Journal of Pure and Applied Algebra 52 (1988) 227–275
North-Holland

ALGEBRAIC LINEARITY FOR AN AUTOMORPHISM OF A SURFACE GROUP

Joan S. BIRMAN*
Department of Mathematics, Columbia University, New York, NY 10027, U.S.A.

Caroline SERIES
Mathematics Institute, University of Warwick, Coventry CV4 7AL, United Kingdom

Communicated by H. Bass
Received 20 October 1986

Let M be a compact surface, $\chi(M)<0$, and let $\Gamma=\pi_1 M$. Let $S(M)$ be the set of isotopy classes of multiple simple loops on M. Each $\Lambda\in S(M)$ determines a family of cyclic words $W=W(\Lambda)$, with associated coinitial graph τ. The finite set of coinitial graphs, obtained as Λ ranges over $S(M)$, is interpreted as a set of 'π_1-train tracks' on M. The linearity theorem asserts that if a topologically induced automorphism ϕ of Γ maps the set of weights W supported on τ to a set supported on τ', then, with appropriate restrictions, the action is linear on the positive linear span of the W's.

Let M be a compact surface of negative Euler characteristic, with genus g and b boundary components and or punctures, and let $\Gamma=\pi_1(M)$. Assume, for the moment, that $b>0$, so that Γ is free. In 1936, in the seminal paper [14], Whitehead studied the question of when a k-tuple $W=\{w_1,\dots,w_k\}$ of cyclic words in the free group Γ can be extended to a basis-up-to-conjugation, or equivalently, when W can be represented geometrically by a family of pairwise disjoint, non-separating simple loops on M. The principal tool introduced by Whitehead was a labelled graph $G(W)$ which has since become known to group theorists as the Whitehead, co-initial or star graph. It may be described as follows. Choosing a fixed basis B_R for Γ, the vertices of $G(W)$ are in one-to-one correspondence with the members of $\Gamma_R=\{x,\bar{x}: x\in B_R\}$. (Here and throughout this paper we write $\bar{x}$ for x^{-1}.) The graph has an edge $E(e,f)$ joining vertices labelled e,f whenever either the two letter sequence $\bar{e}f$ or its inverse $\bar{f}e$ occurs in an element of the set W. Note that it is important to treat the words in W as a cyclic. The edge $E(e,f)$ is labelled by the total number $x_W(e,f)$ of occurrences of either $\bar{e}f$ or $\bar{f}e$ in the set W. For example, if $B_R=\{a,b,c,d\}$, and $W=\{c\bar{d},d,cd\bar{c}\bar{b}\}$, then $x_W(c,\bar{b})=1$, $x_W(\bar{c},\bar{d})=2$, $x_W(b,c)=1$, and so forth. Note especially that $x_W(d,\bar{d})=x_W(\bar{d},d)=1$ in this example. Whitehead's idea was to

* Supported in part by NSF Grant #DMS-8503758.

study the action of Aut Γ on $G(W)$, in order to find an automorphism which would reduce W to a subset of Γ_R.

Let $S(M)$ be the set of isotopy classes of families of pairwise disjoint simple loops on M, subject to the restriction that if $\Lambda \in S(M)$, then no component of Λ is either a loop around a puncture or parallel to a component of ∂M. Each $\Lambda \in S(M)$ is represented by a set of conjugacy classes in Γ, or equivalently, since Γ is free, by a family of cyclic words W_Λ. Our first result is to show that the set of (unweighted) coinitial graphs of the W_Λ, as Λ ranges over $S(M)$, may be reinterpreted as a finite set of canonical 'train tracks' which we call π_1-train tracks on M, where the labels $\{x_W(e,f)\colon e,f \in \Gamma_R\}$ correspond to weights on these tracks (we call this set of labels, ordered in any fixed way, 'π_1-parameters' for W). See Section 1 for the precise definition of a π_1-train track in the special case when Γ is a free group, and see Theorem 1.3 for the statement of this first result, in this special case. See Section 5 for the most general definition (allowing M to be a closed surface) and Theorem 5.4 for the corresponding result.

The second and main result in this paper is a linearity theorem which asserts that if $\varphi \in$ Aut Γ is induced by a diffeomorphism of M and maps a set W of weights supported on one track τ to a set W' supported on the same or any other track τ', then, with some appropriate algebraic restrictions, the action is linear on the positive linear span of the weights W. This theorem is given in three versions. The simplest version is Theorem 2.0. In that version we place restrictions on Γ (we require that Γ be free), on τ (we require that τ be 'orientable') and on $\varphi_* \in$ Aut Γ (we require that φ_* 'preserve τ-orientation'). With all of these restrictions the theorem is relatively easy to prove. In Theorem 4.0 we have removed the restrictions which relate to orientation, at the expense of placing much more subtle restrictions on φ_*. In Theorem 6.0 we remove the restriction on Γ. Theorem 6.0 is the most general version of our work.

While our Theorem 6.0 is related to Thurston's theorem about the piecewise integral linearity of the action of pseudo-Anosov maps on $S(M)$ (cf. [13], and for an excellent exposition [3]) our theorem is not implied by his. On the other hand, our result almost certainly implies Thurston's, although we do not give a proof here.

Our theorem is best appreciated by the working out of concrete examples, in which the linearity seems little short of a miracle. We therefore digress to give an explicit example here. (For a more complicated example, see Example C in Section 7.)

Example A. Take Γ to be the fundamental group of a closed surface M of genus 2, presented as $\langle a,b,c,d; ab\bar{a}\bar{b}cd\bar{c}\bar{d}\rangle$, or if preferred remove a point from M and declare Γ to be free with the same set of generators. Let $W = \{w_1, w_2, \ldots, w_5\}$, where

$$w_1 = b\bar{a}\bar{c}, \qquad w_3 = \bar{c}, \qquad w_5 = \bar{c}\bar{d},$$
$$w_2 = \bar{c}\bar{a}, \qquad w_4 = \bar{c}b.$$

As an example of an element in the positive linear span $\mathrm{Sp}^+(W)$ of W choose:

$$z = \bar{c}\bar{d}\bar{c}\bar{a}\bar{c}\bar{d}\bar{c}b.$$

We can write this word z in the form $w_5w_2w_5w_4$. However '$z = w_2 + w_4 + 2w_5$' in our setting has quite a different meaning. It means that the π_1-parameters associated to the cyclic word z are exactly the sum of those for the cyclic words w_2, w_4 and w_5 (twice). In this example, it means that the 2-letter syllables which occur at the interfaces between w_5 and w_2, w_2 and w_5, w_5 and w_4 and w_4 and w_5 in z, namely $\bar{d}\bar{c}$, $\bar{a}\bar{c}$, $\bar{d}\bar{c}$ and $b\bar{c}$, are exactly those formed from the first and last letters of w_2, w_4 and w_5 (twice), i.e. $\bar{a}\bar{c}, b\bar{c}, \bar{d}\bar{c}, \bar{d}\bar{c}$.

Now, let $\varphi_* \in \mathrm{Aut}\,\Gamma$, $\Gamma = \pi_1 M$, be defined by

$$\varphi_*(a) = \bar{c}\bar{a}, \qquad \varphi_*(b) = \bar{b}\bar{a},$$

$$\varphi_*(c) = \bar{b}\bar{a}d, \qquad \varphi_*(d) = \bar{c}.$$

Calculating, and reducing we see that

$$v_1 = \varphi_*(w_1) = \bar{b}\bar{a}ac\bar{d}ab \sim c\bar{d}a, \qquad v_2 = \varphi_*(w_2) = \bar{d}abac,$$

$$v_3 = \varphi_*(w_3) = \bar{d}ab, \qquad v_4 = \varphi_*(w_4) = \bar{d}ab\bar{b}\bar{a} \sim \bar{d},$$

$$v_5 = \varphi_*(w_5) = \bar{d}abc,$$

$$\varphi_*(z) = \bar{d}abc\bar{d}abac\bar{d}abc\bar{d}ab\bar{b}\bar{a} \sim \bar{d}abc\bar{d}abac\bar{d}abc\bar{d}.$$

The linearity theorem asserts that φ_* maps $\mathrm{Sp}^+(W)$ linearly into $\mathrm{Sp}^+(V)$, where $V = \varphi_*(W) = \{v_1, \ldots, v_5\}$, i.e. in this case we expect that

$$\varphi_*(z) = v_2 + v_4 + 2v_5,$$

as may be verified by adding up the π_1-parameters in the cyclically reduced images. This is the statement of our linearity theorem, viz:

$$\varphi_*(w_2 + w_4 + 2w_5) = v_2 + v_4 + 2v_5.$$

In this example the linearity theorem is fairly straightforward. In more complicated examples the decompositions are much more subtle, because in order to decompose a word like z into a sum like $w_2 + w_4 + 2w_5$, it will be necessary to break apart the w_i's into subwords and recombine syllables in unexpected ways. For example, see the material after the proof of Lemma 7.2, in Section 7 below. The linearity theorem asserts that similar decompositions are possible after φ_* is applied.

As is seen from examination of the example above, summing weights on a Whitehead graph is equivalent to decomposing a family of words into blocks and then regrouping these blocks to form a multiple simple loop. The assertion of our theorem is that an automorphism φ has the property that it carries over this decomposition into blocks: the image words decompose in a similar way and cancellations appear in exactly the right places between end terms in the images of the various

blocks to be regrouped in the image. Geometrically the division points between blocks correspond to intersection points between representative curves on M. As we shall see in detail below, if two curves have weights lying on the same graph, then they have common letters corresponding to each geometric intersection point. These are the points at which the words split into blocks. The further restrictions in the statement of our theorem relate to the relative orientation of the two curves and their images at these intersection points. Our proof depends on looking at the geometrical implications of this splitting of the curves at intersection points and combining this with the surprisingly rigid constraints imposed by the algebra. We were unable, except in the simplest cases, to find a purely algebraic proof, although one suspects this might be very interesting.

We now discuss the relationship of our theorem to the work of Thurston on piecewise linear parameters for $S(M)$ and the PL action of diffeomorphisms. (Cf. Thurston's theorem on the piecewise integral linear action of pseudo-Anosov maps on the Dehn–Thurston parameters of [5] or the traintrack parameters of [3].) First of all, our theorem is algebraic in content, while Thurston's is geometric-algebraic. Thus our parameters are easier to handle than those of either [5] or [3] in that they do not require isotopy of curves into a nice position relative to some pants decomposition of M. Indeed this process is replaced by the quite mechanical process of shortening words algebraically, which except in the case of closed surfaces is just free reduction. One should compare for example the calculations in Penner's thesis [9] and our Examples A and C to see this.

The automorphism φ_* which we defined in Example A is in fact pseudo-Anosov [7]. For pseudo-Anosov maps there is an invariant train track, and in fact we realize it in Example A as the Whitehead graph $G(U)$, where U is the 5-tuple $\{d\bar{c}\bar{a}, \bar{a}\bar{c}, \bar{a}, \bar{d}\bar{a}, \bar{b}\bar{a}\}$. The 3 graphs, $G(W)$, $G(V)$, $G(U)$ are illustrated in Fig. 1, embedded in the cut open surface M. The graphs are all distinct. Our theorem predicts that φ_* acts linearly on $\mathrm{Sp}^+(W)$ mapping into $\mathrm{Sp}^+(V)$ and also (using it in another way) on $\mathrm{Sp}^+(U)$ mapping it linearly onto $\mathrm{Sp}^+(U)$. Thurston's theorem deals with the latter action, but not with the former. Thus we have found larger sets on which φ_* acts linearly. In addition these sets have algebraic meaning and are computable by algebraic means. Our theorem also holds for a wider class of maps than pseudo-Anosovs. We found it surprisingly easy to find examples satisfying the conditions of our theorem. We found we could choose both φ and $\Lambda \in S(M)$ more or less at random, and most of the time get a set of weights on one graph which were mapped to weights on another graph, so that the conditions under which our linearity theorem holds were satisfied.

The invariant lamination of a pseudo-Anosov map is always carried by one of our π_1-train tracks. In Example A above, this train track is $G(U)$, as proved near page 347 of [8]. (Remark: Nielsen does not talk about π_1-train tracks, however the reader who goes back to his paper armed with that concept will have little difficulty in translating his ideas into statements about π_1-train tracks. Indeed, the concept of a π_1-train track is very helpful in revealing the true content of Nielsen's monu-

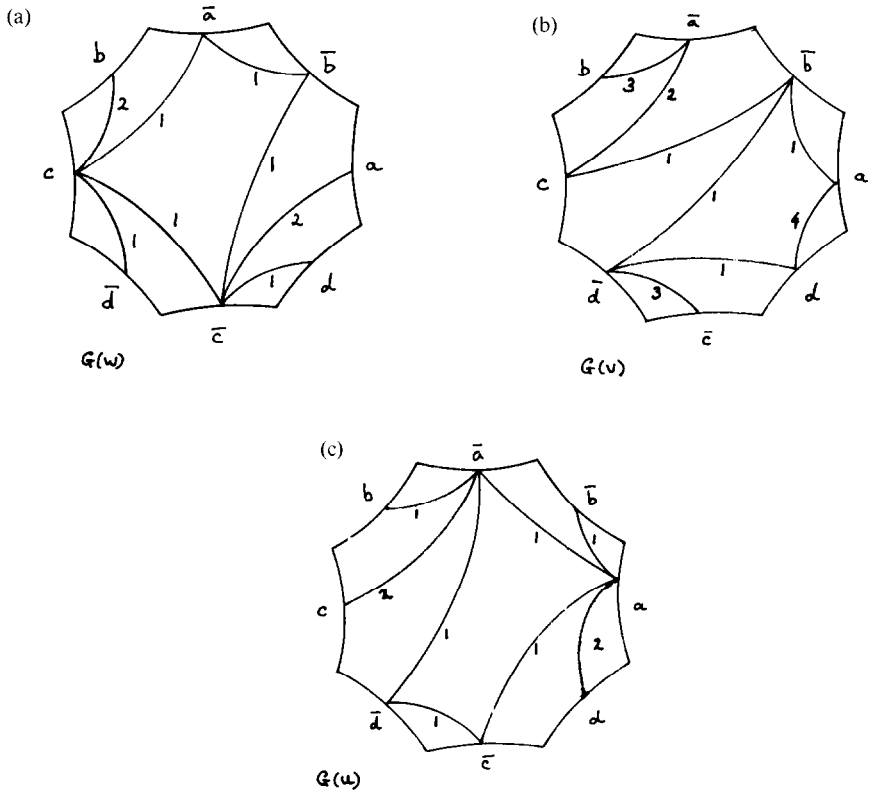


Fig. 1.

mental work.) In this case one verifies easily that $G(\tilde{U}) = G(U)$, where $\tilde{U} = \varphi(U)$. In this example it is also the case that in fact $\tilde{U}$ is contained in the positive linear span of U, so that we may associate to φ acting on $G(U)$ a positive linear matrix Φ_* whose powers represent the iterations of φ (see Example C). Thus the invariant lamination L will be represented by the unique positive eigenvector of A and is supported on $G(U)$. (This in fact gives a representation of leaves of L as 'infinite homotopy classes' or infinite words in Γ. Such representations of laminations are described in detail for the case of the punctured torus in [12].) In Example A it will be seen that in some sense the invariant neighborhood of the lamination we found is very large. The same remark applies to Example C below.

We now explain more fully the interpretation of Whitehead's graphs as train tracks. Fix some hyperbolic metric on M, so that $\Gamma = \pi_1(M)$ may be thought of as a group of isometries of the Poincaré disc $\mathbb{D}$ and $M = U/\Gamma$ where $U \subseteq \mathbb{D}$ is the universal cover of M. Choose a symmetric set of generators Γ_R for Γ which have geometrical meaning as the edge-pairing transformations of the sides of a fundamental region R for the action of Γ in $\mathbb{D}$. (Γ_R is symmetric if $g \in \Gamma_R$ implies $g^{-1} \in \Gamma_R$.)

Label the oriented sides of R by the symbols in Γ_R, so that if $e \in \Gamma_R$ maps a side s_1 to s_2, then s_1 is labelled e on the side interior to R (or $\bar{e}$ on the side exterior to R). In our diagrams (e.g. see Fig. 1) we show the labels on the outside of R. Thus the side labelled e on the outside is carried to the side labelled $\bar{e}$ on the outside by the generator $\bar{e}$. For each $e \in \Gamma_R$, choose a point $P(e)$ on the side whose exterior label is e, in such a way that $\bar{e}P(e) = P(\bar{e})$. Join $P(e)$ to $P(f)$ whenever $x_W(e,f) > 0$ and label this arc $x_W(e,f)$. This gives a labelled graph $G(W)$ with vertex set $P = \{P(e): e \in \Gamma_R\}$. If W represents an element of $S(M)$, then the graph $G(W)$ is such that two edges meet only in the vertices P. We call such a graph *simple*. It may clearly be regarded as a train track in the sense of Thurston (see [3] or [13]) with all of its switches on ∂R, and several branches meeting at each switch.

The set of weights associated to a fixed cyclic word w has a nice interpretation in terms of paths in the universal covering space $U \subseteq \mathbb{D}$. Relative to our fixed hyperbolic metric on M, each simple loop $\tilde{\lambda}$ has a unique smooth geodesic representative. We proved in [1] that for certain special choices of fundamental region there is a remarkable relationship between the path followed by a geodesic and its algebraic representation as a shortest word in the symbols Γ_R. This arises in the following way. Any lift λ of $\tilde{\lambda}$ to $\mathbb{D}$ cuts through a sequence of copies $g_1R, g_2R, \ldots$ of copies of our fixed fundamental region R. The labelling of sides of R described above extends to each of these regions by the action of Γ. If g_iR, $g_{i+1}R$ are adjacent regions, then $e_i = g_i^{-1}g_{i+1} \in \Gamma_R$ is the label of the common side of the two regions on the side interior to $g_{i+1}R$. Thus moving along λ we read off a sequence of labels $\ldots e_i e_{i+1} \ldots$ which is obviously independent of the choice of lift λ.

Moreover, if $\tilde{\lambda}$ is closed this sequence is periodic and its fundamental period represents the free homotopy class of $\tilde{\lambda}$. The crucial fact is that with appropriate choice of R and appropriate restrictions on Γ_R, the fundamental period $w = w(\tilde{\lambda}) = e_1 \ldots e_k$ obtained in this way is cyclically shortest and shortest in its conjugacy class [1]. Thus the geometric path followed by the geodesic representatives of $\Lambda \in S(M)$ determines the weights of the corresponding multiple cyclic word W. The proof of our theorem is based on a detailed exploitation of this interplay between algebra and geometry.

Here is the plan of the paper. In Section 1 we set up notation and review background material, as needed. The linearity theorem is proved in Sections 2–5, beginning with the special case of orientable train tracks and surfaces with free fundamental group (Section 2), then passing to arbitrary train tracks on surfaces with free fundamental group (Sections 3 and 4) and finally to the general case (Sections 5 and 6). In Section 7 we show how to check the orientation condition which is the main restriction in our theorem. In the appendix we give an independent proof of our linearity theorem 2.0, in the special case when M is a once-punctured torus.

We give three examples, which are dicussed at various points in the manuscript. The first, example A, has already been discussed. Example B illustrates a case where the linearity fails; it led us to our condition on orientation at intersection points. Example C is a pseudo-Anosov map. We study it in detail, checking all our orienta-

tion conditions and using it to illustrate the linearity theorem when there is an invariant train track. It will be seen to demonstrate most of the complications in our work, and at the same time to illustrate the non-triviality of the algebraic linearity in very convincing ways. We discovered it early in our investigations. We would have abandoned our work many times, however this one example seemed too amazing to be a matter of chance.

Remark. The reader who has attempted related constructions will be instantly alert to the complications which arise with closed surface groups. These complications are all due to the fact that in closed surface groups certain words (those which contain 'half-relators') have non-unique shortest representatives. As will be seen, almost all of our results hold in this case too, however the proofs required us to deal with myriad technical complications. In this manuscript we focus initially on the cases where these complications do not arise, and later we show how to revise things to handle the exceptional cases involving special words in closed surface groups. See in particular Sections 5 and 6 and the latter part of Section 7 for the exceptional cases. The reader who wishes to do so may simply omit these sections, without losing track of the main arguments.

1. Tight curves, shortest words and π_1-train tracks

In this section we set up notation and introduce our π_1-train tracks and π_1-parameters. We assume throughout that $\Gamma = \pi_1 M$ is a free group, or equivalently, that M is not a closed surface. The modifications in our work which are needed to treat closed surfaces are quite complicated and will be treated separately in Section 5 below. However, we note that almost everything that we say here also holds on closed surfaces, if one excludes the special cases which arise for words containing 'half the defining relator'.

As in the introduction, we fix a hyperbolic metric on M and a representation of Γ as a group of hyperbolic isometries of the Poincaré disc $\mathbb{D}$. (For this and other basic facts about hyperbolic geometry, a good reference is [6].) Choose a fundamental domain R for the action of Γ on $\mathbb{D}$ with all its vertices on $\partial\mathbb{D}$, and let Γ_R be the symmetric set of generators of Γ which pair sides of R. Let N be the set of images of ∂R under Γ. As described in the introduction, each oriented side of N is labelled by an element of Γ_R. We associate to any oriented smooth curve $\tilde{\gamma}$ on M the sequence $\sigma(\tilde{\gamma}) = \ldots e_i e_{i+1} e_{i+2} \ldots$ of labels of sides of N in the order in which they are cut by any lift γ of $\tilde{\gamma}$, choosing always the label on the far side of the edge of N cut by γ. We call $\sigma(\tilde{\gamma})$ the *cutting sequence* of $\tilde{\gamma}$. This sequence will be bi-infinite unless $\tilde{\gamma}$ begins or ends in a boundary component or puncture of M. (The special trivial case in which $\sigma(\tilde{\gamma})$ is empty is uninteresting and will be ignored.) We say that $\sigma(\tilde{\gamma})$ is *shortest* if each finite subword is freely reduced. (We generally prefer the term 'shortest' to 'reduced' in anticipation of the closed surface case.)

It will be convenient to think of the bi-infinite periodic word $\sigma(\tilde{\gamma})$ as a finite *cyclic word*. Clearly the bi-infinite word is shortest if and only if the cyclic word is cyclically shortest, i.e. cyclically reduced. As is well known, two words w, w' which are cyclically shortest represent conjugate elements of Γ if and only if they coincide as cyclic words.

A *multiple cyclic word* is a finite collection of cyclic words. If $\Lambda \in S(M)$ has several components, we write $W(\Lambda)$ for the multiple cyclic word associated to their components. A cyclic or multiple cyclic word is *simple* if it represents an element of $S(M)$.

It will be convenient to extend the class of curves we consider from geodesic curves to what we call *tight* curves. A curve on M, or its lift to $\mathbb{D}$, is *tight* if its cutting sequence is shortest. It is clear that the cutting sequences of tight curves in the same homotopy class are the same. We described in the introduction how to form the graph $G(W)$ of a family of cyclically shortest words W. In the same way we can form the graph of a multiple loop $\Lambda \in S(M)$. Choose any representative of Λ by a family $\tilde{\Lambda}$ of mutually disjoint tight simple curves on M. Let $A(\tilde{\Lambda}) = \pi^{-1}(\tilde{\Lambda}) \cap R$, where $\pi : U \to M$ is the projection of M from the universal cover $U \subseteq \mathbb{D}$. (If M is non-compact we assume also that a tight curve lies always within bounded distance of the geodesic with the same endpoints on $\partial\mathbb{D}$, or equivalently, that its projection on M lies at a bounded distance from some fixed point, say $\pi(0)$, on M.) Let $\tau(\tilde{\Lambda})$ be the weighted graph obtained by collapsing all those arcs which join a given pair e, f of sides of R (exterior labels!) to a single arc from $P(e)$ to $P(f)$, labelled by the number $x_\Lambda(e, f)$ of arcs which were collapsed. If $\Lambda \in S(M)$, the arcs in $A(\tilde{\Lambda})$ will be pairwise disjoint and hence $\tau = \tau(\tilde{\Lambda})$ is simple. Clearly $\tau(\tilde{\Lambda})$ is identical with the graph $G(W)$ of the multiple word $W = \sigma(\tilde{\Lambda})$ and is independent of the particular choice of tight curves representing Λ. Thus we may write $\tau(\Lambda)$ for $\tau(\tilde{\Lambda})$. We note that $\tau(\Lambda)$ is a weighted train track in the sense of Thurston (see, e.g., [3]), having at most one switch on each side of R and no switches in the interior of R.

We now impose conditions on train tracks and weights which ensure that these restricted weights correspond exactly to $S(M)$. If τ is a train track on R and x a weighting on τ, then let $x(e, f)$, $e, f \in \Gamma_R$, be the weight on the branch joining $P(e)$ to $P(f)$.

Condition 1.1 (Switch conditions). These arise because an arc in $A(\tilde{\Lambda})$ leaving R across a side e re-enters across the side $\bar{e}$. Thus we have

$$\sum_{f \in \Gamma_R} x(e, f) = \sum_{f \in \Gamma_R} x(\bar{e}, f) \quad \text{for each } e \in \Gamma_R. \tag{1}$$

We denote by $\Omega(\tau)$ the set of weights on τ satisfying Condition 1.1. If the multiple cyclic word W is such that $x_W = \{x_W(e, f) \in \Omega(\tau)\}$, then we say that W is *supported on* τ, likewise we say that $\Lambda \in S(M)$ is supported on τ if $x_\Lambda \in \Omega(\tau)$.

Now let $x \in \Omega(\tau)$. Replace the branch of τ joining $P(e)$ to $P(f)$ by $x(e, f)$ strands in such a way that the strands corresponding to all possible branches are pairwise

disjoint. There is a unique way to join these strands using the edge pairings of R to form a multiple simple loop on M and a corresponding multiple cyclic word $W(x)$. We denote the collection of all these strands by $A(x)$, and call $W(x)$ the word associated to x. In particular, we note that if W is a multiple cyclic word which is simple, then the set of weights x_W on $\tau(\Lambda)$ determine W uniquely.

Restriction 1.2 (Boundary restrictions). These are needed to ensure that there are no loops parallel to ∂M. We call an edge of τ a *boundary branch* if it joins two sides of R whose ideal endpoints are adjacent points on the circle at infinity, so that its projection on M is parallel to an arc in ∂M, or to an arc which partially surrounds a puncture. The collection of all boundary branches splits into equivalence classes, one for each connected component of ∂M. Let E_B be one such equivalence class, $B \subset \partial M$. We impose the following condition on τ:

> For each connected component of B of ∂M, at least one branch of E_B does not appear on τ. (2)

Condition (2) ensures that the curves determined by τ do not include closed loops which are parallel to B (or a loop surrounding B, if B is a puncture). Let $x \in \Omega(\tau)$ be such that all these branches have non-zero label. Consider, for each branch b in E_B, the arc in $A(x)$ which lies parallel to b and closest to $\partial\mathbb{D}$. When the sides of R are identified, these outermost arcs link to form a loop parallel to B (or a loop surrounding the puncture). Such loops are thus ruled out by condition (2).

The weighted graph $\tau(\Lambda)$, whose weights satisfy Condition 1.1 and Restriction 1.2, is called a *π_1-train track*. (A train track in the sense of [13] is a branched C^1 one-manifold embedded in M.) The *π_1-parameters* of an element $\Lambda \in S(M)$ are its weights on $\tau(\Lambda)$. We sum up what we have said in the following theorem, whose proof is left to the reader.

Theorem 1.3. (i) *Let τ be a π_1-track on R, and let $x \in \Omega(\tau)$. Then $W(x)$ is a multiple simple cyclically reduced word.*

(ii) *Let $\Lambda \in S(M)$. Then $\tau(\Lambda)$ is a π_1-train track and $x_\Lambda \in \Omega(\tau)$.*

(iii) *Let W be a multiple simple cyclic word containing no component which represent loops parallel to ∂M or surrounding punctures. Then the graph $G(W)$ is a π_1-train track τ and $x_W \in \Omega(\tau)$.* □

Notice that our representation of multiple simple words by weights on a graph imposes a 'piecewise linear' structure on $S(M)$. By this we mean that $S(M)$ is covered by the sets $\Omega(\tau)$, each of which is a linear space, and the $\Omega(\tau)$ intersect each other only along lower dimensional subspaces. Several distinct τ are needed to cover $S(M)$, hence the term 'piecewise' linear. The linear structure of $\Omega(\tau)$ arises as follows: the switch condition (Condition 1.1) is linear if we regard the weights as vectors with one component for each branch of τ. Thus $\Omega(\tau)$ has the structure of a positive cone over $\mathbb{N}$. If $x_1, \ldots, x_k \in \Omega(\tau)$ we write

$$\mathrm{Sp}^+\{x_1, \ldots, x_k\} = \left\{ \sum_{i=1}^{k} n_i x_i : n_i \in \mathbb{N} \right\} \subseteq \Omega(\tau).$$

This linear structure leads to a rather curious definition of 'addition' of simple words. To *add* two words w_1, w_2 supported on the same τ, we first add their weights x_{w_1}, x_{w_2} in the vector space $\Omega(\tau)$ and then determine the new simple word obtained by joining the strands in $A(x_{w_1}+x_{w_2})$ as described in Condition 1.1 above. The process is illustrated in Fig. 16(b) of Example C in Section 7.

Orientable train tracks. A train track is *orientable* if each branch can be oriented in such a way that the orientations at each switch are coherent. For example, the train-tracks in Figs. 1(a) and 1(b) are both orientable. In our case this means that the branches of τ with an endpoint at $P(e)$, $e \in \Gamma_R$, must be oriented so that they all point into R or all point out of R at $P(e)$, and so that all branches which have an endpoint at $P(\bar{e})$ point in the opposite sense to those at $P(e)$ relative to R. In particular, if a given letter $e \in \Gamma_R$ appears in a word lying on an oriented train track, then $\bar{e}$ does not.

Recall that an *n-gon* on τ is a region bounded by branches of τ which lifts to a simply connected region B in $\mathbb{D}$ whose boundary contains exactly n switches at which the angle interior to B is zero. It is clear that if τ contains a trigon, or more generally an n-gon with n odd, then τ is not orientable. An example of such a train track is illustrated in Fig. 16(a).

Diffeomorphisms. Obviously $\mathrm{Diff}(M)$ acts on $S(M)$. Let $\varphi \in \mathrm{Diff}(M)$ and $x \in \Omega(\tau)$, for some π_1-train track τ. Then x represents an element $\Lambda(x) \in S(M)$ and $\varphi(\Lambda)$ is again a multiple simple loop and thus an element of $S(M)$. We write $\varphi_*(x) = x_{\varphi(\Lambda)}$. Thus $\varphi_*(x)$ is a weight on $\tau(\varphi(\Lambda))$.

The map φ lifts to a map of $\mathbb{D}$ which extends to a homeomorphism $\bar{\varphi}$ of $\partial\mathbb{D}$. The map $\bar{\varphi}$ is independent of the choice of lift. If λ is a tight curve in $\mathbb{D}$ with cutting sequence σ, with primitive period $W(x)$, and with endpoints $\xi, \eta \in \partial\mathbb{D}$, then we denote by $\varphi_*(\lambda)$ any tight curve in $\mathbb{D}$ whose cutting sequence is periodic with primitive period $W(\varphi_*(x))$ and whose endpoints on $\partial\mathbb{D}$ are $\bar{\varphi}(\xi)$ and $\bar{\varphi}(\eta)$. We shall always assume that diffeomorphisms are orientation preserving ($\varphi \in \mathrm{Diff}^+(M)$), but note in the statement of Theorem 6.0 that our results apply to the orientation reversing case also.

2. The Linearity Theorem, Part I (τ orientable, Γ free)

In this section we come to our main result on the piecewise linearity of the action of $\varphi \in \mathrm{Diff}(M)$ on the π_1-parameters for $S(M)$ set up in Section 1. In order to make the ideas in the proof clear, we begin with the special case in which τ is orientable. Later, in Section 3, we will introduce the modifications needed to handle the general

case, which will be treated in Section 4. In this section and in Sections 3 and 4 we assume, as in Section 1, that M is not closed. However, we stress again that except for special situations everything we say holds equally well on closed surfaces.

Theorem 2.0 (The Linearity Theorem, Part I). *Let M be a hyperbolic surface with boundary and or punctures, and with fundamental region R and generators Γ_R for $\Gamma = \pi_1(M)$ chosen as in Section* 1. *Let τ and τ' be oriented π_1-train tracks on M. Suppose that $\varphi \in \mathrm{Diff}(M)$ and that $x_i \in \Omega(\tau)$, $\varphi_*(x_i) \in \Omega(\tau')$ for $i = 1, \ldots, k$. Suppose further that if $W(x_i)$ is oriented coherently with τ, then $\varphi(W(x_i)) = W(\varphi_*(x_i))$ is oriented coherently with τ'. Then φ acts linearly on $\mathrm{Sp}^+\{x_1, \ldots, x_k\}$.*

We note that the train tracks of Figs. 1(a) and 1(b) in Example A of the introduction are both orientable.

This example illustrates Theorem 2.0.

Remark. Throughout the proof we assume that $\varphi \in \mathrm{Diff}^+(M)$, however, it is easy to see that this assumption is unnecessary, see the remark following Theorem 6.0.

Proof. The proof of Theorem 2.0 will occupy the remainder of Section 2. It will be seen to rest on the observation that the addition of weights in $\Omega(\tau)$ may be interpreted as the topological operation of surgery. The idea is to compare the surgeries of the families of curves associated to $\{x_i\}_{i=1}^k$ and $\{\varphi_*(x_i)\}_{i=1}^k$ and to show that combinatorially they are the same.

Our first object, then, will be to discuss intersecting curves and the interpretation of surgery.

Lemma 2.1. *Let γ, γ' be tight curves in $\mathbb{D}$ supported on τ and suppose that $\gamma \cap \gamma' = P \in gR$, $g \in \Gamma$. Then the segments $\gamma \cap gR$, $\gamma' \cap gR$ meet $\partial(gR)$ on at least one common side.*

Proof. The segments $\gamma \cap gR$ and $\gamma' \cap gR$ lie over branches of τ, and since the two segments intersect, while τ is a simple graph, the branches must either coincide or meet in a common switch on some side of ∂R. Thus $\gamma \cap gR$ and $\gamma' \cap gR$ both meet s. □

We now extend the notion of a tight curve to that of a *tight family*. A family F of tight curves in $\mathbb{D}$ is *tight* if each pair of curves in F intersects at most once and if in addition at most two curves pass through any one point. The family is *simple* if each curve projects to a simple curve on M. We shall only deal with tight families whose projections onto M contain finitely many components. If curves $\gamma, \gamma', \gamma''$ belong to a tight family F and if $\gamma \cap \gamma'$, $\gamma' \cap \gamma''$ and $\gamma'' \cap \gamma$ are all non-empty and distinct, then we denote the triangular region in $\mathbb{D}$ which they determine by $\Delta(\gamma, \gamma', \gamma'')$. The triangle Δ is *minimal for F* if no other curve in F intersects it.

Lemma 2.2. *Let F be a tight family in $\mathbb{D}$, and let $l \in F$. Let $G \subset F$ be the subset of all curves in F which intersect l. Suppose that $\gamma, \gamma' \in G$, with $\gamma \cap \gamma' \neq \emptyset$. Then there is a minimal triangle $\Delta(\mu, \mu', l)$ for G on the same side of l as $\Delta(\gamma, \gamma', l)$.*

Proof. If $\Delta(\gamma, \gamma', l)$ is minimal, we are done. If not, we will construct a nested sequence of triangles $\Delta(\gamma, \gamma', l) = \Delta(\gamma_0, \gamma', l) \supset \Delta(\gamma_0, \gamma_1, l) \supset \Delta(\gamma_1, \gamma_2, l) \supset \Delta(\gamma_2, \gamma_3, l) \supset \cdots$, all on the same side of l as $\Delta(\gamma, \gamma', l)$, with each $\gamma_j \in G$. Since the intersection points of G are isolated, and since at most two curves meet in one point, this will prove the claim. See Fig. 2 for the construction. Assuming $\Delta(\gamma, \gamma', l)$ to be non-minimal, there must be a curve in G which intersects $\gamma_0 = \gamma$, say, between $P_0 = \gamma_0 \cap \gamma'$ and $Q_0 = \gamma_0 \cap l$. See Fig. 2. Let $\gamma_1 \in G$ be the curve intersecting γ_0 closest to Q_0, and let $P_1 = \gamma_1 \cap \gamma_0$, $Q_1 = \gamma_1 \cap l$. Either $\Delta(\gamma_0, \gamma_1, l)$ is minimal or there are curves in G intersecting γ_1 between P_1 and Q_1.

Choose γ_2 to be the one of these curves intersecting γ_1 closest to Q_1, and let $P_2 = \gamma_2 \cap \gamma_1$, $Q_2 = \gamma_2 \cap l$. Again, either $\Delta(\gamma_1, \gamma_2, l)$ is minimal or there are curves in G intersecting γ_2 between P_2 and Q_2; hence we can construct curves γ_3, and points P_3, Q_3 as above, and so on. The point Q_{j+1} is always between Q_j and Q_{j-1} because, if Q_{j+1} were on the other side of Q_j, then Q_{j+1} would not meet γ_j between P_j and Q_j. The construction ends with a minimal triangle $\Delta(\mu, \mu', l)$ on the same side of l as $\Delta(\gamma, \gamma', l)$. □

Tight families F, F' in $\mathbb{D}$ are said to be *related* if there is a 1–1 correspondence $j : F \to F'$ such that the restriction of j to endpoints on $\partial \mathbb{D}$ is order-preserving. The

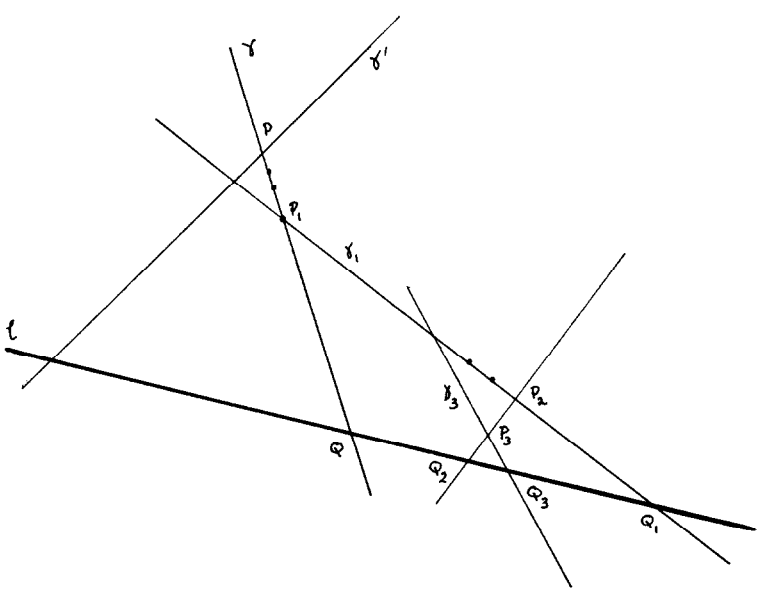


Fig. 2.

correspondence j preserves the *combinatorial pattern* if, for each $\gamma \in F$, it preserves the order $\ldots \gamma_{\mu_i}, \gamma_{\mu_{i+1}} \ldots$ in which curves in the family intersect γ, reading intersection points along γ.

Let $F = \{\gamma_i\}_{i+1}^{\infty}$ be a tight simple family supported on τ, and oriented coherently with τ. The *surgery of F relative to τ*, $S_\tau(F)$, is the family of disjoint curves on $\mathbb{D}$ obtained by surgery on F, where at each intersection point we move onto the other curve following the direction indicated by the orientation of τ. The projection of $S_\tau(F)$ to M, also denoted $S_\tau(F)$, is thus an element of $S(M)$. (We could of course equally well first project F to M and carry out the surgery on M.)

We wish to show that the weight of $S_\tau(F)$ is the sum of the weights of the curves in F. We do this by pushing the intersection points of curves in F which lie in R systematically close to ∂R, where it will be easier to understand the effect of surgery.

Lemma 2.3. *Let F be a tight simple family supported on an oriented π_1-train track τ. Then $S_\tau(F)$ is supported on τ and $x_{S_\tau(F)} = \sum_{\gamma \in F} x_\gamma$. (In the sum on the right we of course choose only one representative of each Γ equivalence class in F.)*

Proof. Let $B_\varepsilon(\partial R)$ denote an ε tubular neighborhood of those components of ∂R which do not cover components of ∂M. Assume first that M is compact. Choose ε small enough so that the components of $B_\varepsilon(\partial R)$ are disjoint. We begin by showing that we can alter F equivariantly to a related tight simple family F' such that F, F' have the same combinatorial pattern, but such that F' has all its intersection points inside the Γ-translates of $B_\varepsilon(\partial R)$.

To prove this assertion, suppose that $\gamma, \gamma' \in F$, with $\gamma \cap \gamma' \in R - B_\varepsilon(\partial R)$. By Lemma 2.1 the segments $\gamma \cap R$, $\gamma' \cap R$ have at least one pair of endpoints on a common side, say s, of ∂R. Let β be any other curve in F which intersects the triangle $\Delta(\gamma, \gamma', s)$. Then $\beta \cap R$ has one endpoint on s, for if not, the three branches of τ supporting $\gamma \cap R$, $\gamma' \cap R$, $\beta \cap R$ would form a trigon, which is impossible because τ is orientable. Thus we may apply Lemma 2.2 to the family $G = \{\beta \in F\colon \beta \cap \Delta(\gamma, \gamma', s) \neq \emptyset\}$ with $l = s$, to find a minimal triangle $\Delta(\gamma_1, \gamma_2, s)$, $\gamma_1, \gamma_2 \in F$, on the same side of s as $\Delta(\gamma, \gamma', s)$. We may clearly replace $\gamma_1 \cap R$, $\gamma_2 \cap R$ by segments with the same endpoints on ∂R such that $\gamma_1 \cap \gamma_2 \in B_\varepsilon(s)$ without altering the intersection pattern of curves in $F \cap R$. We may extend this alteration equivariantly to $\mathbb{D}$. Without loss of generality, it may be assumed that the new intersection point is arbitrarily close to s, say in $B_\delta(s)$. Removing $B_\delta(s)$ from $B_\varepsilon(s)$, we obtain a modification R_1 of R with modified neighborhood $B_{\varepsilon_1}(\partial R_1)$ in which we can repeat the argument, seeking a minimal triangle in R_1. Induction on the number of intersection points completes the proof that F can be moved to F'. Since F, F' have the same endpoints on $\partial \mathbb{D}$, they are clearly related, and so our assertion holds.

For non-compact M the proof is nearly identical. Let M' be a compact subsurface which supports $\pi(F)$, and let $R' = R \cap \pi^{-1}(M')$. Choose ε small enough so that the components of $B_\varepsilon(\partial R) \cap R'$ are disjoint. Proceed as before, pushing intersection points into $B_\varepsilon(\partial R) \cap R'$, one at a time.

Since F, F' have the same combinatorial patterns, the surgeries $S_\tau(F)$, $S_\tau(F')$ are the same, up to homotopy. Moreover, we have, for each pair $e, e' \in \Gamma_R$, $x_F(e,e') = x_{F'}(e,e')$, where $x_F(e,e')$, $x_{F'}(e,e')$ denote the number of strands in F and F' joining side e to side e' of ∂R. Thus it will suffice to prove the proposition for F'.

For $e \in \Gamma_R$, let s be the side of ∂R with interior label e. Let $B_\varepsilon^+(e), B_\varepsilon^-(e)$ denote the components of $\partial B_\varepsilon(s)$ on the same side of s as the labels $e, \bar{e}$ respectively. Thus $B_\varepsilon^-(e) = B_\varepsilon^+(\bar{e})$. All intersections of curves in F' occur in $\bigcup \{B_\varepsilon(s) \,|\, s \text{ is a side of } \partial R\}$. Suppose τ is oriented pointing away from side e of s. For each intersection in $B_\varepsilon(s)$, the two intersecting strands enter $B_\varepsilon(s)$ across $B_\varepsilon^-(e)$ and leave across $B_\varepsilon^+(e)$. Suppose that in total $k(e)$ strands cross $B_\varepsilon(s)$. Then the local effect of surgery on these strands is to produce a family of $k(e)$ 'parallel' strands joining $B_\varepsilon^-(e)$ to $B_\varepsilon^+(e)$. The remaining part of F' in $R - B_\varepsilon(\partial R)$ consists of a disjoint collection of strands, $x_{F'}(\bar{e}, \bar{e}')$ of which join the pair $B_\varepsilon^+(e), B_\varepsilon^+(e')$. The effect of surgery is to link these strands with the parallel strands across each $B_\varepsilon(s)$. It is clear that the resulting multiple loop is the same as the one one would obtain by linking the family consisting of $x_{F'}(\bar{e}, \bar{e}')$ disjoint strands joining the interior sides e, e' of ∂R for each pair $e, e' \in \Gamma_R$.

This family has total weight $\sum_{\gamma \in F} x_\gamma(\gamma)$, which completes the proof. □

Let τ be a π_1-train track and suppose that $x_i \in \Omega(\tau)$, $n_i \in \mathbb{N}$, for $i = 1, \ldots, k$. We denote by $\mathbb{F}(n_1 x_1, \ldots, n_k x_k)$ the tight simple family consisting of all lifts of curves corresponding to $n_i x_i$, $i = 1, \ldots, k$, where by a curve corresponding to $n_i x_i$, we mean n_i 'parallel' but disjoint curves each of which has cutting sequence $W_i = W(x_i)$.

Suppose then that $x_i \in \Omega(\tau)$, $\varphi_*(x_i) \in \Omega(\tau')$, $i = 1, \ldots, k$, as in the statement of Theorem 2.0. Let $\sum_{i-1}^k n_i x_i \in \mathrm{Sp}^+\{x_1, \ldots, x_k\}$, and let $F = \mathbb{F}(n_1 x_1, \ldots, n_k x_k)$, $\varphi_*(F) = \mathbb{F}(n_1 \varphi_*(x_1), \ldots, n_k \varphi_*(x_k))$. Let $\varphi(F)$ denote the topological image of F under some fixed lift of φ to $\mathbb{D}$. Notice that the curves in $\varphi(F)$ are not in general tight.

Our aim is to compare the surgeries $S_\tau(F)$ and $S_{\tau'}(\varphi_*(F))$. The idea is that surgery on a family of oriented curves depends only on the combinatorial pattern in which curves in the family intersect, and that the combinatorial patterns of the families F and $\varphi(F)$ are identical, while those of $\varphi(F)$ and $\varphi_*(F)$ are close enough that $\varphi_*(F)$ may be replaced by a family with the same π_1-parameters whose pattern is identical with that of $\varphi(F)$. To illustrate the main line of argument, we begin with the special case in which the patterns of F and $\varphi_*(F)$ are identical.

By the *segments* of a family of curves G we mean the connected components of $G - \{\gamma \cap \gamma' : \gamma, \gamma' \in G\}$.

Theorem 2.0: special case. *Suppose, with the hypotheses of Theorem* 2.0 *and the notation above, that* $\varphi_* : F = \mathbb{F}(n_1 x_1, \ldots, n_k x_k) \to \varphi_*(F) = \mathbb{F}(n_1 \varphi_*(x_1), \ldots, n_k \varphi_*(x_k))$ *preserves combinatorial patterns. Then*

$$\varphi_*\left(\sum_{i=1}^k n_i x_i\right) = \sum_{i=1}^k n_i \varphi_*(x_i).$$

Proof. The surgery $S_\tau(F)$ can be pictured as sequences of segments of curves in the family F, where we follow along a segment in the direction of the orientation of τ to an intersection point, and then move onto an adjacent segment of the intersecting curve in the direction consistent with the orientation of τ. Since by hypothesis φ_* preserves combinatorial patterns and the orientation of the curves in F, $\varphi_*(F)$ relative to τ and τ', the surgery $S_{\tau'}(\varphi_*(F))$ follows the corresponding sequence of segments of $\varphi_*(F)$ in the same order.

Let λ be a connected component of $S_\tau(F)$ and let $\xi \in \partial\mathbb{D}$ be the positive endpoint of λ. Suppose that the intersection points of curves in F occur in order along λ as $\ldots, P_{i_n} = \gamma_{i_n} \cap \gamma_{i_{n+1}}, P_{i_{n+1}} = \gamma_{i_{n+1}} \cap \gamma_{i_{n+2}}, \ldots$. Along the corresponding component $\bar{\lambda}$ in $S_{\tau'}(\varphi_*(F))$ one will see the intersections $\ldots, \bar{P}_{i_n} = \varphi_*(\gamma_{i_n}) \cap \varphi_*(\gamma_{i_{n+1}})$, $\bar{P}_{i_{n+1}} = \varphi_*(\gamma_{i_{n+1}}) \cap \varphi_*(\gamma_{i_{n+2}}), \ldots$, occurring in the same order, since by hypothesis φ_* preserves the combinatorial patterns of F and $\varphi_*(F)$.

Let $M' \subset M$ be a compact subsurface containing all the curves in F. There is a constant C depending only on M' and φ such that $d(\varphi(\alpha), \varphi_*(\alpha)) \leq C$ for any geodesic curve α in $\pi^{-1}(M')$. (Here d denotes hyperbolic distance.) Since a tight curve is within bounded distance of the geodesic curve with the same endpoints, we obtain the same inequality for any tight curve $\alpha \in F$. In particular, $d(\varphi(P_i), \bar{P}_i) \leq C'$ for each i.

The points $\ldots, \bar{P}_{i_n}, \bar{P}_{i_{n+1}}, \ldots$ lie on $\bar{\lambda}$ which is a component of $S_{\tau'}(\varphi_*(F))$. Since $\bar{\lambda}$ has a cutting sequence which is shortest (cf. e.g. [11, Proposition 4.2]), it has a definite positive endpoint on $\partial\mathbb{D}$, say η, and $\lim_{n\to\infty} \bar{P}_{i_n} = \eta$.

On the other hand, $\lim_{n\to\infty} \varphi(P_n) = \bar{\varphi}(\lim_{n\to\infty} P_n) = \bar{\varphi}(\xi)$. Since $d(\varphi(P_n), \bar{P}_n) \leq C'$ for each n, we have $\eta = \bar{\varphi}(\xi)$. This is equivalent to the statement that $\bar{\lambda} = \bar{\varphi}_*(\lambda)$, which is what we are required to prove. □

In general, the combinatorial patterns of F and $\varphi_*(F)$ will not agree, and the remainder of this section is devoted to circumventing this difficulty. The key to the problem lies in the following situation. Suppose that $\lambda_1, \lambda_2, \lambda_3$ are curves in a tight simple family which intersect in pairs to form a triangle $\Delta(\lambda_1, \lambda_2, \lambda_3)$ as in Fig. 3(a). Altering any one of these curves, say λ_3, to a curve λ_3' as in Fig. 3(b), we obtain a new triplet with the same endpoints on $\partial\mathbb{D}$ but with a different combinatorial pat-

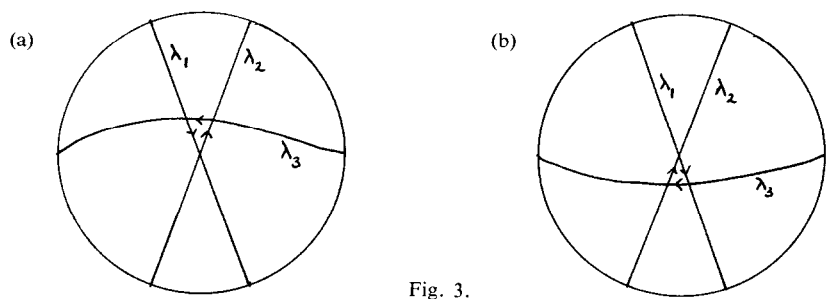


Fig. 3.

tern in which the orders of intersection have been transposed along each of the three curves. We call such an alteration of a triangle a *transposition*. The combinatorial patterns of F and $\varphi_*(F)$ differ by transpositions of triangles. We shall show that by homotopies of curves which do not alter π_1-parameters, triangles may be transposed until the two patterns agree.

The homotopies which we use are achieved by a basic move which we call a *finger move*. Let λ_1, λ_2 be tight curves supported on a train track τ and suppose that $\lambda_1 \cap \lambda_2 = P \in \mathbb{D}$. By Lemma 2.1 the cutting sequences $\sigma(\lambda_1), \sigma(\lambda_2)$ have a common block $B = u_1 u_2 \dots u_k$, $u_i \in \Gamma_R$. The point P occurs in one of the entries of the sequence J of regions $R_0, u_1 R_0, \dots, u_1 \dots u_k R_0$ determined by B, where R_0 is some copy of R. A finger move is an isotopy which pushes P to any other region in J and which replaces λ_1, λ_2 by tight curves λ_1', λ_2' with the same cutting sequences. We discuss the problem of making such moves equivariantly below.

We begin by showing that any triangle $\Delta(\lambda_1, \lambda_2, \lambda_3)$ formed by curves $\lambda_1, \lambda_2, \lambda_3$ supported on τ and intersecting in pairs may be altered by finger moves so that the altered triangle $\Delta(\lambda_1', \lambda_2', \lambda_3')$ is transposed.

We fix notation as follows. For any triangle $\Delta = \Delta(\lambda_1, \lambda_2, \lambda_3)$, let $P_i = \lambda_j \cap \lambda_k$ where i, j, k are distinct. Let $\langle \lambda_i \rangle$ denote the segment of λ_i between P_j and P_k.

Lemma 2.4. *Let $\Delta(\lambda_1, \lambda_2, \lambda_3)$ be a triangle and suppose that $\langle \lambda_i \rangle$, $\langle \lambda_j \rangle$ both intersect some translate gR in R. Then the cutting sequences of $\langle \lambda_i \rangle, \langle \lambda_j \rangle$ between P_k and gR coincide. The point P_k may be moved by a finger move to a point $P_k' \in gR$.*

Proof. Without loss of generality assume $P_k \in R$. The cutting sequence of $\langle \lambda_i \rangle, \langle \lambda_j \rangle$ from P_k to gR are both shortest words representing g. Since Γ is a free group, these sequences must coincide. The intersection point may then be moved by finger moves from P_k through the intervening common regions cut by $\langle \lambda_i \rangle, \langle \lambda_j \rangle$, and into gR. □

Lemma 2.5. *Let $\Delta = \Delta(\lambda_1, \lambda_2, \lambda_3)$ be a triangle as above. Then Δ may be transposed by finger moves on λ_1, λ_2 and λ_3 which do not alter π_1-parameters.*

Proof. We claim that there is a copy gR of R such that $\langle \lambda_1 \rangle, \langle \lambda_2 \rangle, \langle \lambda_3 \rangle$ all intersect gR. For consider the cutting sequences of $\langle \lambda_1 \rangle, \langle \lambda_2 \rangle$ starting at P_3. Either these sequences coincide until we reach one or another of the points P_1, P_2 in which case we are done; or at some point the two paths differ so that there is a region gR such that $\langle \lambda_1 \rangle, \langle \lambda_2 \rangle$ both intersect gR but leave gR across distinct sides s_1, s_2. In this case, P_2, P_3 must be in the half planes bounded by s_1, s_2 and not containing gR. Since gR has no vertices in Int $\mathbb{D}$, it follows that $\langle \lambda_3 \rangle$ crosses from s_1 to s_2 in gR.

By Lemma 2.4, the points P_1, P_2, P_3 may all be moved by finger moves into gR. A further finger move inside gR can obviously be made to transpose Δ. □

The point of the next lemma is that the moves we need to make must be carried out equivariantly.

Lemma 2.6. *Let $\Delta = \Delta(\lambda_1, \lambda_2, \lambda_3)$ be a triangle all of whose sides lie in a tight simple Γ invariant family G, and suppose that Δ is minimal for G. Then the projection of Δ on M is an embedding.*

Proof. If this were not the case, there would be Γ equivalent points P, gP in the triangle Δ.

It is clearly impossible that $P \in \partial\Delta$ and $gP \in \operatorname{Int}\Delta$. Suppose first that both $P, gP \in \partial\Delta$. The points P, gP cannot lie on distinct sides of Δ, for then one of these sides would intersect a translate of the other, contradicting minimality of Δ. If P, gP lay on the same side of Δ, this would contradict the fact that each side of Δ projects to a simple curve on M, unless the arc from P to gP were the fundamental period. If this is the case, and if P, gP are vertices of Δ, with $P = \lambda_1 \cap \lambda_2$, $gP = \lambda_1 \cap \lambda_3$, then $g(\lambda_2) \cap \lambda_1 = P = \lambda_2 \cap \lambda_1$, contrary to the assumption that at most two curves in a tight family meet in a point. Thus there is a point equivalent to but distinct from a vertex of Δ lying between P and gP. Again, translation of the second side through this vertex gives a line in G which is not a side of Δ intersecting Δ. Thus we have shown that we must have $P, gP \in \operatorname{Int}\Delta$.

Now assume that $P, gP \in \operatorname{Int}\Delta$. Consider any geodesic α through P and extend it till it meets $\partial\Delta$ in Q. Let $g\alpha$ extended through gP meet $\partial\Delta$ in Q'. If $d(P, Q) < d(gP, Q')$, then a translate by g of a side of Δ passes through $gQ \in \operatorname{Int}\Delta$, contrary to hypothesis, and likewise if $d(gP, Q') < d(P, Q)$. If $d(P, Q) = d(gP, Q')$, then $Q' = gQ$ and Q, Q' are equivalent points on $\partial\Delta$. If $Q \neq Q'$ this case has been ruled out above, while $Q = Q'$ is impossible since $g \neq \mathrm{id}$ and Γ acts freely in $\mathbb{D}$. □

The proof of Theorem 2.0 will now be completed by the following lemma:

Lemma 2.7. *With the hypothesis of Theorem* 2.0 *and the notation above, let $F = \mathbb{F}(n_1 x_1, \ldots, n_k x_k)$, $\varphi_*(F) = \mathbb{F}(n_1 \varphi_*(x_1), \ldots, n_k \varphi_*(x_k))$. Then there are equivariant finger moves which alter the combinatorial pattern of $\varphi_*(F)$ to coincide with that of F.*

Proof. We proceed by induction on the number of curves we are lifting from M to form F. If there is only one curve, it is by hypothesis simple, and there is nothing to prove. Suppose the assertion holds for families consisting of lifts of m curves, and assume that F contains $m+1$ curves. Fix m of these, and let their lifts be a subfamily G of F. Apply the induction hypothesis to move the curves in $\varphi_*(G)$ by equivariant finger moves until their combinatorial pattern coincides with that of the related family G. Call this new family H. Let the lifts of the remaining curves in F be $\lambda_1, \lambda_2, \ldots$ and let their images in $\varphi_*(F)$ be $\mu_1 = \varphi_* \lambda_1$, $\mu_2 = \varphi_* \lambda_2, \ldots$.

For each i construct a curve $\bar{\mu}_i$ in $\mathbb{D}$ whose endpoints on $\partial\mathbb{D}$ are the same as those of μ_i, and whose combinatorial pattern relative to H is the same as that of λ_i relative to G. Since inside any fundamental region R for Γ we see only finitely many intersection points and thus only need move $\mu_i \cap R$ a finite distance to get $\bar{\mu}_i \cap R$,

and since all the patterns we see in F and $\varphi_* F$ are Γ-invariant, the curves μ_i and $\bar{\mu}_i$ are a bounded distance apart.

Our aim is to move μ_i to $\bar{\mu}_i$ by a finite number of equivariant finger moves which leave the combinatorial pattern of H unchanged.

The only triangles occuring in $\varphi_* F$ which are not formed by curves in H are of the form $\Delta(\gamma, \gamma', \mu)$ where $\gamma, \gamma' \in H$ and $\mu \in \{\mu_i\}$. We say that such a triangle is *wrongly ordered* if it is a transposition of the triangle $\Delta(\gamma, \gamma', \bar{\mu})$. This happens exactly when the intersection $\gamma \cap \gamma'$ lies between μ and $\bar{\mu}$. Thus if there are no wrongly ordered triangles, then μ and $\bar{\mu}$ lie in the same position relative to H and we are done. Hence we may as well assume that there is some $\mu \in \{\mu_i\}$ for which $\Delta(\gamma, \gamma', \mu)$ is wrongly ordered. We now show that there is a minimal such Δ. Let $K = \{\alpha \in H: \alpha$ intersects the region between μ and $\bar{\mu}\}$. By hypothesis $K \neq \emptyset$. Since μ and $\bar{\mu}$ have the same endpoints on $\partial \mathbb{D}$ and since $\varphi_*(F)$ is tight, each curve in K intersects $\bar{\mu}$. Apply Lemma 2.2 with $F = K$ and $l = \mu$, to find a triangle $\Delta_1 = \Delta(\gamma_1, \gamma_2, \mu)$ formed by curves $\gamma_1, \gamma_2 \in K$ with $\gamma_1 \cap \gamma_2$ lying between μ and $\bar{\mu}$ and such that $\gamma' \cap \Delta_1 = \emptyset$ for any $\gamma' \in K$, $\gamma' \neq \gamma_1, \gamma_2$. Clearly, Δ_1 is again wrongly ordered. If Δ_1 is not minimal in $\varphi_* F$, then the only curves in $\varphi_* F$ which intersect it are lifts of $\pi(\mu)$. Let $\mu_{i_1}, \mu_{i_2}, \ldots, \mu_{i_k} = \mu$ be the lifts of $\pi(\mu)$ cutting γ_1 between $P = \gamma_1 \cap \gamma_2$ and $Q = \gamma_1 \cap \mu$. Since $\pi(\mu)$ is simple, the curves $\mu_{i_1}, \ldots, \mu_{i_k}$ intersect γ_2 in the same order as γ_1, moreover each of the triangles $\Delta(\gamma_1, \gamma_2, \mu_{i_j})$ is wrongly ordered. In particular, $\Delta(\gamma_1, \gamma_2, \mu_{i_1})$ is wrongly ordered and minimal in $\varphi_* F$.

Now, by Lemma 2.5, we can reverse the order of $\Delta_1 = \Delta(\gamma_1, \gamma_2, \mu_{i_1})$ by finger moves. By Lemma 2.6, Δ_1 is embedded on M. Hence, carrying out these moves in a small neighborhood of $\pi(\Delta_1)$ on M, we may assume we do not disturb the order of other strands in $\varphi_*(G)$, and also that the changes may be extended equivariantly to all of $\mathbb{D}$.

Clearly there are at most finitely many minimal triangles, up to Γ equivalence. An induction on the number which are wrongly ordered completes the proof. □□

3. Intersection of curves and τ-orientability

The linearity theorem may fail when the condition that φ_* maps words oriented coherently with τ into words oriented coherently with τ' fails. Here is an example.

Example B. Let M be a torus with one puncture, $\Gamma = \pi_1(M) = \langle a, b \,|\, \rangle$ with fundamental domain R as pictured in Fig. 4. Choose $\varphi \in \mathrm{Aut}\,\Gamma$ so that $\varphi(a) = ab^{-4}$, $\varphi(b) = b$. (Geometrically, φ is the fourth power of a Dehn twist about the image on M of the axis of b). Let $w_1 = ab^2$, $w_2 = b$. Then w_1, w_2 are defined by weights on τ and oriented coherently with τ and $w_1 + w_2 = ab^3$. Now, $\varphi(w_1) = ab^{-2}$, $\varphi(w_2) = b$, so that $\varphi(w_1), \varphi(w_2)$ are supported on τ', also $\varphi(w_1) + \varphi(w_2) = ab^{-3}$. However, $\varphi(w_1)$ and $\varphi(w_2)$ are oppositely oriented relative to τ. We have $\varphi(w_1 + w_2) = \varphi(ab^3) = ab^{-1}$, while $\varphi(w_1) + \varphi(w_2) = ab^{-3}$, so that φ does not act linearly on the positive linear span of w_1, w_2. This example is discussed in more detail in Section 7.

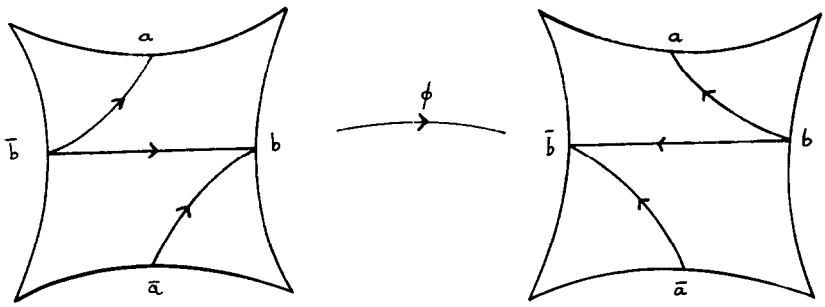


Fig. 4.

More generally, Theorem 2.0 says nothing about the case of non-orientable τ. Our next goal therefore is to extend the linearity theorem to this situation. In the light of the example just given, we need some new restrictions on the behavior of φ relative to the orientations of curves on τ and τ'. Our idea is that, while τ is non-orientable, there is nevertheless a local notion of relative orientation of two curves carried by τ near points where they intersect. This notion will take over the role which the orientability of τ played in our proof in Section 2. In this section we will set up the new ideas which we need.

We saw in Lemma 2.1 that if γ, δ are tight simple curves in $\mathbb{D}$ which are supported on τ, and if $P \in \gamma \cap \delta$ is in gR, then γ, δ meet a common side of gR. Using this common side, we may have *coherent orientations* on γ, δ near P, *relative to* τ by orienting γ, δ so that both curves point toward (or away from) s. If γ, δ cut two common sides s, s' of gR, it clearly will not matter whether we use s or s' to fix coherent orientations. More generally, if γ, δ cut sides $s_1, s_2, \ldots, s_k$ in $g_1R, g_2R, \ldots, g_kR$, and if $P \in g_iR$ is pushed by equivariant finger moves through intermediate copies to $P' \in g_jR$, it will not matter whether we orient γ, δ near P or near P'.

Suppose that $\varphi \in \mathrm{Diff}(M)$ and let γ, γ' be tight curves in $\mathbb{D}$ supported on τ, with $\gamma \cap \gamma' = \{P\}$. Since $\bar{\varphi}$ preserves order on $\partial\mathbb{D}$ and since $\varphi_*(\gamma)$ and $\varphi_*(\gamma')$, being tight, can intersect at most once, $\varphi_*(\gamma)$ and $\varphi_*(\gamma')$ intersect in exactly one point Q. Orient γ and γ' coherently at P relative to τ. The map φ_* induces orientations on $\varphi_*(\gamma)$ and $\varphi_*(\gamma')$. We say that φ_* *preserves* τ*-orientation at* P if whenever γ and γ' are oriented coherently at P relative to γ, the induced orientations on $\varphi_*(\gamma), \varphi_*(\gamma')$ are coherent near Q relative to τ'. We say that φ_* *preserves* τ*-orientation of* F if it preserves τ orientation at each intersection point $\gamma \cap \gamma'$, $\gamma, \gamma' \in F$. Notice that it is exactly this condition which fails in Example B above. A detailed discussion of how to check this condition in practice is given in Section 7.

To prove the linearity theorem we need to impose one further condition on φ_*, also relating to local orientation.

Let $\lambda_1, \lambda_2, \lambda_3$ be the curves in a tight simple family F which determine a triangle $\Delta = \Delta(\lambda_1, \lambda_2, \lambda_3)$ in $\mathbb{D}$. As in the proof of Lemma 2.5, we may find equivariant finger moves which push the three intersection points P_1, P_2, P_3 to points P_1', P_2', P_3' all con-

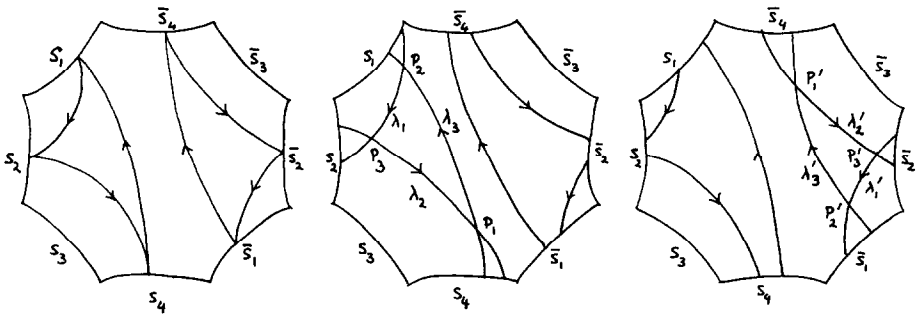


Fig. 5.

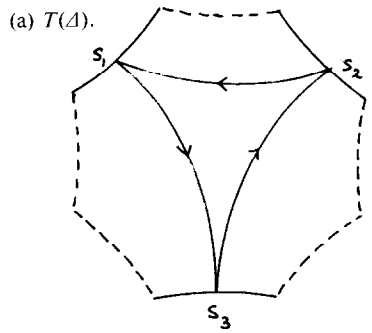
tained in the same copy gR of R and determining a triangle $\Delta' = \Delta'(\lambda_1', \lambda_2', \lambda_3') \subseteq gR$. Since all curves in F are supported on τ, the edge $\langle \lambda_i' \rangle$ of Δ' collapses onto an edge E_i of τ. If the edges E_i, $i = 1, 2, 3$, are all distinct, i.e. if they join distinct switches of τ, then they form a trigon $T(\Delta)$ on τ. We say that Δ' and Δ *lie over* $T(\Delta)$.

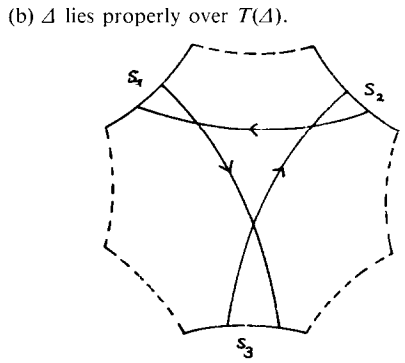
Notice that it may happen that a different set of lifts of the three curves $\pi(\lambda_i)$ on M may lie over a different trigon on τ. An example is given in Fig. 5. Here we take M to be a surface of genus two with fundamental region a symmetrical octagon of interior angle $\frac{1}{4}\pi$, with generators s_i pairing opposite sides of R. Thus $\Gamma = \langle s_1, s_2, s_3, s_4 \,|\, s_1 \bar{s}_2 s_3 \bar{s}_4 \bar{s}_1 s_2 \bar{s}_3 s_4 \rangle$. The three words $w_1 = \bar{s}_1 s_2$, $w_2 = \bar{s}_2 s_4$ and $w_3 = \bar{s}_4 s_1$ are supported on the same train track. See Fig. 5a. The lifts $\lambda_1, \lambda_2, \lambda_3$ (Fig. 5b) intersect in P_1, P_2, P_3 to form a triangle Δ. On the other hand, we can push the intersection points over into P_1', P_2', P_3' and find lifts $\lambda_1', \lambda_2', \lambda_3'$ lying over Δ' (Fig. 5c). Clearly, Δ and Δ' lie over distinct trigons on τ.

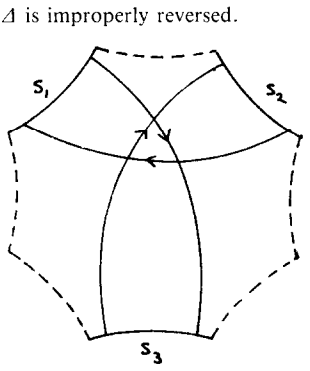
By an *orientation* on a triangle Δ we mean an orientation of $\partial\Delta$. Assume that $\Delta = \Delta(\lambda_1, \lambda_2, \lambda_3)$ lies over a trigon T. The orientation on $\partial\Delta$ induces an orientation on the three sides $\langle \lambda_i \rangle$ which in turn induces an orientation on the corresponding edges of $T(\Delta)$. The orientations of Δ and $T(\Delta)$ may or may not coincide, see Fig. 6. In the first case (Fig. 6(b)) we say that Δ lies *properly* over $T(\Delta)$, while in the latter case (Fig. 6(c)) we say that Δ is *improperly reversed.* Notice that by finger moves we can always place Δ in one copy of R and then transpose Δ without altering π_1-parameters. It will be important later to ensure that neither of the families F and $\varphi_*(F)$ in the proof of Theorem 2.0 contain improperly reversed triangles.

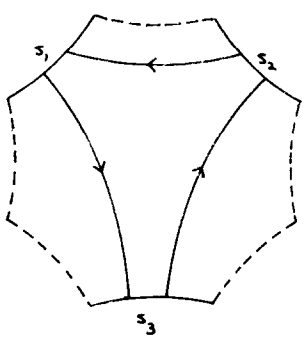
Of course not all triangles in $\mathbb{D}$ lie over trigons. This is illustrated in Fig. 7, in which all sides of the triangle formed by the intersection of the arcs labelled $bc\bar{d}\bar{d}$, $ac\bar{d}\bar{d}$ and $\bar{d}c\bar{d}\bar{c}$ (see lower figure) lie over the same edge $E(c, d)$ of τ. The following lemma shows that those triangles which lie over trigons are preserved under diffeomorphisms which preserve τ-orientation:

Lemma 3.1. *Suppose that simple tight families $F, \varphi_* F$ are supported on π_1-train tracks τ, τ' where $\varphi \in \mathrm{Diff}(M)$. Suppose that φ_* preserves τ-orientation of F. Let*









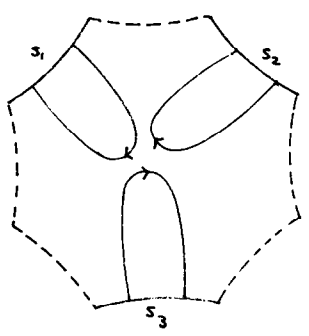


Fig. 6.

$\lambda_1, \lambda_2, \lambda_3 \in F$ form a triangle lying over a trigon on τ. Then $\varphi_*\lambda_1, \varphi_*\lambda_2, \varphi_*\lambda_3$ form a triangle lying over a trigon on τ'.

Proof. Since $\lambda_1, \lambda_2, \lambda_3$ intersect in pairs, so do the image curves $\varphi_*\lambda_1, \varphi_*\lambda_2, \varphi_*\lambda_3$ and hence the image curves form a triangle. Since $\lambda_1, \lambda_2, \lambda_3$ lie over a trigon, we may orient them so that the τ-orientations of each pair are incoherent at each intersection point. Since φ_* preserves τ-orientation of F, the same will be true at each

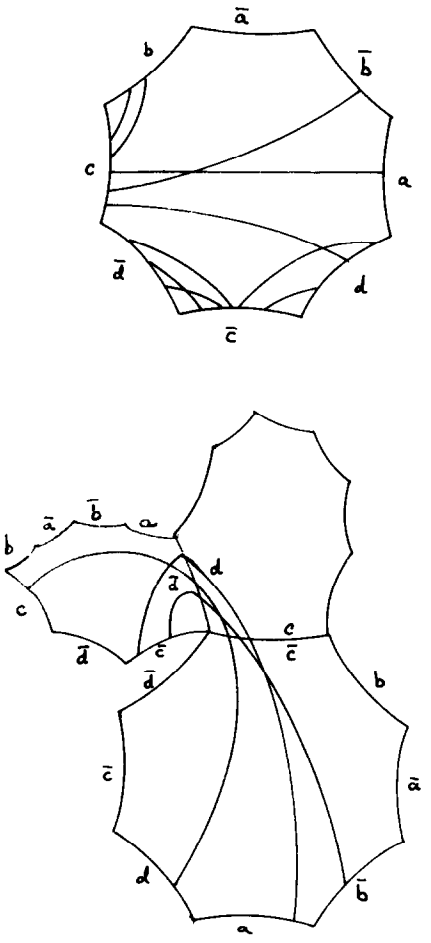


Fig. 7.

intersection of the image pairs. But this means that the image pairs lie over a trigon. □

Definition 3.2. Let $F, \varphi_* F$ be tight families supported on π_1-train tracks τ, τ', where $\varphi \in \mathrm{Diff}(M)$, and suppose φ_* preserves τ-orientation of F. Then φ_* *preserves orientation of F-trigons* if whenever $\lambda_1, \lambda_2, \lambda_3 \in F$ form a triangle lying properly over a trigon T in τ, and $\varphi_*\lambda_1, \varphi_*\lambda_2, \varphi_*\lambda_3$ form a triangle lying properly over a trigon T' in τ', then the orientation on $\lambda_1, \lambda_2, \lambda_3$ which comes from an anticlockwise orientation on ∂T induces an anticlockwise orientation on $\partial T'$.

Remark 3.3. The somewhat convoluted definition above seems to be necessary because it is not clear how to ascribe meaning directly to the image of a trigon on

τ unless it is covered by curves forming a triangle in F. Further, it is not clear that if two different sets of curves cover the same trigon, that the 'image trigons' are the same. We discuss how to check condition 3.2 in Section 7.

4. The Linearity Theorem, Part II (τ arbitrary, Γ free)

In this section we prove the linearity theorem for general (non-orientable) τ, still keeping the assumption that Γ is free.

Theorem 4.0. *Let τ, τ' be π_1-train tracks on M. Assume that $\Gamma = \pi_1 M$ is free. Suppose $\varphi \in \mathrm{Diff}^+(M)$. Suppose that for $i = 1, \ldots, k$, $x_i \in \Omega(\tau)$ and $\varphi_*(x_i) \in \Omega(\tau')$. Let F be the family of lifts of curves corresponding to x_i, and suppose that φ_* preserves τ-orientation of F and orientation of F-trigons. Then φ_* acts linearly on* $\mathrm{Sp}^+\{x_1, \ldots, x_k\}$.

Proof. The added features we have to deal with are the notion of surgery for curves supported on non-orientable train tracks and the existence of trigons. First, we deal with surgery.

Our definition of surgery in Section 2 depended on the orientation of τ; namely, at each intersection point of curves in F we moved onto the intersecting curve in the direction of the orientation of τ. We now replace this with motion consistent with relative τ-orientation at each intersection point; that is, we move from a curve onto the intersecting curve in such a way that this motion along the two curves is consistent with coherent τ-orientation at the intersection point. We again denote the surgery of a family F supported on τ by $S_\tau(F)$.

Suppose that F contains curves which intersect in an improperly reversed triangle. It is easy to see that surgery on such a triangle does not even give tight curves, so that Lemma 2.3 fails, see Fig. 6(c), (e). However, this is the only problem.

Lemma 4.1. *Let F be a tight simple family supported on a π_1-train track τ, containing no improperly reversed triangles. Then $S_\tau(F)$ is supported on τ and*

$$x_{S_\tau}(F) = \sum_{\gamma \in F} x_\gamma .$$

Proof. We may apply the proof of Lemma 2.3 noting that the hypothesis that τ contains no trigons may be replaced by the fact that F contains no improperly reversed triangles. □

The following lemma shows that improperly reversed triangles may always be removed:

Lemma 4.2. *Let F be a tight simple family supported on a π_1-train track τ. Then there exists a family F' with the same π_1-parameters as F containing no improperly reversed triangles.*

Proof. The idea is to push the ends of all strands of F close to the midpoints of sides of F and use the fact that triangles with their vertices close to these midpoints cannot be improperly reversed. (This follows by continuity, since a triangle with vertices at these midpoints is certainly not improperly reversed.)

To do this we construct an isotopy $f: M \to M$ as follows: f fixes vertices and midpoints of sides of ∂R but pushes any other point on ∂R towards the midpoint of the side containing it; f extends smoothly to the rest of M. We say such an f is an (ε, δ)-*contraction* if whenever $x \in s$, $d(x, \partial s) > \varepsilon$, then $d(f(x), P(s)) < \delta$ for all sides s of R.

Let F be a tight family on τ. Replace the family $f(F)$ by the family F' in which each segment of a curve in $f(F)$ which joins sides s, s' of R is replaced by the geodesic with the same endpoints on ∂R. Clearly F' is a tight family with the same π_1-parameters as F. Choosing ε so small that no curves in F lie within ε of a vertex of R, and choosing δ so small that curves whose endpoints lie within δ of the midpoints of sides do not form improperly reversed triangles, we find a family F' as required. □

Thus from now on we may assume that both families $F, \varphi_* (F)$ are replaced by families with the same π_1-parameters but containing no improperly reversed triangles. The proof of Lemma 2.3 now carries over, provided we note that the sequence of segments of $F, \varphi_* F$ to be followed in doing the surgeries $S_\tau(F)$ and $S_{\tau'}(\varphi_* F)$ are now determined not by the orientations of τ, τ', but by relative τ-orientations at intersection points, which by hypothesis are preserved by φ_*.

Following the scheme of proof in Section 2, to complete the proof of Theorem 4.0 we must simply show that the combinatorial pattern of $\varphi_* F$ may be altered by finger moves until it coincides with that of F. We must ensure that our moves do not reintroduce improperly reversed triangles into $\varphi_* F$.

Let Δ be a triangle in $\varphi_* F$. Just as in Section 2 we may apply finger moves to the vertices of Δ until they all lie in a common copy of R. The three sides of Δ either all lie over the same branch of τ, or Δ lies over a trigon T. In the first case, Δ may be transposed if necessary without altering π_1-parameters exactly as in Lemma 2.5. In the second, by hypothesis Δ is not improperly reversed and hence has the same orientation as the trigon T. Moreover, the corresponding triangle Δ' in F lies over a trigon T' by Lemma 3.1, and by hypothesis is also not improperly reversed. By the hypothesis that φ_* preserves orientation of τ trigons, the orientations of T and T' correspond. Hence the orientations of Δ and Δ' agree and so Δ cannot be, in the terminology of Lemma 2.7, a wrongly ordered triangle. The same method as in Lemma 2.7 now completes the proof of Theorem 4.0. □

5. The closed surface

In this section we introduce the extra technicalities needed to deal with the case in which M is a closed surface, so that Γ is no longer free. The actual proof of the linearity theorem in this situation will be completed in Section 6.

As mentioned in the introduction, we need to make heavy use of the results in [1] in which we proved that under appropriate hypotheses on R and Γ_R, the cutting sequences of geodesics are shortest words. We also make use of the restrictions, also investigated in detail in [1], imposed on a shortest word by the requirement that it be simple. For convenience, we will summarize all the definitions and results we need from [1] here.

Choose a hyperbolic metric on M and a realisation of M as $\mathbb{D}/\Gamma$ for some Fuchsian group Γ. Let R be a fundamental domain for Γ acting in $\mathbb{D}$, and let N be the set of images of ∂R under Γ. *We assume that N is a union of complete geodesics in* $\mathbb{D}$, and further that R has at least five sides. In the language of [1], R has *even corners*. Notice that the tiling of $\mathbb{D}$ by symmetric $4g$-gons of interior angle $\pi/2g$ gives such an N for the closed surface of genus g. Our results will apply to any choice of R, Γ with the above property, whether or not $\mathbb{D}/\Gamma$ is a closed surface.

As before, we choose as generators of Γ hyperbolic isometries which pair corresponding sides of R. The convention for labelling sides of R and of its Γ-translates is as in Section 1.

We assume, as in [1], that the generating set Γ_R is *alternating*, i.e. there is a map $\varrho : \Gamma_R \to \{+1, -1\}$ such that $\varrho(x) = \varrho(x^{-1})$ and so that $\varrho(x) = -\varrho(x')$ wherever x, x' are adjacent labels on ∂R. This condition is satisfied for standard choices of R, Γ_R. For details, consult [1]. We need this hypothesis in order to avoid certain complications in the solution to the conjugacy problem in Γ.

Cutting sequences are defined as before, with a qualification. Let $V(N)$ be the set of Γ-translates of vertices of R in Int $\mathbb{D}$. Let $\hat{\gamma}$ be a curve on M, and let γ be a lift of $\hat{\gamma}$ to D.

If for some $v \in V(N)$, $v \in \gamma$, but γ is not coincident with a side of N, then we deform γ in a neighborhood of v to pass round one or another side [1, Fig. 1a]. Notice that the two possible cutting sequences we obtain correspond to the two complimentary halves of the relation in Γ corresponding to v. If γ coincides with a net edge, then we move it slightly into a homotopic curve which runs to one side of and roughly parallel to the original curve [1, Fig. 1b]. Moving to the other side would give a conjugate word in Γ. With this modification, the cutting sequence $\sigma(\gamma)$ is always well defined, although for curves which pass through vertices of N they are non-unique. If γ is a geodesic, we call the primitive period of $\sigma(\gamma)$ the *geodesic word* of γ.

The main result of [1] is that, if R has even corners, then geodesic words are cyclically shortest in the word metric of Γ, Γ_R. In other words, geodesics are tight in the sense of Section 1. In [1] we characterized all shortest words, and showed that

if in addition Γ_R is alternating, then a cyclically shortest word is shortest in its conjugacy class. To state all this precisely we need some further definitions.

Orient ∂R anticlockwise. Suppose that $e \in \Gamma_R$ is such that the side of R with exterior label e has initial vertex $v \in \text{Int } \mathbb{D}$. Let $\psi(e) = \psi_w(e)$ be the label of the next side of N in clockwise order round v. If $2n(v)$ sides of N meet at v, then $e \cdot \psi(e) \dots \psi^{2n(v)-1}(e) = \mathscr{R}_v(e)$ is a relation in Γ. (Notice that $\mathscr{R}_v(e)$ is not the same as the order of labels around ∂R; thus in the group of Fig. 1, $\mathscr{R}_v(e) = ab\bar{a}\bar{b}cd\bar{c}\bar{d}$ while the order round ∂R is $a\bar{b}\bar{a}bcd\bar{c}\bar{d}$.) We call any subword of $\mathscr{R}_v(e)$ or $\mathscr{R}_v(e)^{-1}$ of length at least 2 a *cycle*, and a cycle of length $n(v)$ we call a *half cycle*. A cycle of length $n(v)+1$ we call *long* and a cycle of length $n(v)-1$ we call *short*. If u is a cycle and if $uw = \mathscr{R}_v^{\pm 1}$, then $\bar{w}$ is the *complement* of u. A *half-cycle switch* in a word w replaces a half cycle by its complement. Obviously this operation preserves word length. More generally, a *cycle switch* replaces a cycle by its complement. If $u = e \cdot \psi(e) \dots \psi^{n(v)-2}(e)$ is a short cycle, then we call the short cycle $w = \psi^{-2}(e)\psi^{-3}(e) \dots \psi^{-n(v)}(e)$ its *opposite*. Likewise $\bar{w}$ is the opposite of $\bar{u}$.

Let v, w be vertices of R adjacent in anticlockwise order round ∂R. The clockwise cycles $\dots e\psi_v(e), f\psi_w(f) \dots$ at v, w are *consecutive* if $\psi_v^2(e) = \psi_w^{-1}(f)$. We make a similar definition for anticlockwise cycles. A sequence of consecutive cycles $C_1 C_2 \dots C_k$ we call a *chain*. If C_1, C_k are half cycles and $C_2, \dots, C_{k-1}$ are short cycles, then the chain is *long*. Note that the vertices corresponding to such a sequence of cycles occur along one side of N. In this situation we may successively switch the cycles $C_1, C_2, \dots, C_k$ to obtain a chain $C_1' \dots C_k'$ of short cycles in which $C_2', \dots, C_{k-1}'$ are the opposites of $C_2, \dots, C_{k-1}$. This new chain obviously has shorter length; in particular, a word containing a long chain cannot be shortest.

If $C_1 \dots C_k$, $C_1' \dots C_k'$ are two chains which are equal as elements of Γ, then we say that $C_1 \dots C_k$, $C_1' \dots C_k'$ are *complementary*. It is shown in [1] that in this situation for each i, C_i and C_i' are cycles at the same vertex v_i, and that one chain may obtained from the other by a succession of cycle switches at the vertices v_i. We call such a replacement process a *chain switch*.

The following summarises the main results of [1]:

Theorem 5.1 (Birman and Series [1, Theorem 2.12]). *Let R be a fundamental region for Γ with at least 5 sides*[1] *and even corners, and suppose that the associated generating set Γ_R is alternating. Then*

(i) *Two shortest words representing the same element of Γ differ only in a number of disjoint blocks, which are complementary halves of chains.*

(ii) *A cyclic word in Γ_R is cyclically shortest if and only if it contains no long cycles or long chains.*

(iii) *Geodesic words are cyclically shortest words.*

(iv) *If w, z are cyclically shortest conjugate words, then they have the same length. Either they agree up to half-cycle switches and cyclic permutations, or $z = C_1 \dots C_k$,*

[1] This condition is not quite the most general. See [1] for details.

$w = C_1' \dots C_k'$ where each C_i is a short cycle and C_i' is the opposite of C_i for $1 \le i \le k$, and $eze^{-1} = w$ for some $e \in \Gamma_R$ where $eC_1, C_k'e$ are half cycles so that $eC_1 \dots C_k$ and $C_1' \dots C_k'e$ are complementary chains. In this last case w, z are the geodesic words associated to a side C of the net N. □

We now turn to the additional condition we need to impose on train tracks to ensure that multiple simple words which are defined by a set of weights on the track are cyclically shortest. We form weighted graphs of multiple cyclic words and multiple loops in $S(M)$ exactly as before. The switch conditions (Condition 1.1) and boundary conditions (Restriction 1.2) on weights and train tracks remain in force.

A cyclic word w is said to contain a *generalised cycle* of length $k > 0$ if there is a cycle $= e\psi(e) \dots \psi^k(e)$ such that for each $j = 0, \dots, k-1$ the two letter sequence $\psi^j(e)\psi^{j+1}(e)$ or its inverse occurs in w. Likewise w contains a *generalised chain* if all two letter sequences which occur in some chain appear, possibly inverted, in w. Notice that the 2-letter sequences in question in w need not be coherently oriented, i.e. some may be oriented as in C, others as in C^{-1}.

We call an edge of a train track τ on R a *corner branch* if it joins the points $P(\bar{e}), P(\varphi_v(e))$, corresponding to a cycle of length two. We say that τ supports a generalised cycle of length k if it contains the $k-1$ corner branches corresponding to the $k-1$ adjacent pairs in some generalised cycle. Likewise τ supports a generalised chain if it contains branches corresponding to each of the adjacent pairs of sides in a generalised chain. We call branches which join the end of one cycle to the beginning of the next, *linking branches*.

Our *cycle and chain restriction* is:

Condition 5.2. τ supports no generalised long cycles or long chains.

Remark. Checking Condition 5.2 for generalised long chains is really only a finite check. Since the sequence of vertices occurring along a side of N is necessarily periodic, so is the corresponding sequence of consecutive cycles. Thus we only have to fill in all the corner branches of τ at each of these vertices, together with all possible linking branches, to make the check.

Any train track on R satisfying Condition 1.1, Restriction 1.2 and Condition 5.2 is called a *π_1-train track*. It is clear that any word supported on such a track is shortest. For any non-shortest word must, by Theorem 5.1(ii), contain either a long cycle or a long chain, which would obviously violate Condition 5.2. Notice that, in contrast to the Thurston school, we do *not* exclude the possibility of bigons on our train tracks. In fact bigons occur frequently, corresponding to cycle switches at the vertices $V(N)$.

It is somewhat more subtle to see that *any* simple words can be supported on a train track satisfying Condition 5.2. To see this, we need to make use of the extra restrictions which apply to shortest words which are in addition simple. These are discussed in detail in [1, Section 3].

Recall that a word is *simple* if it represents an element of $S(M)$. We shall say that a simple word w is *strongly simple* if it is cyclically shortest and if tight curves in $\mathbb{D}$ with cutting sequence $\ldots www\ldots$ project to simple curves on M. This last condition is vacuous if either $V(N)=\emptyset$ (implying Γ free) or if $\ldots ww\ldots$ contains no half cycles. If w is geodesic and simple, then it is obviously strongly simple, however the converse is not true. A non-geodesic word obtained from a simple geodesic word by cycle or chain switching may give rise to a curve in $\mathbb{D}$ which does not project to a simple loop.

The following result generalises [1, Theorem 3.1]. The statement in [1] refers to simple words; in our situation it will be enough to restrict to strongly simple w.

Theorem 5.3. *Let w be a strongly simple word in Γ_R. Then w contains no generalised long cycles or long chains.*

Proof. Let $A(w)$ denote a collection of arcs joining pairs of sides of R, where $x_w(e,f)$ arcs join the side containing $P(e)$ to the side containing $P(f)$, $e,f\in\Gamma_R$. Since w is strongly simple, the curve obtained by linking up the arcs in $A(w)$ according to the glueing pattern of R has cutting sequence w.

Suppose that w contains a generalised long chain made up of adjacent pairs of letters in a sequence of consecutive cycles $C_1\ldots C_k$ at vertices $v_1,\ldots,v_k$. We may assume that $k>1$, for otherwise we have the case of a long cycle already dealt with in [1]. Choose k to be minimal among generalised long chains. Let e_i, f_i be the final and initial letters in C_i, C_{i+1} respectively, for $1\le i\le k-1$.

Consider the arc in $A(w)$ which meets the side of R containing $P(\bar{e}_i)$, $1\le i\le k-1$, which is closest to the vertex v_i. We claim that this arc is parallel to the linking branch of τ joining $P(\bar{e}_i)$ to $P(f_i)$. For if not, the arc closest to v_i runs parallel to the corner edge from $P(\bar{e}_i)$ to $P(\psi(e_i))$. But then $C_1C_2\ldots C_i\psi(e_i)^{-1}$ is a shorter generalised long chain in w, contrary to choice of k. Likewise the arc β_i which meets the side labelled $P(f_i)$ closest to v_{i+1} is parallel to the linking branch from $P(\bar{e}_i)$ to $P(f_i)$; and in fact we must have $\alpha_i=\beta_i$. Now consider the arcs α_i, $1\le i\le k-1$, together with the arcs parallel to the corner edges in each of the cycles $C_1,\ldots,C_k$ which are closest to the vertices $v_1,\ldots,v_k$. When R is glued up to form M, all these arcs link forming a long chain in w. But this means w was not shortest, contrary to hypothesis. □

We can now summarize our results in a generalization of Theorem 1.3. Once again, we leave the proof to the reader.

Theorem 5.4. (i) *Let τ be a π_1-train track and let $x\in\Omega(\tau)$. Then $W(x)$ is a multiple strongly simple word.*

(ii) *Let Λ be a multiple simple loop whose cutting sequence is tight. Then $\tau(\Lambda)$ is a π_1-train track and $x_\Lambda\in\Omega(\tau)$.*

(iii) *Let W be a strongly simple multiple word containing no components which*

surround a puncture or which are parallel to ∂M. Then the graph $G(W)$ is a π_1-train track τ and $x_W \in \Omega(\tau)$. □

Suppose that Λ and Λ' are multiple simple loops consisting of tight curves which represent the same element of $S(M)$. Component by component, Λ and Λ' are freely homotopic, and thus the corresponding words are conjugate in Γ. If $\partial M \neq \emptyset$, so that Γ is free, this means that $w_\Lambda, w_{\Lambda'}$ differ only by cyclic permutation, so that $\tau(\Lambda) = \tau(\Lambda')$ and $x_\Lambda = x_{\Lambda'}$. For closed surface groups, Theorem 5.1(iv) guarantees that any component of Λ which is not homotopic to the projection of a side of N differs from the corresponding component of Λ' only by cyclic permutations and cycle-switching. It can then happen that $\tau(\Lambda) \neq \tau(\Lambda')$ (see Fig. 8). This possibility motivates our next definition: strongly simple cyclic words W, W' differ by a *chain switch on* τ if W, W' differ by a chain switch and if both x_W and $x_{W'} \in \Omega(\tau)$. Such words are said to be τ-equivalent and we regard their π_1-parameters as the same. There is a subtle point here: while a chain switch may always be achieved by a sequence of cycle switches, a chain switch on τ may not be realizable as a sequence of cycle switches on τ. For this reason we regard chain-switching on τ, not cycle-switching, as the basic operation.

We shall slightly strengthen the definition of a tight simple family by requiring that all curves in it have strongly simple cutting sequences. We also require that all the curves are deformed so that they do not contain vertices in $V(N)$ and are not coincident with sides of N.

Suppose that $\varphi \in \mathrm{Diff}(M)$ and that $\Lambda \in S(M)$. We write $\varphi_*(x)$ for any weight on any π_1-train track τ' which represents $\varphi(\Lambda)$. In general there may be several different choices of $\varphi_*(x)$, and they may not even all lie on the same π_1-track. In what follows we assume that a definite choice of $\varphi_*(x)$ is made which satisfies all the stated conditions.

If γ is a tight curve in $\mathbb{D}$ whose cutting sequence is $W(x)$, $x \in \Omega(\tau)$, then by $\varphi_*(\gamma)$ we mean any curve in $\mathbb{D}$ whose edge path is $W(\varphi_*(x))$, $\varphi_*(x) \in \Omega(\tau')$, and whose endpoints on $\partial\mathbb{D}$ are $\bar{\varphi}(\partial\gamma)$, where $\partial\gamma$ are the endpoints of γ on $\partial\mathbb{D}$. This makes sense because chain-switching does not change the endpoints at infinity of a curve.

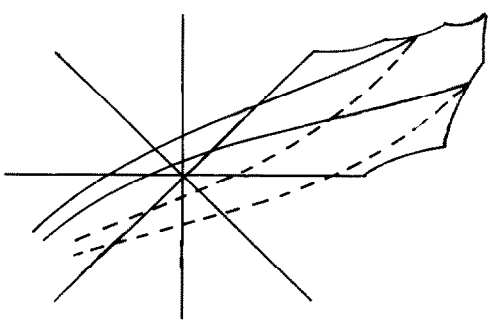


Fig. 8.

6. The Linearity Theorem (general case)

In this section we show how to modify the proof of Theorem 4.0 to include the case when M is a closed surface. We assume throughout that R and Γ_R are chosen as in Section 5 and that our train tracks satisfy Restriction 1.2 and Condition 5.2.

The most general statement of the linearity theorem is then as follows:

Theorem 6.0. *Let τ, τ' be π_1-train tracks on M. Let $\varphi \in \mathrm{Diff}(M)$ and let $x_i \in \Omega(\tau)$, $\varphi_*(x_i) \in \Omega(\tau')$ for $i = 1, \ldots, k$. Let F be the family of lifts of curves corresponding to x_i, and suppose that φ preserves τ-orientation of F. Then if either* (i) $\varphi \in \mathrm{Diff}^+(M)$ *and preserves orientation of F-trigons or* (ii) $\varphi \in \mathrm{Diff}^-(M)$ *and reverses orientation of F-trigons, then φ_* acts linearly on* $\mathrm{Sp}^+\{x_1, \ldots, x_k\}$.

Remark. The extension of our result to orientation reversing maps φ is trivial. The proofs are modified in that we now have to see that the combinatorial patterns of F and $\varphi_*(F)$ coincide after suitable moves. Triangles which do not lie over trigons can be transposed at need as before and our hypothesis implies that the orientation of triangles over trigons behaves correctly. Notice that the condition that φ preserves τ-orientation of F depends only on *relative* orientations of pairs of curves, and is thus unchanged.

We now embark on the admittedly tedious task of modifying the work in Sections 2–4 to cover the closed surface case. We have to take into account the existence of bigons on τ, and trigons which are not contained in one fundamental region, see Fig. 9.

We begin by collecting some results about n-gons on τ for $n = 0, 1, 2, 3$.

Lemma 6.1. *Let τ be a π_1-train track. Then*

(i) *τ contains no nullgons ($n = 0$), or monogons ($n = 1$).*

(ii) *Each bigon ($n = 2$) on τ represents the complementary halves of a chain.*

(iii) *If B is a bigon on τ, then there are no trigons in the closed region enclosed by the sides of B.*

Proof. (i) A nullgon or monogon on τ would represent a non-trivial loop or arc which passed twice through the same copy of R, or once, with both of its ends on the same switch. Its cutting sequence would then represent the identity. However, this is impossible because any path on τ is a shortest word in Γ.

(ii) A bigon on τ represents two curves with the same initial and final points whose cutting sequences differ. Since both cutting sequences are by hypothesis shortest, they differ, according to Theorem 5.1, by a chain switch on τ.

(iii) The bigon B is associated to a chain $C_1 C_2 \ldots C_r$ and its complement $C_1' C_2' \ldots C_r'$, where each C_i and each C_i' is a cycle about a vertex v_i in $V(N)$. Adjacent vertices v_i, v_{i+1} are joined by an edge of N, and this edge might contain an interior switch p_i on τ. The points $p_1, \ldots, p_{r-1}$ are the only possible switches of τ

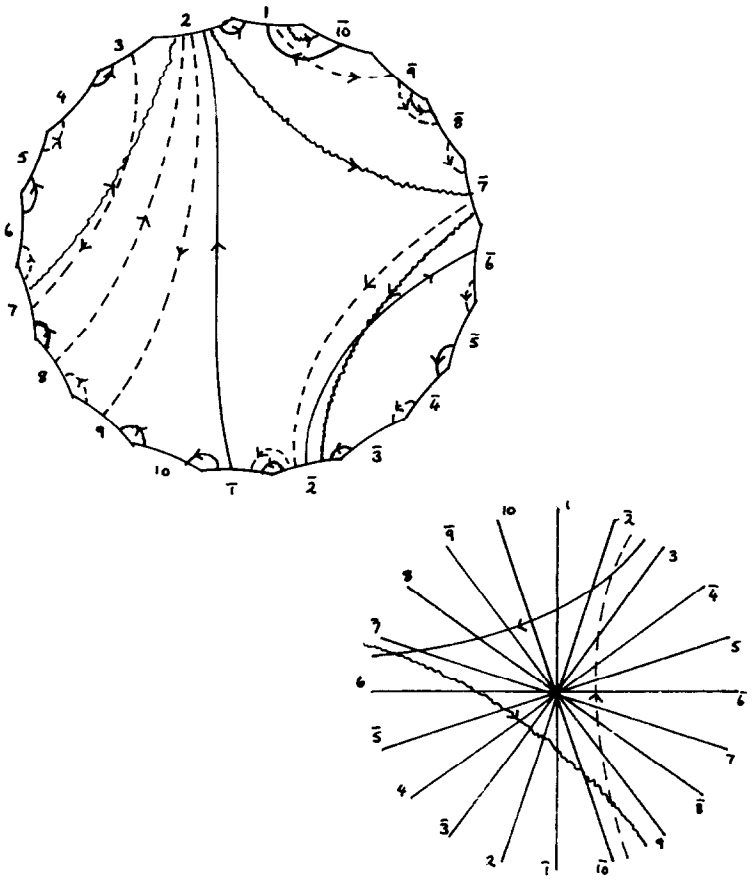


Fig. 9.

interior to B, hence the only possible edges of τ interior to B are the four corner edges at each p_i. Label these $\alpha_i, \beta_i, \gamma_i, \delta_i$, in cyclic order about p_i, where α_i, β_i are corner edges in the cycle about v_i and γ_i, δ_i are corner edges in the cycle about v_{i+1} (see Fig. 10). The switch conditions on τ imply that α_i (resp. β_i) occurs if and only if γ_i (resp. δ_i) occurs. We obtain a trigon on τ if and only if both β_i and γ_i, or both α_i and δ_i occur for some i. However, if both α_i and β_i occur (or if both γ_i and δ_i occur), then τ will contain a generalized long cycle about v_i (or v_{i+1}). □

We now examine the results of Section 2. Everything up to and including the proof of the special case of Theorem 2.0 goes through without change. The only additional feature that we need to introduce is that the moves needed to change the combinatorial pattern of $\varphi_*(F)$ into that of F must now include not only finger moves but also chain switches on τ. Such moves do not (by definition) change π_1-parameters. The following lemmas extend Lemmas 2.4 and 2.5.

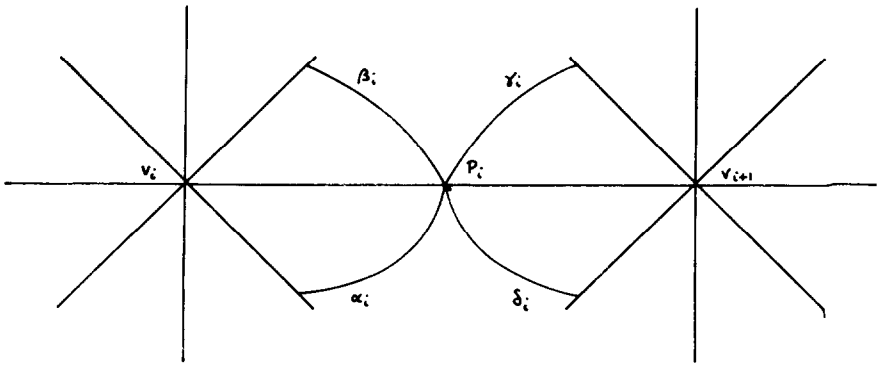


Fig. 10.

Lemma 6.2 (cf. Lemma 2.4). *Let $\Delta = \Delta(\lambda_1, \lambda_2, \lambda_3)$ be a triangle formed by curves in a tight simple family supported on a π_1-train track τ. Suppose that $\langle \lambda_i \rangle$, $\langle \lambda_j \rangle$ both intersect some translate gR of R. Then the cutting sequences of $\langle \lambda_i \rangle$, $\langle \lambda_j \rangle$ between P_k and gR either coincide or differ by a chain switch on τ. This chain switch may be carried out without changing the orientation of Δ.*

Proof. The proof is the same as for Lemma 2.4, except that we use Theorem 5.1(i) to see that the cutting sequences of $\langle \lambda_i \rangle$, $\langle \lambda_j \rangle$ from P_k to gR differ, if at all, by chain switches. These switches lie on τ because both λ_i and λ_j do. The last statement is clear. □

Lemma 6.3 (cf. Lemma 2.5). *Let $\Delta = \Delta(\lambda_1, \lambda_2, \lambda_3)$ be a triangle formed by curves in a tight simple family supported on a π_1-train track and suppose that $\Delta \cap V(N) = \emptyset$. Then Δ may be transposed by finger moves on λ_i.*

Proof. As in the proof of Lemma 2.5, we consider the cutting sequences of $\langle \lambda_1 \rangle$, starting from P_3. Either all three sides intersect a common region gR, or the points P_1, P_2 lie in half planes bounded by sides s_1, s_2 of gR and away from gR. If $\langle \lambda_3 \rangle \cap gR = \emptyset$, then Δ contains those vertices of $\partial(gR)$ which lie between s_1 and s_2, contrary to hypothesis.

We note also that, since $\Delta \cap V(N) = \emptyset$, the cutting sequences of $\langle \lambda_1 \rangle$ and $\langle \lambda_2 \rangle$ from gR to P_3 must coincide. The remainder of the proof is as in Lemma 2.5. □

It follows from Lemma 6.3 that triangles with $\Delta \cap V(N) = \emptyset$ may be treated exactly as before. When $\Delta \cap V(N) \neq \emptyset$ we have to look a little more closely.

Assume again that $\Delta = \Delta(\lambda_1, \lambda_2, \lambda_3)$ is a triangle formed by curves in a tight simple family. Working on the lines of Lemma 2.4, we push the intersection points P_i of the λ_i by finger moves (but *not* cycle switches) until each P_i lies in a region gR in which $\langle \lambda_j \rangle$, $\langle \lambda_k \rangle$ leave gR across distinct edges s_j, s_k (here $\langle \lambda_j \rangle$, $\langle \lambda_k \rangle$ are oriented

pointing away from P_i). Thus in gR, $\langle\lambda_j\rangle$ and $\langle\lambda_k\rangle$ lie over distinct branches of τ which meet at a switch Q_i on ∂R. The three branches of τ defined by the edges $\langle\lambda_i\rangle$ and running between the Q_i enclose, since $\Delta \cap V(N) \neq \emptyset$, a non-empty simply connected region in $\mathbb{D}$. Using Lemma 6.1 we see immediately that this region must be either a bigon or a trigon. (In the case of orientable τ, the second case is of course excluded.)

We say that Δ *lies over* a bigon or a trigon, depending on which case we are in.

Lemma 6.4 (cf. Lemma 2.5). *Let $\Delta = \Delta(\lambda_1, \lambda_2, \lambda_3)$ be a triangle formed by curves in a tight simple family F supported on a π_1-train track τ, and suppose that Δ lies over a bigon on τ. Then Δ may be transposed by finger moves and chain switches on τ.*

Proof. We shall show that Δ may be replaced by a triangle containing no vertices in $V(N)$. The proof is then completed using Lemma 6.3 above.

Let $\langle\lambda_3\rangle$ be the side of Δ so that the vertices of the bigon covered by Δ lie at the ends of $\langle\lambda_3\rangle$. The other side of the bigon is formed by $\langle\lambda_1\rangle$ followed by $\langle\lambda_2\rangle$, say, denote this $\langle\lambda_1\rangle\langle\lambda_2\rangle$. Since the cutting sequences of $\langle\lambda_3\rangle$ and $\langle\lambda_1\rangle\langle\lambda_2\rangle$ are both shortest and have the same endpoints, they can differ only by a chain switch. Thus $\langle\lambda_3\rangle$ may be chain switched on τ to a path whose cutting sequence coincides with that of $\langle\lambda_1\rangle\langle\lambda_2\rangle$; in particular the new triangle we obtain in this way contains no vertices of $V(N)$. □

The remainder of Section 2 now goes through as before. Notice that a triangle in F either lies over a single branch of τ, in which case Lemma 6.3 applies, or over a bigon, which is covered in Lemma 6.4. Nullgons and monogons do not occur in virtue of Lemma 6.1(i), while trigons do not occur because τ is assumed orientable.

We now arrive at Section 3. We must first check that the concept of coherent orientation relative to τ is invariant under chain switching on τ.

Lemma 6.5. *Let γ, δ be curves in a tight simple family which are supported on a train track τ. Suppose that $P = \gamma \cap \delta$, and that γ, δ have been oriented coherently near P. Let δ' be obtained from δ by a chain switch on τ, and let $P' = \gamma \cap \delta'$. Let δ' have the orientation induced by δ. Then γ, δ' are oriented coherently near P'.*

Proof. Let C be the segment of δ which corresponds to the chain which is to be switched, and let C' be the complementary chain. Then $\delta' = \delta - C + C'$. If $P = \gamma \cap \delta$ does not lie on C, there is nothing to prove, so we assume $P \in C$. Since δ, δ' have the same endpoints at infinity, it is clear that δ' also intersects γ, say at $Q \in C'$. Now if the orientations of γ, δ' are not coherent near Q, then the path on τ from the point R where δ, δ' diverge along δ to P, then along γ to Q, then back along δ' to R, is a monogon on τ, which is impossible. □

We now need to verify that the notion of orientation of trigons is invariant under chain switching on τ.

Lemma 6.6. *Suppose that curves λ_i, $i=1,2,3$ in a tight simple family form a triangle Δ lying properly over a trigon T on a π_1-train track τ. Let λ_1' be obtained from λ_1 by a chain switch on τ. Then the curves $\lambda_1, \lambda_2, \lambda_3$ lie over a trigon T', and the orientations of ∂T and $\partial T'$ given by an orientation of λ_2, λ_3 and a common orientation of λ_1 and λ_1' coincide.*

Proof. Orient the λ_i so that $\partial\Delta$ and hence ∂T are oriented anticlockwise. The sides λ_i are oriented incoherently at each intersection P_i. By Lemma 6.5 the same is true at the intersection points of λ_1', λ_2 and λ_3. This means that the curves $\lambda_1', \lambda_2, \lambda_3$ also lie over a trigon T'.

Let B be the bigon on τ bounded by the branches of τ corresponding to the two complementary halves of the chain in λ_1 and λ_1'. The trigons T and T' differ only by B, that is, $T-T'\cup T'-T\subset B$. By Lemma 6.1(iii), at most two vertices of T lie in B. Let P be the third vertex. Then P lies on the same side of the two curves λ_1 and λ_1' in $\mathbb{D}$, further, P must also be a vertex of T'. From this it follows easily that the orientations of ∂T and $\partial T'$ are the same. □

The proof of Lemma 3.1, and Definition 3.2, now make sense and carry through unchanged. Notice that if $\Delta\cap V(N)\neq\emptyset$, then Δ cannot be improperly reversed. For let $v\in V(N)\cap\Delta$. Then v lies to the same side of each side λ_i of Δ and the corresponding branch of τ, so that the orientations of Δ and the trigon it covers must be the same.

Finally, in the proof of Theorem 4.0, at the end of Section 4, we must add the possibility of triangles containing bigons. These are dealt with by Lemma 6.4. The proof of Theorem 6.0 is complete. □

7. Detecting intersections

The main result of this paper, the general case of the linearity theorem (Theorem 6.0) has been proved. Examples are in order, but confront us with an immediate problem: how can one check in specific cases whether the conditions imposed on φ_* in Theorem 6.0 are satisfied? The goal of this section is to show how this can be done. Our approach as before, is to concentrate on the case when Γ is free or when, if Γ is not free, the words we are considering contain no half cycles. The exceptional case is treated either in comments in square brackets [ ] or deferred to the end of the section. All the words we discuss in this section should be assumed to be strongly simple.

Detecting intersections between curves

Our starting point is Lemma 2.1, which implies that if γ, γ' are tight simple curves which intersect in $\mathbb{D}$ and are both supported on the same π_1-train track, then the cyclic words $\sigma(\gamma)$, $\sigma(\gamma')$ when appropriately oriented have a common block B. However, cyclic words may have a common block even though the corresponding words do not intersect. To determine whether or not a common block in a pair of words corresponds to an intersection, we proceed as follows.

Let w, w' be cyclically shortest words with a common letter $u_0^{\pm 1}$ and assume w, w' are oriented so that u_0 occurs with the same orientation in both. Let γ, γ' be curves whose cutting sequences are the doubly infinite periodic words $E = \ldots www \ldots$, $E' = \ldots w'w'w' \ldots$, positioned so that both paths leave the region R across the common (exterior) edge u_0. The endpoints of γ, γ' on $\partial\mathbb{D}$ are exactly the fixed points of w, w'.

Let $B = u_{-m} \ldots u_{-1} u_0 u_1 \ldots u_n$ denote the maximal block along which E, E' are coincident, and let S be the chain of regions $(u_{-m} \ldots u_{-1})^{-1}R, \ldots, u_1^{-1}R, R, u_0 R, \ldots, u_0 u_1 \ldots u_n R$. Then $\gamma \cap \gamma' \subset S$, for if γ and γ' both intersected the region $gR \notin S$ there would be two shortest paths from, say, $u_0 \ldots u_n R$ to gR, which is impossible.

(In the closed surface case, assume all cycles and chains in E or E' which may be switched on τ are anticlockwise. The two shortest paths can differ at most by cycle switches, however since both paths have all the cycles and chains with the same orientation they must coincide.)

Suppose that $E = \ldots eBf \ldots$, $E' = \ldots e'Bf' \ldots$, where by assumption $e \neq e'$ and $f \neq f'$. Since $\gamma \cap \gamma' \subset S$, the order of the endpoints of γ and γ' on $\partial\mathbb{D}$ is the same as the order of the (exterior) sides f and f' of $u_0 \ldots u_n R = R_n$ and $\bar{e}$, $\bar{e}'$ of $(u_{-m} \ldots u_{-1})^{-1} R = R_{-m}$ around ∂S (see Fig. 11). Thus γ and γ' have an intersection corresponding to the common letter u_0 if and only if the pairs of sides f of R_n, $\bar{e}$ of R_{-m} and f' of R_n, $\bar{e}'$ of R_{-m} separate each other round ∂S.

Now one can detect which of these situations occurs mechanically, by a rule explained in detail in [2][2], as follows. Order the symbols in Γ_R in the anticlockwise order in which they appear around the exterior of R. (Thus in Example A we have $\Gamma_R = \{a, \bar{b}, \bar{a}, b, c, \bar{d}, \bar{c}, d\}$.) This is a cyclic ordering, well defined up to cyclic permutation. Let $e, f, g \in \Gamma_R$ be distinct. We say that *e preceeds f relative to g*, written $e <_g f$, if e occurs before f in the cyclic order of Γ_R starting at the exterior label g. For example, referring to Example A, Fig. 1(a), we have $b <_{\bar{b}} \bar{d}$ while $\bar{d} <_b \bar{b}$.

Writing $E = \ldots eBf \ldots$, $E' = \ldots e'Bf' \ldots$ as above, γ and γ' have an intersection corresponding to u_0 only if the order of $\bar{e}, \bar{e}'$ relative to u_{-m} is the same as the order of f, f' relative to $\bar{u}_n$. In this situation we say that w and w' *intersect across B.*

We illustrate this with examples B and C.

[2] Similar ideas have been exploited in [4] to count self-intersection numbers of non-simple curves.

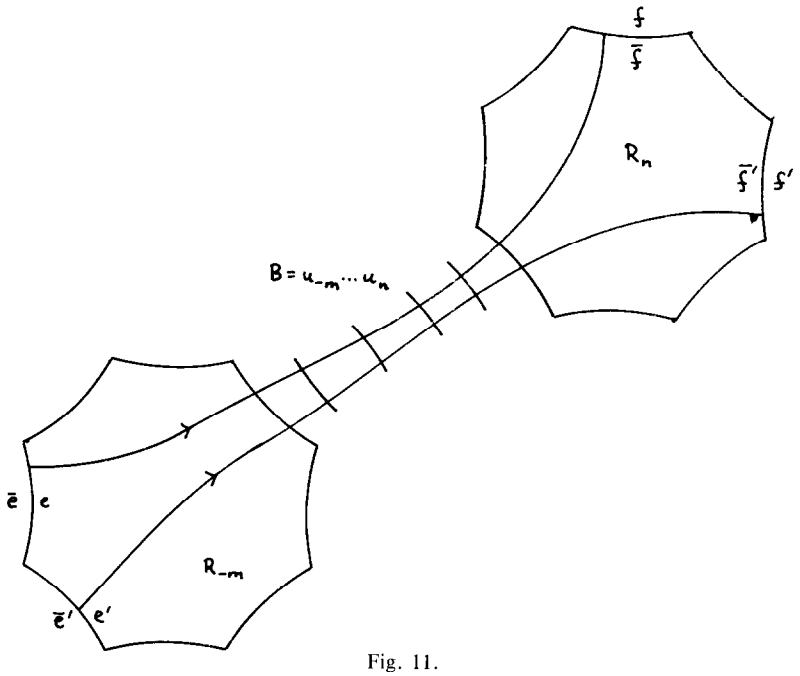


Fig. 11.

Example B (cf. Section 3). With $w_1 = ab^2$, $w_2 = b$ we have $E = E(w_1) = \ldots abbabbabb \ldots$, $E' = E(w_2) = \ldots bbb \ldots$. These have a maximal common block $B = bb$, thus $eBf = a(bb)a$, $e'Bf' = b(bb)b$. Since $\bar{b}$ preceeds $\bar{a}$ as one goes around the boundary of R starting at b in Fig. 4, we have $\bar{a} >_b \bar{b}$. Similarly, $a >_{\bar{b}} b$, hence there is an intersection across B. (This can also be checked by drawing the two curves carefully on R.)

Example C. Let M, Γ be as in Example A, so that Γ_R is the cyclically ordered set $(a, \bar{b}, \bar{a}, b, c, \bar{d}, \bar{c}, d)$. Let $w_1 = b$, $w_2 = bc\bar{d}\bar{c}$, $w_3 = d$, $w_4 = c\bar{d}$, $w_5 = a\bar{b}c\bar{d}$, $w_6 = cd\bar{c}a\bar{b}$. We claim that the intersection matrix for tight representatives $\gamma_1, \ldots, \gamma_6$ of these six shortest words, $\gamma_i = \pi(\gamma(w_i))$, is as given in Table 1. We check $\gamma_2 \cap \gamma_4$, leaving the remaining cases to the tireless reader. We have $E = E(w_2) = \ldots bc\bar{d}\bar{c}\, bc\bar{d}\bar{c} \ldots$, $E' = E(w_4) = \ldots c\bar{d}c\bar{d}c\bar{d} \ldots$, $E'' = E(\bar{w}_4) = \ldots d\bar{c}d\bar{c}d\bar{c} \ldots$. There are two common blocks to investigate, namely $c\bar{d}$ in E, E' and $\bar{c}$ in E, E''. Tackling E, E' first, we have $B = c\bar{d}$, $u_{-m} = c$, $\bar{u}_n = d$, $eBf = b(c\bar{d})\bar{c}$, $e'Bf' = \bar{d}(c\bar{d})c$. Then $\bar{b} >_c d$, $\bar{c} >_d c$, hence an intersection occurs. On the other hand, if we look at E and E'' we have $B = \bar{c}$, so $u_{-m} = \bar{c}$, $\bar{u}_n = c$, $eBf = \bar{d}(\bar{c})b$, $e''Bf'' = d(\bar{c})d$, and $d <_{\bar{c}} \bar{d}$, $b >_c d$, so no intersection occurs. Therefore γ_2, γ_4 intersect once, across the common block $c\bar{d}$ in w_2, w_4.

Remark. Notice that, as in the example above, it is necessary to compare both of the two possible relative orientations of the words in question.

Table 1

	γ_1	γ_2	γ_3	γ_4	γ_5	γ_6
γ_1		0	0	0	1	1
γ_2			0	1	0	1
γ_3				1	1	0
γ_4					0	1
γ_5						1
γ_6						

Checking that φ_* preserves τ-orientation

We now turn to the question of verifying that a diffeomorphism preserves τ-orientation of intersecting curves. Assume that words w, w' lie on a π_1-train track τ while $\varphi_* w$, $\varphi_* w'$ lie on τ', and that w and w' intersect across a common block B, corresponding to an intersection of appropriate lifts $\gamma(w)$, $\gamma(w')$ in $\mathbb{D}$. Then the image curves $\varphi_*(\gamma)$, $\varphi_*(\gamma')$ intersect and hence $\varphi_*(w)$, $\varphi_*(w')$ have a common block B'. The problem is to identify which of the common blocks in $\varphi_*(w)$ and $\varphi_*(w')$ corresponds to the intersection of w, w' across B.

Orient w and w' so that the common block B appears as B (and not B^{-1}) in both words. After cyclically permuting w and w' if necessary, the two sequences $ww\ldots$, $w'w'\ldots$ may both be taken to begin with B. Let the primitive periods of these sequences be u, u', where now u, u' are thought of as ordinary (not cyclic) words. Thus the initial letters of u and u' are the same as the initial letter of B. One can identify the block B' by applying the following lemma:

Lemma 7.1. *In the situation above, suppose that $\varphi(u) = xv\bar{x}$, $\varphi(u') = x'v'\bar{x}'$, where $v = \varphi_*(w)$, $v' = \varphi_*(w')$ are cyclically reduced and supported on τ'; x, x' are words in Γ_R and where $xvv\ldots$, $x'v'v'\ldots$ are shortest. Then there exist $\varepsilon, \varepsilon' = \pm 1$ and a word z so that*

(i) *z is the initial block of both $xv^{\varepsilon}v^{\varepsilon}\ldots$ and $x'v'^{\varepsilon'}v'^{\varepsilon'}\ldots$.*

(ii) *$|z| > \max(|x|, |x'|)$ (here $|z|$ denotes the word length of z).*

(iii) *The part of z contained in $v^{\varepsilon}v^{\varepsilon}\ldots$ and $v'^{\varepsilon'}v'^{\varepsilon'}\ldots$ is part of the common block B' of $\varphi_* w$, $\varphi_* w'$ corresponding to the intersection of w, w' across B.*

Further, φ_* preserves τ-orientation of w, w' at B if and only if $\varepsilon\varepsilon' = -1$.

As the statement of the lemma is somewhat involved, we give two examples before continuing with the proof.

Example B (continued). We showed above that the words $w_1 = abb$, $w_2 = b$ of Example C intersect once across the common block $B = bb$. To put ourselves in the situation of the lemma, replace w_1 by its cyclic permutation $u = bba$ and set $u' =$

$w_2 = b$. Thus $uu\ldots = bbabb\ldots$ and $u'u'\ldots = bbb\ldots$ both begin with B. Notice that in this case B occupies more than one period of $u'u'\ldots$.

Now computing *without* cyclic reduction we find $\varphi(u) = b^2a\bar{b}^4 = b^2(a\bar{b}^2)\bar{b}^2$ and $\varphi(u') = b$. Thus with the notation of the lemma,

$$x = b^2, \qquad v = a\bar{b}^2, \qquad x' = \emptyset \quad \text{and} \quad v' = b.$$

Thus

$$xvv\ldots = b^2a\bar{b}^2a\bar{b}^2\ldots, \qquad x\bar{v}\bar{v}\ldots = b^4\bar{a}b^2\bar{a}b^2\ldots,$$

$$x'v'v'\ldots = bbb\ldots, \qquad \bar{x}'v'v'\ldots = \bar{b}\bar{b}\bar{b}\ldots.$$

Both $xvv\ldots$ and $x\bar{v}\bar{v}\ldots$ begin with the same block b^2 as $x'v'v'\ldots$. However, only for $x\bar{v}\bar{v}\ldots$ does the block extend into $\bar{v}\bar{v}\ldots$. Thus $\varepsilon = -1$, $\varepsilon' = +1$, $z = b^4$, $B' = b^2$, and since $\varepsilon\varepsilon' = -1$, φ_* reserves τ-orientation at B.

Example C (continued). First note (Fig. 16(a)) that $G(W)$ is a simple graph, hence a π_1-train track, τ. Note that it is non-orientable.

Let $\varphi \in \mathrm{Aut}\,\Gamma$ be defined by

$$\varphi(a) = a\bar{b}\bar{b}cd\bar{c}, \qquad \varphi(c) = c\bar{d}\bar{c}bc\bar{d},$$

$$\varphi(b) = c\bar{d}\bar{c}bbad\bar{c}bb\bar{a}, \qquad \varphi(d) = d\bar{c}b\bar{a}dd\bar{c}\bar{b}cd\bar{c}.$$

We claim that φ_* preserves τ-orientation at each of the eight intersection points $\gamma_i \cap \gamma_j$, and verify this assertion at the point $\gamma_2 \cap \gamma_4$ studied earlier. As in the previous example we replace $w_2 = bc\bar{d}\bar{c}$ by its cyclic permutation $u = c\bar{d}\bar{c}b$ and let $u' = w_4 = c\bar{d}$, so that u and u' both begin with the common block $c\bar{d}$ across which γ_2 intersects γ_4. Applying φ, we obtain

$$\varphi(c\bar{d}\bar{c}b) = c\bar{d}\bar{c}bc\bar{d}c\bar{d}\bar{c}bc\bar{d}\bar{c}bb\bar{a}, \qquad \varphi(c\bar{d}) = c\bar{d}\bar{c}bc\bar{d}c\bar{d}\bar{c}bc\bar{d}\bar{d}a\bar{b}c\bar{d}.$$

Our words are cyclically reduced as written, so $x = x' = \emptyset$ and our first word is v, the second v'. These begin with the common block $z = c\bar{d}\bar{c}bc\bar{d}c\bar{d}\bar{c}bc\bar{d}$. Thus $\varepsilon = \varepsilon' = +1$, and φ_* preserves τ-orientation of $\gamma_2 \cap \gamma_4$.

Proof of Lemma 7.1 (Fig. 12(a)). Let $\gamma(w)$ and $\gamma(w')$ be tight curves in $\mathbb{D}$ which intersect across B. Consider the endpoint $\xi \in \partial\mathbb{D}$ of the path $xvv\ldots$ starting from 0. This is fixed by xvx^{-1}, and hence $\xi = \bar{\varphi}(\gamma(w)_\infty)$, where $\gamma(w)_\infty$ is the positive endpoint of $\gamma(w)$ on $\partial\mathbb{D}$. The endpoints of $x\bar{v}\bar{v}\ldots$ and $x'v'^{\pm 1}v'^{\pm 1}\ldots$ are similarly identified. The curves $\gamma(\varphi_*(w))$, $\gamma(\varphi_*(w'))$ with paths $\ldots vv\ldots$, $\ldots v'v'\ldots$, join these endpoints and, because $\bar{\varphi}$ is order preserving on $\partial\mathbb{D}$, intersect in a region gR across the block B. For some choice of $\varepsilon, \varepsilon' = \pm 1$, the region gR is connected to 0 by both the paths $xv^{\varepsilon}v^{\varepsilon}\ldots$ and $x'v'^{\varepsilon'}v'^{\varepsilon'}\ldots$, in particular, $g = xz = x'z'$ in Γ where $z \subset v^{\varepsilon}v^{\varepsilon}\ldots$, $z' \subset v'^{\varepsilon'}v'^{\varepsilon'}\ldots$, and $z, z' \neq \emptyset$. Since Γ is a free group and the words xz, $x'z'$ are reduced, we must have $xz = x'z'$ as words in Γ. Moreover, z contains at least one of the common letters of the block B'. The condition $\varepsilon\varepsilon' = +1$ is exactly the condition that B' appear with the same orientation in $\ldots vv\ldots$ and $\ldots v'v'\ldots$.

[If Γ is a free group, then it is always possible to write $\varphi(u), \varphi(u')$ in the required form. For the closed surface, this is not necessarily the case. However, if it is possible, then the lemma applies as stated except that we now find words z, z' which are the initial blocks of $xv^{\varepsilon}v^{\varepsilon} \ldots$ and $x'v'^{\varepsilon'}v'^{\varepsilon'} \ldots$ so that $xz = x'z'$ up to cycle switches. Possible configurations are illustrated in Figs 12(b and c). A further treatment of the closed surface appears at the end of the section.]

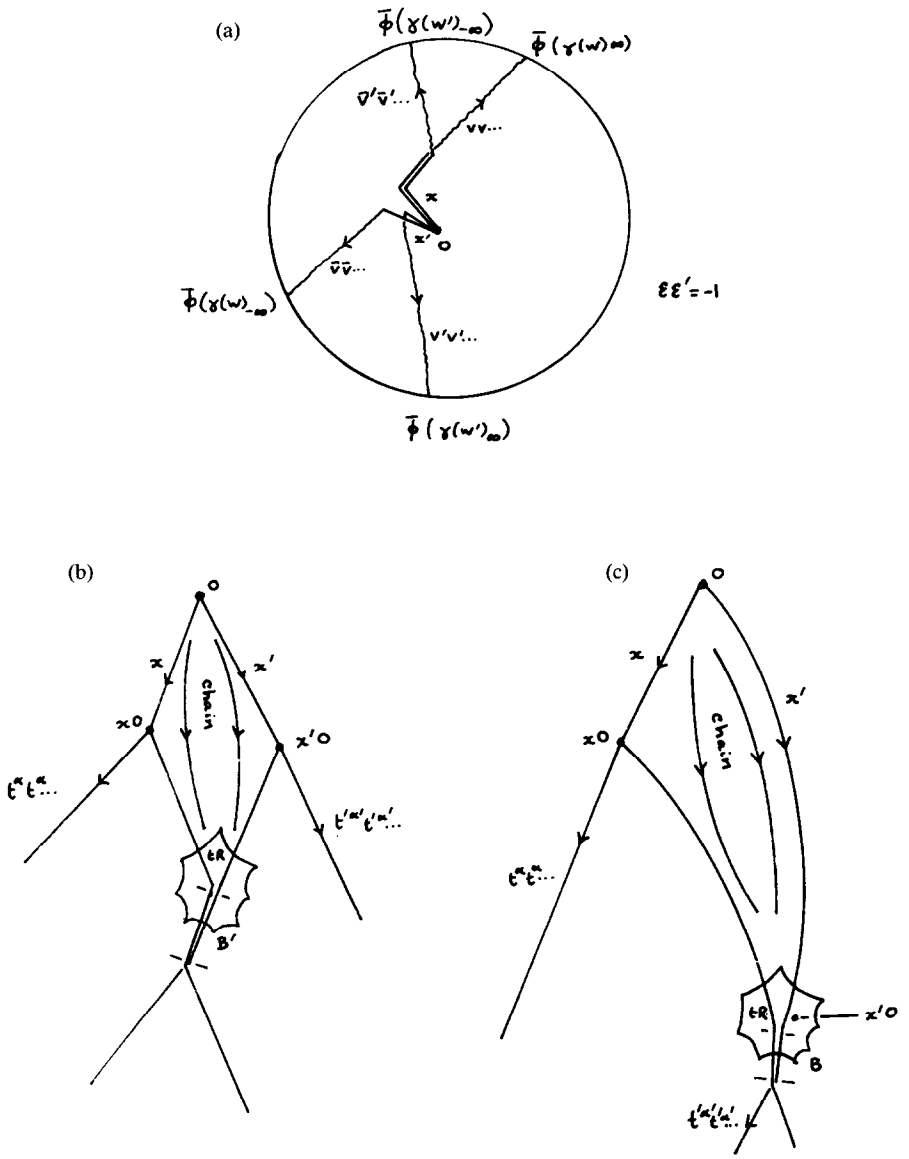


Fig. 12.

Recognising trigons

In order to check the condition that φ_* preserve F-trigons, we need to be able to recognize trigons. In the light of the previous work, this is not hard to do. We can certainly pick out those triplets of words w_1, w_2, w_3 which intersect in pairs and identify their common blocks. Let B_i be the maximal common block associated to the intersection of w_j, w_k. If there are lifts γ_i of w_i, $i=1,2,3$, which lie over a trigon, then, orienting the three curves appropriately, there are (possibly empty) blocks $A_i \subset w_i w_i \dots w_i$ such that

$$B_{i+1} A_i B_{i-1}^{-1} \subset w_i w_i \dots w_i \quad \text{for each } i$$

(i is here defined mod 3), as in Fig. 13. If $A_i = \emptyset$ for each i, then the trigon lies in one copy of R, joining the three (exterior) sides labelled by the inverses of the last letters in the blocks B_i. This allows one to determine the orientation of the trigon. More generally, $A_1 A_2 A_3$ is a non-self intersecting closed loop in the graph Γ, and hence $A_1 A_2 A_3 = \mathrm{id}$ in Γ. The orientation of this relation determines the orientation of the trigon. Of course, this last situation can only arise for non-free Γ. It is illustrated by Fig. 9. Notice that two different trigons may be associated to the same triple of common blocks B_i, for example, one can have $w_i = B_i B_{i+1}^{-1}$, $i=1,2,3$ forming one trigon, while after cyclic permutation and inversion we find another trigon $w_i' = B_{i+1}^{-1} B_i$. This situation is illustrated in Fig. 5. There may, of course, be triangles which do not lie over trigons at all, as illustrated in Fig. 7. Here it is impossible to orient w_2, w_4 and w_6 so that the common blocks are simultaneously oppositely oriented at each intersection point. In fact, this is the case for all the intersecting triplets of Example C. Finally, in Fig. 14 we illustrate a trigon in which the product $A_1 A_2 A_3$ is more than just the relation in the group.

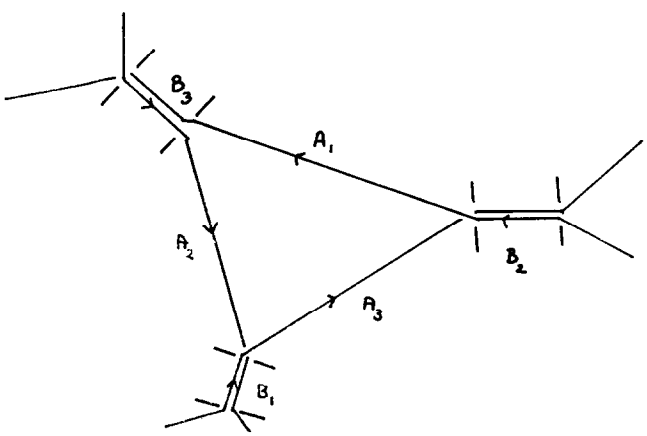


Fig. 13.

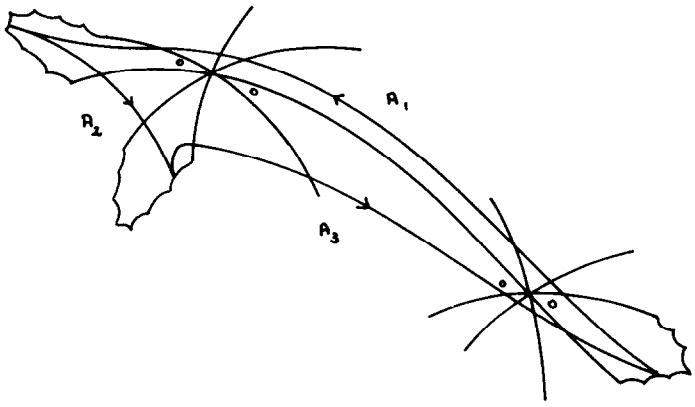


Fig. 14. ○ is an omitted corner edge on τ.

Having identified trigons it is a simple matter to compute the 'image trigon' under φ_* and check the orientation condition as required. We were unable to construct examples in which this condition fails, although we see no reason to suppose they do not exist.

A priori the blocks A_i could be arbitrarily long and thus there could be an infinite family of trigons associated to a given triple of blocks B_i. This would be highly undesirable as far as checking the conditions of our theorem is concerned. In fact this situation cannot arise, as shown by the following lemma, which is illustrated by Fig. 15.

Lemma 7.2. *Let B_1, B_2, B_3 be maximal blocks across which words w_1, w_2, w_3 intersect in pairs. Then there are at most two ways in which the blocks B_i can be fitted together to form a trigon (possibly with intervening blocks A_i). These trigons, if they both occur, have opposite orientations.*

Proof. Fix lifts γ_1, γ_2 of w_1, w_2, whose cutting sequences $\ldots w_1 w_1 \ldots$, $\ldots w_2 w_2 \ldots$ intersect across the block B_3. Let $P = \gamma_1 \cap \gamma_2$.

A triangle could be formed by lifts of w_3 cutting γ_1, γ_2 on either side of P. We shall show that there is at most one lift which forms a trigon on each side of P.

Suppose to the contrary that γ_3 and γ_3' are distinct curves both with cutting sequences $\ldots w_3 w_3 \ldots$ which together with γ_1, γ_2 lie over two distinct trigons on the same side of P, as in Fig. 15. Let S and T denote the endpoints of the common blocks of γ_3 with γ_1 and γ_2 which lie closest to P, and define S', T' similarly. The points S, S' and T, T' occur in the same order along γ_1, γ_2 since γ_3, γ_3' are both lifts of the simple curve w_3. Now consider the two paths from S' to T', one along γ_3' and the other from S' to S along γ_1, then S to T along γ_3, and T to T' back along γ_2. These paths both lie on τ and this situation is therefore ruled out if Γ is free. [For Γ non-free, they differ by at most a chain switch. Thus the path from S' to T' on

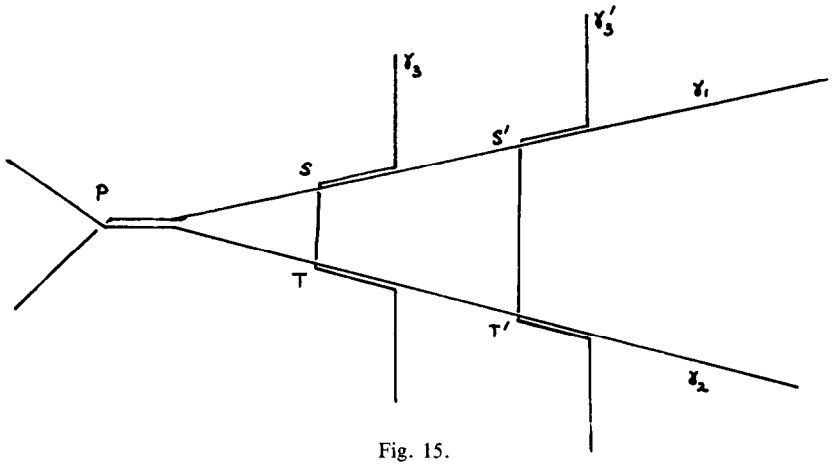


Fig. 15.

γ_3 could be switched on τ to the other path, but this would produce longer common blocks B_1, B_2 with w_1 and w_2, contrary to the hypothesis that B_i were maximal common blocks.] This proves our result. □

Example. We are finally in a position to illustrate our theorem by working out in full the details of Example C. The six images of $w_1, \ldots, w_6$ under φ_* are

$$v_1 = \varphi_*(w_1) = c\bar{d}\bar{c}bb\bar{a}d\bar{c}bb\bar{a},$$

$$v_2 = \varphi_*(w_2) = \bar{c}bb\bar{a}c\bar{d}\bar{c}bc\bar{d}c\bar{d}\bar{c}bc\bar{d},$$

$$v_3 = \varphi_*(w_3) = d\bar{c}b\bar{a}dd\bar{c}\bar{b}cd\bar{c},$$

$$v_4 = \varphi_*(w_4) = c\bar{d}\bar{c}bc\bar{d}c\bar{d}\bar{c}bc\bar{d}\bar{d}a\bar{b}c\bar{d},$$

$$v_5 = \varphi_*(w_5) = a\bar{b}\bar{b}cd\bar{c}a\bar{b}\bar{b}cda\bar{b}cdc\bar{d}\bar{c}bcd\bar{d}a\bar{b}c\bar{d},$$

$$v_6 = \varphi_*(w_6) = d\bar{c}\bar{b}cd\bar{c}d\bar{c}\bar{b}cd\bar{c}a\bar{b}\bar{b}cd\bar{c}a\bar{b}\bar{b}c.$$

Let $W = \{w_1, \ldots, w_6\}$, $V = \{v_1, \ldots, v_6\}$. An easy check, using Fig. 16, shows that $G(V) = G(W)$, hence all conditions for Theorem 6.0 are satisfied[3] and the action is linear.

The linearity will be seen to illustrate some interesting points. The map φ_* of Example C is pseudo-Anosov, having been constructed by the method of [10]. It maps a cell in measured lamination space into itself, and in fact we will see that this cell may be taken to be $\mathrm{Sp}^+(W)$. To prove this, we must show that $V \subset \mathrm{Sp}^+(W)$, a fact easily verified with a little linear algebra. Choose an ordering (in any way) of the edges which occur in $G(W)$, e.g. see Fig. 16, where the ten edges have been

[3] In this case the condition that φ_* preserves orientation of F-trigons is vacuous, because as noted earlier there are no triangles which lie over trigons.

(a) (b) $w_2 + w_4 = \bar{c}\bar{b}cd\bar{c}d$.

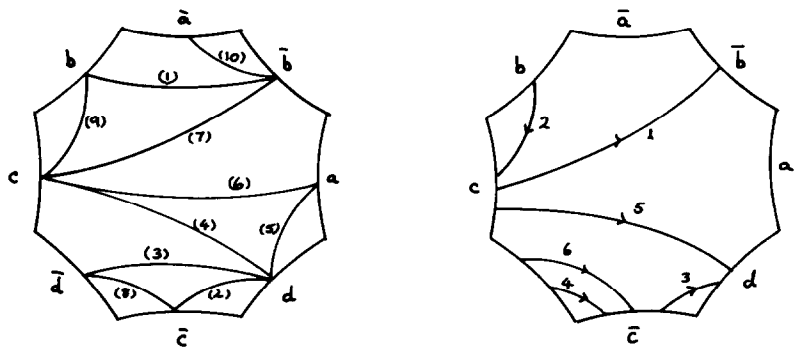


Fig. 16.

labelled 1 to 10. Using this ordering, one verifies that the π_1-parameters for the w_j's are $x_{w_1} = (1,0,0,0,0,0,0,0,0,0)$, $x_{w_2} = (0,1,0,0,0,0,1,1,1,0)$ and so forth, the key point being that the six weights are linearly independent, so that $w_1, \ldots, w_6$ are linearly independent. Since there are four edge-pairing relations which relate the ten weights, it is clear that the dimension of $\mathrm{Sp}^+(W)$ is at *most* six, hence it is exactly six and these six curves are a basis.

Remark. (i) In this case, $\mathrm{Sp}^+(W)$ actually coincides with the set of all weights supported on τ, but that will not be true for arbitrary pseudo-Anosov maps, which will in general have as invariant cells proper (even very small) subsets of the weights supported on τ.

(ii) We have deliberately bypassed the question of how we found the six words $w_1, \ldots, w_6$ for this pseudo-Anosov map. Our method was basically experimental, i.e. iterate the map and investigate the blocks which occur repeatedly in the image, then look at their images, correcting as you go. This idea is due to Nielsen (see [7]). We consider the matter of finding a general working technique sufficiently deep to require a separate investigation.

Continuing, it is now an easy matter to express the v_j's as linear combinations of the w_j's, by counting up how many times each of the six key 2-letter syllables occurs in a given v_j. This data is incorporated into a positive matrix that we call Φ_*, the rows being indexed by $w_1, \ldots, w_6$ and the columns by $v_1, \ldots, v_6$.

$$\Phi_* = \begin{bmatrix} 2 & 1 & 0 & 0 & 2 & 2 \\ 0 & 2 & 1 & 2 & 1 & 2 \\ 0 & 0 & 1 & 1 & 1 & 0 \\ 0 & 1 & 1 & 2 & 1 & 1 \\ 1 & 0 & 1 & 1 & 3 & 0 \\ 1 & 1 & 0 & 0 & 1 & 2 \end{bmatrix}.$$

The reader familiar with [3] will realize that the largest positive real eigenvalue of this matrix is the 'stretch factor' of φ, and the associated eigenvector determines the invariant lamination, which can now be re-interpreted in terms of 'infinite homotopy classes' in Γ.

The square of Φ_* is strictly positive which gives an independent check that φ is indeed pseudo-Anosov.

It will be worthwhile to study at least one of the words v_j, say v_6, to see how it happens that $v_6 = 2w_1 + 2w_2 + w_4 + 2w_6$. (A similar, but much more trivial calculation was done in the introduction.) We first subdivide v_6 into blocks by parenthesis:

$$v_6 = [d(\bar{c}\bar{b}cd)\bar{c}]\ [d\bar{c}\bar{b}c]\ [dca(\bar{b})\bar{b}c]\ [dca(\bar{b})\bar{b}c].$$

The first square bracket is $\bar{w}_2 + \bar{w}_4$, the $\bar{w}_2$ being in the round brackets in the middle. The second square bracket is $\bar{w}_2$. The third and fourth are both $w_6 + \bar{w}_1$ occuring inside the round brackets. Counting up, we get $w_1^{\pm 1}$ twice, $w_2^{\pm 1}$ twice and so forth. The magic in this calculation is, however, not just this decomposition, but also the fact that the 2-letter syllables at the interfaces add up exactly the same way as the 2-letter syllables which would have been obtained by completing each $w_i^{\pm 1}$ (as it occurs in v_j) to a cyclic word! That is, the 2-letter syllables at the interfaces between square or round brackets in v_6, (i.e. $d\bar{c}, d\bar{c}, \bar{c}d, cd, a\bar{b}, \bar{b}\bar{b}, cd, a\bar{b}, \bar{b}\bar{b}$), together with the last and first letter of v_6, (i.e. cd), have the same total weights as the set of 2-letter syllables obtained by completing each $w_i^{\pm 1}$ (as it occurs in v_6) to a cyclic word, (namely $d\bar{c}, \bar{c}d, cd, cd, \bar{b}\bar{b}, dc, \bar{b}\bar{b}, d\bar{c}, a\bar{b}, a\bar{b}$) and the 2-letter syllables which were missed in v_i when the 3 words in square brackets were split apart to insert the words in round brackets, (i.e. $d\bar{c}, a\bar{b}, a\bar{b}$).

We demonstrate the linearity theorem by verifying directly that

$$x_{\varphi_*(w_2)} + x_{\varphi_*(w_4)} = x_{\varphi_*(w_2 + w_4)}.$$

We have $w_2 + w_4 = \bar{c}\bar{b}cd\bar{c}d$ (Fig. 16(b)). By direct computation,

$$\begin{aligned}
\varphi_*(\bar{c}\bar{b}cd\bar{c}d) &= \varphi(\bar{c})\varphi(\bar{b})\varphi(c)\varphi(d)\varphi(\bar{c})\varphi(d) \\
&= \{d\bar{c}\bar{b}cd\bar{c}\}\ \{a\bar{b}\bar{b}cda\bar{b}\bar{b}cd\bar{c}\}\ \{c\bar{d}\bar{c}bc\bar{d}\}\ \{d\bar{c}b\bar{a}dd\bar{c}\bar{b}cd\bar{c}\} \\
&\quad \{d\bar{c}\bar{b}cd\bar{c}\}\ \{d\bar{c}b\bar{a}dd\bar{c}\bar{b}cd\bar{c}\} \\
&= [d\bar{c}\bar{b}c]\ [dca(\bar{b})\bar{b}c]\ [d\bar{c}\bar{b}c]\ [d\bar{c}]\ [d\bar{c}\bar{b}c]\ [d\bar{c}]\ [d\bar{c}\bar{b}a]\ [d]\ [d\bar{c}\bar{b}c]\ [d\bar{c}] \\
&= w_2 + (w_6 + w_1) + w_2 + w_4 + w_2 + w_4 + w_5 + w_3 + w_2 + w_4 \\
&= \varphi_*(w_2) + \varphi_*(w_4),
\end{aligned}$$

as one verifies by consulting the matrix above.

The closed surface case

Our discussion above applies equally well to the closed surface case, except that we need to discuss the situation in which the hypotheses of Lemma 7.1 do not apply;

that is, in which, with the notation of that lemma, it is impossible to find words x, x', v, v' so that $\varphi(u) = xvx$, $\varphi(u') = x'v'x'$, so that v and v' are supported on τ' and so that $xvv\ldots$ and $x'v'v'\ldots$ are shortest.

There are two distinct problems here. Of course we may always write $\varphi(u) = yt\bar{y}$, where t is shortest and cyclically reduced.

(i) The word t may not be cyclically shortest, so that $ttt\ldots$ is not shortest. This will occur only if there is a long chain across the join of t and t.

(ii) The word t we find in this way may not coincide with the representative of $\varphi_* w$ which is known to lie on τ'.

This situation is partially remedied by the following lemma:

Lemma 7.3. *Suppose that $g \in \Gamma$ is conjugate to the cyclically shortest word t. Then there exists a word x and a cyclic permutation t^α of t so that $g = xt^\alpha\bar{x}$ and $xt^\alpha t^\alpha\ldots$ is shortest.*

Proof. Clearly we have $g = yt\bar{y}$, where y may be taken to be a word in shortest form. Let γ be a path through $B = y0$ whose cutting sequence is $tt\ldots$ (Fig. 17), and let ξ be the endpoint of γ on $\partial\mathbb{D}$. If $P, Q \in \gamma$ we write $P > Q$ if P lies between ξ and Q.

If the path $ytt\ldots$ is not shortest, then there is a long cycle or chain at B. (A cancellation cannot occur because t is cyclically reduced.) Replacing this chain by its complement we obtain a path from 0 to a point $B_1 \in \gamma$ strictly shorter than the path via B. Let the path from 0 to B_1 have cutting sequence y_1. Clearly, $g = y_1 t^{\alpha_1} \bar{y}_1$ where t^{α_1} is a cyclic permutation of t.

If $y_1 t^{\alpha_1} t^{\alpha_1}\ldots$ is not shortest, we may repeat the process and hence inductively find a sequence of points $B_0 = B, B_1, B_2, \ldots$ along γ, cyclic permutations $t = t^{\alpha_0}, t^{\alpha_1}, t^{\alpha_2}, \ldots$ of t, and shortest words $y = y_0, y_1, y_2, \ldots$, so that $B_i = y_i 0$, $g = y_i t^{\alpha_i} \bar{y}_i$ and so that $|y_i| < |y_{i-1}| + |B_{i-1} B_i|$ for $i = 0, 1, 2, \ldots$.

This process of length reduction cannot continue indefinitely. For suppose it continued $k = 2y_0 + 1$ times. Choose $P \in \gamma$ with $P > B$.

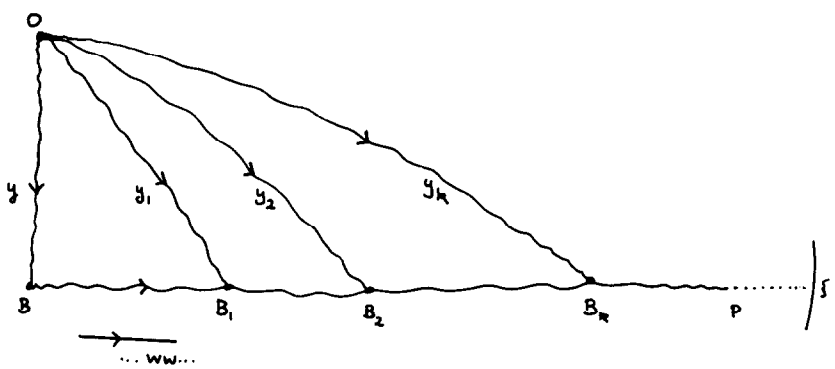


Fig. 17.

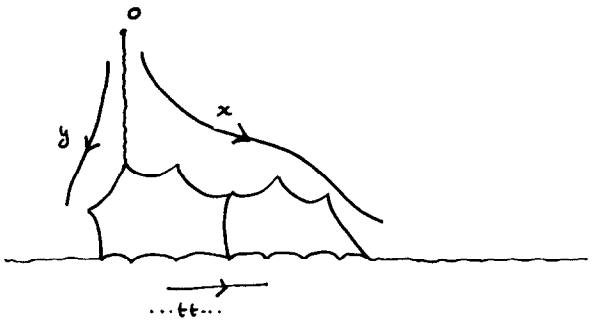


Fig. 18.

Then $|B_kP|+|y_k|<|y_0|+|B_0P|-2|y_0|$, while by the reverse triangle inequality

$$|y_k|\geq|B_0B_k|-|y_0|=|B_0P|-|y_0|-|PB_k|.$$

These two inequalities are incompatible, so that the shortening process stops after at most $2|y_0|$ steps. □

Remark. This proof gives a constructive method of finding x, t^α once the initial conjugating element y is known. It is also possible to find y: starting from a shortest, but not cyclically shortest, representative of g we may repeatedly cyclically permute and reduce in length until we find a cyclically shortest conjugate g'. We then apply Theorem 5.1(iv) to see the relation between g' and the given cyclically shortest conjugate t.

In an alternating presentation of Γ (see Section 5 and also [1]), if t is cyclically shortest, then $tt\ldots$ is shortest. Thus the lemma resolves the difficulties (i) and (ii) above, in that we may apply it with $g=\varphi(u)$ and choose t to be the cyclically shortest representative of $\varphi_*(w)$ which by hypothesis lies on τ'. We find $\varphi(u)=xt^\alpha\bar{x}$, where t^α is supported on τ' and $xt^\alpha t^\alpha\ldots$ is shortest. Obviously we may equally well find $\tilde{x}$ and t^β so that $\varphi(u)=\tilde{x}t^\beta\tilde{x}^{-1}$ and $\tilde{x}\bar{t}^\beta\bar{t}^\beta\ldots$ is shortest. However, it is not in general possible to satisfy the requirement that $xt^\alpha t^\alpha\ldots$ and $x\bar{t}^\alpha\bar{t}^\alpha\ldots$ be the shortest simultaneously, as for example in Fig. 18. Nevertheless, knowledge of the x, t^α such that $\varphi(u)=xt^\alpha\bar{x}$ and $xt^\alpha t^\alpha\ldots$ is shortest, allows one to correctly position $\gamma(\varphi_*(w))$ as a curve with cutting sequence $\ldots tt\ldots$ passing through $x0$. Likewise we may determine the position of $\gamma(\varphi_*(w'))$. The precise position of the intersecting block B' may now be determined graphically. There are undoubtedly very strong constraints on the relation of the sequences $xt^\alpha t^\alpha\ldots$ and $x't'^{\alpha'}t'^{\alpha'}\ldots$ but it becomes tedious to enumerate them all. Some possible configurations are illustrated in Fig. 19.

Appendix: The punctured torus

In the special case of the punctured torus T^* the linearity theorem is a straightforward consequence of the facts that a simple curve is uniquely determined by its

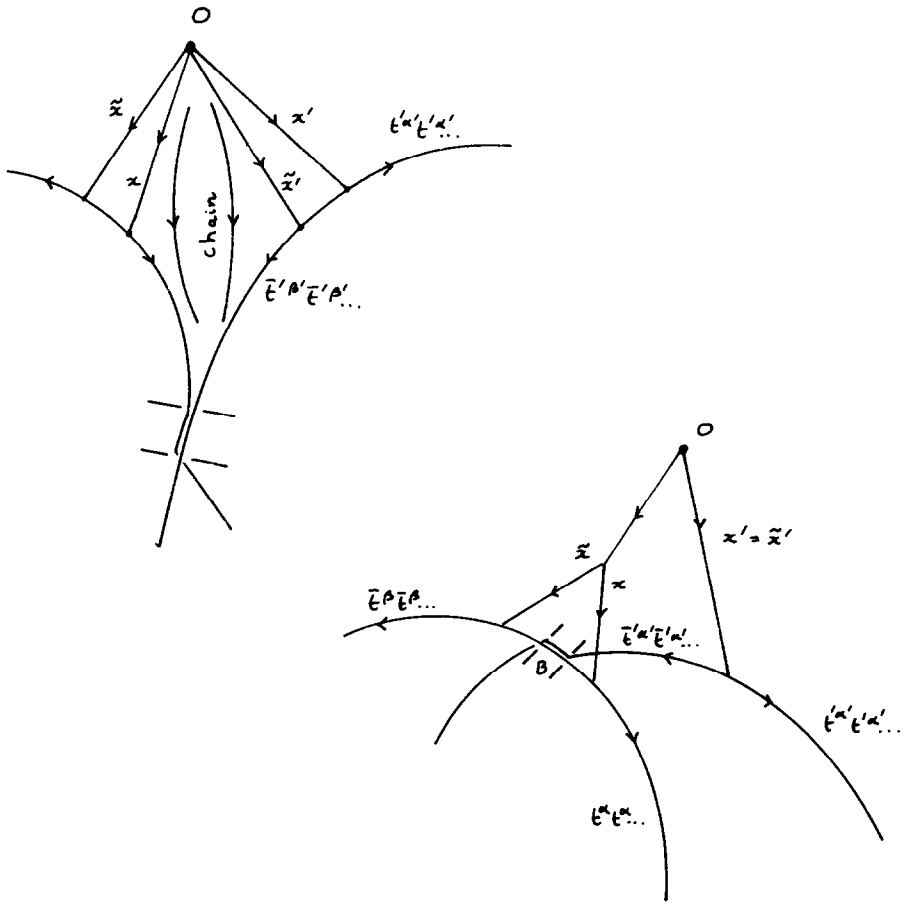


Fig. 19.

representative in the first homology group $H_1(T^*)$ and that $\varphi \in \mathrm{Diff}(T^*)$ induces a linear map φ_1 on H_1.

We begin by determining all π_1-train tracks for the punctured torus. With a fundamental region R as in Fig. 4, *a priori* any π_1-train track on R has five branches. However, the boundary conditions (Restriction 1.2) imply that one of the four corner edges is absent. The switch conditions (Condition 1.1) then imply further that the opposite corner edge to the missing one has weight zero while the other pair of opposite edges have equal weights. Thus any weight on τ has only two independent parameters. For example on the train track in Fig. 4(a) one has $x(b, \bar{b}) = m$, $x(\bar{a}, b) = x(a, \bar{b}) = n$. Taking into account symmetries and orientation, there are in all eight distinct oriented π_1-train tracks τ_i for T^*. These are shown arranged round the circle S^1 in Fig. 20, in accordance with the images of the curves they carry in $H_1 = \mathbb{Z}^2$. The track in the example above is τ_1.

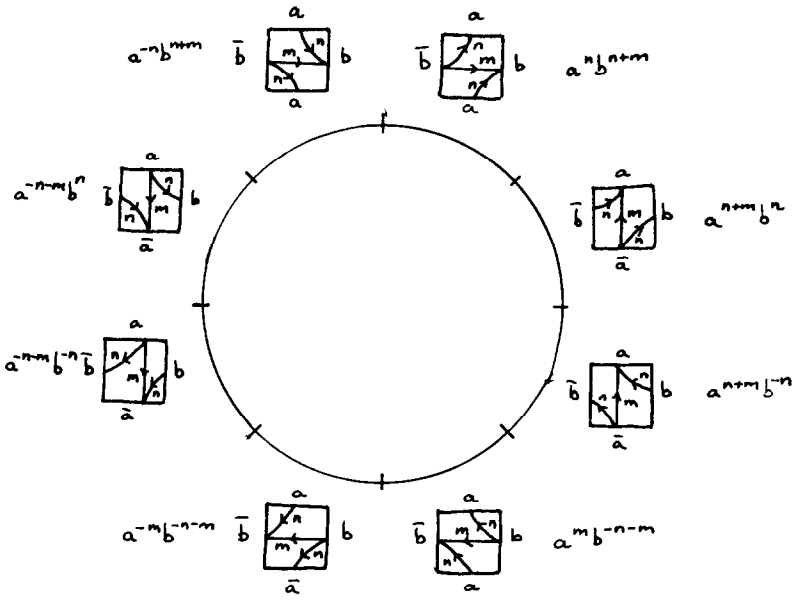


Fig. 20.

Let $\Omega^+(\tau_i)$ denote the space of oriented words obtained by taking a weight $x \in \Omega(\tau_i)$ to define a word oriented coherently with τ_i. The map T_i associating to each word its image in H_1 is linear and bijective onto its image. Thus for example,

$$T_1(x) = a^n b^{n+m},$$

where $x(b, \bar{b}) = m$, $x(\bar{a}, b) = x(a, \bar{b}) = n$, and H_1 is identified with $\mathbb{Z}^2$ by $a^r b^s \to (r, s)$.

Suppose that φ maps $\{x_1, \ldots, x_k\} \in \Omega^+(\tau_i)$ onto $\{\varphi_*(x_1), \ldots, \varphi_*(x_k)\} \in \Omega^+(\tau_j)$. Then the restriction of φ to $\mathrm{Sp}^+\{x_1, \ldots, x_k\}$ is $T_j^{-1}\varphi_1 T_i$, and hence is linear. It is worthwhile studying Example B in this context, to see why linearity fails.

Notice that the arrangement of the train tracks τ_i around S^1 in Fig. 20 shows directly that the space of projective measured laminations on T^* has dimension $6g - 7 + 2b = 1$. It would be nice to have a direct proof that in general our π_1-train tracks fit together properly to form a sphere of the right dimension.

Acknowledgment

We would like to thank the referee for his or her very careful reading of the manuscript and detailed comments.

References

[1] J.S. Birman and C. Series, Dehn's algorithm revisited, Combinatorial Group Theory and Topology, Annals of Mathematics Studies 111 (Princeton University Press, Princeton, NJ, 1987) 451–478.

[2] J.S. Birman and C. Series, An algorithm for simple curves on surfaces, J. London Math. Soc. (2) 29 (1984) 331–342.

[3] A. Casson and S. Bleiler, Automorphisms of Surfaces after Nielsen and Thurston, Student Lecture Notes Series of the London Mathematical Society (Cambridge University Press, Cambridge), to appear.

[4] M. Cohen and M. Lustig, Paths of geodesics and geometric intersection numbers I and II, Combinatorial Group Theory and Topology, Annals of Mathematics Studies 111 (Princeton University Press, Princeton, NJ, 1987) 479–544.

[5] A. Fathi, F. Laudenbach, Poénaru et al., Travaux de Thurston sur les surfaces, Asterisque 66–67 (1979).

[6] L.R. Ford, Automorphic Functions (Chelsea, New York, 1951).

[7] J. Gilman, Determining Thurston classes using Nielsen types, Trans. Amer. Math. Soc. 272(2) (1982) 669–675.

[8] J. Nielsen, Untersuchungen zur Topologie der geschlossenen Zweiseitigen Flachen, Acta Math. 50 (1927).

[9] R. Penner, Ph.D. Thesis, MIT, 1982.

[10] R. Penner, A construction of pseudo-Anosov homeomorphisms, Preprint.

[11] C. Series, The infinite word problem and limit sets in Fuchsian groups, Ergodic Theory Dynamical Systems 1 (1981) 337–360.

[12] C. Series, The geometry of Diophantine approximation, Math. Intelligencer 7 (1985) 20–29.

[13] W. Thurston, On the geometry and dynamics of diffeomorphism of surfaces, Preprint, Princeton University, 1978.

[14] G.H.C. Whitehead, On certain sets of elements in a free group, Proc. London Math. Soc. 41 (1936) 48–56.