SYMBOLIC DYNAMICS

Lectures by Marston Morse
1937-1938

Notes by Rufus Oldenburger

Edition with preface, 1966

Not previously published

The Institute for Advanced Study
Princeton, New Jersey

Copies of these notes are available in the Mathematics Libraries of The Institute for Advanced Study and of Princeton University.

PREFACE

These lectures have never been published. Their limited distribution at this time has been at the request of several mathematicians. It has been prompted also by the relevancy of these lectures to diverse developments in mathematics which have occurred since 1937 which indicate the central role played in many disciplines by what may be termed the algebra and geometry of recurrence.

The recent powerful attacks on problems of stability by such mathematicians as Moser, Arnol'd, and others of the Russian School, have not yet involved symbolic dynamics to any considerable degree. This is expected to change as the global topological complexity of stability problems comes more into play.

The author's first contribution to the field of symbolic dynamics was made in 1917. In his thesis the author verified a conjecture of Poincaré as presented by George Birkhoff in [1]. As finally formulated in terms of Birkhoff's "minimal sets" of motions, the conjecture was that dynamical systems of completely "discontinuous type" occur very generally. A dynamical system is said to be of discontinuous type if it possesses non-periodic recurrent sets Ω of motions, and if in phase space the only continua in a set Ω are subarcs of a motion. See page 35.

The Poincaré problem was approached by way of a new kind of symbolic dynamics. In these lectures the methods used by the author [2] in verifying the Poincaré conjecture are given a systematic treatment.

Shortly after these lectures were given in 1937 the author, while on a visit to Germany, solved the problem of the existence of unending games of chess (German rules). This result and algebraic extensions were published in [3]. As pointed out by R. P. Dilworth [3] the methods and the recurrent symbol used by

Morse in 1917 make it possible to construct a nilpotent semi-group S generated by three elements such that the square of every element in S is zero.

The recurrent symbol, so useful for the above purposes, was apparently discovered much earlier by Axel Thue [4]. Thue was not concerned with the above problems of dynamics, chess or semi-groups. Essentially the same symbol was discovered independently by a Russian in 1934 and used by Novikov [5] in his disproof of what is sometimes called a Frobenious-Burnside conjecture in group theory [5].

A fundamental characterization of this recurrent symbol has been presented recently by Hedlund and Gottschalk in [6].

There are, however, many types of recurrent sequences other than the periodic or the one discovered by Thue, Morse and others. One of these, the so-called Sturmian, is related in a precise way to the separation and comparison theorems that are attributed to Sturm. The Sturmian sequences, as introduced in [7], characterize the interrelations of the sequences of zeros of a solution of a second order linear differential equation with periodic coefficients.

In view of the extensive treatment of "ergodicity" from the point of view of measure theory, the related integral-valued ergodic function

$$r \longrightarrow \varphi(r) \qquad\qquad r = 1,2,\ldots$$

introduced in §8. to characterize a transitive symbol is of interest. See [8] and [9].

The second half of these lectures concerns the symbolic representation of geodesics on a compact Riemannian manifold Σ of constant negative curvature -1, and of genus $p > 1$. The covering manifold of Σ is represented (with Poincaré) by a hyperbolic plane in which the straight lines are the circular arcs in the disc

$$D = (x,y \mid x^2 + y^2 < 1) \tag{1.1}$$

which are othogonal to the circle C bounding D.

The manifold Σ is represented first by a symmetric polygon $P \subset D$ with the origin as center. P is bounded by $4p$ circular arcs (arcs of hyperbolic straight lines). The disc D, regarded as covering manifold of Σ, is the union† of a countable number of images Q of P under hyperbolic transformations of a complex variable z which leave C invariant. The resulting Fuchsian group g has $2p$ generators

$$a_1 b_1, \ldots, a_p b_p$$

between which there is one relation.

The word problem. Let P_o be a polygon which is an image of P under some element of g. The images Q of P under elements of g cover† the disc D a 1-1 way. A problem which is fundamental both for the symbolic dynamics of hyperbolic lines and for group theory is to characterize a minimal sequence of polygons Q which form a simple polygonal path leading from P to P_o. This is a geometric form of the so-called word problem for g. It is solved by a theory of convex regions, each the union of a set of polygons Q and leads to an appropriate theory of recurrence and transitivity of hyperbolic lines regarded as geodesics on Σ.

This symbolism, as developed for geodesics on a surface of constant negative curvature, is appropriate for the characterization of unending minimizing geodesics on an arbitrary compact differentiable manifold of genus $p > 1$. See [10] for the case where the representation is biunique and [11] for the general case.

Revised copies of these lectures will be placed in the Mathematics Libraries of the Institute for Advanced Study and Princeton University.

January 1966

---

†With proper conventions as to the boundary of P.

REFERENCES

[1] Birkhoff, G., Quelques théorèmes sur le mouvement des systèmes dynamique, Bull. Soc. Math. France, 40 (1912), 305-323.

[2] Morse, M., Recurrent geodesics on a surface of negative curvature, Trans. Amer. Math. Soc., 22 (1921), 84-110.

[3] Hedlund, G. A., and Morse, M., Unending chess, symbolic dynamics, and a problem in semi-groups, Duke Math. J., 11 (1944), 1-7.

[4] Thue, Axel, Über Die Gegenseitige Lage Gleicher Teile Gewisser Zeichenreihen, 1912. Kristiania Uldenskapsselskapets Skrifler I. Mat. Naturo. Klasse.

[5] Novikov, P. S., Doklady-Akademii-Nauk-SSSR, vol. 127 (1959), 749-752.

[6] Gottschalk, W. H., and Hedlund, G. A., A characterization of the Morse minimal set, Proc. Amer. Math. Soc., 15 (1964), 70-74.

[7] Hedlund, G. A., and Morse, M., Sturmian sequences, Amer. J. Math., 61 (1940), 1-42.

[8] Martin, Monroe., A problem in arrangements, Bull. Amer. Math. Soc., 40 (1934), pp. 859-864.

[9] Hedlund, G. A., and Morse, M., Symbolic dynamics, Amer. J. Math., 60 (1938), 815-866.

[10] Morse, M., Instability and transitivity, J. Math. Pures Appl. (Paris), 14 (1935), 49-71.

[11] Morse, M., A fundamental class of geodesics on any closed surface of genus p greater than one, Trans. Amer. Math. Soc., 26 (1924), 25-61.

1. Introduction. The study of dynamics is the study of a system

$$\frac{dx_i}{dt} = X^i(x_1,\ldots,x_n), i = 1,\ldots,n \tag{1.1}$$

of ordinary differential equations where the X^i are suitably restricted single-valued functions of the real variables $x_1,\ldots,x_n$. The functions X^i will be, in general, assumed to satisfy a Lipschitz condition. The system (1.1) may be merely the local representation of a system of ordinary differential equations defined on a coordinate manifold R on which the x's are local coordinates. The X^i are then contravariant tensors. In general R will be assumed complete. We shall be concerned with functions $x_i = x_i(t)$ which satisfy these equations. The set $x_i = x_i(t), i = 1,\ldots,n$ is said to be a solution, motion, or trajectory satisfying (1.1).

In the field of dynamics there are two points of view: the well-known geometrical, and the symbolic. The latter was introduced by Hadamard in 1898, used by Morse in his doctoral dissertation, and studied by Birkhoff, Martin, Hedlund, Robbins, and others. The symbolic method reduces certain aspects of dynamics to number theory and group theory.

For background the reader is referred to:

Hadamard, Les surfaces à courbures opposées et leur géodésiques, Journal Math. Pures et Appl., series V, vol. 4 (1898), 27-73;

Birkhoff, Dynamical systems, pp. 198-225;

Morse, Representation of geodesics, American Journal of Mathematics, vol. 43 (1921), pp. 33-51;

Morse, Recurrent geodesics on a surface of negative curvature, Transactions of the American Mathematical Society, vol. 22 (1921), pp. 84-100.

A motion transitive with respect to R is one whose closure is R. We shall consider some examples:

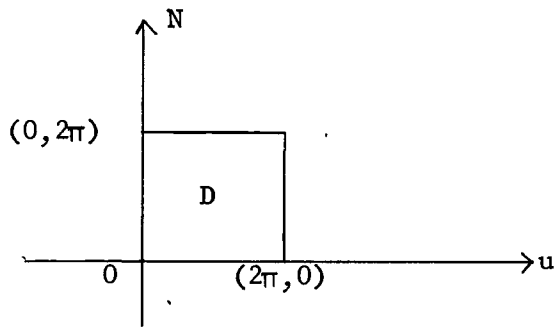
Example 1. Torus. The torus may be represented by parameters u, v with the understanding that pairs (u', v') and (u, v) which are congruent under transformations of the group

$$u' = u + 2\pi n, \qquad v' = v + 2\pi m, \qquad (m,n = \ldots,-1,0,1,\ldots) \tag{1.2}$$

represent the same point on the torus.

It is to be noted that the square D

(Fig. 1.1)



with vertices $(0,0)$, $(2\pi,0)$, $(0,2\pi)$, $(2\pi,2\pi)$ can be mapped topologically on a torus if we identify opposite sides.

Let the differential equations be

$$\frac{du}{dt} = a, \quad \frac{dv}{dt} = b, \tag{1.3}$$

where $a,b \neq 0$, Equations (1.3) have the solution

$$u(t) = at + c, \quad v(t) = bt + d \tag{1.4}$$

where c and d are constants. The solution (1.4) is periodic if corresponding to t there exists a value t' and integers m, n such that

$$u(t) + 2\pi n = at' + c, \qquad v(t) + 2\pi m = bt' + d. \qquad (t' \neq t) \tag{1.5}$$

Combining (1.4) and (1.5), an elementary computation yields

$$\frac{n}{m} = \frac{a}{b}.$$

A motion (1.4) is thus periodic if and only if $\frac{a}{b}$ is rational.

If a solution (1.4) for which $\frac{a}{b}$ is irrational is represented on the square D making use of the congruence relations (1.2), it can be shown that the closure of the solution is the square. The motion is accordingly transitive with respect to the torus.

Example 2. The equations of geodesics on a manifold can be written in the well-known form

$$\frac{d^2u^i}{ds^2} = f^i(u^1,\ldots,u^n;\ \dot{u}^1,\ldots,\dot{u}^n),\ i = 1,\ldots,n, \tag{1.6}$$

where $u^1,\ldots,u^n$ are local coordinates,

$$\dot{u}^i = \frac{du^i}{dt},\ i = 1,\ldots,n,$$

and s denotes arc length. Introducing new variables $z^i = \dot{u}^i$, the equations (1.6) assume the form

$$\frac{dz^i}{ds} = f^i(u,z),$$

$$\frac{du^i}{ds} = z^i,$$

involving first order differential equations. The notation (u,z) is used for $(u^1,\ldots,u^n,z^1,\ldots,z^n)$.

Let

$$ds^2 = g_{ij}(u)du^i du^j$$

by the element of arc on the given manifold, and suppose that

$$g_{ij}(u)\dot{u}^i\dot{u}^j = 1.$$

The set $(u,\dot{u})$ will then be said to be admissible.

A geodesic will be regarded as composed of its "elements" $(u,\dot{u})$. In applying the terms transitive and recurrent to geodesic motion we refer to

the space of elements $(u,\dot{u})$ and not the space of coordinates (u). Thus a transitive geodesic is one whose set of elements $(u,\dot{u})$ has for its closure the set of all admissible elements $(u,\dot{u})$. In geodesic motion on a sphere great circles are not transitive. In fact, for a surface of genus 0, an example of a transitive geodesic is not known. This is one of the great unsolved problems of the theory.

It will be shown that there are transitive geodesics on surfaces of negative curvature of certain types. We refer to surfaces with negative Gaussian curvature at all except a finite number of points.

Let h, k be two curves on a given manifold. Set up a 1-1 sense-preserving correspondence T between points of h and k. Let d_T be the maximum of the distances between corresponding points. The greates lower bound of d_T for all T is called the Fréchet distance between h and k.

Let e, L be given positive numbers. A curve h forms an (e,L) approximation of a motion T, if corresponding to each arc k of T of length L, there is a sub-arc k' of h such that

$$\text{Fréchet distance } kk' < e.$$

A motion T is recurrent if corresponding to each (e,L) there exists a positive number H such that each arc of T of length H forms an (e,L) approximation of T.

The classes of recurrent motions and transitive motions are not identical since simple examples of periodic motions (which are obviously recurrent) can be given which are not transitive. An example of a non-periodic recurrent, non-transitive motion of general type was first given by Morse in the Transactions paper referred to above.

2. Surfaces of negative curvature. Let $z = u(x,y)$ be a real single-valued function of real variables x, y with continuous derivatives up to and including those of second order. We shall use the notation

$$p = u_x \ , \quad q = u_y \ ,$$

$$r = u_{xx} \ , \quad s = u_{xy} \ , \quad t = u_{yy} \ .$$

The Gaussian curvature K is given by the formula

$$K = \frac{rt - s^2}{(1 + p^2 + q^2)^2} \ .$$

Suppose u is harmonic. Then $r = -t$ and $rt - s^2 = -t^2 - s^2$. Evidently $K < 0$ unless $r = s = t = 0$.

Lemma 2.1. If the Gaussian curvature of a function $u(x,y)$ harmonic in a region T is zero on a curve in T, then $u \equiv ax + by + c$ for all points in T.

Let v be a harmonic conjugate of u. Write $f(w) = u + iv$, where $w = x + iy$. Since $f''(w) = u_{xx} + iv_{xx}$, and u, v satisfy the Cauchy-Riemann equations, $K = 0$ on a curve in T only if $f''(w) = 0$ on the curve. Since $f''(w)$ is analytic in T, $f''(w) \equiv 0$ in T, and $f(w) = Aw + B$, where A, B are constants. Hence $u = ax + by + c$.

Since u is harmonic it follows that if $K = 0$ at some points of T, either $K \equiv 0$ throughout T, or $K = 0$ on isolated points or on a curve of T. Lemma 2.1 now implies the following:

Lemma 2.2. If $u(x,y)$ is harmonic in a region T and not of the form $ax + by + c$, the Gaussian curvature of u is less than zero throughout T except at isolated points in T.

That the curvature may be zero at isolated points follows from the example $u = 3x^2y - y^3$ for which $K = 0$ at the origin.

We shall give some examples of surfaces of negative curvature.

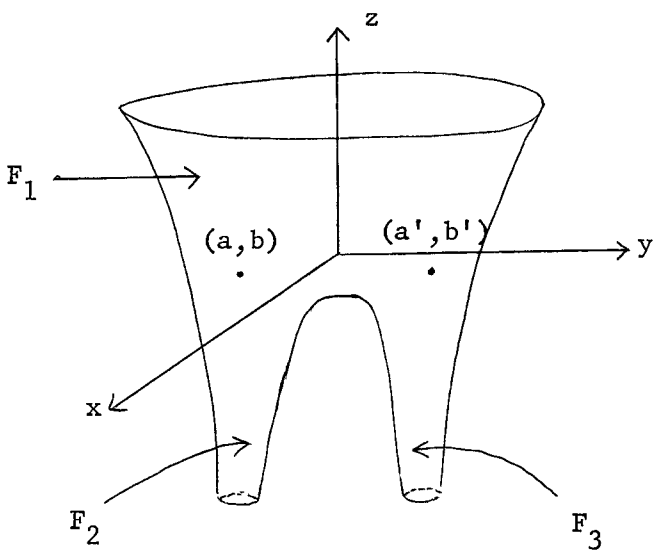
Example 2.1. The surface $z = u(x,y) = \log r$, where $r = \sqrt{(x-a)^2 + (y-b)^2}$, is a surface of negative curvature.

Example 2.2. Since a sum of harmonic functions is harmonic, $z = \log r + \log r'$ is a surface of negative curvature, where r is defined as in Example 1, and

$$r' = \sqrt{(x-a')^2 + (y-b')^2} \quad .$$

Example 2.3. Two finite curves joining the same two points on a surface will be said to be of the same topological type if one can be continuously deformed into the other, holding the end points fast. Two unending curves on a given simply connected surface are said to be of the same topological type if one can be deformed into the other, moving each point P a distance less than some finite number N independent of P. A surface will be called a funnel if it is a homeomorph of half of a circular cylinder cut from the cylinder by a plane perpendicular to its axis.

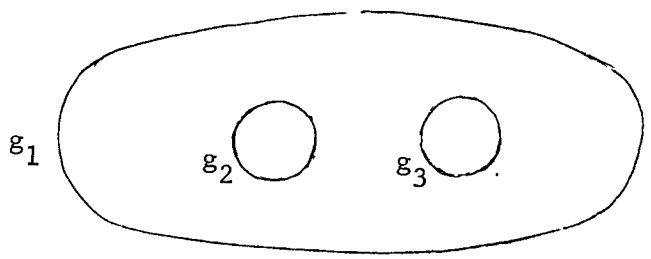
The surface of Example 2.2 is a surface with 3 funnels F_1, F_2, F_3, as indicated in the following figure:



(Fig. 2.1)

On the funnel F_1 the cross-sections (simple closed curves) by horizontal planes $z = c$ increase in length as $c \longrightarrow + \infty$. On each of the funnels F_2, F_3 these sections decrease in length as $c \longrightarrow - \infty$. Hadamard replaced the funnels F_1, F_2 by funnels F_1', F_2' respectively, whose horizontal cross-sections increase as $c \longrightarrow - \infty$. Let us denote the surface thus obtained by s. By taking plane sections sufficiently remote from the xy-plane we obtain on each funnel F of s a simply closed curve g such that there exists a geodesic of minimum length in the class of geodesics on F of the same type as g. Cut s along the minimum geodesic on each funnel, and remove from s the part of F_1 above the minimum geodesic of F_1, and the parts of F_2', F_3' below the minimum geodesic on each of these funnels. The surface s is now reduced to a surface S in a finite region of space, whose total boundary has a projection on the xy-plane of the form

(Fig. 2.2)



One can prove the following:

Theorem 2.1. (See books by Carathéodory and Bolza on the Calculus of Variations.) Given any curve C joining two points on a compact surface Q bounded by a finite number of non-intersecting closed geodesics. There exists a geodesic on Q joining the same two points and of the type of C.

Theorem 2.2 (Gauss). Given any simply-connected geodesic polygon P (closed polygon whose sides are geodesics and finite in number) on a surface Q. Let $a_1, \ldots, a_n$, $n \geq 2$, be the interior angles of P, and let $d\sigma$ be an element

of the surface Q. If $K \leq 0$ on P

$$(2.1) \qquad -\iint_P K d\sigma = [(n-2)\pi - (a_1 + \cdots + a_n)] .$$

If there were a geodesic polygon of two sides on a surface of negative curvature we would have a contradiction, since then

$$-a_1 - a_2 = -\iint_P K d\sigma \leq 0 .$$

Theorem 2.3. On a compact surface of negative curvature bounded by closed geodesics there is exactly one geodesic of a given topological type joining any two given points.

There exist infinitely many unending geodesics on the Hadamard non-analytic surface S described above. The surface can be made simply connected as follows. Let the 3 bounding geodesics of S be denoted by g_1, g_2, g_3 respectively as indicated in Fig. 3. Take an arbitrary simple curve c_1 joining points P_1 and P_2 on g_1, g_2 respectively. By Theorems 1, 3 there is a unique geodesic a joining P_1, P_2 of the type of c_1. Take a simple curve c_2 joining points P_3, P_4 on g_2, g_3 respectively and not intersecting a. Let b denote the unique geodesic of the type of c_2 joining P_3, P_4. As is well-known, two distinct geodesics of finite length intersect in at most a finite number of points. If b and a are tangent at a point they are coincident by a well-known theorem of elementary differential equations theory. If b intersects a, for topological reasons there is a sub-arc of b and a sub-arc of a of the same type joining the same two points, contrary to Theorem 3. The geodesics a and b do not, therefore, intersect. Cutting S along the geodesics a, b yields a simply connected surface M.

We now construct a covering surface N of S. Let there be provided an infinite set of copies of M. Let M_1 be a first copy of M. To each boundary edge of M we join a new copy of M, joining different copies to different boundary edges obtaining thereby a surface M_2 consisting of M_1 and four copies of M adjoined to M_1. We apply this same procedure to M_2 adjoining different copies of M to different boundary edges of M_2 obtaining thereby a surface M_3. Proceeding in this way we obtain a sequence

$$M_1, M_2, \ldots$$

of simply connected surfaces each composed of a finite number of copies of M joined as stated. Let N be the infinitely sheeted surface so constructed, composed of copies of M, which contains each of the surfaces M_n and every point of which is on one of the surfaces M_n. The surface N will be called the universal covering surface of S.

N is simply connected. Therefore on N two curves joining two fixed points P and G are deformable into one another holding P and G fast.

A portion D of a surface is geodesic-convex if there exists an ϵ such that any two points on the boundary of D, apart by a distance less than ϵ, can be joined by a geodesic lying in D.

The surface S is geodesic-convex. Since we cut S to get M_1, the angles between bounding geodesics are less than π, and M_1 is geodesic convex. Continuing, we find that M_n is geodesic convex for each n.

Let P, G be points on M_n. Since M_n is geodesic convex there is at least one geodesic on M_n joining P to Q. By Theorem 2.2 it follows that there is one and only one geodesic on N joining P to G. Moreover, this geo-

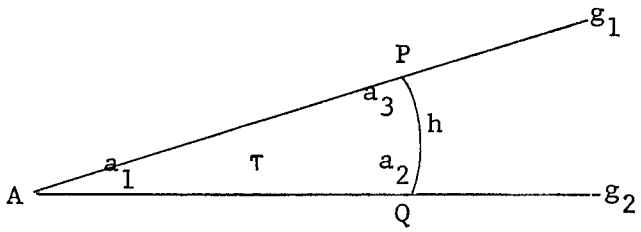
desic minimizes the length among curves which join its end points.

We shall now prove the following:

Theorem 2.4. On the simply connected surface N, two unending geodesics g_1 and g_2 of the same topological type are identical.

Case I. Assume first that g_1, g_2 intersect. By (2.1), g_1 and g_2 can intersect in only one point A.

(Fig. 2.3)



Let P be a point on g_1 which recedes indefinitely from A in one sense on g_1. Since g_1 is of the topological type of g_2, the distance of P from g_2 is at most a finite constant K_0 independent of P. There exists a geodesic arc h from P to g_2 which gives the shortest path from P to g_2. The arc h is orthogonal to g_2 at a point Q.

We shall show that Q recedes indefinitely from A on g_2 as P recedes indefinitely from A on g_1. If B is any point of N, let AB denote the shortest distance from A to B on N. Then

$$AP < AQ + QP \leqq AQ + K_0 ,$$

so that AQ becomes infinite with AP.

We shall now prove that PQ tends to zero as P recedes from A. Suppose first that $-K \geqq d > 0$ for some fixed d, where K is the curvature of N. By (2.1) we have

$$(2.2) \qquad d \iint_{\tau} d\sigma \leqq - \iint_{\tau} K d\sigma \leqq \pi - a_1 - a_2 - a_3 \leqq \pi ,$$

where the symbols are defined as in Fig. 2.3. Since by (2.2) the area of the triangle τ is bounded, PQ tends to zero as P, Q recedes indefinitely. A similar argument holds if $K = 0$ at isolated points.

When P is arbitrarily near to Q, the geodesic angles of h and g_1 at P must be arbitrarily near $\frac{\pi}{2}$ (otherwise g_1 and g_2 would have to intersect once more) so that by the Gauss formula the left member must tend to $-a_1$, which is impossible unless $a_1 = 0$; whence g_1, g_2 have a contact point at A and are identical.

Case II. If g_1, g_2 do not intersect, take any geodesic k joining points of g_1, g_2 such that the sum of the angles e_1, e_2 which k makes with g_1, g_2 on at least one side of k is not less than π. Take a second geodesic h joining points P and Q as in Case I, such that h and k are sides of a quadrilateral in which a_2, a_3 are interior angles. Upon varying P and G as in Case I, using the Gauss formula (2.1) for quadrilaterals, we obtain a contradiction unless g_1 and g_2 are identical.

Let geodesic crossings of a by g_1 in the two senses be denoted by a' and a" respectively. Relative to b let b' and b" be similarly defined. A succession of symbols of the form

$$\cdots c_{-1}c_0c_1 \cdots, \tag{2.3}$$

where the c's are taken† from the set a', a", b', b", and in which a' and a" are never consecutive, nor b' and b", will be termed a symbolic trajectory. If g is an unending geodesic on N, and if the successive crossings of the cuts a, b correspond to successive symbols in (2.3), they "determine" a symbolic trajectory of the form (2.3). We shall show that a second unending geodesic g' which determines the same symbolic trajectory (2.3) is identical with g. For one sees that g and g' lie in the same

†Each c represents a cut and a sensed crossing of the cut.

sequence of copies of M on N, so that g' and g are of the same topological type. Hence g and g' are identical in accordance with the preceding theorem.

Lemma 2.3. Corresponding to each symbolic trajectory T of the form (2.3) there exists an unending geodesic on N with crossings prescribed by T.

Corresponding to T there exists a sequence Z of copies of M on N of the form

$$\dots M^{-1}M^{0}M^{1} \dots \tag{2.4}$$

in which M^i is joined to M^{i-1} along edges corresponding to c_{i-1}, and to M^{i+1} along edges corresponding to c_i, but is otherwise disjoint from copies of M belonging to Z. Each finite block

$$M^{-n} \dots M^{n} \qquad (n > 0) \tag{2.5}$$

of Z is geodesic convex. Let P_n be an arbitrary inner point of M^n. The domain (2.5) on Z being geodesic convex, P_{-n} can be joined to P_n by a geodesic g_n on (2.5) and hence on Z. The geodesic g_n will have the crossings

$$c_{-n} \dots c_{n-1} \tag{2.6}$$

all occurring at inner points of these cuts.

Let E_n be the element $(x, y, z; x', y', z')$ on g_n at its intersection with c_0. The elements E_n will have a limit element E, and the geodesic g bearing the element E will have the crossings (2.3) as required. For it follows from the property of continuous variation of a segment of a geodesic of bounded length with its initial element that g has the crossings (2.6) for each n, and hence the crossings (2.3).

The proof of the lemma is complete.

We combine these results in the following theorem:

Theorem 2.5. There is a 1-1 correspondence between symbolic trajectories T of the form (2.3) and unending geodesics on N in which the crossings of a geodesic g are given by the corresponding T of g.

It is naturally understood that two symbolic trajectories are to be regarded as equivalent if one is obtained from the other by an advance of the subscripts.

We have given an example in which there are three boundaries and two cuts. It is clear that there exist similar examples with m boundaries and m-1 cuts, provided only that $m > 2$. If we had $m = 2$, the two boundary geodesics with the single cut would form a geodesic quadrilateral in which the sum of the angles would be 2π. This would make the Gauss integral of (2.1) zero, which is impossible.

There is a sense in which the correspondence affirmed to exist in the preceding theorem is continuous as we shall see later.

3. Symbolic sequences. Having seen the significance of symbolic trajectories we shall now proceed to develop their theory on an independent basis. Let there be given a finite set S of symbols. If a_n is a symbol of the set S, sequences of the form

(3.1) $$\cdots a_{-1}a_0a_1 \cdots$$

(3.2) $$a_r a_{r+1} a_{r+2} \cdots$$

(3.3) $$a_{r+1} \cdots a_{r+n}$$

will be termed indexed sequences or I-sequences. More particularly the I-sequences of (3.1), (3.2), (3.3) respectively, will be termed I-trajectories, I-rays, and I-blocks.

The I-block (3.3) will be said to have the length n. Infinite I-sequences will be said to be of infinite length.

Two I-sequences A and B will be regarded as <u>identical</u>, i.e. $A = B$, if the ranges of their indices are the same, and if symbols with the same index represent the "same value" in the basic set S. Let D_r represent the operation of adding r to the index of each symbol. The I-sequence obtained from A by operating on A with D_r will be denoted by D_rA. We call two I-sequences A, B <u>similar</u> if there exists an r such that $D_rA = B$, written $A \sim B$. If $r = 0$, $D_rA = A$. In general, if $r \neq 0$, $D_rA \neq A$.

The class of I-sequences similar respectively to an I-trajectory, I-ray, or I-block will be termed a <u>trajectory</u>, <u>ray</u>, or <u>block</u>, and will be represented by the sequences of values involved without indices. For example if S consists of the integers 1 and 2, the unending sequence

$$\ldots\ 1212\ \ldots$$

is a trajectory, but not an I-trajectory.

Corresponding to an I-sequence A we may refer to <u>sub-I-rays</u> or <u>sub-I-blocks</u>. We shall thereby mean subsets of consecutive symbols of A indexed as in A with successive indices differing by unity. Thus, if A is of the form

$$a_1a_2a_3a_4,$$

then a_2a_3 is a sub-I-block, but a_1a_3 is not a sub-I-block.

Let (a) be an I-trajectory and a_r a particular element of (a). The pair (a) and a_r will be termed an I-<u>element</u> of <u>index</u> r <u>based on</u> (a), and will be denoted by E(r,a). If E(m,b) is a second I-element of index s based on an I-trajectory (b), we shall regard E(r,a) and E(s,b) as identical, and write

$$E(r,a) = E(s,b)$$

if and only if (a) = (b), r = s. We shall say that E(r,a) is <u>similar</u> to E(s,b), written $E(r,a) \sim E(s,b)$, if

$$D_{-r}(a) = D_{-s}(b)\ .$$

The infinite class of I-elements similar to a given I-element $E(r,a)$ will be termed an element E represented by $E(r,a)$ based on (a).

Example. We suppose S composed of the numbers 1 and 2. We write the values of the symbols directly above the symbols. Consider an I-trajectory of the form

$$\begin{array}{ccccccccc} \cdots & 2 & 1 & 2 & 1 & 2 & \cdots \\ \cdots & a_{-2} & a_{-1} & a_0 & a_1 & a_2 & \cdots \end{array}$$

Consider also the I-trajectory

$$\begin{array}{ccccccc} \cdots & 1 & 2 & 1 & 2 & \cdots \\ \cdots & b_{-2} & b_{-1} & b_0 & b_1 & \cdots \end{array}$$

Evidently, $E(0,a) \neq E(0,b)$. However, $E(0,a) \sim E(1,b)$. It is to be observed that for even r and s, $E(r,a) \sim E(s,a)$. A similar relation holds when r and s are odd.

We shall assign a metric to the space of all elements E. In the block

$$c_{r-m} \cdots c_r \cdots c_{r+m}$$

c_r is said to be the middle symbol. Let E_1 and E_2 be given elements represented by indexed elements $E(r,a)$ and $E(s,b)$ respectively. Let m be the length of the longest I-blocks of (a) and (b) which are similar ("corresponding" symbols have the same value) and in which a_r and b_s are respectively the middle symbols. To E_1 and E_2 as well as to $E(r,a)$ and $E(s,b)$ we assign a distance

$$E_1E_2 = E(r,a)E(s,b) = \frac{1}{m} \quad .$$

The number m may be zero or infinite, and the distances will then be taken as $+\infty$ and 0 respectively. It is clear that the distance E_1E_2 is independent of the particular I-elements $E(r,a)$ and $E(s,b)$ chosen to represent E_1 and E_2 respectively.

The distances so defined satisfy the usual metric axioms, namely the conditions:

(3.4) $$E_1E_2 = 0 \quad \text{if and only if} \quad E_1 = E_2 ,$$

(3.5) $$E_1E_2 = E_2E_1 ,$$

(3.6) $$E_1E_3 \leq E_1E_2 + E_2E_3 .$$

Relation (3.6) follows from the stronger inequality

(3.7) $$E_1E_3 \leq \text{max. } (E_1E_2, E_2E_3) .$$

To prove this let h, k, m denote the reciprocals of E_1E_3, E_1E_2, and E_2E_3 respectively. Assume that $k \geq m$. Evidently $h \geq m$, whence $h \geq \text{min. } (k, m)$. The latter inequality yields $\frac{1}{h} \leq \text{max. } (\frac{1}{k}, \frac{1}{m})$, except in the cases where $m = 0$ or ∞ in which cases (3.7) holds trivially.

We shall continue with a proof of the following

Theorem 3.1. The space M of all elements is compact.

A metric space is said to be compact if each infinite sequence of points of M contains at least one subsequence which converges to a point of M relative to the metric of M.

Suppose that the element[†] E_n is represented by the I-element $E(0,a^n)$ where (a^n) is an I-trajectory

$$\cdots a^n_{-1} a^n_0 a^n_1 \cdots \qquad n = 1,2,3,\ldots$$

We shall define a matrix

(3.8)
$$\begin{matrix} T_{11} & T_{12} & T_{12} & \cdots \\ T_{21} & T_{22} & T_{23} & \cdots \\ T_{31} & T_{32} & T_{33} & \cdots \\ \vdots & \vdots & \vdots & \end{matrix}$$

in which the elements are I-trajectories (a^{n_i}) chosen as follows. Since S is

[†]Of an infinite sequence (E_n) of elements.

finite, for some value s of S there exists an infinite set of symbols $a_0^{n_i}$ which represent s. Let the first row in (3.8) be a subsequence $(a^{n_1})(a^{n_2})\ldots$ in each trajectory ( ) of which the symbol of index zero has the value s. Again, since S is finite, there is an infinite subsequence of the sequence $T_{11}, T_{12}, \ldots$ such that the pairs $a_{-1}^{n_i}$, $a_1^{n_i}$ have the same values in each T_{ij} chosen; take this for the second row of (3.8). Proceeding inductively we suppose that the first k-1 rows of (3.8) have been determined. The k-th row shall then be a subsequence of the (k-1)st row, composed of the T-trajectories T_{kr} written

$$\ldots\ b_{-1}^r b_0^r b_1^r\ \ldots,$$

such that the symbols

$$b_{-k+1}^r \ \ldots\ b_{k-1}^r \tag{3.9}$$

have the same values for all I-trajectories of the k-th row. The matrix (3.8) is thereby determined.

Let (c) denote an I-trajectory in which c_{-k+1} and c_{k-1} have the values of symbols with indices $-k+1$ and $k-1$ respectively in the k-th row of (3.8). Then (c) has the symbols (3.9) in common with I-trajectories of the k-th row.

Let F, F_n be the elements defined by $E(0,c)$ and $E(0,T_{nn})$ respectively. Then

$$FF_{+n} \leq \frac{1}{2n-1}\ .$$

As $n \longrightarrow \infty$, $FF_n \longrightarrow 0$, so that F_n converges to F. Moreover (F_n) is a subsequence of (E_n) so that the proof of the theorem is complete.

Let T be a trajectory represented by an I-trajectory (a). If there exists a positive integer r such that

$$a_{n+r} = a_r \tag{3.10}$$

for all n in the set $(\ldots,-1,0,1,\ldots)$, T will be said to have the period r.

A trajectory of period r is determined by each of its blocks (period blocks) of length r. There are accordingly at most a countable infinity of distinct periodic trajectories.

An element E will be termed a limit element of a set of elements H if there is an element of H distinct from E in every "neighborhood" (neighborhood taken relative to the metric defined above) of E. We understand that an element E which is based on an I-trajectory (a) is "on" the trajectory defined by (a).

With this understood, let Σ be an arbitrary set of trajectories. A trajectory T will be termed a limit trajectory of Σ if the elements on trajectories of Σ include among their limit elements at least one element of T. The set Σ can consist of a single trajectory, for example of the trajectory T given by

$$\ldots (111)0(11)0\ 1\ 0(11)0(111) \ldots, \tag{3.11}$$

where the parentheses are introduced to show the law of formation. The trajectory

$$\ldots 1111 \ldots$$

is a limit trajectory of the trajectory (3.11).

A periodic trajectory has no limit trajectory because it bears at most a finite set of elements. Let T be a non-periodic trajectory and (a) an I-trajectory representing T. No two I-elements $E(r,a)$, $E(n,a)$, $n > r$ based on (a) are similar, because a similarity relation

$$E(r,a) \sim E(n,a)$$

would imply that (a) had the period $n-r$, contrary to hypothesis. The elements

$$E(1,a)\ E(2,a) \ldots$$

must have at least one limit element E. We term such elements positive limit elements of T. Similarly, the elements

$$E(-1,a)\ E(-2,a)\ \ldots$$

must have at least one limit element and we term such elements negative limit elements of T. Trajectories determined by positive or negative limit elements of T will be termed positive or negative limit trajectories of T.

A trajectory can have several distinct limit trajectories. For example, a trajectory of the form

$$\ldots\ 2222111221221112222\ \ldots$$

has

$$\ldots\ 111\ \ldots$$

and

$$\ldots\ 222\ \ldots$$

as limit trajectories. One can write down a trajectory which has each periodic trajectory as a limit trajectory. Let

$$A_1A_2\ \ldots \tag{3.12}$$

be a set of blocks which includes each finite block. Since there is at most a countable number of finite blocks, such a sequence exists. Let A^r denote the block A repeated r times. The trajectory

$$\ldots\ A_3A_2A_1A_2A_3\ \ldots$$

will have each periodic trajectory

$$\ldots\ AAA\ \ldots$$

as a limit trajectory; for corresponding to each positive integer r, the block A^r will appear in (3.12) in some place.

4. Minimal trajectories. Let s be a trajectory and n an arbitrary positive integer. Since the given set S of symbols is finite, the number of different blocks of length n in s is finite. It will be denoted by

$$P_s(n),$$

and termed the n-th permutation number for s. If t is a limit trajectory of s we have

$$P_s(n) \geq P_t(n) . \tag{4.1}$$

A trajectory s such that

$$P_s(n) = P_t(n) \tag{4.2}$$

for every n and every limit trajectory t of s will be termed minimal.

Periodic trajectories are minimal. They satisfy (4.2) vacuously since they have no limit trajectories.

Theorem 4.1. Among the positive (or negative) limit trajectories of a non-periodic trajectory s there is at least one minimal trajectory.

We shall give the proof for the case of positive limit trajectories. The proof for the case of negative limit trajectories is similar.

We proceed to give an inductive definition of certain symbols.

H_o = the set of limit trajectories of s.

p_1 = the minimum of $P_r(1)$ for r in H_o.

H_1 = the subset of H_o for which $P_r(1) = p_1$.

. . . .

p_n = the minimum of $P_r(n)$ for r in H_{n-1}.

H_n = the subset of H_{n-1} for which $P_r(n) = p_n$.

. . . .

The set H_o and hence H_n is not empty, and

$$H_0 \supset H_1 \supset H_2 \ldots$$

Let r be a limit trajectory of the trajectory s, and let s in turn be a limit trajectory of a trajectory t. It is readily seen that r is a limit trajectory of t. It follows that H_o is closed (i.e., contains its limit trajectories). Proceeding inductively we shall assume that H_{n-1} is closed and shall prove the statement:

(α) The set H_n is closed.

If the trajectories of H_n have no limit trajectory, H_n is closed. Suppose the trajectories in H_n have a limit trajectory t. To show that t is in H_n we first recall that $H_{n-1} \supset H_n$, so that t is a limit of trajectories in H_{n-1}. Since H_{n-1} is assumed to be closed, t is in H_{n-1}. Since the n-th permutation numbers of trajectories in H_{n-1} are at least p_n

$$P_t(n) \geq p_n .$$

On the other hand, since t is a limit trajectory of trajectories of H_n whose n-th permutation numbers are exactly p_n, we have

$$P_t(n) \leq p_n .$$

It follows that

$$P_t(n) = p_n .$$

Thus t belongs to H_n, and H_n is closed, whence (α) is proved.

The sets H_n form a decreasing sequence of closed sets. As follows from the compactness of the space of elements E, the intersection

$$H = H_0 \cdot H_1 \cdot H_2 \ldots$$

of $H_0, H_1, \ldots$ is non-void and closed.

Let t be a trajectory in H. If t is periodic t is minimal. If t is not periodic let r be any limit trajectory of t. Then r and t are both in H_n so that

$$P_t(n) = p_n , \qquad P_r(n) = p_n .$$

Hence

$$P_t(n) = P_r(n)$$

for every n, and every limit trajectory r of t. Hence t is minimal, and the proof is complete.

Theorem 4.2. The relation between a minimal and a limit trajectory is reciprocal.

That is, if s is a minimal trajectory and t is a limit trajectory of s, then t is minimal and s is a limit trajectory of t.

For every block of length n in s is also in t since s is minimal, so that s is a limit trajectory of t. To show that t is minimal, let r be one of its limit trajectories. Then r is also a limit trajectory of s, and since s is minimal

$$P_r(n) = P_s(n) , \quad P_t(n) = P_s(n) ,$$

whence $P_t(n) = P_r(n)$, so that t is minimal.

The proof of the theorem is thus complete.

The set of limit trajectories of a trajectory s will be termed the derived set of s. If s is a non-periodic minimal trajectory each member of its derived set S has s as a limit trajectory and hence has the same derived set S. That this property is characteristic of non-periodic minimal trajectories is stated in the following theorem.

Theorem 4.3. A set S of trajectories which is the derived set of each of its members is composed of non-periodic minimal trajectories.

If s and t are two trajectories of S

$$P_s(n) \geq P_t(n) \quad , \tag{4.3}$$

since t is a limit trajectory of s. But (4.3) also holds with s and t interchanged, so that

$$P_s(n) = P_t(n) \quad .$$

Hence s and t are minimal.

5. Recurrent trajectories. A trajectory s will be termed recurrent if corresponding to each positive integer n there exists a positive integer m with the following property: each block of s of length n has a copy in each block of s of length m.

If s is recurrent, corresponding to each positive integer n there will be a smallest positive integer $m = R_s(n)$ for which the preceding property holds. $R_s(n)$ will be said to be the recurrency function belonging to s.

An I-trajectory will be termed recurrent if it represents a recurrent trajectory. If s is a periodic trajectory with minimum period W, it is clear that s is recurrent and that

$$R_s(n) = n + W - 1 \quad .$$

A recurrent non-periodic trajectory can be constructed as follows. Let blocks $s_1, s_2, \ldots$ be given as follows:

$$s_1 = 1\ 2\ ,$$

$$s_2 = 1\ 2\ 2\ 1\ ,$$

$$s_3 = 1\ 2\ 2\ 1\ \ 2\ 1\ 1\ 2\ ,$$

$$\cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot\ ,$$

s_{r+1} being obtained from s_r by replacing each 1 in s_r by 1 2 and each 2 in s_r by 2 1. Let

$$a_0\ a_1\ a_2 \ldots$$

be an I-ray in which the symbols of the block

$$a_0\ a_1 \ldots a_h \qquad (h = 2^r - 1)$$

take the values given by s_r in the order written. Let

$$a_{-n-1} = a_n \qquad (n = 0,1,\ldots)\ .$$

The I-trajectory (a) can be shown to be recurrent and non-periodic.

Theorem 5.1. The relation between a recurrent trajectory s and any limit trajectory t of s is reciprocal.

That is, if s is a recurrent trajectory and t a limit trajectory of s, then t is a recurrent trajectory and s is a limit trajectory of t.

To show that s is a limit trajectory of t, it is sufficient to show that an arbitrary block u_n of s of length n has a copy in t.

Let $R_s(n)$ be the recurrency function belonging to s. Each block of t is found in s, since t is a limit trajectory of s. Hence each block of t of length $R_s(n)$ will contain a copy of u_n, and t is a limit trajectory of s.

To see that t is recurrent note that each block of length n or $R_s(n)$ in t is also in s so that t must be recurrent with s.

Having proved Theorem 5.1, we now turn to the following fundamental theorem:

Theorem 5.2. A necessary and sufficient condition that a trajectory s be recurrent is that it be minimal.

We shall first prove that s is recurrent if it is minimal. To that end we assume that s is minimal but not recurrent, and we shall arrive at a contradiction.

Let (a) be an I-trajectory representing s. If s is not recurrent there exists a positive integer p with the following property. For each positive integer q no matter how large there exists in (a) an I-block A_n of length n greater than q which fails to contain a sub-I-block similar to at least one I-block of (a) of length p. Since there are at most a finite number of dissimilar I-blocks of length p there must exist at least one, say u_p, together with an infinite sequence of I-blocks

$$A_n \qquad (n = n_1, n_2, \ldots),$$

of length n, where $n_1 < n_2 < n_3 < \ldots$, such that u_p is similar to no I-block in A_n.

Without loss of generality we may assume that n_i is an odd integer so that A_{n_i} has a middle symbol r_i.

We continue with a division into two cases.

Case I. Infinitely many of the I-elements

$$E(r_i, a) \qquad (i = 1,2,\ldots) \tag{5.1}$$

are dissimilar.

Since the space of elements is compact, the elements (5.1) have a limit element E. Let t be the trajectory on which E is based. Since t is a limit trajectory of s and s is minimal, t and s contain the same blocks. But as defined, t does not contain u_p. From this contradiction we infer that s is recurrent.

Case II. Only a finite subset of the elements (5.1) are dissimilar.

An infinite number of the elements (5.1) are similar to one of these elements, say $E(r_1,a)$. Since the blocks A_{n_i} with center r_1 do not contain a block similar to u_p, (a) contains no block similar to u_p. From this contradiction we again infer that s is recurrent.

We shall now show that s is minimal if it is recurrent. If s is periodic it is minimal since it has no limit trajectories. Assume then that s is not periodic. Let t be a limit trajectory of s. Then

$$P_s(n) \geq P_t(n) \quad . \tag{5.2}$$

By Theorem 5.1 s is a limit trajectory of t. Hence (5.2) holds with the inequality reversed; whence

$$P_s(n) = P_t(n) \quad ,$$

and s is minimal.

The proof of the theorem is now complete.

One important consequence of Theorem 5.2 is that we may replace the word "minimal" by "recurrent" in Theorem 4.3.

6. Properties of the space M of elements. A metric space M is termed totally discontinuous if its only continua are points. Cf. Hausdorff, Mengenlehre, 1935, p. 152. We shall prove the following theorem:

Theorem 6.1. The space M of all elements E is totally discontinuous.

We rely on the fact, cf. Hausdorff, p. 151, that on a continuum C every distance is assumed which is less than pq where p and q are points of C. In case the space is M the distances are of the form $\frac{1}{n}$, where n takes on integral values only, so that M can contain no continuum with more than one point.

The derived set S' of a set S of elements of M is the set of all limit elements of the elements of S. A set is perfect if it is identical with its derived set.

Theorem 6.2. The space M of all elements E is perfect.

We know that M is compact. To prove that M is perfect it is necessary only to show that in every neighborhood of an indexed element E(0,a) there is an I-element E(0,b) different from E(0,a). We can define (b) so that it has a prescribed central block in common with (a) but with all larger control blocks different from the corresponding blocks of (a).

Theorems 3.1, 6.1, 6.2 are summarized in the following statement.

Theorem 6.3. The space M of all elements E is compact, perfect, and totally discontinuous.

We shall now prove the following theorem:

Theorem 6.4. If S is the derived set of a recurrent non-periodic trajectory, the elements based on trajectories of S form a compact, perfect, totally discontinuous set N.

N is a subset of the set M of all elements, and N is closed in M. Hence N is compact, being in a compact set M. Since M is totally discontinuous, N is totally discontinuous.

To prove that N is perfect we need only show that in every neighborhood of an element E of N there is an element E' of N distinct from E. Let E be represented by the indexed element E(0,a). By Theorems 4.3 and 5.2 the

I-trajectory (a) is recurrent. Consider a central block

(6.1) $$a_{-m} \cdots a_o \cdots a_m$$

of (a) where m is any arbitrary positive integer. Now (a) is recurrent. Hence there is a block

$$a_{r-m} \cdots a_r \cdots a_{r+m} \ ,$$

with $r \neq 0$ in (a) which is a copy of (6.1). Now $E(0,a)$ and $E(r,a)$ are distinct since (a) is not periodic, and the distance between them is not greater than

(6.2) $$\frac{1}{2m+1} \ .$$

Since m is arbitrarily large, the fraction (6.2) is arbitrarily small, and the theorem is proved.

From Hausdorff, p. 160, we have the following corollary:

Corollary. M and S are homeomorphs of the Cantor perfect nowhere-dense linear set.

Theorem 6.5. The derived set S of a recurrent non-periodic trajectory has the power of the continuum.

By the corollary above the set of different elements on the trajectories in S has the power N of the continuum. On each trajectory in S there is a denumerably infinite number N_o of distinct elements. Let a be the cardinal number of the set of distinct trajectories in S. Therefore

$$N_o a = N \ .$$

Hence, according to a well-known theorem of the theory of sets, cf. Hausdorff, pp. 30-31,

$$N_o a = a \ .$$

Hence $N = a$.

7. Geodesic elements. We suppose that we have given a two-dimensional Riemannian manifold R with overlapping coordinate systems with a metric given locally by

$$ds^2 = g_{ij}(x)dx^i dx^j \qquad (i,j = 1,2) .$$

We suppose that the coefficients $g_{ij}(x)$ are of class C^3. We assume that R is compact as a point set, and that R is bounded by v non-intersecting closed geodesics with $v > 2$. We suppose that R is of genus zero and orientable. Then making $v-1$ simple cuts

$$c_1, \cdots, c_{v-1} .$$

joining the geodesic boundaries in some order yields a simply connected surface S. Let Σ be a universal covering surface belonging to R defined as earlier with the aid of S.

Given an arbitrary point A on a geodesic g of R, we suppose that there is no point B on g such that A and B are mutually conjugate in the ordinary calculus of variations sense. It follows that any two points of A and B on Σ can be joined by one and only one geodesic. That there is at least one geodesic g joining A and B affording an absolute minimum to the distance on R from A to B follows from the fact that R is geodesic convex. We shall show that there can be at most one such geodesic.

Let angles at A with a suitable reference direction be denoted by φ. Let (φ,s) denote the point P on the geodesic g issuing from A with the direction φ at a distance s from A, where s is measured along g. The pair (φ,s) determines a point (u,v) with

$$u = s \cos \varphi ,$$

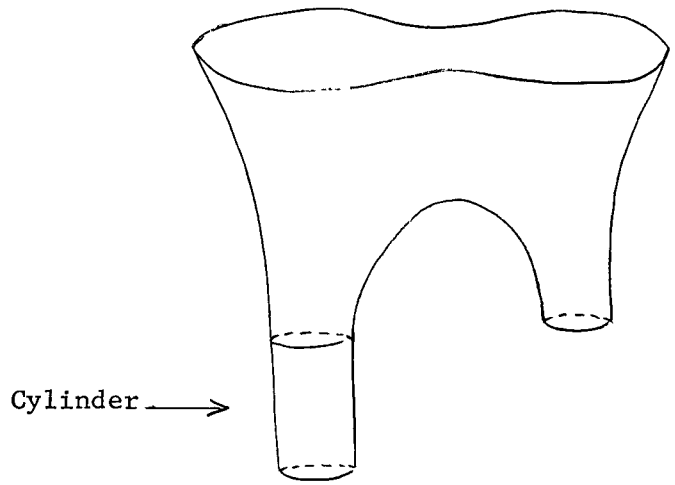
$$v = s \sin \varphi ,$$

in an auxiliary (u,v) plane. The relation between the points P on R and the points (u,v) determined by the same pair (φ,s) is locally one-to-one. Since Σ is simply connected it follows from topological considerations (involving a continuation process) that it is one-to-one in the large. Thus a point P is determined by at most one pair (u,v) so that there is at most one geodesic from A to P.

We now make a new and independent hypothesis.

Hypothesis of unicity. There is at most one unending geodesic on R of a given topological type (see p. 6).

We have seen that this hypothesis is satisfied if R is a surface of negative curvature K. It is not satisfied merely because of the absence of conjugate points. This may be seen by starting with a surface of negative curvature bounded by closed geodesics, with one of these closed geodesics approximately a circle at the end of a funnel. If the funnel is modified so as to be a cylinder near its end, any two right sections of the cylinder will define geodesics of the same topological type



Sufficient conditions less stringent than the condition that the surface have negative curvature have been given by Morse in the paper cited below.

Instability and transitivity, Journal de Mathématiques, vol. 14 (1935), pp. 49-71.

Let g be a given geodesic and let $K(s)$ be the curvature of R in terms of the arc length s on g. We consider the so-called equation of variation

$$\frac{d^2w}{ds^2} + K(s)w = 0 \ . \tag{7.1}$$

A solution $w(s)$ of (7.1) such that

$$w^2(a) + w'^2(a) = 1$$

will be said to be normalized at a. We shall say that g is of <u>unstable type</u> if for every such normalized $w(s)$

$$\lim_{x,y\to\infty} \{|w(a-x)| + |w(a+y)|\} = +\infty \ . \tag{7.2}$$

If $K(s) = 0$ along g, g is not of unstable type because the solution $w = 1$ does not satisfy (7.2). On the other hand if $K < 0$ (7.2) is satisfied. It can be shown by examples that it is by no means necessary that $K \leq 0$. It is merely necessary that $K < 0$ for sufficiently many points.

If the limit (7.2) exists uniformly with respect to all geodesics g on R independent of the choice of a, the geodesics on R will be said to be <u>uniformly unstable</u>. Morse proved the following theorem in the reference cited above.

<u>Theorem</u> 7.1. <u>If the geodesics which remain on</u> R <u>are uniformly unstable the unicity hypothesis is fulfilled on</u> R.

The problem of determining a set of necessary and sufficient conditions for unicity is still unsolved.

By virtue of the conjugate point hypothesis we can suppose that the cuts

$$c_1, \ldots, c_{r-1}$$

are geodesics. We suppose moreover that these cuts have been so made as to have no intersections. With each cut c_k we associate a positive crossing which we denote by c_k. A crossing of c_k in the opposite sense will be denoted by c_k'. Let g be a geodesic which remains on R however continued. The successive crossings of g will define a symbolic trajectory in which the symbols are in the classes (c_k), (c_k'), but in which c_k, c_k' are adjacent for no k. We term such a symbolic trajectory admissible. As in the case of surfaces of negative curvature each admissible trajectory gives the crossings of a geodesic g. That g is uniquely determined by its crossings follows from the unicity hypothesis.

Symbolic elements based on admissible symbolic trajectories will be termed admissible. Let E be such an element. Suppose E based on a trajectory T in which c is the preferred symbol. For $E(r,a)$ the symbol of the I-trajectory (a) with index r is the "preferred" symbol. Let g be the geodesic determined by T. The line element H on g at the crossing corresponding to c will be regarded as an image of E. Two different symbolic elements E determine different line elements H since a line element uniquely determines a geodesic. Line elements H determined by admissible symbolic elements will be termed admissible. The following lemma is obvious:

Lemma 7.1. The relation between admissible symbolic elements E and admissible line elements H is one-to-one.

We define a distance HH' between two line elements H and H' as follows. Let d_1 be the distance on R between the points A and A' of H and H' respectively. Let h be a sensed minimizing geodesic joining A to A'. Let φ and φ' measure the angles which H and H' make with h at A and A' respectively. Let d_2 be the smallest of the angles

$$|\varphi' - \varphi + 2n\pi| \qquad (n = 0,1,\ldots) . \tag{7.3}$$

If there are several minimal geodesics h joining A to A' we take d_2 as the smallest of the angles obtained from all such arcs by the above construction. We set

$$HH' = d_1 + d_2 .$$

We note that the space of admissible line elements is compact. This follows from the property of the continuous variation of a geodesic segment of fixed length with its initial element. The space of line elements will be termed the phase space corresponding to R.

We continue with a proof of the following theorem.

Theorem 7.2. The correspondence between admissible line elements and admissible symbolic elements is one-to-one and continuous.

That the correspondence is one-to-one has already been noted in Lemma 7.1. We have also seen that the space (H) of line elements is compact. That the map of the space (H) onto the space (E) of admissible elements (E) is continuous follows immediately from the continuous variation of geodesic segments of fixed length with their initial elements, and the definition of distance in the space (E). That the inverse map is continuous now follows from the compactness of (H) and the one-to-one character of the correspondence. See Kerekjarto, Vorlesungen über Topologie, p. 34.

Recall that g is termed a limit geodesic of a set S of geodesics if some line element of g is a limit element of line elements on geodesics of S. Let g* be the symbolic trajectory determined by g, and S* the set of symbolic trajectories determined by geodesics of S. If g is a limit geodesic of S, then in the symbolic sense g* will be a limit trajectory of the symbolic trajectories S*, and conversely. Thus in our transition from symbolic to geometrical dynamics admissible symbolic trajectories and elements correspond to admissible geodesics and line elements in a one-to-one manner and limit relations

are preserved. The proofs of the following theorems all depend upon this fact, and the continuous variation of a finite geodesic arc with its initial line segment.

A geodesic will be termed minimal if it is periodic or if it is a member of a set M every geodesic of which has M as a derived set.

Theorem 7.3. A necessary and sufficient condition that a geodesic g be minimal is that the corresponding symbolic trajectory be minimal.

This follows from the definition of a minimal geodesic and the theorem that a minimal symbolic trajectory is either periodic or else a member of a set D every trajectory of which has D as a derived set.

Theorem 7.4. A necessary and sufficient condition that a geodesic g be minimal is that it be periodic or possess the following property. If h is an arbitrary limit geodesic of g each finite arc of g has an arbitrarily small Fréchet distance e in phase space from some subarc of h dependent on e.

Theorem 7.4 is a consequence of Theorem 7.3 and the definition of a symbolic minimal trajectory.

The following theorems are proved in a similar manner.

Theorem 7.5. A necessary and sufficient condition that a geodesic be minimal is that it be recurrent in the sense of the definition of geometric dynamics. (Cf. p. 4.)

Theorem 7.6. The positive or negative limit geodesics of each non-periodic geodesic includes at least one recurrent geodesic.

A non-periodic recurrent geodesic and its limit geodesics are said to form a minimal set.

Consider arcs of geodesics belonging to a minimal set M. Represent M in the corresponding phase space H. Let P be an arbitrary point on M in H. Let N be a neighborhood of P and N* the intersection of the closure of N with M. If for N sufficiently small the only continua on N* are individual arcs of M, the minimal set M is said to be discontinuous at P. A minimal set which is discontinuous at each point P is said to be of discontinuous type.

In Section 1, Example 1, we have exhibited a differential system (1.3) whose solution for a/b irrational forms a minimal set of motions on a torus. The corresponding set of points includes all points on the torus. Clearly this minimal set is not of the discontinuous type.

On the surfaces of negative curvature studied above, each minimal set M is of the discontinuous type. To see this let P be an arbitrary point on a geodesic g of M, and E the line element on g at P. The elements on M sufficiently near E determine geodesics all of which cross some one of the geodesic cuts on the covering surface. Let b denote this geodesic cut. The elements on M with intersection points on b vary continuously with the elements on M near E. But the elements on M with initial points on b form a totally disconnected set by Theorem 6.3. It follows that M is of the discontinuous type.

8. Transitivity. We have seen that there exist rays

(8.1) $$a_1 a_2 \cdots$$

which are transitive; that is, which contain copies of every possible block. This is true even in the case where each symbol has an inverse and inverse

symbols are not permitted to be adjacent. In case (8.1) is a transitive ray there is a least integer $\varphi(r)$ corresponding to each positive integer r such that

$$a_1 \cdots a_{\varphi(r)}$$

contains each block of length r. We term $\varphi(r)$ the ergodic function belonging to the ray (8.1).

We propose the question: "Is there a ray with a best ergodic function $\varphi(r)$; that is, one such that for every other ergodic function $\psi(r)$

$$\psi(r) \geqq \varphi(r) \ ?$$

That the answer is negative at least in the absence of inverses may be seen as follows.

Let there be n symbols in the given set from which our blocks are constructed. As we shall see later, for each positive integer r there exists a block H_r of length $n^r + r - 1$ which contains each r block once and only once. If there are but two symbols 1 and 2, H_1 is 1 2 or 2 1, H_2 is

$$1\ 2\ 2\ 1\ 1 \quad \text{or} \quad 2\ 1\ 1\ 2\ 2\ ,$$

while a particular choice of H_3 is

$$1\ 1\ 2\ 2\ 2\ 1\ 2\ 1\ 1\ 1\ .$$

That H_r exists in general will appear later. We term H_r an r-covering. In general there exists no r-covering which contains a j-covering in the left-hand position for each $j < r$. In particular this is true if $r = 3$. For an H_3 which started with an H_1 would start with 1 2 or 2 1, say with 1 2. The only H_2 which starts with 1 2 is

$$H_2 = 1\ 2\ 2\ 1\ 1\ .$$

To obtain an H_3 containing this H_2 in the left-hand position, we find that we

must continue with 2 1 2 or 1 2 obtaining the blocks

$$1\ 2\ 2\ 1\ 1\ 2\ 1\ 2\ , \qquad 1\ 2\ 2\ 1\ 1\ 1\ 2\ ,$$

and that further continuation to obtain a block H_3 of length 10 containing all blocks of length 3 is impossible. Thus an H_3 of the required type does not exist.

For a given r there exists a transitive ray which starts with an H_r. The ergodic function $\varphi(r)$ for such a ray will be such that

$$\varphi(r) = n^r + r - 1 \tag{8.2}$$

for the given r. Hence a best ergodic function $\varphi(n)$ would be one for which (8.2) holds for all r. But we have just seen that no such ray exists with

$$\varphi(1) = 2\ , \quad \varphi(2) = 5\ , \quad \varphi(3) = 10\ .$$

Thus a "best" ergodic function $\varphi(r)$ does not exist.

From this point on we shall continue with the case where the set of admissible symbols

$$c_1, \ \ldots, \ c_n \tag{8.3}$$

is even in number and the symbols in (8.3) can be grouped into $\frac{1}{2}n$ distinct pairs so that the symbols in the same pair are inverses. As above we shall term a block H_r which contains each admissible block of length r an r-covering. We shall prove the following theorem.

Theorem 8.1. There exists an r-covering corresponding to each positive integer r.

To establish the theorem we shall form a block H_r in accordance with the following rules first given by Martin for the case where the symbols (8.3) possess no inverses. Cf. Monroe Martin, A problem in arrangements, Bulletin of the American Mathematical Society, vol. 40 (1934), pp. 859-864.

Rule 1. Start with $r-1$ symbols c_1, forming a block c_1^{r-1} of length $r-1$.

Rule 2. Continue successively adjoining the symbol in (8.3) of highest index such that no r-blocks are repeated.

Lemma. A necessary and sufficient condition that the application of Rules 1 and 2 yield an r-covering is that c_1 and c_2 in (8.3) shall not be inverses.

Let M_r be the block obtained by applying Rules 1 and 2 as long as possible. If A is any admissible block of length $r-1$, the following blocks of length r

$$Ac_n, \ldots, Ac_1 \tag{8.4}$$

appear if at all in M_r in the order written. One of the blocks (8.3) is inadmissible since the last symbol of A cannot be followed by its inverse. Note that if Ac_1 appears in M_r all other admissible blocks in (8.4) must appear in M_r by Rule 2. If Ac_1 is not admissible, Ac_2 is admissible and appears in M_r only if all of its predecessors in (8.4) appear in M_r. We write $c_1^{r-1} = E$. We shall continue with a proof of the following (a) and (b).

(a) The block M_r ends with E.

If the block M_r ended with a block A of length $r-1$ different from E, A would not occupy the first position in M_r by Rule 1. A could then appear in M_r at most $n-1$ times; otherwise there would be an admissible block of the form c_jA which would appear twice. Upon its $(n-1)$-st appearance A can be followed by c_1, or c_2 if Ac_1 is not admissible, contrary to the hypothesis that M_r ends with A.

(b) The block M_r ends with Ec_1.

Suppose M_r ends with a block $B \neq c_1E$ of length r. Set $B = bE$ in accordance with (a). We suppose that $b \neq c_1$. The block E appears n times

in M_r for otherwise M_r could be continued. Hence Ec_1 appears in M_r but not in the last position. But Ec_1 cannot be continued since it is preceded by each admissible block Ec_j, and itself ends with E. From this contradiction we infer the truth of (b).

We shall now make use of the assumption that c_1 and c_2 are not inverses and show that M is an r-covering.

We first observe that each admissible block Ec_j appears in M_r. Hence M_r contains n copies of E. It follows that M_r contains each admissible block c_jE. Let

(8.5) $$b_1 \ldots b_r$$

be an arbitrary admissible block of length r. We shall show that M_r contains (8.5).

Suppose that M_r does not contain (8.5). Set

$$b_2 \ldots b_r = D.$$

Then $D \neq E$ by virtue of (a). Hence D appears at most n-2 times in M_r; otherwise each admissible block c_jD would appear in M_r. Hence Dc_1 does not appear in M_r (or Dc_2 if Dc_1 is not admissible).

We now apply the same reasoning to Dc_1 (or Dc_2) that we have applied to (8.5), and infer that

(8.6) $$b_3 \ldots b_r c_1 c_1 \quad \text{or} \quad b_3 \ldots b_r c_2 c_1$$

does not appear in M_r. Continuing, we arrive at the conclusion that c_1^r does not appear in M_r contrary to (b).

The condition of the lemma is hence sufficient.

To prove that the condition is necessary we suppose that c_1 and c_2 are inverses. If M_r were an r-covering, c_2^r would appear in M_r, but not in the last position by virtue of (a). But c_2^r cannot be continued in accordance with Rule 2 since the blocks

$$c_2^{r-1} c_j \qquad j = 2,\ldots,n$$

have already appeared. This completes the proof of the lemma.

A function $\psi(r)$ will be said to be asymptotically at least a positive function $\varphi(r)$ if

$$\lim_{r \longrightarrow \infty} \frac{\psi(r)}{\varphi(r)} \geqq 1 \ .$$

The expression "is asymptotically equal" will be indicated by the symbol $\sim$.

There are $n(n-1)^{r-1}$ different admissible blocks of length r so that an r-covering has a length at least

$$n(n-1)^{r-1} + r - 1 \ . \tag{8.7}$$

Set

$$A(n,r) = \frac{(n-1)n(n-1)^{r-1}}{(n-2)} \ .$$

Theorem 8.2. In the case where our symbols possess inverses, each ergodic function is asymototically at least $n(n-1)^{r-1}$, and there exists a transitive ray whose ergodic function is asymptotically not greater than $A(n,r)$.

Let H_r be an r-covering formed in accordance with the preceding rules and let H_r^* be the block obtained from H_r by omitting the first r-1 symbols c_1. The ray

$$H_1 H_2^* H_3^* \ldots \tag{8.8}$$

is clearly transitive since H_r^* is preceded by c_1^{r-1} in (8.8) for each $r > 1$. Moreover (8.8) is admissible. For this ray the ergodic function $\varphi(r)$ is at most the length of the block

$$H_1 H_2^* \ldots H_r^* \ .$$

That is

$$\varphi(r) \leq \sum_{m=1}^{r} n(n-1)^{m-1} \ .$$

Hence

$$\varphi(r) \leq n \left[\frac{1 - (n-1)^r}{1 - (n-1)} \right] \sim A(n,r) \quad .$$

Upon referring to (8.7) we see that

$$\varphi(r) \geq n(n-1)^{r-1} + r - 1 \sim n(n-1)^{r-1} \quad ,$$

which completes the proof of the theorem.

That each ergodic function is asymptotically at least n^r follows from (8.7), and the proof of the theorem is complete.

9. Hyperbolic geometry. We shall use the Poincaré representation of 2-dimensional hyperbolic geometry. The points will be the subset of points of the ordinary Euclidean xy-plane for which $y > 0$. The group G of motions which will preserve the metric of this geometry (this metric will be defined later) will be the group of all fractional linear transformations of the complex variable $z = x + iy$ into $w = u + iv$ such that the domain $y > 0$ of the xy-plane corresponds to the domain $v > 0$ of the uv-plane.

Lemma 9.1. The group G consists of the set of transformations T of the form

$$w = \frac{az + b}{cz + d} \qquad ad - bc > 0 \tag{9.1}$$

in which a, b, c, d are real.

For transformations T

$$\frac{dw}{dz} = \frac{ad - bc}{(cz+d)^2} \ . \tag{9.2}$$

That each transformation T is in G follows from the fact that T carries the real axis into the real axis, preserves sense on these axes since $\frac{dw}{dz}$ is positive, and further by virtue of the direct conformality of T carries the domain $y > 0$ into $v > 0$. Conversely, each transformation H of G can be represented in the form (9.1). For H carries three distinct points x_1, x_2, x_3 on the real axis $y = 0$ into three distinct points u_1, u_2, u_3 on the real axis $v = 0$ (one of these points may be ∞). If z corresponds to w under H, we have the following equality between the cross ratios

$$CR(z, x_1, x_2, x_3) = CR(w, u_1, u_2, u_3) \ . \tag{9.3}$$

The relation (9.3) between w and z reduces at once to (9.1). Moreover,

$ad - bc > 0$ if $y > 0$ corresponds to $v > 0$.

Lemma 9.2. Under transformations T of G

$$\frac{d\sigma}{v} = \frac{ds}{y} , \qquad v \neq 0, \quad y \neq 0, \tag{9.4}$$

where $d\sigma$ and ds are corresponding differentials of arc length in the w- and z-planes respectively and (x,y) corresponds to (u,v).

Note that under T

$$w = \frac{az + b}{cz + d} \cdot \frac{c\bar{z} + d}{c\bar{z} + d} = \frac{az\bar{z} + adz + bc\bar{z} + bd}{|cz + d|^2} , \qquad \bar{z} = x - iy,$$

from which we obtain

$$v = \frac{(ad - bc)y}{|cz + d|^2} , \tag{9.5}$$

while

$$\left| \frac{dw}{dz} \right| = \left| \frac{d\sigma}{ds} \right| = \frac{ad - bc}{|cz + d|^2} . \tag{9.6}$$

Relation (9.4) follows from (9.5) and (9.6).

The length of a curve in our hyperbolic geometry will now be defined by the integral

$$\int \frac{ds}{y} = \int \frac{\sqrt{x'^2 + y'^2}}{y}\, dt, \qquad y > 0 ,$$

taken along this curve represented by the equations $x = x(t)$, $y = y(t)$. The length of a curve is invariant under transformations of G regarded as transforming our hyperbolic plane into itself, as follows from the preceding lemma.

Lemma 9.3. With length so defined, the shortest paths are open semi-circles orthogonal to the axis of reals.

Any circle perpendicular to the axis of reals can be transformed into any other such circle under a transformation of G, since it is necessary merely to transform the end points of the circles into each other under a transformation of the form (9.3).

Moreover the straight lines perpendicular to the x-axis may be included in this set of circles. For the transformation

$$w = \frac{1}{z}$$

is in G and carries the semicircle with end points at $z = 0$, $z = 1$ into the half-line with end at $w = 1$ and perpendicular to $v = 0$. Transformations of the form $w = z + b$ carry this half-line into any other half-line perpendicular to the real axis.

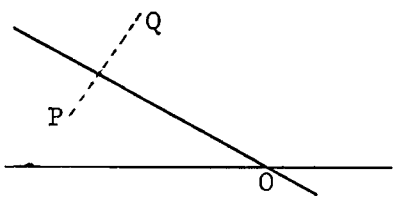
To prove the lemma it is accordingly sufficient to show that each finite segment of the y-axis, on which $y > 0$, is a shortest path in our geometry. To that end let $C : [x(t),y(t)]$ be an arbitrary rectifiable curve on $y > 0$ joining two points $(0,y_1)$, and $(0,y_2)$ with $y_1 < y_2$. If $x(t)$ and $y(t)$ are absolutely continuous

$$\int_C \frac{\sqrt{x'^2 + y'^2}}{y}\, dt \geq \int_C \frac{\sqrt{y'^2}}{y}\, dt \geq \int_{y_1}^{y_2} \frac{dy}{y} ,$$

and the equality holds if and only if C is identical with the segment $y_1 \leq y \leq y_2$ of the y-axis.

The lemma follows directly.

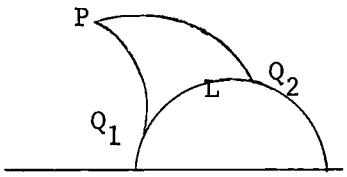
The semicircles perpendicular to the axis of reals including the half-lines will now be called <u>hyperbolic straight lines</u>, or more briefly H-<u>lines</u>. The points on the axis of reals are not regarded as belonging to the H-lines. We shall refer to such points as the <u>ideal end points</u> of the H-lines. Given any two points P and Q in our H-plane, there exists one and only one connecting H-line L. Its Euclidean center lies at the intersection of the x-axis with the Euclidean line equidistant from P and Q.



The H-length of L between P and Q will be termed the H-distance between P and Q.

It is clear that these H-lines and H-distances satisfy all the axioms of Euclidean geometry save the parallel axiom. We say now that two H-lines are parallel if they have a common ideal end point. If L is a given H-line and P a point not on L, there exist two H-lines through P parallel to L.

The parallels to L through P can be obtained as the limiting position of H-lines drawn from P to points Q_1 and Q_2 on L as Q_1 and Q_2 tend respectively to the ideal end points of L.



The fixed points of the transformations T of G given by (9.1) satisfy the relation

$$cz^2 + (d-a)z - b = 0.$$

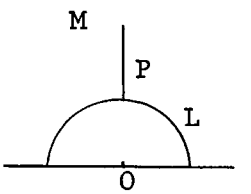
Since the coefficients are real, the fixed points are either real or conjugate imaginary. T is thus one of the following three types.

I. If there are two real fixed points A and B, the circles through A and B must be transformed into themselves since the x-axis is so transformed, and the transformation is conformal. Such transformations of G are termed hyperbolic.

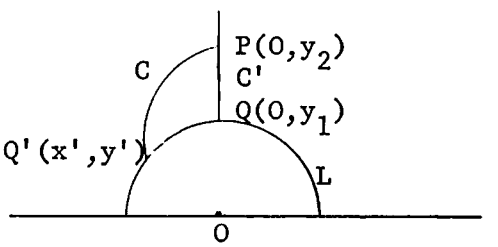
II. If there is just one real fixed point $A \equiv B$, the transformation is parabolic and the family of invariant circles tangent at A must include the real axis.

III. If the fixed points A and B are conjugate imaginary, the circles C orthogonal to the circles through A and B include the axis of reals as an invariant circle. Since the upper half plane is carried into itself the circles C must be invariant and the transformation is elliptic.

If L is an arbitrary H-line and P a point not on L, there is a unique H-line M through P perpendicular to L. Since any H-line can be brought by a transformation (9.1) into any other H-line, we can take L as a semicircle C with center at the origin, and then by using a hyperbolic transformation with fixed end points at the ideal end points of C carry P into a point on the y-axis. Then M goes into a half-line perpendicular to L and the x-axis.



The statement is then obvious, the only H-line perpendicular to L through P being the y-axis. In this position it is also clear that M affords the shortest H-path from P to L. We employ the figure



where C' denotes the H-segment from P to L intersecting L at $Q(0,y_1)$, and C is an arbitrary rectifiable curve on $y > 0$ joining P to $Q(x',y')$ on L, where $Q \neq Q'$. Let C be represented by $[x(t),y(t)]$, where $x(t)$ and $y(t)$ are absolutely continuous. Then

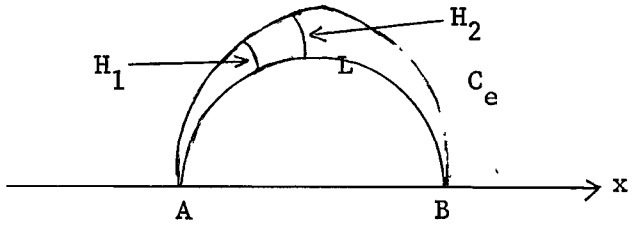
$$\int_C \frac{\sqrt{x'^2 + y'^2}}{y}\, dt \geq \int_{y_1'}^{y_2} \frac{dy}{y} > \int_{y_1}^{y_2} \frac{dy}{y} \quad .$$

But

$$\int_{y_1}^{y_2} \frac{dy}{y}$$

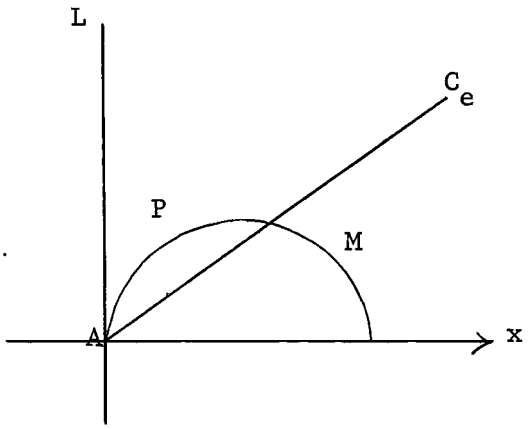
is the H-length from P to Q measured along C'. Hence M affords the shortest H-path from P to L, and this result is general by virtue of the invariance of H-length.

The locus of points at a constant H-distance e from an H-line L is represented by a Euclidean circular arc C_e with end points at the ideal end points A and B of L.



For the H-perpendiculars, e.g. H_1, H_2, from points of C_e to L can be carried into each other by hyperbolic transformations with fixed points A and B and so have the same H-length.

Let M be an H-line parallel to L with A as a common ideal end point of L and M. We shall show that the H-distance of M from L tends to zero as A is approached. Without loss of generality we can take L and M as in the following figure:



Let C_e be a straight line through A and intersecting M, and whose points are an H-distance e from L. If a point P on M is sufficiently near A in the Euclidean sense P will be between C_e and L and hence have an H-distance at

most e from L. Thus the H-distance of M from L tends to zero as A is approached.

It is clear that the upper half-plane as the space of our H-geometry can be replaced by the interior of a unit circle. An explicit transformation which carries the upper half-plane into the interior of the unit circle follows.

$$w = \frac{z-i}{z+i} \quad . \tag{9.7}$$

For z real z-i and z+i are conjugate imaginary so that when z is real $|w| = 1$. Moreover, the point $z = i$ corresponds to $w = 0$ so that the upper half-plane goes into the interior of the unit circle.

Lemma 9.4. If $d\sigma$ and ds are corresponding differentials of arc length in the w- and z-planes, and the point (x,y) corresponds to (u,v) under (9.7), then

$$\frac{ds}{y} = \frac{2d\sigma}{1-u^2-v^2} \quad , \qquad y \neq 0. \tag{9.8}$$

The inverse of (9.7) takes the form

$$z = \frac{-i(w+1)}{(w-1)} \quad .$$

If we indicate conjugates by the bars, we have

$$z = -i\left[\frac{(w+1)(\bar{w}-1)}{(w-1)(\bar{w}-1)}\right] = -i\frac{[w\bar{w}+\bar{w}-w-1]}{|w-1|^2} = -i\frac{[u^2+v^2-2iv-1]}{|w-1|^2} \quad .$$

Now

$$y = \frac{1-u^2-v^2}{|w-1|^2} \quad . \tag{9.9}$$

But

$$\frac{dz}{dw} = \frac{2i}{(w-1)^2} \quad ,$$

and

$$\frac{ds}{d\sigma} = \frac{2}{|w-1|^2} \quad . \tag{9.10}$$

By (9.9) and (9.10)

$$\frac{ds}{d\sigma} = \frac{2y}{1-m^2-v^2} \quad ,$$

from which (9.8) follows.

Representing our hyperbolic plane by the points $u^2 + v^2 < 1$ in the (u,v)-plane, the H-differential ds of arc length has the form

$$ds^2 = \frac{4(du^2+dv^2)}{(1-u^2-v^2)^2} \tag{9.11}$$

by (9.8). We shall show that the Gaussian curvature K of our H-plane is a negative constant. To that end we make use of a formula [cf. Darboux, Leçons sur la théorie générale des surfaces, vol. 2, p. 397] according to which

$$ACK = -\frac{\partial}{\partial u}\left(\frac{1}{A}\frac{\partial C}{\partial u}\right) - \frac{\partial}{\partial v}\left(\frac{1}{C}\frac{\partial A}{\partial v}\right) \quad , \tag{9.12}$$

when $ds^2 = A^2du^2 + C^2dv^2$. By (9.11)

$$A = C = \frac{2}{(1-u^2-v^2)} \quad ,$$

whence

$$\frac{\partial C}{\partial u} = \frac{4u}{r^2} \quad , \qquad \frac{\partial A}{\partial v} = \frac{4v}{r^2} \quad ,$$

where $r = 1 - u^2 - v^2$. Formula (9.12) now takes the form

$$\frac{4K}{r^2} = \frac{\partial}{\partial u}\left(\frac{2u}{r}\right) - \frac{\partial}{\partial v}\left(\frac{2v}{r}\right) = \frac{-4}{r^2} \quad .$$

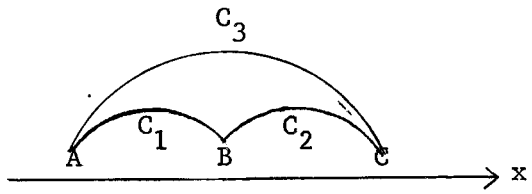
Thus $K = -1$.

A particular consequence is that the sum of the interior angles of a geodesic triangle Δ is less than π. By Theorem 2.2 (Gauss)

$$\iint_{\Delta} K dA = \alpha_1 + \alpha_2 + \alpha_3 - \pi \quad ,$$

where dA is the element of area, and $\alpha_1, \alpha_2, \alpha_3$ are the interior angles of Δ. That the sum of the interior angles of Δ may be arbitrarily small is seen from the example of a geodesic triangle bounded by arcs of H-lines C_1, C_2, C_3 whose points of intersection A, B, C are arbitrarily close to the x-axis as indicated

in the following figure:



We shall need further details regarding the group G' of linear fractional transformations of the disc $u^2 + v^2 < 1$ into itself. Let G be the group of linear fractional transformations leaving the upper half-plane $y > 0$ invariant, and let T be a fractional linear transformation carrying the upper half-plane into the disc $u^2 + v^2 < 1$. If A is a transformation of G the so-called transform

$$T A T^{-1}$$

of A by T is hyperbolic, elliptic, or parabolic according to the character of A. We shall make use of this fact in characterizing the transformations of G'.

Lemma 9.5. For a parabolic or elliptic transformation of G (or G') the greatest lower bound of the H-distances between congruent points is zero.

This is obvious for elliptic transformations since there then exists a fixed point on $y > 0$, and the distance from this fixed point to itself is zero.

To establish the fact for parabolic transformations we consider a parabolic transformation of this domain $y > 0$ into itself. If the fixed point is at infinity this transformation must be of the form

$$w = z + b, \qquad z = x + iy, \qquad (b \neq 0)$$

where b is real. The point (a,c) is transformed into the point $(a+b,c)$ and the H-distance between these points is less than the integral $\int \frac{ds}{y}$ taken along the line $y = c$ joining these two points, that is less than

$$\int_a^{a+b} \frac{dx}{c} = \frac{b}{c} .$$

Now c can be taken arbitrarily large so that b/c will be arbitrarily small.

A hyperbolic transformation T of G has two fixed points A and B on the unit circle. The transformation T moves points on the invariant circles through A and B in one sense, say from A towards B. We term A and B respectively the positive and negative fixed points of T. The sensed H-line joining A to B is called the axis of T.

Lemma 9.6. If a subgroup H of G (or G') contains two hyperbolic transformations with one and only one fixed point in common, H also contains a parabolic transformation.

We again take the space $y > 0$ as representative of the hyperbolic plane. Without loss of generality we can take the common fixed point as the point at infinity and the two given hyperbolic transformations in the form

T: $$w = az \qquad (a \neq 0,1)$$

S: $$w - c = b(z-c) \qquad (b \neq 0,1, c \neq 0),$$

where a, b, and c are real. We have chosen T so that the origin is the other fixed point of T, and S so that (c,0) is its second fixed point. We shall show that the product transformation

$$S\ T\ S^{-1}\ T^{-1},$$

applied starting with T^{-1}, is parabolic. The inverse of S takes the form

S^{-1}: $$z = \frac{w-c}{b} + c.$$

The transformations of our product applied successively [from right to left] to a complex number z lead to the points

$$\frac{z}{a}, \quad \frac{z-ac}{ab} + c, \quad \frac{z-ac}{b} + ac, \quad z + c(b-1)(a-1).$$

But $w = z + c(b-1)(a-1)$ is parabolic, thus proving the lemma.

10. The group g. We shall now define a discontinuous subgroup g of the group G. This group is called the Fuchsian group. See Morse, A fundamental class of geodesics on any closed surface of genus greater than one, Transactions

of the American Mathematical Society, vol. 26 (1924), pp. 25-60; Poincaré, Théorie des groupes fuchsiens, Acta Mathematica, vol. 1 (1882), pp. 1-62. Let p be any positive integer greater than one. Let C_r represent the circle $u^2 + v^2 = r^2$ with $r > 1$. On C_r we consider $4p$ equidistant points taking one of these points as the point $(r,0)$. With these points as centers draw $4p$ circles orthogonal to the unit circle $u^2 + v^2 = 1$. Now let r and the common radius of the $4p$ circles vary, the circles remaining orthogonal to the unit circle, and having centers placed as above. As r becomes arbitrarily large, the $4p$ circles will approach straight lines passing through the origin. The interiors of none of these circles will include the origin. The origin will be in a circular polygon Q of $4p$ sides (circular arcs) with interior angeles which tend to $\pi - \frac{2\pi}{4p}$ as r becomes infinite.

If, however, r decreases to unity, for some value $r > 1$, these $4p$ circles will be tangent to each other at their successive intersections with the unit circle. That is, the interior angles of Q will diminish from $\pi - \frac{2\pi}{4p}$ to zero. But for $p > 1$

$$\pi - \frac{2\pi}{4p} = \frac{2\pi}{4p}\,[2p-1] > \frac{2\pi}{4p} = \frac{\pi}{2p} \quad ,$$

so that for a properly chosen value of r, Q will have interior angles of magnitude equal to $\frac{\pi}{2p}$. Denote the corresponding polygon by S. The sides of S will be segments of H-lines. The sum of the interior angles will be 2π.

Let P_1 and P_2 be successive vertices. Let C_1, C_2, C_3 be three successive H-lines such that C_1, C_2 intersect at P_1, and C_2C_3 in P_2. Now C_1, C_3 do not intersect. Otherwise C_1, C_2, C_3 would define a geodesic triangle with vertices P_1P_2 and R, where R is the intersection of C_1, C_2 on $u^2 + v^2 < 1$. The interior angles of such a triangle would have the magnitudes

$$\pi - \frac{\pi}{2p}$$

at P_1 and P_2 and therefore have a sum at least

$$2\pi - \frac{\pi}{p} = \pi + (\pi - \frac{\pi}{p}) > \pi ,$$

which is impossible.

Let the boundary of S be assigned the counterclockwise sense as a positive sense. The $4p$ sensed sides of S, taken in counterclockwise order starting with the vertex on the u-axis will be labeled

$$A_1, B_1, C_1, D_1, A_2, B_2, C_2, D_2, \ldots, A_p, B_p, C_p, D_p .$$

We now introduce the generators of the group g. We term A_k conjugate to C_k, and B_k conjugate to D_k. We reflect the (u,v)-plane in the radical axis of A_k and C_k, and follow this by a reflection (in ordinary Euclidean sense) with respect to C_k. Let C_k^{-1} denote the arc C_k with sense changed. The product a_k of the above reflections is directly conformal and hence fractional linear. It carries the unit circular disc into itself and hence belongs to G'. Finally it carries A_k into C_k^{-1}. It will carry S into a polygon S_o exterior to S, but incident with S along C_k. Similarly there exists a transformation b_k of G which carries B_k into D_k^{-1}, carrying S into a polygon exterior to S but incident with S along D_k. We use a_k, b_k, $k = 1,\ldots,p$ to generate the group g.

Under transformations of g, H-distances are invariant, and S transforms into an H-congruent polygon. If T is a transformation of g, the image of a point set U under T will be denoted by $T\,U$ and the image of $T\,U$ under a transformation T_1 of g will be denoted by $T_1\,T\,U$, etc. In the product $T_1\,T$, the transformation T is thus to be applied first. With this understood we see that the polygons

$$a_k\ S,\quad b_k\ S,\quad a_k^{-1}\ S,\quad b_k^{-1}\ S$$

are indicdent with S along

$$C_k,\quad D_k,\quad A_k,\quad B_k$$

respectively. Suppose that $U = T\,S$, where T is contained in g, and that U is incident with S along a side of S. Let Z be any transformation of g. Z will carry U into a polygon

$$U' = Z\,U = Z\,T\,S \quad . \tag{10.1}$$

We shall use the principle exemplified by (10.1) to show that starting with S and proceeding in clockwise direction the neighborhood of the initial point P_1 of A_1 is covered by a sequence of polygons congruent to S under transformations of g. Thus the transformation a_1^{-1} carries S into the first polygon S_1 following S. Under a_1^{-1}, C_1^{-1} is congruent to A_1 so that D_1 transforms under a_1^{-1} to an H-segment D_1' issuing from P_1 and adjacent to A_1 on S_1. Since $b_1\,S$ is incident with S along D_1, $a_1^{-1}\,b_1$ carries S into a polygon S_2 incident with S_1 along D_1' in accordance with the principle exemplified by (10.1).

Under $a_1^{-1}\,b_1$, C_1 is congruent to an H-segment C_1' issuing from P_1 and adjacent to D_1' on S_2. Proceeding in this way we see that the successive transformations

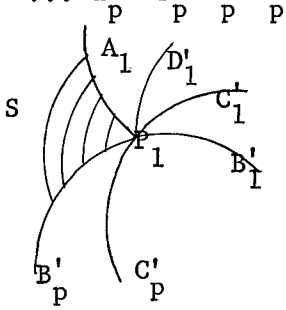
$$
\begin{aligned}
&a_1^{-1},\\
&a_1^{-1}\, b_1,\\
&a_1^{-1}\, b_1\, a_1,\\
&a_1^{-1}\, b_1\, a_1\, b_1^{-1},\\
&a_1^{-1}\, b_1\, a_1\, b_1^{-1}\, a_2,\\
&\cdots\cdots\cdots\cdots\cdots\\
&a_1^{-1}\, b_1\, a_1\, b_1^{-1}\, a_2 \cdots a_p^{-1}\, b_p\, a_p\, b_p^{-1}
\end{aligned}
\tag{10.2}
$$

carry S into a sequence of polygons each incident with its predecessor along an H-segment issuing from P_1. The lines of incidence are successively the images under the preceding transformations of

$$A_1\, D_1\, C_1\, B_1 \cdots A_p\, D_p\, C_p\, B_p\, A_1 \tag{10.3}$$

with initial points at P_1. The last transformation is accordingly the identity, that is,

$$a_1^{-1}\, b_1\, a_1\, b_1^{-1} \cdots a_p^{-1}\, b_p\, a_p\, b_p^{-1} = I.$$



The set of polygons incident as above with P_1 will be called the star Σ_1 incident with P_1. Let T be a transformation of g. If

$$P = T\, P_1, \qquad \Sigma = T\, \Sigma_1\ ,$$

Σ will be termed a star incident with P. The respective transformations (10.2) carry each vertex of S into P_1 as follows from the fact that the sides (10.3) each have images with initial points at P_1. We hereby adopt the convention that

S shall include the inner points of the sides A_k, B_k but not of the sides C_k, D_k, and one vertex. With this understood we shall establish the following theorem:

Theorem 10.1. The images of S under the respective transformations of g cover the H-plane once and only once.

We first note the following. If Q is an arbitrary point of S and P a vertex on the boundary of S at a minimum H-distance from Q, Q will be at least a positive H-distance d independent of Q from the boundary of the star of polygons incident with P and containing S.

(a) The images of S under the transformations of g cover the plane at least once.

Let O be the center of S, and M an arbitrary point of the H-plane. The points O and M can be connected by a curve C of finite H-length. Let s be the H-length measured along C from O. We shall cover C by a process Π defined as follows. We start by taking the point set S. If all points of C for which $s < s_0$ are thereby covered by inner points of S, and s_0 is maximal, we assign S as a neighborhood to each such point. Suppose that we have covered all points of C for which $s < s_1$, and assigned these points neighborhoods simply covered by polygons or stars of polygons. The point s_1 lies on the closure of one of our polygons S_1 already used and is at a minimum H-distance from some vertex P on the boundary of S_1. We add the star Σ of polygons incident with P and containing S_1. Suppose that the segment $s_1 \leq s < s_2$ of C is covered by inner points of Σ and s_2 is maximal. We assign Σ as a neighborhood to each point $s_1 \leq s < s_2$ not already assigned a neighborhood. If s_2 is not the last point of C, $s_2 - s_1 \geq d$, where d is the fixed constant previously defined. The process Π will thus lead after a finite number of

steps to a covering of C, and (a) is proved.

(b) The process Π applied to each curve of finite length leading from 0, leads to a unique covering of the H-plane $y > 0$.

The process Π can be regarded as a continuation process mapping the H-plane K on a new H-plane K', in a possibly multiple manner. The new plane K' is however simply connected, and the inverse mapping admits a locally single-valued continuation along any curve. It follows from the monodromy law that the inverse mapping is single-valued, and the proof of (b) and the theorem is complete.

The H-line through fixed points A, B of a hyperbolic transformation is called the axis A B of T.

Lemma 10.1. An axis A B of a hyperbolic transformation T of g is carried into an axis of a hyperbolic transformation by a transformation U of g.

In fact the points U(A) and U(B) are fixed points of the transformation

$$U\,T\,U^{-1}\,.$$

By an H-rotation about a point A through an angle a is meant an elliptic transformation in G in which the fixed points are A and its inverse B relative to the unit circle, and in which the circles through A and B have their initial directions at A rotated through the angle a in the counter-clockwise sense. Let S' be a polygon obtained from S by a transformation T of g. Under T the center 0 of S goes into a point A in S' which we shall term the center of S'. The radial lines leading from 0 to the vertices of S are H-congruent to H-lines leading from A to the corresponding vertices of the polygon S' containing A. These radial lines pass through the inverse B of A where this inverse is taken with respect to the unit circle. The vertices of S' are at constant H-distances

from A. An H-rotation through an angle π about A will carry S' into a polygon covering S', and the net N of all polygons obtained from S by transformations of g will be carried into a net N' covering N, polygon for polygon. This follows from the fact that any one polygon X of N determines the net. That is, given X, N can be constructed by the process of reflecting X in its sides, and of further reflections on the sides of the resulting polygons.

Lemma 10.2. If an H-transformation T of G carries the polygon net N into a net covering N, polygon for polygon, some power of T belongs to g.

Let X_m be the image of S under T^m. Let U_m denote the transformation of g which carries X_m into a polygon covering S. Then $U_m T^m(S)$ is a polygon covering S. The transformation $U_m T^m$ advances the vertices of S an integral number r of times, possibly zero. But r can have only the values $0,1,\ldots,4p-1$, so that for some two values of m, say m_1 and m_2, we must have

$$U_{m_1} T^{m_1} \equiv U_{m_2} T^{m_2} ,$$

or

$$U_{m_2}^{-1} U_{m_1} \equiv T^{m_2 - m_1} .$$

Thus a power of T is in g, whence the lemma is proved.

Theorem 10.2. The H-line joining any two centers P and Q of polygons of the net N is the axis of a transformation of g.

Let L be the H-line through P and Q and A and B its ideal end points. Let T be an H-rotation about P through 180°. T carries L into itself interchanging A and B. Under T, the net N is carried into a covering net. Let U be a similar rotation about Q. The product UT carries N into a covering net. It leaves A and B fixed and is accordingly hyperbolic with axis L. By virtue of the preceding lemma some power of UT is in the group g and the proof is complete.

Corollary. There exists a hyperbolic transformation with fixed points arbitrarily close to two given points on the unit circle C.

The H-diameters of the polygons of N are constant so that their Euclidean diameters tend to zero as their Euclidean distances from C tend to zero. Each point of C is a cluster point of centers of polygons of N. If A and B are points of C, the H-lines through centers sufficiently near A and B respectively will tend to the H-line with ideal end points A and B, and the ideal end points will tend to A and B. The proof is hence complete.

Corresponding to an ordered pair of polygons X, Y of N incident along a side, we introduce a generator of g or its inverse as follows. Under a unique transformation of g, X goes into S. Then Y goes into a polygon

$$Y' = T(S)$$

adjacent to S. The transformation T is thus uniquely determined by X and Y. We make XY correspond to T. Any pair X'Y' obtained from X, Y by a transformation of g will determine the same transformation T. More generally, let

(10.4) $$X_0, X_1, \ldots, X_n,$$

be a sequence of polygons each of which, excepting X_0 is incident with its predecessor along a side. If $X_i X_{i+1}$ corresponds to T_{i+1} we say that the sequence (10.4) determines the sequence

(10.5) $$T_1 \ldots T_n .$$

Regarded as a transformation of g, the product (10.5) is applied beginning with T_n, and proceeding to the left.

Lemma 10.3. If the sequence of polygons (10.4) determines the sequence of transformations (10.5) and begins with S, then

(10.6) $$X_n = T_1 \ldots T_n(S) \quad .$$

The lemma is true of $n = 1$ by virtue of the definition of T_1. We assume therefore that

$$X_{n-1} = T_1 \ldots T_{n-1}(S),$$

and seek to prove (10.6). By virtue of the definition of T_n

$$(T_1 \ldots T_{n-1})^{-1} X_n = T_n(S)$$

from which (10.6) follows.

11. Minimum polygon paths. A sequence of polygons of the form (10.4) will be called a polygon path joining X_o to X_n. We understand that each polygon except X_o is incident with its predecessor along one side as noted in the preceding section. The number $n+1$ will be called the length of the path. If for X_o and X_n fixed, the path has the minimum length, it will be called an M-path.

If an M-path M_1 belongs to a vertex star Σ, its length L is at most $2p + 1$. If $L < 2p + 1$, the polygons of M_1 are uniquely determined in Σ, and follow one another in one of the circular orders of the polygons of Σ. If $L = 2p + 1$, the end polygons are H-diametrically opposite one another with respect to the vertex, and admit two M-paths in Σ. We term these two paths alternate vertex passes. They contain only their end polygons in common, and the circular orders of their polygons are opposite in Σ. On the boundary of a vertex pass there are two sides incident with the vertex P of the star. They make an angle of $\pi + \frac{\pi}{2p}$ interior to the pass. Such an interior angle will be called a pass angle.

Note that an H-line k, an arc h of which forms a side of a polygon, is composed of a succession of arcs each of which constitutes sides of polygons. This is true for arcs of k incident with a vertex P of h by virtue of the H-symmetry of a vertex star with respect to its vertex, and is then seen to be generally true. The H-line k thus divides the polygon net into two subnets composed of polygons.

Theorem 10.3. A polygon path R belonging to a vertex star with length $L \leq 2p + 1$ is an M-path and is unique except when $L = 2p + 1$. In case $L = 2p + 1$ the end polygons admit no other joining M-path save the alternate vertex passes.

Let P be the given vertex. The $4p$ polygon sides issuing from P, if extended divide the polygon net into $4p$ subnets. A polygon path joining the end polygons of R must have polygons in L of these subnets. If such a path leaves the vertex star it must leave on the same subnet and contain at least $L + 1$ polygons. The theorem follows at once.

Theorem 10.4. The boundary of an M-path M is a simple closed curve C.

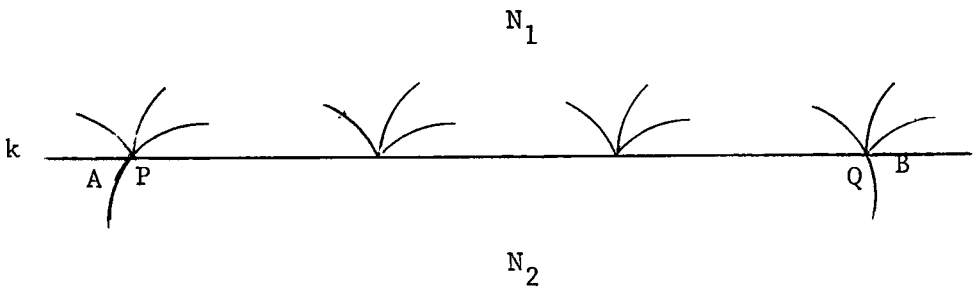
Let P be a vertex on the boundary of M. By virtue of the preceding lemma, M contains at most $2p + 1$ of the polygons incident with P, and must contain these in one of the circular orders around P. The point P accordingly cannot be multiple on g and the theorem follows directly.

Let P be a vertex on the boundary B of an M-path. Suppose the sides incident with B at P make an angle a interior to the path. The angle a will be termed convex, straight, or concave, according as a is less than, equal to, or greater than π. We shall refer to convex or concave angles as corners of B.

Theorem 10.5. The corners on the boundary B of an M-path M are either convex or pass angles. No two successive corners can be pass angles.

The first statement follows from Theorem 10.3.

Let P and Q be successive corners on the boundary of M. The boundary of M between P and Q (in one of its senses) consists of an arc h of an H-line k. The H-line k divides the polygon net into two nets N_1 and N_2. If P and Q are pass corners, the given M-path contains as a subpath M_o all of the polygons incident with h in one of the two nets N_1 and N_2, say N_1. Preceding M_o, M contains a polygon A in N_2 incident with P and k, and following M_o, M contains a polygon B in N_2 incident with Q and k. But A and B can be joined by a shorter path consisting of all polygons in N_2 incident with h. Hence P and Q cannot both be pass corners.



12. Pseudo-convex regions. For present purposes it will be convenient to regard the polygons of our net as closed. The point set sum of a finite or infinite set of such polygons will be called a net region. An M-path is in particular a net region. As we have seen, the boundary of an M-path satisfies the two conditions:

(12.1) Its corners are either convex or pass corners.

(12.2) No two successive corners on a boundary are pass corners.

We shall say that a net region R is polygon connected if any two polygons of R can be joined by a polygon path on R. A net region whose boundary satisfies (12.1) and (12.2) and which is polygon connected, will be termed pseudo-convex.

Lemma 12.1. A pseudo-convex region R whose boundary is a simple closed curve g consists of the polygons interior to g.

Let W be the region of polygons on the outside of g. Suppose that W is pseudo-convex. The corners of g would then satisfy the conditions (12.1) and (12.2) relative to W. Let $a_1,\ldots,a_n$ be the angles at the corners of g on the inner side of g. Let $a = \pi/2p$. By virtue of (12.1)

$$a_i = \pi + ma \ , \tag{12.3}$$

where $m = -1$ at a pass corner and $0 < m < 2p$ at a convex corner. Now condition (12.2) implies that there are at least as many corners at which m is positive as negative so that

$$\Sigma\, a_i \geq n\pi > (n-2)\pi \ ,$$

contrary to Gauss's Theorem [Theorem 2.2] applied to the interior of g.

Lemma 12.2. The boundary g of a connected pseudo-convex net region R is composed of simple arcs possibly infinite in length.

If g has any multiple points there exists a subarc h of g which is a simple closed curve whose interior I is free from arcs of g, and which has just one of the multiple points of g on it. Let this point be denoted by P. Let α be the interior angle at P, and let $a_1,\ldots,a_n$ be the remaining interior angles on h. Suppose that R is not within h. Again, let $a = \pi/2p$. Then $a_i \geqq \pi - a$, and in fact $a_i \geqq \pi + a$ at least at alternating angles a_i. Hence

$$\Sigma\, a_i \geqq n\pi - a \ ,$$

$$\alpha + \Sigma\, a_i \geqq n\pi \ ,$$

whereas by Gauss's Theorem

$$\alpha + \Sigma\, a_i < [(n+1) - 2]\pi = (n-1)\pi \ .$$

We infer that R is within h, and the lemma is proved.

Let R be a pseudo-convex region. By a <u>boundary path</u> B of R is meant a polygon path either wholly exterior or wholly interior to R, consisting of successive polygons incident with successive sides and vertices of one of the boundary arcs b of R. It is understood that a mapping of the polygons B onto the edges or vertices of b is thereby given in which each polygon <u>corresponds</u> to an incident side or vertex of b in such a manner that the order of the polygons of B is the same as the order of the corresponding edges and vertices of b. Suppose, in particular, that the end polygons X' and X'' of B correspond on b to sides b' and b'' on b respectively, and that B is within R. There will exist a second boundary path B^* of R exterior to R with polygons incident and corresponding to the same sides and vertices of b as B, several polygons possibly going into the same vertex.

The length of B^* is at least that of B except possibly in the case where the arc h of b from b' to b'' contains one more pass corner than convex corner. In the latter case, and when there are just $2p-1$ polygons at

each convex corner

(12.4) $$L(B^*) = L(B) - 2$$

where $L(X)$ means the length of X. In this case

$$L(X' B^* X'') = L(B)$$

and the paths $X' B^* X''$ and B will be termed conjugate.

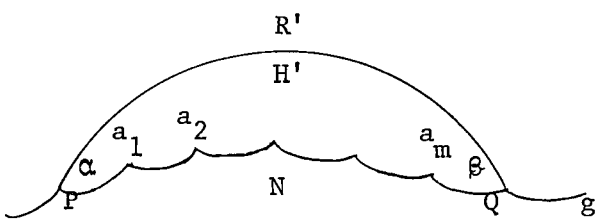
A polygon path joining two polygons A and B will be termed minimizing if its length is a minimum among lengths of polygon paths joining A to B.

Theorem 12.1. If A and B are two polygons on a pseudo-convex region R the minimizing polygon paths joining A to B include at least one path on R.

The case where R is the whole hyperbolic plane is trivial. In any other case the boundary B of R consists of a set of simple curves finite or infinite, and non-intersecting. Let g be one of these curves. The curve g divides the polygon net into two nets N and N'. If g is closed, one of these nets, say N, is finite and includes R. In case g is not closed we again suppose that N includes R.

Let H be the set of hyperbolic rays of the net of N' with first end point on g, starting with a side of a polygon on N' not on N. That no one of these rays meets g a second time can be established by means of Gauss's Theorem as follows.

We assume that a ray R' issuing from g at a point P meets g again at a second point Q. Then g and R' form a geodesic polygon H'. Let the interior angles of H' be labeled $\alpha, a_1, \ldots, a_n, \beta$ as indicated in the following figure:



The interior angles of H' are ordered in such a manner that as one traverses g from P to Q one passes the vertices of the angles $\alpha, a_1, \ldots, a_n, \beta$ in this order. As in the proof of Lemma 12.2 we have

$$\Sigma\, a_i \geqq n\pi - a \quad ,$$

where $a = \pi/2p$, whence

$$\alpha + \beta + \Sigma\, a_i \geqq n\pi - a + \alpha + \beta \quad . \tag{12.5}$$

Since $\alpha \geqq a$, $\beta \geqq a$ we have $\alpha + \beta \geqq 2a$, and (12.5) implies the following inequality

$$\alpha + \beta + \Sigma\, a_i \geqq n\pi + a \quad .$$

By Gauss's Theorem

$$\alpha + \beta + \Sigma\, a_i \leqq n\pi \quad ,$$

whence we have arrived at a contradiction. ¶ These rays H may be given a circular order consistent with the order of the vertices of g from which they emanate, and when emanating from a common vertex P an order consistent with their angular order about P. So ordered it is seen that consecutive rays do not intersect, nor rays emanating from the same or consecutive vertices except possibly at the initial vertex. It follows that no two or these rays intersect except at initial vertices. The rays of H thus divide N' into a set of subnets N_i which follow one another in the order of the rays of H.

Let z be a path joining a polygon A to a polygon B of N but not lying wholly in N. There must exist a subpath x of z which lies wholly in N' except for end polygons X' and X''. By virtue of the manner in which the rays divide up N' into successive polygon nets N_i there must be a boundary path B^* of N' which contains one polygon in each net N_i which contains a polygon of x. The path

(12.6) $X' B^* X''$

is thus as short as x. But, as we have seen, the path (12.6) is longer than a boundary path of N joining X' to X'' unless (12.6) is conjugate to this boundary path in which event the two paths are of the same length.

Thus the minimum paths joining A to B include one in N. Since an arbitrary minimum path joining A to B meets at most a finite number of the bounding arcs of R, the theorem follows:

Theorem 12.2. The corner conditions (12.1) and (12.2) are necessary and sufficient that a polygon path be a minimum path.

We have already seen in §11 that these conditions are necessary. That they are sufficient follows from Theorem 12.1 by identifying this path with R as a pseudo-convex net region.

A polygon path finite or infinite, every finite sub-path of which is an M-path, will also be termed an M-path.

We shall say that a polygon connected net region R is r-convex, $0 < r \leq 2p$, if the number of consecutive polygons of R incident with a boundary vertex is at most r.

We follow with some examples of M-paths.

Example 1. Let g be an H-line which emanates from the origin and bisects two opposite sides of the central polygon S. The H-line g will meet no vertex of the net but will continue through successive centers of polygons of the net. The polygons incident with g will form an M-path not only pseudo-convex but 2-convex.

Example 2. If g is an arbitrary H-line of the net, the set of polygons incident with g on one side will form an M-path. The second boundary curve g' will be 2-convex, the whole M-path 2p-convex.

Example 3. Let g_1 and g_2 be two H-rays of the net emanating from a point P and forming at P an angle $\pi + \pi/2p$. Let g be the curve composed of g_1 and g_2. The polygons of the net on an arbitrary side of g will form an M-path. Those on the side of the pass angle will form a pseudo-convex region, as well, of course, as those on the other.

Example 4. Let g be an infinite sequence of finite arcs of H-rays of the net making angles on one side which are alternatingly $\pi \pm \pi/2p$. The polygons incident with g on either side will form an M-path.

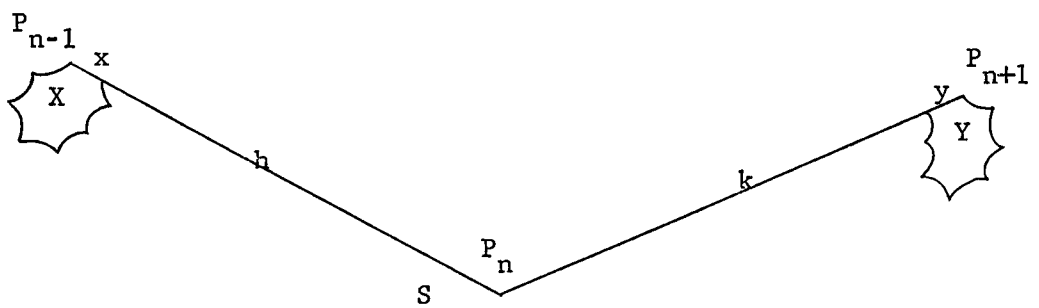
13. Combinatorially convex net regions. A boundary arc b of a 2p-convex net region will be termed convex at infinity if it possesses no infinitely long geodesic subarcs. A net region whose boundary arcs are 2p-convex and convex at infinity will be termed combinatorially convex, written co-convex. A net region R with a property A will be termed a minimal region with property A if no proper sub-net region of R possesses the property A.

Theorem 13.1. Corresponding to a pseudo-convex net region S there exists a minimal co-convex net region R containing S.

The case where S contains every polygon is trivial. We begin with the case where each boundary arc b of S is of infinite H-length and contains no infinite subarc without corners. Let

(13.1) $$\cdots P_{-1} P_0 P_1 \cdots$$

be the successive corners of S on b. Suppose that P_n is a pass corner and let h and k be respectively the subarcs of b determined by the point pairs $P_{n-1}P_n$ and P_nP_{n+1}. On the arc hk let x be the first net side and y the last net side.



Let X and Y be respectively polygons with sides x and y. Let B be the interior boundary path of S which joins X to Y. As we have seen in §12, there is a conjugate M-path of the same length joining X to Y of the form

(13.2) $$W = X\, B^*\, Y\,.$$

More specifically, if X^* and Y^* are the polygons exterior to S with x and y as

sides respectively B^* joins X^* to Y^* and contains all polygons not in S incident with P_n.

We add B^* to S to form a new net region S^*. We say that S^* is derived from S, corresponding to the pass corner P_n. We term B^* the canonical addition corresponding to P_n. S^* is infinite with S.

We see that S^* is pseudo-convex. In fact the corners of S^* are identical in position and character with those of S except for the new corners of S^* on B^*, and corners at P_{n-1} and P_{n+1}. The new corners of S^* on B^* are at most 2-convex. The vertices P_{n-1} and P_{n+1} are at most $(2p-1)$ convex relative to S since P_n is a pass corner. Hence S^* is at most $2p$-convex at P_{n-1} and P_{n+1} so that S^* is pseudo-convex. S^* retains the pass corners of S excepting P_n. It is clear that any $2p$-convex net region which contains S must contain all polygons incident with P_n, hence all polygons of B^*, and accordingly all polygons of S^*.

We shall now construct a minimal co-convex net region R containing S. If S possesses no pass corners we can take R to be S. Suppose then that there are pass corners P_i on the boundary arc b. We order these pass corners in the order of the absolute values of their subscripts taking a positive i before a negative i. Corresponding to the first such pass corner P_r on b, let S_1 be the derived net-region on which P_r is eliminated. Let b_1 be the boundary of S_1 replacing b. Corresponding to the first pass-corner P_s of S_1 on b_1 let S_2 be the net-region derived from S_1 so as to eliminate P_s.

Proceeding inductively, let

$$S\ S_1\ S_2\ \ldots$$

be a sequence (finite or infinite) of pseudo-convex net regions of which S_n is derived from S_{n-1} so as to eliminate the pass corner P_i of lowest order

on the boundary of S_{n-1}. On the net region

$$S(b) = S + S_1 + S_2 + \ldots$$

the boundary b is replaced by a boundary with convex corners. Moreover, any minimal 2p-convex net region containing S must contain S(b).

The sum R of the net regions S(b) corresponding to the different boundary arcs b of S will be a minimal 2p-convex net region containing S.

The case where S is bounded by a single closed arc b is similarly treated. In the case where a pass corner P_n is preceded (or followed) by an unending boundary geodesic h the canonical addition corresponding to P_n shall include all polygons exterior to S and incident with h, and including a set of polygons defined as before and incident with the geodesic boundary arc k on the other side of P_n in b.

There are also "canonical" additions corresponding to a point or points at infinity on a boundary arc b. These arise when b is a geodesic h or contains a subarc h which is an infinite geodesic arc with a finite convex terminal corner P. The net B^* shall then consist of all polygons exterior to S and incident with h including, however, only that one of the exterior polygons incident with P (if P exists) which has an edge on h.

In the general case one constructs R as previously, concluding with canonical additions corresponding to the points at infinity on the respective boundary arcs so as to obtain an R which is "convex at infinity". Q.E.D.

The minimal co-convex net region R containing an unending M-path M has the following properties.

(a) The interior boundary paths of R are M-paths (which may coincide).

(b) Each polygon of R is incident with M.

(c) R is the sum of its two interior boundary paths.

(d) If M' is an arbitrary unending M-path on R, R is the minimal co-convex net region containing M'.

Proof of (a). We begin by showing that if B^* is a canonical addition to M corresponding to a finite pass corner P_n and if we set $S^* = M + B^*$ as above, then the interior boundary path M^* of S^* which contains B^* is an M-path.

To that end we first note that W in (13.2) is pseudo-convex. In particular W is $(2p-1)$-convex at P_n, and P_n is preceded and followed respectively by two pass corners of W, namely the last vertex of x preceding P_n and the first vertex of y following P_n. The case where x terminates at P_n or y begins at P_n is exceptional but offers no difficulty. The remaining corners on the boundary of W are at most 2-corners so that W is pseudo-convex as stated.

We have seen in the preceding proof that S^* is pseudo-convex. By virtue of Theorem 12.1 the path M^* is an M-path if each finite sub-path H of M^* is minimizing relative to paths on S^*. But H could be shortened on S^* only if some sub-path of B^* failed to be an M-path, and this is impossible since B^* is pseudo-convex. Hence M^* is an M-path.

The same conclusion can be drawn when M^* is obtained from M by a canonical addition corresponding to a point at infinity.

Given a boundary arc b of M the successive canonical additions incident with b will yield successive boundary M-paths

(13.2) $\qquad M_1^*, M_2^*, \ldots \qquad$ (possibly finite in number)

as interior boundary paths. The minimal co-convex region R will have an interior boundary path M^*, each finite sub-path of which will be a sub-path of each of the paths of (13.2) for r exceeding some integer r_o. Hence M^* will be an M-path.

This completes the proof of (a).

Proof of (b). Each canonical addition is incident with one of the two boundary arcs of M and hence with M.

Proof of (c). R is the union of the polygons of M and of the canonical additions. If each of these polygons has a vertex on the boundary βR of R (c) is true.

A polygon of M has at least 8 vertices and at least 4 which are 1-convex on M. A 1-convex vertex becomes at most a 3-convex because of the canonical additions and so must be on βR.

A polygon A of a canonical addition B^*, adjoined to a pseudo-convex-region S to form S^*, has at least four 1-convex vertices on S^* which become at most 2-convex by virtue of subsequent additions and so are on βA.

Hence (c) is true.

Proof of (d). Let R' be the minimal co-convex net region containing M'. Since R' is minimal, $R' \subset R$. It remains to show that $R \subset R'$.

Suppose then that R' is a proper sub-region of R. One of the interior boundary paths M^* of R must fail to be wholly contained in R'. Since R' is co-convex and contains every finite M-path whose end polygons are in R', an infinite subsequence H of polygons of M^* must be exterior to R'.

Each polygon A of H is incident with one of the two boundary arcs b of R' by virtue of property (c). Hence H is an exterior boundary path of R'. But H is pseudo-convex and R' co-convex. Hence b can have at most one corner incident with H. Thus b contains an infinite geodesic arc contrary to the fact that R is convex at infinity. Thus $M^* \subset R'$ and (d) follows.

Note that property (d) depends definitely upon the conventions of convexity at infinity without which (d) would be false.

A minimal co-convex net region containing an M-path will be called a ribbon R. According to the preceding argument the ribbon R is uniquely determined by each of its M-paths.

14. Ribbons and geodesics. We shall prove the following theorem:

Theorem 14.1. There is a 1-1 correspondence between geodesics and ribbons on the hyperbolic plane in which each geodesic corresponds to the ribbon in which it is contained.

We begin by proving the following:

(i) Each ribbon R contains one and only one geodesic g.

Let M be an M-path in R with successive polygons

$$\cdots X_{-1} X_0 X_1 \cdots$$

Let P_i be a point on X_i. Since R is 2p-convex, there exists a geodesic arc g_n on R joining P_{-n} to P_n. Let E_n be the element on g_n at a point on g_n nearest X_0. The elements E_n have at least one cluster element E. The geodesic g determined by E will clearly be in R.

Moreover, if g' is a second geodesic in R, $g' = g$. For two H-lines which remain at a finite distance from one another are identical.

The proof of (a) is complete.

We continue with a proof of the following:

(ii) Each geodesic lies in one and only one ribbon R.

Suppose that g is sensed. We refer to the two sides of g as positive and negative respectively. Corresponding to g we shall define a "positively related" M-path M which contains g.

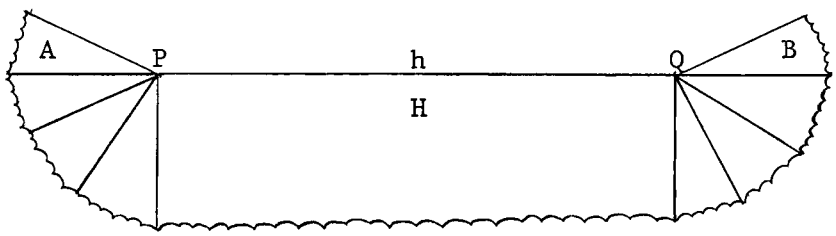
The path M shall include each polygon whose interior is met by g, and each polygon on the positive side of g whose boundary is met by g. These polygons will be given an order consistent with their intersections with g. In case g includes the side of a polygon, M will consist of all polygons incident with g and on the positive side of g. M will then be an M-path and its polygons will be correspondingly ordered. In case g passes through a vertex but contains no side incident with P, the polygons of M incident with P will be taken in M in their circular order about P. The polygons of M are then simply ordered, and M is a path. We finish by proving the following:

(A) The path M is an M-path.

We shall establish A by showing that M is pseudo-convex. To that end recall that a polygon, and a vertex star of polygons, are both convex. Let V be a vertex star into whose interior g enters. Each set of r consecu-

tive polygons of V is convex provided $r \leq 2p$. Hence when g does not pass through the vertex O of the star V, g can intersect at most 2p+1 polygons of V. If g passes through O and is not a net geodesic, g meets at most two polygons of V in points other than O. In each case M contains at most 2p+1 polygons of V.

Two successive corners P and Q on the boundary of M cannot be pass corners. To prove this let h be the geodesic arc joining P and Q and let H be the sub-path of polygons of M incident with h. Let t be the geodesic of which h is a sub-arc. The geodesic t divides the net into two subnets in one of which, say N, the two end polygons A and B of H lie, and in the other of which the remaining polygons of H lie (assuming that P and Q are pass corners).



Now g cannot coincide with t since M has polygons on both sides of t. It has points a and b on A and B respectively, of which at least one, say a, is not on t. But A and B can be joined by a convex path X in N containing the polygons incident with h and in N. The segment ab of g thus meets the polygon of X which follows A in an inner point so that X must be in M contrary to hypothesis.

We infer that no two successive corners of M are pass corners. The path M is accordingly an M-path, and the proof of (A) is complete.

The geodesic g thus lies in the ribbon R containing M. It cannot be in any other ribbon R'. For g is necessarily interior to R', so that R' would contain M. By virtue of (d), R' = R, and the proof of (ii) is complete.

Theorem 14.1 follows from (i) and (ii).

15. Right M-paths. An unending sensed simple curve b with ideal end points A and C on the unit circle will divide the H-plane into two regions. Of these regions the one which contains the infinite neighborhood of an ideal end point B such that A B C follow in counterclockwise order, will be termed the right region, and the other region the left region. Suppose that the two boundaries of an M-path M or ribbon R are consistently sensed. The boundary b in whose left region R or M lie, will be termed the right boundary. The left boundary is similarly defined.

Let M be an M-path on a ribbon R. Suppose the boundary arcs a and b of R have been similarly sensed and lead from an ideal end point A to an ideal end point B. M will be said to be similarly sensed if its polygons are successively numbered so that these numbers become positively or negatively infinite as A and B respectively are approached. A ribbon whose boundaries and M-paths have been similarly sensed will be said to be sensed.

A sensed M-path whose right boundary is co-convex will be termed a right M-path. A left M-path is similarly defined. A sensed M-path may be both a right and left M-path. We shall prove the following theorem:

Theorem 15.1. There is a 1-1 correspondence between sensed ribbons R and right (or left) M-paths M in which M and R have the same right boundary.

Given a sensed ribbon R the inner boundary path which has the same right boundary as R is a right M-path in R. Conversely, given a right M-path M the usual construction of the ribbon R which contains M shows that M and R have the same right boundary. Moreover, R uniquely determines M and

M uniquely determines R.

We shall now give a symbolic characterization of right M-paths or rather of the set of right M-paths congruent to a given right M-path. To that end let X Y be an arbitrary pair of polygons with a common side. Recalling that S_0 is our central polygon, let S_0S be a pair congruent to X Y. The transformation T such that

$$S = T(S_0)$$

will be used to represent the pair X Y and all pairs congruent to X Y. Observe that S S_0 is represented by T^{-1} since S S_0 is carried by T^{-1} into the pair

$$S_0,\ T^{-1}(S_0) .$$

Y X is thus represented by T^{-1}.

More generally a path

(15.1) $$\cdots X_{-1}\ X_0\ X_1 \cdots$$

containing a finite or infinite number of polygons, will be represented by the sequence

(15.2) $$\cdots t_{-1}\ t_0\ t \cdots$$

of generators of the group g [see §10] and their inverses in which t_i represents the pair $X_i\ X_{i+1}$. In particular, the polygons incident with a vertex taken in the clockwise circular order form a path for which the corresponding transformations are in the circular order

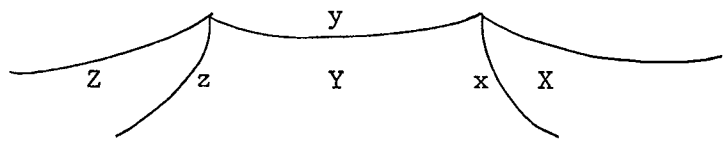
(15.3) $$a_1^{-1}b_1a_1b_1^{-1} \cdots a_p^{-1}b_pa_pb_p^{-1} .$$

See (10.2). We term (15.1) the circular order C. Taken in the anti-clockwise order the same polygons lead to the circular order C^{-1}, namely

(15.4) $$b_pa_p^{-1}b_p^{-1}a_p \cdots b_1a_1^{-1}b_1^{-1}a_1 .$$

We shall have occasion to refer to paths composed of three polygons X Y Z, of which X and Z are incident with Y along sides x and z respectively

which are separated on the boundary of Y by just one side y.



If $x\ y\ z$ appear on the boundary of Y in counterclockwise order, the transformation representing $X\ Y\ Z$ must be one of the pairs

$$a_k a_k,\quad b_k b_k,\quad a_k^{-1} a_{k+1}^{-1},\quad b_k^{-1} b_{k+1}^{-1} \tag{15.5}$$

where k is reduced mod p to one of the numbers $1,\ldots,p$. Cf. §12.

In case x, y, z appear in clockwise order on Y, (15.5) must be replaced by the inverse pairs

$$a_k^{-1} a_k^{-1},\quad b_k^{-1} b_k^{-1},\quad a_{k+1} a_k,\quad b_{k+1} b_k\ . \tag{15.6}$$

Suppose that (15.1) is a right path and (15.2) the corresponding group representation. A maximal sub-block of (15.2) in which the elements appear in the circular order C (or C^{-1}) will be called a C-block (or C^{-1}-block). Two successive C-blocks

$$(t_r \ldots t_s)(t_{s+1} \ldots t_m)$$

in which $(t_s t_{s+1})$ is one of the pairs (15.5) will be termed <u>related</u>. Related C^{-1}-blocks are similarly defined by using (15.6).

<u>The conditions on</u> (15.2) <u>that it represent a right</u> M-<u>path</u> are that $t_i t_{i+1} = I$ for no i, where I denotes identity, together with the following:

(1) C^{-1}-blocks have lengths at most $2p$.

(2) There exists no finite subsequence of successively related C^{-1}-blocks in which the initial and final C^{-1}-blocks are $2p$-blocks while the intermediate C^{-1}-blocks are $(2p-1)$-blocks (or absent).

(3) C-blocks have length at most $2p-1$.

(4) There exists no infinite sequence of successively related C-blocks each of which is a $(2p-1)$-block.

Conditions (1) and (2) insure that the left boundary of the M-path be pseudo-convex, while conditions (3) and (4) insure that the right boundary be co-convex.

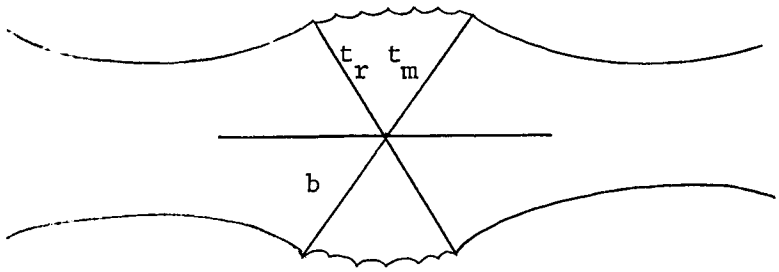
16. Symbolic elements. Let H be a right path and (H) the class of right paths which are congruent to H under operations of g. The class (H) determines a symbolic trajectory T whose indexed representations (t) have the form

$$\cdots t_{-1}\, t_0\, t_1 \cdots \tag{16.1}$$

where t_i is a generator of the group or its inverse. We admit only those indexed trajectories which are determined by right paths. As in §3, an element e on (t) is determined by (t) and a preferred symbol t_r. We then write $e = e(t,r)$. The net sides separating the successive polygons of H can be indexed so that the r-th side separates two polygons represented by t_r. We shall refer to this net side as the side t_r. The element $e(t,r)$ will be represented by H and the preferred net side t_r.

The elements e are assigned the metric used in §3.

Let R be a sensed ribbon determined by H. A net side t_r of H which reaches from boundary to boundary of R will be called a transversal of R. If t_r does not reach the left boundary of R, it shall be extended to the left boundary of R by a net side b. If several choices of b are possible, we shall take that net side whose terminal point is furthest advanced on the left boundary of R. The net side t_r followed by b will also be called a transversal of R. Other net sides t_m may terminate in the same end point as t_r, and will then be followed by b to make another transversal of R.



We shall have occasion to identify points which are congruent under transformations of g. The hyperbolic plane M thereby reduces to a manifold M^* of genus p. Corresponding to each geodesic h on M there will be a geodesic h^* on M^*, and corresponding to each element E on M an element E^* on M^*. In general the asterisk added to a symbol x will indicate the correspondent of x on M^*.

17. The geodesic element F(e). Let e be a symbolic element e(t,r) based on (t) and let R be a sensed ribbon corresponding to (t) with a net side t_r corresponding to the operator t_r. Let h be the geodesic on R, and on h let E be the geodesic element at the intersection of h with the transversal bearing t_r. We term the geodesic element E^* the geodesic element corresponding to e and write $E^* = F(e)$. Different symbolic elements e based on (t) may determine transversals with common final net side and so may determine the same geodesic element E^*. The relation $E^* = F(e)$ is accordingly not one-to-one in general. See figure.

We shall prove the following:

(a) The geodesic element F(e) varies in a uniformly continuous manner with the symbolic element e.

Let $e = e(t,o)$ and $e' = e(t',o)$ be two symbolic elements with the symbolic trajectories (t) and (t') represented by ribbons R and R'

respectively. Without loss of generality we can suppose that the net sides t_o and t'_o are the same. Let m be a positive integer. If the distance

$$e\ e' < \frac{1}{2m+1}$$

the net sides $t_{-m},\ldots,t_m$ will be identical with the net sides $t'_{-m},\ldots,t'_m$. Let D be the H-diameter of a polygon. The geodesics h and h' in R and R' respectively will remain at an H-distance at most $2D$ from each other for an arbitrarily large distance L from $t_o = t'_o$ provided m is sufficiently large. But if L is sufficiently large the distance $F(e)\ F(e')$ between the geodesic elements $F(e)$ and $F(e')$ will be less than a preassigned constant η. Thus $F(e)\ F(e') < \eta$ provided $e\ e'$ is sufficiently small, and the proof of (a) is complete.

Recall that the distance between two elements $e(t,r)$ and $e(t',r')$ is infinite unless the preferred symbols t_r and t'_r are equal. With this understood we state the following:

(b) <u>For the subset Σ of elements e at a finite distance from a given element e_o the relation $E^* = F(e)$ is one-to-one without exception</u>.

Let $e(t,o)$ and $e'(t',o)$ represent any two symbolic elements in the set Σ. Let R and R' be respectively ribbons determined by (t) and (t') with common net side $t_o = t'_o$. Let h and h' be the transversals determined respectively by t_o and t'_o, and let E and E' be the elements with initial points on h and h' determined respectively by e and e'. We wish to show that E and E' are congruent only when $(t) = (t')$.

Suppose then that E and E' are congruent. If the initial points of E and E' are on $t_o = t'_o$, $E = E'$ and $(t) = (t')$. But the transversal extensions of t_o and t'_o have no mutually congruent points save their end points, so that the initial points of E and E' in $\mathring{R}$ and $\mathring{R}'$ must be on $t_o = t'_o$.

Statement (b) follows.

The function inverse to $F(e)$ is discontinuous at certain elements E defined as follows. Let a right path H whose left boundary contains an infinite geodesic arc be termed special. We apply the same term to any symbolic trajectory (t) determined by H, to any symbolic element e on (t), or to the geodesic element $F(e)$. The function $F(e)$ is discontinuous at each special geodesic element E. This may be briefly explained as follows: Let h be the geodesic determined by a special right path. Any finite segment of h can be approximated arbitrarily closely by a geodesic k for which the corresponding right path has arbitrarily long geodesic arcs. Elements $F(e)$ on k can accordingly be arbitrarily close to an element $F(e_o)$ on h without e tending to e_o. This difficulty can be circumvented as follows.

A right path on whose right boundary there are at most m successive 2p-vertices will be said to be in the class m. The corresponding geodesic, and symbolic trajectory t and elements e on t will also be said to be in class m. It is clear that admissible elements e of class m form a compact set. The map $E^* = F(e)$ is locally 1-1 and $F(e)$ is continuous. We accordingly have the following theorem.

Theorem 17.1. The relation $E^* = F(e)$ between symbolic elements e of class m and their images $E^* = F(e)$ is locally 1-1 and bicontinuous

This follows from the theorem that a 1-1 map of a compact set A into a metric space B which maps A continuously on B also maps B continuously on A. See Kerekjarto, Verlesungen über Topologie, p. 34.

Theorem 17.2. A necessary and sufficient condition that a geodesic h^* be periodic is that the corresponding symbolic trajectory t be periodic.

Suppose t is represented by the indexed trajectory (t)

$$\cdots\ t_{-1}\ t_0\ t_1\ \cdots \tag{17.1}$$

and by the right path

$$\cdots\ S_{-1}\ S_0\ S_1\ \cdots \tag{17.2}$$

Suppose that $t_{n+r} = t_r$ for a fixed $n > 0$ and every r. The pairs S_0S_1 and S_nS_{n+1} are represented by t_0 and t_n respectively. But $t_0 = t_n$ so that these pairs are congruent under a transformation T. Thus

$$S_n = T(S_0), \qquad S_{n+1} = T(S_1)\ .$$

Similarly, the pairs S_1S_2 and $S_{n+1}S_{n+2}$ are congruent. It follows that

$$S_{n+2} = T(S_2)\ .$$

Proceeding inductively we find that

$$S_{n+r} = T(S_r) \qquad [r = \ldots,-1,0,1,\ldots]. \tag{17.3}$$

Let R be the ribbon containing the path (17.2), and let h be the geodesic in R. It follows from (17.3) that $R = T(R)$. But the geodesic $T(h)$ lies in $T(R)$, and hence in R. Hence $h = T(h)$ and the condition is proved sufficient.

Conversely, we suppose that h is a periodic geodesic contained in a ribbon R; that is, the image of a closed geodesic h^* on M^*. There accordingly exists a transformation T such that $h = T(h)$. The geodesic $T(h)$ is in $T(R)$ and in R. Hence $R = T(R)$. It follows that the symbolic trajectory t determined by h is periodic, and the proof of the theorem is complete.

Lemma 17.1. An admissible symbolic trajectory t which is in no finite class m has a periodic limit trajectory (a).

Let $h_1,\ldots,h_{2p}$ be the periodic geodesics determined by the 2p pairs of congruent sides of the central polygon S_0. Let π_i be the left boundary path determined by h_i, and T_i the corresponding symbolic trajectory. Let π be a right boundary path determined by t. Corresponding to each positive integer m there exists a sequence of m+1 net sides on the right boundary of π with (2p)-vertices forming a geodesic arc congruent to a subarc of one of the geodesics h_i. The sub-block s_m of t representing the inner net sides of π incident with these (2p)-vertices will be a sub-block of T_i. Since there are only 2p trajectories T_i there exists a subsequence (S'_n) of the blocks s_m and a trajectory T_k such that each S'_n is a sub-block of T_k. The symbolic trajectory T_k is then a limit trajectory of t and the proof of the lemma is complete.

Lemma 17.2. A geodesic h in no finite class m has a periodic limit geodesic.

By definition h is in no finite class m if the corresponding symbolic trajectory t is in no finite class m. By virtue of the preceding proof, t has a periodic limit trajectory represented by a left boundary path π' whose right boundary arc is a periodic geodesic k. The ribbon R containing h contains convex sub-paths congruent to arbitrarily long sub-paths of π'. It follows that there are elements on h arbitrarily near elements of k and the proof of the lemma is complete.

Theorem 17.3. A necessary and sufficient condition that a geodesic h^* on M^* be minimal is that it determine a minimal symbolic trajectory t.

Recall that a geodesic is termed minimal if it is periodic or if it is a member of a set of geodesics M every geodesic of which has M as a derived set. Similarly a symbolic trajectory is termed minimal if it is periodic or if it is a member of a set of symbolic trajectories M' every trajectory of which has M' as a derived set. Cf. Theorem 4.3.

Case 1. If h^* is periodic, t is periodic, and conversely, by virtue of Theorem 17.2.

Case 2. The non-periodic case. Suppose then that h^* is minimal but not periodic. Let t be the corresponding symbolic trajectory. By virtue of the preceding lemma, t is in some finite class m. Otherwise h^* would have a periodic limit geodesic. But the relation between symbolic elements e of class m and their image elements F(e) is locally 1-1 and bicontinuous. Inasmuch as their correspondence preserves limit relations and minimal symbolic trajectories and minimal geodesics are similarly defined in terms of limit relations, we conclude that t is minimal.

Similarly, we suppose t is minimal but not periodic. It is then in a finite class m by virtue of Lemma 17.1, and as previously we infer that the corresponding geodesic is minimal.

Theorem 17.4. A necessary and sufficient condition that a geodesic h^* be recurrent is that its symbolic trajectory t be recurrent.

Case 1. If h^* is periodic, t is periodic, and conversely.

Case 2. Here we are concerned with a geodesic h^* and a trajectory t of a finite class m as in the earlier proof. The theorem follows from the definitions of recurrence and the locally one-to-one bicontinuous character of the relation $E^* = F(e)$.

Theorem 3 and 4 have the

Corollary. A necessary and sufficient condition that a geodesic h^* be recurrent is that it be minimal.

Theorem 17.5. The limit trajectories of each geodesic h^* include at least one recurrent trajectory.

Case I. If h^* is in no finite class m it follows from Lemma 17.2 that h^* has a periodic limit geodesic, and the proof of the theorem is complete.

Case II. Suppose that h^* is in a finite class m. Let t be the corresponding symbolic trajectory. By virtue of Theorem 4.1 there exists a sequence

$$e_1, e_2, \cdots$$

of symbolic elements based on t converging to a limit element e defining a recurrent symbolic trajectory t'. Since the class m is compact, e is in the class m, t' is admissible and defines a geodesic λ. The geodesic λ is recurrent since t' is recurrent, in accordance with Theorem 17.4.

The proof of the theorem is complete.

Theorem 17.6. There exist non-periodic recurrent geodesics.

We need only show that there exist non-periodic recurrent symbolic trajectories which define right paths. A symbolic trajectory t will certainly be admissible if for no i, $t_i\, t_{i+1} = 1$, and if t contains no $(2p-1)$-blocks in the orders C and C^{-1}. In particular a trajectory made up exclusively of the generators a_1, b_1 will satisfy these conditions. Other pairs such as $a_1 a_1$, $b_1 a_2^{-1}$, etc. can obviously be used. We can then use the recurrent trajectory defined on page 24, and substitute a_1 for 1, and b_1 for 2, obtaining thereby an admissible non-periodic recurrent trajectory.

Theorem 17.7. A necessary and sufficient condition that a geodesic h^* be transitive is that the corresponding symbolic trajectory t be transitive.

If t is transitive, h^* is transitive. For corresponding to an arbitrary symbolic element e there is a sequence of elements e_n based on t which have e as a limit element. The geodesic elements $F(e_n)$ on h^* will then have $F(e)$ as a limit element since $F(e)$ is continuous without exception. Thus h^* is transitive.

Conversely, suppose that h^* is transitive. Since h^* is transitive the set of elements on h^* of the form $F(e)$ is everywhere dense in the set of elements $F(e)$. But the relation $E^* = F(e)$ is locally 1-1 and bicontinuous for elements of class m. Moreover, the elements of finite class are everywhere dense among the elements e of infinite class. For a symbolic element $e(t,n)$ of infinite class can be approximated arbitrarily closely by a symbolic element of finite class by modifying the remote symbols of t. Hence symbolic elements e based on t are everywhere dense in the set of all symbolic elements.

The proof of the theorem is complete.

In a sequence aeb of three generating symbols, e will be said to be unrelated to a and b if

$$ae \neq I, \qquad eb \neq I,$$

and neither ae nor eb are in the orders C or C^{-1}.

Lemma 17.3. If a and b are arbitrary elements there exists an element e unrelated to a and b.

The element e is to be chosen from the elements

$$a_1,\ldots,a_p,\ b_1,\ldots,b_p \qquad\qquad p > 1$$

and their inverses. That is, from at least 8 elements. To be unrelated to a and b, the generator e cannot equal a^{-1}, nor b^{-1}, nor the predecessors or successors of a or b in C. There are then at least two choices of e.

With the aid of this lemma it is easy to prove the existence of admissible transitive symbolic trajectories. Hence we have the following theorem.

Theorem 17.8. There exist transitive geodesics.