

Contracting loops on a Riemannian 2-surface

Gregory R. Chambers and Regina Rotman

November 10, 2013

Abstract

Let M be a Riemannian 2-disc and q a point on its boundary. In this paper we will show that, for any $\varepsilon > 0$, any homotopy of its boundary to a point over closed curves of length at most L can be modified so that the resulting homotopy is a fixed point homotopy over the loops based at q of length at most $L + 2d + \varepsilon$, where d is the diameter of M .

We also have a similar result for general Riemannian 2-surfaces. Let γ be any simple closed curve on such a surface that can be contracted to a point over closed curves of length at most L . Then, for any $\varepsilon > 0$ and any $q \in \gamma$, there exists a homotopy that contracts γ to q over loops that are based at q and have length bounded by $3L + 2d + \varepsilon$. Here, d is the diameter of the surface.

Introduction.

The central question that we study in this paper is the following. Assume that M is a smooth surface, possibly with a boundary, and that γ is a simple (non-self-intersecting) closed curve on M . Assume that γ can be contracted to a point via free loops (closed curves) of length at most L . We would like to contract γ over closed curves based at a point $q \in \gamma$ so that the maximal length of these curves is as small as possible. Can we estimate the required maximal length in terms of L and the diameter d of M ? Such a result would have a large number of applications, some of which will be discussed at the end of introduction. This problem is already interesting when M is a 2-disc endowed with a Riemannian metric and γ is its boundary. The simple example of a Riemannian metric on M that looks like a long thin finger shown in Figure 1 demonstrates that we cannot estimate the required length in terms of L alone, and this example suggests that at the very least we need to add a summand equal to $2d$. In this example, we can contract the boundary of the disc via short closed curves to a point p far from ∂M (see Figure 1(a)). To replace this homotopy by one composed of loops based at a point $q \in \partial M$, we

connect p and q by a minimizing geodesic τ . In the course of our new homotopy, we travel along τ to one of the closed curves in the original homotopy, travel along this curve, and then return back along τ (Figure 1(b)). At some moment we end up at a loop that consists of two copies of τ traversed in opposite directions. This loop can then be contracted to q along itself.

Of course, there are other, more complicated Riemannian metrics on the 2-disc, such as the metric depicted in Figure 2. There is also the family of Riemannian metrics considered in [FK]. For these metrics, the connection between the length of curves in the “best” free loop homotopy and the length of curves in the “best” fixed point homotopy is not so evident.

Our first theorem asserts that adding the summand $2d$ and an arbitrarily small ε to L will always suffice. It is quite possible that one does not need ε , but this does not seem to follow from a compactness argument, as when $\varepsilon \rightarrow 0$, our homotopies can become wigglier and wigglier. (More formally, we have not established a control over the Lipschitz constants of our homotopies as $\varepsilon \rightarrow 0$.)

Theorem 0.1 *Let M be a Riemannian manifold with boundary diffeomorphic to the standard disc of dimension 2. Denote its diameter by d . Suppose there exists a homotopy connecting the boundary ∂M of M to some point $p \in M$ such that the length of every closed curve in this homotopy does not exceed a real number L .*

Then, for any $q \in \partial M$ and for any $\varepsilon > 0$, there exists a fixed point homotopy that connects ∂M with q , and passes through loops that are based at q and have length not exceeding $L + 2d + \varepsilon$.

Our second theorem deals with the general case of a simple contractible curve on a surface endowed with a Riemannian metric.

Theorem 0.2 *Let M be a closed Riemannian surface of diameter d . Let $\gamma : [0, 1] \rightarrow M$ be a simple closed curve in M , and q a point on γ . If there exists a homotopy between γ and a point that passes through closed curves of length not exceeding L , then there exists a homotopy that contracts γ to q through loops that are based at q and have length $\leq 3L + 2d + \varepsilon$.*

Our proof of Theorem 0.1 and our proof of Theorem 0.2 each uses the following theorem of independent interest proven by Gregory R. Chambers and Yevgeny Liokumovich in [CL]:

Theorem 0.3 *(G. R. Chambers, Y. Liokumovich [CL]) Let M be a Riemannian surface. Let $\gamma(t) : S^1 \rightarrow M$ be a closed curve in M . Suppose there exists a homotopy $H(t, \tau) : S^1 \times [0, 1] \rightarrow M$ such that $H(t, 0) = \gamma(t)$; $H(t, 1) = p \in M$, and the length of $\gamma_\tau = H(*, \tau)$ is at most L for all $\tau \in [0, 1]$. Then,*

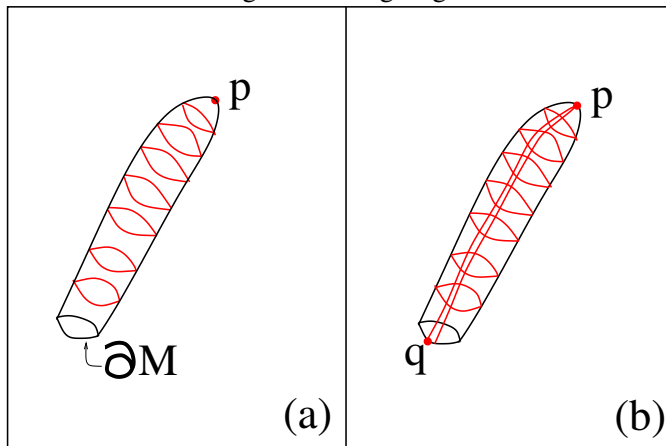
for any $\varepsilon > 0$, there exists a homotopy $\tilde{H}(t, \tau) : S^1 \times [0, 1] \rightarrow M$, such that $\tilde{H}(t, 0) = \gamma(t)$; $\tilde{H}(t, 1) = p$; the length of $\tilde{\gamma}_\tau = \tilde{H}(*, \tau)$ is at most $L + \varepsilon$, and $\tilde{\gamma}_\tau(t)$ is a simple closed curve for every $\tau \in [0, 1]$.

This result immediately reduces both Theorem 0.1 and Theorem 0.2 to the particular case of when the original homotopy contracting ∂M or γ is known to pass through *simple* closed curves of length $\leq L$. (Note that we did not make this assumption in the text of the theorem.) Let us state the particular case of Theorem 0.1 that we are going to prove below, and which implies the general case:

Proposition 0.4 *Let M be a Riemannian manifold with boundary diffeomorphic to the 2-dimensional disc. Let $H(t, \tau) : S^1 \times [0, 1] \rightarrow M$ be a homotopy between $\gamma(t) = \partial M$ and $p \in M$ which passes through simple closed curves of length at most L . Then, for any $\varepsilon > 0$ and any $q \in \partial M$, there exists a homotopy $H_q(t, \tau) : S^1 \times [0, 1] \rightarrow M$ over closed curves of length at most $L + 2d + \varepsilon$, where d is the diameter of M , such that $H_q(t, 0) = \gamma(t)$, $H_q(t, 1) = q$, and $H_q(s_0, \tau) = q$ for all $\tau \in [0, 1]$ and a fixed base point $s_0 \in S^1$.*

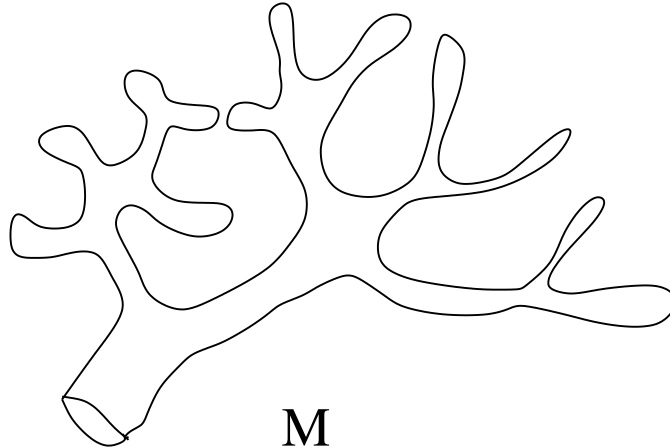
To reiterate, the first step in producing a “short” fixed point homotopy out of a free loop homotopy is to continuously deform all of the curves in the original homotopy so that the lengths of the resulting curves do not increase by much, and so that the resulting curves no longer have self-intersections. This is also the first step in proving Theorem 0.2; the other main steps of its proof will be briefly discussed towards the end of the introduction.

Figure 1: Long finger



The questions considered in this paper fall within the realm of questions of investigating geometric properties of “optimal” homotopies.

Figure 2: “Cactus” metric on the disc

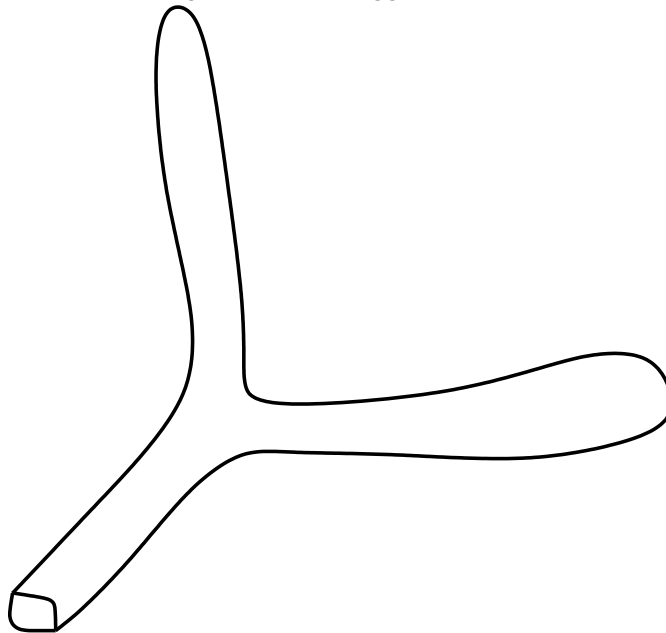


Another example of such a question was a long-standing question of S. Frankel and M. Katz posed at the end of their paper [FK]. They asked if one contract the boundary ∂M of a Riemannian 2-disc M so that the length of curves in the homotopy is majorized above in terms of the diameter d , area A , and the length of the boundary of the manifold? Note that this question is a modification of an earlier question asked by M. Gromov ([Gr], p. 100). The positive answer to this question was given by Y. Liokumovich, A. Nabutovsky and the second author of this paper in [LNR]. In particular, it was shown that ∂M can be contracted over curves of length at most $|\partial M| + 200d \max\{1, \ln \frac{\sqrt{A}}{d}\}$, where $|\partial M|$ is the length of ∂M . This estimate is optimal up to a multiplicative factor in the second term. When $\frac{\sqrt{A}}{d} \ll 1$, [LNR] provides a better bound of $2|\partial M| + 2d + 686\sqrt{A}$. This was improved to the asymptotically tight upper bound $|\partial M| + 2d + O(\sqrt{A})$ by P. Papasoglu in a recent paper [P]. Note that it is impossible to bound the length of curves in the best homotopy solely in terms of the area of M , as an example of a three-legged star fish with long tentacles depicted in Figure 3 demonstrates. It is also impossible to bound the length solely in terms of the diameter of M , as was proven by Frankel and Katz in [FK], answering the original version of the question of M. Gromov mentioned above.

A closely related family of questions deals with establishing the existence of various upper bounds on the maximal length of optimal sweep-outs and slicings of surfaces either by closed curves or, more generally, by cycles. For example, it was shown by Y. Liokumovich that there does not exist a universal diameter bound for the maximal length of curves or cycles in an optimal sweep-out of a closed Riemannian surface (see [L1] and [L2]). On the other hand, F. Balacheff

and S. Sabourau have found an upper bound for the maximal length of a cycle in an optimal sweep-out of a surface in terms of the genus and the area of the surface (see [BS]). Also, a “short” sweep-out of a Riemannian 2-sphere is possible if one assumes that there is no “short” geodesics of index 0. This follows from the results of C. B. Croke in [C]. In this case, the maximal length of a cycle can be bounded by the area or the diameter of the surface.

Figure 3: Three legged star fish



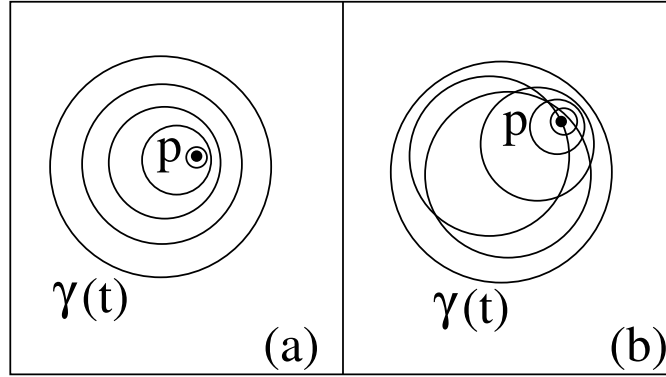
Another interesting problem concerning optimal homotopies is the following. Assume that the boundary of a Riemannian 2-disc can be contracted to a point via *simple* closed curves of length $\leq L$. Let us call such a homotopy *monotone* if the 2-discs bounded by the closed curves of this homotopy “decrease” in the sense of the definition below. Is it always possible to find a *monotone* homotopy contracting the boundary of the 2-disc via closed curves of length $\leq L$? In order to pose this question more formally we need the following definition.

Definition 0.5 Let M be a Riemannian manifold with boundary diffeomorphic to that of the 2-disc. Let $H(t, \tau) : S^1 \times [0, 1] \rightarrow M$ be a smooth map such that $H(t, 0) = \gamma(t) = \partial M$, $H(t, 1) = p \in M$, and $H(t, \tau) = \gamma_\tau(t)$ is a simple closed curve parametrized by t for each $\tau \in [0, 1]$. We will say that H is a *monotone* homotopy if closed 2-discs $D_\tau \subset M$ bounded by γ_τ satisfy the inclusion $D_{\tau_2} \subset$

D_{τ_1} for every τ_1 and τ_2 with $\tau_1 < \tau_2$. If these discs satisfy a stronger condition that $D_{\tau_2} \subset \text{int}D_{\tau_1}$, where $\text{int}D_{\tau_1}$ is the interior of D_{τ_1} , then the homotopy will be called strictly monotone.

Figure 4(a) depicts a strictly monotone homotopy of $\gamma(t)$ to the point p , while Figure 4(b) depicts a homotopy that is not monotone. This definition can be trivially extended to homotopies connecting a simple closed contractible curve to a point on any closed Riemannian surface not diffeomorphic to the 2-disc. There is one technicality, however; if M is diffeomorphic to S^2 , then there is an ambiguity due to non-uniqueness of D_τ . We agree to resolve it by allowing any possible choice of the system of discs D_τ that is continuously dependent on τ , and that has the monotonicity (or strict monotonicity) property.

Figure 4: Monotone and non-monotone homotopies



The following theorem provides an answer to the above question about the existence of monotone homotopies passing through “short curves”.

Theorem 0.6 (Monotonicity Theorem) *Let M be a Riemannian manifold with boundary diffeomorphic to the 2-dimensional disc D . Let $H(t, \tau) : S^1 \times [0, 1] \rightarrow M$ be a homotopy between the boundary of M and some point $p \in M$ passing through simple closed curves $\gamma_\tau(t)$ of length at most L . Then, for any $\varepsilon > 0$, there exists a strictly monotone homotopy $\tilde{H}(t, \tau) : S^1 \times [0, 1] \rightarrow M$ between $\gamma(t)$ and $\tilde{p} \in M$ over curves of length at most $L + \varepsilon$.*

From a technical point of view this is the central result of the paper, as it can be used to give the following short proof of Proposition 0.4 (which implies Theorem 0.1, as we explained above).

Proof of Proposition 0.4. By Theorem 0.6, there exists a strictly monotone homotopy H between $\gamma(t)$ and $\tilde{p} \in M$ over simple curves of length at most $L + \varepsilon$. Fix a

point q on ∂M , and let $\alpha(s) : [0, 1] \rightarrow M$ be a minimal geodesic connecting q to \tilde{p} . The length of α is at most d . For each $\tau \in [0, 1]$, there is exactly one $\tau' \in [0, 1]$ such that the curve $H(*, \tau')$ goes through $\alpha(\tau)$. Let this curve be denoted by H_τ . Note that, if α intersects a curve in H multiple times, then we will be able to find multiple values for τ that result in the same curve.

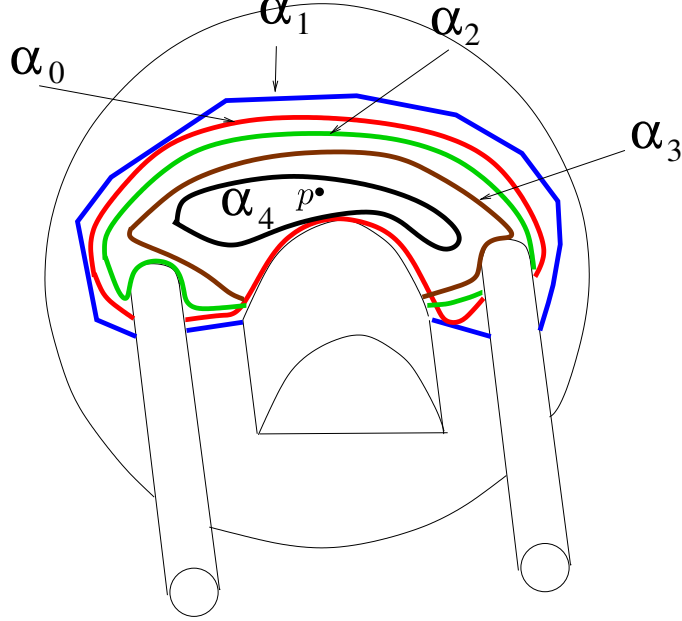
Our new contraction $G : S^1 \times [0, 1] \rightarrow M$ of γ through curves based at q is now defined as follows. For each $\tau \in [0, 1]$, define $G(*, \tau)$ to be the curve

$$\alpha|_{[0, \tau]} * H_\tau * \overline{\alpha|_{[0, \tau]}},$$

where $\overline{\alpha|_{[0, \tau]}}$ is the segment of α traversed from τ to 0. Each curve in this homotopy is bounded in length by $L + 2d + \varepsilon$. Furthermore, it ends at $\alpha|_{[0, 1]} * \overline{\alpha|_{[0, 1]}}$, which can obviously be contracted to q through curves based at q of length no more than $2d$. This completes the proof. \square

Note that the Monotonicity Theorem *does not* necessarily hold when one has a simple curve on a surface, or even a simple closed curve in a disc that is not assumed to be the boundary of that disc. This fact makes proving Theorem 0.2 more difficult than Theorem 0.1, and is the reason for the appearance of the extra $2L$ in our upper bound. We are grateful to Yevgeny Liokumovich for first attracting our attention to this fact in conjunction with the example shown in Figure 5. This figure depicts a metric on a Riemannian 2-disc and a curve α_0 such that the optimal homotopy contracting the curve to a point is not monotone. Notice that there are three bumps depicted in this figure: two of them are long and thin and the one in the middle is short and asymmetric. It takes less length to go under the middle bump than over it. The original curve α_0 winds around the two thin bumps, and goes over the short one. In order to contract α_0 to a point, it has to be stretched over the thin bumps but, because they are long, the length of the curve will necessarily increase in the process. Thus, to begin with, it makes sense to first homotope α_0 to α_1 , which runs below the middle bump. α_1 is shorter than α_0 , so we can “spend” this “excess” length on dragging the curve over the two thin bumps one at a time. This corresponds to the curves α_2 and α_3 in Figure 5. We now have to push the curve over the middle bump. This homotopy results in α_4 , which can then be easily contracted to a point. The resulting homotopy is not monotone. Note that α_0 is not the boundary of the disc, and so this example does not contradict Theorem 0.6.

For surfaces, we will need a weaker form of the Monotonicity Theorem. We will not prove that γ can be contracted by a strictly monotone homotopy via “short” curves, but only that there exists a strictly monotone homotopy contracting another curve $\tilde{\gamma}$ via curves of length $\leq L + \varepsilon$ so that γ will be contained in the disc formed by the image of this monotone homotopy. More formally:

Figure 5: α_0 cannot be contracted to p via a “short” monotone homotopy

Definition 0.7 Let γ be a simple closed contractible curve on a Riemannian surface M . We will say that a homotopy $H(t, \tau) : S^1 \times [0, 1] \rightarrow M$ covers γ if the following conditions are satisfied:

- (1) $H(t, \tau)$ is a strictly monotone homotopy between some simple closed curve $\tilde{\gamma} = H(*, 0)$ and a point $p \in M$;
- (2) The disc $D_{\tilde{\gamma}} = H(S^1 \times [0, 1])$ generated by the image of this homotopy contains γ . (Note that $\partial D_{\tilde{\gamma}} = \tilde{\gamma}$.)

Theorem 0.8 Let M be a two-dimensional Riemannian manifold or a Riemannian manifold with boundary. Let $H(t, \tau) : S^1 \times [0, 1] \rightarrow M$ be a homotopy between a simple closed contractible curve γ and a point $p \in M$ over simple closed curves of length at most L . Then, for any $\varepsilon > 0$, there exists a homotopy $\tilde{H}(t, \tau)$ over curves of length at most $L + \varepsilon$ that covers γ (in the sense of the above definition).

Note that if M is a 2-disc, then this theorem is equivalent to the Monotonicity Theorem as the homotopy that covers $\gamma = \partial M$ can only begin with γ , and so Theorem 0.8 implies Theorem 0.6. The above theorem will be proven in the next section. In order to deduce Theorem 0.2 from Theorem 0.8 we need the following theorem which, in our opinion, is of independent interest:

Theorem 0.9 *Let M be a closed Riemannian manifold of diameter d . Let γ be a simple closed contractible curve in M . Assume that there exists a homotopy $H(t, \tau) : S^1 \times [0, 1] \rightarrow M$ over curves of length at most L that covers γ . Then, for any point $q \in \gamma$ and for any $\varepsilon > 0$, there exists a contraction $\tilde{H}(t, \tau) : S^1 \times [0, 1] \rightarrow M$ of γ over loops of length at most $3L + 2d + \varepsilon$ that are based at q . Furthermore, there is a specific point $q^* \in \gamma$ such that we can find a contraction of γ through loops based at q^* of length bounded by $2L + 2d + \varepsilon$.*

This theorem will be proven in the last section. It is obvious that Theorem 0.2 immediately follows from Theorems 0.8 and 0.9.

Applications. Theorem 0.2 and Theorem 0.6 have many immediate and potential applications to geometry of loop spaces on Riemannian 2-spheres, to questions about the lengths of geodesics as well as to problems about optimal sweep-outs.

In particular, Theorem 0.2 provides a canonical way of obtaining a “short” based loop homotopies out of the “short” free loop homotopy on Riemannian surfaces. The second author has encountered this problem many times, and each time it was solved on an ad hoc basis. Specifically, Theorems 0.2 and 0.6 can be applied in the following situations. The details of each of these applications will appear elsewhere.

(1) **Lengths of geodesics on Riemannian 2-spheres.** Let p, q be an arbitrary pair of points on a Riemannian 2-sphere M of diameter d . A. Nabutovsky together with the second author have demonstrated that there exist at least k geodesics joining them of length at most $22kd$ (see [NR2]). The bound will be $20kd$, when $p = q$ (see [NR1]). We have noticed that an application of Theorem 0.2 can dramatically decrease the complexity of proofs in [NR1] and [NR2], and improve the bounds in [NR2] to $16kd$, (the bounds of [NR1] will be improved to $14kd$). These improvements are due to the fact that the main technical difficulty of [NR1] and especially [NR2] is a formation of possible intersections between various closed curves in the homotopies between a closed curve and a point.

(2) **Geometry of the loop spaces on Riemannian 2-spheres.**

(a) Application of Theorem 0.2 will immediately help to generalize Theorem 1.1 of [NR3] to the *free* loop space on a Riemannian 2-sphere M . To be more precise, one can show that any map $f : S^m \rightarrow \Lambda M$, where ΛM is a free loop space on M , is homotopic to a map $\tilde{f} : S^m \rightarrow \Lambda M$ that passes through curves of length that does not exceed $L = L(m, k, d)$, where d is the diameter of M , k is the number of distinct non-trivial periodic geodesics on M of length at most $2d$, and L is a function of m, k, d that can be written down explicitly. Moreover, one can explicitly majorize the lengths of loops in an “optimal” homotopy connecting f and \tilde{f} in terms of n, k, d and $\sup_{x \in S^m} \text{length}(f(x))$.

(b) The first author and Y. Liokumovich have some concrete ideas of how appli-

cations of Theorems 0.3 and 0.6 can lead to an answer of the following question posed by N. Hingston and H.-B. Rademacher: Let α be a generator of $H_1(\Lambda M)$, where ΛM is the free loop space of a manifold M diffeomorphic to S^2 . Is it possible for the minimax level of 2α (with respect to the length functional on ΛM) to be lower than the minimax level of α ?

(3) **A lower bound for the diastole on a Riemannian 2-sphere of area A and diameter d .** Y. Liokumovich points out that by using Monotonicity Theorem together with the examples of Frankel and Katz from [FK] one can construct an example of Riemannian metrics on S^2 with $A \gg d^2$ that cannot be sliced into simple closed curves of length $\leq \frac{2}{\ln 3} d \ln \frac{\sqrt{A}}{d}$. This result complements Theorem 1.3 in [LNR].

1 Proof of Theorem 0.8

1.1. Scheme of the proof. Our proof of Theorem 0.8 goes as follows. First, we construct a finite set of “short” monotone homotopies that pass through simple closed curves which are homotopic to γ and are of length bounded by $L + \varepsilon$. These monotone homotopies will either overlap or “quasi-overlap” in a certain way. This fact will enable us to “glue” (or, more precisely, “process”) them one by one to obtain a monotone homotopy that passes through simple curves of length $\leq L + \varepsilon$, ends at a point, starts at a simple closed curve homotopic to γ , and the disc generated by this homotopy contains γ . Note that the gluing process is not straightforward. The image of the resulting homotopy is, in general, not the union of images of the two overlapping homotopies that we are trying to combine.

The homotopy that we are going to obtain will only be monotone and not strictly monotone. The loss of monotonicity will however happen only along a finite set of curves. Therefore, one can then obtain a strictly monotone homotopy by inserting collars along each of these curves, regarding each collar as a collection of copies of the original curve, and disengaging curves in the monotone homotopy using the different copies. We will omit the rather obvious but tedious details of this step of the proof and construct only a monotone homotopy.

1.2. Filling contractible closed curves by discs. A simple contractible closed curve α on a surface always bounds a 2-disc. This disc is unique, unless the surface is diffeomorphic to S^2 . In this case one has two choices for such a disc. Let γ be a simple closed curve and H a homotopy that contracts γ to a point through simple closed curves. In this case, we use the following method to assign discs D_{γ_τ} bounded by γ_τ to all curves γ_τ that appear during this homotopy. When γ_τ becomes very small for some value of τ , we choose the smaller of two discs. We then extend the choice of the disc by continuity. Note that making a choice between

two discs bounded by a curve is equivalent to orienting this curve. Thus, we have just oriented all curves appearing in the homotopy H between γ and a point. Our choice of orientation has nothing to do with how the closed curves are parametrized in the homotopy.

Note also that if a (smooth) simple closed curve α has already been oriented, then we can orient each simple closed curve β in a tubular neighbourhood of α by choosing D_β so that $D_\beta \setminus D_\alpha$ is contained in the same tubular neighbourhood of α .

If α_1 and α_2 are two contractible simple closed curves that have already been oriented, and the closure of D_{α_2} is contained in the closure of D_{α_1} , then we say that α_2 lies *inside* α_1 . Further, we call the closure of $D_{\alpha_1} \setminus D_{\alpha_2}$ an *oriented annulus* bounded by α_1 and α_2 and refer to α_2 as to its *inner boundary* and α_1 as its *outer boundary*. Note that α_1 and α_2 are allowed to touch each other, and if they do, the oriented annulus is not diffeomorphic to $S^1 \times [0, 1]$.

Given an oriented annulus, we can orient each simple closed curve β that both lies in this annulus and is homotopic to its inner boundary α_1 by choosing D_β so that it contains D_{α_1} .

When M is diffeomorphic to S^2 and we are given a simple closed curve γ and a homotopy H that contracts it to a point through simple closed curves, we will automatically adopt the conventions about orienting simple closed curves indicated above to obtain a *continuous* orientation on a class of simple curves that will include all closed curves relevant to our proof.

1.3. Simple intersections.

Definition 1.1 Let $\beta_1(t) : [0, 1] \rightarrow M$ and $\beta_2(t) : [0, 1] \rightarrow M$ be two closed curves in a Riemannian manifold M .

We will say that $\beta_1(t)$ and $\beta_2(t)$ satisfy the simple intersection property if, for every two points of intersection of β_1 and β_2 , they are consecutive on β_1 if and only if they are consecutive on β_2 .

1.4. Monotone homotopy covers.

Let γ be a contractible simple closed curve on M , and suppose we are given a homotopy H from γ to a point that passes through simple closed curves. If M is diffeomorphic to S^2 , then this homotopy induces orientations on a family of closed curves as described in section 1.2. If M is not a 2-sphere, all contractible curves automatically bound unique discs.

Consider an oriented annulus (in the sense of definition given in section 1.2). Assume that we can choose a monotone homotopy that starts at its outer boundary and ends at its inner boundary. Such an annulus together with the monotone homotopy will be called an *oriented monotone annulus*. Let A_1 and A_2 be two

oriented monotone annuli. If the inner boundary of A_1 is contained inside of the outer boundary of A_2 , we say that the annuli A_1 and A_2 are *related*. Note that this definition is *not* symmetric with respect to the order of A_1 and A_2 . Also note that related annuli can, in principle, be disjoint as A_1 can be contained inside the disc bounded by the inner boundary of A_2 .

Assume that for some n , a collection of n oriented monotone annuli has the property that for each $i = 1, \dots, n-1$, A_i and A_{i+1} are related. Assume further that the inner boundary of A_n is a point. If γ is the outer boundary of A_1 and D_γ is equal to the disc bounded by the outer boundary of A_1 , we say that A_1, \dots, A_n form a *monotone homotopy cover* of γ corresponding to H . If, instead, the disc D_γ bounded by γ is contained in the 2-disc bounded by the outer boundary of A_1 , we say that A_1, \dots, A_n form a *weak monotone homotopy cover* of γ corresponding to H .

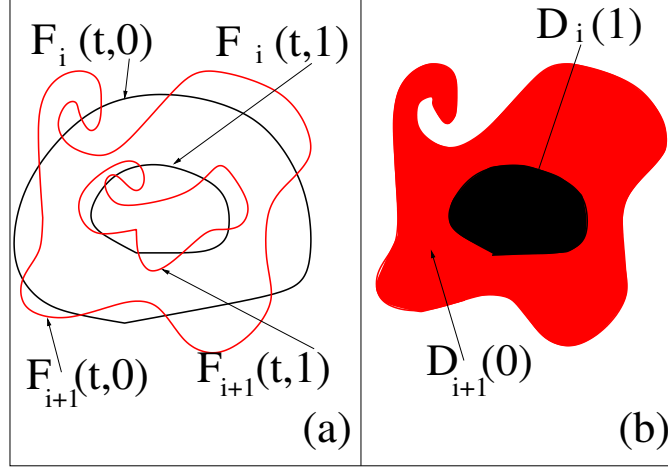
For the convenience of the reader, we give the following self-contained definitions of monotone homotopy cover and weak monotone homotopy cover that do not use the terminology introduced in section 1.2.

Definition 1.2 *Let $\gamma_\tau(t), \tau \in [0, 1]$ be a continuous 1-parametric family of simple closed curves starting from a closed curve $\gamma = \gamma_0(t)$ and ending with a point p . Assume that all curves γ_τ were continuously oriented in the sense of section 1.2, that is, there is a continuous family of discs D_{γ_τ} bounded by γ_τ . (This condition is relevant only if M is diffeomorphic to S^2 .) We will say that n homotopies $F_i : S^1 \times [0, 1] \rightarrow M, i = 1, \dots, n$, form a monotone homotopy cover of γ corresponding to the family $\gamma_\tau(t)$ if the following conditions are satisfied:*

1. $F_0(t, 0) = \gamma_0(t)$ and $F_n(t, 1) = \tilde{p}$ for some point $\tilde{p} \in M$.
2. For each i there is a continuous 1-parametric family $D_i(x)$ of closed 2-discs in M satisfying the condition that the boundary of $D_i(x)$ is the closed curve $F(S^1, x)$. $D_1(0)$ coincides with D_{γ_0} . For each i , $D_i(x_1)$ is contained in $D_i(x_2)$ if $x_1 > x_2$. (The last condition means that, for each i , F_i is a monotone homotopy of closed curves.)
3. These n families of discs satisfy the following condition: For all $i \in \{1, \dots, n-1\}$ $D_i(1) \subset D_{i+1}(0)$. Figure 6(a) and Figure 6(b) depict this condition.

If, instead of assuming that $\gamma_0 = \partial D_1(0)$ and $D_{\gamma_0} = D_1(0)$, we assume only that D_{γ_0} is contained in $D_1(0)$, we say that the n homotopies F_i form a *weak monotone homotopy cover* of γ corresponding to the family γ_τ .

Figure 6: Monotone homotopy cover



Definition 1.3 Let $\{F(t, x)\}_{i=0}^n$ be a (weak) monotone homotopy cover. We will call $M = \max_{i \in \{0, \dots, n\}} \sup_{x \in [0, 1]} \text{length}(F_i(S^1, x))$ its length.

1.5. The existence of monotone homotopy covers.

Here is our first major technical result:

Lemma 1.4 1. Let M be a Riemannian manifold with boundary diffeomorphic to the 2-disc. Let $\gamma_\tau(t)$, $\tau \in [0, 1]$ be a 1-parameter family of simple smooth curves in M parametrized by their arclength with $\gamma_0 = \gamma = \partial M$. If the length of curves in this family is bounded by L , then for any $\varepsilon > 0$, there exists a monotone homotopy cover of γ_τ of length at most $L + \varepsilon$.

2. Let M be a Riemannian surface. Let γ be a closed simple contractible curve on M . Let $\gamma_\tau(t)$, $\tau \in [0, 1]$ be a 1-parameter family of simple closed curves of length at most L , such $\gamma_0(t) = \gamma$, and $\gamma_1(t) = p \in M$. Then, for any $\varepsilon > 0$, there exists a weak homotopy cover of γ corresponding to $\gamma_\tau(t)$ of length at most $L + \varepsilon$.

Proof. The proof of the lemma consists of the following steps. First we discretize the given homotopy. We choose an increasing sequence of n points $\tau_i \in [0, 1]$, $i = 1, \dots, n$, so that the curves γ_{τ_i} satisfy the following conditions:

1. The first curve in this collection is γ_0 (that is, $\tau_1 = 0$).
2. The last curve in this collection is very short and is contained in a very small neighbourhood of p . In particular, it is smaller than the injectivity radius of

M . We choose the radius of this neighbourhood to be so small that γ_n could be contracted to a point inside the small disc that it bounds via curves of length $< \text{length}(\gamma_n) + \varepsilon$.

3. For all values of i , the curves γ_{τ_i} and $\gamma_{\tau_{i+1}}$ are normal variations of each other. This means that:
 - (a) The curve $\gamma_{\tau_{i+1}}$ is inside the r_i -neighbourhood of γ_{τ_i} , where r_i denotes the injectivity radius of the normal exponential map for γ_{τ_i} . Similarly, γ_{τ_i} must be in the r_{i+1} -neighbourhood of $\gamma_{\tau_{i+1}}$.
 - (b) There exists a function $\alpha(t)$ such that $\gamma_{\tau_{i+1}}$ is a reparametrization of the curve $\gamma_{\tau_i}(t) + \alpha(t)n(\gamma_{\tau_i}(t))$, where $n(\gamma_{\tau_i}(t))$ denotes the outer unit normal to γ_{τ_i} at $\gamma_{\tau_i}(t)$. Similarly, there exists a function $\beta(t)$ such that, up to a reparametrization, γ_{τ_i} coincides with $\gamma_{\tau_{i+1}}(t) + \beta(t)n(\gamma_{\tau_{i+1}}(t))$.

To ensure this final property, we simply observe that the infimum r of the injectivity radii of the normal exponential map for γ_τ , $\tau \in [0, \tau_n]$ is positive. As such, we just need to choose τ_i so that γ_{τ_i} and $\gamma_{\tau_{i+1}}$ are $r/2$ -close in C^0 -metric and sufficiently close in C^1 -metric.

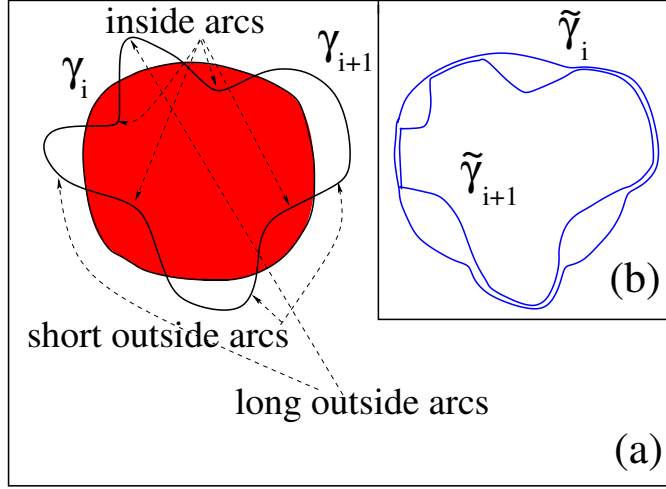
Let us redenote curves γ_{τ_i} by γ_i . Property (3) easily implies that if γ_i and γ_{i+1} intersect, then they satisfy the simple intersection property (see section 1.3). If they do not intersect, then one of them is inside the other. In other words, the situation depicted in Figure 8(c) is impossible. (Here, we are assuming that all the relevant curves were oriented as described in section 1.2, if M is a 2-sphere.)

Note that the n th (last) piece of a (weak) monotone cover will contract γ_n to a point inside the small disc bounded by γ_n . It remains to construct the remaining pieces of the (weak) monotone homotopy cover so that γ_n is outside of the inner boundary of the annulus generated by the $(n - 1)$ th piece.

This construction has two stages. In the first of these stages, for each pair of curves γ_i and γ_{i+1} , we construct two new curves $\tilde{\gamma}_i$ and $\tilde{\gamma}_{i+1}$ such that they are still very close to each other, and $\tilde{\gamma}_{i+1}$ is inside $\tilde{\gamma}_i$. In the second stage we connect these pairs of closed curves by monotone homotopies F_i . For every i , $F_i(t, 0) = \tilde{\gamma}_i(t)$, and $F_i(t, 1) = \tilde{\gamma}_{i+1}(t)$.

To describe the first of these two stages, consider the case of when γ_i and γ_{i+1} intersect. Property (3) above implies that they have the simple intersection property. Consider two cyclically ordered sequences of all points $\{t_j\}_{j=0}^m$ and $\{s_j\}_{j=0}^m$ such that $\gamma_i(t_j) = \gamma_{i+1}(s_j)$.

The curves $\tilde{\gamma}_i$ and $\tilde{\gamma}_{i+1}$ are constructed from arcs of γ_i and γ_{i+1} in the following way. Let us consider each arc $\gamma_i|_{[t_j, t_{j+1}]}$ and $\gamma_{i+1}|_{[s_j, s_{j+1}]}$. First, let us ask whether $\gamma_{i+1}|_{[s_j, s_{j+1}]}$ lies inside the closed disc bounded by $\gamma_i(t)$, or outside of it, as in Figure 7(a). (If M is a 2-sphere, then the choice of the closed disc is determined

Figure 7: Constructing $F_i(t, 0)$ and $F_i(t, 1)$ 

by the chosen orientation of the curve; see section 1.2.) If this arc lies inside the disc, then the arc of γ_i parametrized by the interval $[t_j, t_{j+1}]$ and the arc of γ_{i+1} parametrized by the interval $[s_j, s_{j+1}]$ will remain unchanged and will constitute arcs of $\tilde{\gamma}_i$ and $\tilde{\gamma}_{i+1}$, respectively, as in Figure 7(b). If, however, $\gamma_{i+1}|_{[s_j, s_{j+1}]}$ lies outside the disc, we will then ask which is longer, $\gamma_{i+1}|_{[s_j, s_{j+1}]}$ or $\gamma_i|_{[t_j, t_{j+1}]}$ (see Figure 7(a)). If the former arc is shorter, then both $\tilde{\gamma}_i$ and $\tilde{\gamma}_{i+1}$ will be equal to $\gamma_{i+1}|_{[s_j, s_{j+1}]}$, if the latter arc is shorter, then both curves will become $\gamma_i|_{[t_j, t_{j+1}]}$, as shown in Figure 7(b).

Note that it is possible that the set of intersection points of γ_i and γ_{i+1} is empty. In that case γ_{i+1} either lies inside the disc that γ_i bounds as in Figure 8(a), or outside of it as in Figure 8(b). (Note that, as described above, the situation depicted in Figure 8(c) is impossible.) If γ_{i+1} lies in the disc bounded by γ_i , we will let $\tilde{\gamma}_i = \gamma_i$ and $\tilde{\gamma}_{i+1} = \gamma_{i+1}$. That is, everything will remain unchanged. If γ_{i+1} lies outside of the disc bounded by γ_i , then we choose the shorter of γ_i and γ_{i+1} , and set both $\tilde{\gamma}_i$ and $\tilde{\gamma}_{i+1}$ to this shorter curve. This case is pictured in Figure 8(b).

Now, after we have defined the map $F_i(t, x)$ at the endpoints of the interval $[0, 1]$, we would like to extend the map to its interior. Note that we can ensure that the curves $F_i(t, 0)$ and $F_i(t, 1)$ are constructed from γ_i and γ_{i+1} in such a way so that the C^0 and C^1 distances between them are only slightly larger than the C^0 and C^1 distances between γ_i and γ_{i+1} . In particular, we can ensure that these distances are no more than twice the original distances between γ_i and γ_{i+1} , which are uniformly bounded. To extend $F_i(t, *)$ to the whole interval, we will proceed as follows.

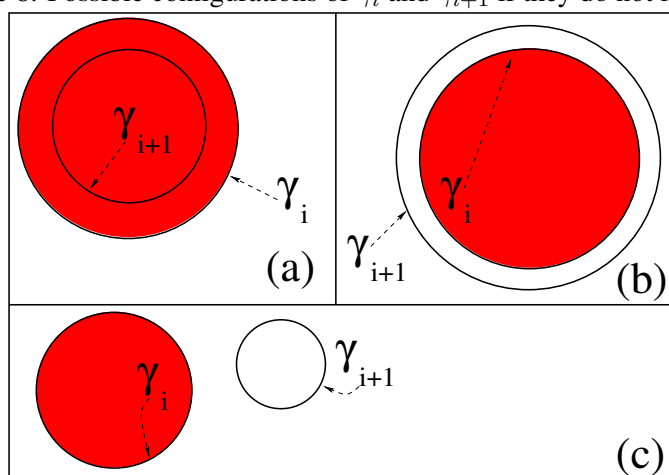
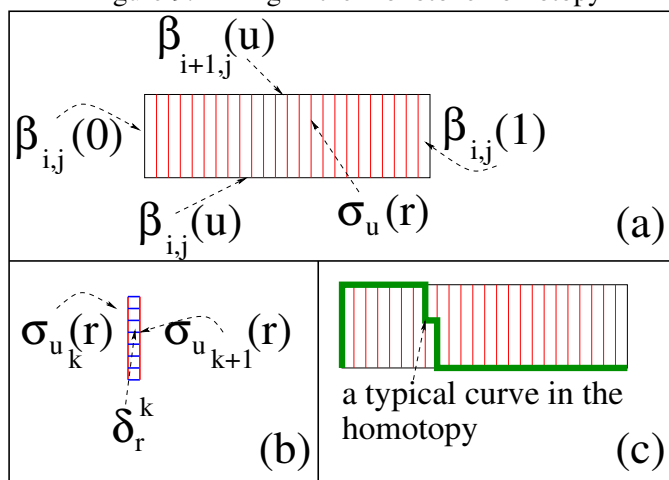
Figure 8: Possible configurations of γ_i and γ_{i+1} if they do not intersect

Figure 9: Filling in the monotone homotopy



Let us consider those arcs of $F_i(t, 0)$ and $F_{i+1}(t, 1)$ that are distinct. Parametrize each pair of these arcs proportionally to their arclength by the unit interval $[0, 1]$. Let us denote them as $\beta_{i,j}(u)$ and $\beta_{i+1,j}(u)$, as in Figure 9(a).

For any positive δ , we can choose our original points τ_i close enough together so that $d(\beta_{i,j}(t), \beta_{i+1,j}(t)) \leq \delta$. In particular, by making δ sufficiently small, we can insure that there exists a family of minimal geodesics $\sigma_u(r)$ of length at most δ continuously connecting the points on $\beta_{i,j}(u)$ to the points on $\beta_{i+1,j}(u)$, as shown in Figure 9(a). We then construct the path homotopy between $\beta_{i,j}$ and $\beta_{i+1,j}$ as follows.

Note that the lengths $l_{i,j}$ and $l_{i+1,j}$ of $\beta_{i,j}$ and $\beta_{i+1,j}$ are close to each other, (since δ is small), and let $0 = u_0 < u_1 < \dots < u_N = 1$ be a subdivision of $[0, 1]$ such that $\beta_{i,j}$ and $\beta_{i+1,j}$ are subdivided into equal number of small segments of length at most some small positive number $\tilde{\delta}$. We are assuming that the subdivision is fine enough to satisfy the following condition: for each pair of $\sigma_{u_k}, \sigma_{u_{k+1}}$, $k = 1, \dots, N - 1$, there is a continuous family of minimal geodesics δ_r^k of length at most $\tilde{\delta}$ that connects them, as shown in Figure 9(b). Moreover, we are assuming that the subdivision is sufficiently fine to ensure that if δ_r^k intersects β_i or β_{i+1} , then the intersection is simple in the sense of Definition 1.1.

The homotopy between $\beta_{i,j}$ and $\beta_{i+1,j}$ can be described as follows. It progresses through the sequence of curves $\beta_{i+1,j}|_{[0,u_1]} * \bar{\sigma}_{u_1} * \beta_{i,j}|_{[u_1,1]}$, $\beta_{i+1,j}|_{[0,u_2]} * \bar{\sigma}_{u_2} * \beta_{i,j}|_{[u_2,1]}$, etc. In general, it goes through curves of the form $\beta_{i+1,j}|_{[0,u_l]} * \bar{\sigma}_{u_l} * \beta_{i,j}|_{[u_l,1]}$. Here, $\bar{\sigma}_{u_l}$ refers to σ_{u_l} traversed in the opposite direction. The basic step is to construct a homotopy between $\beta_{i+1,j}|_{[0,u_l]} * \bar{\sigma}_{u_l} * \beta_{i,j}|_{[u_l,1]}$ and $\beta_{i+1,j}|_{[0,u_{l+1}]} * \bar{\sigma}_{u_{l+1}} * \beta_{i,j}|_{[u_{l+1},1]}$ for each l . The idea here is to use the curves $\beta_{i+1,j}|_{[0,u_l]} * \sigma_{u_l}|_{[r,1]} * \delta_r^l * \sigma_{u_{l+1}}|_{[0,r]} * \beta_{i,j}|_{[u_{l+1},1]}$ for all $r \in [0, 1]$. Figure 9(c) depicts the typical curve in this homotopy. This idea does not quite work, as we want a monotone homotopy, but δ_r^l can potentially have non-trivial intersections with the segments of either $\beta_{i,j}$ or $\beta_{i,j+1}$ between u_l and u_{l+1} . The idea is that, if this happens, we replace all arcs of δ_r^l outside of the curvilinear quadrilateral formed by $\sigma_{u_l}, \beta_{i,j}|_{[u_l, u_{l+1}]}, \beta_{i+1,j}|_{[u_l, u_{l+1}]}, \sigma_{u_{l+1}}$ by the arc of $\beta_{i,j}$ (or $\beta_{i+1,j}$) with the same endpoints. The simple intersection property between δ_r^l with $\beta_{i,j}$ and $\beta_{i+1,j}$ implies that using the replacement does not lead to discontinuities in the resulting family of curves. It is easy to see that the lengths of curves in the resulting monotone homotopy exceed L by a summand that can be made arbitrarily small by choosing sufficiently small parameters in the above algorithm.

Note that although this homotopy is monotone but not *strictly* monotone, it is not difficult to modify this construction to obtain a strictly monotone homotopy.

This completes the construction of the desired (weak) monotone homotopy covering.

Now it remains to check that, for every i , the inner boundary of the annulus corresponding to the i th homotopy is inside the outer boundary of the annulus corresponding to the $i + 1$ st homotopy. (Of course, these curves are allowed to wholly or partially coincide.) To explain this, note that when we were replacing each pair of curves γ_i and γ_{i+1} by the pair $\tilde{\gamma}_i$ and $\tilde{\gamma}_{i+1}$ in the second stage of the proof of the lemma, segments of γ_{i+1} were only moved inwards to obtain $\tilde{\gamma}_{i+1}$, and segments of γ_i were only moved outwards to obtain $\tilde{\gamma}_i$. Now note that both the inner boundary of the annulus corresponding to F_i and the outer boundary of the annulus corresponding to F_{i+1} were both obtained from the same curve γ_{i+1} on two different steps of the process. First, it was the inner curve and could move only inwards, then it was the outer curve and could only move only outwards. As such, the second curve cannot be inside of the first one.

A similar argument demonstrates that γ_n will be outside of the inner boundary of F_{n-1} that was obtained from γ_n , where some arcs of γ_n were allowed to move inwards.

□

1.6. Combining monotone homotopies. The second major ingredient of our proof of Theorem 0.8 is the following lemma that enables one to “combine” monotone homotopies:

Lemma 1.5 *Let M be a 2-dimensional Riemannian manifold (with or without boundary). Let $\gamma_x(t) = G(t, x) : S^1 \times [0, 1] \rightarrow M$ and $\beta_x(t) = H(t, x) : S^1 \times [0, 1] \rightarrow M$ be two monotone homotopies through simple closed curves of length at most L , and $D_{\gamma, x}, x \in [0, 1]$, $D_{\beta, x}, x \in [0, 1]$ two families of closed 2-discs that continuously depend on x and satisfy the following conditions: First, $\partial D_{\gamma, x} = G(*, x)$ for all $x \in [0, 1]$, and $D_{\gamma, x_1} \supset D_{\gamma, x_2}$ for all $x_1 < x_2$. Similarly, $\partial D_{\beta, x} = H(*, x)$ for all $x \in [0, 1]$, and $D_{\beta, x_1} \supset D_{\beta, x_2}$ for all $x_1 < x_2$. Second, $D_g = D_{\gamma, 1} \subset D_{\beta, 0} = D_h$.*

Then, for every positive ε , there exists a monotone homotopy $F(t, x) : S^1 \times [0, 1] \rightarrow M$ and a corresponding 1-parametric family of discs $D_F(x)$ continuously depending on $x \in [0, 1]$, such that:

1. $\partial D_F(x) = F(*, x)$ for all $x \in [0, 1]$.
2. $D_F(x_1) \supset D_F(x_2)$ for all $x_1 < x_2$.
3. For every x , the length of the simple closed curve $\partial D_F(x)$ is at most $L + \varepsilon$.
4. $D_F(0) \supset D_{\gamma, 0}$ and $D_F(1) \subset D_{\beta, 1}$.

Proof. Without loss of generality, let us assume that the homotopies G and H are strictly monotone and pass through smooth curves. Also assume that ∂D_h and ∂D_g are disjoint. If necessary, both of these properties can be achieved by an arbitrarily small perturbation of G and H . We will construct the homotopy in two steps:

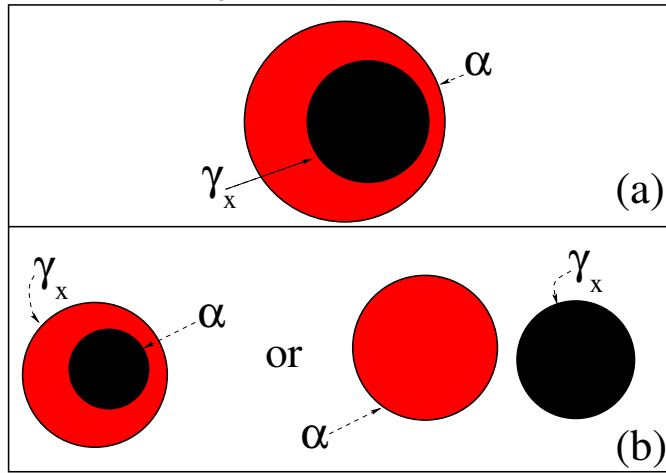
Let us consider the closure of the annulus $D_h \setminus D_g$. Let α be the shortest closed curve among all closed curves in this closed annulus that are homotopic to ∂D_h . Note that the length of α is at most L . It is easy to see that α is a simple closed curve. Let D_α be a closed domain diffeomorphic to the 2-disc that has α as its boundary and is contained in D_h .

Step 1. We will modify the homotopy $G(t, x)$ to obtain a new homotopy $\tilde{G}(t, x)$ so that the new homotopy will be monotone, $\tilde{G}(t, 0)$ will be outside of the disc bounded by $G(t, 0)$, and $\tilde{G}(t, 1) = \alpha(t)$.

First, we will construct a 1-parameter family of simple closed curves $\tilde{\gamma}_x(t)$ that can possibly have a finite number of discontinuities (as a function of the parameter x). We will then remove the discontinuities.

The new curves $\tilde{\gamma}_x(t)$ will be constructed as follows. The general principle is to “push” (arcs of) γ_x outside of the interior of the disc D_α . More specifically, if $D_{\gamma_x} \subset D_\alpha$ as in Figure 10(a), then we will let $\tilde{\gamma}_x(t) = \alpha(t)$. If $D_\alpha \subset D_{\gamma_x}$ or $D_\alpha \cap D_{\gamma_x} = \emptyset$ as in Figure 10(b), then we will let $\tilde{\gamma}_x(t) = \gamma_x(t)$.

Figure 10: γ_x relative to α



Now suppose that $D_{\gamma_x} \cap D_\alpha \neq \emptyset$ and that neither one is a subset of the other. Let us consider $0 = t_0 < t_1 < \dots < t_n = 1$, a subdivision of $[0, 1]$ such that $\gamma_x(t_i) = \alpha(s_{j(i)})$ and $\gamma_x \neq \alpha$ otherwise.

Let us consider those arcs of γ_x that are inside D_α . Clearly those arcs will be

inside D_h as well. However, also, by the monotonicity of the homotopy G , these arcs will be outside of D_g . Thus, they will lie in the closed annulus $cl(D_h - D_g)$, the closure of $D_h - D_g$.

Consider one such arc of γ_x between $\gamma_x(t_i)$ and $\gamma_x(t_{i+1})$. These two points coincide with points $\alpha(s_{j(i)})$ and $\alpha(s_{j(i+1)})$ which subdivide the curve α into two arcs. Let us select the arc A of α between these two points which is path homotopic to the arc of γ_x with the same endpoints which lies inside the closed annulus between D_h and D_g . Now replace this arc of γ_x by A . The resulting curve will not be contractible in the closed annulus between D_g and D_h . Also, the length of A does not exceed the length of the arc of γ_x that it replaced, as otherwise α would not be the shortest non-contractible curve in the closed annulus. Therefore, the resulting curve will have length bounded by the length of the original curve. Note that $\tilde{G}(t, 0)$ is outside of $G(t, 0)$ as (arcs of) curves in the homotopy G can move only outside. Also, by the hypotheses of lemma, $D_{\gamma,1} \subset D_\alpha$. As a result, $\tilde{G}(t, 1) = \alpha$.

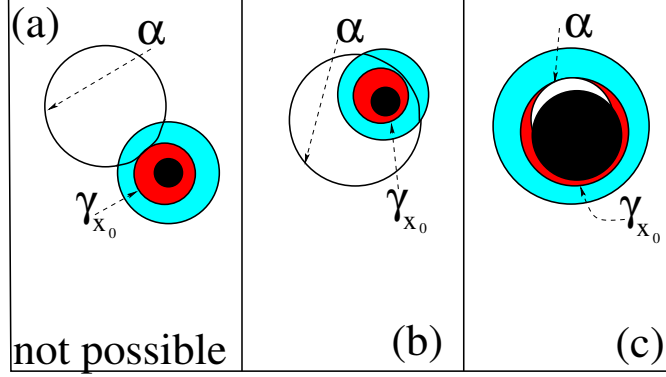
Additionally, note that the resulting map $\tilde{G}(t, x)$ regarded as a 1-parametric family of closed curves depending on the parameter x is monotone, but its dependence on x is not necessarily continuous.

The possible discontinuities may only occur at those curves γ_x that are tangent to α at some points. (Here it is convenient for us to think that α is smooth. While, in general, α will be only piecewise smooth, we can perturb it into an arbitrarily close smooth curve of arbitrarily close length. This replacement of α by a very close smooth curve adds a summand to our estimates that can be made arbitrarily small. Therefore, without loss of generality we can assume that α is smooth.) Without loss of generality, we can also assume that, for each value of x , there is at most one point where γ_x and α are tangent, and that all of these tangencies are non-degenerate. (If not, we can perturb G so that the resulting homotopy has these properties, and so that the curves increase in length only by an arbitrarily small amount.) Let us modify the family $\tilde{G}(t, x)$ so that the singularities at tangential points are resolved.

In order to do this, let us consider x_0 such that the curve γ_{x_0} is tangent to α at the point $\gamma_{x_0}(s_{x_0})$. Let us consider the following two cases:

Case A. The set of intersections of γ_{x_0} with the curve α consists of the one point $\gamma_{x_0}(s_{x_0}) = \alpha(s_{x_0})$. First, consider the situation in which $intD_{\gamma, x_0} \cap intD_\alpha = \emptyset$ as in Figure 11(a), where int denotes the interior of the set. This situation is not possible since the homotopy G is monotone, and since $D_g \subset D_\alpha$. Hence, there are two possibilities: the curve $D_{\gamma, x_0} \subset D_\alpha$, as in Figure 11(b), or $D_\alpha \subset D_{\gamma, x_0}$, as in Figure 11(c). In the former case, we let $\tilde{\gamma}_{x_0} = \alpha$, and in the latter case, we let $\tilde{\gamma}_{x_0} = \gamma_{x_0}$.

In both of these cases, the problem with continuity of the newly constructed

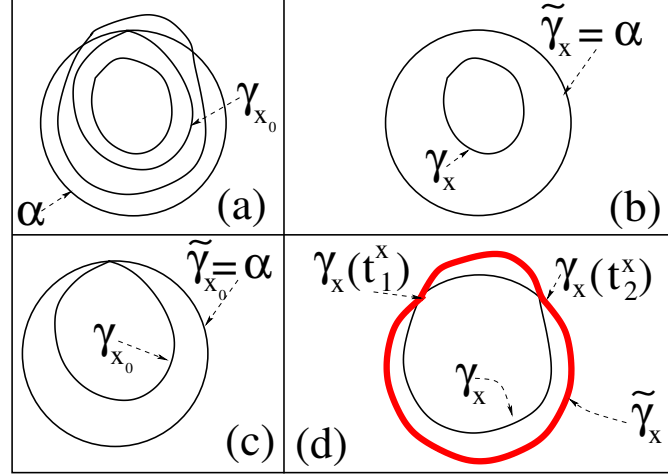
Figure 11: One intersection between α and γ_{x_0} 

1-parametric family of curves at x_0 cannot arise. Indeed, let us consider the first case, when $D_{\gamma, x_0} \subset D_\alpha$ as in Figures 11(a) and 12(b). The second situation will be analogous. Let us consider a family of curves $\gamma_x(t)$, $x \in (x_0 - \delta, x_0 + \delta)$. We can choose δ to be small enough so that there are no points where γ_x and α are tangent for any $x \neq x_0$ in the interval $(x_0 - \delta, x_0 + \delta)$. By the monotonicity of the homotopy G , $D_{\gamma, x} \subset D_\alpha$ if $x \in (x_0, x_0 + \delta)$, so $\tilde{\gamma}_x = \alpha$ for each $x \in (x_0, x_0 + \delta)$, as depicted in Figures 12(b) and 12(c). Since $\gamma_{x_0} = \alpha$, $\tilde{\gamma}_x$ is continuous at x_0 from the right. For the other direction, each curve $\gamma_x(t)$ with $x \in (x_0 - \delta, x_0)$ has exactly two points of intersection with α that we will denote $\gamma_x(t_1^x)$ and $\gamma_x(t_2^x)$, where $t_1^x < t_2^x$. Note that $\gamma_x(t_1^x) = \alpha(s_1^x)$ and $\gamma_x(t_2^x) = \alpha(s_2^x)$, where $s_1^x < s_{x_0} < s_2^x$ by the monotonicity of the homotopy G , and both $\alpha|_{[s_2^x, s_1^x]}$ and $\gamma_x|_{[t_2^x, t_1^x]}$ vary continuously with x as x approaches x_0 from the left, and approach $\alpha(s_{x_0})$ and $\gamma_{x_0}(t_{x_0})$, respectively. Note that the segment $\gamma_x|_{[t_2^x, t_1^x]}$ is inside of D_α . Hence, our algorithm continuously replaces this segment by the segment $\alpha|_{[s_2^x, s_1^x]}$. Thus, we obtain continuity from the left, as shown in Figure 12(d).

Case B. Let us now assume that $\gamma_{x_0}(t)$ intersects α at $2k + 1$ points for some $k \geq 1$. Let $0 \leq t_1 < \dots < t_{2k+1} \leq 1$ be a partition of the unit interval such that $\gamma_{x_0}(t_j) = \alpha(s(t_j)) = \alpha(s_j)$, and where all the intersections are transverse except for $\gamma_{x_0}(t_{j_0})$, where it touches the curve α at $\alpha(s_{j_0})$.

Let us consider the arcs $a = \gamma_{x_0}|_{[t_{j_0-1}, t_{j_0}]}$ and $b = \gamma_{x_0}|_{[t_{j_0}, t_{j_0+1}]}$. There are two possibilities to consider: either both a and b lie inside D_α , or both of them lie outside. It is not possible for one of the arcs to be inside, and for the other to be outside, because that would imply that either the intersection at $\gamma_{x_0}(t_{j_0})$ is transverse, or the tangency at this point is degenerate, and either of these options contradicts our assumptions.

First, assume that both arcs a and b are outside of D_α . Then, according to our

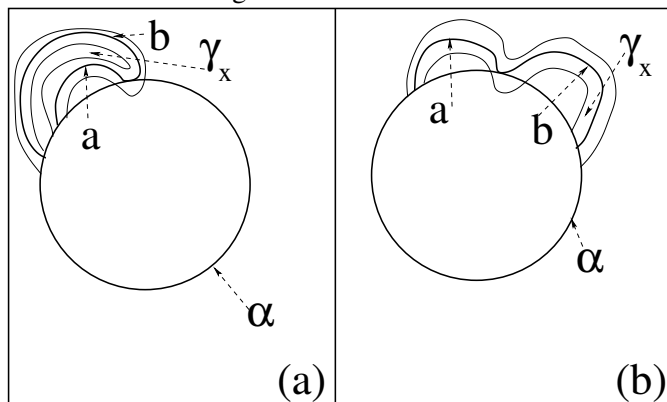
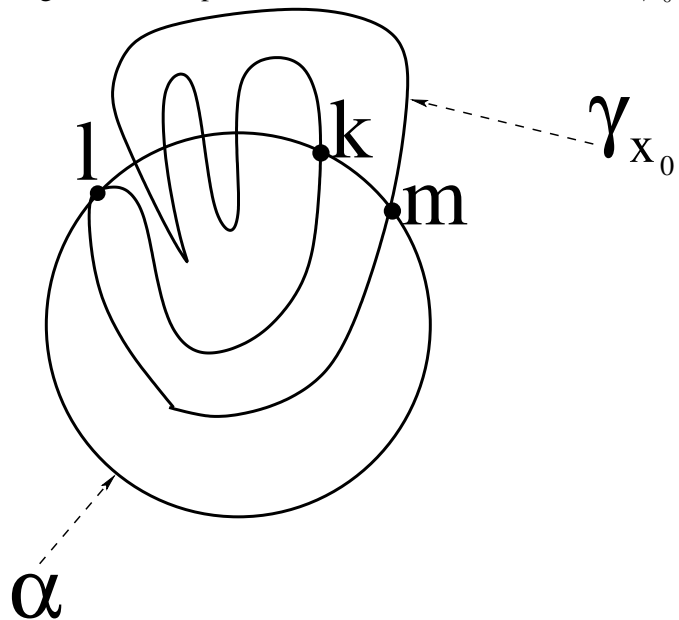
Figure 12: $\tilde{\gamma}_x$ for x near x_0 

algorithm, the arcs of the new curve $\tilde{\gamma}_{x_0}$ on the interval $[t_{j_0-1}, t_{j_0+1}]$ will remain unchanged. Let us consider $\gamma_x(t)$, $x \in (x_0 - \delta, x_0 + \delta)$, where δ is selected so that the curves in the family do not have any additional tangential intersections, as in Figures 13(a) and 13(b). Then, for any nearby curve, the corresponding arc $\gamma_x|_{[t_{j_0-1}^x, t_{j_0+1}^x]}$ is either outside D_α , or has two more additional transverse intersections with α in small neighbourhood of $\gamma_{x_0}(t_{j_0})$. Let us denote these intersections as $t_*^x < t_{j_0}$ and $t_{**}^x > t_{j_0}$. Thus, $\tilde{\gamma}_x$ will locally be formed by continuously replacing $\gamma_x|_{[t_*^x, t_{**}^x]}$ by an arc of α between the same endpoints. Hence, in this case, $\tilde{\gamma}_x$ changes continuously when x is near x_0 . Note that while s_{j_0-1} , s_{j_0} , and s_{j_0+1} can come in any order as is indicated in Figure 13, this fact does not affect the above analysis.

Thus, the only problematic case is when the arcs a and b are inside D_α as in Figures 15 and 16. Let m , l , and k be three points of intersection between γ_{x_0} and α , where l is the tangential point, and m and k are its neighbours. By “neighbours” we will mean points of intersection that are the closest to the tangential point along γ_{x_0} . That is, if $l = \gamma_{x_0}(t_{j_0})$, then $m = \gamma_{x_0}(t_{j_0-1})$ and $k = \gamma_{x_0}(t_{j_0+1})$. It is, however, quite possible that along α there are intersection points that are closer to l than m and k as in Figure 14.

Each pair of points subdivides the curve α into two segments connecting them. Let us denote the two arcs connecting m with l as a_{ml} and \tilde{a}_{ml} , the two arcs connecting m and k as a_{mk} and \tilde{a}_{mk} , and the two arcs connecting l and k as a_{lk} and \tilde{a}_{lk} . Also, let us select a_{ml} , a_{mk} , and a_{lk} so that their interiors do not intersect. Without loss of generality, let a be the arc of γ_{x_0} between m and l in D_α , while b

Figure 13: Outside arcs

Figure 14: Complicated intersections between α and γ_{x_0} 

is the arc between l and k in D_α .

Since a and b are inside D_α , our algorithm requires that we change them to the corresponding arcs of α . The possible discontinuity is a result of the following situation. Let us define two options for replacing arcs of γ_{x_0} with arcs of α . Note that, as before, the arc of α that we are replacing an arc of γ_{x_0} with is the one that is path homotopic to the arc of γ_{x_0} inside the closed annulus $cl(D_h \setminus D_g)$.

Option 1. Separately replace a and b with the appropriate arcs of α .

Option 2. Let $c = a * b$. Replace c with the appropriate arc of α .

As we will see, the two different options will some time result in the same curve, and some times not. Our algorithm, at point x_0 , always uses Option 1. Let us, however, consider the curves that are formed by both options.

Let us first consider Option 1. Without loss of generality, suppose that arc a is changed to a_{ml} . (The other option is \tilde{a}_{ml} .) Now there are two possibilities for the arc b . It will either be changed to a_{lk} or to \tilde{a}_{lk} .

Now let us consider Option 2. Note that if b was changed to a_{lk} , then c must be changed to $\tilde{a}_{mk} = a_{ml} * a_{lk}$. Thus, in this case, it does not matter whether we used Option 1 or Option 2 (see Figure 15).

However, if b was changed to \tilde{a}_{lk} , then c must be changed to a_{mk} . While $a_{mk} \neq a_{ml} * \tilde{a}_{lk}$, we have that $a_{ml} * \tilde{a}_{lk} = a_{ml} * \bar{a}_{ml} * a_{mk}$. (Recall that \bar{a}_{ml} denotes a_{ml} traversed in the opposite direction, from l to m .) Thus, a_{mk} is path homotopic to $a_{ml} * \tilde{a}_{lk}$ by simply contracting $a_{ml} * \bar{a}_{ml}$ to m along itself. Observe that the length of the curve during this homotopy changes monotonically (see Figure 16).

One can see that while the former situation does not create a discontinuity, the latter situation does. Let $\delta > 0$ be once again small enough so that γ_x and α do not have any additional tangential points on the interval $(x_0 - \delta, x_0 + \delta)$.

It is possible that, when $x' \in (x_0 - \delta, x_0)$, $\tilde{\gamma}_{x'}$ will approach the curve obtained using Option 2 as x' approaches x_0 , and for $x'' \in (x_0, x_0 + \delta)$, $\tilde{\gamma}_{x''}$ will approach the curve obtained using Option 1 as x'' approaches x_0 (or the other way around). This is shown in Figure 17.

If we include the homotopy between the curves formed by Option 1 and Option 2 as described above, the resulting family of curves $\tilde{\gamma}_x$ will become continuous and we will be done.

Note that some of the curves obtained in the procedure described above could have self-intersections, however, this happens only when they include arcs of α traversed twice in opposite directions. It is easy to see that one can make all closed curves in \tilde{G} simple using an arbitrarily small perturbation.

Step 2. We will modify the homotopy $H(t, x)$ to obtain $\tilde{H}(t, x)$, a monotone homotopy with the following properties: $\tilde{H}(*, 0) = \alpha$, and $\tilde{H}(*, 1)$ is contained inside the disc bounded by $H(*, 1)$. Moreover, the maximal length of curves in

Figure 15: Inside arcs

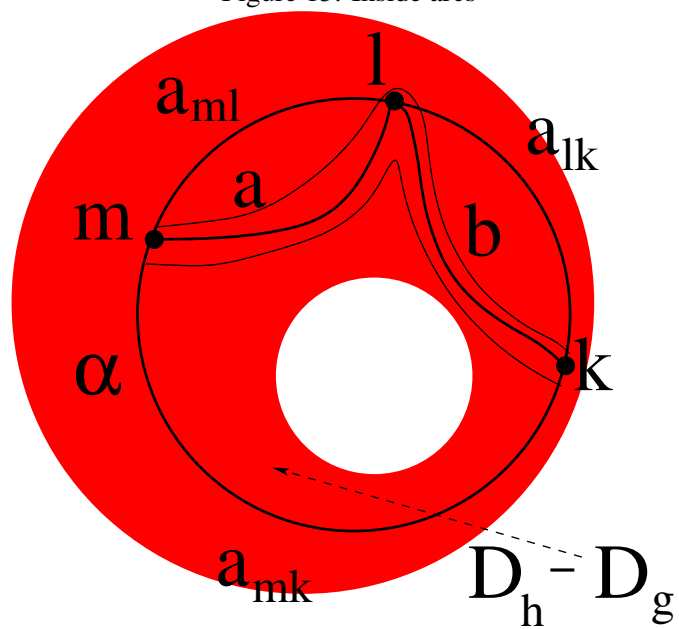


Figure 16: Two ways to change arcs

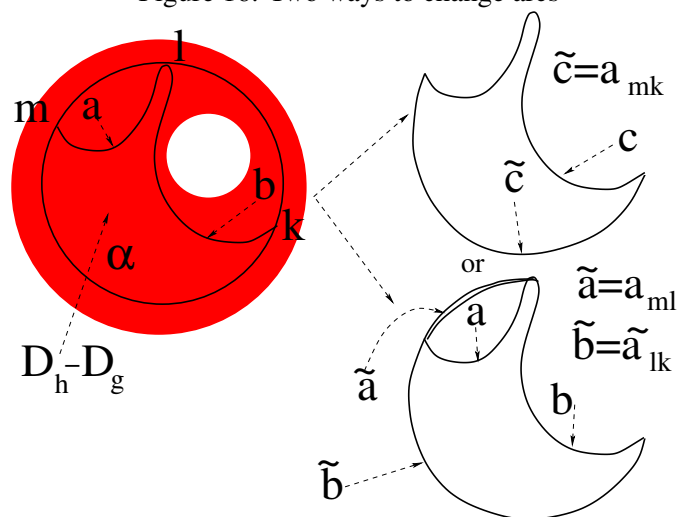
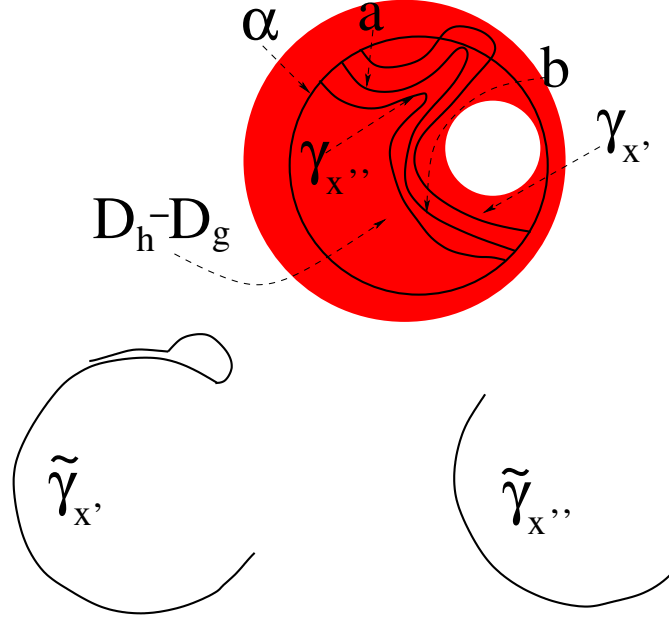


Figure 17: Homotopy between the curves formed by Option 1 and Option 2



the new homotopy will increase by not more than a summand that can be made arbitrarily small.

By analogy with Step 1, the new curves in the homotopy will be constructed by “pushing in” those segments of β_x that lie outside the disc D_α . Let $D_{\beta,x}$ be the closed disc that has β_x as its boundary, as in the hypotheses of the lemma. It will be a procedure that is dual to the one in Step 1.

We will denote the curves in the new homotopy by $\tilde{\beta}_x(t) = \tilde{H}(t, x)$. In particular, $D_\alpha \subset D_{\beta,0}$, so $\tilde{\beta}_0 = \alpha$.

Now, let us describe the curve $\tilde{\beta}_1$. If $D_{\beta_1} \subset D_\alpha$, then we will let $\tilde{\beta}_1(t) = \beta_1(t)$. If $D_\alpha \subset D_{\beta_1}$, then we will let $\tilde{\beta}_1 = \alpha(t)$. If $D_\alpha \cap D_{\beta_1} = \emptyset$, then we will let $\tilde{\beta}_1(t)$ be some point \tilde{p} , where \tilde{p} is obtained as follows. Let $\tilde{x}_0 = \sup\{x \in [0, 1] \text{ such that } D_{\beta,x} \cap D_\alpha \neq \emptyset\}$. Let $\tilde{p} = D_{\beta,\tilde{x}_0} \cap D_\alpha$.

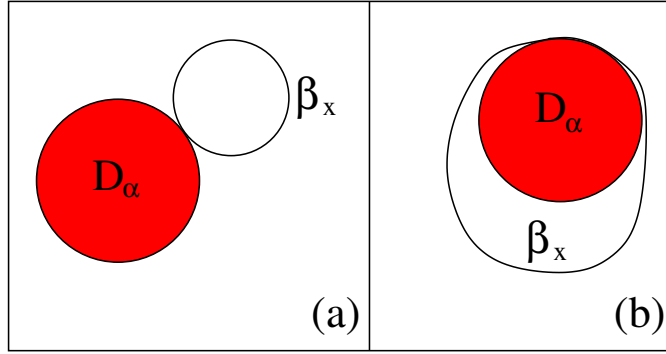
Finally, suppose that $D_\alpha \cap D_{\beta_1} \neq \emptyset$, but that one is not a subset of the other. In this case, $\tilde{\beta}_1$ is constructed as follows. Let us consider arcs of $\beta_1(t)$ that are outside D_α . That is, let $0 = t_0 < t_1 < \dots < t_n = 1$ be a subdivision of the unit interval, such that $\beta_1(t_i) = \alpha(s_{j(i)})$ for some $s_{j(i)} \in [0, 1]$, and $\beta_1 \neq \alpha$ otherwise. Let us consider each $\beta_1|_{[t_i, t_{i+1}]}$ that lies outside D_α . By the monotonicity of $H(x, t)$, it lies in the annulus bounded by $H(0, t)$ and $G(1, t)$. The points $\alpha(s_{j(i)})$ and $\alpha(s_{j(i+1)})$ subdivide α into two arcs. Exactly one of these arcs has the property

that, if it is replaced by $\beta_1|_{[t_i, t_{i+1}]}$, then the resulting curve will be non-contractible in the annulus. We replace $\beta_1|_{[t_i, t_{i+1}]}$ by this arc of α . When this is done for all arcs, we obtain $\tilde{\beta}_1$. Clearly, $D_{\tilde{\beta}_1} \subset D_{\beta_1}$.

$\tilde{\beta}_x$ is constructed in a completely analogous manner for an arbitrary $x \in [0, 1]$. Note that while the length of curves in the one-parameter family $\tilde{\beta}_x$ with $x \in [0, 1]$ has not increased compared to β_x , at this stage there can arise some discontinuities with respect to x . As in Step 1, those discontinuities can only occur at the points where α and β_x are tangent, and only when β_x touches α from the outside of D_α , since only arcs of β_x outside D_α are to be replaced.

Note also that if α and β_{x_0} intersect at only one point, as in Figures 18(a) and 18(b), then continuity at x_0 remains intact.

Figure 18: One point of intersection



Thus, let us assume that there are $2k + 1$ intersection points at $0 \leq t_1 < t_2 < \dots < t_{2k+1} \leq 1$, $k \geq 1$. Let $\beta(t_{j_0}) = \alpha(s_{j_0}) = l$ be the point of tangency. Once again, let m and k be its neighbours with respect to β_{x_0} . Let us denote the arc of β_{x_0} that connects m with l by a and the arc that connects l with k by b , as in Figures 19(a) and 19(b).

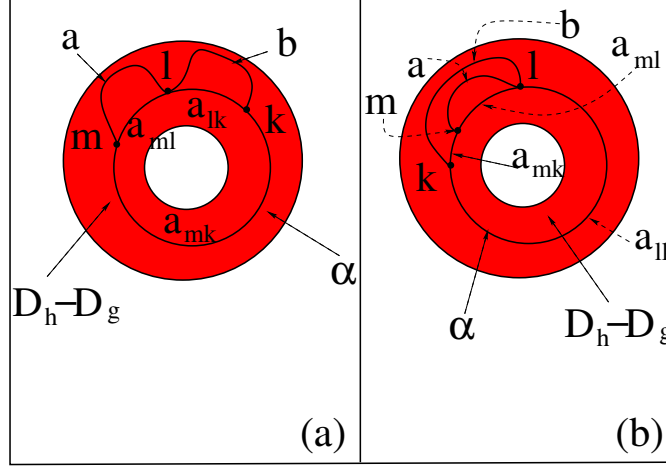
m , l , and k subdivide α into three non-intersecting arcs that will be denoted as a_{ml} , a_{mk} , and a_{lk} , indexed by the pair of points that each segment connects. Again, we are assuming that these three arcs have disjoint interiors. Their complements in α will be denoted as \tilde{a}_{ml} , \tilde{a}_{mk} , and \tilde{a}_{lk} , respectively.

As in Case 1, a discontinuity can only arise if replacing a followed by b yields a different curve than that obtained by replacing $a * b$ as a single arc.

It is easy to see that, when β_{x_0} touches α from the outside of D_α , there are two situations to consider.

The first situation is depicted in Figure 19(a). In this case, we replace a by a_{ml} and b by a_{lk} . The other approach would be to replace the arc $c = a * b$ by the arc

Figure 19: Outside arcs



$\tilde{a}_{mk} = a_{ml} * a_{lk}$. Thus, in this case, the two different options of replacing arcs lead to the same result. In this situation, no discontinuities are formed.

The second situation is depicted in Figure 19(b). In this case, we replace a by a_{ml} and b by \tilde{a}_{lk} , while c is replaced by a_{mk} . While $a_{mk} \neq a_{ml} * \tilde{a}_{lk}$, we observe that $a_{ml} * \tilde{a}_{lk} = a_{ml} * \bar{a}_{ml} * a_{mk}$. Thus, a_{mk} is path homotopic to $a_{ml} * \tilde{a}_{lk}$ by simply contracting $a_{ml} * \bar{a}_{ml}$ along itself to the point m . This forms a discontinuity in our homotopy \tilde{H} , but if we also include this contraction (extended to the whole curve as a part of our 1-parameter family $\tilde{\beta}_x$), the discontinuity will be resolved. This is analogous to the method used to resolve discontinuities in Case 1.

Step 3. After we have constructed $\tilde{G}(t, x)$ and $\tilde{H}(t, x)$, we can concatenate \tilde{G} and \tilde{H} ($\tilde{G} * \tilde{H}$) to produce the desired homotopy.

□

Proof of Theorem 0.8. We begin with the original homotopy $H(t, \tau) : S^1 \times [0, 1] \rightarrow M$ and use Lemma 1.4 to obtain a corresponding weak monotone homotopy cover $\{F_i(t, x)\}_{i=1}^n$.

Now the idea of the proof is to glue together the homotopies $F_i(t, x)$ to obtain a monotone homotopy of a contractible simple closed curve γ to some point \tilde{p} that covers the curve $H(*, 0)$. The procedure is inductive with respect to the index i of the homotopies in the cover $F_i(t, x)$. The homotopy $F_1 = G_1$ constitutes the base of the induction. On the k th step we apply Lemma 1.5 to $G = G_{k-1}$ and $H = F_k$ to obtain a monotone homotopy that we redenote as G_k . It is easy to see that, for each k , G_k covers $H(*, 0)$. In particular, G_{n-1} covers $H(*, 0)$, and $G_{n-1}(*, 1)$ will be inside $F_n(*, 0)$. As we already noted in the course of the proof

of Lemma 1.4, this fact implies that $G_{n-1}(*, 1)$ will be contained in a metric ball in M that can be made arbitrarily small by choosing appropriate parameters in the algorithm that we used to prove Lemma 1.4. This curve can now be contracted to a point through monotone curves so that the lengths of these curves can exceed the length of $G_{n-1}(*, 1)$ only by a summand that can be made arbitrarily small. Hence, G_{n-1} can easily be modified so that it covers $H(*, 0)$ (in the sense of Definition 0.7). This completes the proof. \square

2 Proof of Theorem 0.9

In this section, we prove Theorem 0.9. We are given a simple closed contractible curve γ in M , and a contraction $H(t, \tau) : S^1 \times [0, 1] \rightarrow M$ which covers γ , and which consists of simple closed curves of length no more than L .

Given a point $q \in \gamma$ and an $\varepsilon > 0$, we wish to construct a contraction of γ through curves based at q with the property that all curves are bounded in length by

$$3L + 2d + \varepsilon$$

where d is the diameter of the manifold. We will also show that there is a specific point $q^* \in \gamma$ such that there is a contraction of γ through curves based at q^* of length bounded by

$$2L + 2d + \varepsilon.$$

Throughout this proof, we produce curves of length less than or equal to $Q + \varepsilon$ for some $Q > 0$, where $\varepsilon > 0$ is chosen to be arbitrarily small. When we combine two curves of length bounded in this way, we simply write that the result has length bounded by $2Q + \varepsilon$. Although not strictly true, since we chose the original ε to be as small as desired, we can just go back and choose it to be $\frac{\varepsilon}{2}$, in which case the new inequality $2Q + \varepsilon$ holds. To improve readability, we do not mention this argument when it is used.

We will also be using the terms *interior* and *exterior* of H_τ and of γ , which we redefine here for clarification:

Definition 2.1 *Since H_τ is a monotone contraction, there is a disc $\mathbb{D} \subset M$ defined by the set of all points that are in the image of some curve in H . For each point τ , H_τ is simple and is contained in \mathbb{D} , and as such divides M into two open regions. Exactly one of these regions is entirely contained in \mathbb{D} . This region is the interior of H_τ , and the other region is the exterior. Similarly, since H covers γ , γ is contained in \mathbb{D} . Since it is simple, γ divides M into 2 regions, exactly one of which is entirely*

contained in \mathbb{D} . This region is the interior of γ , and the other region is the exterior of γ .

We will prove two lemmas which, when combined, will allow us to prove this theorem. For each, we assume that $\varepsilon > 0$ is fixed.

Lemma 2.2 *There exists a point $x \in \gamma$ and a point τ^* such that there exists a homotopy \tilde{H} from γ to either a curve formed by slightly perturbing H_{τ^*} or to the point x through curves of length at most $2L + \varepsilon$. Additionally, x lies on every curve in the homotopy \tilde{H} .*

Since the point $x \in \gamma$ has the aforementioned properties, \tilde{H} is a based loop homotopy. Our second lemma takes \tilde{H} and transforms it into a contraction of γ through curves based at x of length at most $2L + 2d + \varepsilon$.

Lemma 2.3 *If Lemma 2.2 does not contract γ to x , then there exists a contraction of the curve formed by slightly perturbing H_{τ^*} through loops based at x of length bounded by $2L + 2d + \varepsilon$. We denote this contraction by K .*

We will first demonstrate how these two lemmas can be used to prove Theorem 0.9, and will then prove each of them in turn.

Proof of Theorem 0.9. Let H be our original homotopy, \tilde{H} be the homotopy generated by Lemma 2.2, and let K be the homotopy generated by Lemma 2.3. By Lemma 2.2, either \tilde{H} contracts γ to the point x , or it homotopes γ to a slight perturbation of H_{τ^*} . If it contracts γ to the point x , then we are done. If it doesn't, then we have to use Lemma 2.3. We do this by concatenating \tilde{H} and K to get a contraction of γ through curves based at x of length at most $2L + 2d + \varepsilon$, as desired. Hence, the point x is the special base point $q^* \in \gamma$ mentioned above. Furthermore, this will complete the proof of the theorem: if we choose any point $q \in \gamma$, then we can build the appropriate contraction based at q as follows. Let α be an arc of γ from q to x of length at most $\frac{L}{2}$, and let $-\alpha$ be the same arc, but with opposite orientation. Lastly, let β be the curve formed by concatenating α with $-\alpha$. We can then take our contraction of γ based at q^* , and for each curve γ_τ in this contraction, we replace γ_τ with the curve that is formed by traversing α , then γ_τ , then $-\alpha$. In this way, we produce a homotopy from γ to β which is based at q , and which consists of curves of length at most $3L + 2d + \varepsilon$. Since β can be contracted through loops based at q of length at most L , this completes the proof. \square

We are left now with proving each of the two lemmas outlined above.

2.1 Proof of Lemma 2.2

To prove Lemma 2.2, we will adopt an approach that will be very similar to that used by Chambers and Liokumovich in [CL]. To begin with, we would like to perturb the homotopy H so that only finitely many non-transverse intersections between H and γ occur, and so that they do not occur concurrently.

Lemma 2.4 (Perturbation Lemma) *For any $\varepsilon > 0$, we can perturb H , obtaining a new homotopy \overline{H} and points*

$$0 = \tau_0 < \cdots < \tau_n = 1$$

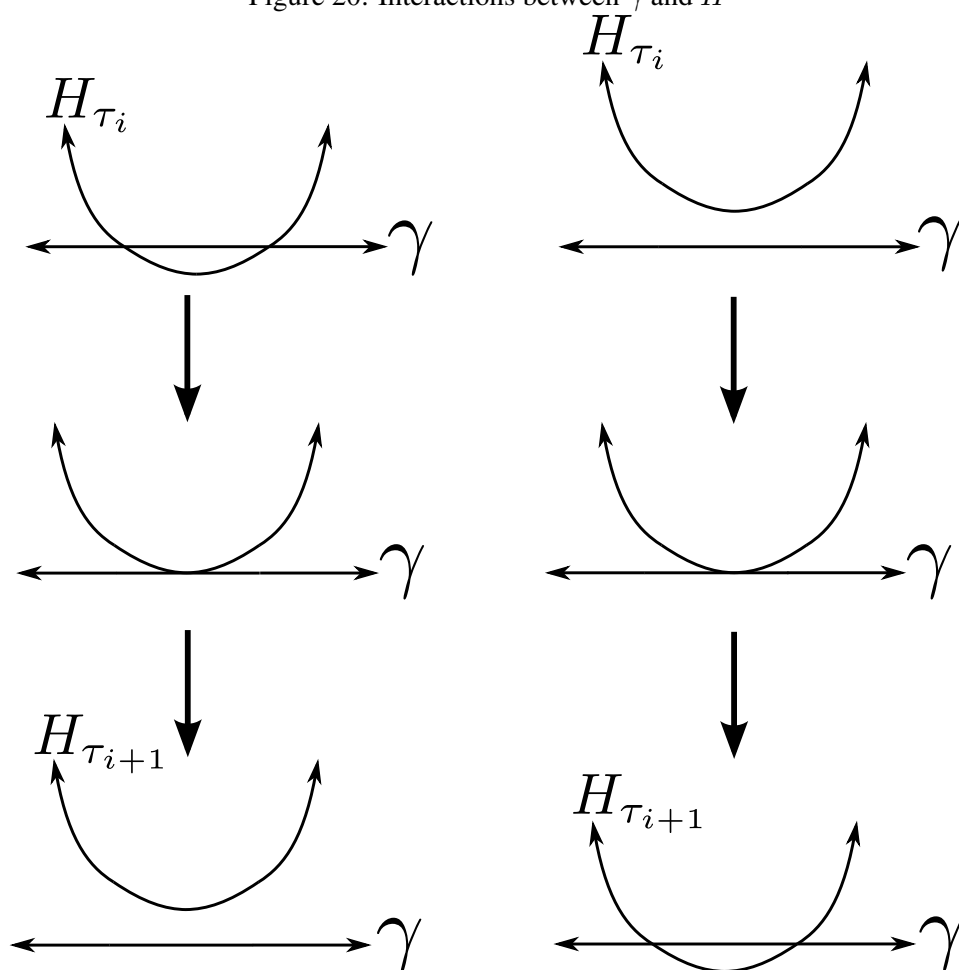
such that, for all $\tau \in [\tau_i, \tau_{i+1}]$, all intersections between H_τ and γ are transverse, except for exactly one intersection at one point τ . The two possible interactions are shown in Figure 20. \overline{H} also has the following additional properties:

1. \overline{H} is a contraction that covers γ .
2. \overline{H} is monotone.
3. \overline{H} consists of curves of length at most $L + \varepsilon$.

To prove this lemma, we use the same technique as in Proposition 2.1 from [CL]; we apply the parametric version of Thom's Multijet Transversality Theorem to the submanifold of the 2-fold 1-jet bundle corresponding to curves with singularities to show that a perturbation is possible which satisfies the above criteria. This approach does not rule out other singular behaviour which involves self-intersections in γ or in H_τ , however, since both of these are simple, they have no self-intersections, and so the interactions between the two curves are limited to the isolated tangential intersections shown in Figure 20. We use the term *Reidemeister move* to describe this behaviour, this term being derived from the obvious relationship between this singularity and the knot moves used in Reidemeister's Theorem. We also note that, since \overline{H} is a contraction that covers γ , $n \geq 2$. In other words, there must be at least 2 Reidemeister moves, once where \overline{H} transitions from a curve which lies completely in the exterior of γ to one which only partly lies in the exterior, and a Reidemeister move in which H goes from being a curve which only partly lies in the interior of γ , to a curve which lies either entirely in the interior of γ , or entirely in the exterior of γ .

To simplify this exposition, we will assume that H has already been perturbed, and so it already has all of the properties described in Lemma 2.4.

We now want to prove Lemma 2.2 for H and γ . We will define the point x and the point τ^* , and then prove that these points satisfy all of the required criteria.

Figure 20: Interactions between γ and H 

Definition 2.5 *Let x be the last point at which H and γ intersect, and let τ^* be the point at which this intersection occurs. Note that $\tau^* \in (\tau_{n-1}, \tau_n)$, and H_{τ^*} and γ intersect tangentially at x .*

The idea to prove that Lemma 2.2 holds for these values of x and τ^* is similar to that used by Chambers and Liokumovich in [CL]. We construct a certain graph Γ where the vertices represent curves, and the edges represent homotopies between curves. We then show that this graph contains a certain path which represents a homotopy that easily implies the existence of the desired homotopy.

Vertices

We begin to construct this graph Γ by defining its vertices. As above, each vertex will correspond to a certain curve. For each $i \in \{1, \dots, n-1\}$, consider

$$U_i = \gamma \cup H_{\tau_i}.$$

We will begin by identifying certain closed curves whose images lie in U_i . We will then eliminate some of these curves based on several criteria. For each curve that remains, we will add a vertex. We begin by defining our large set of closed curves. We will call these curves *subcurves* at τ_i .

Definition 2.6 (Subcurves at τ_i) *Choose any pair (p_1, p_2) of distinct intersection points between H_{τ_i} and γ . We can write γ as the disjoint union of p_1, p_2 , and two open segments ϱ_1 and ϱ_2 . Each of these segments can be used to join p_1 to p_2 . Similarly, H_{τ_i} can be written as the disjoint union of p_1, p_2 , and two open segments σ_1 and σ_2 . Each of these segments can also be used to join p_1 to p_2 .*

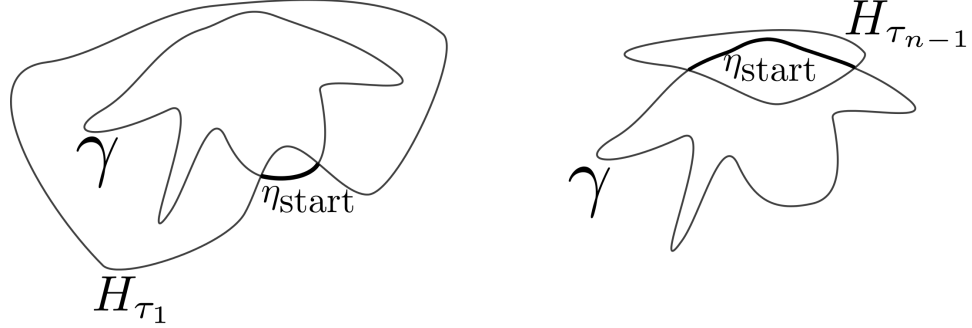
For any piecewise smooth closed curve α whose image lies in U_i , if we can find such a pair (p_1, p_2) of intersection points such that α can be written as the disjoint union of p_1, p_2, σ_i and ϱ_j for $i, j \in \{1, 2\}$, then we say that α is a subcurve at τ_i .

We say that such a subcurve has endpoints p_1 and p_2 , and we will denote the segment of the curve that comes from γ as ϱ , and the segment that comes from H_{τ_i} as σ . Both are open, contiguous segments of their respective curves.

Before we define which subcurves we will use to generate vertices, we will need a few definitions first. To start, we want to define two open, disjoint, contiguous segments of γ , which we will call η_{start} and η_{end} . Note that the monotonicity of H guarantees that they are disjoint.

Definition 2.7 (η_{start} and η_{end}) *We define the segment η_{start} as the segment of γ that is not contained in the closure of the interior of H_{τ_1} . Since there are exactly two intersection points between H_{τ_1} and γ , this segment is well defined.*

We define η_{end} as the open segment of γ that is contained in the interior of $H_{\tau_{n-1}}$. Since $H_{\tau_{n-1}}$ and γ intersect in exactly two points, this segment is well defined. These are shown in Figure 21.

Figure 21: η_{start} and η_{end} 

We have a simple property of η_{start} and η_{end} which results from the monotonicity of H :

Lemma 2.8 *For every point $\tau \in [\tau_1, \tau_{n-1}]$, and for any intersection point p between H_τ and γ , p lies neither in η_{start} , nor does it lie in η_{end} .*

We can now begin to define the set of subcurves that we will use to produce our vertices; we will define whether or not a subcurve *respects* γ .

Definition 2.9 (Respects γ) *We say that a subcurve α at τ_i respects γ if the segment ϱ of α (the segment that came from γ) has the following two properties:*

1. $\eta_{\text{start}} \cap \varrho = \emptyset$
2. $\eta_{\text{end}} \subset \varrho$

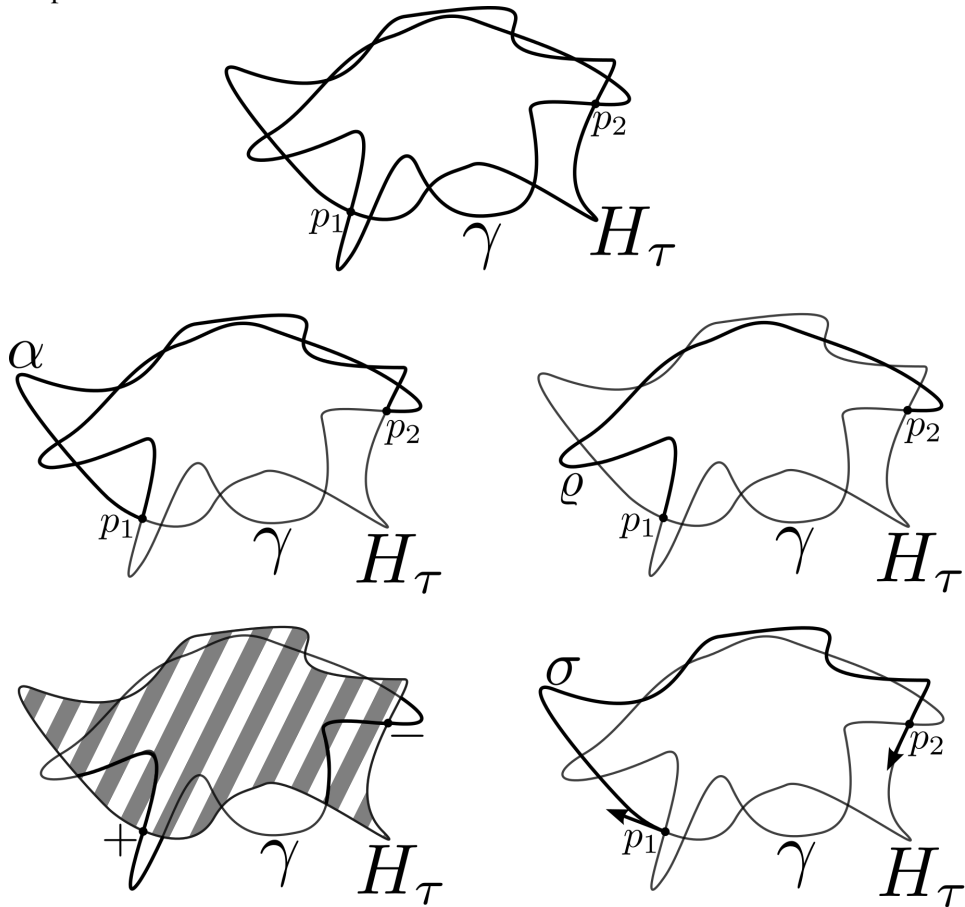
For every subcurve α at τ_i that respects γ , we give each endpoint of this curve a sign, either a $+$, or a $-$. Let p be an endpoint of γ . Orienting γ , we can list the order in which we encounter intersection points. Let q and r be the intersection points which we encounter immediately before and after p , which may be the same point. Since γ is oriented, we can also produce two contiguous segments of γ : the segment traversed from q to p , and the segment traversed from p to r (with respect to the orientation of γ). Neither segment contains any intersection points. Let them be β_1 and β_2 .

We also see that exactly one of β_1 and β_2 must be contained in the interior of H_{τ_i} since p is a transverse intersection point of H_{τ_i} and γ . Let this component be β_j . Furthermore, recalling that ϱ is the segment of α that comes from γ , exactly one of β_1 and β_2 must be contained in ϱ . Let this component be β_k .

If $k = j$, then we assign a $+$ sign to p . If not, then we assign a $-$ sign to p . Note that the sign of a point does not depend on how we orient γ .

Figure 22 depicts curves H_τ and γ . It also depicts a subcurve α which respects γ and its endpoints. Here, γ is the same curve that appears in Figure 21, and we assume that η_{start} and η_{end} are as in this figure. The segment ϱ of α is shown, and the signs of both endpoints are displayed as well. Lastly, σ is shown with the tangent vector at each of its endpoints. We see that the directions of these tangents with respect to the interior of γ do not agree with the signs of both intersection points, as per Definition 2.10. Hence, α is not a valid subcurve.

Figure 22: From top to bottom, left to right: H_τ and γ , a subcurve α that respects γ , ϱ , the signs of the endpoints of α , and σ with the tangent vector at each of its endpoints



We can now define the set of subcurves which we want to use to produce vertices. We call such a subcurve *valid*.

Definition 2.10 (Valid subcurve) We say that a subcurve α at τ_i is valid if it respects γ , and if the following additional properties are true of σ .

To begin, let p_1 and p_2 be the endpoints of α . We can parametrize σ so that it goes from p_1 to p_2 . Since H_{τ_i} and γ intersect transversely at p_1 and at p_2 , we can categorize the tangent vector of σ at p_1 and at p_2 as being into the interior of γ , or into the exterior of γ .

We then require that, at p_1 , the tangent of σ points into the interior of γ if the sign at p_1 is $+$, and that it points into the exterior of γ if it is $-$. We also require that, at p_2 , the tangent of σ points into the exterior of γ if the sign at p_2 is $+$, and that the tangent points into the interior at p_2 if the sign is $-$. Note that this definition is independent of the order in which we choose the endpoints of α ; a subcurve is valid with respect to one order of endpoints if and only if it is valid with respect to the other order.

For each valid subcurve at τ_i with $i \in \{1, \dots, n-1\}$, we add a vertex v to the graph Γ . We say that this vertex is generated from τ_i . We also have a length bound for each valid subcurve, as a result of it being composed of a segment of γ and a segment of H_{τ_i} :

Lemma 2.11 (Length bound for valid subcurves) For each valid subcurve α , the length of α is at most

$$2L + \varepsilon.$$

Edges

We now add edges to this graph. The idea will be that, for each $i \in \{1, \dots, n-2\}$, we will add a set of edges, denoted by E_i . We will specify an algorithm which takes any vertex v generated from τ_i or from τ_{i+1} , and produces a different vertex w , also generated from τ_i or from τ_{i+1} . This algorithm is symmetric in that, if given vertex w , it will produce vertex v . We then join each pair of vertices produced by this algorithm by an edge. E_i will be the collection of these edges.

To define this algorithm, fix a vertex v in Γ generated from τ_i . We will define the algorithm in two parts, depending on whether the resulting vertex w is generated from τ_{i+1} (a “vertical” edge), or if it is generated from τ_i (a “horizontal” edge).

Throughout the definition of this algorithm, we say that two intersections p and q between H and γ at τ_i are “involved” or “deleted” in the Reidemeister move between τ_i and τ_{i+1} . By this, we mean the following. Let the point at which H and γ become tangent to each other be τ' , with $\tau_i < \tau' < \tau_{i+1}$. Since all intersections between H and γ are transverse on (τ_i, τ') , we can trace the path of p and q forward to τ' . When we do this, we see that p gets traced to the tangential intersection at

τ' (which is deleted), and q gets traced to the same intersection point. We use the same terminology to describe intersection points between H and γ at τ_{i+1} that can be traced backwards to the tangential intersection at τ' .

Vertical Edges Recall that, between τ_i and τ_{i+1} , there is exactly one Reidemeister move. This move involves two intersection points between H and γ ; it either creates two intersection points, or it deletes two of them. Let α be the valid subcurve at τ_i that produced the vertex v and let p_1 and p_2 be the two distinct endpoints of α . If neither of these points is involved in the Reidemeister move, then the algorithm to find the vertex w generated from τ_{i+1} is simple. Since neither p_1 nor p_2 are deleted from τ_i to τ_{i+1} , they both follow continuous paths from τ_i to τ_{i+1} . Let \tilde{p}_1 and \tilde{p}_2 be the points which we reach at τ_{i+1} . We also see that we can follow σ and ϱ from τ_i to τ_{i+1} in a similar fashion, arriving at $\tilde{\sigma}$ and $\tilde{\varrho}$. Let $\tilde{\alpha}$ be the subcurve formed by following $\tilde{\sigma}$ from \tilde{p}_1 to \tilde{p}_2 , followed by $\tilde{\varrho}$ from \tilde{p}_2 back to \tilde{p}_1 . If $\tilde{\alpha}$ is a valid subcurve, then it corresponds a vertex w . We will show that it is indeed valid; w is the vertex that is produced by the algorithm.

As a result of Lemma 2.8, $\tilde{\varrho}$ has all of the required inclusion/exclusion properties with respect to η_{start} and η_{end} , and so $\tilde{\alpha}$ respects γ . To show that it is valid, we notice that the sign of p_1 is the same as that of \tilde{p}_1 , and the sign of p_2 is the same as that of \tilde{p}_2 . Let us orient σ from p_1 to p_2 , and $\tilde{\sigma}$ from \tilde{p}_1 to \tilde{p}_2 . We then have that the direction of the tangent vector of σ at p_1 with respect to the interior of γ is the same as the direction of the tangent vector of $\tilde{\sigma}$ at \tilde{p}_1 with respect to the interior of γ . Similarly, the direction of the tangent vector at p_2 is the same as that at \tilde{p}_2 . Hence, $\tilde{\alpha}$ is a valid subcurve, and so we are done.

If v is instead generated from τ_{i+1} , and neither of the endpoints of α are involved in the Reidemeister move between τ_i and τ_{i+1} , then we follow the exact same procedure as above, but in reverse.

Horizontal Edges Again, let the vertex v be generated from τ_i , let α be the valid subcurve which corresponds to v , and let p_1 and p_2 be the endpoints of α . If neither p_1 nor p_2 are involved in the Reidemeister move between τ_i and τ_{i+1} , then we use the algorithm described above. In this component of the algorithm, we determine the resulting vertex w if p_1 or p_2 are involved in the move. Furthermore, let τ' be the point between τ_i and τ_{i+1} at which $H_{\tau'}$ is tangent to γ . This is the point at which the Reidemeister move ‘‘occurs’’.

We first rule out the possibility that both p_1 and p_2 are involved in the Reidemeister move:

Lemma 2.12 p_1 and p_2 cannot both be deleted in the Reidemeister move between τ_i and τ_{i+1} .

Proof. Assume that they are both involved in the Reidemeister move. As in the definition of subcurves, we can break γ into two contiguous segments, each

with endpoints p_1 and p_2 . We do this by starting at p_1 , and then by traversing γ to p_2 in each of the two possible directions. Let these two components be β_1 and β_2 . If both p_1 and p_2 are involved in the Reidemeister move, then at least one of these two segments would have to contain no intersection points. This is because no intersection points are deleted between τ_i and τ' , and there is no way for intersection points to move through each other. As such, until τ' , the order of intersection points as we traverse γ remains the same. Hence, if there were intersection points in both β_1 and β_2 , then there would be no way for p_1 and p_2 to be deleted together, as there would have to be an interaction between at least one other pair of intersection points first. Let this intersection-free segment be κ .

We can also choose this segment κ so that, for every $s \in \kappa$, s is an intersection point between H_{τ_s} and γ for some $\tau_s \in [\tau_i, \tau']$.

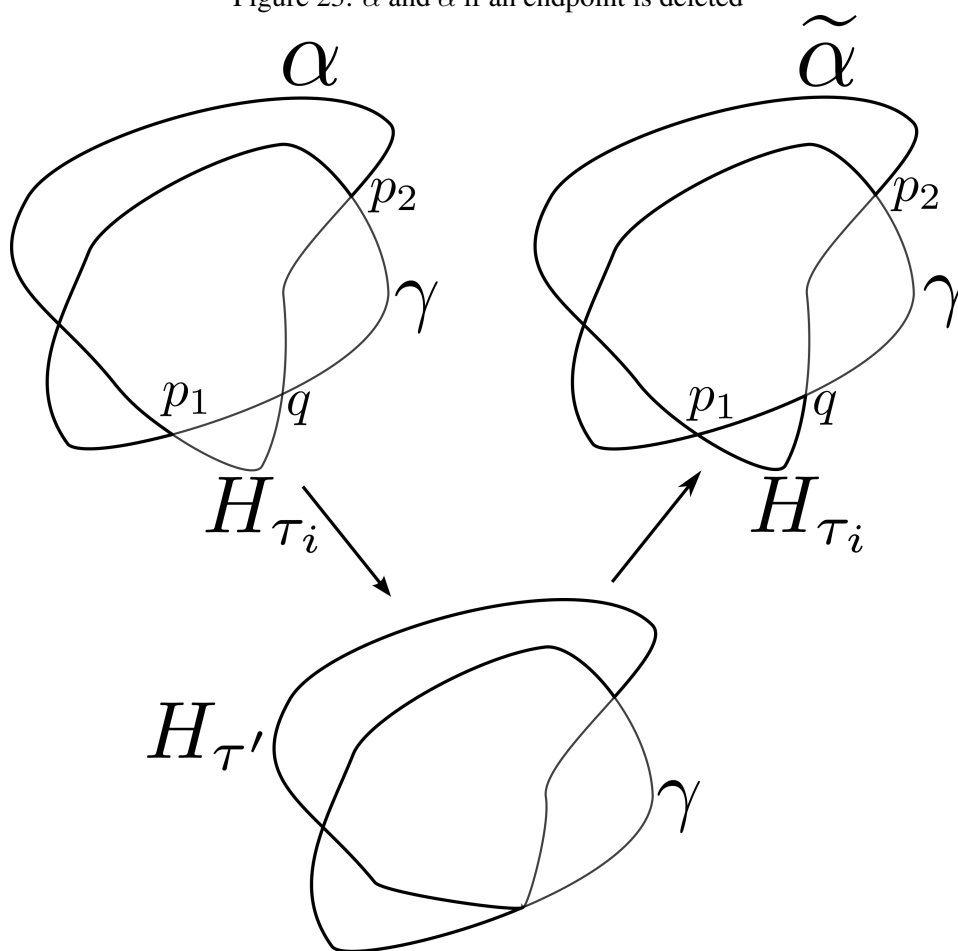
Furthermore, since α is a valid subcurve, it respects γ , and so we see that exactly one β_j must contain η_{start} , and the other must contain η_{end} . Hence, κ must contain one of these curves. By Lemma 2.8, there are thus points in κ that are not realized as intersection points between τ_i and τ_{i+1} . This is a contradiction, and so p_1 and p_2 cannot both be deleted in the Reidemeister move between τ_i and τ_{i+1} . \square

Let us now move to the case where just one of p_1 or p_2 is deleted at τ' . Without loss of generality, let us assume that it is p_1 , and let q be the other intersection point at τ_i which is deleted with p_1 in the Reidemeister move. We adopt a similar approach to when we added vertical edges. We can trace the path of p_1 forward until τ' , and we also trace the path of q forward until τ' . We notice that, since both are deleted at τ' , they merge at this point. We can thus trace a path from p_1 to q by first going forward to τ' , tracing the path of p_1 forward, and then we can go backward, tracing the path of q backwards.

This path from p_1 to q induces a homotopy from α to a subcurve $\tilde{\alpha}$ at τ_i with endpoints q and p_2 . $\tilde{\alpha}$ is formed from the segment $\tilde{\varrho}$ of γ and the segment $\tilde{\sigma}$ of H_{τ_i} . The first is found by following ϱ forward to τ' , then by going backwards to τ_i , using q as an endpoint instead of p_1 as we go backwards; $\tilde{\sigma}$ is found by doing the same, but with σ . This is shown in Figure 23.

The question, as before, is if $\tilde{\alpha}$ is a valid subcurve. We see that, since α is a valid subcurve, it respects γ , and so ϱ has the proper inclusion/exclusion properties with respect to η_{start} and η_{end} . Lemma 2.8 then implies that $\tilde{\varrho}$ has similar properties, and so $\tilde{\alpha}$ respects γ .

To show that $\tilde{\alpha}$ is valid, we must show that $\tilde{\sigma}$ agrees with the signs of q and p_2 . We first observe that the sign of q with respect to $\tilde{\alpha}$ is opposite to the sign of p_1 with respect to α . On the other hand, the sign of p_2 remains unchanged. If we orient σ from p_1 to p_2 , and $\tilde{\sigma}$ from q to p_2 , then we see that the tangent vector of $\tilde{\alpha}$ at p_2 has the same direction with respect to the interior of γ as the tangent vector

Figure 23: α and $\tilde{\alpha}$ if an endpoint is deleted

of σ at p_2 , and so this endpoint meets the necessary criteria. In terms of q , we see that the direction of the tangent vector of $\tilde{\sigma}$ at q with respect to the interior of γ is opposite to that of the tangent vector of σ at p_1 . Hence, this endpoint meets the necessary criteria as well, and so $\tilde{\alpha}$ is valid. The rigorous proof of this is a case-by-case analysis on the segment of $H_{\tau_i} \cup \gamma$ around q and p_1 . The cases are formed by considering all possible interiors of H_{τ_i} , and all possible arcs ϱ . This is shown in Figure 24. Since $\tilde{\alpha}$ is valid, it corresponds to a vertex w , which is the desired vertex.

If v is generated from τ_{i+1} , then we follow the above procedure, but in reverse. That is, if the Reidemeister move between τ_i and τ_{i+1} creates two intersection points of which one is an endpoint of α , then we follow the above steps to produce a vertex w . Note that for reasons analogous to those presented in the proof of Lemma 2.12, both endpoints of α cannot be created by the Reidemeister move between τ_i and τ_{i+1} .

Before we complete the proof of Lemma 2.2, we prove some important properties of Γ :

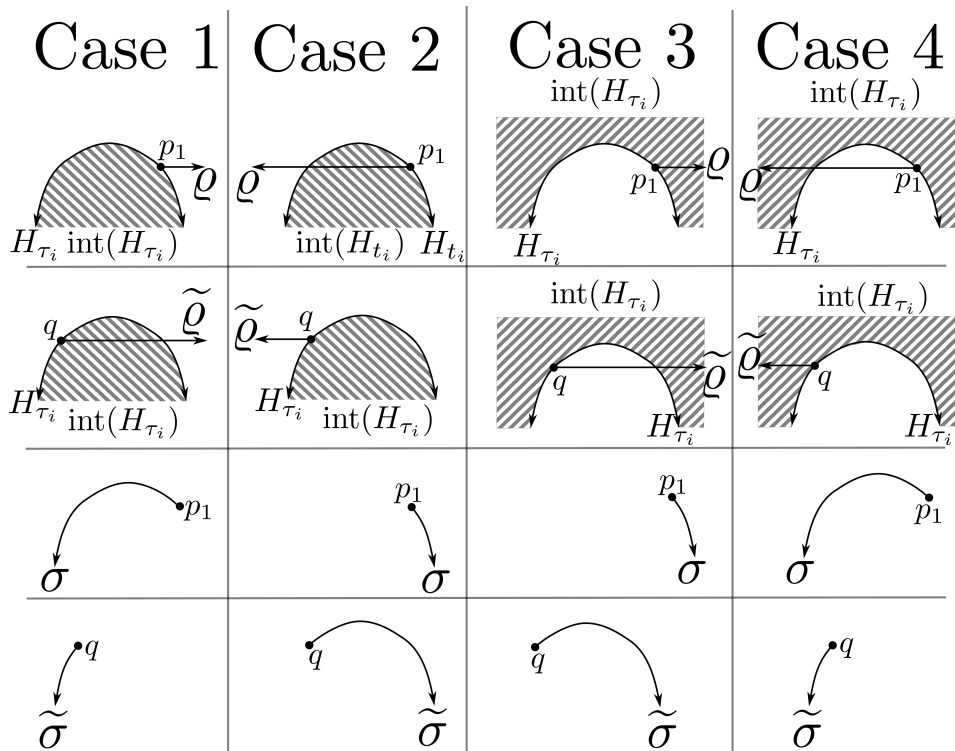
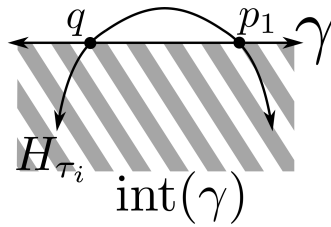
Lemma 2.13 (Properties of Γ) *The graph Γ has the following properties:*

1. *For each set of edges E_i , $i \in \{1, \dots, n-2\}$, and for each vertex v generated from τ_i or τ_{i+1} , v is the endpoint of exactly one edge in E_i .*
2. *All vertices generated from τ_1 and all vertices generated from τ_{n-1} have degree 1; all other vertices have degree 2.*
3. *There is exactly one vertex generated from τ_1 , and one vertex generated from τ_{n-1} , and they correspond to the curves shown in Figure 25.*
4. *If two vertices are joined by an edge, then there is a homotopy of closed curves between the curves corresponding to the vertices through closed curves of length at most $2L + \varepsilon$. Furthermore, all of these curves contain η_{end} .*

Proof. The first statement is a result of the fact that the algorithm used to add edges takes any vertex v generated from τ_i or from τ_{i+1} and produces a vertex w , $v \neq w$. Since we use this algorithm to add edges, there is an edge between v and w . Additionally, it is easy to check that this algorithm is symmetric in that the vertex w will produce the vertex v . Hence, each vertex is the endpoint of exactly one edge in E_i .

The second statement results from the fact that, for each vertex v that is generated from τ_1 or from τ_{n-1} , v is the endpoint of exactly one edge from E_1 or E_{n-1} , respectively, and there is no other set E_j which contains an edge that has

Figure 24: In order of rows, top to bottom: intersections q and p_1 with the interior of γ shaded, q with the interior of H_{τ_i} shaded, \tilde{q} with the interior of H_{τ_i} shaded, σ , and $\tilde{\sigma}$



v as an endpoint. Thus, the degree is 1. For any vertex v generated from τ_i with $i \in \{1, \dots, n-1\}$, v is the endpoint of an edge from E_i , and is also the endpoint of an edge from E_{i+1} . Hence, it has degree 2.

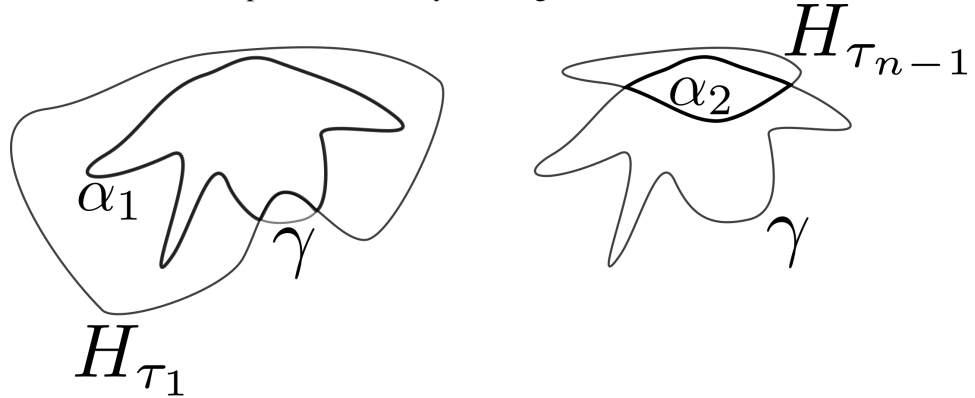
The third statement follows from looking at the set of all valid subcurves at τ_1 and τ_{n-1} . At each of these points, there are exactly two intersections between H and γ , and so it is a simple exercise to look at each of the four subcurves and to show that the only ones that are valid are the ones depicted in Figure 25.

The last statement comes from examining how we add edges. In all of the cases, we are tracing two intersection points back or forth, and keeping track of one segment of H_τ that connects these 2 points and one segment of γ that also connects these two points, which generates a continuous homotopy. Since both γ and H_τ are bounded in length by $L + \varepsilon$, taking a segment of one and joining it with a segment from the other has length at most

$$2L + \varepsilon.$$

The fact that they all contain η_{end} is a result of two observations. First, all subcurves at any τ_i respect γ , and so contain η_{end} . Second, as a result of Lemma 2.8, no intersections between H and γ lie in η_{end} for any $\tau \in [\tau_1, \tau_{n-1}]$. \square

Figure 25: The curve that corresponds to the only vertex generated from τ_1 , and the curve that corresponds to the only vertex generated from τ_{n-1}



We can now prove Lemma 2.2.

Proof of Lemma 2.2. From Lemma 2.13, we have that there is only 1 vertex v generated from τ_1 , and one vertex w generated from τ_{n-1} . Additionally, they have degree 1, and all other vertices in Γ have degree 2. As a result, we have that there

is a path in Γ from v to w . Let α_1 and α_2 be the subcurves that correspond to v and w , respectively. Due to the property of Γ that edges represent homotopies over closed curves of length at most $2L + \varepsilon$, there is thus a homotopy from α_1 to α_2 over such curves. Furthermore, every curve in this homotopy contains η_{end} .

We now observe that γ is homotopic to α_1 over curves of length at most $2L + \varepsilon$, and so we can homotope γ to α_2 over closed curves with the same length bound. All of these curves also contain η_{end} .

The rest of the proof depends on whether H contracts γ to a point inside γ or outside γ . If it contracts γ to a point outside γ , then we see that α_2 can be contracted to the point $x \in \eta_{\text{end}}$ on γ through curves that contain x . Since η_{end} is contained in all curves in this homotopy up to α_2 , we conclude that x is contained in every curve in this entire contraction.

If H contracts γ to a point inside γ , then recalling that τ^* is the last point at which H intersects γ , and x is the point of tangential intersection at τ^* , we can homotope α_2 to H_{τ^*} through curves containing x and which are bounded in length by $2L + \varepsilon$. Since γ can also be homotoped to α_2 through such curves, this gives us a desirable homotopy from γ to H_{τ^*} . This completes the proof. \square

Finally, we illustrate this process using an explicit homotopy. This is shown in Figure 26.

2.2 Proof of Lemma 2.3

We now prove Lemma 2.3. Given a curve H_{τ^*} and a point $x \in \gamma \cap H_{\tau^*}$, we want to show that we can contract H_{τ^*} through curves based at x , and of length at most $2L + 2d + \varepsilon$. The idea here is to employ a method similar to that used in this article to produce a contraction of the boundary of a Riemannian disc from a monotone contraction of that boundary. To do this, let c be the point that H contracts H_{τ^*} to. Join x to c via a minimal geodesic, and let y be the last point of intersection between this geodesic and H_{τ^*} . This is depicted in Figure 27.

Our homotopy now works as follows. One should refer to Figure 28 for a visual reference. Let β be the segment of the minimal geodesic that connects y to c entirely in the interior of H_{τ^*} . Let the length of β be B ; we of course have that $B \leq d$, where d is the diameter of the manifold. Let α be a segment of H_{τ^*} that connects x to y which is of length at most $\frac{L+\varepsilon}{2}$. We now produce our contraction of H_{τ^*} in three parts. The first part is a homotopy from H_{τ^*} to the curve formed by traversing α from x to y , following by traversing the entirety of H_{τ^*} from y to y , and then by traversing $-\alpha$ from y back to x . This homotopy consists of curves bounded in length by $2L + \varepsilon$. Let us call this curve η .

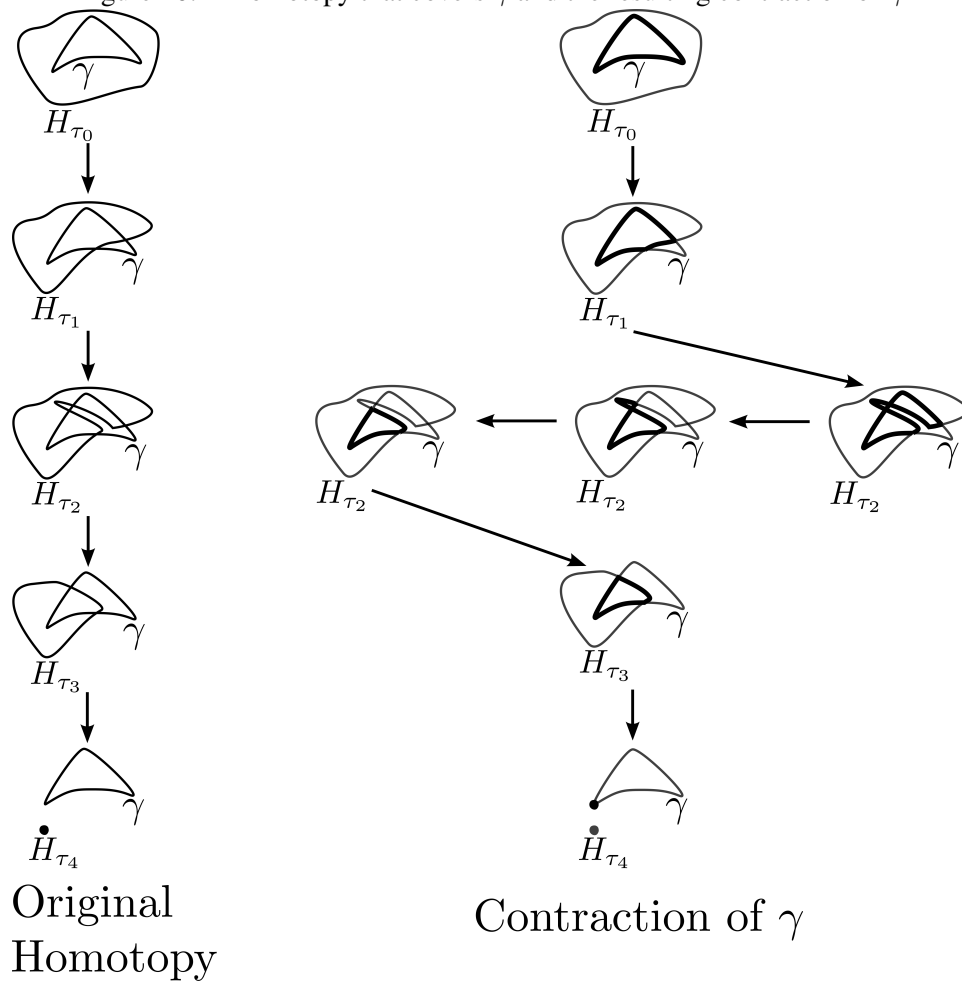
Figure 26: A homotopy that covers γ and the resulting contraction of γ 

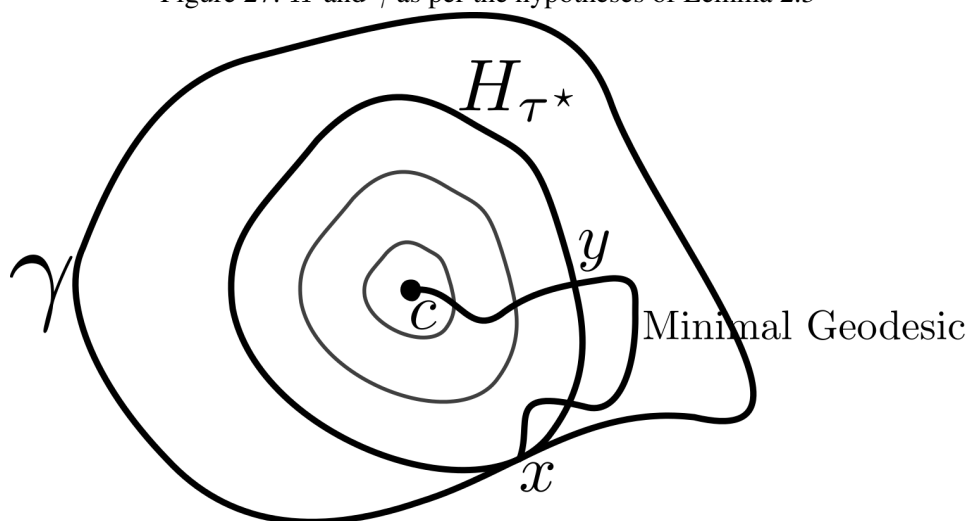
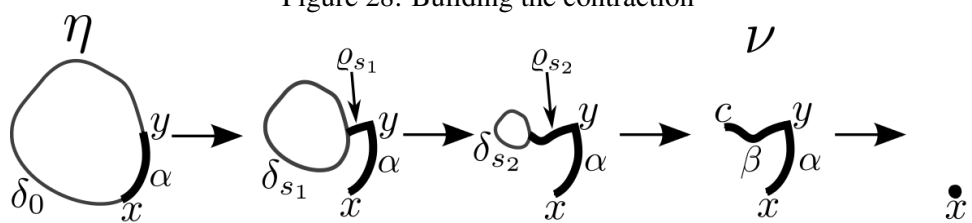
Figure 27: H and γ as per the hypotheses of Lemma 2.3

Figure 28: Building the contraction



The second step is a homotopy from η to the curve formed by traversing α from x to y , then β from y to c , then $-\beta$ from c back to y , then $-\alpha$ from y back to x . Let us call this curve ν . This homotopy, P , is defined on the interval $[0, B]$, where (as above) B is the length of β . For each $s \in [0, B]$, let ϱ_s be the segment of β from y which has a length of s . Since H is monotone, there is exactly one curve δ corresponding to a curve in the homotopy H which has the property that it goes through y if $s = 0$, and that it goes through the endpoint of ϱ_s which is not y if $s > 0$. Now, we define $P(s)$ to be the curve formed by traversing α , then ϱ_s , then δ , then $-\varrho_s$, then $-\alpha$. Since H is monotone, this produces a continuous homotopy of piecewise smooth simple curves of length at most $2L + 2d + \varepsilon$.

The third step is that we homotope ν to the point x by contracting it in the obvious way; since it is a curve traversed forward from x to y to c and then backward from c to y to x , it is obvious how to do this without exceeding a length bound of $L + 2d + \varepsilon$.

By concatenating the above homotopies, we get a homotopy of closed curves with the desired properties, completing the proof.

Acknowledgements: This work was supported in part by an NSERC Discovery Grant (Rotman), by an NSERC Postgraduate Scholarship (Chambers), and by an Ontario Graduate Scholarship (Chambers). This paper was partially written during the authors' visit of the Max-Planck Institute for Mathematics in Bonn. The authors would like to thank the Max-Planck Institute for its kind hospitality.

References

- [BS] F. BALACHEFF, S. SABOURAU, Diastolic and isoperimetric inequalities on surfaces, *Ann. Sci. École Norm. Sup.* 43 (2010), 579-605.
- [CL] G. CHAMBERS, Y. LIOKUMOVICH, Converting homotopies to isotopies and dividing homotopies in half in an effective way, preprint, math arXiv:1311.0779.
- [C] C. B. CROKE, Area and the length of the shortest closed geodesic, *J. Diff. Geom.* 27 (1988), 1-21.
- [FK] S. FRANKEL, M. KATZ, The Morse landscape of a Riemannian disc, *Ann. Inst. Fourier, Grenoble* 43, 2 (1993), 503-507.
- [Gr] M. GROMOV, Asymptotic invariants of infinite groups in *Geometric Group Theory*, v. 2, 1-295, London Math. Soc. Lecture notes Ser., vol. 182, Cambridge Univ. Press, Cambridge, 1993.
- [LNR] Y. LIOKUMOVICH, A. NABUTOVSKY, R. ROTMAN, Contracting the boundary of a Riemannian 2-disc, preprint, math arXiv:1205.5474.

- [L1] Y. LIOKUMOVICH, Spheres of small diameter with long sweep-outs, Proc. Amer. Math. Soc. 141 (2013), 309-312.
- [L2] Y. LIOKUMOVICH, Surfaces of small diameter with large width, preprint, math arXiv:1307.2306, to appear in Journal of Topology and Analysis, 2013.
- [NR1] A. NABUTOVSKY, R. ROTMAN, Linear bounds for lengths of geodesic loops on Riemannian 2-spheres, J. Diff. Geom. 89 (2011), 217-232.
- [NR2] A. NABUTOVSKY, R. ROTMAN, Linear bounds for lengths of geodesic segments on Riemannian 2-spheres, J. Topol. Anal. 05, 409 (2013), 409-438.
- [NR3] A. NABUTOVSKY, R. ROTMAN, Length of geodesics and quantitative Morse theory on loop spaces, Geom. Funct. Anal., vol. 23 (2013), 367-414.
- [P] P. PAPASOGLU, Contracting thin disks, preprint, math arXiv:1309.2967.

Gregory R. Chambers
Department of Mathematics
University of Chicago
Chicago, Illinois 60637
USA
e-mail: chambers@math.uchicago.edu

Regina Rotman
Department of Mathematics
University of Toronto
Toronto, Ontario M5S 2E4
Canada
rina@math.utoronto.ca