ISRAEL JOURNAL OF MATHEMATICS, Vol. 51, Nos. 1-2, 1985

INTERSECTIONS OF CURVES ON SURFACES

BY

JOEL HASS[a] AND PETER SCOTT[b]

[a]*Mathematics Department, University of Michigan, MI 48109, USA;*
and [b]*Department of Pure Mathematics, University of Liverpool,*
P.O.Box 147, Liverpool L69 3BX, England

ABSTRACT

The authors consider curves on surfaces which have more intersections than the least possible in their homotopy class.

THEOREM 1. *Let f be a general position arc or loop on an orientable surface F which is homotopic to an embedding but not embedded. Then there is an embedded* 1*-gon or* 2*-gon on F bounded by part of the image of f.*

THEOREM 2. *Let f be a general position arc or loop on an orientable surface F which has excess self-intersection. Then there is a singular* 1*-gon or* 2*-gon on F bounded by part of the image of f.*

Examples are given showing that analogous results for the case of two curves on a surface do not hold except in the well-known special case when each curve is simple.

Let C_1 and C_2 be simple closed curves on the annulus A. It is easy to show that if C_1 and C_2 intersect and do so transversely, then there must be a 2-disc D in A whose boundary is $\lambda_1 \cup \lambda_2$ where λ_i is a sub-arc of C_i. We call such a disc a *2-gon* between C_1 and C_2. If two simple closed curves C_1 and C_2 on a surface F intersect transversely, we will say that C_1 and C_2 have *excess intersection* if one of them can be homotoped so as to reduce the number of intersection points with the other. The natural generalisation of the above result about two curves on the annulus is that if C_1 and C_2 are simple closed curves on a surface F and if they have excess intersection then there is a 2-gon between C_1 and C_2. This result is fairly well known, but, for completeness, we give a proof at the start of §3.

In this paper, we consider the question of finding analogous results about the intersection of two possibly singular loops on a surface and about the self-intersection of a single loop. Various results in this area have been assumed to be obvious by some authors. However, we give examples which demonstrate that

Received June 19, 1984

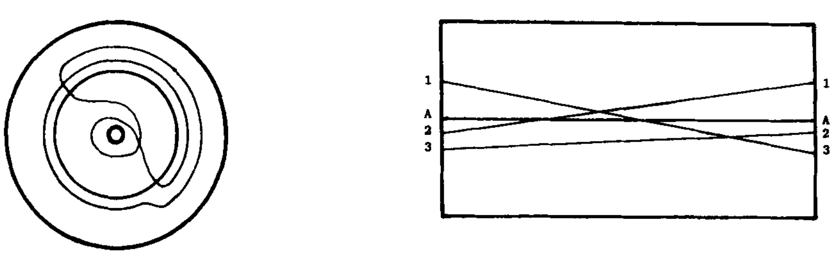
many of these assumptions are incorrect. Further, the results which are true are surprisingly difficult to prove.

One of our original motivations for this work was to obtain a new proof of the results of Freedman, Hass and Scott in [1] and of Hass and Rubinstein in [2] on the intersections and self-intersections of loops on a Riemannian surface, each of which is shortest in its homotopy class. We succeed in obtaining new proofs of the results in [1] and [2] on the self-intersections of a single loop, but it seems that the results on intersections of two loops cannot be proved in this way. We also give, at the end of §2, the first specific examples of filling curves on surfaces. A singular loop on a closed surface F is a *filling curve* if it cannot be homotoped into a proper incompressible subsurface of F.

We will now describe our main results and counter examples. We will consider the intersections of proper arcs on a surface as well as of loops. Throughout this paper, we will assume that all our arcs and loops are immersed and in general position. Let f_1 and f_2 be two proper arcs on a surface F, i.e., $f_i : I \to F$ and $F_i^{-1}(\partial F) = \partial I$. We say that f_1 and f_2 have *excess intersection* if they can be homotoped rel boundary to intersect in less points. We make similar definitions of excess self-intersection of a single arc and of excess intersection and self-intersection of loops.

If two proper arcs f_1 and f_2 have excess intersection, it is easy to show that there are sub-arcs λ_i of f_i such that the endpoints of λ_1 coincide with those of λ_2, and λ_1 is homotopic to λ_2 rel $\partial\lambda_1$. See Lemma 3.2. This can be thought of as asserting the existence of a singular 2-gon between f_1 and f_2. However, it is a surprising fact that no such result holds in the case of two singular loops on a surface. Figure 0.1 shows two examples of loops on the annulus with excess intersection but no (singular) 2-gon between them.

When we consider the self-intersections of a single arc or loop on a surface F, we again find some positive results and some surprising examples. First we note



(a) (b)

Fig. 0.1.

that the example of a figure-eight loop in the plane shows that a loop can have excess self-intersection without there being a 2-gon. Clearly, in this example, we have 1-gons instead. If f is a loop on a surface F, i.e. $f: S^1 \to F$, we say that a 2-disc D in F is an *embedded* 1-*gon for* f if there is a sub-arc α of S^1 such that $f(\alpha) = \partial D$ and $f \mid \operatorname{int} \alpha$ is injective. We say that a 2-disc D in F is an *embedded* 2-*gon for* f if there are disjoint sub-arcs α and β of S^1 such that f embeds α and β so that

$$f(\alpha) \cup f(\beta) = \partial D \quad \text{and} \quad f(\alpha) \cap f(\beta) = f(\partial\alpha) = f(\partial\beta).$$

We make similar definitions if f is an arc on F. Our first result is the following.

THEOREM 2.7. *Let f be a general position loop on an orientable surface F which is homotopic to an embedding but is not an embedding. Then there is an embedded 1-gon or 2-gon for f.*

REMARKS. We cannot expect that we can always find an innermost 1-gon or 2-gon D, i.e., that the image of f does not meet the interior of D. Figure 0.2 shows a singular loop in the plane for which there are no innermost 1-gons or 2-gons.

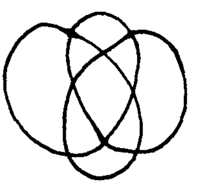
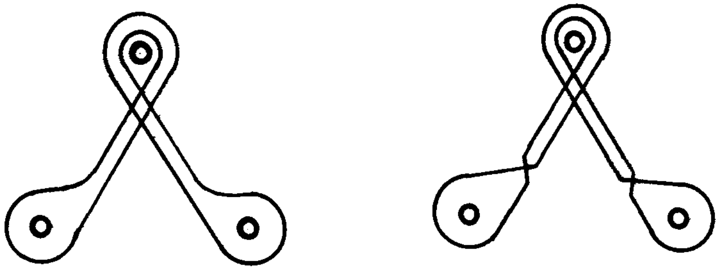


Fig. 0.2.

We are unable to decide whether the above theorem remains true if the orientability hypothesis is removed. However, we show that it does remain true in the case when F is not closed, and we prove a similar result for arcs on a surface F. See the end of §2 for a discussion of the problems here. Finally, the hypothesis that f be homotopic to an embedding is essential as the example in Figure 0.3 shows. This example suggests that one should be able to find a singular 1-gon or 2-gon for f, because a singular 2-gon is apparent in Fig. 0.3(b). If f is a general position loop on a surface F, we say that f has a *singular* 1-*gon* if there is a sub-arc α of S^1 such that f identifies the endpoints of α and $f \mid \alpha$ defines a null-homotopic loop on F. We say that f has a *singular* 2-*gon* if there are disjoint sub-arcs α and β of S^1 such that f identifies the endpoints of α and β and $f \mid (\alpha \cup \beta)$ defines a null-homotopic loop on F.



(a) (b)

Fig. 0.3.

THEOREM 4.2. *Let f be a general position loop on an orientable surface F. If f has excess self-intersection, then f has a singular* 1*-gon or* 2*-gon.*

REMARKS. A similar result holds for arcs. We are unable to decide whether any loop f on an orientable surface which possesses a singular 1-gon must possess an embedded 1-gon or 2-gon. The example in Fig. 0.4 shows that f need not possess an embedded 1-gon.

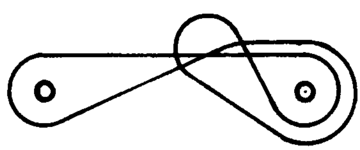
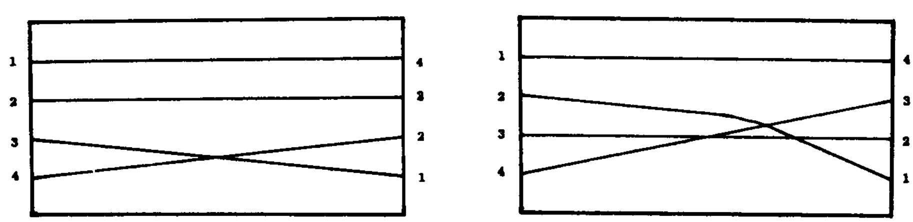


Fig. 0.4.

The hypothesis that F be orientable cannot be omitted from Theorem 4.2 as the example in Fig. 0.5 shows. The crucial point here is our insistence, in the definition of singular 2-gons, that the sub-arcs α and β of S^1 which give rise to a singular 2-gon must be disjoint. If one removes this condition, we finally arrive at a result which holds for any loop with excess self-intersection. We say that a loop



(a) (b)

Fig. 0.5.

f on a surface F has a *weak 2-gon* if there are sub-arcs α and β of S^1 such that f identifies the endpoints of α with those of β and the loop formed by $f \mid \alpha$ and $f \mid \beta$ is null-homotopic on F.

THEOREM 3.5. *Let f be a general position loop on a surface F. If f has excess self-intersection, then f has a singular 1-gon or a weak 2-gon.*

REMARKS. The converse of this result is false as consideration of a degree-two loop on the annulus shows. A similar result holds for arcs.

The organisation of this paper is as follows. In §1, we exhaustively analyse intersections of arcs and loops on the annulus and Moebius band. In §2, we consider arcs and loops which are homotopic to simple curves. In §3, we consider arbitrary arcs and loops on arbitrary surfaces, and then briefly explain how our results can be applied to give new proofs of results of [1] and [2]. Finally, in §4, we consider arcs and loops on orientable surfaces which may not be homotopic to simple curves.

The work in this paper was done while the second author was visiting the University of Michigan and was also partially supported by NSF grant 81-02469. We would also like to thank Jeff Parker for a helpful conversation.

§1. Curves on the annulus and Moebius band

Throughout this paper any map f of a 1-manifold Σ into a surface F will be assumed to be proper and in general position. Thus $f^{-1}(\partial F) = \partial \Sigma$ and the only singularities of f are double points. The number of double points of f is counted in F, i.e., it equals the cardinality of $\{x \in F : f^{-1}(x) \text{ is two points}\}$. We say that a loop f on a surface F represents α in $\pi_1(F)$ if f is freely homotopic to a based loop representing α. Thus f also represents any conjugate of α.

In this section, we analyse exhaustively what can be said about the existence of 1-gons and 2-gons when one has arcs or loops with excess intersection on the annulus $S^1 \times I$ or the Moebius band. The only result in this area which seems trivial is the following for curves on simply connected surfaces.

LEMMA 1.1. *If f is an arc or loop on a simply connected surface F such that f is not simple, then there is an embedded 1-gon for f.*

PROOF. Among all pairs (x_i, y_i) of points in the domain Σ of f for which $f(x_i) = f(y_i)$, choose an innermost pair (x_0, y_0). This means that there is a sub-arc α of Σ with endpoints x_0 and y_0 such that α contains no other pair (x_i, y_i). Clearly

α yields a simple loop on F and hence an embedded 1-gon for f, as any simple loop on F must bound a 2-disc.

Now we consider arcs and loops on the annulus $S^1 \times I$, which we denote by A. We let $\partial_0 A$ and $\partial_1 A$ denote the two boundary components of A.

LEMMA 1.2. *Let f be an arc on the annulus A joining $\partial_0 A$ to $\partial_1 A$. If f is not simple, then there is an embedded 1-gon or 2-gon for f.*

REMARK. Any such arc is homotopic, rel boundary, to a simple arc, so the hypothesis that f is not simple is equivalent to assuming that f has excess self-intersection.

PROOF. We will suppose that f has no embedded 1-gons or 2-gons and will deduce that f is simple.

Let λ be a simple arc in A joining $\partial_0 A$ to $\partial_1 A$ which meets the image of f transversely in the least possible number of points. We cut A along λ to obtain a 2-disc R with arcs λ_0 and λ_1 in ∂R corresponding to λ. The image of f is cut by λ into arcs μ_i in R. As f has no embedded 1-gons, it follows from Lemma 1.1 that each μ_i is simple. As f has no embedded 2-gons, it follows that each pair of μ_i's meets in at most one point. Let μ_0 be the sub-arc of f which meets $\partial_0 A$. If we suppose that λ meets f, then μ_0 must meet λ_0 or λ_1. We suppose that μ_0 meets λ_0 at X. See Fig. 1.3.

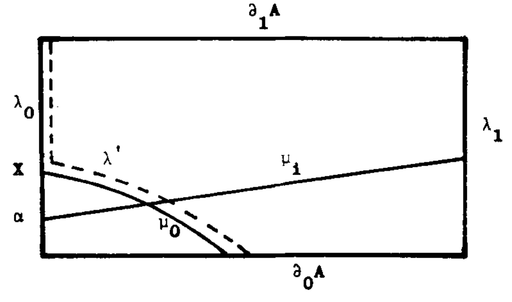


Fig. 1.3.

The point X cuts λ_0 into two sub-arcs and we let α denote the sub-arc with one endpoint on $\partial_0 A$. If some arc μ_i crosses μ_0, it must meet λ_0 in α because μ_0 and μ_i can meet only once. It follows that the arc λ' shown dotted in Fig. 1.3 meets f in less points than λ does, contradicting our definition of λ. It follows that λ does not meet f at all and hence that f equals the simple arc μ_0, completing the proof of the lemma.

LEMMA 1.4. *Let $f : S^1 \to A$ be a loop on the annulus representing a generator of $\pi_1(A)$. If f is not simple, then there is an embedded* 1-*gon or* 2-*gon for f.*

PROOF. As in the previous lemma, we will suppose that f has no embedded 1-gon or 2-gon and will show that f is simple. We again let λ be a simple arc joining $\partial_0 A$ to $\partial_1 A$ which meets the image of f transversely in the least possible number of points. We let R denote the 2-disc obtained by cutting A along λ and let λ_0 and λ_1 be the arcs in ∂R corresponding to λ. Note that λ must meet f as otherwise f would be null-homotopic. Hence λ cuts the image of f into arcs μ_i with boundary in $\lambda_0 \cup \lambda_1$. As in the previous lemma, each μ_i is simple and each pair of μ_i's meets in at most one point. If each arc μ_i joins λ_0 to λ_1, then the number of μ_i's equals the degree of f. In this case, there must be only one μ_i and so it follows at once that f is simple. Otherwise, one of the μ_i's has both endpoints in the same λ_j. We suppose that μ_0 has both endpoints in λ_0 as in Fig. 1.5, and let α denote the sub-arc of λ_0 with the same endpoints as μ_0. If some arc μ_i crosses μ_0, it must meet λ_0 in α, because μ_0 and μ_i can meet only once. It follows that the arc λ' shown dotted in Fig. 1.5 meets f in less points than λ does, contradicting our definition of λ. This contradiction completes the proof of the lemma.

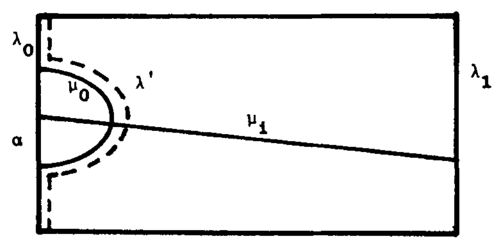


Fig. 1.5.

At this point, we note that the above arguments can equally well be applied to a loop on the Moebius band M which represents a generator of $\pi_1(M)$. One chooses λ to be a non-separating simple arc in M which meets the image of f transversely in the least possible number of points. The result obtained is the following.

LEMMA 1.6. *Let $f : S^1 \to M$ be a loop on the Moebius band M representing a generator of $\pi_1(M)$. If f is not simple, then there is an embedded* 1-*gon or* 2-*gon for f.*

Now we return to our consideration of curves on the annulus.

LEMMA 1.7. *Let λ and μ be two simple arcs on the annulus A each joining $\partial_0 A$ to $\partial_1 A$. If λ and μ are in general position and there are no embedded 2-gons between λ and μ, then the configuration of λ and μ in A is unique up to an ambient isotopy of A which fixes ∂A.*

REMARK. In particular λ and μ intersect in the least possible number of points obtainable by homotoping λ or μ rel boundary.

PROOF. We cut A along λ to obtain a 2-disc R with arcs λ_1 and λ_2 in ∂R, corresponding to λ. As μ is simple, the intersection of μ with R is a collection of disjoint simple arcs μ_i. As there are no embedded 2-gons between λ and μ, no μ_i can have both endpoints on λ_1 or on λ_2. If we choose a path μ' in R joining the endpoints of μ, it is clear that the number of arcs μ_i equals one more than the degree of the loop $\mu \cup \mu'$ on A. The required uniqueness result follows. See Fig. 1.8 for the configuration in R when $\mu \cup \mu'$ has degree 4.

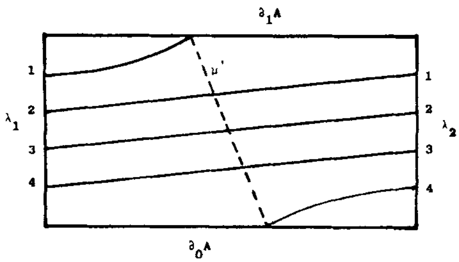


Fig. 1.8.

LEMMA 1.9. *Let $f:(I, \partial I) \to (A, \partial_0 A)$ be a general position arc on A. If f has no embedded 1-gons or 2-gons, the configuration of the image of f is unique up to an ambient isotopy of A which fixes ∂A.*

REMARK. In particular, f has least possible self-intersection.

PROOF. Let α be a simple arc in A with one endpoint on $\partial_1 A$, the other endpoint on $f(I)$ and with its interior disjoint from $f(I)$. Using α, we obtain two new arcs λ and μ from f as shown in Fig. 1.10. As f has no embedded 1-gons or 2-gons, neither do λ and μ. Also there are no embedded 2-gons between λ and μ. Now Lemma 1.2 implies that λ and μ are both simple, and then Lemma 1.7 implies that the configuration of λ and μ in A is unique. There are only two possibilities for the configuration of f, which are shown in Fig. 1.11 in the case when λ and μ are as in Fig. 1.8. As the first possibility implies that f has an embedded 1-gon, the uniqueness result follows.

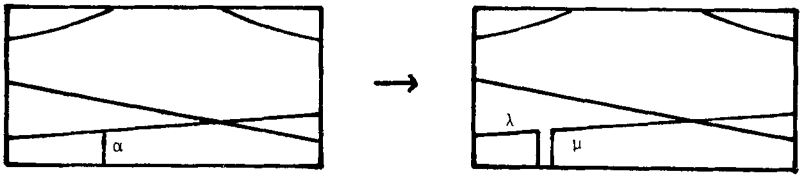


Fig. 1.10.

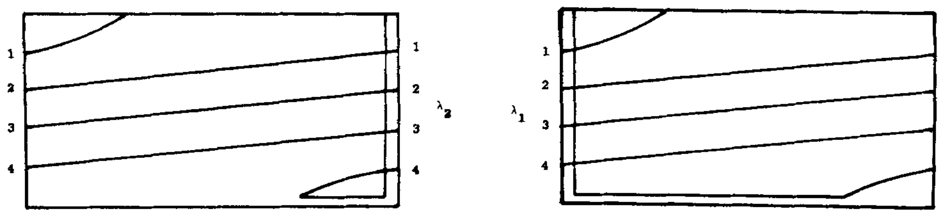


Fig. 1.11.

LEMMA 1.12. *Let $f: S^1 \to A$ be a general position loop on the annulus of degree d. If f has no embedded 1-gons or 2-gons, then the configuration of f in A is unique up to ambient isotopy of A, and f has $(d-1)$ double points.*

PROOF. We argue as in Lemma 1.9. Join $f(S^1)$ to $\partial_0 A$ by a simple arc α whose interior does not meet $f(S^1)$. This yields an arc $l:(I, \partial I) \to (A, \partial_0 A)$. As f has no embedded 1-gons or 2-gons, neither has l. Hence, by Lemma 1.9, the configuration of l is unique up to an ambient isotopy of A fixing ∂A. As in Lemma 1.9, there are two possible configurations for f, one of which contains an embedded 1-gon. The uniqueness result follows, and the configuration in the case d equals 4 is shown in Fig. 1.13. Clearly, f will have $(d-1)$ double points.

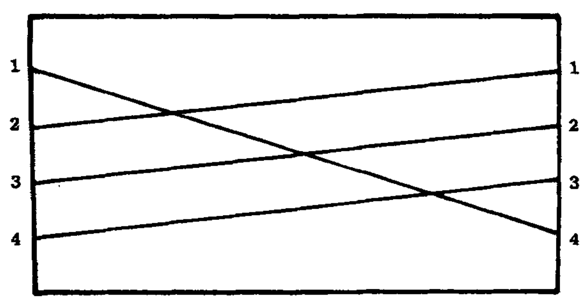


Fig. 1.13.

LEMMA 1.14. *Let f_0 and f_1 be arcs in $(A, \partial_0 A)$ and $(A, \partial_1 A)$ respectively which are in general position. If f_0 and f_1 have no embedded 1-gons or 2-gons and if there are no embedded 2-gons between f_0 and f_1, then f_0 and f_1 must be disjoint.*

PROOF. First suppose that f_0 is simple. Then f_0 separates A and one of the components obtained by cutting A along f_0 is a 2-disc R. If f_1 meets f_0, some sub-arc λ of f_1 must lie in R. By Lemma 1.1, λ must be simple and thus R must contain an embedded 2-gon D bounded by λ and some sub-arc of f_0. This contradiction shows that f_0 and f_1 must be disjoint, when f_0 is simple.

Now suppose that f_0 is not simple. Lemma 1.9 tells us that the configuration of f_0 in A is unique up to ambient isotopy. The image of f_0 cuts A into various regions. We let σ denote the simple loop in $f_0(I)$ which lies in the boundary of the region containing $\partial_1 A$, and let τ denote the simple sub-arc of f_0 shown in Fig. 1.15. (If f_0 has only one double point, we have two choices for τ.) Let R denote the 2-disc obtained from A by cutting along σ and removing a regular neighbourhood of τ. Thus ∂R contains an arc corresponding to σ and two arcs τ_0 and τ_1 corresponding to τ, as shown in Fig. 1.15.

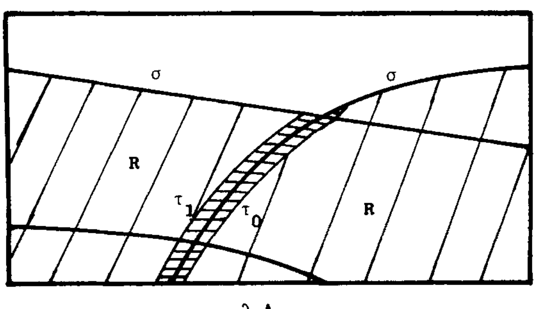


Fig. 1.15.

If f_0 and f_1 are not disjoint, then $f_1(I)$ must meet R. As f_1 has no embedded 1-gons, f_1 will meet R in a collection of simple arcs μ_i. As there are no embedded 2-gons between f_1 and f_2 none of these arcs can have both endpoints in σ or both in τ_0 or both in τ_1. The number of endpoints of μ_i's on τ_0 must equal the number on τ_1, as both are equal to the number of times $f_1(I)$ meets τ. Hence the number of μ_i's joining σ to τ_0 must equal the number of μ_i's joining σ and τ_1. It follows that some μ_i has its endpoints on σ and τ_0. Clearly, this yields an embedded 2-gon between f_0 and f_1. This contradiction shows that f_0 and f_1 must be disjoint as required.

Note that the above proof did not use the hypothesis that f_1 has no embedded 2-gons.

Finally, we consider the self-intersection of a single loop on the Moebius band M. As discussed in the introduction, a loop on M with excess self-intersection need not have an embedded 1-gon or 2-gon or even a singular 1-gon or 2-gon, but we will show that it must have a singular 1-gon or a weak 2-gon. Before proving that, it is interesting to see how easy it is to prove such a result for the annulus.

LEMMA 1.16. *Let f be a general position loop on the annulus A. If f is not simple, then f has an embedded 1-gon or a weak 2-gon.*

REMARK. Note that f need not have excess self-intersection, so this result does not follow from Lemma 1.12.

PROOF. We will suppose that f has no embedded 1-gon or 2-gon and will show that f must possess a weak 2-gon. Let γ be a simple arc in A joining the two boundary components and chosen to intersect f in the least possible number of points. Then γ cuts A into a 2-disc R whose boundary contains two arcs γ_1 and γ_2 corresponding to γ. If $f \cap \gamma$ is empty, Lemma 1.1 shows that f has an embedded 1-gon which is a contradiction. Otherwise, $f \cap R$ consists of arcs λ_i each of which must be simple. Further, the usual argument (see Lemma 1.4) shows that each λ_i must join γ_1 to γ_2. As f is not simple, there must be at least two λ_i's and a point x at which two of the λ_i's intersect. We can now obtain a weak 2-gon for f, i.e., sub-arcs λ and μ of S^1 such that $f(\lambda)$ is homotopic to $f(\mu)$ rel boundary. We find λ and μ by starting at x and following the two sub-arcs of f from x to γ_2 and continuing to follow these arcs around A until they cross again. Of course, these arcs will cross at some point because eventually, they will each traverse all of the circle and both arrive back at x. However, if they failed to cross before traversing the whole circle, we would have a problem as then the "arcs" described will each be all of S^1. But, after traversing the whole circle our two arcs must have interchanged sides, as they cross at x and the orientability of A implies that they must have crossed at some intermediate point. This completes the proof of Lemma 1.16.

LEMMA 1.17. *Let f be a loop with excess self-intersection on the Moebius band M. Then f has a singular 1-gon or weak 2-gon.*

PROOF. If f is an orientation preserving loop, the arguments of the previous lemma apply unchanged to show that f has an embedded 1-gon or weak 2-gon.

As in Lemma 1.16, this result does not use the fact that f has excess self-intersection but only uses the fact that f is not simple.

If f is orientation reversing, we must use the hypothesis that f has excess self-intersection as Fig. 1.18 shows. In this case, we consider the loop $\tilde{f}: \tilde{S} \to A$ where A is the double covering of M and $\tilde{S}$ is the double covering of the circle S. We let τ denote the covering involution on A. As f has excess self-intersection, so does $\tilde{f}$. Hence, by Lemma 1.12, the loop $\tilde{f}$ has an embedded 1-gon or 2-gon. If $\tilde{f}$ has an embedded 1-gon D, we let λ denote the sub-arc of $\tilde{S}$ such that $\tilde{f}(\lambda)$ is the boundary of D. If λ projects to a sub-arc of S, then this sub-arc defines a singular 1-gon for f. Otherwise, the image of λ must equal S, so that $\lambda \cup \tau\lambda = \tilde{S}$. As $\tilde{f}$ is injective on the interiors of λ and $\tau\lambda$ and identifies the endpoints of λ and $\tau\lambda$, the only possibility is that λ and $\tau\lambda$ intersect only in the two points of $\partial\lambda$. But this implies that $\tilde{f}$ is null-homotopic as $\tilde{f}(\lambda)$ and $\tilde{f}(\tau\lambda)$ are null-homotopic loops. This contradiction shows that λ must project to a sub-arc of S as required.

If $\tilde{f}$ has an embedded 2-gon D, we let λ and μ denote the sub-arcs of $\tilde{S}$ which form the boundary of D. These arcs yield a weak 2-gon for f unless the image in S of λ or μ is all of S. Suppose that the image of λ equals S, so that $\lambda \cup \tau\lambda = \tilde{S}$. As λ is disjoint from μ, we must have $\mu \subset \tau\lambda$. But λ and $\tau\lambda$ must overlap, so it follows that $\tilde{f}(\mu)$ meets $\tilde{f}(\tau\lambda)$ and hence that $\tilde{f}(\tau\lambda)$ is a singular arc. As $\tilde{f}(\lambda)$ is simple, this is a contradiction and so completes the proof of Lemma 1.17.

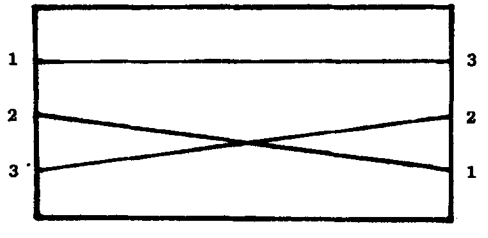


Fig. 1.18.

§2. Curves homotopic to simple curves

We start by considering arcs which are homotopic to simple arcs. We obtain the following result.

THEOREM 2.1. *Let f be a general position arc on a surface F such that f is homotopic rel boundary to a simple arc g on F, but f is not simple. Then f has an embedded 1-gon or 2-gon.*

Before proving this, we will need the following special case.

LEMMA 2.2. *Let f be a general position arc on a surface F such that f is homotopic rel boundary to a simple arc g in ∂F. If f has no embedded 1-gons or 2-gons, then f is simple.*

PROOF. If F is not compact, we can replace F by a compact subsurface which contains the image of the homotopy from f to g. Hence we may assume that F is compact. We will argue by induction on the Euler number $\chi(F)$ of F. As F has non-empty boundary, $\chi(F) \leqq 1$. Further $\chi(F) = 1$ only when F is the 2-disc. In this case, the result of Lemma 2.2 follows from Lemma 1.1.

Now suppose that $\chi(F) < 1$. Then there is a non-separating simple arc γ in F such that γ is disjoint from the simple arc g in ∂F. We choose γ within its isotopy class rel boundary so as to minimise the number of points in $f \cap \gamma$. If $f \cap \gamma$ is empty, then f lies in the surface F' obtained by cutting F along γ and f is homotopic in F' to the simple arc g in $\partial F'$. Hence by our induction hypothesis f is simple and the induction step is complete. It remains to show that $f \cap \gamma$ must be empty.

Let $H: I \times I \to F$ be a homotopy between the arcs f and g with $H_0 = f$, $H_1 = g$. We can arrange that H is transverse to γ and then consider the 1-submanifold $H^{-1}(\gamma)$ of $I \times I$. As g and γ are disjoint, $H^{-1}(\gamma)$ consists of circles and of arcs with endpoints in $I \times \{0\}$. Any circles in $H^{-1}(\gamma)$ can be eliminated by a homotopy of H. It follows that if f meets γ, then there is a sub-arc λ of I such that $f(\lambda)$ meets γ only in its endpoints and $f \mid \lambda$ is homotopic rel $\partial\lambda$ into γ. Further, this homotopy of $f \mid \lambda$ takes place in the surface F'. It follows that $f \mid \lambda$ is an arc in F' which is homotopic rel $\partial\lambda$ to a simple arc in $\partial F'$. Now $f \mid \lambda$ has no embedded 1-gons or 2-gons as f has none. Thus our induction hypothesis implies that $f \mid \lambda$ is simple. As in §1, it now follows that we can find an arc γ' isotopic to γ but meeting f in less points as shown in Fig. 2.3. This uses the fact that any arc μ of $f(I)$ which meets λ must also meet γ in the sub-arc with endpoints $\partial\lambda$, as shown in Fig. 2.3. For otherwise, we would find an embedded 2-gon for f. The existence of γ' contradicts our choice of γ. Thus we deduce that $f \cap \gamma$ is empty, completing the proof of Lemma 2.2.

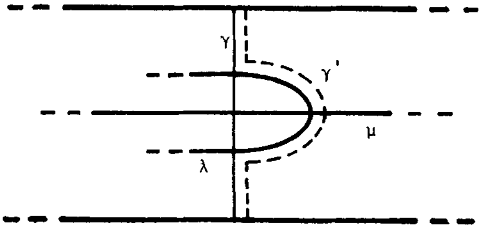


Fig. 2.3.

Now we can prove Theorem 2.1 which we restate.

THEOREM 2.1. *Let f be a general position arc on a surface F such that f is homotopic rel boundary to a simple arc g on F, but f is not simple. Then f has an embedded 1-gon or 2-gon.*

PROOF. We will assume that f has no embedded 1-gons or 2-gons and will show that f is simple. If the arc g is inessential in F, i.e. is parallel to an arc in ∂F, then Lemma 2.2 applies to show that f is simple. If g is essential in F, we choose a simple arc γ in F such that γ can be isotoped rel boundary to be parallel to g, and γ meets f transversely in the least possible number of points. We can arrange that g is disjoint from γ, by isotoping g if necessary. As f can be homotoped rel boundary to be disjoint from γ, the proof of the induction step of Lemma 2.2 shows that $f \cap \gamma$ must be empty. Let F' denote the surface obtained from F by cutting along γ. Then f is homotopic in F' to the simple arc g which is in turn parallel to an arc in $\partial F'$. As f has no embedded 1-gons or 2-gons, Lemma 2.2 implies that f is simple and the proof of Theorem 2.1 is complete.

If f is a loop on a surface F which is homotopic to a simple loop, we use similar arguments to obtain the following result.

THEOREM 2.4. *Let f be a general position loop on a surface F with non-empty boundary. If f is homotopic to a simple loop g but is not simple, then f has an embedded 1-gon or 2-gon.*

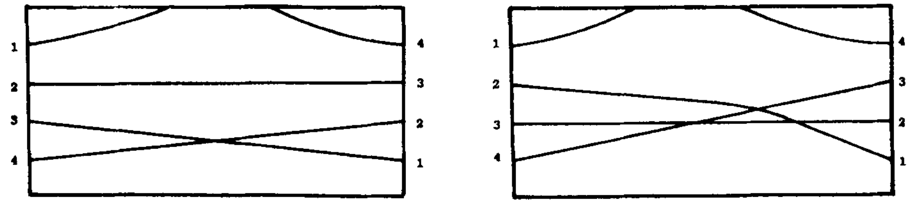
PROOF. As usual, we will assume that F is compact and that f has no embedded 1-gons or 2-gons and will show that f is simple, by induction on the Euler number $\chi(F)$ of F. If $\chi(F)$ equals 1, then F is the 2-disc and the result follows by Lemma 1.1. Now suppose that $\chi(F) \leqq 0$. If there is a non-separating simple arc δ in F disjoint from g, we choose an arc γ isotopic rel boundary to δ so as to meet f in the least possible number of points. We can suppose that g is disjoint from γ, by isotoping g if necessary. As f can be homotoped to be disjoint from γ, the proof of the induction step of Lemma 2.2 shows that $f \cap \gamma$ must be empty. Hence f lies in the surface F' obtained by cutting F along γ. Now our induction hypothesis implies that f is simple as required.

If there is no such arc δ in F, then F has connected boundary C and g is parallel to C. We can find a non-separating simple arc δ in F such that δ meets g in exactly two points. As usual, we choose γ isotopic rel boundary to δ so as to meet f in the least possible number of points and let F' denote the surface obtained by cutting F along γ. Now the arguments of Lemma 2.2 show that f must meet γ in two points. Thus f is cut by γ into two sub-arcs λ and μ. Further the homotopy f_t from f to g can be chosen so that γ meets f_t in two points, for each t. Hence the arcs λ and μ are each homotopic to simple arcs in $\partial F'$ with the homotopy taking place in F', and during the homotopy $\partial\lambda$ and $\partial\mu$ remain in γ. It follows that each of λ and μ is homotopic rel boundary to a simple arc in $\partial F'$. Hence Lemma 2.2 implies that λ and μ are both simple. Now λ together with an arc in $\partial F'$ bounds a 2-disc R. If λ meets μ, then some sub-arc of μ lies in R and yields an embedded 2-gon D between λ and μ. But D would then be an embedded 2-gon for f which contradicts our hypothesis. Hence λ and μ are disjoint arcs in F' and so f is a simple loop in F as required.

Before we can consider loops on closed surfaces, we need the following generalisation of Lemma 2.2.

Lemma 2.5. *Let f be a general position arc on an orientable surface F such that f is homotopic rel boundary into ∂F. If f has no embedded* 1*-gons or* 2*-gons, then f lies in some collar neighbourhood of ∂F.*

Remark. This result is false if we omit the orientability hypothesis on F. For the arc f on the Moebius band M shown in Fig. 2.6(b) has no embedded 1-gons or 2-gons and yet cannot lie in a collar neighbourhood of ∂M. This is because f has excess double points, as is shown by the homotopic arc g in Fig. 2.6(a), and hence if f lay in an annulus, Lemma 1.9 would force f to have an embedded 1-gon or 2-gon.



(a) (b)

Fig. 2.6.

Proof of Lemma 2.5. As usual, we assume that F is compact and argue by induction on the Euler number $\chi(F)$ of F. If $\chi(F)$ equals 1, then F is D^2 and the result is trivial. If $\chi(F)$ equals 0, then F is the annulus and the result is again trivial. (This uses the orientability of F.) Now suppose that $\chi(F)<0$. Let C denote the component of ∂F which contains ∂f. Then there is a non-separating simple arc or loop γ on F which is disjoint from C. We choose γ to meet f in the least possible number of points. If $f \cap \gamma$ is empty, then f lies in the surface F'

obtained by cutting F along γ, and f is homotopic rel boundary into $\partial F'$. Hence our induction hypothesis implies that f lies in a collar neighbourhood of C as required. It remains to show that $f \cap \gamma$ must be empty. If γ is an arc, this follows from the proof of Lemma 2.2, so we suppose now that γ is a circle.

Let $H : I \times I \to F$ be a homotopy with $H_0 = f$ and $H_1(I) \subset C$. We can assume that H is transverse to γ and then consider the 1-submanifold $H^{-1}(\gamma)$ of $I \times I$. As C and γ are disjoint, $H^{-1}(\gamma)$ consists of circles and of arcs with endpoints in $I \times \{0\}$. We can homotop H to eliminate all circles from $H^{-1}(\gamma)$. It follows that if f meets γ, then there is a sub-arc λ of I such that $f(\lambda)$ meets γ only in its endpoints and $f \mid \lambda$ is homotopic rel $\partial\lambda$ into γ. Further if F' denotes the surface obtained from F by cutting along γ, then this homotopy takes place in F'. It follows that $f \mid \lambda$ is an arc in F' which is homotopic rel $\partial\lambda$ into $\partial F'$. Hence our induction hypothesis implies that $f(\lambda)$ lies in a collar neighbourhood of $\partial F'$. Let A be an annulus in F with γ as one of its boundary components such that $f(\lambda)$ lies in A, and let γ_1 denote the other component of ∂A. We know that $f \cap A$ consists of arcs such that each arc has no embedded 1-gons and 2-gons and there are no embedded 2-gons between any two of these arcs. Hence Lemma 1.14 implies that any arc of $f \cap A$ with both endpoints on γ_1 must be disjoint from $f(\lambda)$. Thus, by choosing γ_1 closer to γ, so that A is thinner, we can suppose that no arc of $f \cap A$ has both endpoints on γ_1. It follows that γ_1 meets f in less points than γ does which contradicts our initial choice of γ. This contradiction shows that $f \cap \gamma$ must be empty and completes the proof of Lemma 2.5.

Using this result, we can prove the following.

THEOREM 2.7. *Let f be a general position loop on an orientable surface F. If f is homotopic to a simple loop g but is not simple, then f has an embedded 1-gon or 2-gon.*

PROOF. This result follows from Theorem 2.4 if F is not closed, and follows from Lemma 1.1 if F is the 2-sphere. So we will suppose that F is closed but not S^2. As usual, we will assume that f has no embedded 1-gons and 2-gons and will show that f is simple. We choose γ to be an essential simple loop on F disjoint from g which meets f in the least possible number of points. (If g is essential, we can choose γ parallel to g.) Let F' denote the surface obtained by cutting along γ. If $f \cap \gamma$ is empty, then f is a loop in F' and is homotopic in F' to a simple loop. Hence, Theorem 2.4 implies that f is simple as required. If $f \cap \gamma$ is not empty, the fact that f can be homotoped to be disjoint from γ implies that there is a sub-arc of I such that $f(\lambda)$ meets γ only in its endpoints and $f \mid \lambda$ is homotopic rel $\partial\lambda$ into γ. Further this homotopy takes place in F'. Thus $f \mid \lambda$ is an arc in F'

which is homotopic rel boundary into $\partial F'$. Hence Lemma 2.5 implies that $f(\lambda)$ lies in a collar neighbourhood of $\partial F'$. Now the proof of Lemma 2.5 shows that there is a circle γ_1 in F which is parallel to γ and meets f in less points. This contradicts the choice of γ and shows that $f \cap \gamma$ must be empty. This completes the proof of Theorem 2.7.

The results proved so far say nothing about loops on a closed non-orientable surface, and we are unable to decide whether Theorem 2.7 holds for all such surfaces. However, certain cases can be handled exactly as in the orientable case. For example, let f be a general position loop on a closed non-orientable surface F such that f is homotopic to a simple loop g and f has no embedded 1-gons or 2-gons. Suppose that there is a simple 2-sided loop γ on F such that γ is disjoint from g and the surface F' obtained from F by cutting along γ is orientable. Then the arguments of Theorem 2.7 apply unchanged to show that f is simple. Unfortunately such a loop γ need not always exist, even when the surface involved is the Klein bottle. However, we can show that Theorem 2.7 holds for loops on the Klein bottle, using the fact that there are only four isotopy classes of simple loops on the Klein bottle. One can also show easily that Theorem 2.7 holds for loops on the projective plane P^2.

The arguments used in the proof of Theorem 2.7 and earlier results prove the following statement about loops which need not be homotopic to simple loops.

LEMMA 2.8. *Let f be a loop on an orientable surface F which does not have embedded 1-gons or 2-gons. Let γ be an essential simple arc or loop on F and suppose that f can be homotoped to be disjoint from γ. Then there is a simple loop γ' isotopic to γ such that f is disjoint from γ'.*

PROOF. Choose γ' isotopic to γ and with least possible intersection with f.

COROLLARY 2.9. *Let f be a loop on an orientable surface F which does not have embedded 1-gons or 2-gons. If each component of the surface obtained from F by cutting along $f(S^1)$ is a 2-disc, then f cannot be homotoped into any incompressible subsurface F_1 of F unless $\pi_1(F_1) = \pi_1(F)$.*

PROOF. If the result is false, there is an essential simple arc or loop γ on F such that f can be homotoped disjoint from γ. Lemma 2.8 shows that f is already disjoint from γ, contradicting our hypothesis.

As a consequence of Corollary 2.9, we can exhibit a loop f on the closed orientable surface F_g of genus g, such that f is not homotopic into any proper incompressible subsurface, so long as $g \geqq 2$. Note that in the case $g = 1$, any loop on the torus is homotopic into some sub-annulus. We start with two arcs, as shown in Fig. 2.10, on the surface X obtained by removing the interior of a 2-disc from the torus. One arc is simple, the other not. There are no embedded 1-gons or 2-gons for these arcs. These arcs divide X into five regions each of which is a 2-disc and each region meets ∂X in a single arc or the empty set. We then assemble F_g from g copies of X and a g-holed sphere Y, and construct the required loop f on F_g so that its intersection with each copy of X is as in Fig. 2.10. If g is odd, the intersection of f with Y is as shown in Fig. 2.11. If g is even, the intersection of f with Y is as shown in Fig. 2.12, which is the same as Fig. 2.11 apart from the single double point. As $f \cap Y$ cuts Y into 2-discs and as $f \cap X$ cuts X into 2-discs each meeting ∂X in at most one interval, it follows that f cuts F_g into 2-discs. If f has an embedded 1-gon or 2-gon D, it cannot lie in any copy

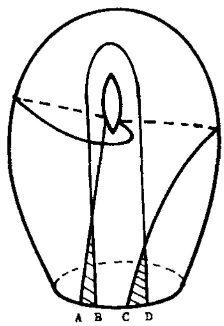


Fig. 2.10.

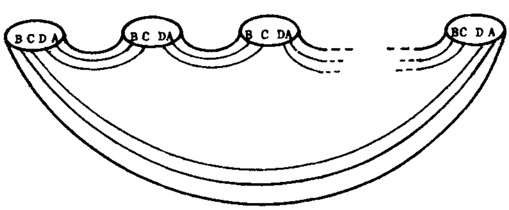


Fig. 2.11.

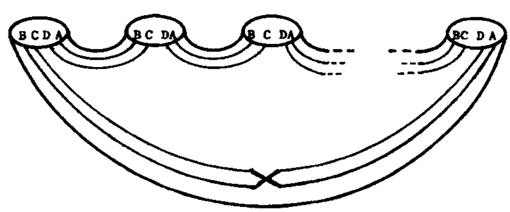


Fig. 2.12.

of X nor in Y by our construction. Hence D must meet ∂Y. If D is a 1-gon, this yields an embedded 2-gon D' with one edge in ∂Y and the other edge a sub-arc of f, which is impossible. If D is a 2-gon, we denote its edges by λ and μ. No arc of $\partial Y \cap D$ can have both endpoints in λ or both in μ as this would again yield an impossible 2-gon D'. Hence every arc of $\partial Y \cap D$ joins λ to μ. Hence ∂Y cuts D into a triangle at each end and some quadrilaterals. Clearly any quadrilaterals must lie in Y so that ∂Y meets D in only one or two arcs. In any case, one of the triangles lies in one of the copies of X and so must be one of the two triangles shaded in Fig. 2.10. Now the way that $f \cap Y$ joins up the various copies of X yields a contradiction in all cases. We conclude that the loop f described in Fig. 2.11 or 2.12 has no embedded 1-gons or 2-gons, so that Corollary 2.9 can be applied to f as required.

§3. The weak results

The arguments in this section have their origin in the proof of the following well-known result whose proof we give for completeness.

LEMMA 3.1. *Let C_1 and C_2 each be a simple arc or loop on a surface F. If C_1 and C_2 have excess intersection, then there is an innermost embedded 2-gon between them, i.e., a 2-gon whose interior does not meet C_1 or C_2.*

PROOF. We consider the pre-images $\tilde{C}_1$ and $\tilde{C}_2$ of C_1 and C_2 in the universal convering $\tilde{F}$ of F. The fact that C_1 can be homotoped rel ∂C_1 so as to meet C_2 in less points implies that there are components S and T of $\tilde{C}_1$ and $\tilde{C}_2$ respectively such that $S \cap T$ consists of more than one point. Hence there is a 2-gon D in $\tilde{F}$ between S and T. This 2-gon need not be innermost, i.e., its interior may meet $\tilde{C}_1$ or $\tilde{C}_2$, but if it is not innermost, there must be a smaller 2-gon inside. This is because no component of $\tilde{C}_1$ or $\tilde{C}_2$ can cross both S and T as this would contradict the fact that C_1 and C_2 are simple curves in F. It follows that there is an innermost 2-gon B in $\tilde{F}$ between some component S_1 of $\tilde{C}_1$ and some component S_2 of $\tilde{C}_2$. We will show that B projects to an embedded 2-gon in F between C_1 and C_2 by showing that B is disjoint from all translates gB, where g is a non-trivial element of $\pi_1(F)$.

Denote the two vertices of B by x and y and suppose that g is a non-trivial element of $\pi_1(F)$ such that gB meets B. Then gx or gy equals x or y. As g cannot fix any point of $\tilde{F}$, we must have $gx = y$ or $gy = x$. It follows that $gS_1 = S_1$ and $gS_2 = S_2$, as S_1 and S_2 are the only components of $\tilde{C}_1$ and $\tilde{C}_2$ passing through x or y. Hence C_1 and C_2 must be loops on F which represent g and intersect in a single point. It follows that g must be orientation reversing. But this means that C_1 and C_2 do not have excess intersection which contradicts our hypothesis. This completes the proof of Lemma 3.1.

It is natural to ask if there is a similar result when two singular arcs or loops have excess intersection. However, the examples in Fig. 0.1 show that there is no such result in the case of two singular loops and similar examples show that there is no such result in the case of a singular arc and loop with excess intersection. In the case of two singular arcs, it is easy to prove the following analogue of Lemma 3.1.

LEMMA 3.2. *Let f and g be two arcs on a surface F which are in general position. If f and g have excess intersection, there is a singular 2-gon between them.*

PROOF. The fact that f and g have excess intersection implies that there are lifts $\tilde{f}$ and $\tilde{g}$ of f and g to the universal covering $\tilde{F}$ of F such that $\tilde{f}$ intersects $\tilde{g}$ in more than one point. If we let x and y denote two points of $\tilde{f} \cap \tilde{g}$ and let λ_1 and λ_2 denote the sub-arcs of $\tilde{f}$ and $\tilde{g}$ respectively with endpoints x and y, then λ_1 is homotopic to λ_2 rel $\partial\lambda_1$ because $\tilde{F}$ is simply connected. Hence λ_1 and λ_2 project to sub-arcs of f and g which define a singular 2-gon between them.

For the rest of this section, we will consider the self-intersections of a single arc or loop. Our first, rather trivial, result is the following.

LEMMA 3.3. *Let f be a general position arc on a surface F. If f has excess self-intersection then f has a singular 1-gon or a weak 2-gon.*

PROOF. Let $\tilde{f}$ denote a lift of f to the universal covering $\tilde{F}$ of F. If $\tilde{f}$ is singular, then Lemma 1.1 implies that $\tilde{f}$ has an embedded 1-gon. This will project to a singular 1-gon for f. If $\tilde{f}$ is simple, the fact that f has excess self-intersection implies that for some $\alpha \in \pi_1(F)$, $\alpha\tilde{f}$ meets $\tilde{f}$ in more than one point. Thus there is an embedded 2-gon in $\tilde{F}$ between $\tilde{f}$ and $\alpha\tilde{f}$. This will project to a weak 2-gon for f, completing the proof of Lemma 3.3.

Our main result about self-intersections of loops is more complicated to prove. First we need the following result.

LEMMA 3.4. *Let f be a general position loop on a surface F, not S^2 or P^2, and suppose that f represents $\beta = \alpha^d$, where α is a non-trivial and primitive element of $\pi_1(F)$. Let $\tilde{F}$ denote the universal covering of F and let F_α denote the quotient of $\tilde{F}$ by the cyclic subgroup of $\pi_1(F)$ generated by α. Let $f_\alpha : S^1 \to F_\alpha$ be the lift of f, and let l denote one of the lines in $\tilde{F}$ above $f_\alpha(S^1)$. Then f has least possible self-intersection if and only if f_α has least possible self-intersection and, for all g in*

$\pi_1(F)$ such that g is not a power of α, the intersection $gl \cap l$ consists of at most one point.

REMARK. It is easy to prove a similar result for arcs on a surface.

PROOF. First we note that there is a general position loop f' homotopic to f such that f'_α has least self-intersection and, for all g in $\pi_1(F)$ such that g is not a power of α, the intersection $gl' \cap l'$ consists of at most one point. To obtain f', we first choose a complete hyperbolic or Euclidean metric on F, as appropriate, and take a geodesic loop representing β. If $d = 1$ we take f' to be this loop. If $d > 1$, we choose a loop f'_α in F_α which is close to the geodesic loop just described and which represents β and has least possible self-intersection. The projection of this loop to F is the required loop f'.

Now we consider the picture in F_β, the quotient of $\tilde{F}$ by the cyclic group generated by β. We let f'_β denote the lift of f'_α to F_β. The loop f'_β is the image of a simple line l' in $\tilde{F}$. The lines gl', for g in $\pi_1(F)$, project to possibly singular lines or loops in F_β. Our choice of f' shows that each such singular line meets the loop f'_β in at most one point. Further each singular loop is disjoint from f'_β except for those loops which are the images of the line $\alpha l', \alpha^2 l', \dots, \alpha^{d-1} l'$. (Unless F is the torus or Klein bottle, these are the only loops which can occur.) Now any homotopy of the loop f' induces a proper homotopy of all these singular lines and circles. In particular, if gl' meets l' in a single point, they must continue to meet after any homotopy of f', because a line in F_β which meets f'_β once cannot be properly homotoped to be disjoint from f'_β. It follows that f' has the least possible self-intersection. Further, if f also has least possible self-intersection then $gl \cap l$ consists of a single point if and only if $gl' \cap l'$ does. Thus it follows that f_α has the same number of double points as f'_α. Now it follows that f has least self-intersection if and only if f_α has least self-intersection and, for all g in $\pi_1(F)$ such that g is not a power of α, the intersection $gl \cap l$ consists of at most one point.

Now we come to the main result of this section.

THEOREM 3.5. *If f is a general position loop on a surface F such that f has excess self-intersection, then f has a singular 1-gon or a weak 2-gon.*

PROOF. We divide the proof into cases.

Case 1. f is null-homotopic or F is the projective plane P^2

If f is null-homotopic, then f has a lift $\tilde{f}$ to the universal cover $\tilde{F}$ of F. If $\tilde{f}$ is singular, it has an embedded 1-gon by Lemma 1.1, and this will descend to a singular 1-gon for f. If $\tilde{f}$ is simple, there must be an element g of $\pi_1(F)$ such that $g\tilde{f}$ meets $\tilde{f}$, because f is singular. Hence there is an embedded 2-gon between $\tilde{f}$ and $g\tilde{f}$. This will descend to a weak 2-gon for f, so that we have the required result when f is null-homotopic.

If f is an essential loop on P^2 which is not simple, we consider the covering map $\tilde{f}$ into S^2. This cannot be simple and so has an embedded 1-gon. Now the arguments in the proof of Lemma 1.17 show that this yields a singular 1-gon for f as required.

From now on we will suppose that f is essential.

Case 2. *F is not closed*

As usual, we can suppose that F is compact. We suppose that f represents $\beta = \alpha^d$ where α is a non-trivial, primitive element of $\pi_1(F)$. We will also assume that f does not have a singular 1-gon or weak 2-gon and will obtain a contradiction. Let F_α denote the quotient of $\tilde{F}$ by the cyclic group generated by α and let $f_\alpha : S^1 \to F_\alpha$ be the lift of f. If f_α possesses a singular 1-gon or weak 2-gon, this will descend to yield a singular 1-gon or weak 2-gon for f which contradicts our hypothesis. Hence f_α does not possess a singular 1-gon or weak 2-gon and so Lemma 1.16 or 1.17 shows that f_α has minimal self-intersection. It also follows that the pre-image in $\tilde{F}$ of $f_\alpha(S^1)$ consists of simple lines. If f_α is orientable, this is clear from Lemma 1.12. If F_α is non-orientable we simply note that, as in the orientable case, we can suppose that f_α is monotone in F_α, i.e., there is a homotopy equivalence $\pi : F_\alpha \to S^1$ s.t. $\pi \circ f_\alpha$ is monotone. We let l denote one of these lines.

As f has excess self-intersection but f_α does not, Lemma 3.4 shows that there must be an element g in $\pi_1(F)$ such that g is not a power of α and gl meets l in more than one point. There is a 2-gon D in $\tilde{F}$ between l and gl. We will show that the two arcs of ∂D descend to sub-arcs of S^1 and this yields a weak 2-gon for f contradicting our assumption. More precisely if $\partial_0 D$ denotes $D \cap l$ and $\partial_1 D$ denotes $D \cap gl$, we will prove that, for all non-zero integers n, $\beta^n(\partial_0 D) \cap \partial_0 D$ and $(g\beta g^{-1})^n(\partial_1 D) \cap \partial_1 D$ are both empty.

We start by choosing disjoint simple arcs a_i such that F cut along the a_i's is a 2-disc R. We choose these arcs so as to minimise their intersection with $f(S^1)$. Then the a_i's cut f into sub-arcs μ_j each lying in R. As f has no singular 1-gons, Lemma 1.1 shows that each μ_j is simple and as f has no embedded 2-gons, the usual arguments show no μ_j can have both ends on the same a_i. Let R denote a lift of R to $\tilde{F}$, the universal covering of F. Then $\tilde{F}$ is tiled by the translates by $\pi_1(F)$ of R. We let Γ denote the graph dual to this tiling. As $\tilde{F}$ is simply connected, Γ is a tree. Abstractly, Γ is the graph of $\pi_1(F)$ with respect to the generating set determined by the a_i's. We let $d(x, y)$ denote the distance function on vertices of Γ given as the least number of edges of a path in Γ from x to y. If β is an element of $\pi_1(F)$, it acts on Γ leaving invariant a unique line, called the axis of β, which contains all vertices x with minimum $d(x, \beta(x))$. This minimum value is called the translation length of β and we denote it $t(\beta)$. It equals the length of β as a word in our generators of $\pi_1(F)$, after cyclic reduction.

We choose a projection $p : \tilde{F} \to \Gamma$ which respects the actions of $\pi_1(F)$. By a homotopy of the μ_j's in R rel boundary, we can arrange that the restriction of p to each μ_j is monotone. Hence the restriction of p to each component of the pre-image of $f(S^1)$ is also monotone. Recall that l is a component of this pre-image stabilised by β. The axis of β as it acts on Γ is $p(l)$, which we denote L. Note that L is also the axis for α.

Now we consider the projection $p(D)$ of the 2-gon D into the graph Γ. Clearly, $p(D) \subset L \cap gL$, because of the monotonicity properties of p. We write λ for $L \cap gL$. We claim that $l(\lambda) < t(\alpha)$, where $l(\lambda)$ denotes the length of λ. This implies that, for all non-zero integers n, the sets $\alpha^n \lambda \cap \lambda$ and $(g\alpha g^{-1})^n \lambda \cap \lambda$ are both empty, and as $\beta = \alpha^d$, it follows that $\beta^n(\partial_0 D) \cap \partial_0 D$ and $(g\beta g^{-1})^n(\partial_1 D) \cap \partial_1 D$ are both empty as required. In order to prove our claim, we suppose that $l(\lambda) \geqq t(\alpha)$ and obtain a contradiction as follows. There must be a vertex x of Γ in λ such that $\alpha(x)$ also lies in λ. Then $\alpha(x)$ must equal $(g\alpha g^{-1})(x)$ or $(g\alpha g^{-1})^{-1}(x)$. As $\pi_1(F)$ acts freely on Γ, we deduce that $g\alpha g^{-1}$ equals α or α^{-1}, which implies that g and α generate a cyclic subgroup of $\pi_1(F)$, contradicting the fact that g is not a power of α. This completes the proof of Case 2 of Theorem 3.5.

Case 3. *F is closed, but is not the torus, Klein bottle,* 2*-sphere or projective plane*

Let f be a loop on F representing $\beta = \alpha^d$ such that f has excess self-intersection but does not have a singular 1-gon or weak 2-gon. As discussed at the start of Case 2, the pre-image in $\tilde{F}$ of $f(S^1)$ consists of simple lines. Further, if l denotes one of these lines with stabiliser β, there is an element g in $\pi_1(F)$ such that g is not a power of α and gl meets l in more than one point. Let G denote the subgroup of $\pi_1(F)$ generated by α and g and let F_1 denote the quotient of $\tilde{F}$ by G. Then Lemma 3.4 shows that the lift $f_1 : S^1 \to F_1$ of f has excess self-intersection, because gl meets l in more than one point. Also f_1 has no singular 1-gons or weak 2-gons, as these would yield singular 1-gons or weak

2-gons for f. Now F_1 cannot be closed, as $\pi_1(F_1)$ can be generated by two elements but F is not the torus or Klein bottle. Hence Case 2 applied to f_1 gives the required contradiction.

Case 4. *F is the torus*

In this case, we will prove that if f is an essential loop on the torus F with excess self-intersection, then f has an embedded 1-gon or 2-gon, which is a stronger result than we need. We will argue as in §2. Let f be a loop on F without embedded 1-gons or 2-gons. Then f represents a power of an essential simple loop γ, and so f can be homotoped to be disjoint from γ. We choose γ so as to meet f transversely in the least possible number of points. The arguments in the proof of Theorem 2.7 show that f must be disjoint from γ and so lies in the complementary annulus. Now Lemma 1.12 shows that f has least possible self-intersection, as claimed.

Case 5. *F is the Klein bottle K*

If f is an orientation reversing loop on K with excess self-intersection, we let $\tilde{f}$ denote the covering map into the orientable double cover T of K. As f has excess self-intersection, so does $\tilde{f}$. Now Case 4 implies that $\tilde{f}$ has an embedded 1-gon or 2-gon. Finally the arguments of Lemma 1.17 show that this yields a singular 1-gon or weak 2-gon for f as required.

If f is an orientation preserving loop on K with excess self-intersection, we proceed as follows. Recall that $\pi_1(K)$ has presentation $\{a, b : b^{-1}ab = a^{-1}\}$, so that every element of $\pi_1(K)$ can be expressed in the form $a^r b^s$. Further a is an orientation preserving element of $\pi_1(K)$ and b is orientation reversing. Thus f must represent an element of $\pi_1(K)$ of the form $a^r b^{2s}$. We let S denote the covering of K corresponding to the subgroup of $\pi_1(K)$ generated by a^r and b^{2s}. Thus S is a torus and is a regular covering of K and f has a lift f_1 into S. Further f_1 is homotopic to a simple loop on S. If f_1 is not simple, Theorem 2.7 shows that f_1 has an embedded 1-gon or 2-gon. This will project to a singular 1-gon or weak 2-gon for f. Otherwise f_1 is simple and so are all its translates in S by the covering group G which acts on S with quotient K. As f has excess self-intersection there must be an element τ of G such that τf_1 and f_1 have excess intersection. Now Lemma 3.1 shows that there must be an embedded 2-gon between f_1 and τf_1, and this will project to a weak 2-gon for f, as required. This completes the proof of Theorem 3.5.

We end this section by explaining how to deduce some of the results of [1] and [2] from Theorem 3.5. A loop on a surface with a metric is called *shortest* if it is shorter than all freely homotopic loops.

THEOREM 3.6 ([1], [2]). *Let F be a surface with a complete Riemannian metric and let α be a non-trivial primitive element of $\pi_1(F)$. If f is a shortest loop on F representing α, then f has least possible self-intersection.*

PROOF. First note that our hypothesis that α is primitive does not allow f to factor through a covering of circles. Thus f is self-transverse. It need not be in general position as there may be triple points, but the result of Theorem 3.5 is still applicable. Hence, if f has excess self-intersection there are sub-arcs λ and μ of S^1 such that $f \mid \lambda$ and $f \mid \mu$ are homotopic rel boundary. Now λ and μ need not be disjoint, so we cannot necessarily do the usual interchange construction to obtain a shorter loop f' homotopic to f. (Consider, for example, the loop in Fig. 0.5.) Instead we argue as follows. Let $\bar{\lambda}$ and $\bar{\mu}$ be sub-arcs of S^1 which contain λ and μ in their respective interiors. As f is a shortest loop on F, $f \mid \bar{\lambda}$ and $f \mid \bar{\mu}$ are each shortest arcs on F rel boundary. Now we define new, possibly singular, arcs $\lambda' = \bar{\lambda} - \lambda + \mu$ and $\mu' = \bar{\mu} - \mu + \lambda$ on F. Clearly $\bar{\lambda}$ is homotopic to λ' and $\bar{\mu}$ is homotopic to μ'. Also

$$l(\lambda') + l(\mu') = l(\bar{\lambda}) + l(\bar{\mu}),$$

and, as $\bar{\lambda}$ and $\bar{\mu}$ are distinct, the arcs λ' and μ' have corners. After rounding these corners we obtain shorter arcs λ'', μ'' so that

$$l(\lambda'') + l(\mu'') < l(\bar{\lambda}) + l(\bar{\mu}).$$

It follows that $l(\lambda'') < l(\bar{\lambda})$ or $l(\mu'') < l(\bar{\mu})$, either of which contradicts the fact that $f \mid \bar{\lambda}$ and $f \mid \bar{\mu}$ are shortest arcs rel boundary. This completes the proof of Theorem 3.6.

If one considers a shortest loop f on an orientable surface F representing a power α^d of α, then [1] shows that f factors through a covering of a loop g representing α. This also can be deduced from the above approach using the fact, Lemma 1.16, that any non-simple loop on the annulus has an embedded 1-gon or weak 2-gon.

§4. Curves on orientable surfaces

In this section, we show how to strengthen the results of §3, when one restricts attention to curves on orientable surfaces. Our first result strengthens Lemma 3.3. As usual, $\tilde{F}$ denotes the universal cover of F.

LEMMA 4.1. *Let f be a general position arc on an orientable surface F. If f has excess self-intersection, then f has a singular 1-gon or 2-gon.*

PROOF. We suppose that f has no singular 1-gons and will show that f must have a singular 2-gon. As in the proof of Lemma 3.3, it follows that the pre-image in $\tilde{F}$ of $f(I)$ consists of simple arcs and that if l denotes one of these arcs, there is g in $\pi_1(F)$ such that gl meets l in more than one point. Let F_g denote the quotient of $\tilde{F}$ by the cyclic group generated by g. The image of l in F_g is a singular arc L, and L has excess self-intersection by the analogue for arcs of Lemma 3.4. As F is orientable, so is F_g and hence Lemmas 1.2 or 1.9 imply that L has an embedded 1-gon or 2-gon. Now an embedded 1-gon for L would project to a singular 1-gon for f, contradicting our assumption. Hence L has an embedded 2-gon. Let λ and μ denote the two edges of the 2-gon. There are arcs $\tilde{\lambda}$ and $\tilde{\mu}$ in l such that under the projection map of l into F_g, $\tilde{\lambda}$ and $\tilde{\mu}$ are sent homeomorphically to λ and μ. In particular, $\tilde{\lambda}$ and $\tilde{\mu}$ are disjoint sub-arcs of l whose projections into F_g are homotopic rel boundary. It follows at once that the projections of $\tilde{\lambda}$ and $\tilde{\mu}$ into F are also homotopic rel boundary and so yield a singular 2-gon for f. This completes the proof of Lemma 4.1.

Our next result strengthens Theorem 3.5. The basic idea of the proof is as for Theorem 3.5.

THEOREM 4.2. *Let f be a general position loop on an orientable surface F. If f has excess self-intersection, then f has a singular 1-gon or 2-gon.*

PROOF. *Case* 1. *f is null-homotopic*

Then f is homotopic to a simple loop. Thus Theorem 2.7 implies that f has an embedded 1-gon or 2-gon.

From now on, we suppose that f is not null-homotopic.

Case 2. *F is closed*

If F is the torus, then Case 4 of Theorem 3.5 shows that f must have an embedded 1-gon or 2-gon, which is a stronger result than is required. If F is not the torus, we will deduce the required result from the case when F is not closed, in exactly the same way as in the proof of Case 3 of Theorem 3.5.

Case 3. *F is not closed*

As usual, we assume that F is compact with boundary. We also assume that f represents $\beta = \alpha^d$ where α is a non-trivial, primitive element of $\pi_1(F)$, and that f has excess self-intersection but does not have singular 1-gons or 2-gons. We will obtain a contradiction.

Exactly as in Case 2 of the proof of Theorem 3.5, the pre-image in $\tilde{F}$ of $f(S^1)$ consists of simple lines. If we let l denote one of these lines stabilised by β, there is g in $\pi_1(F)$ such that g is not a power of α and gl meets l in more than one point. Also by choosing arcs a_i in F which cut F into a 2-disc R, we obtain a tiling of $\tilde{F}$ by copies of R. The graph Γ dual to this tiling is a tree and, by homotoping f in F, we can suppose that we have a projection $p: \tilde{F} \rightarrow \Gamma$ such that the restriction of p to each of the lines gl is monotone.

As gl meets l in at least two points, there is a 2-gon D between l and gl. Our aim is to show that we can choose g and D so that D projects to a singular 2-gon for f. If we let $\partial_0 D$ denote $D \cap l$ and $\partial_1 D$ denote $D \cap gl$, we need to show that $\beta^n(\partial_0 D) \cap \partial_0 D$ and $(g\beta g^{-1})^n(\partial_1 D) \cap \partial_1 D$ are both empty, for any non-zero integer n, and that $g\beta^n(\partial_0 D) \cap \partial_1 D$ is empty, for all n. As in §3, we will achieve this by considering the projection of D into Γ. The axis of β as it acts on Γ is $p(l)$ which we denote L, and $p(D) \subset L \cap gL$. Note that L is also the axis for α.

We will need some technical results about Γ and the action of $\pi_1(F)$ on Γ. Recall that Γ is the graph of $\pi_1(F)$ with respect to a certain set of generators. We say that an edge path in Γ is *reduced* if it does not contain an edge immediately followed by its inverse. We denote the length of a reduced path γ by $l(\gamma)$. Any two vertices of Γ are joined by a unique reduced path.

LEMMA 4.3. *Let α and g be non-trivial elements of $\pi_1(F)$ with distinct axes L and M respectively. Let λ denote $L \cap gL$ and μ denote $L \cap M$. Then*

(i) $l(\lambda) < t(\alpha)$;

(ii) *if μ is empty, λ is also empty*;

(iii) *if μ is non-empty, either $t(g) = l(\mu)$ and λ meets M in a single point or $t(g) \neq l(\mu)$ and $\lambda \subset \mu$.*

PROOF. We start by observing that λ must be connected and hence an interval, as Γ is a tree.

(i) If $l(\lambda) \geqq t(\alpha)$, there is a vertex x of Γ in λ with $\alpha(x)$ also in λ. Also $\alpha(x)$ must equal $g\alpha g^{-1}(x)$ or $(g\alpha g^{-1})^{-1}(x)$. As $\pi_1(F)$ acts freely on Γ, it follows that $g\alpha g^{-1}$ equals α or α^{-1}, so that the group generated by g and α is cyclic. This contradicts our hypothesis that g and α have distinct axes L and M.

(ii) If $\mu = L \cap M$ is empty, we let γ denote the unique shortest path in Γ joining L and M. Thus $g\gamma$ joins gL and M. Let δ denote the unique shortest path in M joining γ to $g\gamma$. Note that δ is a non-trivial path as g is non-trivial. Suppose that there is a vertex x of Γ in $\lambda = L \cap gL$, and let ε and ε' denote the unique reduced paths in L and gL joining x to γ and $g\gamma$ respectively. Then the paths $\varepsilon \cup \gamma$ and $\varepsilon' \cup g\gamma \cup \delta$ are distinct reduced paths in Γ with the same endpoints. See Fig. 4.4. This contradiction shows that λ must be empty, proving part (ii) of the lemma.

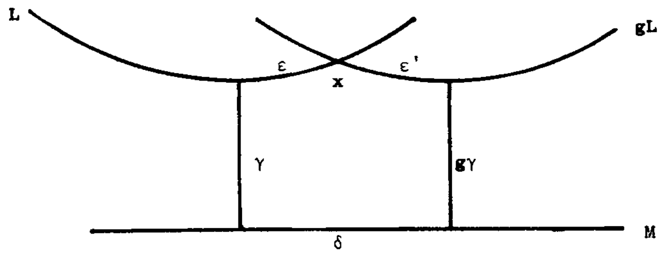


Fig. 4.4.

(iii) If $\mu = L \cap M$ is non-empty, we note that μ must be an interval. Further μ has finite length. In fact $l(\mu)$ must be less than the least common multiple N of $t(\alpha)$ and $t(g)$. For otherwise, choose vertices x and y of μ, distance N apart. There are integers r and s such that $y = g^r x = \alpha^s x$. As $\pi_1(F)$ acts freely on Γ, it follows that $g^r = \alpha^s$. As $\pi_1(F)$ is a free group, it follows that g and α generate a cyclic subgroup. This contradicts our hypothesis that the axes of g and α are distinct.

Now the line L is a union of three intervals I_0, μ and I_1, and thus $gL = gI_0 \cup g\mu \cup gI_1$. We orient M so that g moves points of M in the positive direction and choose I_0 and I_1 so that $I_1 \cap M$ is on the positive side of $I_0 \cap M$. Now it is clear that gI_0 cannot meet I_0 and that gI_1 cannot meet I_0 or I_1. Hence λ must equal $\mu \cap g\mu$, unless $t(g) = l(\mu)$, in which case λ is contained in I_1 and meets M in a single point. (Possibly λ is this single point.) This completes the proof of Lemma 4.3.

LEMMA 4.5. *Let α and g be non-trivial elements of $\pi_1(F)$ with distinct axes L and M respectively. Let λ denote $L \cap gL$ and let μ denote $L \cap M$. Suppose that $t(g) \leqq t(g\alpha^n)$, for all integers n.*

(i) *If $t(g) = l(\mu)$, so that λ meets M in a single point, then $g\alpha^n\lambda \cap \lambda$ is empty, for all integers n.*

(ii) *If $t(g) \neq l(\mu)$, so that $\lambda \subset \mu$, and if $\bar{\lambda}$ is a sub-arc of λ with $l(\bar{\lambda}) \leqq t(g)$, then $g\alpha^n\bar{\lambda} \cap \bar{\lambda}$ is empty for all non-zero integers n.*

PROOF. (i) Recall from the proof of Lemma 4.3 that L can be expressed as a union of intervals I_0, μ and I_1 and that $\lambda \subset gI_0 \cap I_1$. See Fig. 4.6. Let X denote $I_0 \cap M$, and let Y denote the other end of $g^{-1}\lambda$. We will prove that if ν denotes the sub-arc $g^{-1}\lambda \cup \mu \cup \lambda$ of L, then $l(\nu) < t(\alpha)$. This implies the required result as follows. If $\alpha^n\lambda \subset I_1$, then $g\alpha^n\lambda$ cannot meet λ as it cannot even meet I_1. Otherwise $\alpha^n\lambda$ must lie in I_0 and be disjoint from $g^{-1}\lambda$ because $l(\nu) < t(\alpha)$.

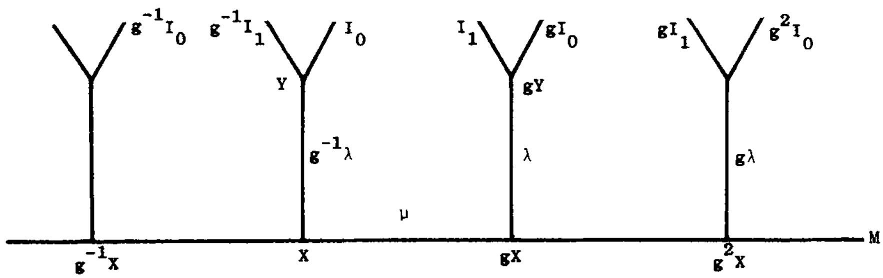


Fig. 4.6.

Hence $g\alpha^n\lambda$ cannot meet λ in this case either. It remains to prove the required inequality.

We assume that L and M are oriented so that α and g translate in the positive direction and that these orientations agree on μ. First note that we must have $t(g) < t(\alpha)$. For otherwise $t(\alpha) \leqq t(g)$ and as $l(\mu) = t(g)$, it follows that $g^{-1}\alpha$ sends X to a point of μ closer to X than $t(g)$. Hence $t(g\alpha^{-1})$ is strictly less than $t(g)$ contradicting our hypothesis on g. Now we will show that $\alpha g^{-1}(\lambda)$ cannot meet λ. As $\alpha(X)$ cannot lie in μ, because $t(\alpha) > l(\mu)$, this will show that $\alpha g^{-1}(\lambda)$ lies in $I_1 - \lambda$ and so will prove that $t(\alpha) > l(\nu)$ as required. Suppose that $\alpha g^{-1}(\lambda)$ does meet λ. As λ and $g^{-1}\lambda$ have the same length, one of $\alpha(X)$ or $\alpha(Y)$ must lie in λ. If $\alpha(X)$ lies in λ, then $g^{-1}\alpha(X)$ lies in $g^{-1}\lambda$ so that $\alpha g^{-1}\alpha X$ lies in L. If d denotes the distance of $\alpha(X)$ from $g(X)$, then d is also the distance of $\alpha g^{-1}\alpha(X)$ from $\alpha(X)$. It follows that $\alpha g^{-1}\alpha(X)$ equals $g(X)$, as these two points lie in L at the same distance and in the same direction from $\alpha(X)$. Hence $\alpha g^{-1}\alpha = g$ as $\pi_1(F)$ acts freely on Γ. Hence αg^{-1} has order two, so that αg^{-1} is trivial, as $\pi_1(F)$ is free. But this contradicts our hypothesis that g and α have distinct axes. If $\alpha(Y)$ lies in λ, we obtain a contradiction in the same way, by showing that $\alpha g^{-1}\alpha(Y) = g(Y)$.

(ii) Now suppose that $t(g) \neq l(\mu)$ so that $\lambda \subset \mu$, and let $\bar{\lambda}$ be a sub-arc of λ with $l(\bar{\lambda}) \leqq t(g)$. If $g\alpha^n\bar{\lambda}$ meets $\bar{\lambda}$, then $t(g\alpha^n) \leqq t(g)$. But the hypothesis of our lemma is that $t(g) \leqq t(g\alpha^n)$, for all n. We deduce that if $g\alpha^n\bar{\lambda}$ meets $\bar{\lambda}$ then $t(g\alpha^n) = t(g)$. This implies that $g\alpha^n\bar{\lambda} \cap \bar{\lambda}$ is a single vertex. We label the endpoints of $\bar{\lambda}$ by v and w so that $g\alpha^n(v) = w$. But g translates vertices of m by distance $t(g)$ so that we must also have w equal to $g(v)$ or $g^{-1}(v)$. As $\pi_1(F)$ acts freely on Γ we deduce that $g\alpha^n$ equals g or g^{-1}, either of which is impossible when n is non-zero. This completes the proof of Lemma 4.5.

Now we return to our proof of Case 3 of Theorem 4.2. Recall that as f has excess self-intersection but no singular 1-gons or 2-gons, the pre-image of $f(S^1)$

in $\tilde{F}$ consists of lines gl, where l is stabilised by β and that there is g in $\pi_1(F)$ which is not a power of α such that gl meets l in more than one point. In this situation we say that $gl \cap l$ is *excess*.

Our first step is to show that there exists g in $\pi_1(F)$ such that $gl \cap l$ is excess, that no power g^n of g, for $n \geqq 2$, has this property, and that $t(g) \leqq t(g\alpha^n)$, for all integers n. We will construct a sequence of elements g_i of $\pi_1(F)$ and show that this sequence must stop at an element of the required type. We start with g_1 such that $g_1 l \cap l$ is excess. If $t(g_1\alpha^r) < t(g_1)$, for some r, we let g_2 be one of the elements $g_1\alpha^r$ of shortest translation length. Note that $g_2 l = g_1 l$ so that trivially $g_2 l \cap l$ is excess. If some proper power g_2^n of g_2 has the same property, we let g_3 denote g_2^n. Now we repeat the above two steps. If this process stops, we are done, but it is not obvious that it should stop. We use our analysis of the intersection $L \cap gL$ carried out in Lemma 4.3. Note that if $gl \cap l$ is excess, then $gL \cap L$ must have length at least one. Now it follows from Lemma 4.3, that if $gL \cap L$ is non-empty, then $g^n L \cap L$ is strictly shorter than $gL \cap L$ for all $n \geqq 2$. Hence in the above construction $l(g_1 L \cap L) > l(g_3 L \cap L) > \cdots$. Hence the construction stops and we obtain the required element g of $\pi_1(F)$.

Consider the image $\pi(l)$ of l in the surface F_g the quotient of $\tilde{F}$ by the infinite cyclic group generated by g. This surface is orientable, and this is the one point in the proof where we use the orientability of F. The fact that $gl \cap l$ is excess implies that we can find a compact annulus A in F_g such that the arc $\pi(l)$ meets A in an arc with excess self-intersection. This uses the analogue for arcs of Lemma 3.4. Now Lemmas 1.2 and 1.9 imply that this arc possesses an embedded 1-gon or 2-gon. But an embedded 1-gon in F_g would lift to an embedded 1-gon in $\tilde{F}$ which is impossible. Hence we have an embedded 2-gon in F_g. Lifting this to $\tilde{F}$ yields an embedded 2-gon D between l and gl such that $gD \cap D$ is empty. (Note that in principle D could be between l and $g^n l$, but this is excluded as $g^n l \cap l$ is not excess, for $n \geqq 2$.) We claim that D descends to a singular 2-gon for f. This is equivalent to asserting that the two arcs $\partial_0 D$ and $\partial_1 D$ of ∂D yield disjoint sub-arcs of the domain of f. A better way to say this is that for every non-trivial element h in $\pi_1(F)$, the arcs $h(\partial_0 D)$ and $\partial_0 D$ never overlap, that $h(\partial_1 D)$ and $\partial_1 D$ never overlap and that $h(\partial_0 D)$ and $\partial_1 D$ never overlap. The only elements of $\pi_1(F)$ which could cause such overlaps are those of the form β^n, $(g\beta g^{-1})^n$, $g\beta^n$ and their inverses.

Recall the projection $p: \tilde{F} \to \Gamma$ and consider the image of D. Clearly $p(D) \subset L \cap gL = \lambda$ and, in particular, λ has length at least one. As $l(\lambda) < t(\alpha)$, by Lemma 4.3, we know that $\alpha^n\lambda \cap \lambda$ and $(g\alpha g^{-1})^n\lambda \cap \lambda$ are empty, for all non-zero integers n. It follows that $\partial_0 D$ and $\partial_1 D$ each yield sub-arcs of the domain of f, as $\beta = \alpha^d$. Now we consider the two cases of Lemma 4.5. In the first case, $g\alpha^n\lambda \cap \lambda = \varnothing$, for all integers n, so it follows at once that D descends to a singular 2-gon for f as required. In the second case, we note that $p(D)$ is a sub-arc $\bar{\lambda}$ of λ. The fact that $gD \cap D$ is empty implies that $l(\bar{\lambda}) \leqq t(g)$ and we can apply the result of Lemma 4.5(ii) to deduce that $g\alpha^n\bar{\lambda} \cap \bar{\lambda}$ is empty, for all non-zero integers n. It follows that D descends to a singular 2-gon for f and we have completed the proof of Theorem 4.2.

References

1. M. H. Freedman, J. Hass and P. Scott, *Closed geodesics on surfaces*, Bull. London Math. Soc. **14** (1982), 385–391.

2. J. Hass and J. H. Rubinstein, *One-sided closed geodesic on surfaces*, University of Melbourne, preprint.