MEASURING SIMILARITY BETWEEN CURVES ON 2-MANIFOLDS VIA HOMOTOPY AREA

Erin Wolf Chambers^{*} and Yusu Wang[†]

ABSTRACT. Measuring the similarity of curves is a fundamental problem arising in many application fields. There has been considerable interest in several such measures, both in Euclidean space and in more general setting such as curves on Riemannian surfaces or curves in the plane minus a set of obstacles. However, so far, efficiently computable similarity measures for curves on general surfaces remain elusive. This paper aims at developing a natural curve similarity measure that can be easily extended and computed for curves on general orientable 2-manifolds. Specifically, we measure similarity between homotopic curves based on how hard it is to deform one curve into the other one continuously, and define this “hardness” as the minimum possible surface area swept by a homotopy between the curves. We consider cases where curves are embedded in the plane or on a triangulated orientable surface with genus g , and we present efficient algorithms (which are either quadratic or near linear time, depending on the setting) for both cases. The results are also extended to comparing simple cycles (simple closed curves) in the plane or on a sphere, although the algorithms become less efficient by a factor of n . We also discuss the case of cycles on surfaces, which remains open.

1 Introduction

Measuring curve similarity is a fundamental problem arising in many application fields, including graphics, computer vision, and geographic information systems. Traditionally, much research has been done on comparing curves embedded in the Euclidean space. However, in many cases it is natural to study curves embedded in a more general space, such as a terrain or a surface.

In this paper, we study a natural measure of curve similarity both in the plane and on more general surfaces. Specifically, given two simple homotopic curves embedded on an orientable 2-manifold (including the plane), we measure their similarity by the minimum total area swept when deforming one curve to the other (the “area” of the homotopy between them), and present efficient algorithms to compute this new measure.

Related work. From the perspective of computational geometry, the most widely studied similarity measures for curves is the Fréchet distance. Intuitively, imagine that a man and his dog are walking along two paths with a leash between them. The Fréchet distance between these two paths is the minimum leash length necessary for them to move from one

^{*}Dept. of Computer Science, Saint Louis University, echambe5@slu.edu

[†]Dept. of Computer Science and Engineering, The Ohio State University, yusu@cse.ohio-state.edu

end of the paths to the other end without back-tracking. Since the Fréchet distance takes the “flow” of the curves into account, in many settings it is a better similarity measure for curves than alternatives such as the Hausdorff distance [5, 6].

Given two polygonal curves P and Q with n total edges in $\mathbb{R}^d$, the Fréchet distance can be computed in $O(n^2 \log n)$ time [4]. An $\Omega(n \log n)$ lower bound for the decision problem in the algebraic computation tree model is known [11], and Alt has conjectured that the decision problem is 3SUM-Hard [2]. Recently, Buchin et al. [12] showed that there is a real algebraic decision tree to solve the Fréchet problem with sub-quadratic depth, suggesting that perhaps this is not the case. They also give an improved algorithm which runs in $O(n^2 \sqrt{\log n} (\log \log n)^2)$ time. Very recently, Agarwal et al. presented a novel approach to compute the discrete version of the Fréchet distance between two polygonal curves in sub-quadratic time [1]. This is the first algorithm for any variant of the Fréchet distance to have a sub-quadratic running time for general curves. No previous algorithm, *exact or approximate*, with running time $o(n^2)$ was known for general curves, although sub-quadratic approximation algorithms for special families of curves are known [6, 7, 22].

While the Fréchet distance is a natural curve similarity measure, it is sensitive to outliers. Variants of it, such as the summed-Fréchet distance, and the partial Fréchet similarity, have been proposed [13, 14, 24], usually at the cost of further increasing the time complexity.

The problem of extending and computing the Fréchet distance to more general metric space has also received much attention. Geodesic distance between points is usually considered when the underlying domain is not $\mathbb{R}^d$. For example, Maheshwari and Yi [31] computed the geodesic Fréchet distance between two polygonal paths on a convex polytope in roughly $O(n^3 K^4 \log(Kn))$ time, where n and K are the complexity of the input paths and of the convex polytope, respectively. Raichel and Har-Peled consider approximating the weak Fréchet distance between simplicial complexes in $\mathbb{R}^d$ [28]. Geodesic Fréchet distance between polygonal curves in the plane within a simple polygon has also been studied [8, 18, 25]. In general, computing (variants of) the Fréchet distance seems to typically induce high computational complexity for more general domains.

One issue with generalizing Fréchet distance directly to surfaces is that the underlying topology is not taken into account; for example, in geodesic Fréchet distance, while the length of the leash varies continuously, the actual leash itself does not. As a result, several measures of similarity have been proposed which take the underlying topology into account. Chambers et al. [15] proposed the so-called homotopic Fréchet distance and gave a polynomial (although not efficient) algorithm for when the curves reside in a planar domain with a set of polygonal obstacles. The extra requirement for this homotopic Fréchet distance is that the leash itself and not just its length has to vary in a continuous manner, essentially restricting the homotopy class which the leash is in. A stronger variant called isotopic Fréchet distance has also been proposed and investigated, although no algorithms at all are known to even approximate this distance [16].

Orthogonal to homotopic Fréchet distance is the concept of the *height* of a homotopy; instead of minimizing the maximum leash length, this measure views the homotopy as tracing a way for the first curve to deform to the second curve, where the goal is to min-

imize the longest intermediate curve length. Introduced independently in two very different contexts [10, 17], it is not even known if the problem is in NP.

Recent work on approximating the homotopy height and the homotopic Fréchet distance has yielded efficient $O(\log n)$ approximation algorithms for both of these problems [27]. A recent paper also considers finding the homotopy area of a curve in the plane where the curve may be self-intersecting, in contrast to our setting where we consider two simple and regular curves that share endpoints [34, 26]. This results in a polynomial time algorithm which runs in time quintic in the number of regions in the decomposition of the plane given by the input curve.

New work. In this paper, we develop a natural similarity measure for curves on general surfaces that can be computed both quickly and exactly, although it requires that the endpoints of the curve be matched (in contrast to Fréchet distance and its variants). Intuitively, we measure distances between homotopic curves based on how hard it is to deform one curve into the other one, and define this “hardness” as the minimal total surface area swept by a homotopy between them, which we call the optimal homotopy area. Our similarity measure is natural, and robust against noise (as the area in a sense captures average, instead of maximum, deviation from one curve to the other). To the best of our knowledge, this is the first similarity measure for curves on general surfaces with efficient polynomial-time algorithms to compute it exactly.

It is worth noting that this definition in a way combines homotopic Fréchet distance with homotopy height; those measures compute the “width” and “height” of the homotopy, while our measure calculates the total area. It is thus interesting that while no exact algorithms are known for either of those measures on surfaces, we are able to provide a polynomial running time for computing the area of a homotopy.

We consider cases both where curves are embedded in the plane, or on a closed, triangulated orientable surface with genus g . In the plane, our algorithm runs in $O(n \log n + I^2 \log I)$ time, where n is the total complexity of input curves and I is the number of intersections between them. On a surface, if the input is a triangulation of complexity N , then our algorithm runs in time $O(I^2 \log I + ng \log n + N)$. While our similarity measure is more expensive to compute for the case of curves in the plane than the Fréchet distance when $I = \omega(n)$, one major advantage is that this measure can be computed on general orientable surfaces efficiently. In fact, the ideas and algorithms behind the planar case form the foundation for the handling of the case on general surfaces. Finally, these results can also be extended to comparing two intersecting simple cycles (closed curves) in the plane or on a sphere, although the time complexity then has an extra linear factor.

The main ideas behind our approach are developed by examining some properties of one natural class of homotopies, including a relation with the winding number of a closed curve in the plane. Specifically, the use of the winding number enables us to compute the optimal homotopy area efficiently, where the homotopy is restricted to be piecewise differentiable and regular. This forms the basis of our dynamic programming framework to compute similarity between curves in the plane. We also show how to build efficient data structures to keep the total cost of the dynamic program low. The algorithm on more general oriented surfaces then follows by lifting the curves from the surface to a planar

covering space, and solving the problem using the planar algorithm.

We remark that the idea of measuring deformation areas has been used before in practice [19, 32]. For example, similarity between two convex polygons can be measured by their symmetric difference [3, 42]; we note that this is not equivalent to homotopy area, although it may be the same value in some situations. In another paper, the area sandwiched between an *x-monotone curve* and another curve is used to measure their similarity [9].

2 Definitions and Background

Paths and cycle. We will assume that we are working on an orientable 2-manifold M (which could be the plane). A *curve* (or a *path*) on a surface M is a map $P : [0, 1] \rightarrow M$; a *cycle* (or a *loop*) is a continuous map $\gamma : S^1 \rightarrow M$ where S^1 is the unit circle. A curve P or a cycle γ is *simple* if $P(t_1) \neq P(t_2)$ (resp. $\gamma(t_1) \neq \gamma(t_2)$) for any $t_1 \neq t_2$.

Homotopy A *homotopy* between two paths P and Q (with the same endpoints) is a continuous map $H : [0, 1] \times [0, 1] \rightarrow M$ where $H(0, \cdot) = P$, $H(1, \cdot) = Q$, $H(\cdot, 0) = P(0) = Q(0)$ and $H(\cdot, 1) = P(1) = Q(1)$. A homotopy describes a continuous deformation between the two paths or curves: for any value $t \in [0, 1]$, we let $H_t = H(t, \cdot)$ be the intermediate curve at time t , where $H_0 = P$ and $H_1 = Q$.

We define the *area* of a homotopy H to be the total area covered by the image of the homotopy on the surface, where an area that is covered multiple times will be counted with multiplicity. More precisely, given a homotopy H whose image is piecewise differentiable,

$$\text{Area}(H) = \int_{s \in [0,1]} \int_{t \in [0,1]} \left| \frac{dH}{ds} \times \frac{dH}{dt} \right| ds dt. \quad (1)$$

The *minimum homotopy area* between P and Q is the infimum of the areas of all homotopies between P and Q , denoted by $\sigma(P, Q)$. If such an infimum is realizable and can be achieved by a homotopy, we call that homotopy an *optimal homotopy*.

We note that it is not immediately clear that this value exists, depending on the curves and underlying homotopy. Minimum area homotopies were considered by Douglas [21] and Rado [35] in the context of Plateau's problem; they noted that not only is the integral improper in general, but the infimum itself may not be continuous. The eventual proof that these exist in $\mathbb{R}^n$ relies on a definition using Dirichlet integrals, which allow further control over the parameterizations of the minimal surface. See the book by Lawson [30] for an overview of this result as well as several extensions to minimal area submanifolds in more general settings.

However, beyond a proof of existence, we are interested in computing such homotopies, or at least measuring their actual area, in much simpler settings such as $\mathbb{R}^2$ or a surface. To this end, we will restrict the class of curves and the types of allowable homotopies to guarantee existence and finiteness, as well as to allow a polynomial running time. In particular, we will have two input curves which each consist of a finite number of piecewise analytic components. We

do not require their union to be a Jordan curve, as our algorithm serves as a constructive proof that in the planar or surface-embedded case we can find a minimum area homotopy even when the boundary is not simple. We will restrict to the class of homotopies which are continuous and piecewise differentiable, so that the integral can be defined; this means they will also be regular homotopies, where at any time t the intermediate curve H_t is *regular* (see [43] for smooth curves and [33] for piecewise-linear curves).

Here, a curve $C : [0, 1] \rightarrow M$ is *regular* if its first derivative $C'(t)$ is defined and not zero for all $t \in [0, 1]$. In our paper, similar to the setting of [23], we allow a curve $C : [0, 1] \rightarrow M$ to be *piecewise-regular*, meaning that C is continuous and the above conditions on C' holds for all t other than potentially a finite number of singular points $\{s_1 < s_2 < \dots < s_m\} \subset [0, 1]$. Furthermore, at each singular point s_i , the right and left limits of C' exists. A *regular homotopy* is a homotopy $H : [0, 1] \times [0, 1] \rightarrow M$ such that each intermediate curve $H(t, \cdot)$ is piecewise-regular.

Our goal is then to compute an *optimal regular homotopy* whose homotopy area (as defined in Eqn (1) is smallest possible among all regular homotopies between input curves P and Q . For example, this means that the deformation is “kink-free” [33], and cannot create or destroy a local loop as shown in the above figure (the singular point in the right curve is a kink). In the remainder of this paper, the term “homotopy” between P and Q refer to such regular homotopy.

Requiring regularity will allow us to use a connection to the winding number, which is the primary tool in our calculation. We also note that regular homotopies between regular curves which are homotopic always exist when the underlying curves have the same winding number [43], and that similar results also hold in Riemannian manifolds [39]. This constraint is also quite natural in another way, since we are interested in homotopies which are in some sense are “good” morphisms, mapping one curve or cycle to the other without introducing unnecessary complexity in the intermediate curves. More general (including non-regular) curves are considered in more recent work on homotopy area, but with considerably worse running times [34, 26].

Decomposing arrangements. Consider two simple piecewise analytic curves P and Q with the same endpoints. Their concatenation forms a (not necessarily simple) closed curve denoted by $C = P \circ \text{rev}(Q)$, where $\text{rev}(Q)$ is the reversal of Q . Let $\text{Arr}(C)$ denote the arrangement formed by C , where vertices in $\text{Arr}(C)$ are the intersection points between P and Q . An edge / arc in $\text{Arr}(C)$ is a subcurve of either P or Q . See Figure 1 (a) for an example. We remark that in the remainder of the paper, for simplicity of illustration, we always draw as if curve Q is a straight line (see Figure 1 (b)). However, note that this view does not change the combinatorial structure of the arrangement $\text{Arr}(C)$, and is purely for illustration purpose.

We can view C as an oriented curve, with the natural orientation inherited from the parametrization of P and $\text{rev}(Q)$ (see Figure 1 (b)). Hence we can talk about the *sidedness* with respect to C at a point $p \in P$. Specifically, a point $x \in \mathbb{R}^2$ is to the *right* of C at p if it is a counter-clockwise turn from the orientation of the vector px to the orientation of (tangent of) C at p (see Figure 1 (b) for an example). Given two oriented curves γ_1 and γ_2 , an intersection point p of them is *positive* if it is a counter-clockwise turn from the

Figure 1: (a) Two curves P and Q . The shaded region represents a cell in the arrangement $\text{Arr}(C)$ formed by $C = P \circ \text{rev}(Q)$. (b) Simplified view where Q is straight: The topological structure of the arrangement remains the same. The shaded region in (b) corresponds to the same shaded cell in (a).

orientation of γ_1 to that of γ_2 at p . For a curve γ and a point $x \in \gamma$, the *index of x* is the parameter of x under the arc-length parameterization of γ . We sometimes use x to represent its index along γ when its meaning is clear from the context. Given two points $x, y \in \gamma$, we will use $\gamma[x, y]$ to denote the unique sub-curve of γ between points x and y .

We say that a homotopy H from P to Q is *right sense-preserving* if it moves only to the “right” or stays fixed at all points, i.e. for any $t, s \in [0, 1]$, we have that either $H_{t+dt}(s) = H_t(s)$ or $H_{t+dt}(s)$ is to the right of the oriented curve H_t at $H_t(s)$. If it is the former case, then we say that $p = H_t(s)$ is a *fixed point* at time t . Similarly, we say that H is *left sense-preserving* if for any $t, s \in [0, 1]$, $H_t(s)$ is either a fixed point or deforms to the left of the curve H_t . A homotopy H is *sense-preserving* if it is either right or left sense-preserving. The sense-preserving property means that we can continuously deform the curve P always in the same direction, without causing local folds in the regions swept. Intuitively, any minimum area homotopy should have this property to some extent, which we will make more precise and prove later.

3 Structure of Optimal Homotopies

Given two simple curves P and Q (with the same start and end point, oriented in the same direction) embedded on an orientable 2-manifold M , let $X = \{\mathbf{x}_1, \dots, \mathbf{x}_I\}$ denote the set of I intersection points between them, sorted by their order along P . Given a homotopy H from P to Q , a point $p \in M$ is called an *anchor point with respect to H* if it remains on $H(t, \cdot) = H_t$ at all times $t \in [0, 1]$. Of course not all intersection points are anchor points. However, if p is an anchor point, then it is necessarily an intersection point between P and Q , as $p \in H_0 = P$ and $p \in H_1 = Q$. We exclude the beginning and ending end points of P and Q from the list of anchor points, as they remain fixed for all homotopies. In what follows, we show that any minimum area homotopy can be decomposed by anchor points such that each of the resulting smaller homotopies has a simple structure.

Specifically, consider an arbitrary homotopy with minimum area H^* . Let $\mathbf{B} = \{\mathbf{b}_1, \dots, \mathbf{b}_k\}$ be the set of anchor points with respect to H^* , the minimum area homotopy. We order the $\mathbf{b}_i$ ’s by their indices along P . It turns out that the order of their indices along

Q is the same:

Lemma 3.1 *The order of $\mathbf{b}_i$'s along P and along Q are the same.*

Proof: Note that H^* is a map from $\square \rightarrow M$, where $\square = [0, 1] \times [0, 1]$ is the unit square and a point $(s, t) \in \square$ will be mapped to $H_t^*(s)$. See the right figure for an illustration. Specifically, the figure shows the domain $\square$, where the bottom segment corresponds to the parametrization for P , while the top for Q – they will be mapped to P and Q , respectively. (Since P and Q share starting and ending endpoint, the left and right sides of $\square$ will be mapped to the same point.) Given an anchor point $\mathbf{b}_i$, let $\mathbf{p}_i$ and $\mathbf{q}_i$ be the parameters of $\mathbf{b}_i$ in H_0^* and H_1^* , respectively; that is, $H_0^*(\mathbf{p}_i) = H_1^*(\mathbf{q}_i) = \mathbf{b}_i$. By definition of anchor points, the pre-image of $\mathbf{b}_i$ under the map H^* necessarily includes a curve in $\square$ connecting $\mathbf{p}_i$ on the bottom edge to $\mathbf{q}_i$ on the top boundary edge of $\square$. Since $\mathbf{b}_i \neq \mathbf{b}_j$, the pre-images of $\mathbf{b}_i$ cannot intersect with that of $\mathbf{b}_j$. Hence no two such curves can intersect each other, which means that $\mathbf{p}_i$ s must be ordered in the same way as $\mathbf{q}_i$ s. ■

This observation implies that we can decompose H^* into a list of sub-homotopies, where H_i^* morphs $P[\mathbf{b}_i, \mathbf{b}_{i+1}]$ to $Q[\mathbf{b}_i, \mathbf{b}_{i+1}]$. Obviously, each H_i^* is necessarily optimal, and it induces no anchor points. The following result states that an optimal homotopy without anchor points has a simple structure, which is sense-preserving. Intuitively, if any point changes its deformation direction at any moment, the deformation will sweep across some area redundantly and thus cannot be optimal.

Lemma 3.2 *If a minimum area homotopy H from P to Q has no anchor points, then it is sense-preserving.*

Proof:

Consider an arbitrary time t in the homotopy and let $\gamma = H_t$. First note that if some portion of γ is left sense-preserving at time t and then reverses its direction and becomes right sense preserving at time t^+ a small amount later, some portion of the domain has been swept twice. Note for example the image at the right, where γ swept forward to γ^+ , but then a portion reversed when it swept to become γ^- . Here, the red section of the area was swept twice. This means that H cannot have minimal area, since we can create a smaller one which stops at time t and moves directly to some intermediate curve at a time greater than t^+ without sweeping this portion twice. This implies that for any point along the curve, the homotopy always moves to the same (local) direction. More precisely, for any $s \in (0, 1)$ if $H_t(s)$ is moving to the right (resp. left) of curve H_t for some t , then for all $t' \in (0, 1)$, $H_{t'}(s)$ is either a fixed point or is moving to the right (resp. left) of $H_{t'}$.

Now for any $s \in (0, 1)$, we mark it '+' if $H_t(s)$ moves to the right for some $t \in [0, 1]$, or '-' if $H_t(s)$ moves to the left for some $t \in [0, 1]$. Note that every s needs to receive either a '+' or '-' as it cannot remain a fixed point for all $t \in [0, 1]$, since this would then imply that we have an anchor point. Suppose the minimum homotopy is not sense-preserving, then there will be both '+' and '-' in the signs received for $s \in (0, 1)$. By continuity of our

homotopy H (and since no point can get both a $+$ and a $-$), as the sign changes from $+$ to $-$ it is necessary that we have some value s remaining a fixed point for all $t \in [0, 1]$. This is not possible, as this means that this point gives rise to an anchor point. Hence all $s \in (0, 1)$ receive either all $+$ or all $-$, and so the minimum area homotopy is sense-preserving if it contains no anchor points. ■

4 Minimum Area Homotopies for Two Curves in the Plane

In this section, we consider the case where the input consists of two simple polygonal curves in the plane. We develop an algorithm to compute the similarity between P and Q in $O(I^2 \log I + n \log n)$ time, where n is the total complexity of input curves and I is the number of intersections. Note that $I = \Theta(n^2)$ in the worst case, although of course it may be much smaller. Although efficient algorithms for comparing curves in the plane exist (such as the Fréchet distance), our planar algorithm will be the fundamental component for comparing curves on general surfaces in the next section.

4.1 Relations to Winding Numbers

We are given two simple (open) curves in the plane which share common endpoints. Previously, we have shown that if an optimal homotopy does not induce anchor points, then it is sense-preserving. The implication of this result is manifested by using the *winding number*, defined for a loop γ in the plane at a base point x .

Intuitively, imagine starting from a point y on γ , and connecting x and y by a string. The winding number at x w.r.t. γ , denoted by $\text{wn}(x; \gamma)$, is an integer measuring how many times this string winds, in a clockwise manner, around x as y traverses γ . More formally, consider an infinite ray f based at x which is generic (so it has a finite set of transversal intersections / crossings with γ). Consider a crossing $\gamma(t)$ between the ray f and γ . This crossing is *positive* if the triangle $x, \gamma(t)$, and $\gamma(t + \epsilon)$ is oriented counterclockwise, and is *negative* if oriented clockwise. Then $\text{wn}(x; \gamma)$ is the number of positive crossings minus the number of negative crossings with respect to any generic ray from x .

We say an oriented curve γ has *consistent winding numbers* if $\text{wn}(x, \gamma)$ is either all non-negative, or all non-positive, for all $x \in \mathbb{R}^2$. Note that for a curve with consistent winding numbers, we can always orient the curve appropriately so that $\text{wn}(x, \gamma)$ is all non-negative. Two examples are shown in the figure on the right, where the second example has consistent winding numbers. Let $\text{Arr}(\gamma)$ denote the arrangement formed by the curve γ . All points in the same cell of the arrangement $\text{Arr}(\gamma)$ of γ have the same winding number, and the winding numbers of two neighboring cells differ by 1. The relation of consistent winding numbers and sense-preserving homotopies is given below:

Lemma 4.1 *If there is a sense-preserving homotopy H from P to Q , then the closed curve $P \circ \text{rev}(Q)$ has consistent winding numbers.*

Proof: Without loss of generality, assume that the map H is left sense-preserving, always deforming an intermediate curve to its left. Consider the time-varying function $F : [0, 1] \times \mathbb{R}^2 \rightarrow \mathbb{Z}$, where $F(t, x) = \text{wn}(x; H_t)$ is the winding number at $x \in \mathbb{R}^2$ with respect to the curve parameterized by H_t . Obviously, $F(0, x) = \text{wn}(x; P \circ \text{rev}(Q))$, and $F(1, x) = 0$. During the deformation, $F(t, x)$ changes by either 1 or -1 whenever the intermediate curve sweep over it. Since the homotopy is left sense-preserving, when an intermediate curve sweeps x , x always moves from the left side of the intermediate curve to its right side. Hence the winding number x decreases monotonically. Since in the end, the winding number at each point is zero, $\text{wn}(x; P \circ \text{rev}(Q)) = F(0, x) \geq 0$.

If the map H is right sense-preserving, then a symmetric argument shows that $\text{wn}(x; P \circ \text{rev}(Q)) \leq 0$ for all $x \in \mathbb{R}^2$. ■

Next, we describe two results to connect the above lemma to the computation of optimal homotopy. First, we define the *total winding number* $\text{Tw}(\gamma)$ of a curve γ as

$$\text{Tw}(\gamma) = \int_{\mathbb{R}^2} \text{wn}(x; \gamma) d\nu(x),$$

where $d\nu(x)$ is the area form¹. The following observation is straightforward.

Lemma 4.2 *For any P and Q in the plane,*

$$\sigma(P, Q) \geq |\text{Tw}(P \circ \text{rev}(Q))|.$$

Proof: Take any regular homotopy H from P to Q . The area of a regular homotopy H in our setting can be reformulated as an integral on the image domain as

$$\text{Area}(H) = \int_{\mathbb{R}^2} \text{deg}_H(x) d\nu(x),$$

where $\text{deg}_H(x)$ is defined as the number of connected components in the pre-image of x under H . In other words, $\text{deg}_H(x)$ is the number of times that any intermediate curve H_t sweeps through x . Now consider the function $F : [0, 1] \rightarrow \mathbb{R}$ defined as $F(t) = \text{Tw}(H_t \circ \text{rev}(Q))$. Obviously, $F(0) = \text{Tw}(P \circ \text{rev}(Q))$, $F(1) = 0$, and F is a continuous function. Furthermore, each time the winding number at a point x changes by 1 for some $t \in [0, 1]$, it means that some intermediate curve $H(t)$ sweeps through it. Hence $|\text{wn}(x)|$ is a lower bound for $\text{deg}_H(x)$. We thus have that

$$|\text{Tw}(P \circ \text{rev}(Q))| \leq \int_{\mathbb{R}^2} |\text{wn}(x)| d\nu(x) \leq \int_{\mathbb{R}^2} \text{deg}_H(x) d\nu(x)$$

for any regular homotopy H , implying that

$$|\text{Tw}(P \circ Q)| \leq \inf_H \text{Area}(H) = \sigma(P, Q).$$

■

¹Note that this allows us to use any Riemannian metric on the plane (including the standard Euclidean metric). This will be necessary later when we use the same algorithm for curves in a universal covering space.

Figure 2: (a) The cell R with highest positive winding number. Its boundary consists of alternating P -arcs (red) and Q -arcs (green). The two cases of relations between $P[p, q]$ and R are shown in (b) and (d), respectively. For case (b), we can deform P to sweep through Ω as shown in (c), and reduce the number of intersections by 2. Similarly, for case (d), we can identify any bigon R' and deform P to reduce the number of intersections by 2 as well.

Lemma 4.3 *Given P and Q , if $\gamma = P \circ \text{rev}(Q)$ has consistent winding numbers, then $\sigma(P, Q) = |\text{Tw}(\gamma)|$.*

Proof: We prove the claim by induction on the number of intersections between P and Q . The base case is when there is no intersection between P and Q . In this case, γ is a Jordan curve which decomposes $\mathbb{R}^2$ into two regions, one inside γ and one unbound. By orienting γ appropriately, every point in the bounded cell has winding number 1 and the claim follows.

Now assume that the claim holds for cases with at most $k - 1$ intersections. We now prove it for the case with k intersections. Let an X -arc denote a subcurve of curve X . Consider the arrangement $\text{Arr}(\gamma)$ formed by $\gamma = P \circ \text{rev}(Q)$. Since P and Q are simple, every cell in this arrangement has boundary edges alternating between P -arcs and Q -arcs. Assume without loss of generality that γ has all non-negative winding numbers. Consider a cell $R \in \text{Arr}(\gamma)$ with largest (and thus positive) winding number. Since its winding number is greater than all its neighbors, it is necessary that all boundary arcs are oriented consistently as shown in Figure 2 (a), where the cell R (shaded region) lies to the right of its boundary arcs.

If R has only two boundary arcs, e from P and e' from Q , respectively, then we can morph P to another simple curve P' by deforming e through R to $-e'$ (where ‘ $-$ ’ means reversing the orientation). See the right figure for an illustration. The area swept by this deformation is exactly the area of cell R . Furthermore, after the deformation, every point $x \in R$ decreases their winding number by 1, and no other point changes its winding number. Since points in this cell initially have strictly positive winding number, the resulting curve $\gamma' = P' \circ \text{rev}(Q)$ still has all non-negative winding number. The number of intersections between P' and Q is $k - 2$. By induction hypothesis, $\sigma(P', Q) = \text{Tw}(\gamma')$. Since $\text{Tw}(\gamma) - \text{Tw}(\gamma') = \text{Area}(R)$, we have that $\text{Tw}(\gamma) = \sigma(P', Q) + \text{Area}(R)$. It then follows from Lemma 4.2 and the fact $\sigma(P, Q) \leq \sigma(P', Q) + \text{Area}(R)$ that $\sigma(P, Q) = \text{Tw}(\gamma)$.

Otherwise, the cell R has more than one P -arc. Take the P -arc e_1 with the smallest index along P , and let p be the ending endpoint of it. Let e_2 be the next P -arc along the boundary of R , and q its starting endpoint, and $\text{rev}(Q)[p, q]$ the Q -arc between e_1 and e_2 ,

denoted by $\bar{e}$ in Figure 2. Obviously, the subcurve $P[p, q]$ cannot intersect R , and $P[p, q]$ and $Q[q, p]$ bound a simple polygon, which we denote by Ω . Either Ω is on the opposite side of the Q -arc $\bar{e}$ from the interior of R (Figure 2 (b)), or they are on the same side (Figure 2 (d)).

Case (1): R and Ω are on opposite sides of $\bar{e}$. In this case, the region Ω is to the right of the oriented arc $P[p, q]$; see Figure 2 (b). Note that P does not intersect the interior of Ω ; as otherwise, P will either intersect itself or intersect $\bar{e}$, neither of which is possible. Hence only Q can intersect Ω . Since Q is also a simple curve, there are no vertices of $\text{Arr}(\gamma)$ contained in the interior of Ω . As a result, every cell of $\text{Arr}(\gamma)$ contained in Ω must have at least one boundary edge coming from $P[p, q]$. This implies that each cell contained in Ω has strictly positive winding number; that is, $\text{wn}(x; \gamma) > 0$ for any $x \in \Omega$. This is because if a cell $\xi \subseteq \Omega$ has winding number 0, then its neighbor across its boundary on the other side of $P[p, q]$ will have winding number -1 , as Ω is to the right of $P[p, q]$. This violates the condition that γ has all non-negative winding numbers and thus cannot happen.

We now deform P to P' by sweeping $P[p, q]$ through Ω to $\text{rev}(Q)[p, q]$. See Figure 2 (c). The cost of this sweeping is $\text{Area}(\Omega)$ and $\text{Tw}(\gamma) - \text{Tw}(P' \circ \text{rev}(Q)) = \text{Area}(\Omega)$. P' is still simple, and the number of intersection points between P' and Q is now $k - 2$. Since $\text{wn}(x; \gamma) > 0$ for any $x \in \Omega$, we have $\text{wn}(x; P' \circ \text{rev}(Q)) \geq 0$ for $x \in \Omega$. No other point will change their winding number after this deformation. Thus the curve $P' \circ \text{rev}(Q)$ has all non-negative winding numbers as well. Hence by induction hypothesis, we have that $\sigma(P', Q) = \text{Tw}(P' \circ \text{rev}(Q))$. Since $\sigma(P, Q) - \sigma(P', Q) \leq \text{Area}(\Omega)$ and $\text{Tw}(\gamma) - \text{Tw}(P' \circ \text{rev}(Q)) = \text{Area}(\Omega)$, it then follows from Lemma 4.2 that $\sigma(P, Q) = \text{Tw}(\gamma)$.

Case (2): R and Ω are both from the same side of $\bar{e}$. We now consider the remaining case as shown in Figure 2 (d). Take the unbounded region $\bar{\Omega} := \mathbb{R}^2 \setminus \Omega$ which is the complement of Ω . This unbounded region lies to the right of the oriented curve $P[p, q]$. Since both P and Q are simple, only Q can intersect Ω and there is no vertices of $\text{Arr}(\gamma)$ contained in the interior of $\bar{\Omega}$. First, observe that it is not possible that $\bar{\Omega} \cap Q = \emptyset$. This is because otherwise, $\bar{\Omega}$ is the unbounded face of $\text{Arr}(\gamma)$ and thus the winding number for all points in $\bar{\Omega}$ is 0. This however is not possible as this will imply that any point y to the immediate left of $P[p, q]$ has winding number -1 , violating our assumption that all cells in $\text{Arr}(\gamma)$ have consistent (non-negative) winding numbers.

Hence $\bar{\Omega} \cap Q \neq \emptyset$, and there is set of arcs from Q intersecting $P[p, q]$. Then there must exist a bigon cell R' bounded by only two arcs, one P -arc $P[a, b] \subseteq P[p, q]$ and a Q -arc β . See Figure 2 (d). Similar to the argument from the previous paragraph, we can show that points in R' must have strictly positive winding number. Now deform P to P' by sweeping $P[a, b]$ through R' to β as shown in Figure 2 (e). P' is still simple, the number of intersection points between P' and Q is now $k - 2$. Only points in R' reduce their winding number by 1, and the resulting arrangement still has consistent winding numbers. As such, by induction hypothesis, we have that $\sigma(P', Q) = \text{Tw}(P' \circ \text{rev}(Q))$. Since $\sigma(P, Q) - \sigma(P', Q) \leq \text{Area}(R')$ and $\text{Tw}(\gamma) - \text{Tw}(P' \circ \text{rev}(Q)) = \text{Area}(R')$, it then follows from Lemma 4.2 that $\sigma(P, Q) = \text{Tw}(\gamma)$. ■

4.2 The Algorithm

Lemma 3.2 and 4.1 imply that if the closed curve $P \circ \text{rev}(Q)$ produces both positive and negative winding numbers, then any minimum area homotopy from P to Q must have at least one anchor point. On the other hand, if it has consistent winding numbers, then by Lemma 4.3 we can compute the optimal cost to deform them by simply computing the total winding number. This leads to a simple dynamic-programming (DP) approach to compute $\sigma(P, Q)$.

Specifically, let $\mathbf{x}_0, \mathbf{x}_1, \dots, \mathbf{x}_I$ denote the intersection points between P and Q , ordered by their indices along P , where $\mathbf{x}_0$ and $\mathbf{x}_I$ are the beginning and ending points of P and Q , respectively. Let $T(i)$ be the cost of the minimum area homotopy between $P[0, \mathbf{x}_i]$ and $Q[0, \mathbf{x}_i]$, and $C[i, j]$ the closed curve formed by $P[\mathbf{x}_i, \mathbf{x}_j] \circ \text{rev}(Q)[\mathbf{x}_j, \mathbf{x}_i]$ if $\mathbf{x}_i$ is before $\mathbf{x}_j$ along Q ; or by $P[\mathbf{x}_i, \mathbf{x}_j] \circ \text{rev}(Q)[\mathbf{x}_i, \mathbf{x}_j]$ if $\mathbf{x}_i$ is after $\mathbf{x}_j$ along Q . We say that a pair of indices (i, j) is *valid* if (1) $\mathbf{x}_i$ and $\mathbf{x}_j$ have the same order along P and along Q ; and (2) the closed curve $C[i, j]$ has consistent winding numbers. We have the following recursion:

$$T(i) = \begin{cases} \text{Tw}(C[0, i]), & \text{If } C[0, i] \text{ has consistent winding numbers;} \\ \min_{k < i, (k, i) \text{ is valid}} \{ \text{Tw}(C[k, i]) + T(k) \}, & \text{Otherwise.} \end{cases}$$

4.3 Time Complexity Analysis

The main components of the DP framework described above are (i) to compute $\text{Tw}(C[i, j])$ for all pairs of i, j s, and (ii) to check whether each pair (i, j) is valid or not. These can be done in $O(I^2 n)$ total time in a straightforward manner. We now show how to compute them in $O(I^2 \log I)$ time after $O(I \log I + n \log n)$ pre-processing time. Specifically, we describe how to compute such information in $O(I \log I)$ time for all $C[r, i]$ s for a fixed $r \in [1, I]$ and all indices $i > r$.

To simplify the description of the algorithm, we extend Q on both sides via two disjoint rays till infinity, and obtain $\hat{Q}$. Now collect all intersection points between P and $\hat{Q}$, $\{\hat{\mathbf{x}}_1, \dots, \hat{\mathbf{x}}_I\}$, which is a super-set of previous intersection points, and sort them by their order along the curve P . The algorithm can be made to work with Q directly, but using $\hat{Q}$ makes the intuition behind our algorithm, as well as its description, much more clear.

Note that $\hat{Q}$ divides the plane into two parts. For *illustration purpose*, we will draw $\hat{Q}$ as a horizontal line, and use the upper and lower half-planes to refer to these two sides of $\hat{Q}$. Another way to see that regarding $\hat{Q}$ as a horizontal line does not cause any loss of generality is that one can always find a homeomorphism from $\mathbb{R}^2 \rightarrow \mathbb{R}^2$ such that the image of $\hat{Q}$ is a horizontal line under this homeomorphism. Note that viewing $\hat{Q}$ as a horizontal line does not change the number of intersections between P and $\hat{Q}$.

Now for a fixed integer r , we traverse P starting from $\hat{\mathbf{x}}_r$. We aim to maintain appropriate data structures so that each time we pass through an intersection point $\hat{\mathbf{x}}_i$ with $\hat{Q}$, we can, in $O(\log I)$ time, (1) check whether (r, i) is valid, and (2) obtain total winding number for $C[r, i]$. Given that $\hat{\mathbf{x}}_i$ s are ordered w.r.t. P , for any $\hat{\mathbf{x}}_i$ and $\hat{\mathbf{x}}_j$, we use $\hat{Q}(\hat{\mathbf{x}}_i, \hat{\mathbf{x}}_j)$ to denote $\text{rev}(Q)[\hat{\mathbf{x}}_j, \hat{\mathbf{x}}_i]$ if $\hat{\mathbf{x}}_i$ is before $\hat{\mathbf{x}}_j$ along $\hat{Q}$, and use it to denote $\text{rev}(Q)[\hat{\mathbf{x}}_i, \hat{\mathbf{x}}_j]$ otherwise.

Figure 3: Illustration of the regions R_u s.

Total winding numbers. We first explain how to maintain the total winding number for the closed curve $C[r, i] = P[\hat{\mathbf{x}}_r, \hat{\mathbf{x}}_i] \circ \hat{Q}(\hat{\mathbf{x}}_i, \hat{\mathbf{x}}_r)$ as i increases. Assume i changes from u to $u + 1$. Since $\hat{\mathbf{x}}_u$ and $\hat{\mathbf{x}}_{u+1}$ are two consecutive intersection points along P , the arcs $P[\hat{\mathbf{x}}_u, \hat{\mathbf{x}}_{u+1}]$ and $\hat{Q}(\hat{\mathbf{x}}_u, \hat{\mathbf{x}}_{u+1})$ form a simple closed polygon which we denote by R_u (shaded region Figure 3). Comparing the arrangement $\text{Arr}(C[r, u + 1])$ with $\text{Arr}(C[r, u])$, regardless of where r is, only points within R_u will change their winding number, either all by $+1$ or all by -1 , depending on whether R_u is to the right side or the left side of the P -arc $P[\hat{\mathbf{x}}_u, \hat{\mathbf{x}}_{u+1}]$, respectively. Hence the change in the total winding number is simply $\alpha_u \text{Area}(R_u)$, where α_u is either $+1$ or -1 . See Figure 3, where all points in R_u will decrease their winding number by 1 as we move from $C[r, u]$ to $C[r, u + 1]$.

We can pre-compute the area of R_u 's for all u in $O(n \log n + I \log I)$ time, by observing that the set of R_u s satisfy the parentheses property: Namely, either R_u and R_v are disjoint in their interior, or one contains the other.

Figure 4: The containment relations of all R_u regions can be represented as a forest structure on the right.

Specifically, first, we compute the arrangement of $\text{Arr}(P + \hat{Q})$ and the area of all cells in it in $O(n \log n + I)$ time. Each R_u is the region bounded between a P -arc $P[\hat{\mathbf{x}}_i, \hat{\mathbf{x}}_{i+1}]$ and a corresponding $\hat{Q}$ -segment $\hat{Q}(\hat{\mathbf{x}}_i, \hat{\mathbf{x}}_{i+1})$. Since no two P -arcs intersect, the containment relationship between such P -arcs satisfies parentheses property. In particular, we can use a collection of trees to represent the containment relation among all R_u regions. See Figure 4 for an illustration. The difference between the region represented at a parent node and the union of regions represented by all its children is a cell in $\text{Arr}(P + Q)$. For example, the shaded cell in the above figure is the difference between R_0 and its children R_2 and R_4 . We can thus compute the area of all R_u s by a bottom-up traversal of these trees. Computing these trees take $O(I \log I)$ time by first sorting all intersection points with respect to their order along $\hat{Q}$. Traversing these trees to compute all R_u s takes $O(I + n)$ time. Putting everything together, we need $O(n \log n + I \log I)$ time.

With the area of all R_u s known, updating the total winding number from $C[r, u]$ to $C[r, u + 1]$ takes only constant time.

Checking the validity of (r, i) s. To check whether (r, i) is valid or not, we need to check whether all cells in the arrangement $\text{Arr}(C[r, i])$ have consistent winding numbers. First observe that for any r and i , $\text{Arr}(P + \hat{Q})$ is a refinement of the arrangement $\text{Arr}(C[r, i])$. That is, a cell in $\text{Arr}(P + \hat{Q})$ is always contained within some cell in $\text{Arr}(C[r, i])$. Hence all points within the same cell of $\text{Arr}(P + \hat{Q})$ always have the same winding number with respect to any $C[r, i]$, and we simply need one point from each cell in $\text{Arr}(P + \hat{Q})$ to maintain the winding number for all cells in $\text{Arr}(C[r, i])$, for any r and i . We now describe how to maintain the winding number for cells of $\text{Arr}(P + \hat{Q})$ (thus for $\text{Arr}(C[r, u])$ s) as we pass each $u > r$, so that we can check whether $C[r, u]$ has consistent winding numbers or not efficiently.

To this end, take four points around each intersection point $\hat{x}_i$ of P and $\hat{Q}$ (shown as stars in the right figure). The collection of such *representative points* hit all cells in $\text{Arr}(P + \hat{Q})$. (It does not matter whether there may be more than one point taken from a cell of $\text{Arr}(P + \hat{Q})$.) Hence $\text{Arr}(C[r, i])$ has consistent winding number if and only if all these representative points have consistent winding numbers. Next, we build a data structure to maintain the winding numbers for these points as i increases. Specifically, let U be the set of representatives that are to the right of $\hat{Q}$, which are the stars above $\hat{Q}$ in the right figure. (Those to the left of $\hat{Q}$ will be handled in a symmetric manner). Each point has a key associated with it which is its index along $\hat{Q}$. We build a standard balanced 1-D range tree on U based on such keys, where each leaf f stores a point from U . Every internal node v is associated with an interval $[l_v, r_v]$, where l_v and r_v are the smallest and largest keys stored in the subtree rooted at v . In other words, all representatives with an index along $\hat{Q}$ within $[l_v, r_v]$ are stored in the subtree rooted at v . At every node v , interior or not, we also store a value $\text{add}W_v$. To compute the winding number for the representative point p_f stored at a leaf node f , we identify the path $\{v_0, v_1, \dots, v_a = f\}$ from the root v_0 to f . The winding number for p_f is simply $\sum_{i=0}^a \text{add}W_{v_i}$. Finally, each internal node v also stores the maximum and minimum winding numbers associated with all leaves in its subtree. At the beginning, all winding numbers are zero. The size of this tree is $O(I)$ with height $O(\log I)$, and can be built in $O(I \log I)$ time once the arrangement $\text{Arr}(P + \hat{Q})$ is known.

Let $\mathbf{q}_i$ denote the index of point $\hat{x}_i$ along $\hat{Q}$. As we move from $C[r, u]$ to $C[r, u + 1]$, cells of $\text{Arr}(P + \hat{Q})$ contained in R_u should either all increase or all decrease their winding number by 1. Note that representatives of these cells are simply those contained in the interval $[\mathbf{q}_u, \mathbf{q}_{u+1}]$ (or $[\mathbf{q}_{u+1}, \mathbf{q}_u]$ if $\mathbf{q}_{u+1} < \mathbf{q}_u$). Hence updating the winding number is similar to an interval query of $[\mathbf{q}_u, \mathbf{q}_{u+1}]$, and the $O(\log I)$ number of nodes in the canonical decomposition of $[\mathbf{q}_u, \mathbf{q}_{u+1}]$ update their $\text{add}W_v$ values by either $+1$ or -1 depending on the sidedness of R_u with respect to the arc $P[\hat{x}_u, \hat{x}_{u+1}]$. The minimum and maximum winding numbers can also be updated $O(1)$ time per visited node. The entire process visits $O(\log I)$ nodes, and thus takes $O(\log I)$ time as i increases from u to $u + 1$. To see whether $C[r, u + 1]$ has consistent winding numbers or not, we only need to check the minimum and maximum

winding numbers stored at the root of the tree, denoted by $w_{\min}$ and $w_{\max}$, respectively. If $w_{\min} \times w_{\max}$ equals to zero, then all winding numbers w.r.t. $C[r, u + 1]$ are either all non-negative or all non-positive. Otherwise, $(r, u + 1)$ is not valid.

Repeat the above process for every $r \in [1, I]$. Overall, after $O((n + I) \log n)$ pre-processing, we can check whether (r, i) is valid or not and compute $\text{Tw}(C[r, i])$ for all $r \in [1, I]$ and all $i > r$ in $O(I^2 \log I)$ time. Putting everything together, we have the following result.

Theorem 4.4 *Given two simple polygonal chains P and Q (with the same endpoints) in the plane of n total complexity, and with I intersection points between them, we can compute the optimal area (regular) homotopy in $O(I^2 \log I + n \log n)$ time and $O(I^2 + n)$ space.*

5 Minimum Area Homotopies for Curves on 2-Manifolds

In this section, we consider computing an minimal area homotopy between curves P and Q on a compact orientable and triangulated 2-manifold M without boundary. Our input is a triangulation K of M with complexity N , and two simple homotopic polygonal curves P and Q sharing endpoints. Edges in P and Q are necessarily edges from the triangulation K . The total complexity of P and Q is n , and there are I number of intersections between them. Note that in this setting, $I = O(n)$. Below we discuss separately the cases when M has non-zero genus and when M is a topological sphere.

5.1 Surfaces with Positive Genus

Given a compact orientable 2-manifold M , let $\mathcal{U}(M)$ be a universal covering space of M with $\phi : \mathcal{U}(M) \rightarrow M$ the associated covering map. Note that ϕ is continuous, surjective, and a local homeomorphism. Informally, being a local homeomorphism means that at any point $x \in \mathcal{U}(M)$, there exists a sufficiently small neighborhood around x such that ϕ is homeomorphism when restricted to this neighborhood. Intuitively, this means that ϕ preserves local (but not necessarily global) structures. (For full details on covering spaces, we refer the reader to topology textbooks that address this area [36, 29]; we will also build on existing algorithmic techniques developed for the computing and working in the universal cover [20, 37].)

For any path γ in M , if we fix the lift (pre-image) of its starting point, then it lifts to a unique path $\tilde{\gamma}$ in $\mathcal{U}(M)$, such that $\phi(\tilde{\gamma}) = \gamma$. Since P and Q are homotopic with common endpoints, the closed curve formed by $C = P \circ \text{rev}(Q)$ is contractible on M , and the lift of C , denoted by $\tilde{C}$, is a closed curve in $\mathcal{U}(M)$. The Homotopy Lifting Property of the universal cover (see for example [29], Proposition 1.30) says that after fixing a lift of an initial curve in a homotopy, the rest of the homotopy lifts uniquely; this immediately gives the following:

Lemma 5.1 *Once we fix the lift of the starting point of P and Q in $\mathcal{U}(M)$, there is a one-to-one correspondence between homotopies between P and Q in M and those between $\tilde{P}$ and $\tilde{Q}$ in $\mathcal{U}(M)$.*

We now impose an area measure in $\mathcal{U}(M)$ by lifting the area measure in M ; this can be done via the map ϕ , which is a local homeomorphism. Now the area of a homotopy in M is the same as the area of its lift in $\mathcal{U}(M)$. As such, we can convert the problem of finding an minimal area homotopy in M to finding one in $\mathcal{U}(M)$. Furthermore, for any orientable compact 2-manifold with genus $g > 0$, its universal cover is topologically equivalent to $\mathbb{R}^2$. Intuitively, this means that we can then apply algorithms and results from previous section to the universal covering space.

More specifically, our algorithm proceeds as follows:

Step 1: Compute relevant portion of $\mathcal{U}(M)$. We will construct a portion of a universal covering space $\mathcal{U}(M)$ made from polygonal schema of M [41, 20]. Specifically, we use the algorithm from [20] to construct a reduced polygonal schema T in $O(N)$ time. The universal covering space consists of an infinite number of copies of this polygonal schema glued together appropriately. We call each copy of the polygonal schema in the constructed universal covering space a *tile*.

Recall that the universal covering space $\mathcal{U}(M)$ is homeomorphic to $\mathbb{R}^2$. We fix a lift of the starting endpoint of P and Q in $\mathcal{U}(M)$ and obtain a specific lift $\tilde{P}$ and $\tilde{Q}$ for P and Q respectively. Since P and Q are homotopic, $\tilde{P}$ and $\tilde{Q}$ form a closed curve, denoted by $\tilde{C} = \tilde{P} \circ \text{rev}(\tilde{Q})$. Note that the number of intersection points between $\tilde{P}$ and $\tilde{Q}$ is at most I , as every intersection point in the lift necessarily maps to an intersection point of P and Q under ϕ , but not vice versa.

Consider the arrangement formed by $\text{Arr}(\tilde{C})$ in the planar domain $\mathcal{U}(M)$. We will construct the portion of the universal covering space $U \subseteq \mathcal{U}(M)$ which is the union of tiles that intersect or are contained inside of $\text{Arr}(\tilde{C})$.

From [20], we know that the lifted curve $\tilde{C}$ passes through $O(n)$ tiles in $\mathcal{U}(M)$. However, while the total number of tiles in the interior of $\text{Arr}(\tilde{C})$ is $O(n)$ for the case where $g > 1$, it can be $\Theta(n^2)$ for the case when $g = 1$. Hence we will separate the case for $g = 1$ and $g > 1$, since we wish to avoid the $O(n^2)$ overhead in the genus 1 case.

For the case $g > 1$, we use the algorithm by Dey and Schipper [20] to compute the relevant portion U of the universal covering space in $O(n \log g + N)$ time. The output contains all $O(n)$ copies of the polygonal schema in U , where each tile is represented by a reduced $4g$ -gon without being explicitly filled with triangles from K . However, recall that P and Q are curves which follow edges of the triangulation; in this construction of the polygonal schema tiles, each edge of K can be broken into $O(g)$ pieces. So in the worst case, we must break each edge in P or Q into $O(g)$ pieces, giving a total complexity for $\tilde{P}$ and $\tilde{Q}$ is $O(ng)$ in this representation of U . Once these are known, we can compute the combinatorial structure of the arrangement of $\tilde{C}$ in U , as well as the description of the set of tiles each cell in $\text{Arr}(\tilde{C})$ intersects or contains, in $O(ng + I \log I)$ time.

For the case $g = 1$, the input manifold is a torus, and the canonical polygon schema for it is a rectangle with oriented boundary arcs $aba^{-1}b^{-1}$. Imagine now that we give the *base polygonal schema* T_0 (which is the tile that contains the lift of the starting point of P and Q) a *coordinate* $(0, 0)$; we must now assign a coordinate for every other copy of the polygon (as shown in Figure 5(a)). Specifically, a copy of polygonal schema T has coordinate (i, j) if

Figure 5: (a) A combinatorial view of the universal covering space $\mathcal{U}(M)$. a and b are the generators and we can give each cell a coordinate. (b) The lift of P (solid curve) and the lift of Q (dashed curve). The heavily shaded region are copies of the polygonal schema contained inside cells of $\text{Arr}(\tilde{C})$, and their total number can be easily computed by a scanning algorithm. R_1 is an essential cell; R_2 and R_3 are two non-essential cells.

the closed loops whose lifts start in T_0 and end in T have the same homotopy type as $a^i b^j$. We can obtain the sequence of the rectangles (and their coordinates) that the curve $\tilde{C}$ will pass through in $O(n + N)$ time [20]. Once these coordinates are known, the combinatorial structure of the arrangement of $\tilde{C}$ in U can also be computed in $O(n + I \log I)$ time. Note that in this case, we have avoided explicitly enumerating the set of tiles fully enclosed within $\text{Arr}(\tilde{C})$ (the shaded tiles in Figure 5 (b)), whose number can be $\Theta(n^2)$ instead of $O(n)$ when $g = 1$.

Step 2: Area of cells in $\text{Arr}(\tilde{C})$. In order to perform our algorithm introduced in Section 4 to the lifted curves $\tilde{P}$ and $\tilde{Q}$, in addition to the combinatorial structure of $\text{Arr}(\tilde{C})$, we also need the area of each cell in $\text{Arr}(\tilde{C})$. We first describe how to compute it for the case $g = 1$.

Take any cell X in $\text{Arr}(\tilde{C})$ and assume the boundary of X intersects m copies of polygonal schema. Even though that X may contain $\Theta(m^2)$ copies of (rectangular) tiles in its interior, we do not need to enumerate these interior tiles explicitly to compute their total area.

Indeed, by a scanning algorithm from left to right, we can compute in $O(m)$ time how many tiles are completely contained inside X (heavily-shaded regions in Figure 5 (b)) (note that the coordinates of each tile traversed by the boundary of X are known). Since the area of every polygonal schema is simply the total area of the input triangulation, we can compute the total area of tiles contained inside X in $O(m)$ time.

Now let $\mathbf{R}$ be the collection of tiles that intersect the boundary of X . It remains to compute the total area of $\mathbf{R} \cap X$. Call each region in $T \cap X$ a *sub-cell*, for any tile $T \in \mathbf{R}$. Let G denote the boundary curves of the polygonal schema T . There are two types of sub-cells: the *essential* ones which contain at least one intersection point between $\tilde{P}$ and $\tilde{Q}$ as their vertices, and the *non-essential* ones which have no intersection; see Figure 5 (b) for examples. (Here, we will use P/Q -arc to refer to either a P -arc or a Q -arc). Note that a non-essential cell is bounded by arcs from G alternating with P/Q -arcs from $\tilde{P}$ or $\tilde{Q}$, since there is no intersection of $\tilde{P}$ and $\tilde{Q}$ along the boundary of a non-essential cell.

Figure 6: (a) We overlay all non-essential sub-cells involving P -arcs into one copy of the polygonal schema. (b) An example of the canonical region T_α is shown for arc α (shaded region in the top-right corner). The shaded region in the middle is a sub-cell X which can be computed as $T_{\beta_0} - T_{\beta_1} - T_{\beta_2} - T_{\beta_3}$, where β_i s are the boundary P - and Q -arcs for X . Among these P/Q -arcs, β_0 is the top-most arc in the containment relation.

First let us consider the collection of non-essential sub-cells formed by alternating G -arcs (boundary arcs of a tile) and arcs from $\tilde{P}$ and $\tilde{Q}$, and compute the area of each such non-essential sub-cells. If we plot all the P -arcs within a single tile T , no two P -arcs can intersect in this tile, since P is a simple curve. Imagine that we pick an arbitrary but fixed point on the boundary G of the polygonal schema T as the origin $\mathbf{o}$. Each P -arc α subdivides T into two regions; we let T_α denote the canonical one excluding $\mathbf{o}$. Note that since P is a simple curve, the set of *canonical regions* T_α s for all P -arcs must satisfy the parenthesis property, and these regions and their areas, called *canonical areas*, can be computed in $O(n \log n + N)$ time using a data structure similar to one used in Section 4.3 to compute the area of R_{us} . See Figure 6 for an illustration. Similarly, we can put all Q -arcs within the same tile and compute the canonical regions / areas associated with all Q -arcs in $O(n \log n + N)$ time. Once these areas are known, the area of each non-essential sub-cell can be computed in $O(s)$ time where s is the number of P -arcs and Q -arcs on the boundary of this sub-cell: Specifically, it is the difference between the canonical area of the top-most P/Q -arc and the union of the canonical areas of all other P/Q -arcs on the boundary of this cell. See Figure 6 (b). Hence the areas of all non-essential sub-cells can be computed in $O(n)$ time once all T_α s are known. The total time complexity required here is thus $O(n \log n + N)$.

What remains is to compute the area of all essential sub-cells. Note that there are $O(I)$ essential sub-cells since each contains an intersection between P and Q . Let a PQ -arc to refer to an arc that starts and ends with points on G (the boundary of the polygonal schema T) and consists of alternating P - and Q -arcs. An essential sub-cell is (i) either completely contained within the interior of a polygonal schema; or (ii) its boundary is formed by alternating PQ -arcs and G -arcs. In case (i), it means that the cell X is contained in the interior of this tile T and this essential sub-cell equals X . An example of case (ii) is the region R_1 in Figure 5 (b), whose boundary consists of a PQ -arc and a G -arc. Now

collect all P -arcs and Q -arcs that are involved in the boundary arcs of those essential sub-cells completely contained within the interior of a tile. Plot them within the same tile T and compute their arrangement A as well as the area for each cell in A . This can be done in $O(n \log n + N)$ time. Since A can have only $O(I)$ vertices in the interior of the tile T , A contains $O(I)$ cells. If an essential sub-cell X is completely contained within a polygonal schema, then it is a union of a set of cells from A . We can simply spend $O(I)$ time to go through cells in A , identify those contained in X and return their total area. Hence it takes $O(I^2)$ time to compute the area of all such $O(I)$ essential sub-cells for all cells X in $\text{Arr}(\tilde{C})$. If an essential sub-cell X has G -arcs on its boundary (case (ii) above), then we need a slightly more complicated way to handle it.

Specifically, for all the remaining essential sub-cells, there can be $O(I)$ number of PQ -arcs along their boundaries, denoted by L . We collect all P -arcs and Q -arcs involved in L and plot them in the same tile T and compute their arrangement $\text{Arr}(L)$. Each PQ -arc $\alpha \in L$ divides the tile T into two regions, and we define T_α to be the canonical one that excluding a specific origin $\mathbf{o}$ on G similar to before. T_α consists of a union of cells from the arrangement $\text{Arr}(L)$, and we can compute the area of T_α in $O(I)$ time since $\text{Arr}(L)$ has $O(I)$ cells. Overall, in $O(I^2)$ time, we can compute the area of all T_α s for all PQ -arcs $\alpha \in L$. Now take an essential sub-cell X that has s number of P -, Q -, or PQ -arcs along its boundary, denoted by $\alpha_1, \alpha_2, \dots, \alpha_s$. Let α_1 be the arc (which can be P -, Q - or PQ -arc) whose endpoints along G spans the largest interval. Then, X can be represented as $X = T_{\alpha_1} - \bigcup_{i \in [2, s]} T_{\alpha_i}$, where T_{α_i} is the canonical region defined by an arc α_i . Since the area of all canonical regions are known (for P -arcs or Q -arcs, we have computed their canonical areas before), X 's area can be computed in $O(s)$ time. Computing the area of all remaining essential sub-cells thus takes $O(I^2 + n)$ time.

Putting everything together, the total time needed to compute the area of all cells in $\text{Arr}(\tilde{C})$ is $O(n \log n + N + I^2)$ when $g = 1$. The case when $g > 1$ is similar but much simpler. Indeed, we now can afford to compute all the tiles contained within any cell of $\text{Arr}(\tilde{C})$ explicitly, as their total number is bounded by $O(n)$ [20, 37]. The areas of essential and non-essential sub-cells are computed using the same algorithm as above. The total time complexity is $O(n \log n + N + I^2)$.

Step 3: Putting everything together. With the combinatorial structure of $\text{Arr}(\tilde{C})$ and the area of each cell computed, we now apply the algorithm from Section 4.2 to compute the minimum area homotopy in $O(I^2 \log I + n \log n)$ time in $\mathcal{U}(M)$, which, by Lemma 5.1, gives the minimum area homotopy between P and Q in M in the same time bound. The total time complexity for the entire algorithm is $O(n \log n + I^2 \log I + N)$.

5.2 The Case of the Sphere

We now consider the remaining case where the input has $g = 0$, so M is a (topological) sphere $\mathbf{S}$. All paths on $\mathbf{S}$ are homotopic. The universal cover of a sphere is itself, and hence is compact. However, the previous algorithm in Section 4.2 works for a domain homeomorphic to $\mathbb{R}^2$ and cannot be directly applied.

We observe that the results in Section 3 still hold. However, as the sphere is compact,

Figure 7: Two ways of sweeping a curve on sphere from base point p .

the winding number is not well-defined. For example, see Figure 7, where there are two ways that the curve γ winds around the point p . In the first case, the winding number at p is 0, while in the second case, the winding number is -1 . In order to use a dynamic programming framework as before to compute the minimum area homotopy between P and Q , we need to develop analogs of Lemma 4.1 and 4.3 for curves on the sphere.

To this end, note that if we remove one point, say $\mathbf{z} \in \mathbf{S}$ from the sphere $\mathbf{S}$, then the resulting space $\mathbf{S}_{\mathbf{z}} = \mathbf{S} - \mathbf{z}$ is homeomorphic to $\mathbb{R}^2$, and the concept of the winding number is well defined for $\mathbf{S}_{\mathbf{z}}$. Specifically, $\mathbf{z}$ can be considered as the point of infinity in $\mathbb{R}^2$. The winding number of $x \in \mathbf{S}_{\mathbf{z}}$ w.r.t. C and $\mathbf{z}$, denoted by $\text{wn}(x; \mathbf{z}, C)$ (C omitted when its choice is clear), is simply the summation of signed crossing numbers for any path connecting x to $\mathbf{z}$. As in the planar case, we say that C is consistent w.r.t. $\mathbf{z}$ if $\text{wn}(x; \mathbf{z}, C)$ is either non-negative, or non-positive for all $x \in \mathbf{S}_{\mathbf{z}}$. Similar to before, we define the total winding number w.r.t. a base point $\mathbf{z}$ as $\text{Tw}(C; \mathbf{z}) = \int_{\mathbf{S}_{\mathbf{z}}} \text{wn}(x; \mathbf{z}, C) dx$. Let $\sigma(P, Q; \Omega)$ denote the minimum cost to morph P to Q within domain Ω .

Lemma 5.2 *If there is a minimum area homotopy between P and Q that does not sweep through some point $\mathbf{z} \in \mathbf{S}$, then we have $\sigma(P, Q; \mathbf{S}) = \sigma(P, Q; \mathbf{S}_{\mathbf{z}})$.*

Lemma 5.3 *Suppose H^* is a minimum area homotopy between P and Q with no anchor points. For any cell R in $\text{Arr}(P + Q)$, if H^* sweeps through one point in its interior, then it sweeps through all points in R .*

Proof: Suppose x and y are two points from the interior of R such that H^* sweeps through x , but not y . Connect x with y by any path γ in the interior of R . This path has to intersect the boundary of the region swept by H^* , and let z be one such intersection point on γ . Obviously, there is a local fold in the minimum area homotopy as it sweeps through z ; namely, some intermediate curve will touch z and immediately trace back. Thus the input homotopy H^* cannot be sense-preserving, contradicting Lemma 3.2. Hence H^* sweeps y as well. ■

Lemma 5.4 *Given a closed curve Γ and any two points $\mathbf{z}, \mathbf{w} \in \mathbf{S}$, we have that: $\text{wn}(x; \mathbf{w}) = \text{wn}(x; \mathbf{z}) + \text{wn}(\mathbf{z}; \mathbf{w})$ (all winding numbers are w.r.t the curve Γ). In particular, for any two points $\mathbf{z}_1, \mathbf{z}_2$ from the same cell of $\text{Arr}(\Gamma)$, we have that $\text{wn}(x; \mathbf{z}_1) = \text{wn}(x; \mathbf{z}_2)$ for all $x \neq \mathbf{z}_1, \mathbf{z}_2$.*

Proof: Let $\gamma(a, b)$ be a path connecting point a to b . Note that the concatenation between

$\gamma(x, \mathbf{z})$ and $\gamma(\mathbf{z}, \mathbf{w})$ is a path from x to $\mathbf{w}$. Since $\text{wn}(x; \mathbf{w})$ is simply the summed signed crossing number of any path from x to $\mathbf{w}$ with respect to Γ , the claim follows immediately. ■

The key result we need is the following lemma which characterizes an optimal homotopy with no anchor points:

Lemma 5.5 *If there is a minimum area homotopy H^* of P and Q with no anchor point, then the image of this optimal homotopy cannot cover all points in $\mathbf{S}$.*

Proof: We prove the lemma by induction on the number of intersections between P and Q . When there is no intersection between P and Q (other than the common endpoints), the Jordan curve $P \circ \text{rev}(Q)$ divides the sphere into two connected components, and the minimum area homotopy is the smaller area of the two. The lemma holds for this base case.

Now assume that the lemma holds for P and Q with at most k intersection points. We wish to show the result for the case where P and Q have $k + 1$ intersection points. Since H^* has no anchor points, this homotopy is sense-preserving by Lemma 3.2. Assign an orientation to the closed curve $C = P \circ Q$ so that locally, every point on the curve P will continuously deform to its right during the optimal homotopy. Now pick an *arbitrary* point $\mathbf{z}$ not on P and Q , and compute the winding number for each cell of $\text{Arr}(P + Q)$ w.r.t. $\mathbf{z}$. Take the cell R with the largest winding number. We assume that $\mathbf{z} \notin R$. Suppose this is not the case and that $\mathbf{z} \in R$. Then we show that we can change the choice of $\mathbf{z}$ to make this hold.

Specifically, if $\mathbf{z} \in R$, then the cell R must have winding number 0. Now take the cell R' of $\text{Arr}(P + Q)$ with the smallest winding number, and let $\mathbf{w}$ be a point from R' . Obviously, $\text{wn}(\mathbf{w}; \mathbf{z}) \leq \text{wn}(x; \mathbf{z}) \leq 0$ for any $x \in \mathbf{S}$. Now we consider the winding numbers w.r.t. to $\mathbf{w}$ instead of $\mathbf{z}$. By Lemma 5.4 we have that $\text{wn}(x; \mathbf{w}) = \text{wn}(x; \mathbf{z}) + \text{wn}(\mathbf{z}; \mathbf{w})$. On the other hand, we have that $\text{wn}(\mathbf{z}; \mathbf{w}) = -\text{wn}(\mathbf{w}; \mathbf{z})$. Hence $\text{wn}(\mathbf{z}; \mathbf{w}) \geq \text{wn}(x; \mathbf{z}) \geq 0$ for any $x \in \mathbf{S}$. In other words, for this new choice of point $\mathbf{w}$, we have that R still has the largest winding number and in this case, $\mathbf{w} \notin R$.

Figure 8: (a) Take Ω as the region to the left of $Q[p, q]$. (b) There always exists a bigon Ω' (possibly Ω is no Q -arc intersects $P[p, q]$).

Similar to the proof of Lemma 4.3, the boundary of this cell consists of alternating arcs from P and from Q , and they necessarily have the orientation as shown in Figure 2 (a) (otherwise, one of the neighboring cell if R will have a larger winding number). Choose

the P -arc e_1 that appears earliest along P , with p being its ending endpoint. Let $P[q]$ be the next intersection between P and R . We have that $P[p, q]$ and $\bar{e} = Q[p, q]$ do not intersect each other. The Jordan curve $P[p, q] \circ (\text{rev}(Q[p, q]))$ bounds two regions on the sphere (instead of a bounded one and an unbounded one in the planar case as shown in Figure 2 (b) and (d)). We consider the region that lies to the right of $P[p, q]$ (thus left of $\bar{e}$), and denote it by Ω . See Figure 8 (a). Since R is to the right of $\bar{e}$, $\Omega \cap R = \emptyset$. As P is simple, only Q can potentially intersect the cell Ω . Hence there always exists a bigon Ω' in Ω which is a cell in $\text{Arr}(P + Q)$. See Figure 8 (b). Note that it is possible that $\Omega' = \Omega$. Let $P[a, b]$ denote the P -arc that bounds the bigon Ω' . Let P' be a new curve obtained by replacing $P[a, b]$ with (slightly above) $Q[a, b]$. Since P deforms always to the right in the optimal homotopy, and in the end, $P[a, b]$ needs to deform to some portion of Q (which is not necessarily $Q[a, b]$ though), one can show that there is a minimum area homotopy between P and Q that consists of first sweeping $P[a, b]$ to $Q[a, b]$ through Ω' , and then optimally morph P' to Q . On the other hand, by the induction hypothesis, there is a minimum area homotopy H' from P' to Q that does not sweep some point, say $\mathbf{z}_1$ in $\mathbf{S}$. There are now two cases:

- (i) If $\mathbf{z}_1 \in \mathbf{S} - \Omega'$, then there is a minimum area homotopy from P to Q that does not sweep $\mathbf{z}_1$ as well. The induction step then holds and the claim follows.
- (ii) Otherwise, $\mathbf{z}_1 \in \Omega'$. Consider the cell $R' \in \text{Arr}(P' + Q)$ that contains $\mathbf{z}_1$. Note that $R' \cap (\mathbf{S} - \Omega') \neq \emptyset$, as there is no vertices of $\text{Arr}(P' + Q)$ contained neither on nor inside Ω' . Hence R' must also contain some point, say $\mathbf{z}_2$, that is outside of Ω' . It then follows from Lemma 5.3 that $\mathbf{z}_2$ is not swept either. This leads us back to case (i), and the induction step again holds.

The claim then follows by induction. ■

Given two homotopic paths P and Q from $\mathbf{S}$ sharing common endpoints, Lemma 5.5 and Lemma 5.2 imply that if P can be morphed to Q optimally without anchor points, then there exists some point $\mathbf{z} \in \mathbf{S}$ such that $\sigma(P, Q; \mathbf{S}) = \sigma(P, Q; \mathbf{S}_{\mathbf{z}})$. For this choice of $\mathbf{z}$, it is necessary that the closed curve $P \circ Q$ has consistent winding numbers in $\mathbf{S}_{\mathbf{z}}$. Once this $\mathbf{z}$ is identified, $\sigma(P, Q; \mathbf{S}_{\mathbf{z}})$ is simply the total winding number of $P \circ Q$ w.r.t. $\mathbf{z}$, as suggested by Lemma 4.3, because $\mathbf{S}_{\mathbf{z}}$ is homeomorphic to the plane. Furthermore, by Lemma 5.3, we only need to pick one point from each cell of $\text{Arr}(P + Q)$ to check for the potential $\mathbf{z}$. Specifically, let $\{\mathbf{z}_1, \dots, \mathbf{z}_l\}$ be a set of such *representatives*, where $l = O(I)$. Define $\text{Tw}^*(\gamma) = \min_{\mathbf{z}} \text{Tw}(\gamma; \mathbf{z})$ where $\mathbf{z}$ ranges over any $\mathbf{z}_i$, $i \in [1, l]$ with respect to whom γ has consistent winding numbers. We then have:

Corollary 5.6 *If there is an optimal area homotopy from P to Q without anchor points, then $\sigma(P, Q) = \text{Tw}^*(P \circ \text{rev}(Q))$.*

Hence if we assume that if there is an *optimal* homotopy between P and Q with *no* anchor points, then we can compute $\sigma(P, Q)$. This will form the base case to handle the general case as we discuss below.

5.3 The Algorithm for the Sphere Case

To compute the minimum area homotopy between P and Q , we follow the same dynamic programming framework as before. If there is no anchor point in a minimum area homotopy, then we use the discussion above to compute the optimal homotopy area. Otherwise, we identify the intersection point that serves as next anchor point, and recurse. The main difference lies in the component of computing $\sigma(i, j) := \sigma(P[\mathbf{x}_i, \mathbf{x}_j], Q[\mathbf{x}_i, \mathbf{x}_j])$, assuming that there is an optimal homotopy from $P' = P[\mathbf{x}_i, \mathbf{x}_j]$ to $Q' = Q[\mathbf{x}_i, \mathbf{x}_j]$ with no anchor points. Previously, this is done by checking whether $P' \circ Q'$ has consistent winding numbers. Now, we need to check the same condition but against $l = O(I)$ number of potential representatives $\{\mathbf{z}_1, \dots, \mathbf{z}_l\}$ as the potential point of infinity. This gives a linear-factor blow-up in the time complexity compared to the algorithm for the planar case. However, we show that this linear blow-up can be tamed down and we can again compute *all* $\sigma(r, j)$ s for all r s and all $j > r$ in $O(I^2 \log I)$ time, after $O(n \log n + N)$ pre-processing time.

Dynamic programming framework. Similar to the planar case, let $\mathbf{x}_0, \dots, \mathbf{x}_I$ denote the intersection points between P and Q , ordered by their indices along P , with $\mathbf{x}_0$ and $\mathbf{x}_I$ being the beginning and ending points of P and Q . Let $T(i)$ denote the optimal homotopy area between $P[0, \mathbf{x}_i]$ and $Q[0, \mathbf{x}_i]$, and $C[i, j]$ the closed curve formed by $P[\mathbf{x}_i, \mathbf{x}_j] \circ Q[\mathbf{x}_j, \mathbf{x}_i]$. However, now we say that a pair of indices (i, j) is *valid* as long as $\mathbf{x}_i$ and $\mathbf{x}_j$ have the same order along P and along Q . This is different from the definition of valid pairs of indices as in the planar case.

Specifically, for any closed curve γ , it turns out that γ can always have consistent winding number for *some* choices of the point of infinity $\mathbf{z}$: that is, there always exists $\mathbf{z} \in \mathbf{S}$ such that $\text{wn}(x, \mathbf{z}; \gamma)$ is consistent for all $x \in \mathbf{S}_{\mathbf{z}}$. We call such choices of $\mathbf{z}$ *consistent representatives w.r.t. γ* . Recall we have defined earlier that $\text{Tw}^*(\gamma) := \min_{\mathbf{z}} |\text{Tw}(\gamma; \mathbf{S}_{\mathbf{z}})|$ where $\mathbf{z}$ ranges over all possible choices of consistent representatives w.r.t. γ . By Corollary 5.6, if there is an optimal homotopy between $P[\mathbf{x}_i, \mathbf{x}_j]$ and $Q[\mathbf{x}_i, \mathbf{x}_j]$ with no anchor points, then the optimal homotopy area is $\text{Tw}^*(\gamma)$. However, different from the planar case, $\text{Tw}^*(C[i, j])$ is defined for all pairs of $i < j$ where $\mathbf{x}_i$ and $\mathbf{x}_j$ have the same order along P and Q . We now define a pair (i, j) to be *valid* if $i < j$ and $\mathbf{x}_i$ and $\mathbf{x}_j$ have the same order along P and Q . (Note that in the planar case, there is one more condition for a pair being valid, that is, $C[i, j]$ needs to have consistent winding numbers in the planar case as well.) We also remark that in general $\text{Tw}^*(C[i, j])$ may not be the optimal homotopy area for $P[\mathbf{x}_i, \mathbf{x}_j]$ and $Q[\mathbf{x}_i, \mathbf{x}_j]$. We now have the following recurrence:

$$T(i) = \begin{cases} 0, & \text{if } i = 0 \\ \min_{k < i \text{ and } (k, i) \text{ is valid}} \{ \text{Tw}^*(C[k, i]) + T(k) \}, & \text{otherwise} \end{cases}$$

As before, the final goal is to compute $T(I) = \sigma(P, Q)$.

Computing Tw^* s. Here we describe how to compute $\text{Tw}^*(C[i, j])$ efficiently. Specifically, we show how to compute all $\text{Tw}^*(C[r, j])$ s for all $j > r$ in $O(I)$ time, for any fixed r , after

$O(n \log n + I \log I + N) = O(n \log n + N)$ preprocessing.

First, similar to the planar case, let us choose the representatives by taking four points around each intersection points $\mathbf{x}_i$ between P and Q as shown in the right figure. Consider only those representatives to the right of Q (which are those above Q in the right figure), and denote them by $Z = \{\mathbf{z}_1, \dots, \mathbf{z}_I\}$. Z is sorted by their indices along Q . Those to the left of it can be handled similarly. The first observation is that for any two consecutive representatives, $\text{wn}(x; \mathbf{z}_i) - \text{wn}(x; \mathbf{z}_{i+1})$ is 1 or -1 , depending on the orientation of P -arc that separating them.

Now to compute which $\mathbf{z}_i$ will give consistent winding numbers, we first compute the winding number of each cell in $\text{Arr}(P + Q)$ for $\mathbf{z}_1$. Next, take the cells R_1 and R_2 with minimum and maximum winding numbers, and assume that $\mathbf{z}_{i_1}$ and $\mathbf{z}_{i_2}$ are their representatives. If there is more than one cell with largest (or smallest) winding numbers, just pick an arbitrary one. By Lemma 5.4, the closed curve $C[r, j] := P[\mathbf{x}_r, \mathbf{x}_j] \circ Q[\mathbf{x}_r, \mathbf{x}_j]$ has all non-negative winding number w.r.t $\mathbf{z}_{i_1}$, and all non-positive winding number w.r.t. $\mathbf{z}_{i_2}$. Hence we simply compute the total winding number $\text{Tw}(C[r, j]; \mathbf{z}_{i_1})$ and $\text{Tw}(C[r, j]; \mathbf{z}_{i_2})$, and return the one with the smaller absolute value as $\text{Tw}^*(C[r, j])$. We refer to the indices i_1 and i_2 as the *wn-min* and *wn-max* indices, respectively, and these two total winding numbers as *valid total winding numbers*. Basically, by Lemma 5.7 below, the smaller of the absolute values of the two valid total winding numbers is simply the minimum cost to deform $P[\mathbf{x}_r, \mathbf{x}_j]$ to $Q[\mathbf{x}_r, \mathbf{x}_j]$ without using anchor points. This improves the time complexity of computing each $\text{Tw}^*(C[r, j])$ to $O(I)$ time, instead of the naive $O(In)$ time by computing all $\text{Tw}(C[r, j]; \mathbf{z}_i)$ s, for $i \in [1, I]$, from scratch.

Lemma 5.7 *Given an arbitrary oriented (not necessarily simple) curve $C = P' \circ \text{rev}(Q')$ on $\mathbf{S}$, let $\text{Arr}(C)$ be the arrangement of C , and $Z = \{z_1, \dots, z_k\}$ a set of representative points from each cell in C . Pick an arbitrary point, say z_1 , and compute the winding number of C w.r.t. each z_i . Let i_1 and i_2 be the *wn-min* and *wn-max* indices. Then the minimum cost to deform P' to Q' with no anchor point is $\min\{|\text{Tw}(C; z_{i_1})|, |\text{Tw}(C; z_{i_2})|\}$.*

Proof: First, call a point z *valid* if $\text{wn}(x; z, C)$ is consistent for all $x \in \mathbf{S}$. The optimal cost to deform P' to Q' with no anchor point is $\min_{\text{valid } z \in Z} |\text{Tw}(C; z)|$. For any base point z , note that by Lemma 5.4, we have $\text{wn}(x; z) = \text{wn}(x; z_1) - \text{wn}(z; z_1)$. In order for the winding number to be consistent, we need that either $\text{wn}(x; z) \geq 0$ for any $x \in \mathbf{S}$, or $\text{wn}(x; z) \leq 0$ for any $x \in \mathbf{S}$. Assume it is the former case. Then $\text{wn}(z; z_1) \leq \text{wn}(x; z_1)$ for all $x \in \mathbf{S}$, implying that $\text{wn}(z; z_1) = \text{wn}(z_{i_1}; z_1)$. Furthermore, note that

$$\begin{aligned} \text{Tw}(C; z) &= \int_{\mathbf{S}} \text{wn}(x; z) dx = \int_{\mathbf{S}} [\text{wn}(x; z_1) - \text{wn}(z; z_1)] dx \\ &= \int_{\mathbf{S}} [\text{wn}(x; z_1) - \text{wn}(z_{i_1}; z_1)] dx = \text{Tw}(C; z_{i_1}). \end{aligned}$$

If it is the latter case, then $\text{wn}(z; z_1) \geq \text{wn}(x; z_1)$ for all $x \in \mathbf{S}$, implying that $\text{wn}(z; z_1) =$

$\text{wn}(z_{i_2}; z_1)$. In this case we have that

$$\begin{aligned}\text{Tw}(C; z) &= \int_{\mathbf{S}} \text{wn}(x; z) dx = \int_{\mathbf{S}} [\text{wn}(x; z_1) - \text{wn}(z; z_1)] dx \\ &= \int_{\mathbf{S}} [\text{wn}(x; z_1) - \text{wn}(z_{i_2}; z_1)] dx = \text{Tw}(C; z_{i_2}).\end{aligned}$$

The optimal cost $\min_{\text{valid } z \in Z} |\text{Tw}(C; z)|$ is thus achieved by the smaller one of the absolute value of $\text{Tw}(C; z_{i_1})$ and $\text{Tw}(C; z_{i_2})$. ■

To further improve the time complexity, we will start with $C[r, r+1]$, and update the winding number information in each cell as well as the valid total winding numbers, as we traverse P and pass through each intersection point $\mathbf{x}_i$. To this end, we use the same range tree data structure as in Section 4.2. Specifically, we use this data structure to maintain the winding number information w.r.t a fixed based point $\mathbf{z}_1$. The wn-min and wn-max indices can be easily maintained by storing at each internal node the minimum and maximum winding number within its subtree. We can also maintain the total winding number w.r.t. the base point $\mathbf{z}_1$. The time complexity for updates remains the same as before (i.e, $O(\log I)$ time per update).

The remaining task is to compute the valid total winding numbers as i increases. Let A denote the total area of topological sphere $\mathbf{S}$. First, observe that for a fixed curve C , by Lemma 5.4, we have $\text{wn}(x; \mathbf{z}, C) = \text{wn}(x; \mathbf{z}_1, C) - \text{wn}(z; \mathbf{z}_1, C)$ with respect the fixed based point $\mathbf{z}_1$. Hence we have that:

$$\begin{aligned}\text{Tw}(C; z) &= \int_{\mathbf{S}} \text{wn}(x; z, C) dx = \int_{\mathbf{S}} \text{wn}(x; \mathbf{z}_1, C) dx - A \cdot \text{wn}(z; \mathbf{z}_1, C) \\ &= \text{Tw}(C; \mathbf{z}_1) - A \cdot \text{wn}(z; \mathbf{z}_1, C).\end{aligned}\tag{2}$$

Assume that i_1 is the wn-min index and i_2 is the wn-max index. Hence we can compute $\text{Tw}(C; \mathbf{z}_{i_1})$ and $\text{Tw}(C; \mathbf{z}_{i_2})$ in $O(1)$ time using Eqn (2), since $\text{Tw}(C; \mathbf{z}_1)$, $\text{wn}(\mathbf{z}_{i_1}; \mathbf{z}_1, C)$ and $\text{wn}(\mathbf{z}_{i_2}; \mathbf{z}_1, C)$ are all maintained as C changes from $C[r, u]$ to $C[r, u+1]$.

Putting everything together, with $O(n \log n + N)$ pre-processing time, we can compute all $\sigma(r, j)$ s for all $j > r$ for any fixed r , in $O(I \log I)$ time, and thus computing all $\text{Tw}^*(C[r, u])$ s for all $r \in [1, I]$ and all $u < r$, in $O(I^2 \log I)$ total time. Putting everything together, the dynamic programming problem can be solved in $O(n \log n + I^2 \log I + N)$ total time.

Now combining both cases ($g > 0$ and $g = 0$) together, we conclude with the following main result.

Theorem 5.8 *Given a triangulation K of an orientable compact 2-manifold M with genus g , let N be the complexity of K . Given two homotopic paths P and Q of total complexity n with I intersection points, we can compute an optimal homotopy and its area $\sigma(P, Q; M)$ in $O(I^2 \log I + n g \log n + N)$ time for $g > 0$, and in $O(I^2 \log I + n \log n + N)$ for $g = 0$.*

6 Minimum Area Homotopies between Cycles

In this section, we extend our algorithms to work for calculating the area between two cycles in the plane or on the sphere, both of which are fairly simple extensions of our algorithms for curves. We conclude by discussing a few observations for the case of cycles on a surface, which remains open.

6.1 Cycles in the Plane

We now consider the case where we have two simple cycles P and Q in the plane. We have the following characterization:

Lemma 6.1 *If the two simple cycles P and Q intersect, then there is an anchor point in the optimal homotopy between them.*

Proof: Suppose that P and Q intersect but there is no anchor point in the optimal homotopy H^* . By Lemma 3.2, we know that H^* must be sense preserving. However, this means that H^* continually moves one curve to the other in one local direction, which means that one curve must be entirely contained within the other, contradicting the assumption that they intersect. ■

At this point, the algorithm for cycles which intersect each other reduces to the one for curves: if we know which intersection point between P and Q is the anchor point, we can simply “break” the cycles into two curves at this point; this will become the start and end point for each of the two curves. Hence our algorithm for cycles will take a multiplicative factor of $O(I)$ extra time than the algorithm for curves, since we need to try each possible intersection point as the required anchor point.

It then follows from Theorem 4.4 that:

Corollary 6.2 *Given two polygonal cycles P and Q in the plane of n total complexity and with $I > 0$ intersection points, we can compute the optimal area homotopy in $O(I(I^2 \log I + n \log n))$ time if $I > 0$.*

The remaining case is that when the two polygonal cycles P and Q are disjoint. If one of the cycles contains the other, then the area of the optimal homotopy is simply the area sandwiched between these two simple cycles. This can be computed in $O(n \log n)$ time easily.

However, if the cycles are disjoint but neither contains the other, then under our definition there is no “optimal” (regular) homotopy between them to achieve the minimum homotopy area while keeps intermediate curves regular. Specifically, the “optimal” free homotopy between them is by first collapsing each curve separately to a point and then deform the points to each other. The associated homotopy area Σ is the sum of the areas bounded by the two closed curves. Indeed, this is the smallest possible by a similar argument as the one in the proof of Lemma 4.2. However, this homotopy is not regular as it collapses each curve to a point. On the other hand, there is a sequence of regular homotopies whose

areas converge to Σ . The minimum area can be achieved in the limit by a sequence of regular homotopies, but there is no corresponding optimal regular homotopy to realize it.

Connection with immersed disks. We remark that one can actually show that if there is a minimum area homotopy between P and Q without anchor points, then $P \circ Q$ is the so-called *interior boundary* [40], which is a notion slightly more relaxed than saying that $P \circ \text{rev}(Q)$ is the boundary of an *immersed disk*, known also as the *self-overlapping curves* in discrete settings [38]. It is known that one can check whether an arbitrary closed polygonal chain C with n edges is a self-overlapping curve in $O(n^3)$ time [38]. Lemma 4.3 suggests that in our case, where C can be decomposed into two simple curves, it can be done in $O(n)$ time by checking winding numbers. (Lemma 4.3 does not hold for arbitrary boundary of immersed disk.)

6.2 Cycles on Spheres

This is very similar to the planar case. In particular, if the minimum area homotopy has an anchor point, then we can break our cycles into two curves that share common start and end points, and then use the algorithm from Section 5.2. If there is no anchor point, then similar to Lemma 6.1, due to the sense-preserving property of the optimal homotopy, it is necessary that the two simple cycles P and Q are disjoint. Note that in this setting, there is always a regular homotopy between the two cycles, as by Jordan Curve Theorem, one of the cycle is always contained within one of the bounded region as produced by the other cycle. The area of this sandwiched region between P and Q may not be optimal, however. The optimal homotopy area can be achieved by either this sandwiched area, or by the sum of the areas of the smaller region bounded by P and by Q respectively. Again, in the planar case, the optimal area is achieved in the limit, without an optimal regular homotopy to realize it.

6.3 Cycles on Surfaces with Non-zero Genus

For more general surfaces, we first note that if the cycles are disjoint and do not bound an annulus, no optimal homotopy exists under our definition, just as in the planar case. (See Section 6.1 for the discussion on this.) This means we only need to consider cycles which are disjoint and bound an annulus, or cycles which intersect at least once and may or may not contain an anchor point in the optimal homotopy.

If the optimal homotopy has an anchor point, then we can break our cycles into curves that share a common start and end point, and use the algorithm from the previous section. However, on a surface the analog of Lemma 6.1 no longer holds. If the two curves are noncontractible, it is now possible that there is no anchor point but that the two cycles do intersect in the manifold M , unlike in the planar case. For example, in the figure to the right, we see a torus with two cycles (one red and one blue); in the small portion of the

universal cover shown, we see a lift of the red curve and two of the blue, where the red curve does not intersect one of the blue lifts at all. In other words, certain lifts of the two cycles do not intersect in the universal covering space, although P and Q intersect in M . In order to deal with all of the possibilities, we must consider infinitely possible lifts of the two cycles.

We conjecture that the homotopy area can still be calculated in polynomial time, since we can fix a lift of one of the curves and then rule out lifts of the second which are far away. However, we are unable to bound this as a function of n , and hence the case for cycles on surfaces remains open.

7 Conclusion

In this paper, we propose a new curve similarity measure which captures how hard it is to deform from one curve to the other based on the amount of total area swept. It is robust to noise (as it is area-based), and can be computed efficiently. Two immediate next questions are (i) Can improve the dependency on I for the planar case, where $I = \Theta(n^2)$ in the worst case. Can we improve the time complexity to $O(n^2 \text{polylog} n)$ instead? (ii) Can we improve the time complexity of our algorithm for cycles to remove the extra linear or quadratic factor that we currently need?

Currently, we assume that two input paths are simple paths which share starting and ending points, which makes it easier to define homotopy equivalence. This leads to two natural questions, namely how to handle curves which do not share endpoints and how to deal with non-simple curves. Another interesting problem is to compute optimal *isotopy area* where we require that any intermediate curve during the deformation is also simple.

Measuring similarity of curves on surfaces is an interesting problem, and many open areas remain. Geodesic Fréchet-based measures ignore the topological constraints of underlying surface, while the homotopic Fréchet distance, homotopy height, and our method require identification of a homotopy which optimizes some cost. Future directions include developing efficient curve simplification algorithms based on this measure, and studying similarity between curves from more general simplicial complexes than considered in this paper (such as a manifold with boundary or holes, or non-manifolds).

Acknowledgment. The authors would like to thank Joseph O’Rourke, Rephael Wenger, Michael Davis, and Tadeusz Januszkiewicz for useful discussions at the early stage of this work, and David Letscher, Brody Johnson, and Bryan Clair for helpful discussions at the later stage of this paper. We would also like to thank the anonymous reviewers for their comments. This work is partially supported by the National Science Foundation under grants CCF-0747082, CCF-1054779, and IIS-1319573.

References

- [1] P. K. Agarwal, R. B. Avraham, H. Kaplan, and M. Sharir. Computing the discrete Fréchet distance in subquadratic time. In *Proc. 24th Ann. ACM-SIAM Sympos. Discrete*

- Alg. (SODA)*, 2013.
- [2] H. Alt. The computational geometry of comparing shapes. In *Efficient Algorithms*, volume 5760 of *Lecture Notes in Computer Science*, pages 235–248. Springer Berlin / Heidelberg, 2009.
 - [3] H. Alt, U. Fuchs, G. Rote, and G. Weber. Matching convex shapes with respect to the symmetric difference. *Algorithmica*, 21:89–103, 1998.
 - [4] H. Alt and M. Godau. Computing the Fréchet distance between two polygonal curves. *International Journal of Computational Geometry and its Applications*, 5:75–91, 1995.
 - [5] H. Alt and L. J. Guibas. Discrete geometric shapes: Matching, interpolation, and approximation. In J.-R. Sack and J. Urrutia, editors, *Handbook of Computational Geometry*, pages 121–153. Elsevier Science Publishers B. V. North-Holland, Amsterdam, 2000.
 - [6] H. Alt, C. Knauer, and C. Wenk. Comparison of distance measures for planar curves. *Algorithmica*, 38(1):45–58, 2004.
 - [7] B. Aronov, S. Har-Peled, C. Knauer, Y. Wang, and C. Wenk. Fréchet distance for curves, Revisited. In *Proc. of European Symposium on Algorithms*, pages 52–63, 2006.
 - [8] S. Bespamyatnikh. An optimal morphing between polylines. *International Journal of Computational Geometry and Applications*, 12(3):217–228, 2002.
 - [9] P. Bose, S. Cabello, O. Cheong, J. Gudmundsson, M. van Kreveld, and B. Speckmann. Area-preserving approximations of polygonal paths. *Journal of Discrete Algorithms*, 4:554–566, 2006.
 - [10] G. R. Brightwell and P. Winkler. Submodular percolation. *SIAM J. Discret. Math.*, 23(3):1149–1178, July 2009.
 - [11] K. Buchin, M. Buchin, C. Knauer, G. Rote, and C. Wenk. How difficult is it to walk the dog? In *Proc. 23rd Europ. Workshop Comput. Geom.*, pages 170–173, 2007.
 - [12] K. Buchin, M. Buchin, W. Meulemans, and W. Mulzer. Four Soviets walk the dog—with an application to Alt’s conjecture. *arXiv/1209.4403*, 2012.
 - [13] K. Buchin, M. Buchin, and Y. Wang. Partial curve matching under the fréchet distance. In *Proc. ACM-SIAM Sympos. Discrete Alg. (SODA)*, 2009.
 - [14] M. Buchin. *On the Computability of the Fréchet Distance Between Triangulated Surfaces*. PhD thesis, Dept. of Comput. Sci., Freie Universität Berlin, 2007.
 - [15] E. W. Chambers, É. Colin de Verdière, J. Erickson, S. Lazard, F. Lazarus, and S. Thite. Homotopic Fréchet distance between curves or, walking your dog in the woods in polynomial time. *Computational Geometry*, 43(3):295 – 311, 2010. Special Issue on 24th Annual Symposium on Computational Geometry (SoCG’08).

- [16] E. W. Chambers, T. Ju, D. Letscher, and L. Liu. Isotopic fréchet distance. In *CCCG11*, 2011.
- [17] E. W. Chambers and D. Letscher. On the height of a homotopy. In *CCCG'09*, pages 103–106, 2009.
- [18] A. F. Cook and C. Wenk. Geodesic Fréchet distance inside a simple polygon. In *Proc. 25th Internat. Sympos. Theoret. Asp. Comp. Sci.*, pages 193–204, 2008.
- [19] R. G. Cromley. *Digital Cartography*. Prentice Hall, Englewood Cliffs, NJ, 1992.
- [20] T. K. Dey and H. Schipper. A new technique to compute polygonal schema for 2-manifolds with application to null-homotopy detection. *Discrete and Computational Geometry*, 14(1):93–110, 1995.
- [21] J. Douglas. Solution of the problem of plateau. *Trans. of the American Mathematical Society*, 33:263–321, 1931.
- [22] A. Driemel and S. Har-Peled. Jaywalking your dog: computing the fréchet distance with shortcuts. In *Proc. 23rd Annual ACM-SIAM Symposium on Discrete Algorithms, SODA '12*, pages 318–337. SIAM, 2012.
- [23] N. Dym, A. Shtengel, and Y. Lipman. Homotopic morphing of planar curves. *Computer Graphics Forum*, 34(5):239–251, 2015.
- [24] A. Efrat, Q. Fan, and S. Venkatasubramanian. Curve matching, time warping, and light fields, new algorithms for computing similarity between curves. *J. Math. Imaging Vis.*, 27(3):203–216, 2007.
- [25] A. Efrat, S. Har-Peled, L. J. Guibas, J. S. Mitchell, and T. Murali. New similarity measures between polylines with applications to morphing and polygon sweeping. *Discrete and Computational Geometry*, 28:535–569, 2002.
- [26] B. Fasy, S. Karakoç, and C. Wenk. On minimum area homotopies. In *Computational Geometry: Young Researchers Forum*, pages 37–40, 2016.
- [27] S. Har-Peled, A. Nayyeri, M. Salavatipour, and A. Sidiropoulos. How to walk your dog in the mountains with no magic leash. In *Proc. 28th Symposium on Computational Geometry, SoCG '12*, pages 121–130, New York, NY, USA, 2012. ACM.
- [28] S. Har-Peled and B. Raichel. The fréchet distance revisited and extended. In *Proc. 27th Annual ACM symposium on Computational geometry, SoCG '11*, pages 448–457, New York, NY, USA, 2011. ACM.
- [29] A. Hatcher. *Algebraic Topology*. Cambridge University Press Press, 2002.
- [30] H. Lawson. *Lectures on minimal submanifolds*, volume 1 of *Mathematics lecture series*. Publish or Perish, 1980.
- [31] A. Maheshwari and J. Yi. On computing Fréchet distance of two paths on a convex polyhedron. In *European Workshop on Computational Geometry*, pages 41–44, 2005.

- [32] R. B. McMaster and K. S. Shea. *Generalization in Digital Cartography*. Association of American Cartographers, Washington DC, 1992.
- [33] K. Mehlhorn and C.-K. Yap. Constructive Whitney-Graustein Theorem: Or how to untangle closed planar curves. *SIAM J. Comput.*, 20(4):603–621, 1991.
- [34] Z. Nie. On the minimum area of null homotopies of curves traced twice. <http://arxiv.org/pdf/1412.0101v2.pdf>.
- [35] T. Rado. On plateau’s problem. *Annals of Mathematics*, 31:457–469, 1930.
- [36] J. J. Rotman. *An Introduction to Algebraic Topology*. Graduate Texts in Mathematics; 119. Springer-Verlag New York Inc., 1988.
- [37] H. Schipper. Determining contractibility of curves. In *Proc. Sympos. on Computational Geometry*, pages 358–367, 1992.
- [38] P. W. Shor and C. J. V. Wyk. Detecting and decomposing self-overlapping curves. *Comput. Geom.*, 2:31–50, 1992.
- [39] S. Smale. Regular curves on riemannian manifolds. *Transactions of the American Mathematical Society*, 87(2):492–512, 1958.
- [40] C. Titus. The combinatorial topology of analytic functions of the boundary of a disk. *Acta Mathematica*, 106(1-2):45–64, 1961.
- [41] G. Vegter and C. K. Yap. Computational complexity of combinatorial surfaces. In *Proc. Sympos. on Computational Geometry*, pages 102–111, 1990.
- [42] R. C. Veltkamp. Shape matching: similarity measures and algorithms. In *Proc. Shape Modeling International*, pages 188–199, 2001.
- [43] H. Whitney. On regular closed curves in the plane. *Compositio Mathematica*, 4:276–284, 1937.