

$$(8) \qquad\qquad \prod_{j=1}^{n} \lambda_j^2 \geqq \Delta \cdot e^{F_n(2\beta,\rho)} > \Delta \cdot e^{F(2\beta,\rho)} ,$$

where F_n *and* F *are defined above and* β *and* ρ *are defined by* (1), (2), (3), *and* (5). *On the other hand, given* S *such that* $\Delta > 0$ *and* $F_n(2\beta, \rho)$ *is defined, there exist real positive definite matrices* $P_1, P_2, \cdots, P_n$ *such that* (7) *holds and* (8) *holds with equality in the first inequality.*

REMARKS. Combining Lemma 1 with the last part of Theorem 1 and (2), (3), and (5), we get the nasc stated in the summary that the factorization (7) be possible: if $n = 1$, then $\beta = 0$ so $c = b$ and $a + d \geqq 2$, that is, T (and hence S) is positive definite (this case is trivial, but we include it for the sake of completeness); if $n = 2$, then $\beta = 0$ if $\rho = 0$ and if $\rho > 0$ then

$$| \beta | < \arccos \left(\frac{2}{p} \right) ,$$

that is, $a + d = p \cos \beta > 2$ or else $T = I$, and these are just the cases where T (and likewise S) is diagonable with positive characteristic values; if $n = 3$, then

$$| \beta | < \left(\frac{\pi}{2} \right) + \arccos \left(\frac{2}{p} \right)$$

which is equivalent to the condition (always $p \geqq 2$)

$$| c - b | = p \sin | \beta | > 2 \text{ whenever } a + d = p \cos \beta \leqq 0 ;$$

if $n = 4$, then $| \beta | < \pi + \arccos (2/p)$, which precludes only the case $p = 2, | \beta | = \pi$, that is, the case $T = -I$; if $n \geqq 5$, then β and p are unrestricted (aside from $| \beta | \leqq \pi, p \geqq 2$).

Theorem 1 is derived (i.e., proved) in the same way Theorem 2.4

of [1] is derived, using the fact, shown in [1], that the factorization (7) (which induces the factorization

$$T = Q_n \cdots Q_2 Q_1 \,,$$

where Q_j is the unimodular part of $P_j, j = 1, 2, \cdots, n$) corresponds under (4) to an n-gonal arc (in the Poincaré hyperbolic metric) in the unit disc of the z-plane. As shown in [1], the area of this arc is 2β and its chordlength is 2 argcosh $(p/2)$, i.e., is the number ρ given by (5). The unimodular case of (8) is thus derived as in the cited theorem, and the general case follows trivially from the unimodular case.

3. **Geometric interpretation.** In [1] it is mentioned that $a + d$ and $c - b$ (hence ζ, given by (3)) form a complete system of rotation-similarity invariants of T. We may thus expect to get a clearer picture of the situation by considering in more detail the correspondence between polygonal arcs in the z-plane and their images in the ζ-plane. It is clear that these ζ-curves are excluded from the unit disc ($|\zeta| \geqq 1$ since $p \geqq 2$) and all start at $\zeta = 1$ (corresponding to $T = I$), so all the z-curves start at $z = 0$. One sees that positive definite T correspond to those ζ which are real and $\geqq 1$, negative definite T correspond to ζ which are real and $\leqq -1$, $T = -I$ corresponds to $\zeta = -1$, the matrix T of a rotation by the angle β corresponds to $\zeta = e^{i\beta}$ (and to $z = 0$, for all β), T has distinct positive characteristic values if and only if Re $\zeta > 1$, T has both characteristic values $= 1$ if and only if Re $\zeta = 1$, both characteristic values are nonreal if and only if $|\text{Re}\,\zeta| < 1$, and both are negative if and only if Re $\zeta \leqq -1$.

The ζ-arcs corresponding to polygonal z-arcs are piecewise analytic arcs. Each piece, when extended indefinitely in both directions, is an analytic curve which is asymptotic to two rays from the origin making an angle > 0 and $< \pi/2$ (one sees this by noting that $\beta = \arg \zeta$ and 2β is the area of the corresponding geodesic in the z-plane), except for those which lie entirely on one ray through the origin (the extensions of radial pieces). These latter are analytic except where they meet the unit circle; they come in from ∞, "bounce" off the unit circle, and go back to ∞ along the same ray. One sees that, through a given point in the ζ-plane exterior to the unit circle, and in a given nonradial direction at that point, there is one and only one of these analytic curves, and the nonradial pieces cut each ray from the origin at most once. It is also clear that the nonradial pieces do not meet the unit circle at all. (All this can be seen without any detailed calculation. One might suspect from the above properties that each analytic curve which extends a nonradial piece is a branch of a hyper-

bola with center at $\zeta = 0$. Calculation shows that this is indeed the case and that further the semi-transverse axis a and the semi-conjugate axis b of each such hyperbola satisfy $a^2 - b^2 = 1$. Thus one sees that also the extension of each radial piece is one branch of a degenerate hyperbola with center at $\zeta = 0$ and with $a = 1$ and $b = 0$, the limiting case of the nonradial hyperbolas.)

The mapping from the z-plane into the ζ-plane is a mapping of arcs into arcs, and is not a single-valued mapping of points into points, not even locally. But as one can see from the formulas in [1], the mapping is locally conformal (when $z \neq 0$, i.e., when $|\zeta| > 1$), in the sense that the angle of a "corner" in a z-arc is the same as the angle of the corresponding corner in the image ζ-arc. One can further see that radial pieces in the z-plane correspond to (and only to) radial pieces in the ζ-plane.

Thus we see that a 1-gonal z-arc (starting at $z = 0$ always) maps onto a segment of the real ζ-axis, starting at $\zeta = 1$ and going to the right (i.e., outward). Thus the first piece of the image of a 2-gonal z-arc is a segment lying radially outward from $\zeta = 1$ and the terminal point of the second piece may be any point to the right of the line Re $\zeta = 1$ (or may be the point $\zeta = 1$), and only such a point. The terminal point of the image of a 3-gonal z-arc may be any point ζ such that $|\zeta| \geq 1$ and such that

$$|\operatorname{Im} \zeta| > 1 \text{ whenever Re } \zeta \leq 0 ,$$

and only such a point. Thus we see that the images of 3-gonal z-arcs are bounded away from the negative real ζ-axis (hence the corresponding T-arcs are bounded away from the set of negative definite symmetric matrices). The terminal point of the image of a 4-gonal z-arc may be any ζ such that $|\zeta| \geq 1$ and $\zeta \neq -1$ (and only such ζ), and the terminal point of the image of an n-gonal arc with $n \geq 5$ may be any ζ such that $|\zeta| \geq 1$.

4. **Analytic formulas for F_n and F.** $F_n(A, \rho)$ is given implicitly by the system of equations

$$2n\, a = F_n(A, \rho) ,$$

$$|A| = n(\pi - 2\alpha) - 2 \arcsin\left[\cos \alpha \coth a \tanh\left(\frac{\rho}{2}\right)\right] ,$$

$$\sinh\left(\frac{\rho}{2}\right) = (\sinh a)\, U_{n-1}(\sin \alpha \cosh a) ,$$

where (in order that the regular n-gonal arc be simple)

$$\sin \alpha \cosh a \geq \cos\left(\frac{\pi}{n}\right) ,$$

and where the principal branch (from $-\pi/2$ to $\pi/2$) of the arcsine is taken, except when

$$\sin\,\alpha\,\cosh\,a < \cos\frac{\pi}{2n}\,,$$

in which case the second branch (from $\pi/2$ to $3\pi/2$) is taken. U_{n-1} is the Chebyshev Polynomial of the Second Kind of degree $n-1$. (The geometric meaning of α is half the angle included between the sides of the regular n-gonal arc.) Of course n, A, and ρ must satisfy the condition given in Lemma 1. For some values of n, A, and ρ we can give $F_n(A, \rho)$ explicitly. For example, corresponding to a regular closed polygon of n sides, $n \geqq 3$, we have, for $|A| < (n-2)\pi$,

$$F_n(A, 0) = 2n\ \mathrm{argcosh}\left[\left(\cos\frac{\pi}{n}\right)\sec\frac{|A| + 2\pi}{2n}\right].$$

(In [1] $F_n(A, 0)$ was called $F_n(A)$.) Corresponding to half of a regular closed polygon of $2n$ sides ($n \geqq 2$) we have in terms of A, for

$$|A| < (n-1)\pi\,,$$

$$F_n\!\left(A, 2\ \mathrm{argcosh}\left[\left(\cot\frac{\pi}{2n}\right)\tan\frac{|A| + \pi}{2n}\right]\right)$$

$$= 2n\ \mathrm{argcosh}\left(\left(\cos\frac{\pi}{2n}\right)\sec\frac{|A| + \pi}{2n}\right),$$

and in terms of ρ, for $\rho \geqq 0$,

$$F_n\!\left(-\pi + 2n\ \arctan\left(\tan\frac{\pi}{2n}\cosh\frac{\rho}{2}\right), \rho\right)$$

$$= 2n\ \mathrm{argsinh}\left(\sin\frac{\pi}{2n}\sinh\frac{\rho}{2}\right).$$

$F(A, \rho)$ is given implicitly by the system

$$s = F(A, \rho)\,,$$

$$A = \kappa s - 2\ \arcsin\left(\kappa\ \tanh\frac{\rho}{2}\right)$$

$$(9)\qquad\qquad \sinh\frac{\rho}{2} = \frac{1}{\nu}\sin\left(\frac{\nu s}{2}\right),$$

$$(10)\qquad\qquad \nu^2 = \kappa^2 - 1\,,$$

where κ is always real (ν need not be real) and (so that the circular arc be simple)

$$\nu^2 s^2 \leqq 4\pi^2\,,$$

and where the principal branch of the arcsine is taken except when $\nu^2 s^2 > \pi^2$, in which case the second branch is taken. (The expression on the right side of (9) is an even function of ν, hence by (10) an entire function of κ and s, and is real for real κ and s). The geometric meaning of κ is the (constant) geodesic curvature of the circular arc. For some values of A and ρ we can give $F(A, \rho)$ explicitly. For example, corresponding to a full circle we have

$$F(A, 0) = (A^2 + 4\pi \mid A \mid)^{1/2} .$$

Corresponding to a semicircular arc we have in terms of A

$$F\left(A, 4 \arg \sinh \left(\frac{\mid A \mid}{2\pi}\right)^{1/2}\right) = (A^2 + 2\pi \mid A \mid)^{1/2} ,$$

and in terms of ρ, for $\rho \geqq 0$,

$$F\left(2\pi \, \sinh^2 \left(\frac{\rho}{4}\right), \rho\right) = \pi \sinh \left(\frac{\rho}{2}\right) .$$

REFERENCE

1. C. S. Ballantine, *A semigroup generated by definite matrices*, J. Analyse Math. **12** (1964), 257-275.