ISSN 1464-8997 **201**

Geometry & Topology Monographs
Volume 2: Proceedings of the Kirbyfest
Pages 201–213

Configurations of curves and geodesics on surfaces

Joel Hass
Peter Scott

Abstract We study configurations of immersed curves in surfaces and surfaces in 3–manifolds. Among other results, we show that primitive curves have only finitely many configurations which minimize the number of double points. We give examples of minimal configurations not realized by geodesics in any hyperbolic metric.

AMS Classification 53C22; 57R42

Keywords Geodesics, configurations, curves on surfaces, double points

Let f and g be general position immersions of a manifold M into the interior of a manifold N. We will say that f and g *have the same configuration* if there is a regular homotopy from f to g through general position immersions. Equivalently, there is an ambient isotopy of N moving $f(M)$ to $g(M)$. This defines an equivalence relation on general position immersions, and an equivalence class will be called a *configuration.* All the immersions in a configuration "look the same" in a precise sense. In this paper we will be interested in the cases when the dimension of M is 1 or 2 and the dimension of N is 2 or 3. Our aim is to explore the question of how many configurations a given homotopy class can have. For primitive curves on a surface, we show that the number is finite if one restricts to immersions with the least possible number of double points, but little can be said for curves with excess intersections. We then consider the possible configurations of closed geodesics on a surface equipped with a hyperbolic metric. It is well known that geodesics in a hyperbolic metric minimize the number of double points in their homotopy class. It was shown by Shephard [5] and Neumann-Coto [4] that any curve configuration (not necessarily connected) which minimizes the number of double points is realized by shortest geodesics in some metric. We construct examples which show that some configurations cannot be realized by closed geodesics in a hyperbolic metric.

We say that a map of the circle into a surface is *primitive* if it is not homotopic to a proper power of some other map. Our first and most general result is the following.

Theorem 1 *Let $f\colon\ S^1 \to F$ be a primitive map of the circle into an orientable surface. Then the general position immersions which are homotopic to f and have the minimal possible number of double points belong to only finitely many configurations.*

Proof If f is nullhomotopic in F, then any general position immersion which is homotopic to f and has the minimal possible number of double points must be an embedding. Further such an embedding will bound a disc. It follows that there are two configurations possible for such maps, one for each orientation of the curve. Thus the theorem holds for the 2–sphere. If f is homotopically essential in F, and is homotopic to an embedding, there is only one configuration possible among embeddings homotopic to f. For the torus T, the assumption that $f\colon\ S^1 \to T$ is primitive implies that f is homotopic to an embedding and so has a unique configuration.

Assume now that $\chi(F) < 0$ and pick a hyperbolic metric on F. The pre-image of $f(S^1)$ in the universal cover $\mathbb{H}^2$ consists of a line (in the topological sense) l and its translates $\{gl\}$, $g \in G$. These lines will not be geodesics in general, but each will lie in a bounded neighborhood of a unique geodesic. As f represents a primitive element of $\pi_1(F)$, no two of these lines have the same endpoints. The minimality of the number of double points of f implies that any two of these lines meet in at most one point, as is the case with hyperbolic geodesics. Let p_{ij} denote the point of intersection of distinct translates l_i and l_j, with the convention that p_{ij} does not exist if l_i and l_j are disjoint.

Claim 2 *If we know the side of l on which p_{ij} lies for all i, j, then the configuration of f is determined.*

Proof Note that the assumption in the claim implies that for each l_k we know the side of l_k on which p_{ij} lies for all i, j. We will construct the configuration of lines one at a time, starting with $l = l_1$. Assume that the lines $l_1, \ldots, l_{n-1}$ have a unique configuration. We will establish that the configuration of the lines $l_1, \ldots, l_n$ is also unique. Consider the choices when we add the additional line l_n. Two disjoint lines in $\mathbb{H}^2$ cannot be interchanged by a homotopy of f, unless they have the same endpoints, as they do not lie within a bounded distance of one another. But the assumption that the curve f is primitive implies that no two lines have the same endpoints. Hence if l_n is disjoint from l_i then the side of l_i on which it lies is determined. Suppose that l_n crosses some l_k. The points $p_{ik}, i < n$, in which the previous lines meet l_k, divide l_k into several arcs. As we know on which side of l_i the point p_{kn} lies, we know in which of these arcs p_{kn} lies. It follows that up to an isotopy of the lines $l_1, \dots, l_n$, there is at most one possible way in which to add l_n. Now induction on n shows that the collection of all translates of l is determined up to ambient isotopy of $\mathbb{H}^2$. Further, if we have two immersions f and g of S^1 in F such that the corresponding families of lines in $\mathbb{H}^2$ are ambient isotopic, we claim that the isotopy can be chosen to be equivariant under the action of $\pi_1(F)$ on $\mathbb{H}^2$, so that f and g must have the same configuration as claimed. The way to do this is first to ensure that the isotopy is equivariant when restricted to the intersection points of the two families of lines, then to ensure equivariance of the isotopy when restricted to the union of the lines and finally to ensure that the entire isotopy is equivariant by defining it equivariantly on each of the regions into which the union of the lines divides the hyperbolic plane. □

Claim 3 *Let γ be a closed geodesic in some hyperbolic metric on F, so that l and its translates are geodesics in the hyperbolic plane $\mathbb{H}^2$. Fix a line l_i which crosses l. Then the number of lines which cross both l and l_i is finite.*

Proof The entire configuration of lines projects to a closed curve in F, which must have only finitely many double points, and it follows that there are only finitely many values for the angles between any two lines l_i and l_j which meet. In particular, the angles are bounded uniformly away from zero. This yields an upper bound to the lengths of the sides of any triangle formed by these lines. If the number of l_j's which cross both l and l_i is not bounded, then since the set of all lines cannot accumulate, there must be triangles of unbounded size, a contradiction. □

Now we can complete the proof of Theorem 1. We return to the general situation where no metric is assumed. As any two of the lines forming the pre-image of f intersect in at most one point, the intersections of these lines correspond to the intersections of the corresponding geodesics in $\mathbb{H}^2$. It follows that the conclusion of the preceding claim applies, so that the number of l_j's which cross both l and l_i is a finite number m_i. Now the translates of l which cross l fall into $2n$ orbits under the action of the stabiliser of l, where n denotes the number of double points of f. Let $l_1, ..., l_{2n}$ denote one representative from each orbit. The total number of points $p_{ij}, 1 \leq i \leq 2n$, such that l_j crosses l and l_i is $m = m_1 + m_2 + ... + m_{2n}$. For each l_j which crosses both of l and l_i, there are at most two choices for which side of l the point p_{ij} occurs. Hence the total number of choices for which side of l these m points lie is bounded by 2^m. But these choices determine completely on which side of l every p_{ij} lies.

Now Claim 2 implies that there are at most 2^m possible configurations. This completes the proof of Theorem 1. □

Remark The condition that the number of double points be minimal is essential for Theorem 1.

Even if one restricts the number of double points to two, there is a curve on a surface whose homotopy class contains infinitely many distinct configurations. An example, as shown in Figure 1, can be obtained by beginning with a simple closed curve C on a surface F, choosing a simple arc λ on F which meets C only in its endpoints, and isotoping a small arc of C at one end of λ until it runs back and forth along λ and cuts C twice near the other end of λ. For most surfaces F, the relative homotopy class of λ can be chosen in infinitely many different ways, yielding infinitely many distinct configurations with two double points which are all homotopic to the initial embedding.

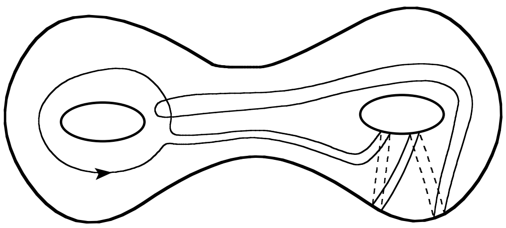

Figure 1: Extra double points on a surface of genus two

We will now consider some examples which show that configurations of curves on a surface with minimal self-intersection cannot always be realized by a geodesic in a hyperbolic metric.

Lemma 4 *Let g_t be a family of Riemannian metrics on a closed manifold, let γ be a closed curve in M, and let γ_t be a shortest closed geodesic homotopic to γ in the metric g_t. If γ_0 is the unique geodesic in its homotopy class then γ_t varies continuously with t at $t = 0$. If each γ_t is unique, then the whole family is continuous.*

Proof Let N_ϵ be an ϵ–neighborhood of γ_0 in the metric g_0. If there are γ_{t_i} not entirely contained in N_ϵ for a sequence $t_i \to 0$, then a subsequence of these converges by an application of Ascoli's Theorem, and the limit will be a geodesic not entirely contained in N_ϵ, but homotopic to γ_0 and having the same length. Thus γ_t lies inside N_ϵ for t sufficiently small, and the family of geodesics varies continuously at $t = 0$. □

Our first example, for simplicity of construction, considers intersections of three simple curves. We then describe a similar, but more complicated example which uses a single singular curve.

Example 5 There are three simple closed curves on a punctured torus F which have several minimal intersection configurations, of which only one is achieved by geodesics in any hyperbolic metric on F.

Let a and b be a basis for $\pi_1(F)$ representing a longitude and meridian, and let α, β and γ be closed geodesics representing a, b and ab. Each of these curves is simple and each pair cross in a single point. See Figure 2.

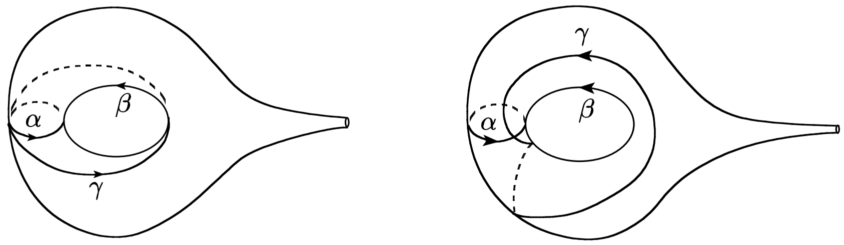

Figure 2: Only the first configuration can be realized by hyperbolic geodesics.

The punctured torus has an involution $\tau\colon F \to F$ which fixes three points and such that $\tau(a) = a^{-1}, \tau(b) = b^{-1}$. So α and β are preserved by the involution. We have

$$\tau(ab) = a^{-1}b^{-1} = a^{-1}(b^{-1}a^{-1})a = a^{-1}(ab)^{-1}a$$

so that ab is taken to a conjugate of its inverse, and the geodesic γ is also preserved. Hence each curve is invariant, but reversed, and so its image contains two fixed points of τ. For any pair of the three curves, the unique point at which they intersect must be fixed by the involution.

If all three of α, β and γ intersect at a common point x then this point is fixed by τ, as are three additional and distinct points, one on each of α, β and γ. This would result in more than three fixed points for τ, a contradiction. So $\alpha \cap \beta, \beta \cap \gamma$ and $\alpha \cap \gamma$ are three distinct points on F, as in Figure 2. Now suppose that there is more than one possible configuration in the homotopy class of the three curves, realized by two distinct hyperbolic structures T_1 and T_2. We can connect the two structures in Teichmuller space by a path of hyperbolic structures T_t. By Lemma 4, closed geodesics in a given homotopy class vary continuously on the surface as we follow a path of hyperbolic metrics in Teichmuller space. The above argument shows that for each metric T_t, the unique geodesics in the homotopy classes a, b and ab have no triple points. It follows that we cannot change configurations. However there is a complementary region of these three curves which is a triangle—in fact two of them are. So topologically it is possible to alter the configuration by sliding one of the edges of this triangle across the opposite vertex. The resulting configuration still minimizes the number of intersection points but cannot be realized in any hyperbolic metric. See Figure 2. The same example can be put into any surface, by constructing it inside a subsurface homeomorphic to a torus with a disk removed.

Example 6 A connected closed curve on a punctured torus with several minimal intersection configurations, of which only one is achieved by a geodesic in a hyperbolic metric.

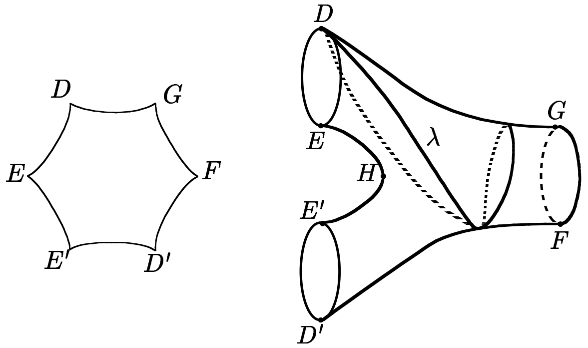

Figure 3: An all right hexagon and a geodesic arc in a pair of pants

We start with an all right angled hyperbolic hexagon $DEE'D'FG$, then double it along the edges $EE', D'F$ and GD to obtain a pair of pants X with a hyperbolic metric, as in Figure 3. Thus X admits a reflection involution σ which interchanges the two hexagons. It also admits an orientation preserving involution τ which fixes a single point H of X, where H is the midpoint of the arc EE'. Now choose a geodesic loop λ on X based at D as shown in Figure 3. This loop is not a closed geodesic, as there will be a corner at D. It is freely homotopic to the square of the boundary component which contains F and G, so it cannot be simple. However, it can be realized with only one double point and hence has exactly one double point. As each boundary component of X is preserved by σ but reversed in orientation and as D is fixed by σ, it follows that λ is preserved by σ but with reversed orientation. Hence λ must look as shown in Figure 3 with its single double point on the arc $D'F$.

Now form a once punctured torus T from X by gluing together the two boundary components containing D, E and D', E' so that D is glued to D' and E is glued to E'. Then τ induces an orientation preserving involution on T which we will continue to denote by τ, which fixes $D = D', E = E'$ and H. We will be interested in the closed loop γ on T defined by $\gamma = \lambda \cup \tau\lambda$. As τ is a rotation through π in a neighborhood of D, it follows that γ is a closed geodesic. See Figure 4 which shows that γ has seven double points. The loop γ has two innermost triangles, and using one of these triangles we can change the configuration. However, we claim that no such triangle move can be realized by the closed geodesics in a family of hyperbolic metrics. For any hyperbolic metric on X can be obtained from some all right angled hyperbolic hexagon by doubling, so the preceding argument applies to show that γ will always have seven distinct double points and no triple points. Thus the configuration of γ cannot alter as the hyperbolic metric changes continuously.

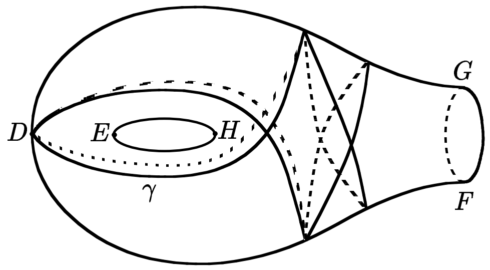

Figure 4: A curve with a unique configuration in any hyperbolic metric

Next we discuss another example of a loop on a surface F with several minimal intersection configurations, of which only one is achieved by a geodesic in a hyperbolic metric.

Example 7 A unique configuration on a thrice-punctured S^2.

Let Σ denote a thrice-punctured S^2 equipped with a complete hyperbolic metric of finite area. Let α denote the element of $\pi_1(\Sigma)$ represented by the first loop shown in Figure 5. We will use the fact that Σ admits an action of $\mathbb{Z}_3$ by isometries which cycles the three ends of Σ to show that the configuration of the closed geodesic representing α must be the first one shown in Figure 5.

The proof is to consider the geodesic loop λ shown in Figure 6, whose corner is at v, one of the two points fixed by the action of $\mathbb{Z}_3$, and show that $\theta > \pi/3$. Clearly the union of the translates of λ under the action of $\mathbb{Z}_3$ forms a loop

Figure 5: A forced configuration and an impossible configuration

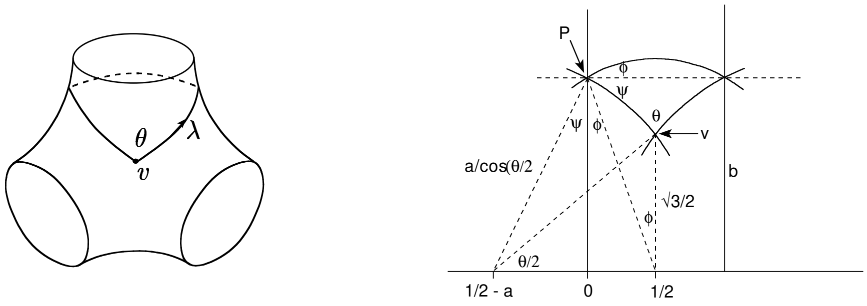

Figure 6: Calculating θ

representing α. If $\theta = \pi/3$, this loop will be a closed geodesic and so the geodesic representing α will have a triple point. If $\theta < \pi/3$, the geodesic representing α would have the second configuration shown in Figure 5. The proof that $\theta > \pi/3$ involves some straightforward hyperbolic geometry to show that $\theta = 2\tan^{-1}\left(\frac{\sqrt{3}}{2}\right)$, which is approximately 81.79 degrees. See Figure 6, which shows an ideal triangle with vertices at $0, 1$ and ∞ in the upper half plane model of the hyperbolic plane. If we regard Σ as the double of this triangle, there is a natural quotient map from Σ to the triangle and the image of λ is the piecewise geodesic triangular loop shown. It has the properties that the exterior angles between λ and the geodesics $x = 0$ and $x = 1$ are all equal. Thus the angles marked ϕ and ψ must be equal. We let r denote the Euclidean radius of the circle which forms the hyperbolic geodesic joining v and P and a denote the width of the projection to the x–axis of the radial segment of length r connecting the center of this circle to v. Then $r = \dfrac{a}{\cos\theta/2}$ and the circle is centered at $(1/2 - a, 0)$. In these coordinates, the rotation of the hyperbolic plane which sends 0 to 1 to ∞ is the Mobius transformation $z \to \dfrac{1}{1-z}$. Recall that v is fixed by this map. It follows that v is the point

$\frac{1}{2}+i\frac{\sqrt{3}}{2}$. Hence $\tan(\theta/2)=\frac{\sqrt{3}}{2a}$. Also, if b denotes the Euclidean height of P above the x-axis, then $\tan\phi=\frac{1/2}{b}$ and $\tan\psi=\frac{a-1/2}{b}$. As ϕ and ψ are equal, we have $\tan\phi=\tan\psi$, so that $a=1$. It follows that $\tan(\theta/2)=\frac{\sqrt{3}}{2}$, so that $\theta=2\tan^{-1}\left(\frac{\sqrt{3}}{2}\right)$, as claimed.

Remark Ian Agol has pointed out that the second configuration can also be eliminated by a direct calculation in hyperbolic geometry. There is a hexagon in the complement of the arcs, as well as a triangle. If $\theta_1,\theta_2,\theta_3$ are the three interior angles of the triangle, then the hexagon has exterior angles $\theta_1,\theta_2,\theta_3,\theta_1,\theta_2,\theta_3$. This is a contradiction since we must have $\theta_1+\theta_2+\theta_3<\pi$ and $2\theta_1+2\theta_2+2\theta_3>2\pi$. Moreover, Agol's observation applies more generally in any complete negatively curved metric on the three punctured sphere.

Now we give an example of non-uniqueness of configurations.

Example 8 Non-unique configurations realized by a hyperbolic geodesic.

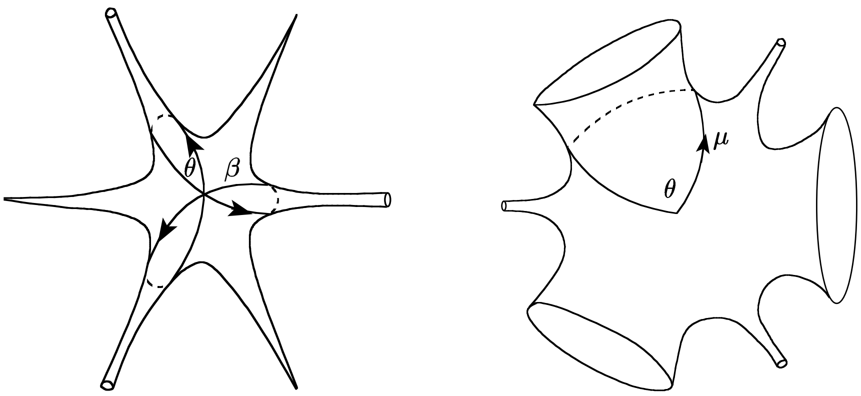

Figure 7: Two six punctured spheres with a hyperbolic metric

Let Ω denote S^2 with six points removed equipped with a complete hyperbolic metric having three cusp ends and three ends of infinite area and admitting an action of $\mathbb{Z}_3$ which cycles the two types of end among themselves. Let β denote the element of $\pi_1(\Omega)$ represented by the loop shown in Figure 7. As before we consider the arc μ shown in Figure 7, and the angle θ. Note that the union of the three translates of μ by the action of $\mathbb{Z}_3$ forms a loop representing β. We will show that the closed geodesic representing β has at least two configurations which can be realized by closed geodesics for some hyperbolic structure on Ω.

To see this, start with a metric in which all the ends are cusps and there is an action of $\mathbb{Z}_6$ on Ω which cycles the ends. In this case, it is not as easy to calculate θ, but it is clear that $\theta < \pi/3$. Now alter the metric on Ω, by enlarging the three infinite area ends. Clearly $\theta \to 2\pi/3$ as the lengths of the three closed geodesics tends to infinity. Hence by continuity, there is a metric where $\theta = \pi/3$, and so the closed geodesic representing β has a triple point. Distinct configurations will be obtained for metrics near to this one for which $\theta < \pi/3$ and $\theta > \pi/3$.

Now we consider surfaces immersed in 3–manifolds. There is a natural analog of Theorem 1. The statement, which we give below, uses the 1–line property, which is a property of a map of a surface into a 3–manifold. See sections 1 and 2 of [3] for the definition and basic results.

Theorem 9 *Let N be a closed, P^2–irreducible 3–manifold, let F be a closed surface and let $f\colon F \to N$ be a 2–sided π_1–injective map. Then the general position immersions which are homotopic to f, have the 1–line property, whose double curves are primitive on F and have the least possible number of double points for their homotopy classes, belong to only finitely many configurations.*

Remark It is not assumed that f has the 1–line property. Further, there may be no immersions homotopic to f with the required properties. In this case, the result is trivial, but not interesting!

Proof Homotopic maps with the 1–line property have precisely the same double curves up to homotopy. Our hypothesis that the double curves have the least number of double points means that we can use Theorem 1 to deduce that there are only finitely many configurations for the double curves of the maps homotopic to f which have the 1–line property and the other properties which we are assuming. Finally, the proof of Lemma 4.1 of [3] shows that each configuration of double curves determines only one configuration for a map $F^2 \to N$, so the result follows. □

For any map g of a surface into N, double points of the double curves of g are triple points of g. Thus if g satisfies all the hypotheses of Theorem 9, then it must have the least possible number of triple points in its homotopy class. However, the following example due to Casson shows that a map may have the least possible number of triple points in its homotopy class, while its double curves do not have the least possible number of double points. In fact, *football regions*, complementary regions homeomorphic to balls and bounded by three 2–gons, must occur in Casson's example.

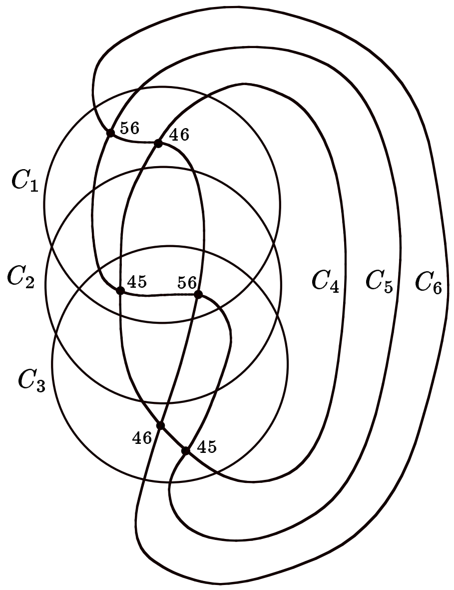

Figure 8: Any disks in a ball which have these curves as boundary must have a football region between them.

Example 10 (Casson) A collection of surfaces which must contain a football region in any configuration.

We consider six simple closed curves $C_1, ..., C_6$ on the 2–sphere S^2 as shown in Figure 8, so that each pair intersects transversely in two points. Each C_i bounds a 2–disc D_i properly embedded in the 3–ball B^3, and we assume that these discs are chosen in general position. Further, by choosing these discs to be least area in some metric, we can assume that any pair of these discs intersect in a single arc, ie, there are no circles of intersection. The surprising property of this picture is that there must be a football region W in B, ie, a sub-ball W of B bounded by the union of three discs each lying in some D_i, such that each pair of discs intersects in an arc. In particular, it is impossible to embed the six discs D_i in B so that the double arcs minimize their number of double points. Note that we are not claiming that W is a component of the complement of the six discs. It is quite possible that some of the discs can cut across W.

Before starting on the proof, we remark that if one considers three simple closed curves on S^2 which are in general position and such that each pair intersect in exactly two points, then there are only two possible configurations, as shown in Figure 9. In the first configuration shown in Figure 9, which we refer to as the prism case, the discs can be chosen so that each pair intersects in a single arc and there is no triple point. In this case, the three discs cut B^3 into seven regions, one of which meets S^2 in two triangular regions. This region is referred to as the prism region. In the second configuration shown in Figure 9, which we refer to as the triple point case, the discs can be chosen so that each pair intersects in a single arc and there is exactly one triple point. In the triple point case there must always be at least one triple point however the discs are embedded. In Figure 8 , the configuration of C_1, C_2, C_3 is of the prism type, and the configuration of C_4, C_5, C_6 is of the triple point type.

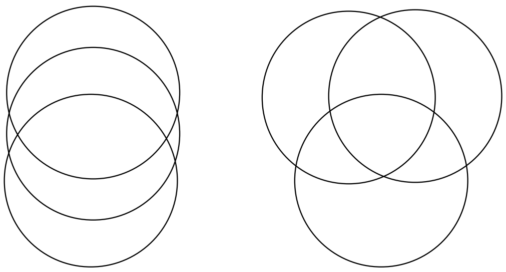

Figure 9: Two configurations of three curves on a sphere

Now suppose that we have an embedding of the D_i's in B^3 such that any two double lines of the D_i's intersect in at most one point. Figure 8 shows that the two ends of the double curve $D_5 \cap D_6$ (labelled 56 in the picture) lie on the same side of D_1, and that this is on the opposite side of D_1 from the prism region P formed by D_1, D_2 and D_3. Similarly the two ends of the double curve $D_4 \cap D_5$ (labelled 45 in the picture) lie on the same side of D_3, and this is on the opposite side of D_3 from P. Finally the two ends of the double curve $D_4 \cap D_6$ (labelled 46 in the picture) lie on the same side of D_2, and this is on the opposite side of D_2 from P. This implies that the three arcs in question cannot have a common point, as the intersection of the sides of D_1, D_2 and D_3 which do not contain the prism region P is empty. This contradicts the fact that the configuration of D_4, D_5, D_6 is of the triple point type, so we conclude that for any embedding of the D_i's in B some pair of double lines l and m must intersect in at least two points. For notational simplicity, suppose that $l = D_1 \cap D_2$ and $m = D_1 \cap D_3$. Thus there are 2–gon regions in D_1 bounded by sub-arcs of l and m. We choose one X which is innermost in the sense that its interior is disjoint from l and m, and let λ and μ denote the sub-arcs of l and m respectively which form the boundary of X. Let $n = D_2 \cap D_3$ and let ν denote the sub-arc of n which has ends at $\lambda \cap \mu$. Then $\lambda \cup \nu$ bounds a 2–gon Y in D_2 and $\mu \cup \nu$ bounds a 2–gon Z in D_3 and $X \cup Y \cup Z$ bounds a football region W in B^3.

Acknowledgements

Research of Joel Hass was partially supported by NSF grant DMS-9704286. Research of Peter Scott was partially supported by NSF grant DMS-9626537.

References

[1] **R Gulliver**, **P Scott**, *Least area surfaces can have excess triple points*, Topology, 26 (1987) 345–359.

[2] **J Hass** , **P Scott**, *Intersections of curves on surfaces*, Israel Journal of Math. 51 (1985) 90–120.

[3] **J Hass** , **P Scott**, *Homotopy equivalence and homeomorphism of 3–manifolds*, Topology, 31 (1992) 493–517.

[4] **M Neumann-Coto**, *A characterization of shortest geodesics in surfaces*, preprint (1999)

[5] **M Shephard**, PhD Thesis, UC Berkeley (1990)

Department of Mathematics, University of California
Davis, CA 95616, USA
and
Department of Mathematics, University of Michigan
Ann Arbor, MI 48109, USA

Email: `hass@math.ucdavis.edu, pscott@math.lsa.umich.edu`

Received: 22 March 1999 Revised: 23 August 1999