APPENDIX : THE DEHN-NIELSEN THEOREM

JOHN STILLWELL

1. Introduction

One important theorem attributed to Dehn does not appear in his published work, and his proof of it seems to be lost. This is the theorem that every automorphism of the fundamental group $\pi_1(S)$ of a closed orientable surface S can be induced by a homeomorphism of S. The earliest known proof appears in Nielsen [1927], and consists of two parts. The first, which is in §9 of Nielsen's paper and expounded below as Theorem 5, is quite elegant and is attributed to Dehn. The second, in §23, is agonisingly long and is apparently Nielsen's own, though he still credited the theorem to Dehn. As was mentioned in the introduction to Paper 7, Nielsen in 1931 found a replacement for §23 which was much simpler, but he did not publish it. We give a similar proof below as Theorem 6.

It is a pity that Nielsen did not publish his 1931 proof, as his [1927] proof was made obsolete by the much shorter one of Seifert [1937], with the result that Dehn's contribution to the theorem was buried along with Nielsen's. Moreover, since Seifert's proof was purely topological, it became forgotten that the roots of the Dehn-Nielsen theorem were in hyperbolic geometry.

With the recent revival of interest in hyperbolic geometry, it seems timely to reconstruct the Dehn-Nielsen proof and its geometric background. This not only fills a gap in the record of Dehn's work, it also brings to light some little known but influential work of Poincaré. It is fairly well known that Poincaré discovered the connection between hyperbolic geometry and surface topology in his work on fuchsian groups in the early 1880's (for an English translation of these papers, see Poincaré [1985]). In a less well known (though often cited) paper, Poincaré [190 took up the idea again for purely topological motives. He obtained results about simple curves on surfaces by a method which turned out to be crucial to the Dehn-Nielsen proof - the use of geodesics as canonical representatives for homotopy classes of curves. This method is explained in sections 3 and 4 below.

It was Poincaré [1904] which inspired Dehn to use hyperbolic geometry in his solution of the word and conjugacy problems for surface groups, Dehn [1912a] (Paper 4 in this volume). In the 1920's, Dehn was still a little under the influence of Poincaré, judging by the remarks at the beginning of Dehn [1921], but he was gravitating towards purely topological methods. The torch of hyperbolic geometry in topology was picked up by Nielsen in his [1927], [1929], [1932], a monumental series of works which have only recently been assimilated, in the reworking and extension of Nielsen's theory by Thurston. In surveying the development of geometric methods in topology, the Dehn-Nielsen proof seems an excellent vantage point from which to look back to Poincaré and forward to Thurston.

It should be mentioned that the hyperbolic metric used in the Dehn-Nielsen proof can be replaced by a metric on elements of the fundamental group, the purely combinatorial "word metric" introduced by Dehn [1912a] (see the Introduction to Paper 4). A generalisation of Dehn's Theorem 5 is proved using the word metric by Floyd [1980], and it seems clear that the concepts in the second part of the Dehn-Nielsen proof (Theorem 6) could similarly be replaced by combinatorial ones. What this probably means, however, is not that hyperbolic geometry is irrelevant, but that the most relevant features of hyperbolic geometry are the combinatorial properties of its regular tesselations.

2. The special case of the torus

Among the closed orientable surfaces of genus $\geqslant 1$, the torus T (genus 1) is distinguished by having a natural euclidean structure. This is one reason for a separate discussion of T. The other is that T enables us to review standard facts about canonical curves, universal covering and Cayley diagram in a context which is free of geometric difficulties. When we move on, to a surface S of genus > 1, it will then be possible to concentrate on the geometric difficulties which arise from the hyperbolic structure on S.

The torus T contains a pair of canonical curves a,b with a characteristic property: they are simple curves which meet at exactly one point, 0, where they cross. Cutting T along a,b yields the canonical polygon T_0 for T, which is topologically a square whose edges originate from a and b as indicated in Fig. 1.

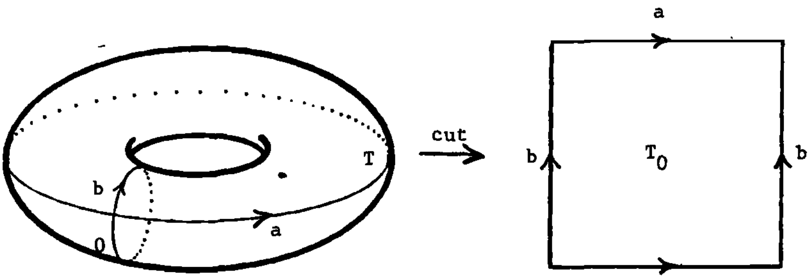


Fig.1

This follows from the classification theorem for surfaces (Dehn and Heegaard [1907]), and does not depend on the special position of a,b, as Fig.1 might suggest. Namely, the cut along b leaves a connected orientable surface with two boundary curves (the two "sides" of the cut, which are connected by the curve a). The claesification theorem saye that any such surface is topologically a cylinder, and cutting a cylinder along a simple arc a joining its boundaries yields the square shown.

Thus if a', b' are any other curves with the characteristic property of canonical curves, then cutting T along a', b' yields a polygon T_0' homeomorphic to T_0, and with correspondingly labelled sides. For example, Fig. 2.

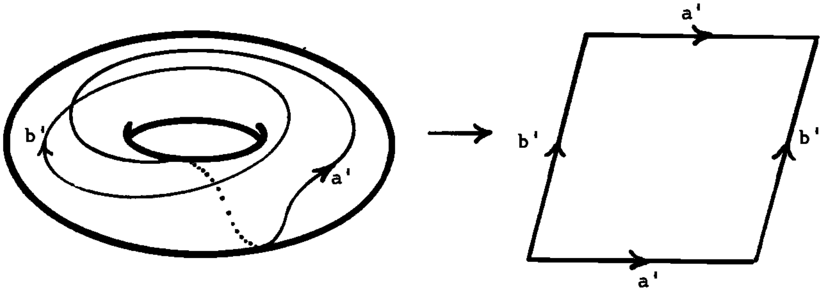


Fig. 2

It follows that *there is a homeomorphism* $\tau: T \to T$ *which maps* a *onto* a' *and* b *onto* b'. One only needs to construct a homeomorphism of T_0 onto T_0' which maps equivalent points in the boundary of T_0 to equivalent points in the boundary of T_0'. A more intuitive way to construct τ is to paste the equivalent sides of P' together again to form a torus on which a', b' "look like" a, b (Fig. 3).

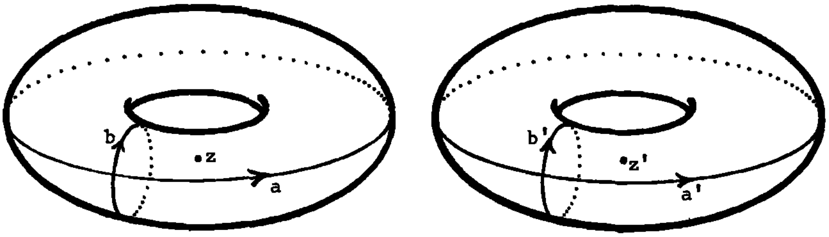


Fig 3.

Then the map $\tau: T \to T$ which sends a, b onto a', b' respectively is defined by sending each point z to the similarly positioned z'.

To summarise, we have

<u>Theorem 1</u>: If a, b are simple curves on T which meet at a single point, where they cross, and if a', b' is another pair with the same property, then there is a homeomorphism $\tau: T \to T$ which maps a onto a' and b onto b'. □

It follows that if $\tau_*: \pi_1(T) \to \pi_1(T)$ is the automorphism of the fundamental group induced by τ, then τ_* sends the homotopy class [a] of a to the homotopy class [a'] of a', and the homotopy class [b] of b to the homotopy class [b'] of b'. We are therefore able to induce any automorphism $I: [a] \mapsto [a'], [b] \mapsto [b']$ of $\pi_1(T)$ for which [a'], [b'] are representable by a pair of canonical curves. To prove the Dehn-Nielsen theorem for the torus, then, it suffices that any automorphism preserves the characteristic property of canonical curves.

<u>Theorem 2.</u> Any automorphism $I: \pi_1(T) \to \pi_1(T)$ can be induced by a homeomorphism $\tau: T \to T$.

<u>Proof</u>: The fundamental group $\pi_1(T)$ is the free abelian group on two generators (corresponding to a fixed pair of canonical curves a,b), so we can identify it with the lattice of integer points (m,n) in the plane

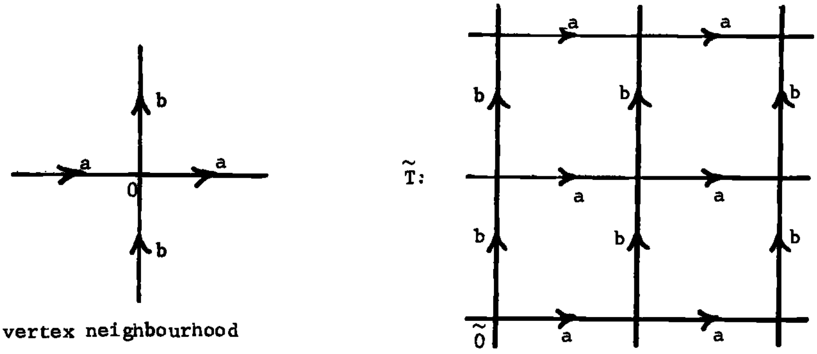
under vector addition. An automorphism I: $\pi_1(T) \to \pi_1(T)$ is then a one-to-one linear map of the lattice onto itself, hence given by an integer matrix

$$I = \begin{pmatrix} p_1 & p_2 \\ q_1 & q_2 \end{pmatrix}$$

with determinant ± 1. I sends the generators (1,0) to (p_1,p_2) and (0,1) to (q_1,q_2), where p_1, p_2 are coprime, and so are q_1,q_2, since det I = $\pm$ 1. It can then be checked by brute force that the corresponding homotopy classes $[a^{p_1}b^{p_2}]$ and $[a^{q_1}b^{q_2}]$ can be represented by simple curves which meet, and cross, at a single point, whence the theorem follows from Theorem 1. □

A much clearer view of the above proof can be obtained by looking at the universal cover of T. I shall now sketch this approach, since it points the way to go in the case of higher genus, where the automorphisms of the fundamental group are no longer so accessible.

The universal cover $\tilde{T}$ of T is a plane obtained by pasting together copies of the canonical polygon T_0 for T so that the neighbourhood of each vertex looks like the neighbourhood of the single vertex 0 on T (Fig. 4).



vertex neighbourhood

Fig.4

A point $\tilde{P}$ on $\tilde{T}$, relative to $\tilde{0}$, represents the homotopy class $[c]$ of an arc c on T with fixed endpoints 0 and P. $\tilde{P}$ is found by lifting c to an arc $\tilde{c}$ on $\tilde{T}$ with initial point $\tilde{0}$, and its association with the homotopy of class is due to the fact that a homotopy of c with endpoints fixed lifts to a homotopy of $\tilde{c}$ with endpoints fixed. Fig. 5 shows an example in which $\tilde{c}$ is deformed into the straight line segment between $\tilde{0}$ and $\tilde{P}$.

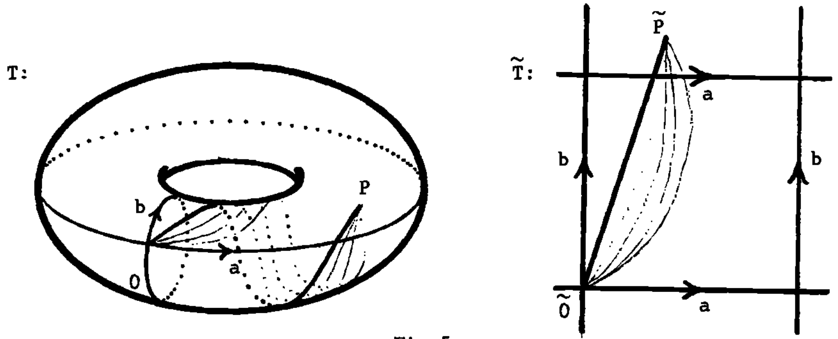


Fig.5

If we pull back the euclidean metric from the plane $\tilde{T}$ to the torus T, then line segments on $\tilde{T}$ cover geodesics on T, and we can represent each non-trivial homotopy class by a unique geodesic. If we imagine that T is a self-contained world in which light rays travel along geodesics, then $\tilde{T}$ is the <u>actual view</u> of T seen by its inhabitants.

In particular, the vertices, which we take to be situated at the integer lattice points, are the different images of the origin eeen along closed geodesics. Thus the lattice points represent the homotopy classes of closed curves based at 0, the elements of $\pi_1(T)$, with (m,n) corresponding to the element $[a^m b^n]$. This recovers the representation of $\pi_1(T)$ as the integer lattice under vector addition, and shows that the net of labelled edges on $\tilde{T}$ is the Cayley diagram of $\pi_1(T)$.

An automorphism $I: \pi_1(T) \to \pi_1(T)$ can then be viewed as a one-to-one linear map $\tilde{f}$ of the lattice onto itself, and we can extend $\tilde{f}$ to a one-to-one linear map of the whole plane $\tilde{T}$. This extended map sends points (x,y) and $(x,y)+(m,n)$ over the same point P of T to points $\tilde{f}(x,y)$ and $\tilde{f}(x,y) + \tilde{f}(m,n)$ (by linearity of $\tilde{f}$) which again lie over the eame point, $f(P)$ say, since $\tilde{f}(m,n)$ is a lattice point by hypothesis. Thus $\tilde{f}$ covers a homeomorphism $f: T \to T$, and f induces I because

$$[f(a^m b^n)] = \tilde{f}(m,n) = I[a^m b^n]$$

by definition of $\tilde{f}$.

Fig. 6 shows the $\tilde{f}$ and f associated with the automorphism I: $[a] \mapsto [ab]$, $[b] \mapsto [ab^2]$.

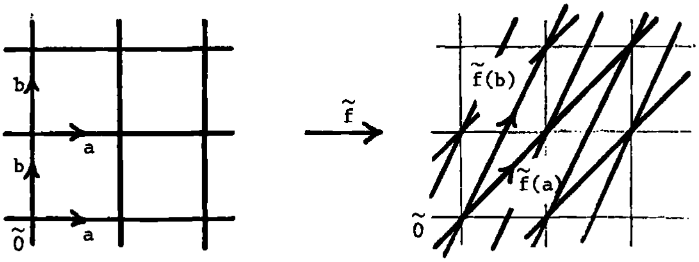
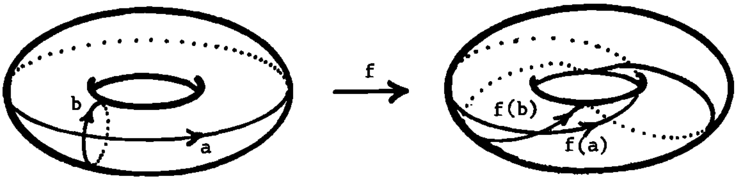


Fig.6



Although the above proof does not require an explicit determination of automorphisms of $\pi_1(T)$, it depends heavily on the ability of the euclidean plane to admit linear maps which are not rigid motions. This ability can be expressed algebraically by saying that any automorphism of the group of lattice translations extends to an automorphism of the group of all translations. In the hyperbolic plane there are no linear maps except rigid motions*, and hence there is no obvious way to extend the automorphism of the "lattice" which represents the fundamental group of a surface of genus > 1 to a map $\tilde{f}$ of the whole plane. In the absence of a map $\tilde{f}$ we have to make a rather exacting analysis of simple curves and intersections to show that an automorphism sends one canonical geodesic curve system to another.

*This can be seen from the projective model. In this model, the hyperbolic plane is the interior of the unit disc, "lines" are portions of ordinary straight lines within the unit disc, and rigid motions are all collineations of the ordinary plane which map the unit disc onto itself. Hence there are no non-rigid maps of the hyperbolic plane which map "lines" to "lines".

3. Facts from group theory and hyperbolic geometry

Hyperbolic geometry enters the theory of surfaces when one attempts to construct the universal cover $\tilde{S}$ of a closed surface S of genus $\geqslant 2$ in a metrically regular way. For example, cutting the surfaoe of genus 2 along its canonical curves yields an octagon S_0 as canonical polygon (Fig. 7).

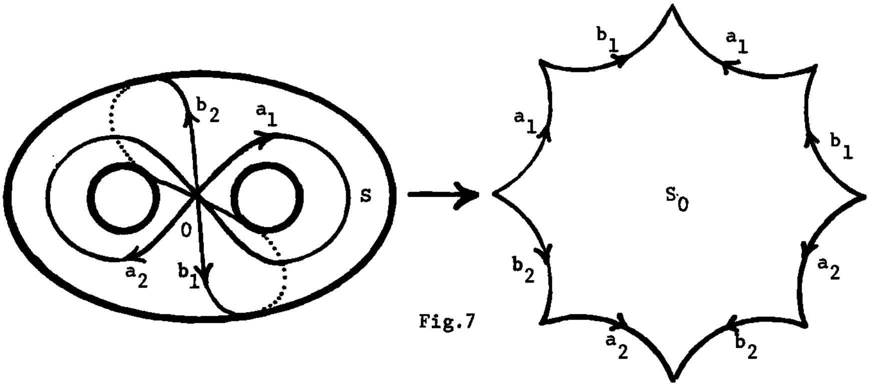


Fig.7

and $\tilde{S}$ is constructed by pasting copies of S_0 together so that each vertex has a neighbourhood like the neighbourhood of 0 on S, namely, Fig. 8

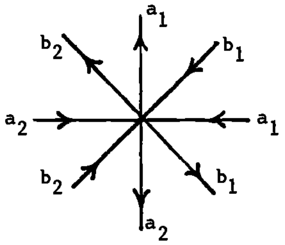


Fig. 8

Thus eight octagons meet at each vertex on $\tilde{S}$. This is not possible for regular euclidean octagons, but it is possible for regular octagons in the hyperbolic plane, where the angle sum of a polygon can take any value between 0 and the euclidean value. Thus one has regular octagons with corner angle $\frac{\pi}{4}$, and one can fit them together to form $\tilde{S}$ as a tessellation of the hyperbolic plane, the labelled edges of which form the Cayley diagram of $\pi_1(S)$. The construction is similar for any genus $p \geqslant 2$ - one has $4p$ $4p$-gons meeting at each vertex. Since $\pi_1(S)$ is the automorphism group of its own Cayley diagram, this gives a faithful representation of $\pi_1(S)$ as a group of hyperbolic motions. The motion associated with the element $g \in \pi_1(S)$ shifts the vertex $\underline{1}$ representing the identity of $\pi_1(S)$ to the vertex representing g, and hence the vertex representing h, for any $h \in \pi_1(S)$, to the vertex representing gh.

The topological properties of Cayley diagrams are said by Dehn 1922 to underly the proof of Dehn's theorem. To reconstruct such a proof I generalise the Cayley diagram construction to any curve system $\{\alpha_i\}$ which cuts S down to a simply connected piece. The curves α_i then lift to a net which partitions $\tilde{S}$ into cells congruent to P. Since P is simply connected, each element of $\pi_1(S)$ is represented by a closed path through the union of the α_i on S, hence by a vertex of the net on $\tilde{S}$ and by the corresponding motion of the net onto itself. To avoid making continual pedantic distinctions between elements $g \in \pi_1(S)$ and the vertices or motions which represent them, I shall speak of "the element g", "the vertex g" or "the motion g" as the occasion requires.

Of course, one can define $\tilde{S}$ and its motions purely combinatorially, but the lesson to be learned from Poincare, Dehn and Nielsen is that hyperbolic metric concepts have great heuristic value, even if they can theoretically be eliminated. We shall therefore assume the basic facts of hyperbolic geometry, including the Poincare disc model, and develop only the less familiar facts which are pertinent to surface theory. The best sources for the facts we assume seem to be works on complex analysis, such as Siegel [1971], and Zieschang [1981].

Fact 1 (see Siegel [1971]). The hyperbolic plane can be modelled by the open unit disc $D \subseteq \mathbb{R}^2$ when "motions" are interpreted as conformal (equivalently, circle-preserving) maps of $\mathbb{R}^2$ which map the unit circle, ∂D, onto itself. Hyperbolic straight lines are then circular arcs in D orthogonal to ∂D, hyperbolic angle equals euclidean angle, and the motions are of three types corresponding to the following families of euclidean circles in D (Fig. 9. (i), (ii), (iii))

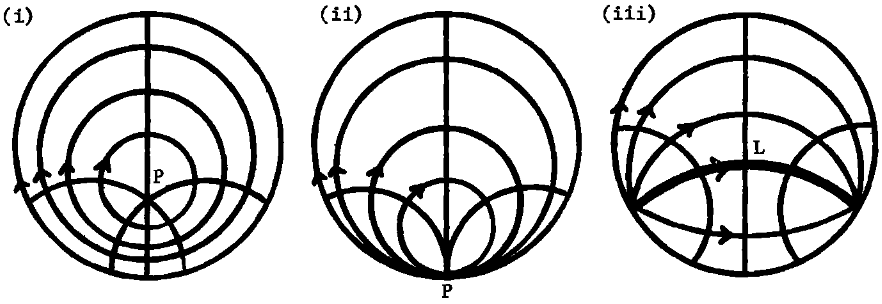


Fig.9

(i) Circles orthogonal to the pencil of hyperbolic lines through a point $P \in D$ are the hyperbolic circles with hyperbolic centre P. A motion which leaves P fixed maps each circle with hyperbolic centre P onto itself, permutes the hyperbolic lines through P, and is called a rotation about P.

(ii) Circles orthogonal to the pencil of hyperbolic parallels with limit point $P \in \partial D$ are called hyperbolic limit circles with centre P. A motion which leaves P fixed maps each limit circle onto itself, permutes the parallels with limit P, and is called a limit rotation about P. Any two hyperbolic parallels approach each other arbitrarily closely in the hyperbolic metric, and under a limit rotation there are points which are moved through arbitrarily small hyperbolic distancas.

(iii) Circles orthogonal to the hyperbolic perpendiculars to a given line $L \subseteq D$ are the *curves at constant hyperbolic distance from* L (distance curves of L). A motion which fixes the end points of L on ∂D maps each distance curve of L onto itself, permutes the perpendiculars to L, and is called a *translation*, or *displacement*, with *axis* L. Each point of L is sent a constant hyperbolic distance, λ, called the *displacement length* of the motion, and the segment between any point of L and ita image is called a *displacement segment*. Each point on a distance curve C of L is also sent a constant hyperbolic distance, but the constant is $> \lambda$ and increases with the distance of C from L. □

Fact 2. Each non-identity motion $g \in \pi_1(S)$ is a translation.

Proof. Since the motion g maps the Cayley diagram of $\pi_1(S)$ onto itself non-trivially, it must move every point through at least the diameter of a cell in the tessellation. Thus there is a non-zero lower bound to the distance moved by each point under the motion g, whence g is a translation by Fact 1. □

As Dehn [1910] remarks, the non-existence of points on $\tilde{S}$ which move through arbitrarily small distances reflects the non-existence of small curves on S which do not contract to a point. This is why translations are important for topology, and we shall analyse them in detail. Following Nielsen [1944], we call the fixed points of a translation g its *fundamental points*; g moves points of D away from the <u>negative</u> fundamental point, called $U(g)$, towards the <u>positive</u> fundamental point, called $V(g)$, along the "streamlines" which are the axis of g, called $A(g)$, and its distance curves.

The next fact, also observed by Dehn [1910], gives a way to find $U(g)$, $V(g)$, and hence $A(g)$, from the Cayley diagram.

<u>Fact 3.</u> The vertices $\ldots, g^{-1}, 1, g, g^{2}, \ldots$ of the Cayley diagram lie on a distance curve through $U(g)$, $V(g)$. Moreover $g^{n} \to V(g)$ and $g^{-n} \to U(g)$ (in the euclidean metric), as $n \to \infty$.

<u>Proof.</u> The motion g, by definition, maps each member of the sequence of vertices $\ldots, g^{-1}, 1, g, g^{2}, \ldots$ onto its successor. Hence the vertices must lie along a streamline of the motion, i.e. along a distance curve of $A(g)$, by Fact 1. Since they are equally spaced along the distance curve, by Fact 1, the hyperbolic distances of g^{n} and g^{-n} from 1 tend to ∞ as $n \to \infty$, hence $g^{-n} \to U(g)$ and $g^{n} \to V(g)$ in the euclidean metric. □

<u>Fact 4.</u> If f, g are any two translations, then fgf^{-1} is a translation of the same length as g, and

$$A(fgf^{-1}) = A(g) \cdot f^{-1} \quad \text{(the } f^{-1}\text{-image of } A(g))$$

<u>Proof.</u> Consider the f^{-1}-images, $U(g) \cdot f^{-1}$ and $V(g) \cdot f^{-1}$, of $U(g)$ and $V(g)$, and their images in turn under fgf^{-1}.

$$U(g) \cdot f^{-1} \cdot fgf^{-1} = U(g) \cdot gf^{-1} = U(g) \cdot f^{-1}$$

since $U(g)$ is fixed by g. Thus $U(g) \cdot f^{-1}$ is a fixed point of fgf^{-1}, and similarly so is $V(g) \cdot f^{-1}$. Hence these are the fundamental points of fgf^{-1} and $A(g)\cdot f^{-1} = A(fgf^{-1})$. Moreover, a point $P \in A(g)$ and its g-image $P\cdot g \in A(g)$ are sent by f^{-1} to $P\cdot f^{-1} \in A(fgf^{-1})$ and $P\cdot gf^{-1}$ = image of Pf^{-1} under fgf^{-1}, hence the translations on the two axes are of equal length. □

The remaining facts are proved following Nielsen [1927].

Fact 5. In a group of translations, two non-commuting elements have no common fundamental point.

Proof. Suppose on the contrary that f, g are non-commuting elements with a common fundamental point. By replacing one of f, g by its inverse, if necessary, we can assume that $V(f) = V(g)$. Now $V(fgf^{-1}) = V(g)$ by Fact 4, thus fgf^{-1} and g^{-1} have parallel axes (with limit $V(fgf^{-1}) = V(g) = U(g^{-1})$) and displacement lengths which are equal (by Fact 4) but opposite in direction. It then follows from the continuity of motions that by choosing a point P on $A(fgf^{-1})$ sufficiently close to $A(g^{-1})$, the point $P\cdot fgf^{-1}\cdot g^{-1}$ can be made arbitrarily close to P. Thus the non-identity group element $fgf^{-1}f^{-1}$ moves points through arbitrarily small distances, contrary to the properties of translations (Fact 1). □

Fact 6. In a group of translations, two elements commute if and only if they have the same axis.

Proof. If f, g have the same axis then
fg = [translation of length $\lambda_f + \lambda_g$ with axis $A(f)$] = gf.
Conversely, if $fg = gf$ then $fgf^{-1} = g$, hence $A(fgf^{-1}) = A(g)$. But $A(fgf^{-1}) = A(g)\cdot f^{-1}$ by Fact 4, hence f^{-1} must map the

hyperbolic line A(g) onto itself. This line is therefore the axia of f^{-1}, and hence of f, by Fact 1. □

For the la■t fact we u■e Nielaen'■ notation I for an automorphi■m and g_I for the I-image of g.

Fact 7. An automorphi■m of a group G of tran■lation■ induce■ a permutation of the fundamental point■ of G.

Proof. An automorphi■m ■end■ $ghg^{-1}h^{-1}$ to 1 if and only if $ghg^{-1}h^{-1} = 1$, hence $f_Ig_I = g_If_I$ if and only if $fg = gf$. Since I permute■ the element■ of G, it then follows from Fact 6 that I induce■ a permutation of the axe■ of G. But no two di■tinct axes have a common endpoint by Fact 5, hence the permutation of axes i■ in fact a permutation of fundamental point■. □

4. Axes, simple curves and intersections

If we attempt to generalise the analysis of the torus given in section 2 to a surface S of higher genus, then several generalisations of the canonical geodesic curve system suggest themselves. For reasons which will appear later, it is good to take a system conaisting of simple closed geodesics with no multiple intersections, and such that cuts along these geodesics render S simply connected. In fact, on the surface S of genus p there are geodesics $\alpha_1, \alpha_2, \dots, \alpha_{2p}$ with the following properties

(i) Each α_i meets α_{i+1} at exactly one point, where they cross.

(ii) There are no other multiple points; in particular, each α_i is simple.

We shall see in Theorem 3 below that cutting along such curves α_i renders S simply connected, so this system of curves meets all our requiramonts.

Such a system was used by Dehn's student Baer, [1927], and a similar one appears in Nielsen [1929, §4], though neither notes that the α_i can be realised by geodesics. Certainly it is clear that there are simple <u>curves</u> α_i with properties (i), (ii), e.g. as in Fig. 10, but to show that

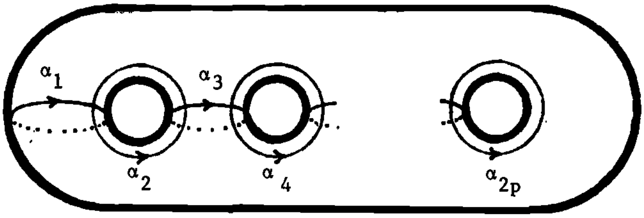


Fig. 10

topologically realisable incidences of curves are also realised by gecdesics one needs the following ideas of Poincare [1904].

The metric on S is inherited from the hyperbolic metric on $\tilde{S}$, hence the geodesics on S are the curves covered by hyperbolic straight lines on $\tilde{S}$, and intersections of geodesics lift to intersections of hyperbolic linea. In turn, hyperbolic lines L_1, L_2 intersect if and only if the end points of L_1 on $\partial\tilde{S}$ separate the endpoints of L_2. Poincaré makes this trite remark topologically potent by combining it with the observation that a deformation of S lifts to a deformation of $\tilde{S}$ through a bounded hyperbolic distance, which therefore, deforms any line into a curve with the same endpoints on ∂S.

Lemmas 1-5 below ayatematise Poincaré's rather fragmentary applications of these remarks, and they suffice in particular to show the existence of geodesics α_i with propertiea (i), (ii). In stating the proofs it is convenient to extend the notation of section 3 to allow closed curves on S in place of their homotopy classes, the elements of $\pi_1(S)$. Thus when a closed curve c on S is lifted to $\tilde{c}$ on $\tilde{S}$ with initial point $\underline{1}$ we shall call the final point c. It is a vertex of the Cayley diagram, and the translation which carries $\underline{1}$ to c will be called the motion c, its axis $A(c)$, and the closed curve on S covered by a displacement segment $\delta(c)$ of $A(c)$ will be called $\alpha(c)$. (Of course, when we are discussing elementa $g \in \pi_1(S)$ directly, the old notation will still be used, and $\alpha(g)$ will denote the curve on S covered by a displacement segment of $A(g)$.)

In what follows, "equivalent" sets on $\tilde{S}$ are sets with the same projection on S, i.e. translates of each other by motions in $\pi_1(S)$.

Lemma 1. The closed geodesics on S are precisely the curves $\alpha(g)$, $g \in \pi_1(S)$.

Proof. A closed curve c on S lifts to an arc $\tilde{c}$ on $\tilde{S}$ with endpoints equivalent under the motion c. If the curve c is a geodesic then $\tilde{c}$ is a hyperbolic line segment which is collinear with its own translate under the motion c. That is, $\tilde{c}$ is a displacement segment of $A(c)$.

Conversely, a displacement aegment $\delta(g)$ of $A(g)$, $g \in \pi_1(S)$, has a closed projection $\alpha(g)$ on S, since the endpoints of $\delta(g)$ are equivalent (under g). And every segment of $\alpha(g)$ is geodesic because successive translates of $\delta(g)$ by g are collinear and equivalent (It follows that any displacement segment of $A(g)$ has the same projection $\alpha(g)$, if we do not distinguish an initial point.) □

In general, the curve on S covered by a hyperbolic line segment with equivalent endpoints is a geodesic monogon with a corner where the two endpoints project, rather than a closed geodesic. For example, the curve covered by an edge labelled a_1 in the Cayley diagram of $\pi_1(S)$ has a right angled corner at the basepoint O on S, since adjacent edgea labelled a_1 meet at right angles; thus segments of the curve which contain O are not geodesic. This monogon does however extend to a closed "figure eight" geodesic covered by adjacent edges a_1, a_2^{-1}, since such edges are collinear. In particular, the segment labelled $a_1a_2^{-1}$ issuing from $\underline{1}$ on $\tilde{S}$ is a displacement segment of $A(a_1a_2^{-1})$. (Cf. Figs. 7 and 8).

Since successive a_1 edges in the Cayley diagram are not collinear, the distance curve of $A(a_1)$ through the vertices $\underline{1}$, a_1, a_1^2,.. is not a hyperbolic line, and hence the curve a_1 on S cannot be deformed into a geodesic with 0 fixed. This leads us to consider <u>free homotopies</u> on S, i.e. deformations without fixed basepoint, and their lifts to $\tilde{S}$ which are not fixed at the vertices of the Cayley diagram.

<u>Lemma 2</u> Any closed, non-contractible curve c on S is freely homotopic to a unique closed geodesic, namely $\alpha(c)$.

<u>Proof</u>. Consider the lift $\tilde{c}$ of c which runs betwen the vertices $\underline{1}$ and c in the Cayley diagram. Since the curve c ia non-contractible on S, these vertices are distinct and, since they lie on a distance curve of $A(c)$, we can drop perpendiculars from them to the endpoints of a displacement segment $\delta(c)$ of $A(c)$. If we deform $\tilde{c}$ into the distance curve segment between $\underline{1}$ and c, and then descend through the intervening distance curve segments onto $\delta(c)$ (Fig.11), then the endpoints remain equivalent throughout, and hence the whole deformation projects to a free homotopy on S which converts c to $\alpha(c)$.

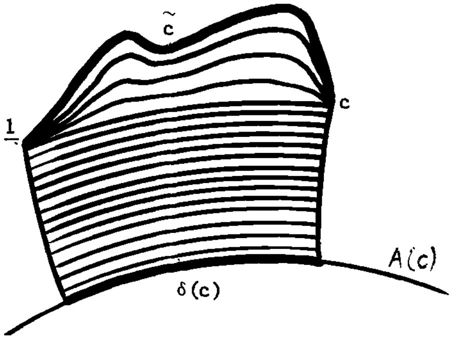


Fig. 11

The uniqueness of $\alpha(c)$ can be sean as follows. By Lemma 1, any deformation of c to a closed geodesic lifts to a deformation $\tilde{c}_t$, $0 \leq t \leq 1$, from $\tilde{c}_0 = \tilde{c}$ to $\tilde{c}_1 =$ some $\delta(g)$, in which the endpoints of $\tilde{c}_t$ are equivalent for each t. Initially, when $\tilde{c}_t = \tilde{c}$, the endpoints are equivalent under the motion c, hence they must remain so (since $\pi_1(S)$ ia a discontinuous group, e.g. by Fact 2). But the only axis which containa points equivalent under the motion c is $A(c)$, hence $\tilde{c}_1 = \delta(c)$, which projects to $\alpha(c)$ as required. □

The next three lemmas exploit the separation properties of points on $\partial\tilde{S}$ and the fixture of these points under deformations on S.

<u>Lemma 3</u>. A curve c on S is freely homotopic to a simple curve $\Longleftrightarrow$ $\alpha(c)$ is simple.

<u>Proof</u>. ($\Leftarrow$) is immediate from Lemma 2. To prove ($\Rightarrow$), suppose that $\alpha(c)$ is not simple. A self-intersection of $\alpha(c)$ is covered by an intersection of $A(c)$ with one of its translates $A(c)\cdot g$, $g \in \pi_1(S)$.

Case 1 : $A(c)$ crosses $A(c).g$. Then the endpoints of $A(c)$ separate those of $A(c)\cdot g$ on $\partial\tilde{S}$. We know from Lemma 2 that there is a deformation on S of $\alpha(c)$ to c, and this deformation lifts to a deformation on $\tilde{S}$ of $A(c)$ to a curve $\tilde{c}_\infty$ with the same endpoints on $\partial\tilde{S}$. The deformation similarly lifts to a deformation of $A(c)\cdot g$ to $\tilde{c}_\infty\cdot g$, with fixed endpoints. The endpoints therefore continue to separate each other on $\partial\tilde{S}$, and hence $\tilde{c}_\infty$ crosses $\tilde{c}_\infty\cdot g$ (e.g., by the Jordan curve theorem). This crossing covers a self intersection of c on S.

Case 2. $A(c) = A(c)\cdot g$, where g necessarily has axis $A(c)$ although it is not a power of c, since a double point on c lifts to points on $\tilde{S}$ which are inequivalent under the motion c. The corresponding deformations then yield $\tilde{c}_\infty$ and $\tilde{c}_\infty \cdot g$ with the same endpoints as $A(c)$. Since $\tilde{c}_\infty$ is periodic under the translation c along $A(c)$, it has points at maximum distance to left and right of $A(c)$, and since $\tilde{c}_\infty . g$ is congruent to $\tilde{c}_\infty$ by the translation g along $A(c)$, $\tilde{c}_\infty . g$ attains the same maximum distances to left and right of $A(c)$. It follows that $\tilde{c}_\infty . g$ cannot lie entirely to one side of $\tilde{c}_\infty$, and hence they meet. Since g is not a power of c, an intersection of $\tilde{c}_\infty$ and $\tilde{c}_\infty . g$ covers a double point of c. □

<u>Lemma 4</u>. If c, d are disjoint simple curves on S, then $\alpha(c)$, $\alpha(d)$ are either disjoint or identical.

<u>Proof</u>. Axes, and hence geodesics, cannot be tangential unless they coincide, so if $\alpha(c)$, $\alpha(d)$ are neither disjoint nor identical they must cross, and hence $A(c)$ crosses $A(d)$ on $\tilde{S}$. Thus the endpoints of $A(c)$ on $\partial\tilde{S}$ separate those of $A(d)$. Deformations of $\alpha(c)$, $\alpha(d)$ to c, d respectively lift to deformations of $A(c)$, $A(d)$ to $\tilde{c}_\infty$, $\tilde{d}_\infty$, where $\tilde{c}_\infty$ has the same endpoints as $A(c)$ and $\tilde{d}$ has the same endpoints as $A(d)$. Then $\tilde{c}_\infty$ crosses $\tilde{d}_\infty$, and this crossing covers a point on S where c meets d, which is a contradiction. □

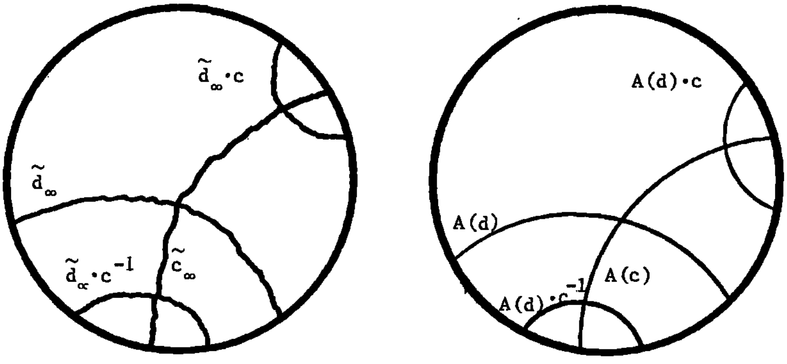
The reader may easily guess that $\alpha(c) = \alpha(d) \Longleftrightarrow c, d$ are freely homotopic, though we shall not need this result. It is worth pointing out, however, that the criterion $\alpha(c) = \alpha(d)$ then leads to an algorithm for deciding when two given curves are freely homotopic (Dehn [1910]), just as Lemma 3 leads to an algorithm for deciding whether c is homotopic to a simple curve (Poincaré [1904]).

Lemma 5. If c,d are simple curves which meet at a single point, where they cross, than $\alpha(c)$, $\alpha(d)$ have the same property.

Proof. We know from Lemma 3 that $\alpha(c)$, $\alpha(d)$ are simple.

To establish the crossing property, let $\tilde{c}_\infty$, $\tilde{d}_\infty$ denote the curves on $\tilde{S}$ which cover c,d and have the same endpoints es A(c), A(d). The single crossing X of c by d lifte to a single croseing $\tilde{X}$ of $\tilde{c}_\infty$ by $\tilde{d}_\infty$, together with a single crossing $\tilde{X}\cdot c^n$ of c_∞ by eech translate $\tilde{d}_\infty \cdot c^n$ of $\tilde{d}_\infty$ by a power of the motion c. The latter correspond to the encounters with X which result from successive treversels of c, and since c,d do not meet in any other way, $\tilde{c}_\infty$ does not meet any other trenelate $\tilde{d}_\infty \cdot g$, where $g \in \pi_1(S)$. We want to show that A(c) is similarly crossed - by A(d) and its translates $A(d)\cdot c^n$, and no other translatee $A(d)\cdot g$ - as thie will imply a single crossing of $\alpha(c)$ by $\alpha(d)$.

Since d is simple, all the translates $\tilde{d}_\infty \cdot c^n$ are disjoint, hence $\tilde{d}_\infty$ cennot lie elong A(c) by the argument of Lemma 3, case 2. Thus $A(d) \neq A(c)$ and the two axes must cross, otherwise there could not be a single crossing of $\tilde{c}_\infty$ by $\tilde{d}_\infty$. We therefore have a situation like that shown in Fig. 12.



Trenslates of $A(d)$ by the motion c then also croes $A(d)$, but translates by any other $g \in \pi_1(S)$ do not, since thie would imply a crossing of $\tilde{c}_\infty$ by $\tilde{d}_\infty .g$. □

It follows easily from Lemmae 3-5 that the curves $\alpha_1,\ldots,\alpha_{2p}$ on S with the properties (i), (ii) stated at the beginning of this section can be taken as geodesics. (The only worry is that disjoint curves might have coincident geodeaic representatives, e.g. α_i and α_{i+2}. But this is ruled out beceuse one of α_i, α_{i+2} crosses a curve which the other doea not croee.) Any geodesics $\alpha_1,\ldots,\alpha_{2p}$ with properties (i), (ii) will be eaid to conetitute a *canonical geodesic system*. The proofs of Lemmas 1-5 show that $\alpha_1,\ldots,\alpha_{2p}$ cen be represented as $\alpha(g_1),\ldots,\alpha(g_{2p})$ for some $g_1,\ldots,g_{2p} \in \pi_1(S)$ whose axes have intersections which reflect the intersections of $\alpha_1,\ldots,\alpha_{2p}$. We combine these results into the following.

Theorem 3. For the surface S of genus p there are elements $g_1,\ldots,g_{2p} \in \pi_1(S)$ euch that $\alpha(g_1),\ldots,\alpha(g_{2p})$ is a canonicel geodesic system on S. This meane

(i) For each i,$A(g_i)$ crossee all translates $A(g_{i+1}).g_i^n$ of $A(g_{i+1})$ by the motion g_i, but no other translates $A(g_{i+1}).g$ where $g \in \pi_1(S)$

(ii) Except for the trivial coincidence of $A(g_i)$ with its own translates under the motions g_i^n, the different translates $A(g_i).g$, $1 \leqslant i \leqslant 2p$, $g \in \pi_1(S)$, have no other intersections. □

The above theorem does not appear explicitly in the literature, to my knowledge, but it would surely have been clear to Poincaré. Since Poincaré [1904] also proves the existence of homeomorphisms mapping curve systems with given intersections onto other systems with the same intersections, it also seems reasonable to attribute the following theorem to him.

Theorem 4. If $\alpha_1,\ldots,\alpha_{2p}$ and $\alpha_1',\ldots,\alpha_{2p}'$ are canonicel geodesic systems on S, then there is a homeomorphism $\tau : S \to S$ such that $\tau(\alpha_i) = \alpha_i'$ for each i.

Proof. It is easily seen by induction on i thet cutting S along $\alpha_1,\ldots,\alpha_{2i-1}$ gives a connected surface, with two boundaries that are joined by α_{2i}, and that cutting along α_{2i} gives a connected surface with one boundary. An Euler characteristic calculation then shows that when $i = p$ the cut surface becomes simply connected, i.e. a polygon P.

The boundary ∂P of P is divided into 8p-4 segments identified in pairs : one pair for each of α_1,α_{2p} and two for each α_i, $2 \leq i \leq 2p - 1$ since each of the latter α_i is subdivided at two points, by α_{i-1} and α_{i+1}. We label and orient these segments by the names of the curves from which they originate, calling the pieces of α_i, $2 \leq i \leq 2p - 1$, α_{i1} and α_{i2}. The latter pieces are determined by the orientation of α_i if we let α_{i1} run from α_{i-1} to α_{i+1} and α_{i2} from α_{i+1} to α_{i-1} (See Fig. 14 below for the case $p = 2$.)

It can then also be checked by induction that the cyclic sequence of labels on ∂P is unique, hence the polygon P can be mapped onto the polygon P' which results from cutting S along the α_i' in such a way that each segment in ∂P is mapped onto the corresponding primed segment in $\partial P'$. Adjusting the map by a deformation if necessary, we can arrange that equivalent point pairs on ∂P are sent to equivalent point pairs on $\partial P'$, so that we have a homeomorphiam $\tau : S \to S$ with $\tau(\alpha_i) = \alpha_i'$. □

Informally speaking, both P and P' can be pasted together to "look like" the standard picture of Fig. 10. This shows that there are homeomorphisms of S which send both the α_i and α_i' systems to standard position, and hence there is a homeomorphiam of one system onto the other.

Let us now compare our situation to where we were with the torus efter Theorem 1. Then it was cleer, because the basepoint for $\pi_1(T)$ could be taken at the intersection of the two canonical curves a,b, that an automorphiam I of $\pi_1(T)$ was "given" by the images a',b' of e,b, and it only remained to ehow that the I-images of a,b were canonical curves for any I.

Now, in avoiding multiple intersectione, we seem to have lost a basepoint for $\pi_1(S)$, since there is no point common to all of $\alpha_1,\ldots,\alpha_{2p}$. However, since the α_i cut S down to a simply connected piece, it turns out that an automorphiam I of $\pi_1(S)$ *is* in fact "given" by its effect on the α_i. The main problem is showing that I sends one canonical geodesic system to another, when we know nothing about the automorphisms I.

This is where the avoidence of multiple intersections becomes important. We only have to show that I preserves the single intersection of α_i with α_{i+1}, for each i, and the non-intersection of α_i with α_j, $j \neq i - 1, i + 1$, which in turn reduces to showing that I preserves intersections and non-intersections of axes. Dehn proved this result very elegantly by looking at the behaviour of $\pi_1(S)$ at infinity. His idea is developed in the next section.

5. Mappings of $\partial\tilde{S}$ induced by automorphiams of $\pi_1(S)$.

As just mentioned above, the key step towards the Dehn-Nielsen theorem is to prove that an automorphism I of $\pi_1(S)$ preserves intersections and non-intersections of axes. We let I ect on the set of axes by sending $A(g)$ to $A(g_I)$ for each $g \in \pi_1(S)$. Since $A(g)$ is determined by $U(g)$ and $V(g)$, it is equivalent to let I act on the set of fundamental points by sending $U(g)$ to $U(g_I)$ and $V(g)$ to $V(g_I)$. We know that this map is well-defined (in fact a bijection) by Fact 7, and it is appropriate to view the action of I on axes as really taking place on $\partial\tilde{S}$. In $\tilde{S}$ itself, I naturally acts on the vertices of the Cayley diagram by sending vertex h to vertex h_I, and this can only be interpreted as an action on axes by going to infinity (letting $h^{\pm n} \to U(h),V(h)$) as in Fact 3, we get $h_I^{\pm n} \to U(h_I),\ V(h_I)$).

It was Dehn's idea to investigate the action of I on axes by approximating axes by edge paths in the Cayley diagram. One then needs nothing but obvious properties of automorphiams in order to conclude that the map vertex $h \mapsto$ vertex h_I is a "quasi-isometry" of the Cayley diagram, and hence that nothing drastic can happen at $\partial\tilde{S}$, where the axes and their intersections are determined.

The approximation of axes by edge paths is implicit in Dehn's [1912a,b] solution of the conjugacy problem (see Nielsen [1927, §7(g)]), but Dehn's analysis of automorphisms seems to be lost, and the proof which follows is extracted from Nielsen [1927, §9], where credit is given to Dehn for the essential ideas.

Theorem 5 (Dehn). $U(g)$, $V(g)$ separate $U(h)$, $V(h)$ on $\partial\mathfrak{S}$ $\Longleftrightarrow$ $U(g_I)$, $V(g_I)$ separate $U(h_I)$, $V(h_I)$.

Proof. The first step of the proof is to approach $U(h)$, $V(h)$ by a doubly infinite edge path $p(h)$ through the vertices h^n, $n = 0, \pm 1, \pm 2, \ldots$. We know from Fact 1 that these vertices lie on e distence curve of $A(h)$. The segments of $p(h)$ between them can be chosen with identical sequences of edge labels (spelling out a word for h), and then $p(h)$ lies within a bounded hyperbolic distance of $A(h)$. We cen similerly approach $U(g)$, $V(g)$ by polygonal paths through the vertex sequences $\{g^{-n} | n \geq N\}$ and $\{g^n | n \geq N\}$, for any positive integer N. Assuming $U(g)$, $V(g)$, $U(h)$, $V(h)$ are distinct, Fact 3 says we can choose N so that the latter paths are arbitrarily far away from $p(g)$, and then if $U(g)$, $V(g)$ eeparate $U(h)$, $V(h)$ we can also join their ends g^{-N}, g^N by a path through successive vertices $g^{-N} = g_0, g_1, \ldots, g_i = g^N$ which are all at least as far from $p(h)$ as g^{-N}, g^N themselves. The resulting path $p(g)$ is then far from $p(h)$.

We cen also interpret "far" in terms of the word metric on the Cayley diagram, which is the minimum number of edges connecting given vertices. Vertices "far apart" in the hyperbolic metric are also "far apart" in the word metric, since their hyperbolic distance is $\leq$ (number of edges × maximum length of an edge), end conversely, since only finitely many vertices lie within a given hyperbolic redius.

To simplify notation we denote the successive vertices of $p(g)$ by $\ldots, g_{-1}, g_0, g_1, g_2, \ldots$ and the successive vertices of $p(h)$ by $\ldots, h_{-1}, h_0, h_1, h_2, \ldots$. Fig. 13 shows how these paths might look.

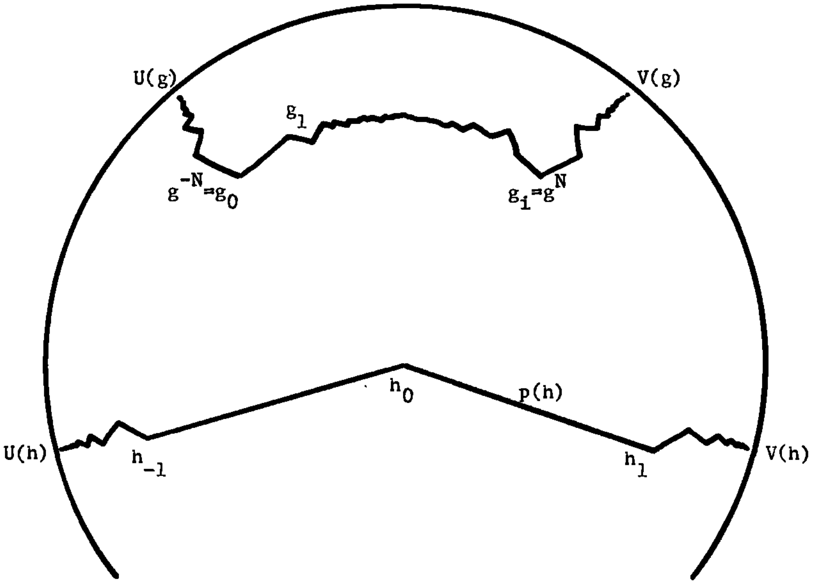


Fig.13

We now show that there ere similar disjoint edge paths, $p(h_I)$ between $U(h_I)$ and $V(h_I)$, and $p(g_I)$ between $U(g_I)$ and $V(g_I)$, by showing that I cennot send vertices which are far apart to vertices which are close together. (This will show that $U(g_I)$, $V(g_I)$ separate $U(h_I)$, $V(h_I)$ if $U(g)$, $V(g)$ separate $U(h)$, $V(h)$, and the converse follows by considering I^{-1}.)

In fact, letting $d(u,v) = \text{length}\,(u^{-1}v)$ denote the minimum number of edges in a path between the vertices u, v in the Cayley diagram (the word metric) we shall show that there ere constants $k, K > 0$ such that

$$d(u_I,v_I) \leqslant Kd(u,v) \quad \ldots \qquad (1)$$

$$d(u_I,v_I) \geqslant kd(u,v) \quad \ldots \qquad (2)$$

This says that I is a "quasi-isometry", while (2) in perticular says that for any distance $d > 0$ in the word metric there is a $D > 0$ such that vertices $> D$ apart are sent to vertices $> d$ apart.

To prove (1), let w be a word of minimal length for $u^{-1}v$. Then

$$\begin{aligned} d(u,v) &= \text{length } (w) \\ \text{while} \quad d(u_I,v_I) &= \text{length } (u_I^{-1}v_I) \\ &= \text{length } ((u^{-1}v)_I) \\ &= \text{length } (w_I) \\ &\leqslant K \text{ length } (w) = Kd(u,v) \end{aligned}$$

where $K = \max \text{ length } \{(a_1)_I,\ldots,(b_p)_I\}$, since a word for w_I is obtained by replacing each generator a_i by $(a_i)_I$ and b_i by $(b_i)_I$. The inequality (2) now follows from

$$d(u,v) \leqslant K'd(u_I,v_I) \ \ldots \ (1)' \ ,$$

which is proved by applying the same argument to I^{-1}, and teking $k = 1/K'$.

Now to obtain a path $p(g_I)$ through the vertices $(g_i)_I$, and e disjoint path through the vertices $(h_i)_I$, we first notice, from (1), that each vertex $(g_i)_I$ can be connected to $(g_{i+1})_I$ by a path of length $\leqslant K$, since g_{i+1} is one edge awey from g_i by definition. We construct $p(g_I)$, and similarly $p(h_I)$, as the union of such subpaths Then to ensure that $p(g_I)$ does not meet $p(h_I)$ it is enough to ensure that each $(g_i)_I$ is distance $\geqslant 2K + 1$ from each $(h_j)_I$. By (2), this can be achieved if each g_i is distance $\geqslant D$ from each h_i, for a suitable D, and D in turn can be achieved by choosing a sufficiently large N in the construction of $p(g)$. □

The Dehn-Nielsen theorem now follows from Theorems 3, 4, 5.

Theorem 6. (Dehn-Nielsen) For any automorphism $I : \pi_1(S) \to \pi_1(S)$ there is a homeomorphism $\tau : S \to S$ which induces I.

Proof. Theorem 3 gives canonicel geodesics $\alpha_i = \alpha(g_i)$ on S, and specifies intersections of the axes $A(g_i).g$, $g \in \pi_1(S)$, on $\tilde{S}$ which cover them. The axes $A((g_i)_I.g_I$, $g_I \in \pi_1(S)$, which cover the geodesics $(\alpha_i)_I = \alpha((g_i)_I)$, are eeen to have the same intersections, by Theorem 5, when one notes that they result from the $A(g_i).g$ by the I action.

Namely $A(g_i).g = A(g^{-1}g_i\ g)$ by Fact 4, and

$$A((g^{-1}g_i\ g)_I) = A((g_I)^{-1}(g_i)_I\ g_I) = A((g_i)_I).g_I.$$

Thus the $(\alpha_i)_I$ also form a canonical geodesic system on S, and there is a homeomorphism $\tau : S \to S$ with $\tau(\alpha_i) = (\alpha_i)_I$ by Theorem 4.

In fact, the proof of Theorem 4 tells ue that the polygons P, P_I which result from cutting S along the α_i and $(\alpha_i)_I$ systems respectively have isomorphically labelled boundaries, hence the generalised Cayley diagrams $\mathcal{D}$, $\mathcal{D}_I$ which result from lifting the α_i and $(\alpha_i)_I$ systems respectively to $\tilde{S}$ are isomorphic. Collinear edges of $\mathcal{D}$ unite to form the axes $A(g_i).g$, $g \in \pi_1(S)$, and collinear edges of $\mathcal{D}_I$ unite to form the axes of $A((g_i)_I).g_I$, $g_I \in \pi_1(S)$.

Fig. 14 shows the surface of genus 2 being cut along the α_i, and Fig.15 showe the cell of the Cayley diagram $\mathcal{D}$ which results.

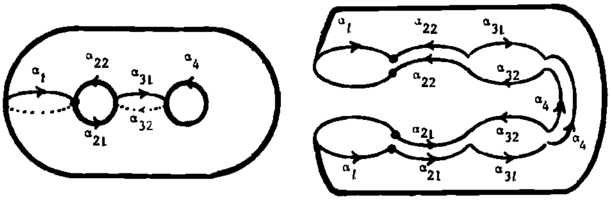


Fig. 14.

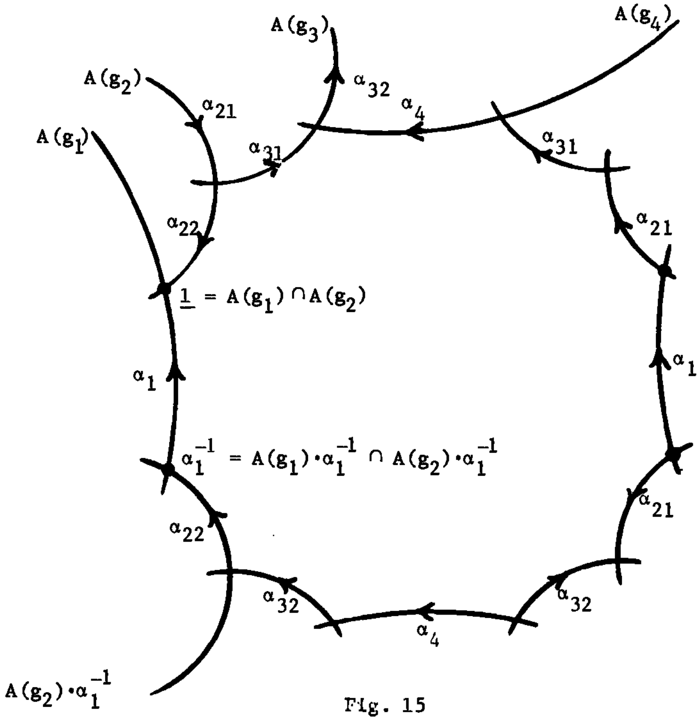


Fig. 15

If we let the identity vertex of $\mathcal{D}$ be $\underline{1} = A(g_1) \cap A(g_2)$ and let the identity vertex of $\mathcal{D}_I$ be $\underline{1}_I = A((g_1)_I) \cap A((g_2)_I)$, then the isomorphism $\mathcal{D} \to \mathcal{D}_I$ means that if an axis $A(h)$ is reached by following a certain sequence of α_i, α_{i1} or α_{i2} edges from $\underline{1}$ in $\mathcal{D}$, then $A(h_I)$ is reached by following the corresponding sequence of edges in $\mathcal{D}_I$ from $\underline{1}_I$. But the vertex of $\mathcal{D}$ representing $g \in \pi_1(S)$ is just

$$A(g_1)\cdot g \cap A(g_2)\cdot g = A(g^{-1}g_1 g) \cap A(g^{-1}g_2 g)$$

and hence g is determined by a common path to these two axes. The corresponding path from $\underline{1}_I$ in D_I leads to

$$\begin{aligned} A((g^{-1}g_1g)_I) \cap A((g^{-1}g_2g)_I) &= A((g_I)^{-1}(g_1)_I g_I) \cap A((g_I)^{-1}(g_2)_I g_I) \\ &= A((g_1)_I)\cdot g_I \cap A((g_2)_I)\cdot g_I \end{aligned}$$

which is the vertex g_I of $\mathcal{D}_I$.

In other words, a closed path from $\alpha_1 \cap \alpha_2$ on S representing $g \in \pi_1(S)$ is mapped by τ onto a closed path from $\tau(\alpha_1 \cap \alpha_2) = (\alpha_1)_1 \cap (\alpha_2)_1$ on $\tau(S)$ representing $g_1 \in \pi_1(S)$. Thus τ induces the automorphism $I : \pi_1(S) \to \pi_1(S)$. | |