Computing the Geometric Intersection Number of Curves*

Vincent Despré[†] Francis Lazarus[‡]

November 28, 2019

Abstract

The geometric intersection number of a curve on a surface is the minimal number of self-intersections of any homotopic curve, i.e. of any curve obtained by continuous deformation. Given a curve c represented by a closed walk of length at most ℓ on a combinatorial surface of complexity n we describe simple algorithms to (1) compute the geometric intersection number of c in $O(n + \ell^2)$ time, (2) construct a curve homotopic to c that realizes this geometric intersection number in $O(n + \ell^4)$ time, (3) decide if the geometric intersection number of c is zero, i.e. if c is homotopic to a simple curve, in $O(n + \ell \log \ell)$ time. The algorithms for (2) and (3) are restricted to orientable surfaces, but the algorithm for (1) is also valid on non-orientable surfaces.

To our knowledge, no exact complexity analysis had yet appeared on those problems. An optimistic analysis of the complexity of the published algorithms for problems (1) and (3) gives at best a $O(n + g^2\ell^2)$ time complexity on a genus g surface without boundary. No polynomial time algorithm was known for problem (2) for surfaces without boundary. Interestingly, our solution to problem (3) provides a quasi-linear algorithm to a problem raised by Poincaré more than a century ago. Finally, we note that our algorithm for problem (1) extends to computing the geometric intersection number of two curves of length at most ℓ in $O(n + \ell^2)$ time.

*This work was supported by the LabEx PERSYVAL-Lab ANR-11-LABX-0025-01.

[†]Équipe INRIA-GAMBLE, Université de Nancy, France, Vincent.Despre@inria.fr

[‡]GIPSA-Lab, CNRS, Grenoble, France, Francis.Lazarus@grenoble-inp.fr

Contents

1 Introduction 3

2 Historical notes 5

3 Our strategy for counting intersections 7

4 Combinatorial framework 9
- 4.1 Combinatorial curves on surfaces . . . 10
- 4.2 Systems of quads . . . 11
- 4.3 Geodesics . . . 14

5 Computing intersection numbers on oriented surfaces 16
- 5.1 Crossing double-paths . . . 18
- 5.2 Proof of Theorem 1 in the primitive and orientable case . . . 19

6 Computing intersection numbers: the general case 22
- 6.1 Thick double-paths . . . 23
- 6.2 Proof of Theorem 1: the primitive case . . . 25
- 6.3 Non-primitive curves . . . 26
- 6.4 End of proof of Theorem 1 . . . 27

7 Combinatorial perturbations 29

8 Computing a minimal immersion 31
- 8.1 Bigons and monogons . . . 32
- 8.2 Proof of Theorem 2 . . . 34

9 Detecting simple curves 35
- 9.1 Some preparatory lemmas for Theorem 3 . . . 35
- 9.2 The unzip algorithm . . . 38

10 Concluding remarks 48

A Proof of the direct implication of Proposition 28 53

B Proof of Lemma 29 55

1 Introduction

Let S be a surface. Two closed curves $\alpha, \beta : \mathbb{R}/\mathbb{Z} \to S$ are **freely homotopic**, written $\alpha \sim \beta$, if there exists a continuous map $h : [0,1] \times \mathbb{R}/\mathbb{Z} \to S$ such that $h(0,t) = \alpha(t)$ and $h(1,t) = \beta(t)$ for all $t \in \mathbb{R}/\mathbb{Z}$. The curves α and β being homotopic or not, their number of intersections is

$$|\alpha \cap \beta| = |\{(t,t') \mid t,t' \in \mathbb{R}/\mathbb{Z} \text{ and } \alpha(t) = \beta(t')\}|.$$

Their **geometric intersection number** only depends on their free homotopy classes and is defined as

$$i(\alpha,\beta) = \min_{\alpha' \sim \alpha, \beta' \sim \beta} |\alpha' \cap \beta'|$$

This minimum is finite and obtained with curves that intersect transversally. Likewise, the number of self-intersections of α is given by

$$\frac{1}{2}|\{(t,t') \mid t \neq t' \in \mathbb{R}/\mathbb{Z} \text{ and } \alpha(t) = \alpha(t')\}|,$$

and its minimum over all the curves freely homotopic to α is its **geometric self-intersection number** $i(\alpha)$. Note the one half factor that comes from the identification of (t,t') with (t',t).

The geometric intersection number is an important parameter that allows to stratify the set of homotopy classes of curves on a surface. The surface is usually endowed with a hyperbolic metric, implying that each homotopy class is identified by its unique geodesic representative. Extending a former result by Mirzakhani [Mir08], Sapir [Sap15, Mir16] has recently provided upper and lower bounds for the number of closed geodesics with bounded length and bounded geometric intersection number. Chas and Lalley [CL12] also proved that the distribution of the geometric intersection number with respect to the word length approaches the Gaussian distribution as the length grows to infinity. Other more experimental results were obtained with the help of a computer to show the existence of length-equivalent homotopy classes with distinct geometric intersection numbers [Cha14]. Hence, for both theoretical and practical reasons, various aspects of the computation of geometric intersection numbers have been studied in the past including the algorithmic ones. Nonetheless, all the previous approaches rely on rather complex mathematical arguments and to our knowledge no exact complexity analysis has yet appeared. In this paper, we make our own the words of Dehn who noted that the metric on words (on some basis of the fundamental group of the surface) can advantageously replace the hyperbolic metric [DLH10]. We propose a combinatorial framework that leads to simple algorithms of low complexity to compute the geometric intersection number of curves or to test if this number is zero. Our approach is based on the computation of canonical forms as recently introduced for the purpose of testing whether two curves are homotopic [LR12, EW13]. Canonical forms are instances of combinatorial geodesics who share nice properties with the geodesics of a hyperbolic surface. On hyperbolic surfaces each homotopy class contains a unique geodesic that moreover minimizes the number of self-intersections. Although a combinatorial geodesic is generally not unique in its homotopy class, it must stay at distance one from its canonical representative and a careful analysis of its structure leads to the first result of the paper.

Theorem 1. *Given two curves represented by closed walks of length at most ℓ on a combinatorial surface of complexity n we can compute the geometric intersection number of each curve or of the two curves in $O(n + \ell^2)$ time.*

As usual the complexity of a combinatorial surface stands for its total number of vertices, edges and faces. A key point in our algorithm is the ability to compute the primitive root of a canonical curve c in linear time. This is a curve r that is not homotopic to a proper power of any other curve and such that $c \sim r^k$ for some integer k. We next provide an algorithm to compute an actual curve immersion – its combinatorial description is part of our combinatorial framework – that minimizes the number of self-intersections in its homotopy class. While the combinatorial surface in Theorem 1 may be non-orientable, the next two results only apply to orientable surfaces.

Theorem 2. *Given a curve c represented by a closed walk of length ℓ on an orientable combinatorial surface of complexity n we can compute a combinatorial immersion with $i(c)$ crossings in $O(n + \ell^4)$ time.*

We also propose a nearly optimal algorithm that answers an old problem studied by Poincaré [Poi04, §4]: decide if the geometric intersection number of a curve is null, that is if the curve is homotopic to a simple curve.

Theorem 3. *Given a curve represented by a closed walk of length ℓ on an orientable combinatorial surface of complexity n we can decide if the curve is homotopic to a simple curve in $O(n + \ell \log \ell)$ time*[1]*. In the affirmative we can construct an embedding of c in the same amount of time.*

We emphasize that our results represent significant progress with respect to the state of the art. No precise analysis appeared in the previously proposed algorithms [BS84, Chi69, Chi72, CL87, Lus87, dGS97, Pat02, GKZ05] concerning Theorems 1 or 3. An optimistic analysis of what seems the most efficient approach [Lus87, Th. 3.7], although particularly complex, gives at best a quadratic time complexity for computing the geometric intersection number on an orientable genus g surface without boundary, assuming that the curves are primitive and expressed as words in a canonical basis of the fundamental group. Note that these assumptions are not required in Theorems 1, 2 or 3. Schaefer et al. [SSS08] propose an efficient computation of the geometric intersection number of a set of curves represented by normal coordinates in a triangulated surface. However, their approach is limited to *simple* input curves. Apart from a recent algorithm by Aretinnes [Are15], which is restricted to surfaces with *nonempty* boundary, we know of no polynomial time algorithm for Theorem 2. Finally, Theorem 3 states the first quasi-linear algorithm for detecting homotopy classes of simple curves since the problem was raised by Poincaré more than a century ago [Poi04, §4]. A related problem was tackled by Chang *et al.* [CEX15, Th. 8.2] (see also [AAET17] for the planar case) who describe an algorithm to decide if a given closed path on a combinatorial surface is *weakly simple*, i.e. admits an infinitesimal perturbation that is an embedding. Note that a path may go several times through the same vertex or edge and still have infinitesimal perturbations that are simple. However,

[1] In the preliminary version of this work, as it appeared in the proceddings of the 33rd International Symposium on Computational Geometry (SoCG 2017), the announced complexity had an extra $\log \ell$ factor.

the algorithm in [CEX15] does not always detect if the path is homotopic to a simple curve since it is only authorized an infinitesimal perturbation.

In the next section we review some of the previous relevant works. Section 3 presents our general simple strategy to compute the geometric intersection number. This strategy will be applied in a combinatorial framework introduced in Section 4. The proof of Theorem 1 restricted to primitive curves on orientable surfaces is then given in Section 5. The general case for curves that may be non-primitive on possibly non-orientable surfaces is treated in Section 6. We remark that the number of crossings of two given combinatorial curves is not always well-defined as parts of curves may overlap along shared edges. We resolve this ambiguity with the help of combinatorial perturbations as presented in Section 7. The computation of a minimally crossing perturbation is then presented in Section 8 with the proof of Theorem 2. We finally propose a simple algorithm to detect and embed curves homotopic to simple curves (Theorem 3) in Section 9.

2 Historical notes

In the fifth supplement to its *Analysis situs* Poincaré [Poi04, §4] describes a method to decide whether a given closed curve γ on a surface can be continuously deformed to a simple curve. For this purpose, he considers the surface as the quotient $\mathbb{D}/\mathit{\Gamma}$ of the Poincaré disk $\mathbb{D}$ by a (Fuchsian) group $\mathit{\Gamma}$ of hyperbolic transformations. The endpoints of a lift of γ in the Poincaré disk are related by a hyperbolic transformation whose axis is a hyperbolic line L representing the unique geodesic homotopic to γ. He concludes that γ is homotopic to a simple curve if and only if all the transforms of L by $\mathit{\Gamma}$ are pairwise disjoint or equal. This method was turned into an algorithm by Reinhart [Rei62] who worked out the explicit computations in the Poincaré disk using the usual representation of hyperbolic transformations by two-by-two matrices. The entries of the matrices being algebraic, the computation could indeed be performed accurately on a computer. The ability to recognize curves that are **primitive**, i.e. whose homotopy class cannot be expressed as a proper power of another class, happens to be crucial in this algorithm though computationally expensive.

Birman and Series [BS84] subsequently proposed an algorithm for the case of surfaces with nonempty boundary that avoids manipulating algebraic numbers. While their arguments appeal to a hyperbolic structure, their algorithm is purely combinatorial. Intuitively, a surface with boundary deform retracts onto a fat graph (in fact a fat bouquet of circles) whose universal covering space embeds as a fat tree in the Poincaré disk. The successive lifts of a curve γ trace a bi-infinite path in this tree. The limit points of this path belongs to the circle $\partial\mathbb{D}$ at infinity and coincide with the ideal endpoints of the axis of the hyperbolic transformation corresponding to any of the lifts of the curve in the path. The question as to whether two lifts give rise to intersecting axes can thus be reduced to test if the corresponding bi-infinite paths have separating limit points on $\partial\mathbb{D}$. As for Reinhart, Birman and Series assume that the homotopy class of γ is given by a word W on some given set of generators of the fundamental group of the surface. In turn, the above test on bi-infinite paths boils down to consider the cyclic permutations of W and W^{-1} in a cyclic lexicographic order and to check if this ordering is well-parenthesized with respect to some pairing of the words.

Cohen and Lustig [CL87] further observed that the approach of Birman and Series on surfaces with boundary could be extended to count the geometric intersection number of curves. In a second paper Lustig [Lus87] tackles the case where the curves are taken on a surface without boundary. Like Poincaré he considers a closed surface with negative Euler characteristic as the quotient of the Poincaré disk by a group of transformations isomorphic to the fundamental group of the surface. The main contribution of the paper is to define a canonical representative for every free homotopy class α given as a word in some fixed system of loops generating the fundamental group. Lustig first notes that there is no obvious way of choosing a canonical form among the words representing α. In particular, the shortest words are far from unique. He rather represents (a lift of) α by a path in the union $G \cup N \cup H$ of three tessellations of $\mathbb{D}$, where the edges of G are all the lifts of the generating loops, N is the dual tessellation, and the edges of H joins the vertices of G with the vertices of N. (Although the graphs of G, N and H are embedded their union is not as every edge of G crosses its dual edge in N.) Lustig gives a purely combinatorial characterization of canonical paths and argues that the method in his first paper [CL87] can be applied to the canonical representative of α. Overall the two papers add up to 60 pages with essential arguments from hyperbolic geometry and the complexity analysis of the whole procedure remains to be done.

Other approaches were developed without assuming any hyperbolic structure. Based on the notion of winding number, Chillingsworth [Chi69, Chi72] provides an algorithm to test whether a curve is homotopic (this time with fixed basepoint) to a simple curve on a surface with nonempty boundary. He also proposed an algorithm for determining when a given set of *simple* closed curves can be made disjoint by (free) homotopy [Chi71]. While the winding number relies on a differentiable structure of the surface, Zieschang [Zie65, Zie69] appeals to the topological structure only. He used the connection between the automorphisms of a topological surface and the automorphisms of its fundamental group in order to detect the homotopy classes of simple closed curves. We also mention some works by Schaefer et al. [SSS08] to compute the geometric intersection number of two *simple* curves given by their normal coordinates in a triangulated surface. Their algorithm is based on repeated applications of Dehn twists and is claimed to have polynomial time complexity.

A related work by Hass and Scott [HS85] is concerned with curves that have excess intersection, i.e. that can be homotoped so as to reduce their number of intersections. Hass and Scott introduce various types of monogons and bigons that can be either embedded, singular or weak. A *singular monogon* of a curve $\gamma : \mathbb{R}/\mathbb{Z} \to S$ is a contractible subpath of γ whose endpoints define a self-intersection of γ. A *bigon* of γ is defined by two self-intersections joined by two homotopic subpaths (with fixed endpoints) $\gamma|_\sigma$ and $\gamma|_{\sigma'}$ with $\sigma, \sigma' \subset \mathbb{R}/\mathbb{Z}$. The bigon is said *singular* if the defining segments σ, σ' are disjoint and *weak* otherwise. Hass and Scott prove that a curve with excess self-intersection on an orientable surface must have a singular monogon or a singular bigon. Their result directly suggests an algorithm to compute the geometric intersection number of a curve: iteratively remove monogons or untie bigons until there are no more. The final configuration must have the minimal number of self-intersections. Designing an efficient procedure to find monogons and bigons remains the crux of this approach. Arettines [Are15] has proposed an algorithm based on this approach to compute a minimal configuration of a single curve on an orientable surface with *nonempty* boundary. Note that the method cannot be extended to compute the geometric intersection number of a curve on a non-orientable surface since Hass and Scott give a counter-example of a curve with excess self-intersections but with no singular monogon or bigon. The method also cannot be extended to compute the geometric intersection number of two curves on an orientable surface. Indeed, Hass and Scott give two counter-examples to the fact that two curves with excess intersections should have a singular bigon. (A singular bigon between two curves is a pair of homotopic subpaths, one of each curve.) One of their counter-examples contains a curve with excess self-intersections and the other one contains a non-primitive curve. Our counter-example, Figure 1, shows that even assuming each curve to be primitive and in minimal configuration, we may have excess intersections without singular bigons. Nonetheless, it was proved [HS94, dGS97, Pat02] that starting from any configuration of curves one may reach a configuration with a minimal number of intersections by applying a finite sequence of elementary moves involving monogons, bigons and trigons similar to the Reidemeister moves in knot theory. A surprising consequence was obtained by Neumann-Coto [NC01]. Define a *cut and paste* on a family of curves by cutting the curves at some of their intersection points and glueing the resulting arcs in a different order. Neumann-Coto proves that any set of primitive curves can be brought to a homotopic set with minimal (self-)intersections by a set of cut and paste operations. Note that each intersection of two curve pieces can be re-arranged in three different manners, including the original one, by a cut and paste. Hence, if the curves have I (self-)intersections we may find a minimal configuration out of the 3^I possible re-configurations!

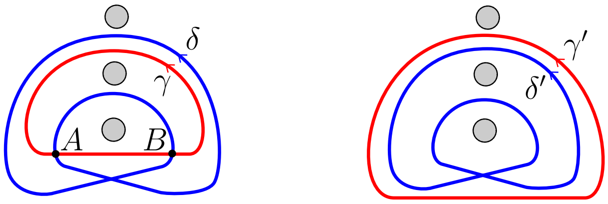

Figure 1: The plain circles represent non-contractible curves. The two curves γ and δ on the left have homotopic disjoint curves γ' and δ'. The curves γ and δ thus have excess intersection although there is no singular bigon between the two. If $A = \gamma(0) = \delta(0)$ and $B = \gamma(u) = \delta(v)$ we nonetheless have $\delta|_{[0,v]} \sim \gamma|_{[0,1+u]}$ where $\gamma|_{[0,1+u]}$ is the concatenation of γ with $\gamma|_{[0,u]}$. In particular, $\gamma|_{[0,1+u]}$ wraps more that once around γ.

We also mention the algebraic approach of Gonçalves et al. [GKZ05] based on previous works by Turaev [Tur79] who introduced intersection forms over the integral group ring of the fundamental group of the surface. This approach is based on the transformation of each curve into a certain algebraic sum of homotopy classes. A simple analysis of the time complexity for the computation of this sum leads to $O(\ell^5)$ operations for a curve with ℓ self-intersections. See [CL87, GKZ05] for more historical notes.

3 Our strategy for counting intersections

Here, we assume some familiarity with basic hyperbolic geometry in the Poincaré disk and with the notion of (universal) covering of a surface. We refer the reader to Stillwell [Sti93], especially Chapter 1 and 6, for a gentle introduction. Following Poincaré's original approach we represent a surface S with negative Euler characteristic as the hyperbolic quotient surface $\mathbb{D}/\mathit{\Gamma}$ where $\mathit{\Gamma}$ is a discrete group of hyperbolic motions of the Poincaré disk $\mathbb{D}$. We denote by $p : \mathbb{D} \to \mathbb{D}/\mathit{\Gamma} = S$ the universal covering map. Any closed curve $\alpha : \mathbb{R}/\mathbb{Z} \to S$ gives rise to its infinite power $\alpha^\infty : \mathbb{R} \to \mathbb{R}/\mathbb{Z} \to S$ that wraps around α infinitely many times. A **lift** of α^∞ is any curve $\tilde{\alpha} : \mathbb{R} \to \mathbb{D}$ such that $p \circ \tilde{\alpha} = \alpha^\infty$ where the parameter of $\tilde{\alpha}$ is defined up to an integer translation (we thus identify the curves $t \mapsto \tilde{\alpha}(t + k)$, $k \in \mathbb{Z}$). **For conciseness, we shall write from now on "a lift of α" where we actually mean a lift of α^∞.** Note that $p^{-1}(\alpha)$ is the union of all the images $\mathit{\Gamma} \cdot \tilde{\alpha}$ of $\tilde{\alpha}$ by the motions in $\mathit{\Gamma}$. The curve $\tilde{\alpha}$ has two limit points on the boundary of $\mathbb{D}$ which can be joined by a unique hyperbolic line L. The projection $p(L)$ wraps infinitely many times around the unique geodesic homotopic to α. In particular, the set of pairs of limit points of all lifts of α only depends on the homotopy class of α.

No two motions of $\mathit{\Gamma}$ have a limit point in common unless they are powers of the same motion. This can be used to show that when α is primitive, its lifts are uniquely identified by their limit points [FM12]. Let α and β be two primitive curves. We fix a lift $\tilde{\alpha}$ of α and denote by $\tau \in \mathit{\Gamma}$ the hyperbolic motion sending $\tilde{\alpha}(0)$ to $\tilde{\alpha}(1)$. Note that τ leaves $\tilde{\alpha}$ globally invariant. Let $\mathit{\Gamma} \cdot \tilde{\beta}$ be the set of lifts of β. We consider the subset

$$B = \{\tilde{\beta}' \in \mathit{\Gamma} \cdot \tilde{\beta} \mid \text{ the limit points of } \tilde{\beta}' \text{ and } \tilde{\alpha} \text{ alternate along } \partial\mathbb{D}\},$$

and we denote by B/τ the set of equivalence classes of lifts generated by the relations $\tilde{\beta}' \approx \tau(\tilde{\beta}')$. In other words, two lifts are equivalent if one is the image of the other by some power of τ.

Lemma 4 ([Rei62])**.** *$i(\alpha, \beta) = |B/\tau|$.*

Proof. Put $I(\alpha, \beta) = \{(t, t') \mid t, t' \in \mathbb{R}/\mathbb{Z} \text{ and } \alpha(t) = \beta(t') \text{ and if } \alpha = \beta : t \neq t'\}$. Define a map $\varphi : I(\alpha, \beta) \to (\mathit{\Gamma} \cdot \tilde{\beta})/\tau$ as follows. Given $(u \bmod 1, v \bmod 1) \in I(\alpha, \beta)$ there is, by the unique lifting property of coverings, a unique lift $\tilde{\beta}'$ of β that satisfies $\tilde{\beta}'(v) = \tilde{\alpha}(u)$. We set $\varphi(u \bmod 1, v \bmod 1)$ to the class of this lift. Note that changing u to $u + k$, $k \in \mathbb{Z}$, leads to the lift $\tau^k(\tilde{\beta}')$, so that φ is well-defined. We have $B/\tau \subset \varphi(I(\alpha, \beta))$. Indeed, if $\tilde{\beta}' \in B$ then $\tilde{\beta}'$ and $\tilde{\alpha}$ must intersect at some point $\tilde{\beta}'(v) = \tilde{\alpha}(u)$. It follows that $\varphi(u \bmod 1, v \bmod 1)$ is the class of $\tilde{\beta}'$. As an immediate consequence,

$$|I(\alpha, \beta)| \geq |B/\tau|$$

When α and β are geodesics all their lifts are hyperbolic lines and φ is a bijection onto its image B/τ. We conclude that $|I(\alpha, \beta)|$ is minimized among all homotopic curves, so that $i(\alpha, \beta) = |B/\tau|$. □

When α and β are hyperbolic geodesics, their lifts being hyperbolic lines have alternating limit points exactly when they have a non-empty intersection and they have a unique intersection point in that case. This point projects to a crossing of α and β that actually identifies the corresponding element of B/τ. When α and β are not geodesic the situation is more ambiguous and their lifts may have multiple intersection points. Those intersection points project to crossings of α and β, so that the elements of B/τ are now identified with subsets of crossing points (with odd cardinality) rather than single crossing points. The induced partition is generated by the following relation: two crossings are equivalent if they are connected by a pair of homotopic subpaths of α and β, namely one of

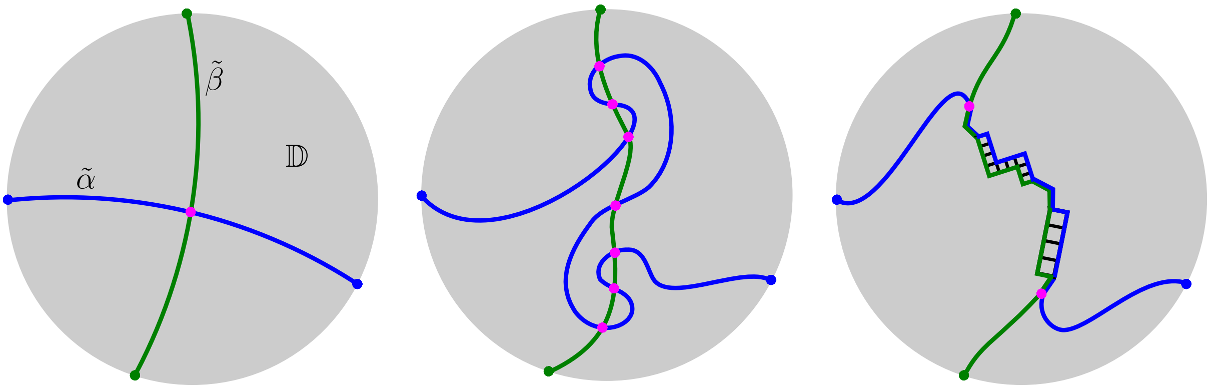

Figure 2: Left, two intersecting hyperbolic lines corresponding to lifts of geodesic curves, say α and β. Middle, lifts of non-geodesic curves homotopic to α and β may intersect several times. Right, lifts of combinatorial geodesics.

the two subpaths of α and one of the two subpaths of β cut by the two crossings. Indeed, if the two crossings are projections of intersections of two lifts, then the paths between those intersections in each lift project to homotopic paths. Conversely, homotopic paths lift to paths with common endpoints that can be seen as subpaths of two intersecting lifts. In order to compute the above partition, we thus essentially need an efficient procedure for testing if two paths are homotopic. This homotopy test can be performed in linear time according to Theorem 8. Since a combinatorial curve of length ℓ may have $O(\ell^2)$ crossings, we directly obtain an algorithm with time complexity $O(\ell^5)$ to compute the above partition.

In practice, we shall work in a combinatorial framework as presented in the next section. In this framework, curves are identified with paths in the 1-skeleton of a combinatorial surface. It appears more convenient to assume that all the faces of the combinatorial surface are quadrilaterals. We thus need to reduce the given surface to a system of quads[2] as described in Section 4.2. Each curve may now be replaced by a homotopic combinatorial geodesic in such a system of quads. The situation becomes somehow intermediate between the ideal hyperbolic case and the most general situation; the unique intersection point of hyperbolic lines is replaced by an elongated pair of homotopic paths. See Figure 2. By taking advantage of the specific structure of combinatorial geodesics in the system of quads we can indeed avoid computing the above partition and directly identify B/τ with certain pairs of homotopic subpaths of the combinatorial geodesics. See Proposition 18. This leads to a more efficient algorithm with quadratic complexity.

4 Combinatorial framework

In order to study the computational aspects of the intersection number we need to specify the encoding of curves on surfaces. A common representation of the homotopy class of a curve on a surface consists of a word in some set of generators of the fundamental group of the surface. Typically, when the fundamental group is given by a combinatorial presentation with a single relation, one can build a two dimensional complex that is

[2] Some topologists would call it a *square-tiled surface*.

homeomorphic to the surface and such that the word for the curve corresponds to a homotopic closed path in this complex. We shall use a more general notion of combinatorial surfaces where curves a represented by closed walks. We introduce below this combinatorial framework. In order to mimic as closely as possible hyperbolic geometry we further restrict our framework to system of quads which are combinatorial surfaces composed only of quadrilaterals. We finally study the structure of geodesics in such systems.

4.1 Combinatorial curves on surfaces

Combinatorial surfaces. As usual in computational topology, we model a surface by a cellular embedding of a graph G in a compact topological surface S. Such a cellular embedding can be encoded by a **combinatorial surface** composed of the graph G itself together with a rotation system [MT01] that records for every vertex of the graph the **direct** cyclic order of the incident arcs. When S is orientable, this order can be chosen consistently for all vertices. We call this order clockwise. Nonorientable surfaces can be encoded by providing an additional signature with every edge that indicates if the direct orientations at its endpoints are consistent or not. The edge is said **half-twisted** when the orientations are inconsistent. The **facial walks** are obtained from the rotation system by the face traversal procedure as described in [MT01, p.93]. In order to handle surfaces with boundaries we allow every face of G in S to be either an open disk or an annulus (open on one side). In other words G is a cellular embedding in the closure of S obtained by attaching a disk to every boundary of S. We record this information by storing a boolean for every facial walk of G indicating whether the associated face is perforated or not. All the considered graphs may have loops and multiple edges. A directed edge will be called an **arc** and each edge corresponds to two opposite arcs. We denote by a^{-1} the arc opposite to an arc a. Every combinatorial surface Σ can be reduced by first contracting the edges of a spanning tree and then deleting edges incident to distinct plain, i.e. non-perforated, faces. The resulting **reduced surface** has a single vertex. A reduced surface without boundary has a single face, but in general a reduced surface may have several faces. The combinatorial surface Σ and its reduced version encode different cellular embeddings on a same topological surface.

Combinatorial curves. Consider a combinatorial surface with its graph G. A combinatorial curve (or path) c is a walk in G, i.e. an alternating sequence of vertices and arcs, starting and ending with a vertex, such that each vertex in the sequence is the target vertex of the previous arc and the source vertex of the next arc. We generally omit the vertices in the sequence. A combinatorial curve is closed when additionally the first and last vertex are equal. When no confusion is possible we shall drop the adjective combinatorial. A closed curve is **2-sided** if its number of half-twisted arcs is even. It is otherwise **1-sided**. The **length** of c is its total number of arc occurrences, which we denote by $|c|$. If c is closed, we write $c(i)$, $i \in \mathbb{Z}/|c|\mathbb{Z}$, for the vertex of index i of c and $c[i, i+1]$ for the arc joining $c(i)$ to $c(i+1)$. A **double point** of c is a pair of indices $(i, j) \in \mathbb{Z}/|c|\mathbb{Z} \times \mathbb{Z}/|c|\mathbb{Z}$ such that $i \neq j$ and $c(i) = c(j)$. Likewise, given another closed curve d, we define a double point of (c, d) as a pair $(i, j) \in \mathbb{Z}/|c|\mathbb{Z} \times \mathbb{Z}/|d|\mathbb{Z}$ with $c(i) = d(j)$.

For convenience we set $c[i+1, i] = c[i, i+1]^{-1}$ to allow the traversal of c in reverse direction. In order to differentiate the arcs from their occurrences we denote by $[i, i \pm 1]_c$ the corresponding occurrence of the arc $c[i, i \pm 1]$ in $c^{\pm 1}$, where c^{-1} is obtained by traversing $c^1 := c$ in the opposite direction. More generally, for any non-negative integer ℓ and any sign $\varepsilon \in \{-1, 1\}$, The sequence of indices

$$(i, i + \varepsilon, i + 2\varepsilon, \dots, i + \varepsilon\ell)$$

is called an **index path** of c of length ℓ. The index path can be **forward** ($\varepsilon = 1$) or **backward** ($\varepsilon = -1$) and can be longer than c so that an index may appear more than once in the sequence. We denote this path by $[i \xrightarrow{\varepsilon\ell}]_c$. Its **image path** is given by the arc sequence

$$c[i \xrightarrow{\varepsilon\ell}] := (c[i, i + \varepsilon], c[i + \varepsilon, i + 2\varepsilon], \dots, c[\varepsilon(\ell - 1), \varepsilon\ell]).$$

The image path of a length zero index path is just a vertex. There is an evident notion of inclusion between index paths: a forward index path $[i \xrightarrow{\ell}]_c$ is included in a forward index path $[j \xrightarrow{k}]_c$ if there exists a natural number s such that $s + \ell \leq k$ and $i = j + s$. This inclusion relationship extends to both forward and backward paths, replacing each backward index path $[i \xrightarrow{-\ell}]_c$ by $[i - \ell \xrightarrow{\ell}]_c$ to perform the comparison.

A closed curve is **contractible** if it is homotopic to a trivial curve (i.e., a curve reduced to a single vertex). We will implicitly assume that a homotopy has fixed endpoints when applied to paths and is free when applied to closed curves. If α, β are two closed paths, we shall sometimes write $\alpha \overset{*}{\sim} \beta$ to denote homotopy with fixed basepoint as opposed to $\alpha \sim \beta$ for the free homotopy relation.

4.2 Systems of quads

Reduction to a system of quads. Let Σ be a combinatorial surface with negative Euler characteristic. Following Lazarus and Rivaud [LR12] we describe the reduction of Σ to a standard quadrangulation, called a **system of quads** by Erickson and Whittlesey [EW13]. We first consider the case of a surface without boundary, i.e., when Σ has no perforated faces. After reducing Σ to a surface Σ' with a single vertex v and a single face f this system of quads is obtained by adding a vertex w at the center of f, adding edges between w and all occurrences of v in the facial walk of f, and finally deleting the edges of Σ'. See left and middle Figure 3. The graph of the resulting system of quads, called the **radial graph** [LR12], is bipartite. It contains two vertices, namely v and w, and $4g$ edges (resp. $2g$ edges) if the surface is orientable (resp. non-orientable), where g is the genus of Σ. All its faces are quadrilaterals. A similar construction applies when Σ has perforated faces. In this case, the reduced surface Σ' may have several plain faces, each surrounded by perforated faces. We essentially perform the same sequence of operations to each plain face. However, when deleting edges of Σ' incident to both a perforated and a plain face, we merge the two faces to obtain a larger perforated face[3]. In the end, the plain faces are quadrilaterals and every vertex is incident to a perforated face. In order to get a bipartite graph, we eventually subdivide the remaining loop edges by introducing a vertex in the middle.

[3] In general, we cannot afford merging the plain faces with the adjacent perforated faces in a systematic manner to get a surface with perforated faces only. Indeed, when transforming a path of Σ into a homotopic path of this reduced version this may increase the complexity of the path by a factor of g.

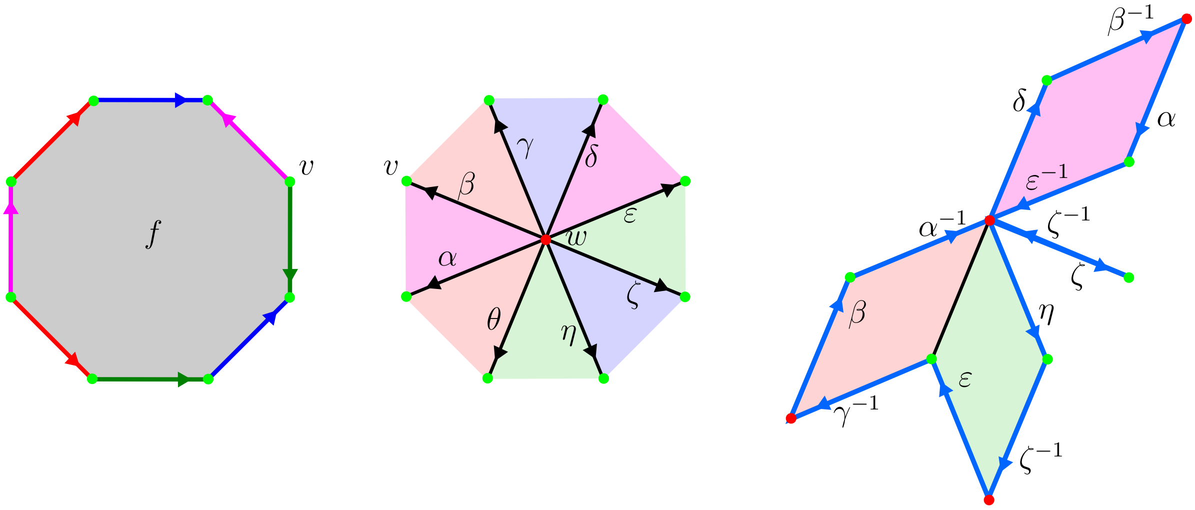

Figure 3: Left, a planar unfolding of a reduced orientable surface of genus two after cutting along its edges. Middle, the corresponding system of quads with its radial graph. Each face is a quadrilateral obtained by gluing two triangular faces along a deleted edge of the reduced surface. Right, a disk diagram for the contractible curve $(\beta, \alpha^{-1}, \delta, \beta^{-1}, \alpha, \varepsilon^{-1}, \zeta, \zeta^{-1}, \eta, \zeta^{-1}, \varepsilon, \gamma^{-1}, \beta)$. The unbounded face corresponds to the perforated face in Δ.

Note that the system of quads can be deduced from Σ by a sequence of edge contractions, deletions, insertions and subdivisions. Every cycle c of Σ can thus be modified accordingly to give a cycle $\varphi(c)$ in the system of quads. Clearly, there exist cellular embeddings ρ, ρ' of the graphs of Σ and of the system of quads in a same topological surface such that $\rho(c) \sim \rho'(\varphi(c))$.

Lemma 5 ([DG99, LR12]). *Let n be the number of edges of the combinatorial surface Σ. The above construction of a system of quads can be performed in $O(n)$ time so that for every closed curve c of length ℓ in Σ, we can compute in $O(\ell)$ time a curve c' of length at most 2ℓ in the system of quads such that $c' \sim \varphi(c)$.*

For the rest of the paper, unless stated otherwise, we shall assume that **all surfaces have negative Euler characteristic**.

Diagrams. A **disk diagram** over the combinatorial surface Σ is a combinatorial sphere Δ with one perforated face together with a labelling of the arcs of Δ by the arcs of Σ such that

1. opposite arcs receive opposite labels,

2. the facial walk of each plain face of Δ is labelled by the facial walk of some plain face of Σ.

The diagram is **reduced** when no edge of Δ is incident to two plain faces labelled by the same facial walk (with opposite orientations) of Σ. An **annular diagram** is defined similarly by a combinatorial sphere with two distinct perforated faces. A vertex of a diagram that is not incident to any perforated face is said **interior**.

Lemma 6 (van Kampen, See [EW13, Sec. 2.4])**.** *A cycle of Σ is contractible if and only if it is the label of the facial walk of the perforated face of a reduced disk diagram over Σ. Two cycles are freely homotopic if and only if the facial walks of the two perforated faces of a reduced annular diagram over Σ are labelled by these two cycles respectively.*

Figure 3 (right) shows a disk diagram for a contractible curve. Note that two plain faces that are adjacent and consistently oriented in a reduced diagram must be labelled by adjacent faces that are consistently oriented in Σ. Moreover, the degree of an interior vertex of the diagram is a multiple of the degree of the corresponding vertex in Σ. In the sequel, **all the considered diagrams will be supposed reduced**.

Spurs, brackets and canonical curves. Thanks to Lemma 5 we may assume that our combinatorial surface Σ is a system of quads. Moreover, the construction of this system of quads with the assumption on the Euler characteristic implies that **all interior vertices have degree at least 6**. For surfaces without boundary, this follows from the fact that the two vertices of Σ have degree $4g$ in the orientable case and $2g$ otherwise. For surfaces with non-empty boundary, the claim is trivial since the system of quads has no interior vertices. Let (a_1, a_2) be a pair of arcs sharing their origin vertex v and such that the face corners between a_1 and a_2 in the direct cyclic order around v belong to plain faces. Following the terminology of Erickson and Whittlesey [EW13], we define the **turn** of (a_1, a_2) as the number of those face corners. Hence, if v is an interior vertex of degree d in Σ, the turn of (a_1, a_2) is an integer modulo d that is zero when $a_1 = a_2$. If v is a boundary vertex and one of the faces between a_1 and a_2 (in the direct order) is perforated, while none of the faces between a_2 and a_1 is perforated, then the turn of (a_1, a_2) is minus the number of face corners between a_2 and a_1. The turn is undefined when one of the faces between a_1 and a_2 and one of the faces between a_2 and a_1 are perforated. When Σ is oriented, the **turn sequence** of a subpath $(a_i, a_{i+1}, \ldots, a_{i+j-1})$ of a closed curve of length ℓ is the sequence of $j+1$ turns of $(a_{i+k}^{-1}, a_{i+k+1})$ for $-1 \leq k < j$, where indices are taken modulo ℓ. The subpath may have length ℓ, thus leading to a sequence of $\ell+1$ turns. On a non-necessarily oriented surface the sign of the turn of $(a_{i+k}^{-1}, a_{i+k+1})$ should be changed according to the parity of the number of half-twisted arcs among $(a_i, a_{i+1}, \ldots, a_{i+k})$. This amounts to untwisting the path $(a_i, a_{i+1}, \ldots, a_{i+k})$ before computing the turn sequence. A **spur** in a curve is a subpath of the form (a, a^{-1}). Hence, the turn of $(a_{i+k}^{-1}, a_{i+k+1})$ is zero precisely when (a_{i+k}, a_{i+k+1}) is a spur. A **bracket** is any subpath whose turn sequence has the form 12^*1 or $\bar{1}\bar{2}^*\bar{1}$ where t^* stands for a possibly empty sequence of turns t and $\bar{x}$ stands for $-x$.

The next theorem follows from Lemma 6 and a combinatorial version of Gauss-Bonnet theorem [GS90].

Theorem 7 ([GS90, EW13])**.** *A nontrivial contractible closed curve on a system of quads must have either a spur or four brackets. Moreover, if the curve is the label of the boundary walk (i.e., of the facial walk of the perforated face) of a disk diagram with at least one interior vertex, then the curve must have either a spur or five brackets.*

Lazarus and Rivaud [LR12] have introduced a canonical form for every nontrivial free homotopy class of closed curves in an oriented system of quads. In particular, two curves are freely homotopic if and only if their canonical forms are equal (up to a circular shift of their vertex indices). This canonical form is defined as the rightmost generator of a certain cylindrical covering surface of the system of quads. It was further characterized by Erickson and Whittlesey [EW13] in terms of turns and brackets. The **canonical form** (of the free homotopy class) of a curve is the homotopic curve that contains no spurs or brackets and whose turning sequence contains no -1's and contains at least one turn that is not -2.

Theorem 8 ([LR12],[EW13])**.** *Up to a circular shift of its vertex indices, the canonical form of a combinatorial closed curve of length ℓ on an oriented system of quads is uniquely defined. It can be computed in $O(\ell)$ time.*

4.3 Geodesics

A **combinatorial geodesic** of length ℓ is a curve that contains no spurs or brackets, except maybe a bracket of length $\ell - 1$ if the curve is *exceptional*. See the notion of exceptional curve after Corollary 11. Given a curve on a system of quads, orientable or not, we can compute a homotopic geodesic by iteratively removing spurs and brackets as much as possible. The resulting curve are said *reduced* in [EW13]. The proof therein applies to orientable systems of quads but trivially extends to the non-orientable case.

Theorem 9 ([EW13, Lem. 4.3])**.** *Given a combinatorial closed curve of length ℓ on a system of quads, a homotopic geodesic can be computed in $O(\ell)$ time.*

On an oriented surface, the canonical form is an instance of combinatorial geodesic; this is the rightmost homotopic geodesic. The definitions of a geodesic and of a canonical form extend naturally to paths: a **geodesic path** is a path that contains no spurs or brackets. It is **canonical** (or in canonical form) if its turning sequence contains no -1's. Theorem 8 extends trivially to paths: every path has a unique homotopic geodesic in canonical form.

Although we cannot claim in general the uniqueness of geodesics in a homotopy class, homotopic geodesics are almost equal and have the same minimal length. Specifically, define a **(quad) staircase** as a planar sequence of quads obtained by stitching an alternating sequence of rows and columns of quads to get the shape of a staircase. Assuming that the staircase goes up from left to right, we define the **initial tip** of a quad staircase as the lower left vertex of the first quad in the sequence. The **final tip** is defined as the upper right vertex of the last quad. See Figure 4. A **closed staircase** is obtained by identifying the two vertical arcs incident to the initial and final tips of a staircase.

Theorem 10. *Let c, d be two nontrivial homotopic combinatorial geodesics. If c, d are closed curves, then they label the two boundary cycles of an annular diagram composed of a single closed staircase or of an alternating sequence of paths (possibly reduced to a vertex) and quad staircases connected through their tips. Likewise, if c, d are paths, then the closed curve $c \cdot d^{-1}$ labels the boundary of a disk diagram composed of an alternating sequence of paths (possibly reduced to a vertex) and quad staircases connected through their tips.*

Proof. We only detail the proof when c, d are paths. See Figure 4. The similar case of closed curves is covered in [EW13]. By Lemma 6, $c \cdot d^{-1}$ is the label of the facial walk

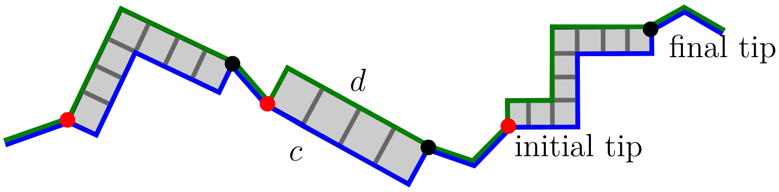

Figure 4: A disk diagram for two homotopic paths c and d composed of paths and staircases.

of the perforated face of a disk diagram Δ. This diagram has a cactus-like structure composed of 2-cells subdivided into quads and connected by trees. A vertex v such that $\Delta \setminus v$ is not connected is called a *cut vertex*. We also consider as cut vertices the endpoints of c in Δ. A 2-cell of Δ must have more than one cut vertex on its boundary. Otherwise, this boundary is entirely labelled by a subpath of either c or d and Theorem 7 implies the existence of four brackets, one of which (in fact two) must avoid the cut vertex, hence be contained in the interior of this subpath. This would contradict that c and d are geodesic. Moreover, because c and d have no spur, Δ cannot have more than two degree one vertices, namely the common endpoints of c and d. It follows that Δ is an alternating sequence of paths and 2-cells. In particular, each 2-cell has exactly two cut vertices. No 2-cell in this sequence has an interior vertex. For otherwise, by the second part of Theorem 7, the boundary of this 2-cell would contain five brackets one of which would be contained in the interior of either c or d and this would again contradict that c and d are geodesic. It follows that the dual of a 2-cell, viewed as an assembling of quads, is a tree. We finally remark that this tree must be a path with a staircase shape. Indeed, any other shape would imply the existence of a bracket in either c or d. □

Since the tips of a staircase divide its boundary into two sides of equal length, and since removing a bracket or a spur shortens a curve, we easily deduce the following.

Corollary 11. *With the hypothesis of Theorem 10, c and d have equal length which is minimal among homotopic curves.*

It will often be useful to think of a disk diagram bounded by homotopic geodesic paths as a subset of the universal cover $\tilde{\Sigma} = \mathbb{D}$ of the system of quads. Indeed, by the remark after Lemma 6, the sequence of faces along a staircase is labelled by a sequence of faces arranged the same way in $\tilde{\Sigma}$. Beware, however, that two successive staircases in a diagram may correspond to opposite orientations in $\tilde{\Sigma}$. Some staircases should thus be flipped in order to get an exact image of the diagram in $\tilde{\Sigma}$. A consequence of this representation is that homotopic geodesic paths have a unique disk diagram. This follows directly from the unique lifting property of coverings.

Exceptional curves A geodesic curve of length ℓ is **exceptional** if it has a bracket of length $\ell - 1$. An exceptional curve must be 1-sided, since otherwise the curve would have a bracket of length 1. Figure 5 shows the lift of an exceptional curve; it is contained in an infinite straight strip of quadrilaterals that covers a Möbius strip infinitely many times.

Corollary 12. *A geodesic has no nontrivial index path whose image path is contractible.*

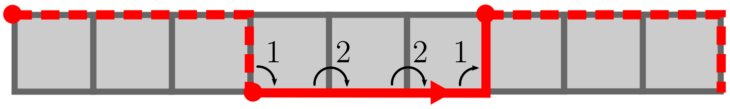

Figure 5: The lift of an exceptional curve of length 4 has brackets of length 3.

Proof. Let c be a geodesic. If c is not exceptional, it follows from the definition of geodesics that the image path of any index path of c has no spur or brackets, hence is also a (non-contractible) geodesic. If c is exceptional, the possible brackets of its index paths can be flattened to get a homotopic non-trivial geodesic. □

The next two remarks follow directly from the characterization of geodesics and canonical forms in terms of spurs, brackets and turns.

Remark 1. The image path of any index path of a combinatorial geodesic that is not exceptional is geodesic. If the combinatorial geodesic is in canonical form, so is the image path.

Remark 2. Likewise, any power c^k of a combinatorial closed geodesic c that is not exceptional is also a combinatorial geodesic. Moreover, if c is in canonical form, so is c^k.

5 Computing intersection numbers on oriented surfaces

Here, we assume that the considered system of quads is orientable and consistently oriented. In this case we can rely on canonical forms to obtain a relatively simple algorithm for the computation of the geometric intersection number. We shall see in Proposition 18 that the set B/τ of Lemma 4, composed of equivalence classes of lifts, can be identified with certain *crossing double paths*. Those are pairs of index paths with coincident images, see Section 5.1. The asymptotic complexity of the resulting algorithm is the same as for the general case, where the system of quads may be non-orientable. The general case leads to a slightly more complicated algorithm and is deferred to Section 6.

The next technical Lemma will be used in Proposition 18 to analyze the intersection of canonical curves. Let c_R and c_L^{-1} be canonical paths such that $c_R \sim c_L$. In other words, c_L is the leftmost geodesic homotopic to c_R. By Theorem 10, there is a disk diagram Δ composed of quad staircases and paths and whose left and right boundaries Δ_L and Δ_R are labelled by c_L and c_R respectively. A **spoke** is a non-boundary edge of Δ.

Lemma 13. *Let v, w be two vertices, one on each side of Δ. Then Δ contains a path p from v to w labelled by a canonical path. Moreover, p can be uniquely decomposed as either $\lambda.\rho$, $\rho.\lambda$, $\lambda.e.\rho$ or $\rho.e.\lambda$, where*

1. *λ is a subpath (possibly reduced to a vertex) of Δ_L or Δ_L^{-1},*
2. *ρ is a subpath (possibly reduced to a vertex) of Δ_R or Δ_R^{-1},*
3. *e is a spoke,*
4. *if λ is a subpath of Δ_L^{-1} of positive length then $p \cap \Delta_L^{-1} = \lambda$,*

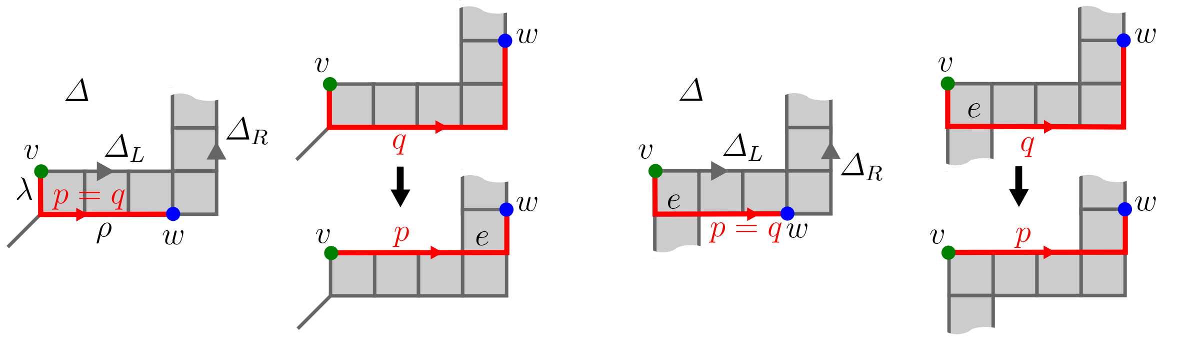

Figure 6: The canonical path p from v to w when v is not incident to a spoke.

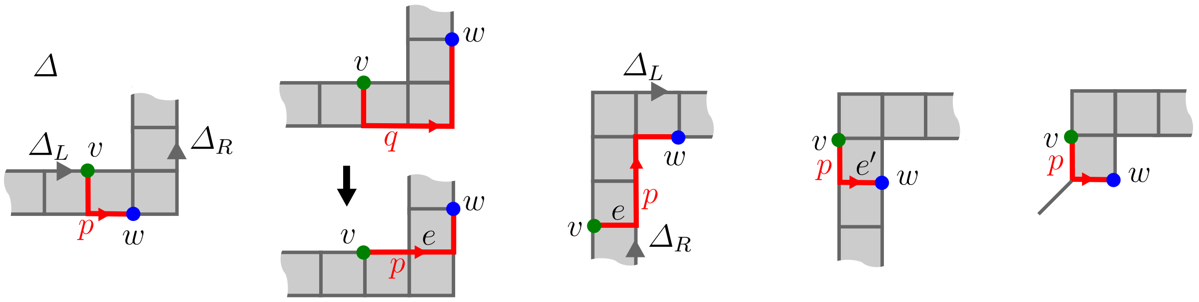

Figure 7: The canonical path p from v to w when v is incident to exactly one spoke.

5. *if ρ is a subpath of Δ_R of positive length then $p \cap \Delta_R = \rho$.*

6. *if ρ is a subpath of Δ_R^{-1} and λ is a subpath of Δ_L then either ρ is reduced to a vertex and $p \cap \Delta_R = \rho$ or λ is reduced to a vertex and $p \cap \Delta_L^{-1} = \lambda$.*

Proof. We assume that v is on the left side and w on the right side of Δ, the other case being symmetric. Let i, j be such that $v = \Delta_L(i)$ and $w = \Delta_R(j)$. We first consider the case where $j \geq i$. If Δ_L and Δ_R coincide at v, then we trivially obtain the desired decomposition as $p = \Delta_L[i \xrightarrow{0}].\Delta_R[i \xrightarrow{j-i}]$. Otherwise, v may be incident to $0, 1$ or 2 spokes.

- If v is not incident to a spoke then $\Delta_L(i-1)$ is either the initial tip of a staircase or is incident to a spoke e. We set $q = \Delta_L[i \xrightarrow{-1}].\Delta_R[i-1 \xrightarrow{j-i+1}]$ in the first case and $q = \Delta_L[i \xrightarrow{-1}].e.\Delta_R[i-1 \xrightarrow{j-i+1}]$ otherwise. The path q has no spurs or -1 turns but may start with a bracket. If not, by the characterization of canonical geodesic paths in Section 4.3, q is already canonical and we can set $p = q$. Otherwise, we short cut the bracket in q to obtain a canonical path p satisfying the above points 1 to 5. See Figure 6.

- If v is incident to exactly one spoke e, then e connects v to either $\Delta_R(i-1)$ or $\Delta_R(i+1)$. In the former case we set $q = e.\Delta_R[i-1 \xrightarrow{j-i+1}]$. As above, q may be canonical or starts with a bracket and we easily obtain the path p with the desired properties. See Figure 7. If e is incident to $\Delta_R(i+1)$ and $j > i$ the path $p = \Delta_L[i \xrightarrow{0}].e.\Delta_R[i+1 \xrightarrow{j-i-1}]$ has the required properties. When e is incident to $\Delta_R(i+1)$ and $j = i$, then either $\Delta_L(i-1)$ is incident to a spoke e' and we must have $p = \varDelta_L[i \xrightarrow{-1}].e'.\varDelta_R[i \xrightarrow{0}]$, or $\varDelta_L(i-1)$ is the initial tip of a staircase and we must have $p = \varDelta_L[i \xrightarrow{-1}].\varDelta_R[i-1 \xrightarrow{1}]$.

- If v is incident to two spokes, then one of them, e_-, connects v to $\varDelta_R(i-1)$ and the other e_+ connects v to $\varDelta_R(i+1)$. We can directly set $p = e_-.\varDelta_R[i-1 \xrightarrow{1}]$ if $j = i$ and $p = e_+.\varDelta_R[i+1 \xrightarrow{j-i-1}]$ if $j > i$.

For Point 6 in the Lemma, we note that we cannot have $p = \lambda.\rho$ or $p = \lambda.e.\rho$ with both λ and ρ being subpaths of positive length of $\varDelta_L$ and $\varDelta_R^{-1}$ respectively. Indeed, p would have a spur or a $\bar{1}$ turn in the first case and a bracket in the other case. Moreover, if ρ is reduced to a vertex and λ is a subpath of $\varDelta_L$, we may assume that the intersection $p \cap \varDelta_R$ is reduced to ρ. Otherwise, the canonical path between any other intersection point and ρ would have to follow $\varDelta_R$ and we could express p so that ρ is a subpath of positive length of $\varDelta_R$. An analogous argument holds to show that we can assume $p \cap \varDelta_L^{-1} = \lambda$ if λ is reduced to a vertex.

We next consider the case $i > j$. When $\varDelta_L$ and $\varDelta_R$ coincide at w, we obtain the desired decomposition as $p = \varDelta_L[i \xrightarrow{j-i}].\varDelta_R[i \xrightarrow{0}]$. Otherwise, w may be incident to $0, 1$ or 2 spokes. Similar arguments as in the case $j \geq i$ allow to conclude the proof. □

5.1 Crossing double-paths

Let c, d be two combinatorial closed curves on a combinatorial surface. A **double-path** of (c, d) of length ℓ is a pair of forward index paths $([i \xrightarrow{\ell}]_c, [j \xrightarrow{\ell}]_d)$ with the same image path $c[i \xrightarrow{\ell}] = d[j \xrightarrow{\ell}]$. If $\ell = 0$ then the double path is just a double point. A double path of c is defined similarly, taking $c = d$ and assuming $i \neq j$. The next Lemma follows from Remark 1.

Lemma 14. *Let $[i \xrightarrow{\ell}]_c$ and $[j \xrightarrow{k}]_d$ be forward index paths of two canonical curves c and d such that the image paths $c[i \xrightarrow{\ell}]$ and $[j \xrightarrow{k}]$ are homotopic. Then $k = \ell$ and $([i \xrightarrow{\ell}]_c, [j \xrightarrow{\ell}]_d)$ is a double path.*

A double path $([i \xrightarrow{\ell}]_c, [j \xrightarrow{\ell}]_d)$ gives rise to a sequence of $\ell+1$ double points $(i+k, j+k)$ for $k \in [0, \ell]$. A priori a double point could occur several times in this sequence. The next two lemmas claim that this is not possible when the curves are primitive.

Lemma 15. *A double path of a primitive combinatorial curve c cannot contain a double point more than once in its sequence. In particular, a double path of c must be strictly shorter than c.*

Proof. Suppose that a double path $\mathbb{P}$ of c contains two occurrences of a double point (i, j). Because the couples (i, j) and (j, i) represent the same double point there are two cases to consider.

- If $\mathbb{P}$ contains the couple (i, j) twice then it must contain a subsequence of length $|c|$ starting with (i, j). We thus have $c[i \xrightarrow{|c|}] = c[j \xrightarrow{|c|}]$. This implies that c is equal to some nontrivial circular permutation of itself. It is a simple exercise to check that c must then be a proper power of some other curve, contradicting that c is primitive.

- Otherwise $\mathbb{P}$ contains (i,j) and (j,i). Let ℓ be the distance between these two occurrences in $\mathbb{P}$. We thus have $c[i \xrightarrow{\ell}] = c[j \xrightarrow{|c|-\ell}]$ from which we deduce that c is a square (and $\ell = |c|/2$), contradicting that c is primitive.

□

Lemma 16. *Let c and d be two non-homotopic primitive combinatorial curves. A double path of (c,d) cannot contain a double point more than once in its sequence. Moreover, the length of a double path of (c,d) must be less than $|c|+|d|-1$.*

Proof. Suppose that a double path of (c,d) contains two occurrences of a double point. After shortening the double path if necessary, we may assume that these two occurrences are the first and the last double points of the double path. Its length must accordingly be a nonzero integer multiple p of $|c|$ as well as a nonzero integer multiple q of $|d|$. It follows that for some circular permutations c' of c and d' of d we have $c'^p = d'^q$. By a classical result of combinatorics on words [Lot97, Prop. 1.3.1] this implies that c' and d' are powers of a same curve, in contradiction with the hypotheses in the lemma. In fact, by a refinement due to Fine and Wilf [Lot97, Prop. 1.3.5] it suffices that c'^p and d'^q have a common prefix of length $|c|+|d|-1$ to conclude that c' and d' are powers of a same curve. This proves the second part of the lemma. □

A double path whose index paths cannot be extended is said **maximal**. As an immediate consequence of Lemmas 15 and 16 we have:

Corollary 17. *The maximal double paths of a primitive curve or of two primitive curves in canonical form induce a partition of the double points of the curves.*

Let (i,j) and $(i+\ell, j+\ell)$ be the first and the last double points of a maximal double path of (c,d), possibly with $c=d$. When $\ell \geq 1$ the arcs $c[i,i-1]$, $d[j,j-1]$, $c[i,i+1]$ must be pairwise distinct because canonical curves have no spurs, and similarly for the three arcs $c[i+\ell, i+\ell+1]$, $d[j+\ell, j+\ell+1]$, $c[i+\ell, i+\ell-1]$. We declare the maximal double path to be a **crossing double path** if the circular ordering of the first three arcs at $c(i)$ and the circular ordering of the last three arcs at $c(i+\ell)$ are either both clockwise or both counterclockwise with respect to the rotation system of the system of quads. When $\ell = 0$, that is when the maximal double path is reduced to the double point (i,j), we require that the arcs $c[i,i-1], d[j,j-1], c[i,i+1], d[j,j+1]$ are *pairwise distinct* and appear in this circular order, or its opposite, around the vertex $c(i) = d(j)$.

5.2 Proof of Theorem 1 in the primitive and orientable case

Let c,d be primitive combinatorial curves such that d is canonical and let c_R and c_L^{-1} be the canonical curves homotopic to c and c^{-1} respectively. We denote by $\varDelta$ the annular diagram corresponding to c_R and c_L. When the two boundaries $\varDelta_R$ and $\varDelta_L$ of $\varDelta$ have a common vertex we implicitly assume that c_R and c_L are indexed so that this vertex corresponds to the same index along $\varDelta_R$ and $\varDelta_L$. We consider the following set of double paths:

- $\mathcal{D}_+$ is the set of crossing double paths of positive length of c_R and d,

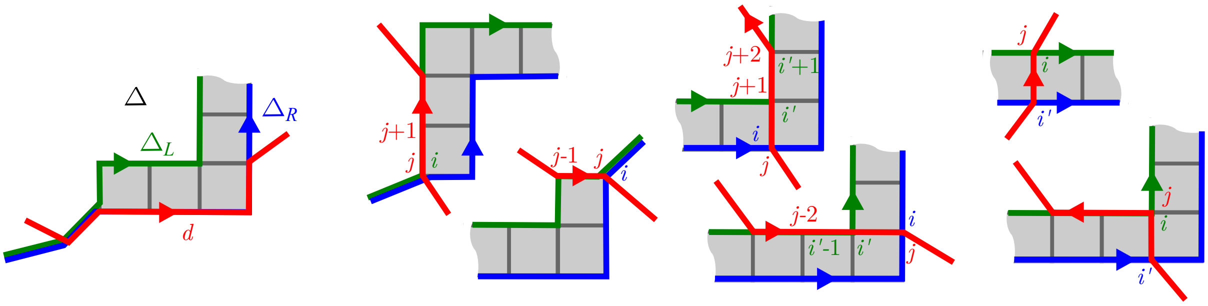

Figure 8: Left, a typical crossing double path in $\mathcal{D}_+$. Middle, four configurations in $\mathcal{D}_0$. Right, two configurations in $\mathcal{D}_-$.

- $\mathcal{D}_0$ is the set of crossing double paths (i, j) of zero length of c_R and d such that either
  - the two boundaries of $\varDelta$ coincide at $\varDelta_L(i) = \varDelta_R(i)$ and $d[j-1, j] = c_L[i-1, i]$ or $d[j, j+1] = c_L[i, i+1]$, or
  - one of $d[j, j-1]$ or $d[j, j+1]$ is the label of a spoke $(\varDelta_R(i), \varDelta_L(i'))$ of $\varDelta$ and $d[j-2, j-1] = c_L[i'-1, i']$ in the first case or $d[j+1, j+2] = c_L[i', i'+1]$ in the other case.
- $\mathcal{D}_-$ is the set of crossing double paths $([i \xrightarrow{\ell}]_{c_L^{-1}}, [j \xrightarrow{\ell}]_d)$ $(\ell \geq 0)$ of c_L^{-1} and d such that *none* of the following situations occurs:
  - the two boundaries of $\varDelta$ coincide at $\varDelta_L^{-1}(i) = \varDelta_R(i')$ and $d[j-1, j] = c_R[i'-1, i']$,
  - the two boundaries of $\varDelta$ coincide at $\varDelta_L^{-1}(i+\ell) = \varDelta_R(i')$ and $d[j+\ell, j+\ell+1] = c_R[i', i'+1]$,
  - $d[j-1, j]$ is the label of a spoke $(\varDelta_L^{-1}(i), \varDelta_R(i'))$ of $\varDelta$ and $d[j-2, j-1] = c_R[i'-1, i']$,
  - $d[j+\ell, j+\ell+1]$ is the label of a spoke $(\varDelta_L^{-1}(i+\ell), \varDelta_R(i'))$ of $\varDelta$ and $d[j+\ell+1, j+\ell+2] = c_R[i', i'+1]$.

In case $c \sim d$ or $c \sim d^{-1}$, which we can detect by comparing their canonical forms, we enforce $c_R = d$, inverting d if necessary. This way, recalling that the index paths of a double path of c must be distinct by definition, the maximal double paths have finite length by Lemma 15. Figure 8 depicts some configurations. Referring to Section 3, we view the underlying surface of the system of quads, call it $\varSigma$, as a quotient $\mathbb{D}/\varGamma$ of the Poincaré disk. The system of quads lifts to a quadrangulation of $\mathbb{D}$ and the lifts of a combinatorial curve in $\varSigma$ are combinatorial bi-infinite paths in this quadrangulation. By Remark 2, if the combinatorial curve is geodesic (resp. canonical) so are its lifts. In this case, each lift is simple by Corollary 12. We fix a lift $\tilde{c_R}$ of c_R and consider the set B/τ of Lemma 4 corresponding to the classes of lifts of d whose limit points alternate with the limit points of $\tilde{c_R}$ along $\partial\mathbb{D}$.

Proposition 18. *B/τ is in 1-1 correspondence with the disjoint union $\mathcal{D}_+ \cup \mathcal{D}_0 \cup \mathcal{D}_-$.*

Proof. Let $\tilde{c_L}$ be the lift of c_L with the same limit points as $\tilde{c_R}$. These two lifts project onto the boundaries of the annular diagram Δ and thus form an infinite strip $\tilde{\Delta}$ of width at most 1 in $\mathbb{D}$ composed of paths and quad staircases (possibly a single infinite staircase). We shall define a correspondence between B/τ and $\mathcal{D}_+ \cup \mathcal{D}_0 \cup \mathcal{D}_-$. To this end we consider a lift $\tilde{d}$ of d whose limit points alternate with those of $\tilde{c_R}$. In other words, $\tilde{d} \in B$. The lift $\tilde{d}$ must cross $\tilde{\Delta}$. Let i and j be respectively the smallest and largest index k such that $\tilde{d}(k)$ is in $\tilde{\Delta}$. By Remark 1, the corresponding subpath $\tilde{d}[i,j]$ of $\tilde{d}$ is canonical. Since $\mathbb{D}$ is simply connected, $\tilde{d}[i,j]$ is homotopic to any path joining the same extremities and we can apply Lemma 13 to show that $\tilde{d}[i,j]$ is actually contained in $\tilde{\Delta}$ and that it can be decomposed as either $\lambda.\rho$, $\rho.\lambda$, $\lambda.e.\rho$ or $\rho.e.\lambda$ where e is a spoke of $\tilde{\Delta}$, $\rho = \tilde{c_R}[a \xrightarrow{r}]$ and $\lambda = \tilde{c_L}[b \xrightarrow{\ell}]$ for some $a, b, r, \ell \in \mathbb{Z}$.

- If $r > 0$ then by Point 5 of Lemma 13 we have $\tilde{d} \cap \tilde{c_R} = \rho$ so that this intersection defines a maximal double path of $\tilde{d}$ and $\tilde{c_R}$. It must be crossing since $\tilde{d} \in B$. Its projection on Σ is a crossing double path of length $r > 0$ of d and c_R, hence in $\mathcal{D}_+$, to which we map $\tilde{d}$.

- If $r = 0$ and $\ell > 0$, then Point 6 of Lemma 13 implies $\tilde{d} \cap \tilde{c_R} = \rho$. As above, ρ must define a crossing double path of length zero of $\tilde{d}$ and $\tilde{c_R}$. We map $\tilde{d}$ to the projection of this crossing double point on Σ and remark that this projection is in $\mathcal{D}_0$.

- Otherwise, we must have $r \leq 0$ and $\ell \leq 0$ by Point 6 of Lemma 13. If $\ell < 0$ then Point 4 of Lemma 13 implies $\tilde{d} \cap \tilde{c_L}^{-1} = \lambda$. This intersection defines a crossing double path of $\tilde{d}$ and $\tilde{c_L}^{-1}$ and we map $\tilde{d}$ to its projection on Σ. If $\ell = 0$ and $r < 0$ then Point 6 of Lemma 13 implies $\tilde{d} \cap \tilde{c_L}^{-1} = \lambda$, which also holds true if $r = \ell = 0$. In both cases λ corresponds to a crossing double path of length zero of $\tilde{d}$ and $\tilde{c_L}^{-1}$ and we map $\tilde{d}$ to its projection on Σ. We finally remark that this last projection or the above one belong to $\mathcal{D}_-$.

Because $\tilde{\Delta}$ is left globally invariant by τ (the hyperbolic motion that sends $\tilde{c_R}(0)$ to $\tilde{c_R}(|c_R|)$), we have $\tau(\tilde{d}) \cap \tilde{\Delta} = \tau(\tilde{d}) \cap \tau(\tilde{\Delta}) = \tau(\tilde{d} \cap \tilde{\Delta})$. It follows that $\tilde{d}$ and $\tau(\tilde{d})$ are mapped to the same crossing double path by the above rules. We thus have a well defined map $B/\tau \to \mathcal{D}_+ \cup \mathcal{D}_0 \cup \mathcal{D}_-$. The uniqueness of the decomposition in Lemma 13 implies that this map is 1-1. In order to check that the map is onto we consider a maximal crossing double path $\mathbb{P} = ([i \mod |c_R| \xrightarrow{\ell}]_{c_R}, [j \mod |d| \xrightarrow{\ell}]_d)$ of c_R and d in $\mathcal{D}_+$. By the unique lifting property of coverings there is a unique lift $\tilde{d}$ of d such that $\tilde{d}(j) = \tilde{c_R}(i)$ and $\mathbb{P}$ lifts to a crossing double path of $\tilde{d}$ and $\tilde{c_R}$. By Lemma 14 this double path is the only intersection of $\tilde{d}$ and $\tilde{c_R}$ so these lifts must have alternating limit points. In other words $\tilde{d}$ is in B and is mapped to $\mathbb{P}$. A similar argument applies to the crossing double paths of $\mathcal{D}_0$ and $\mathcal{D}_-$. □

This leads to a simple algorithm for computing combinatorial crossing numbers.

Corollary 19. *Let c, d be primitive curves of length at most ℓ on an orientable combinatorial surface with complexity n. The crossing numbers $i(c,d)$ and $i(c)$ can be computed in $O(n + \ell^2)$ time.*

Proof. By Lemma 5 we may assume that the surface is a system of quads. By Theorem 8 we may compute the canonical forms of c, c^{-1} and d in $O(\ell)$ time. According to Proposition 18, we have

$$i(c,d) = |\mathcal{D}_+| + |\mathcal{D}_0| + |\mathcal{D}_-|$$

The set $\mathcal{D}_+$ can be constructed in $O(\ell^2)$ time. Indeed, since the maximal double paths of c and d form disjoint sets of double points by Corollary 17, we just need to traverse the grid $\mathbb{Z}/|c|\mathbb{Z} \times \mathbb{Z}/|d|\mathbb{Z}$ and group the double points into maximal double paths. Those correspond to diagonal segments in the grid that can be computed in time proportional to the size of the grid. We can also determine which double paths are crossing in the same amount of time. Likewise, we can construct the sets $\mathcal{D}_0$ and $\mathcal{D}_-$ in $O(\ell^2)$ time. □

The cases where c and d are non-primitive or where the Euler characteristic of the input surface is non-negative are dealt with in Section 6.4.

6 Computing intersection numbers: the general case

Here, we drop the orientability assumption to consider systems of quads that may be non-orientable. We first relate the intersection number of two primitive combinatorial curves with their number of *crossing pairs of index paths* as defined below. See Lemma 20. Let c, d be primitive combinatorial closed geodesics on a system of quads Σ. We assume that neither of them is exceptional as defined above Corollary 12. If c is homotopic to d or to its inverse, which we can detect by considering the canonical forms of their squares (see the discussion before Proposition 25), we enforce $c = d$. As in Section 5.2, we view the underlying surface of Σ as a quotient $\mathbb{D}/\Gamma$ of the Poincaré disk. The system of quads lifts to a quadrangulation of $\mathbb{D}$ and the lifts of c, d in Σ are combinatorial bi-infinite paths in this quadrangulation. Note that enforcing $c = d$ whenever c was initially homotopic to d or its inverse ensures that the lifts of c and d have pairwise distinct limit points, unless those lifts are equal (in which case we obviously have $c = d$). We fix a lift $\tilde{c}$ of c and consider any lift $\tilde{d}$ of d whose limit points alternate with the limit points of $\tilde{c}$ along $\partial\mathbb{D}$. It follows from Remark 2 and Corollary 11 that $\tilde{c}$ and $\tilde{d}$ are geodesics that each minimizes the length between any two of its vertices. Let i be the smallest index along $\tilde{c}$ such that $\tilde{c}(i) \in \tilde{d}$ and let ℓ be maximal such that $\tilde{c}(i+\ell) \in \tilde{d}$. By the preceding discussion $\tilde{c}(i)$ and $\tilde{c}(i+\ell)$ also delimitate a maximal subpath $\tilde{d}[j, j+\varepsilon\ell]$ of $\tilde{d}$ whose endpoints belong to $\tilde{c}$, with $\varepsilon \in \{-1, 1\}$. We are thus in the situation depicted on Figure 2, right, where $\tilde{c}[-\infty, i-1]$ and $\tilde{c}[i+\ell+1, +\infty]$ lie on either side of $\tilde{d}$ in $\mathbb{D}$. In this case we say that the pair of index paths $([i \xrightarrow{\ell}]_c, [j \xrightarrow{\varepsilon\ell}]_d)$ is **crossing**. This happens precisely when $([i \xrightarrow{\ell}]_c, [j \xrightarrow{\varepsilon\ell}]_d)$ is a maximal pair of index paths with homotopic image paths and when the circular ordering of the arcs $\tilde{c}[i, i+1], \tilde{c}[i, i-1]$ and $\tilde{d}[j, j-\varepsilon]$ is the same as the circular ordering of $\tilde{c}[i+\ell, i+\ell-1], \tilde{c}[i+\ell, i+\ell+1]$ and $\tilde{d}[j, j+\varepsilon(\ell+1)]$. This last condition can be checked by comparing the direct circular orderings of the projection of those arcs on Σ, taking into account the parity of the number of half-twisted arcs in $c[i \xrightarrow{\ell}]$ (or equivalently in $d[j \xrightarrow{\varepsilon\ell}]$). We emphasize that the parameters i, j, ℓ, ε are attached to c and d, not to their lifts. Those parameters determine a crossing pair of index paths for c and d if there exist two lifts $\tilde{c}$ of c and $\tilde{d}$ of d such as described above. According to Lemma 4 in the strategy section, we have

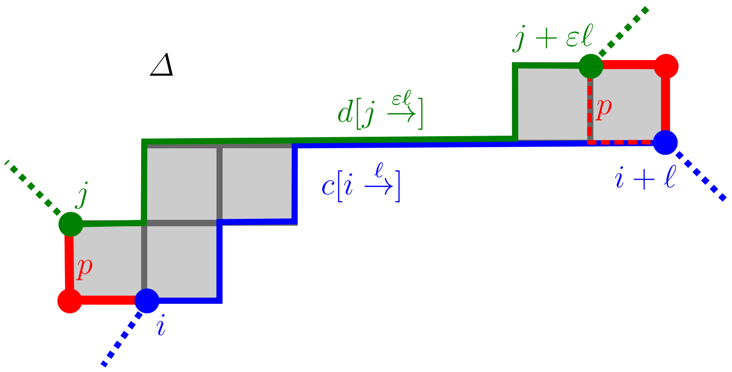
Δ
$j + \varepsilon\ell$
$d[j \xrightarrow{\varepsilon\ell}]$
p
$c[i \xrightarrow{\ell}]$
$i + \ell$
j
p
i

Figure 9: A thick double path extends to a pair of homotopic geodesics and bound a partial diagram. When (i, j) and $(i + \ell, j + \varepsilon\ell)$ are in the same configuration, the corresponding pairs of vertices in Δ are both joined by a path of length zero or two labelled by the same path p.

Lemma 20. *Let c, d be primitive combinatorial closed geodesics, none of which is exceptional. We assume that c is not homotopic to d or its inverse, unless c is precisely equal to d. Then, $i(c, d)$ is equal to the number of crossing pairs of index paths.*

The case where c or d is exceptional can be dealt with separately. See the proof of Lemma 23 and Section 6.3 for further details on the exceptional case.

6.1 Thick double-paths

Following the description of Theorem 10, the two sides $c[i \xrightarrow{\ell}]$ and $d[j \xrightarrow{\varepsilon\ell}]$ of a crossing pair of index paths $([i \xrightarrow{\ell}]_c, [j \xrightarrow{\varepsilon\ell}]_d)$ label a disk diagram composed of a sequence of paths and staircases. In particular, the vertices of the two sides can be put in 1-1 correspondence and corresponding vertices satisfy a local condition: they stay at distance zero or two and bound a same quad in the disk diagram. Hence, when looking for crossing pairs of index paths of (c, d), we can start from any double point (i, j) and walk in parallel along the two sides from i and j, checking the above local condition at each step. However, we do not know a priori if the two sides will meet again to form a pair of homotopic paths and the search may be unsuccessful. In order to analyze the complexity of the search we consider pairs of paths that could potentially be part of a crossing pair of index paths, but are not necessarily so. Formally, a pair $([i \xrightarrow{\ell}]_c, [j \xrightarrow{\varepsilon\ell}]_d)$ of distinct index paths such that $c[i \xrightarrow{\ell}]$ and $d[j \xrightarrow{\varepsilon\ell}]$ are corresponding subpaths of homotopic geodesic paths is called a **thick double path** of length ℓ. In other words, a thick double path is such that its image paths can be extended to form homotopic geodesic paths. A thick double path gives rise to a sequence of $\ell + 1$ index pairs $(i + k, j + \epsilon k)$ for $k \in [0, \ell]$. Let Δ be a disk diagram whose boundary is labelled by homotopic geodesic paths extending the thick double path as illustrated on Figure 9. The restriction of Δ to the part delimited by the thick double path is called a **partial diagram**. In a partial diagram each index pair (i, j) may be in at most one of five **configurations**: either the corresponding vertices coincide in the partial diagram, or they appear diagonally opposite in a quad; in turn this quad may be labelled by one of the (at most) four quads incident to either $c[i, i + 1]$

or $d[j, j+\varepsilon]$. A partial diagram is **maximal** if it cannot be extended to form a partial diagram of a longer thick double path.

Lemma 21. *Let c and d be primitive geodesic curves, none of which is exceptional. Suppose in addition that either $c = d$, or c is not homotopic to d or its inverse. Then each configuration of an index pair may occur at most once in a partial diagram.*

Proof. Suppose for a contradiction that the index pair (i, j) occurs twice in the same configuration in a partial diagram of a thick double path of (c, d). The two occurrences of (i, j) are separated by a path whose length ℓ is an integer multiple of $|c|$, say $\ell = k|c|$, and an integer multiple of $|d|$, say $\epsilon\ell = t|d|$. In the partial diagram, $c(i)$ and $d(j)$ label vertices that are either identical or opposite in a quad. They are thus connected by a path p of length zero or two. See Figure 9. We infer that

$$c[i \xrightarrow{k|c|}] \overset{*}{\sim} p \cdot d[j \xrightarrow{t|d|}] \cdot p^{-1} \tag{1}$$

It ensues that $c^k \sim d^t$, considering free homotopy. Because c and d are primitive curves on a surface with negative Euler characteristic, this implies that c is homotopic to d or its inverse: every homotopy class has indeed a unique primitive root up to orientation on such a surface [Rei62, Lem. 9.2.6], [Bus92, p. 213]. By the hypotheses in the lemma, we thus have $c = d$. We first assume that the two index paths of the thick double path are both forward (i.e., $\varepsilon = 1$). Exchanging the roles of i and j if necessary, we may further assume that $0 < j - i \leq |c|/2$. Equation (1) now writes

$$c[i \xrightarrow{k|c|}] \overset{*}{\sim} p \cdot c[j \xrightarrow{k|c|}] \cdot p^{-1} \overset{*}{\sim} p \cdot c[i,j]^{-1} \cdot c[i \xrightarrow{k|c|}] \cdot c[i,j] \cdot p^{-1}$$

Equivalently, $c'^k \cdot q \overset{*}{\sim} q \cdot c'^k$ where $c' := c[i \xrightarrow{|c|}]$ is a cyclic permutation of c and $q := p \cdot c[i,j]^{-1}$ is a (closed) path of length at most $j - i + 2$. Since they are commuting, c'^k and q must admit a common primitive root [Rei62, Lem. 9.2.6], [Bus92, p. 213]. Whence $q \overset{*}{\sim} c'^t$ for some $t \in \mathbb{Z}$, since c' is itself primitive. By Remark 2, c'^t is geodesic, hence has minimal length in its homotopy class. It follows that $j - i + 2 \geq |t|.|c|$. This is only possible if $t = 0$ or if $|t| = 1$ and $j - i = |c| - 2$ (recall that the graph of the system of quads is bipartite so that indices in a pair have the same parity).

- If $t = 0$, then $q \sim 1$, so that $p \overset{*}{\sim} c[i, j]$. Call Δ the partial diagram for $c[i \xrightarrow{k|c|}]$ and $c[j \xrightarrow{k|c|}]$. The uniqueness of disk diagrams discussed after Corollary 11 implies that the initial point of $c[j \xrightarrow{k|c|}]$ and the point of index $j - i$ along $c[i \xrightarrow{k|c|}]$ (which is also $c[j]$) are mapped to the same point in Δ. However, coincident points in a (partial) diagram must have the same index along each side. This would imply $i = j$, which is impossible since we are only considering pairs of distinct index paths.

- If $|t| = 1$ and $j - i = |c| - 2$, then $|c| = 4$ (recall that $0 < j - i \leq |c|/2$) and thus $j = i + 2$. In the case $t = 1$, we have $c' \overset{*}{\sim} p \cdot c[i, j]^{-1} \overset{*}{\sim} p \cdot c[j \xrightarrow{2}] \cdot c'^{-1}$, whence $(c')^2 \sim p \cdot c[j \xrightarrow{2}]$. This is however impossible as the left member of the homotopy has minimal length 8 in its homotopy class while the right member has length 4. In the case $t = -1$, we have $c' \overset{*}{\sim} c[i, j] \cdot p^{-1}$, whence $p \overset{*}{\sim} c[j \xrightarrow{2}]^{-1}$. Similarly to the case $t = 0$, we infer that the initial point of $c[i \xrightarrow{k|c|}]$ and the point of index 2 along $c[j \xrightarrow{k|c|}]$ are mapped to the same point in the partial diagram, which is also impossible.

We now assume that the thick double path is composed of a forward and a backward index paths. Similarly to the previous case, we infer that $c' \overset{*}{\sim} q \cdot (c')^{-1} \cdot q^{-1}$, where $c' := c[i \xrightarrow{|c|}]$ and $q := p \cdot c[i, j]^{-1}$. Equivalently, c and its inverse are freely homotopic. This is however impossible unless c is contractible. □

Corollary 22. *Each configuration of an index pair may occur at most once in the set of maximal partial diagrams.*

Proof. By the preceding Lemma, we only need to check that a configuration of an index pair cannot occur twice in distinct partial diagrams of maximal thick double paths. This is essentially the unique lifting property of coverings. □

6.2 Proof of Theorem 1: the primitive case

Lemma 23. *Let c, d be primitive curves of length at most ℓ on a system of quads. The crossing number $i(c, d)$ can be computed in $O(\ell^2)$ time.*

Proof. We may assume that c and d are geodesic by Theorem 9. We first consider the case where none of c or d is exceptional. According to Lemma 20, we can compute $i(c, d)$ by enumerating the crossing pairs of index paths of (c, d). Following the previous Section 6.1 we only need to list the maximal partial diagrams and to count those that correspond to crossing pairs of index paths. As described above Lemma 20, this last step can be performed in time proportional to the total length of the maximal partial diagrams, which is $O(\ell^2)$ by Corollary 22. Since we are only interested in partial diagrams whose sides have coincident vertices, namely the endpoints of the crossing pairs of index paths, we can list the set of maximal partial diagrams as follows. Pick any index pair (i, j) such that $c(i) = d(j)$ and form a trivial partial diagram with these vertices in coincident configuration. Extend the partial diagram maximally in both directions. By the unique lifting property of coverings, there is only one way to do so. Mark all the configurations of index pairs found in this diagram and pick a new index pair which is not marked in the coincident configuration. Repeat until all the coincident configurations of index pairs are marked. The total time spent by the procedure is proportional to the total length of the maximal partial diagrams, hence is $O(\ell^2)$.

We next consider the case where d is not exceptional but c is an exceptional 1-sided curve. Its square c^2 is a 2-sided curve with a unique homotopic geodesic whose turn sequence is 2^* or $\bar{2}^*$. In particular, c^2 cannot form a staircase with any curve since both sides of a staircase must have a vertex whose corresponding turn is ± 1. (Note that it forms an infinite strip of width one with itself.) It follows that the crossing pairs of index paths of (c^2, d) are **flat**, i.e., the index paths have identical image paths. This makes the search for crossing pairs of index paths even easier than in the general case so that $i(c^2, d)$ can be computed in $O(\ell^2)$ time. Applying the formulas in Proposition 26 below (with $p = 2$ and $q = 1$) we finally derive $i(c, d) = i(c^2, d)/2$. Similar arguments apply when both c and d are exceptional to compute $i(c, d) = i(c^2, d^2)/4$, including the case where c and d are homotopic. □

6.3 Non-primitive curves

In order to finish the proof of Theorem 1 we need to handle the case of non-primitive curves. Thanks to canonical forms, computing the primitive root of a curve becomes extremely simple (see also Lustig [Lus87, Rem. 5.2]) on orientable surfaces.

Lemma 24. *Let c be a combinatorial curve of length $\ell > 0$ in canonical form on an oriented system of quads. A primitive curve d such that c is homotopic to d^k for some integer k can be computed in $O(\ell)$ time.*

Proof. By Theorem 8, we may assume that c is in canonical form. Let d be a primitive curve such that $c \sim d^k$ for some integer k. We look for d in canonical form. By Remark 2, the curve d^k is also in canonical form. The uniqueness of the canonical form implies that $c = d^k$, possibly after some circular shift of d. It follows that d is the smallest prefix of c such that c is a power of this prefix. It can be found in $O(\ell)$ time using a variation of the Knuth-Morris-Pratt algorithm to find the smallest period of a word [KMP77]. □

We now consider the case of a curve c on a non-orientable system of quads. If c is 1-sided we can reduce the computation of a primitive root to the 2-sided case by noting that c and c^2 have the same primitive roots. If c is 2-sided it can be given one of two canonical forms by fixing one of the two consistent orientations along the curve before removing spurs, brackets and -1's in its turn sequence. Alternatively, we can build and orient the double cover of the system of quads and note that the 2-sided curves are precisely those that lift to closed curves in this cover. The projection of the canonical form of a lift provides one of two canonical forms corresponding to the two possible lifts. By Lemma 24 we can compute in linear time a primitive root of the lift in canonical form. Its projection d is thus a 2-sided geodesic. If d is itself primitive, then d is a primitive root of c. Otherwise, $d \sim e^2$ for some 1-sided primitive geodesic e, which is thus a primitive root of c.

- If e is exceptional then d has a constant turn sequence 2^* or $\bar{2}^*$ and bounds an immersed annulus wrapping twice around a Möbius strip. Consider any non-boundary arc a of this annulus. Let $d(i)$ be the source point of a, and let $\ell = |d|/2$. Then $d \sim (d[i + \ell \xrightarrow{\ell}] \cdot a)^2$ so that e must be homotopic to $d[i \xrightarrow{\ell}] \cdot a$. See Figure 10.

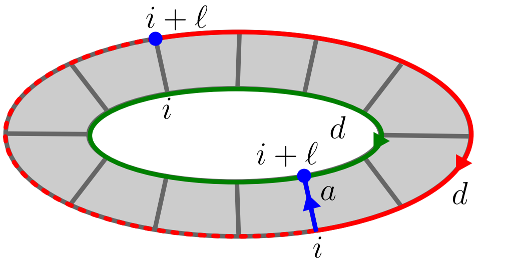

Figure 10: An annular diagram for d with itself.

- If e is not exceptional, then e^2 is geodesic so that $|e| = \ell$ with $\ell = |d|/2$. Consider an annular diagram Δ for d and e^2. As for a thick partial diagram, the boundary loops of Δ can be put in 1-1 correspondence so that corresponding vertices appear in one of five possible configurations (in fact, only three are possible since d is canonical). Let p_0 and p_ℓ be paths of length at most two connecting $d(0)$ and $d(\ell)$ to respectively $e^2(0)$ and $e^2(\ell)$. The possible configurations provide at most five candidates for p_0 and similarly for p_ℓ. For each pair of candidates, say (q_0, q_ℓ), we consider the curve $e' := q_0^{-1} \cdot d[0 \xrightarrow{\ell}] \cdot q_\ell$. We can confirm that the candidate paths are the right ones by checking whether e'^2 is indeed homotopic to d. This takes $O(\ell)$ time by Theorem 8.

To summarize the discussion,

Proposition 25. *Given a curve c of length ℓ on a system of quads, orientable or not, a primitive root of c can be computed in $O(\ell)$ time.*

6.4 End of proof of Theorem 1

We now have all the ingredients to complete the proof of Theorem 1. We essentially need to extend Lemma 23 to non-primitive curves and to handle the case of curves on surfaces of non-negative Euler characteristic. The geometric intersection number of non-primitive curves is related to the geometric intersection number of their primitive roots by simple formulas. The next result is rather intuitive, see Figures 11 and 12, and is part of the folklore although we could only find references in some relatively recent papers.

Proposition 26 ([dGS97, GKZ05])**.** *Let c and d be primitive curves such that c is not homotopic to d or its inverse and let p, q be positive integers. Then,*

$$i(c^p, d^q) = pq \times i(c, d)$$

and

$$i(c^p, c^q) = \begin{cases} 2pq \times i(c) + \min\{p, q\} & \textit{if } c \textit{ is 1-sided and } p \textit{ and } q \textit{ are odd,} \\ 2pq \times i(c) & \textit{otherwise.} \end{cases}$$

Moreover, the geometric self-intersection number of a curve is related to the geometric self-intersection number of its root thanks to the formulas

$$i(c^p) = \begin{cases} p^2 \times i(c) + p - 1 & \textit{if } c \textit{ is 2-sided,} \\ p^2 \times i(c) + \lfloor \frac{p-1}{2} \rfloor & \textit{otherwise.} \end{cases}$$

Proof of Theorem 1. Let c, d and Σ be the two combinatorial curves and the combinatorial surface as in the Theorem. By Lemma 5 we can assume after $O(n)$ time preprocessing that Σ is a system of quads. We first consider the case where Σ has negative Euler characteristic. Thanks to Theorem 9 and Proposition 25 we can determine geodesic primitive curves c' and d' and integers p, q such that $c \sim c'^p$ and $d \sim d'^q$ in $O(\ell)$ time. We then compute $i(c', d')$ and $i(c', c')$ in $O(\ell^2)$ time according to Corollary 23. We finally use the formulas in the previous proposition to deduce $i(c, d)$ and $i(c)$ from $i(c', d')$ and $i(c', c')$.

We next consider the case where Σ has non-negative Euler characteristic.

- If Σ is a sphere or a disk, then every curve is contractible and $i(c, d) = i(c) = 0$.

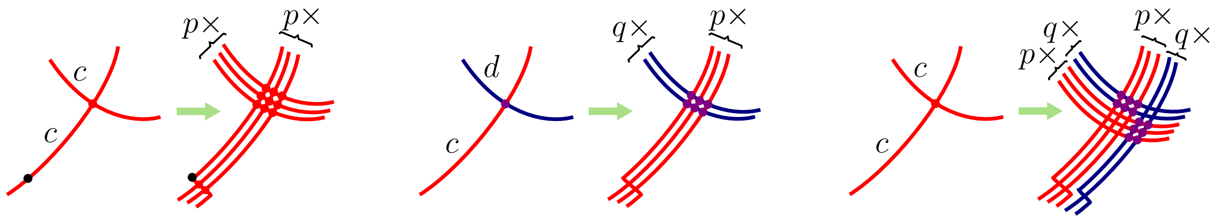

Figure 11: Left, each self-intersection of a 2-sided curve c gives rise to p^2 self-intersections of its pth power obtained by wrapping p times around c in a small tubular neighborhood. $p-1$ additional self-intersections are needed to connect the end of the wrapping curve with the basepoint. Middle, each intersection of c and d gives rise to pq intersections of (homotopic perturbations of) c^p and d^q. Right, when c and d are homotopic and c^p or c^q is 2-sided, one should count $2pq$ intersections per self-intersection of c.

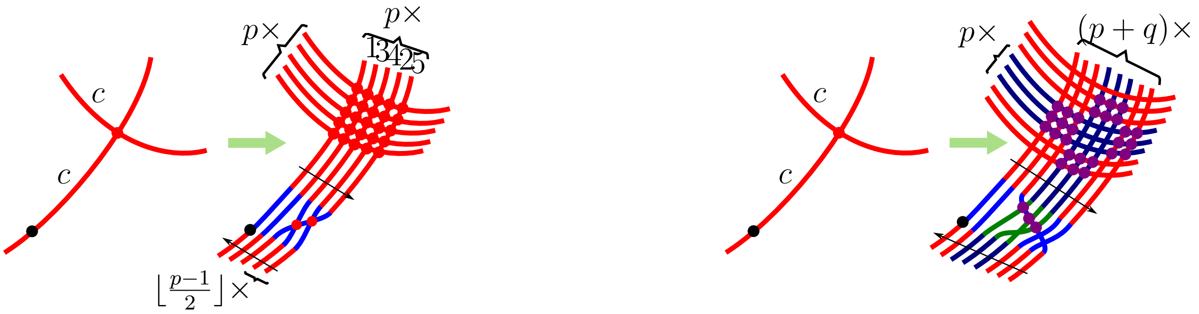

Figure 12: Left, the pth power of a 1-sided curve c is obtained from p copies of c running parallel to c and arriving at the starting basepoint of c in reverse order. By ordering the copies conveniently we get $\lfloor \frac{p-1}{2} \rfloor$ self-intersections to reconnect the copies instead of $p-1$ for a 2-sided curve. Right, a minimal configuration for c^p and c^q with $p=3$ and $q=5$.

- If Σ is a cylinder, then every two curves can be made non crossing so that $i(c,d) = 0$ while $i(c) = p - 1$.

- If Σ is a projective plane, there is only one non-trivial homotopy class, say c', for which $i(c', c') = 1$ and $i(c') = 0$.

- If Σ is a torus, there is no need to build a system of quads; we may just compute the reduced surface which is composed of two loop edges. Those loop edges, say α and β, represent generators of the fundamental group of the torus and we can write $c \sim \alpha^x \cdot \beta^y$ and $d \sim \alpha^{x'} \cdot \beta^{y'}$. It results from classical formulas that: $i(c) = \gcd(x,y) - 1$ and $i(c,d) = |\det\big((x,y),(x',y')\big)|$.

- Finally, if Σ is a Klein bottle, similarly to the torus case the reduced surface is composed of two loops α, β satisfying either the relation $\alpha^2\beta^2 \sim 1$ or $\alpha\beta\alpha\beta^{-1} \sim 1$. We can transform the embedded graph in constant time so that the second relation holds. This relation can be written as a pseudo commutation relation $\alpha\beta = \beta\alpha^{-1}$ that preserves the cumulative sum of the exponents of β and the parity of the cumulative sum of the exponents of α. From this simple remark we easily derive that every conjugacy class in the fundamental group has a unique normal form $\alpha^m\beta^{2n}$, β^{2n+1} or $\alpha\beta^{2n+1}$, where m, n can take any integer value. Table 1 provides all the geometric intersection numbers for the primitive classes, which must have one of the forms β, $\alpha\beta$, or $\alpha^n\beta^{2m}$ with m and n coprime. We can combine this table with the formulas of Propositions 26 to compute the geometric intersection number of any curves.

	$\alpha^k\beta^{2\ell}$	β	$\alpha\beta$
$\alpha^m\beta^{2n}$	$\left\lVert\begin{smallmatrix}k & m\\ \ell & n\end{smallmatrix}\right\rVert + \left\lVert\begin{smallmatrix}k & -m\\ \ell & n\end{smallmatrix}\right\rVert$	$\lvert m\rvert$	$\lvert m\rvert$
β	$\lvert k\rvert$	1	0
$\alpha\beta$	$\lvert k\rvert$	0	1

c	$\alpha^k\beta^{2\ell}$	β	$\alpha\beta$
$i(c)$	$\lvert k\ell\rvert$	0	0

Table 1: Left table, the geometric intersection numbers for the pairs of distinct primitive homotopy classes. Here, k and ℓ are coprime, as well as m and n, and $(k,\ell) \neq \pm(m,n)$. We use the notation $\left\lVert\begin{smallmatrix}x & y\\ z & t\end{smallmatrix}\right\rVert := |xt - yz|$. Right table, the geometric self-intersection numbers of primitive classes. Again, k and ℓ are assumed to be coprime.

□

7 Combinatorial perturbations

A combinatorial curve on a combinatorial surface may be seen as a continuous curve in general position, i.e. with a finite number of transverse pairwise crossings, snapped to the graph G of the combinatorial surface. When doing so several parts of the continuous curve may be mapped to the same edge of G. In order not to lose information, one needs to record their ordering transverse to that edge. Following the notion of a combinatorial set of loops as in [CdVL05], we define a **combinatorial perturbation** of a set $\mathcal{C}$ of combinatorial curves as the data for each arc a in G of a left-to-right order $\preceq_a$ over all the occurrences of a or a^{-1} in the curves of $\mathcal{C}$. The only requirement is that opposite arcs should be associated with inverse orders[4]. Let A_v be the set of occurrences of a or a^{-1} in the curves of $\mathcal{C}$ where a runs over all arcs of G with origin v. A combinatorial perturbation induces for each vertex v of G a circular order $\preceq_v$ over A_v; if $a_1, \ldots, a_k$ is the direct cyclic order of the arcs of G with origin v, then $\preceq_v$ is the cyclic concatenation of the orders $\preceq_{a_1}, \ldots, \preceq_{a_k}$. See Figure 13. A perturbation of $\mathcal{C}$ restricts to a perturbation of each of its curves. Conversely, given a perturbation of each curve in $\mathcal{C}$ we can built a perturbation (in general many) of their union by merging for each arc a the left-to-right orders $\preceq_a$ of each curve perturbation.

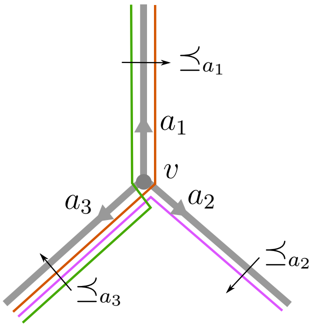

Figure 13: The circular order $\preceq_v$ is the concatenation of the orders $\preceq_{a_1}, \preceq_{a_2}, \preceq_{a_3}$.

Combinatorial crossings. Recall that a double point of a pair (c, d) of combinatorial closed curves is a pair of indices $(i, j) \in \mathbb{Z}/|c|\mathbb{Z} \times \mathbb{Z}/|d|\mathbb{Z}$ such that $c(i) = d(j)$, assuming $i \neq j$ when $c = d$. Let Π be a perturbation of c and d. A double point (i, j) of (c, d) is a **crossing**[5] in Π if the pairs of arc occurrences $([i-1, i]_c, [i, i+1]_c)$ and $([j-1, j]_d, [j, j+1]_d)$ are linked in the $\preceq_{c(i)}$-order, i.e. if they appear in the cyclic order

$$\cdots [i, i-1]_c \cdots [j, j-1]_d \cdots [i, i+1]_c \cdots [j, j+1]_d \cdots,$$

with respect to $\preceq_{c(i)}$ or the opposite order. An analogous definition holds for a self-crossing of a single curve, taking $c = d$ in the above definition. Note that the notion of (self-)crossing is independent of the traversal directions of c and d. The number of crossings of c and d and of self-crossings of c in Π is denoted respectively by $i_\Pi(c, d)$ and $i_\Pi(c)$.

We define the **combinatorial self-crossing number** of c, denoted by $i(c)$, as the minimum of $i_\Pi(c')$ over all the combinatorial perturbations Π of any combinatorial curve c' freely homotopic to c. The **combinatorial crossing number** of two combinatorial curves c and d is defined the same way taking into account crossings between c and d only.

[4] This definition assumes that the combinatorial surface is orientable and consistently oriented. In the general case, opposite arcs should be associated with inverse orders if the corresponding edge is untwisted and with the same order if the edge is half-twisted.

[5] The crossings of a perturbation should not be confused with the crossing double points of a combinatorial curve appearing as crossing double paths of zero length in Section 5.1. Which notion of crossings is used should always be clear from the context.

Lemma 27. *The combinatorial (self-)crossing number coincides with the geometric (self-)intersection number, i.e. for every combinatorial closed curves c and d on a combinatorial surface with graph G:*

$$i(c) = i(\rho(c)) \qquad \text{and} \qquad i(c,d) = i(\rho(c), \rho(d)).$$

where ρ is any cellular embedding of G.

Proof. Every combinatorial perturbation Π of c can be realized by a continuous curve $\gamma \sim \rho(c)$ with the same number of self-intersections as the number of self-crossings of Π. To construct γ we consider small disjoint disks centered at the images of the vertices of G in S. We connect those vertex disks by disjoint strips corresponding to the edges of G. For every arc a of G we draw inside the corresponding edge strip parallel curve pieces labelled by the occurrences of a or a^{-1} in c in the left-to-right order $\preceq_a$. The endpoints of those curve pieces appear on the boundary of the vertex disks in the circular vertex orders induced by Π. It remains to connect those endpoints by straight line segments inside each disk (via a parametrization over a Euclidean disk) according to the labels of their incident curve pieces. The resulting curve γ can be homotoped to $\rho(c)$ and its self-intersections may only appear inside the vertex disks. Clearly, an intersection of two segments of γ in a disk corresponds to linked pairs of arc occurrences, i.e. to a combinatorial crossing, and vice-versa. It follows that $i(c) \geq i(\rho(c))$. To prove the reverse inequality we show that for any continuous curve γ in general position there exists a combinatorial perturbation of a curve c' such that $\rho(c') \sim \gamma$ and the number of self-crossings of c' is at most the number of self-intersections of γ. By an isotopy we can enforce γ to stay in the neighborhood of G composed of the edge strips and vertex disks, and to have all its self-intersections inside the vertex disks. After removing the vertex disks from S, we are left with a set of disjoint and simple pieces of γ lying inside the edge strips. If a piece of curve in a strip has its endpoints on the same end of the strip we can further isotope the piece inside the incident vertex disk. This way all the curve pieces join the two ends of their strip and can be ordered from left-to-right inside each strip (assuming a preferred direction of the strip). Replacing each curve piece by an arc occurrence we thus define a combinatorial perturbation of a curve c' with the required properties. The same constructions apply to a perturbation of two curves showing that $i(c,d) = i(\rho(c), \rho(d))$. □

8 Computing a minimal immersion

For the rest of the paper we assume that **all considered system of quads are orientable** (and consistently oriented) and we only deal with the self-intersection number of a single curve. We thus drop the subscript c to denote an index path $[i \xrightarrow{\ell}]$ or an arc occurrence $[i, i+1]$. *We also write intersections for self-intersections.* By Theorem 1 we can compute the geometric intersection number of a curve efficiently. Here, we describe a way to compute a minimal immersion, that is an actual immersion with the minimal number of intersections. We refer to the framework of the previous Section 7 to describe such an immersion as a combinatorial perturbation. Thanks to Lemma 27, we know that this framework faithfully encodes the topological configurations.

8.1 Bigons and monogons

A **bigon** of a perturbation Π of c is a pair of index paths $([i \xrightarrow{\ell}], [j \xrightarrow{k}])$ whose **sides** $c[i \xrightarrow{\ell}]$ and $c[j \xrightarrow{k}]$ have strictly positive lengths, are homotopic, and whose **tips** (i, j) and $(i+\ell, j+k)$ are combinatorial crossings for Π. See Figure 14. A **monogon** of Π is an index path $[i \xrightarrow{\ell}]$ of strictly positive length such that $(i, i+\ell)$ is a combinatorial crossing and the image path $c[i \xrightarrow{\ell}]$ is contractible.

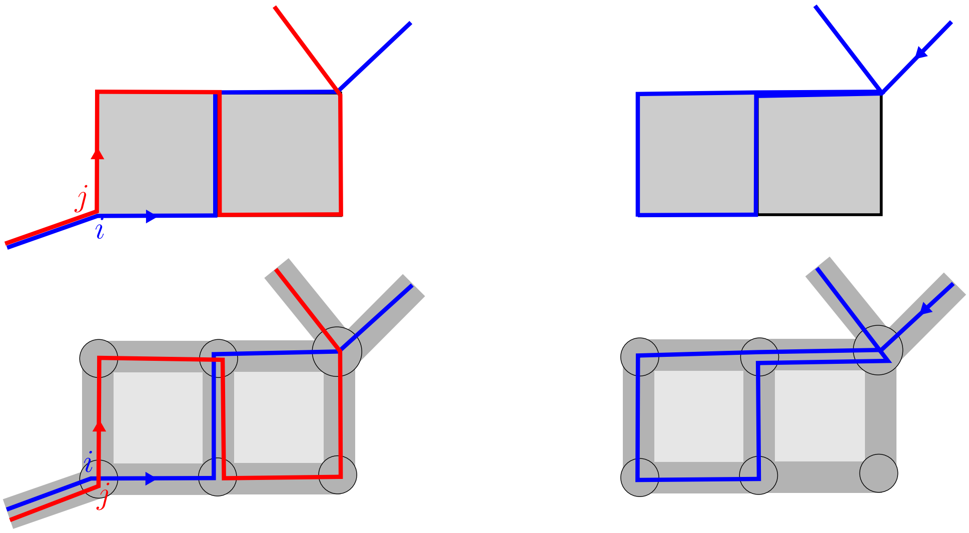

Figure 14: Left, a pair of paths in system of quads and a combinatorial perturbation with three bigons $([i \xrightarrow{2}], [j \xrightarrow{2}])$, $([i+2 \xrightarrow{1}], [j+2 \xrightarrow{3}])$ and $([i \xrightarrow{3}], [j \xrightarrow{5}])$. Right, a path and a combinatorial perturbation with a monogon of length 6.

Proposition 28. *A combinatorial perturbation of a primitive curve has excess self-crossing if and only if it contains a bigon or a monogon.*

Proof. We just need to prove the existence of a bigon or monogon for a continuous realization $\gamma : \mathbb{R}/\mathbb{Z} \to S$ of the combinatorial perturbation. This bigon or monogon corresponds to a combinatorial bigon or monogon as claimed in the Proposition. To prove the topological counterpart, we first note as in Section 3 that the minimal number of intersections is counted by the lifts of γ whose limit points alternate along $\partial\mathbb{D}$. It follows that γ is minimally crossing when its lifts are pairwise disjoint unless the corresponding limit points alternate in $\partial\mathbb{D}$ and they should intersect exactly once in that case. When γ is not minimally crossing there must exist two lifts intersecting more than necessary, thus forming a bigon in $\mathbb{D}$. The projection of such a bigon on S gives the desired result. Note that a statement similar to the one in the Proposition holds for a perturbation of a pair of primitive curves c and d with excess crossing. In Appendix A, we provide a purely combinatorial proof of the direct implication of the Proposition in the case of two curves. □

Suppose that Π is a combinatorial perturbation of a primitive geodesic c. It cannot have a monogon by Corollary 12. Hence, according to Proposition 28 the perturbation Π is minimally crossing unless it contains a bigon. One could eliminate such a bigon

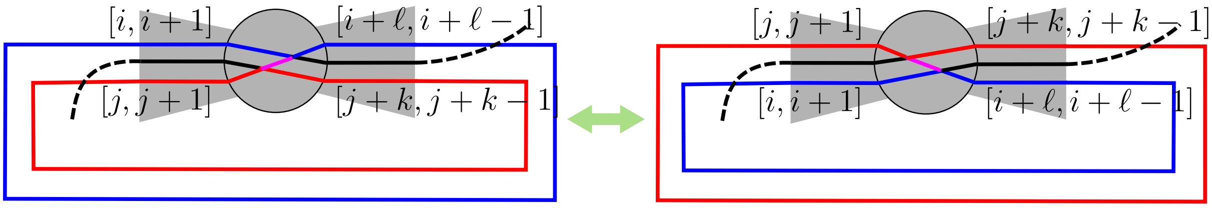

Figure 15: Right, the realization of the bigon $([i \xrightarrow{\ell}], [j \xrightarrow{k}])$ where $j = i + \ell$ and $[i, i-1]$ is in-between $[i+\ell, i+\ell-1]$ and $[j+k, j+k-1]$, and $[j+k, j+k+1]$ is in-between $[i, i+1]$ and $[j, j+1]$. The small purple part is at the same time the beginning of the red side of the bigon and the end of the blue side. Swapping this bigon does not reduce the number of crossings.

by swapping its sides. However, it may happen that the index paths of the bigon share some common part, which would make the swapping ambiguous. In agreement with the terminology of Hass and Scott [HS85], a bigon $([i \xrightarrow{\ell}], [j \xrightarrow{k}])$ is said **singular** if

1. its two index paths have disjoint interiors, i.e. they do not share any arc occurrence;
2. when $j = i + \ell$ the following arc occurrences
$$[i, i-1], [j, j-1], [i, i+1], [j+k, j+k+1], [j, j+1], [j+k, j+k-1]$$
do not appear in this order or its opposite in the circular ordering induced by Π at $c(j)$;
3. when $i = j + k$ the following arc occurrences
$$[i, i-1], [j, j-1], [i+\ell, i+\ell-1], [i, i+1], [i+\ell, i+\ell+1], [j, j+1]$$
do not appear in this order or its opposite in the circular ordering induced by Π at $c(i)$.

Figure 15 shows that when a bigon does not satisfy condition 2 or 3, swapping its sides does not reduce the number of crossings. Note that when c is geodesic and $k = -\ell$ there cannot be any identification between $\{i, i+\ell\}$ and $\{j, j-\ell\}$. For instance, $i = j$ implies $c[i \xrightarrow{\ell}] \sim c[j \xrightarrow{-\ell}] = c[i \xrightarrow{-\ell}]$ so that $c[i-\ell \xrightarrow{2\ell}] \sim 1$, while $j = i + \ell$ implies $c[i \xrightarrow{\ell}] \sim c[j \xrightarrow{-\ell}] = c^{-1}[i \xrightarrow{\ell}]$ so that $c^2[i \xrightarrow{\ell}] \sim 1$. In both cases c would have a monogon which would contradict Corollary 12.

When c is primitive and geodesic, we cannot have $j = i + \ell$ and $i = j + \ell$ at the same time. Otherwise, we must have $k = \ell$ by the preceding paragraph, and c would be a circular shift of $c[i \xrightarrow{\ell}] \cdot c[j \xrightarrow{\ell}]$ and thus homotopic to a square. We also remark that when one of the last two conditions in the definition is not satisfied, the bigon maps to a non-singular bigon in the continuous realization of Π as in Lemma 27. See Figure 15.

When the bigon is singular we can swap its two sides by exchanging the two arc occurrences $[i+p, i+p+1]$ and $[j+\varepsilon p, i+\varepsilon(p+1)]$ in the left-to-right order associated to their supporting arc, for $0 \leq p < \ell$ and $k = \varepsilon\ell$.

Lemma 29. *Swapping the two sides of a singular bigon of a perturbation of a geodesic primitive curve decreases its number of crossings by at least two.*

This is relatively obvious if one considers a continuous realization of the perturbation, performs the swapping and comes back to a combinatorial perturbation as in the proof of Lemma 27. We nonetheless provide a purely combinatorial proof in Appendix B.

Hence, by swapping singular bigons we may decrease the number of crossings until there is no more singular bigons. It follows from the next theorem that the resulting perturbation has no excess crossing.

Theorem 30 (Hass and Scott [HS85, Th. 4.2])**.** *A perturbation of a primitive geodesic curve has excess crossing if and only if it contains a singular bigon.*

Proof. We realize Π by a continuous curve γ with the same construction as in Lemma 27. By Theorem 4.2 in [HS85] the curve γ has a singular bigon. In turn this bigon corresponds to a singular bigon in Π. □

In the next section we describe how to detect bigons in practice.

8.2 Proof of Theorem 2

If a perturbation Π of a primitive geodesic c has a bigon then, by the same arguments as for a crossing pair of index paths in Section 6.1, the two sides label a disk diagram and their vertices can be put in 1-1 correspondence such that corresponding vertices stay at distance zero or two and bound a same quad in the disk diagram. Hence, we can look for bigons among the maximal partial diagrams of c. This leads to a quadratic algorithm for finding a bigon of Π, or deciding that Π has no bigon.

Proof of Theorem 2. By Theorem 8 we can assume that c is geodesic. We first consider the case where c is primitive. Let Π be a randomly chosen perturbation of c. We store the induced vertex orderings of the arc occurrences in arrays so that we can test if a double point is a crossing in constant time using pointer arithmetic. By Theorem 30, if Π has excess crossing it has a singular bigon. This singular bigon defines a thick double path that must appear in some maximal partial diagram of Π as two boundary paths with common extremities. Hence, in order to find a singular bigon we just need to scan in each maximal partial diagram the index pairs that are combinatorial crossings and test whether the two sides between any two consecutive such index pairs define a singular bigon. According to the proof of Corollary 22 in Section 6, the cost for scanning all the maximal partial diagrams is $O(\ell^2)$. By the very definition of a singular bigon, testing whether a pair of index paths define a singular bigon can be performed in constant time. It follows that we can compute a singular bigon, if any, in $O(\ell^2)$ time. Once a singular bigon is found we swap its sides in $O(\ell)$ time. This preserves the geodesic character and, by Lemma 29, the number of crossings is reduced by at least two. Since Π may have $O(\ell^2)$ excess crossings, we need to repeat the above procedure $O(\ell^2)$ times before the curve has no more singular bigon, hence is minimally crossing. This concludes the proof of the theorem in the case of primitive curves. When $c = d^p$, with d primitive, we first find a minimally crossing perturbation for d by the above procedure. We further traverse the perturbation p times duplicating d as many times. As we start each traversal we connect the last arc occurrence of the previously traversed copy with the first occurrence of the next copy. We continue this copy by duplicating each traversed arc occurrence to its right. It is easily seen that the number of crossings of the final perturbation of c, after connecting the last traversed arc with the first one, satisfies the formula in Proposition 26. This concludes the proof for non-primitive curves. □

9 Detecting simple curves

As noticed in Section 3, geodesics on a hyperbolic surface possess a minimal number of self-intersections in their homotopy class but this is no longer true for combinatorial geodesics. In particular, it is easy to construct examples of combinatorial geodesics homotopic to simple curves that nonetheless exhibit crossings. This remains true even if we assume the geodesic to be in canonical form. In Section 9.2, we present a simple and efficient algorithm to recognize curves homotopic to simple curves. Given a combinatorial curve, this algorithm constructs a combinatorial perturbation of some homotopic geodesic. The aim is to obtain, if possible, a perturbation without crossings. After computing the canonical form of the given curve, the construction starts with the perturbation of an empty path and adds the arcs of the canonical geodesic incrementally, each time extending the path perturbation. In order to ensure at each step that this perturbation has no crossings we may need to transform the canonical geodesic into another homotopic geodesic. Those transformations amounts to unzipping the potential bigons of the canonical geodesic. Lemma 36 states that this unzip algorithm can be implemented to run in $O(\ell \log \ell)$ time, where ℓ is the length of the input curve. Although the returned geodesic need not be in canonical form, the algorithm strongly requires that the given curve be in canonical form. The rest of Section 9.2 is devoted to the proof of Proposition 38 asserting that the perturbation returned by the unzip algorithm has no combinatorial crossings if and only if the input curve is homotopic to a simple curve. The main argument is to show that if the unzip algorithm fails to produce a perturbation without crossings then one of its crossings gives rise to a pair of lifts with alternating limit points. This certifies that the crossing number of the input curve is at least one.

9.1 Some preparatory lemmas for Theorem 3

Before studying the recognition of curves homotopic to simple curves, we state some refinements of Lemma 21 and its Corollary that will be useful for the proof of Theorem 3. We first highlight a simple property of the system of quads.

Lemma 31. *The facial walk of a quad cannot contain an arc twice, either with the same or the opposite orientations.*

Proof. If the facial walk of a quad contains two occurrences of an arc with the same orientation, then their identification creates a Möbius strip in the system of quads, in contradiction with its orientability. If the facial walk contains two consecutive occurrences of an arc with opposite orientations, then their common endpoint must have degree one in the system of quads. This contradicts the minimal degree (8 on a closed orientable surface) in our system of quads. Finally, if the facial walk contains non consecutive occurrences of an arc with opposite orientations, then their identification creates a cylinder bounded by loop edges. However, the radial graph being bipartite, this cannot occur. □

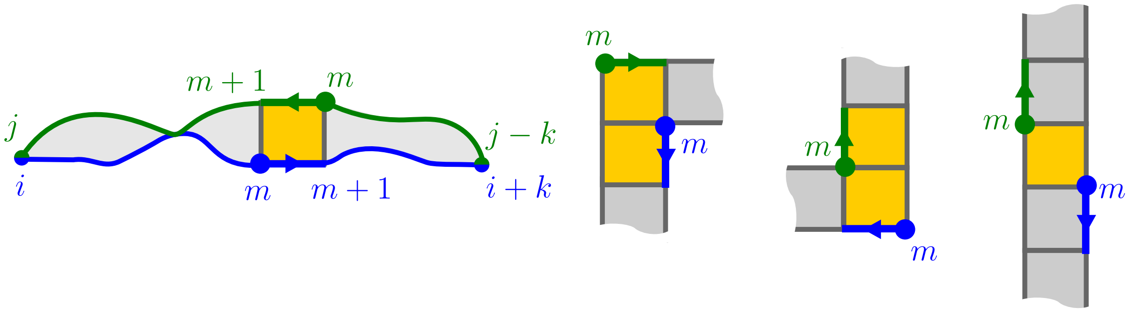

Figure 16: When the index pair (m, m) corresponds to distinct vertices (the thick m dots) in the diagram, the arc $c[m, m+1]$ labels two arcs that may be in one of four possible configurations. In each configuration the yellow quads must be bounded twice by the same arc, contradicting Lemma 31.

Lemma 32. *Let c be a primitive geodesic curve. If a forward index path $[i \xrightarrow{k}]$ and a backward index path $[j \xrightarrow{-k}]$ have homotopic image paths $c[i \xrightarrow{k}] \sim c[j \xrightarrow{-k}]$, then they cannot share any index.*

Proof. We remark that we cannot have $i = j$ for otherwise $c[j-k \xrightarrow{2k}] = (c[j \xrightarrow{-k}])^{-1}.c[i \xrightarrow{k}] = 1$ in contradiction with Corollary 12. Likewise, we cannot have $i+k = j-k$. By way of contradiction, suppose that the index paths $[i \xrightarrow{k}]$ and $[j \xrightarrow{-k}]$ have a common index $i + r = j - t$ for some integers $0 \leq r, t \leq k$. By the previous remark we have $0 < \frac{r+t}{2} < k$. It follows that (m, m) is an interior index pair of $([i \xrightarrow{k}], [j \xrightarrow{-k}])$, where $m = i + \frac{r+t}{2} = j - \frac{r+t}{2}$ (recall that i and j have the same parity). Consider a disk diagram for $c[i \xrightarrow{k}] \sim c[j \xrightarrow{-k}]$. The index pair (m, m) cannot correspond to the same vertex in the diagram since the path $c[i, j] = c[i, m].c[m, j]$ would then label a closed path in the diagram, hence be contractible. This would again contradict Corollary 12. It ensues that (m, m) corresponds to opposite vertices in a quad of the diagram. This quad is part of a staircase where the arc $c[m, m+1]$ labels two arcs on either sides of the partial diagram. Figure 16 depicts the four possible configurations. In each case we infer the existence of a quad bounded twice by the same arc, in contradiction with the previous lemma. □

Lemma 33. *Two distinct index paths of a nontrivial primitive geodesic curve c having homotopic image paths have length distinct from $|c|$ and at most $|c| + 1$.*

Proof. When the index paths have opposite orientations, the previous lemma directly implies the result. Put $\ell = |c|$ and let $[i \xrightarrow{k}], [j \xrightarrow{k}]$ be two index paths such that $c[i \xrightarrow{k}] \sim c[j \xrightarrow{k}]$. Clearly, $k \neq \ell$ since otherwise c would be homotopic to a conjugate of itself by the shorter curve $c[i, j]$, in contradiction with the primitivity of c. For the sake of a contradiction, suppose that $k > \ell + 1$. Consider a diagram Δ with sides Δ_R and Δ_L for the homotopic paths $c[i \xrightarrow{k}] \sim c[j \xrightarrow{k}]$. By Lemma 21 the vertices $\Delta_R(i+\ell)$ and $\Delta_L(j+\ell)$ must be diagonally opposite in a quad of Δ. Using that $\Delta_R[i \xrightarrow{2}]$ and $\Delta_R[i+\ell \xrightarrow{2}]$ are labelled by the same edges and similarly for $\Delta_L[j \xrightarrow{2}]$ and $\Delta_L[j+\ell \xrightarrow{2}]$ we easily deduce that the system of quads has a quad with two occurrences of a same arc in its facial walk, or has a vertex of degree at most 5, or is non-orientable. This would contradict

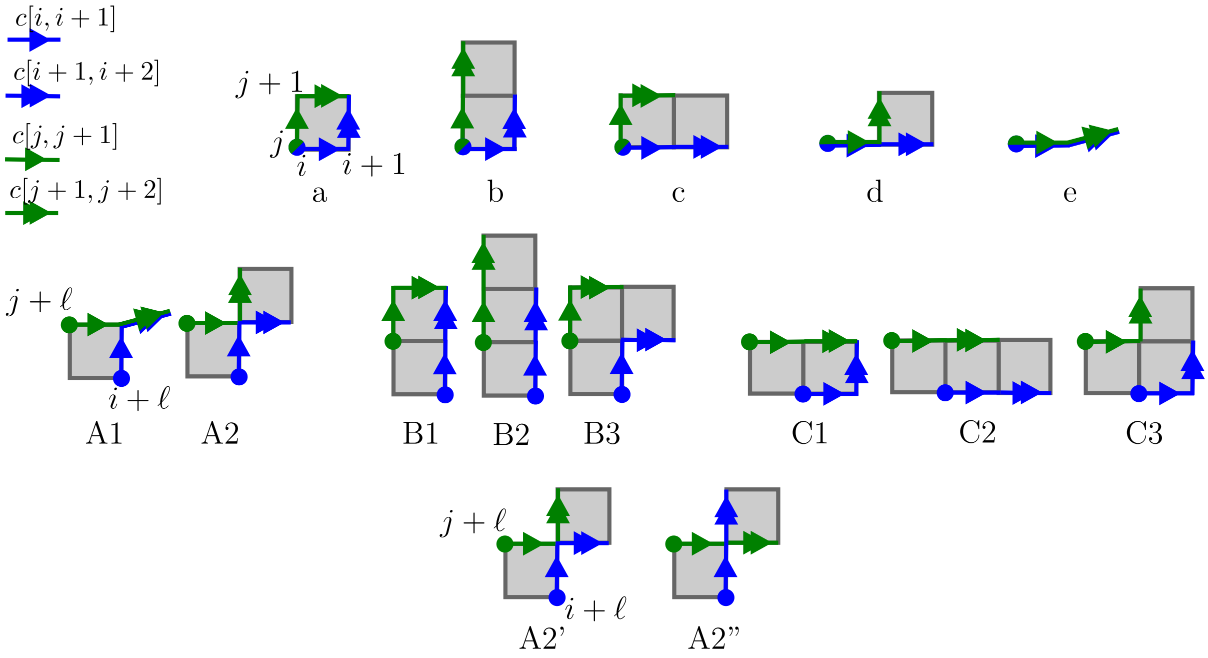

Figure 17: Upper row, the five possible configurations at the beginning of the disk diagram for $c[i \xrightarrow{k}] \sim c[j \xrightarrow{k}]$. Middle row, the eight possible configurations in the neighborhood of the index pair $(i+\ell, j+\ell)$ in this diagram. Lower row, case A2 splits into two subcases as we fix the relative orientations of the two quads.

Lemma 31 or the hypotheses on the system of quads. In details, Figure 17 depicts the five possible configurations (a,b,c,d,e) for $\Delta_R[i \xrightarrow{2}]$ and $\Delta_L[j \xrightarrow{2}]$ in the diagram Δ and the eight possible configurations (A1-2, B1-3, C1-3) for $\Delta_R[i+\ell \xrightarrow{2}]$ and $\Delta_L[j+\ell \xrightarrow{2}]$. We remark that cases C1-3 are symmetric to cases B1-3, so that we only need to consider the five configurations A1-2, B1-3. We denote by (x,Y) $\in$ {a,b,c,d,e} × {A1-2, B1-3} the conjunction of configurations x and Y in Δ. The quads in configurations a,b,c,d are part of a staircase σ_0 in Δ. Similarly, the vertices $\Delta_R(i+\ell)$ and $\Delta_L(j+\ell)$ are contained in a staircase σ_ℓ in Δ (possibly the same). Being drawn in the plane the staircases of Δ inherit an orientation from the plane. This orientation is said to be positive or negative according to whether or not it is consistent with some default orientation of the quad system.

- If the orientations of the two staircases σ_0 and σ_ℓ have the same sign, then the face to the right of $\Delta_L[j, j+1]$ has the same facial walk as the face to the right of $\Delta_L(j+\ell, j+\ell+1)$. The induced identifications imply that for (x,Y) $\in$ {a,b,c,d} × {A1-2, B1-3} $\cup$ {e} × {A1-2,B1} one quad is bounded twice by the same arc (with or without the same orientation) in contradiction with Lemma 31. If (x,Y) $\in$ {e} × {B2-3}, we easily deduce that $c(i+1) = c(j+1)$ has degree four in the system of quads, again a contradiction.

- When the orientations of σ_0 and σ_ℓ have opposite signs the faces to the right of $\Delta_L[j, j+1]$ and to the right of $\Delta_L(j+\ell, j+\ell+1)$ (resp. to the left of $\Delta_R[i, i+1]$ and to the left $\Delta_R[i+\ell, i++\ell+1]$) correspond to the two faces incident to $c[j, j+1]$ (resp. $c[i, i+1]$). We further split case $A2$ into two variants A2$'$ and A2$''$ according to whether the two quads (see Figure 17) have orientations with the same sign or not, respectively. By the induced identifications we derive the following forbidden situations: two occurrences of a same arc in a quad for (x,Y) $\in$ {(a,A1),(e,B1)} $\cup$ {d,e} $\times$ {A1,A2$'$} $\cup$ {a-e} $\times$ {A2$''$}, a degree two vertex (namely $c(j+1)$) for (x,Y) $\in$ {a,c} $\times$ {B1,B3}, a degree three vertex for (x,Y) $\in$ {b} $\times$ {B1-3} $\cup$ {a-c} $\times$ {B2}, a degree four vertex for (x,Y) $\in$ {b,c} $\times$ {A1} $\cup$ {e} $\times$ {B2-3} $\cup$ {(a,A2$'$),(d,B1)} and a degree five vertex for (x,Y) $\in$ {b,c} $\times$ {A2$'$} $\cup$ {d} $\times$ {B2-3}.

□

When the orientations of the staircases σ_0 and σ_ℓ in the above proof have the same sign it is sufficient to use that $\Delta_R[i \xrightarrow{1}]$ and $\Delta_R[i+\ell \xrightarrow{1}]$ are labelled by the same edge, and similarly for $\Delta_L[j \xrightarrow{1}]$ and $\Delta_L[j+\ell \xrightarrow{1}]$, in order to reach a contradiction for cases $A1$ and $A2$. In other words, the two homotopic paths $c[i \xrightarrow{k}] \sim c[j \xrightarrow{k}]$ cannot have length $\ell+1$. This leads to the following refinement.

Lemma 34. *A bigon $([i \xrightarrow{k}], [j \xrightarrow{k}])$ of a perturbation of a primitive geodesic curve c without intermediate crossings (i.e., no index pair $(i+r, j+r)$, $1 < r < k$, is a combinatorial crossing) has length $k < |c|$.*

Lemma 35. *If two forward index paths of a canonical curve have homotopic image paths then they actually have the same image paths, i.e. they form a double path. In particular, bigons composed of two forward paths must be flat.*

Proof. By Theorem 10 the image paths bound a disk diagram composed of paths and staircase. Remark that a staircase with both sides directed forward must have a $\bar{1}$ turn on its left side. Since the curve is canonical we infer that the diagram cannot have any staircase, implying that the two sides coincide. □

9.2 The unzip algorithm

We now turn to the original problem of Poincaré [Poi04, §4], deciding whether a given curve c is homotopic to a simple curve. In the affirmative we know by Lemma 27 that some geodesic homotopic to c must have a (combinatorial) **embedding**, i.e. a perturbation without crossings. This also results from Lemma 29 and Theorem 30 by swapping bigons as much as possible. Rather than swapping the sides of a singular bigon as in Lemma 29 we choose to switch one side along the other side. This will also decrease the number of crossings if the bigon contains no other interior bigons. This time, however, one side of the bigon is left unchanged and this allows us to enforce a given edge of c to stay fixed as we remove crossings: each time the edge is involved in a bigon switch, we switch the side that does not contain the edge. This suggests an incremental computation of an embedding in which a larger and larger part of the curve is fixed as we switch bigons: we assume that c is canonical and consider the trivial embedding of its first arc occurrence $[0,1]$. We next insert the successive arc occurrences incrementally to maintain, if possible, an embedding of the path formed by the already inserted arcs. Note that given an embedding of the subpath $c[0 \xrightarrow{i}]$, either this embedding cannot be extended to an embedding of $c[0 \xrightarrow{i+1}]$, or there is exactly one way to do so. When inserting the occurrence $[i, i+1]$ we need to compare its left-to-right order with each already inserted arc occurrence β of its supporting arc.

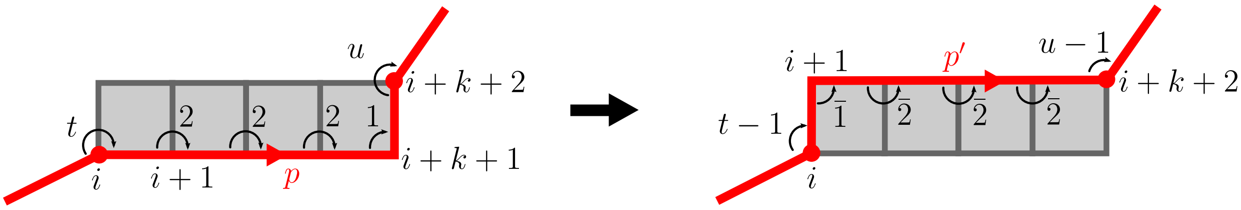

Figure 18: The arc $[i, i+1]$ is switchable.

- If $\beta \neq [0,1]$ we can use the comparison of the occurrence $[i-1,i]$ with the occurrence γ preceding β (or succeeding β if it is a backward occurrence). If $[i-1,i]$ and γ have the same supporting arc, we just propagate their relative order to $[i,i+1]$ and β. Otherwise, we use the circular ordering of the supporting arcs of $[i-1,i]$, γ and $[i,i+1]$ to deduce the relative order of $[i,i+1]$ and β that avoids a crossing at the double point (i,j), where j is the common index of β and γ.

- When $\beta = [0,1]$, we cannot use the occurrence preceding $[0,1]$ as it is not yet inserted. We rather compare $[i,i+1]$ and $[0,1]$ as follows. In the Poincaré disk, we consider two lifts $\tilde{d}_i$ and $\tilde{d}_0$ of c such that $\tilde{d}_i[i,i+1] = \tilde{d}_0[0,1]$. We decide to insert $[i,i+1]$ to the left (right) of $[0,1]$ if one of the limit points of $\tilde{d}_i$ lies to the left (right) of $\tilde{d}_0$. Note that when c is homotopic to a simple curve the two limit points of $\tilde{d}_i$ should lie on the same side of $\tilde{d}_0$.

After comparing $[i,i+1]$ with all the occurrences of its supporting arc, we can insert it in the correct place. If no crossings were introduced this way, we proceed with the next occurrence $[i+1,i+2]$. It may happen, however, that no matter how we insert $[i,i+1]$ in the left-to-right order of its supporting arc, the resulting perturbation of $c[0 \overset{i+1}{\rightarrow}]$ will have a combinatorial crossing. In order to handle this case, we first check if $[i,i+1]$ is **switchable**, i.e. if for some $k \geq 0$ and some turns t,u the subpath $p := c[i \overset{k+2}{\rightarrow}]$ has turn sequence $t2^k1u$ and the index path $[i \overset{k+2}{\rightarrow}]$ does not contain the arc occurrence $[0,1]$. See Figure 18. When $[i,i+1]$ is switchable we can switch p to a new subpath p' with turn sequence $(t-1)\bar{1}\bar{2}^k(u-1)$ such that p and p' bound a diagram composed of a single horizontal staircase. We indeed perform the switch *if some intersection is actually avoided this way*. More formally, this happens when there is a crossing (i,j) such that the turn of $(c[j,j+\varepsilon], c[i,i+1])$ is one with either $\varepsilon = 1$ or $\varepsilon = -1$ as illustrated on Figure 19. In such a case we say that $[j,j+\varepsilon]$ has **triggered** the switch. Note that replacing p by p' leads to a curve c' that is still geodesic and homotopic to c. Moreover, *the non traversed part of c' starting at index $i+1$ is in canonical form* as it contains no $\bar{1}$ turns. We then insert the arc occurrence $[i,i+1]$ and proceed with the algorithm using c' in place of c. The successive switches in the course of the computation untangle c incrementally. We call our embedding procedure the **unzip algorithm**. For further reference we make some observations that easily follows from the above switch procedure and the fact that c is initially canonical.

Remark 3. If $[i,i+1]$ is switchable but not switched at the time of being processed, then it remains not switched until the end of the algorithm, meaning that the vertex of index i is directly followed by a subpath p with a turn sequence of the form 2^k1 (k may get

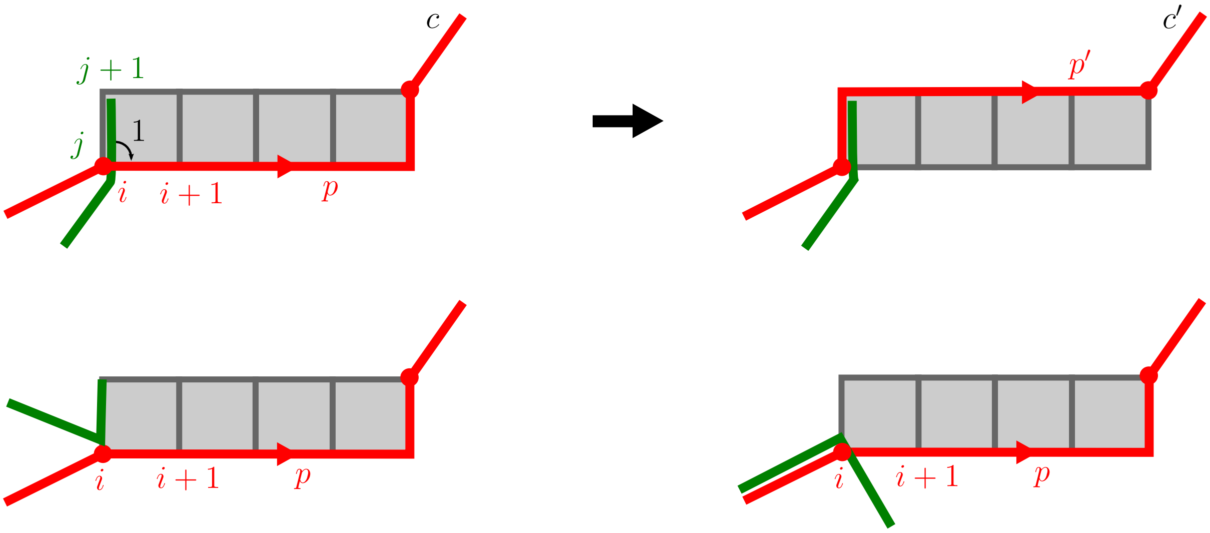

Figure 19: Top, the arc occurrence $[j, j+1]$ triggers a switch. Bottom, $[i, i+1]$ is switchable but not switched in these two situations.

smaller if some other arc of p is switched). Similarly, if $[i, i+1]$ is not switchable because of an inappropriate turn sequence, this remains true until the end of the algorithm.

Remark 4. Let c' be a curve homotopic to c obtained after a certain number of switches during the execution of the unzip algorithm. Then c' cannot contain a subpath with turn sequence $\bar{1}\bar{2}^*$ that ends at index 0, nor can it contain a subpath with turn sequence $\bar{2}^*\bar{1}$ that starts at index 0.

Lemma 36. *The unzip algorithm applied to a canonical primitive curve c of length ℓ can be implemented to run in $O(\ell \log \ell)$ time.*

Proof. We shall describe an implementation of the unzip algorithm with the announced time complexity. This algorithm constructs a combinatorial perturbation of c incrementally adding its arcs one after the other. The proof that this perturbation is an embedding if and only if c is homotopic to a simple curve is deferred to Proposition 38. Before we start the incremental construction, we need to perform some preprocessing, namely marking the switchable arcs of c and precomputing a left-to-right comparison between the first arc of c and the other occurrences of that arc in c.

Marking the switchable arcs. We first traverse c in reverse direction to mark all the switchable arcs. In the course of the algorithm, each time a switch applies to a subpath p we unmark all the arcs in p except the last one that may become switchable (depending on the last turn of p and on the status of the arc following p). It easily follows that an arc can be switched at most twice and that the amortized cost for the switches is linear. Alternatively, we could use the run-length encoded turn sequence of c as defined in [EW13] to detect each switchable arc and update the turn sequence in constant time per switches.

Comparisons with the first arc of c. In the preprocessing phase we also compute the relative order of $[0, 1]$ with all the other occurrences of $c[0, 1]$ as follows. If $c[i, i+1] = c[0, 1]$, the corresponding arc occurrences form a double path of length one and we compute their relative order by extending maximally this double path in the backward direction. Looking at the tip of this double path, say (j, k), we can decide which side is to the left of the other. Indeed, the three arcs $c[j, j-1]$, $c[k, k-1]$ and $c[j, j+1] = c[k, k+1]$ must be pairwise distinct and their circular order about their common origin vertex in the system of quads provides the necessary information as follows from Lemma 35. The computation of the maximal extensions in the backward direction amounts to evaluating the longest common prefix of c^{-1} with all its circular shifts. Overall, this can be done in $O(\ell)$ time thanks to a simple variation of the Knuth-Morris-Pratt algorithm.

The incremental construction. Recall that a combinatorial perturbation of c is the data for each arc of the system of quads of an ordering of the arc occurrences of c equal to that arc or its opposite. In practice, we store each of these orderings into a (balanced) binary search tree to allow the left-to-right comparison of arcs in logarithmic time. Although an edge is made of two opposite arcs, we shall store only one ordering per edge, assuming a default orientation of the edge. The orderings associated to opposite arcs should indeed be opposite, so that there is no need to store both of them.

We start the incremental construction of the combinatorial perturbation by initializing an empty search tree for every edge of the system of quads. We now traverse c in the forward direction starting with $[0, 1]$ and insert each traversed arc occurrence $\alpha = [i, i+1]$ in its tree, possibly after switching α. The switches may modify c so that we denote by c_k the geodesic homotopic to c resulting from the first switches in the algorithm up to and including the insertion of the arc occurrence $[k, k+1]$. The insertion of α is composed of two steps. We check in a first step if α needs to be switched, that is if α is switchable and some already inserted arc occurrence indeed triggers its switch. This last condition happens when one of the occurrences of the arc that makes a one turn with $c_{i-1}(\alpha)$ defines a crossing with α at their common endpoint (see Figure 19). This can be easily tested in $O(\log \ell)$ time using a dichotomy over these occurrences. We perform the switch if the test is positive and obtain a possibly new curve c_i. In a second step we insert α in the tree of its supporting arc $c_i(\alpha)$. If this search tree is empty, we just insert α at the root. Otherwise, we perform the usual tree insertion; as we go down the tree we need to perform $O(\log \ell)$ comparisons, each time comparing α with some already inserted arc occurrence β such that $c_i(\beta) = c_i(\alpha)$. There are three cases to consider.

- If $\beta \neq [0, 1]$ we can use the comparison of the occurrence $[i-1, i]$ preceding α with the occurrence β' preceding β (or succeeding β if it is a backward occurrence) in order to decide the relative order of α and β. If moreover $c_i(\beta') \neq c_i[i-1, i]$ this comparison can be done in constant time: we can index the arcs around each vertex of the system of quads in a prepossessing step and use index arithmetic to make the necessary comparisons between the supporting arcs of β', $[i-1, i]$ and α.

- If $\beta \neq [0, 1]$ as above, while β' and $[i-1, i]$ have the same supporting arc, we can use the search tree of $c_i(\beta')$ to perform the comparison. This would cost $O(\log \ell)$ time for each such β', implying a $O(\log^2 \ell)$ cost for inserting α in its tree. However, we can gather the comparisons of α with all such β''s into a single search. Indeed, if the computed perturbation for the first $i-1$ arc occurrences is an embedding, then the β''s must be contiguous in the left-to-right order of $c_i([i-1, i])$ and must appear in the same order as the corresponding β's in the search tree of $c_i(\alpha)$. It follows that

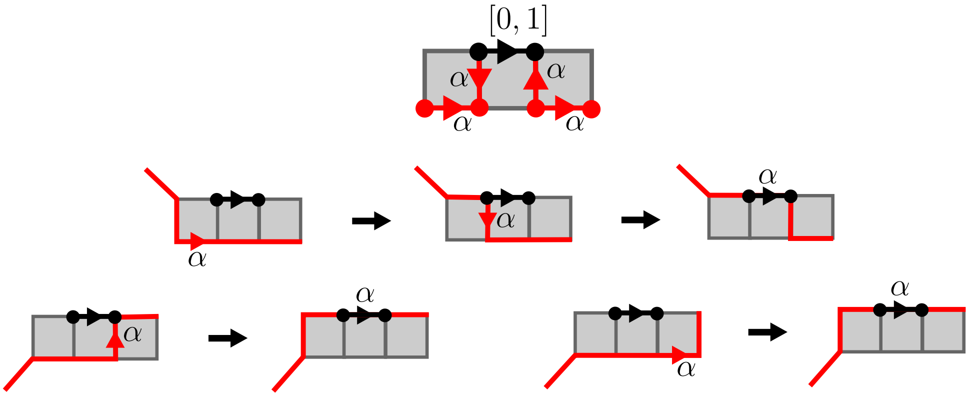

Figure 20: Top: the arc $c(\alpha)$ may be in one of four possible configurations with respect to the first arc $c[0,1]$. Middle and bottom row: from its initial canonical position, α is switched to the same arc as $[0,1]$.

we can determine the relative order of α with all the β's thanks to a single search of $[i-1,i]$ in the search tree of $c_i([i-1,i])$. The overall cost is $O(\log \ell)$. If the computed perturbation for $[0 \xrightarrow{i}]$ happens to be non-simple, then the β''s need not be contiguous and this procedure may return an incoherent answer. However, this would certify that the unzip algorithm has already failed to produce an embedding and we need not worry about incoherent answers.

- When $\beta = [0,1]$ we can use our precomputed comparisons unless α was previously switched, thus not compared with $[0,1]$ during the preprocessing phase. Since c is canonical, arcs may only be switched to their left in one of four possible configurations as on Figure 20 and we infer that α must lie to the right of $[0,1]$ as justified by Lemma 37 below.

In conclusion, α can be inserted in its search tree in $O(\log \ell)$ time. Inserting each arc occurrence in turn thus leads to an $O(\ell \log \ell)$ time algorithm.

Note that after α is inserted we do not try to determine if the current perturbation has a crossing or not. This will be checked in a second step after the unzip algorithm is completed. □

Lemma 37. *Let $\tilde{c}$ and $\tilde{d}$ be two lifts of a primitive canonical geodesic in the Poincaré disk $\mathbb{D}$. The limit points of $\tilde{c}$ cut the boundary circle $\partial\mathbb{D}$ into two pieces. By the* right piece, *we mean the piece of $\partial\mathbb{D}$ that bounds the part of $\mathbb{D} \setminus \tilde{c}$ to the right of $\tilde{c}$. Suppose that $\tilde{c}$ contains an arc a and that $\tilde{d}$ contains an arc b such that a and b are in one of the four relative positions depicted on Figure 21. Then one of the limit points of $\tilde{d}$ is in the right piece of $\partial\mathbb{D}$.*

Proof. We first note that b itself lies to the right of $\tilde{c}$. Indeed, $\tilde{c}$ would have to use a $\bar{1}$ turn to see b on its left or to pass along b, in contradiction with its canonicity. If both limit points of $\tilde{d}$ were to the left of $\tilde{c}$, then $\tilde{c}$ and $\tilde{d}$ would form a bigon with one staircase part containing b. The $\tilde{c}$ side of this staircase would thus see b on its right, which is impossible since $\tilde{c}$ is canonical (here, we have not used that $\tilde{d}$ is canonical but only geodesic). □

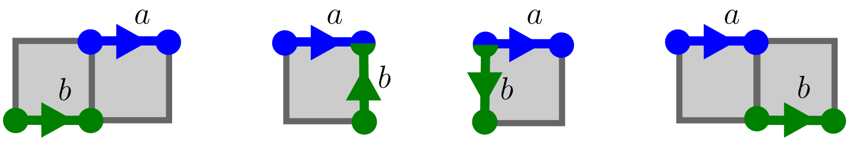

Figure 21: The relative positions of a and b.

Proposition 38. *If $i(c) = 0$ the unzip algorithm returns an embedding of a geodesic homotopic to c.*

Proof. Let Π be the perturbation computed by the unzip algorithm. We denote by c_k the geodesic homotopic to c resulting from the first switches in the algorithm up to and including the insertion of the arc occurrence $[k, k+1]$. Note that $k > r$ implies $c_k[0 \xrightarrow{r+1}] = c_r[0 \xrightarrow{r+1}]$. Suppose that Π has a crossing. For the rest of this proof, we denote by i the smallest index such that the insertion of $[i, i+1]$ creates a crossing, i.e. the restriction of Π to $[0 \xrightarrow{i+1}]$ has a crossing double point while its restriction to $[0 \xrightarrow{i}]$ is an embedding. By convention we set $i = \ell$ if the crossing appeared after the last arc insertion. We shall show that $c_{\ell-1}$ has two lifts whose limit points are alternating on the boundary of the Poincaré disk. It will follow from Section 3 that $i(c) = i(c_{\ell-1}) > 0$ thus proving the Proposition. We first establish some preparatory claims.

Claim 1. *If (i, j) is a crossing of Π with $0 < j < i < \ell$, then the backward arc $c_i[j, j-1]$ and the forward arc $c_i[j, j+1]$ are distinct from the supporting arc $c_i[i, i+1]$ of $[i, i+1]$. Moreover, if $[i, i+1]$ was switched just before its insertion, then $c_i[j, j-1]$ and $c_i[j, j+1]$ are also distinct from the supporting arc $c_{i-1}[i, i+1]$ of $[i, i+1]$ just before its switch.*

Proof. The first part of the claim follows directly from our insertion procedure and the fact that the restriction of Π to $[0 \xrightarrow{i}]$ is an embedding. Moreover, assume that $[i, i+1]$ was switched just before its insertion. By the insertion procedure, this means that the insertion of $[i, i+1]$ before the switch would have induced a removable crossing (i, k) where $0 < k < i$ and both $c_{i-1}[k, k+1]$ and $c_{i-1}[k, k-1]$ are distinct from $c_{i-1}[i, i+1]$ by the first part of the claim. By the same first part we have $c_i[j, j-1], c_i[j, j+1] \neq c_i[i, i+1]$. Since the restriction of Π to $c_i[0 \xrightarrow{i}]$ has no crossing, the length 2 path $c_{i-1}[k-1 \xrightarrow{2}] = c_i[k-1 \xrightarrow{2}]$ separates $c_i[j-1 \xrightarrow{2}]$ from $c_{i-1}[i, i+1]$, implying $c_i[j, j-1], c_i[j, j+1] \neq c_{i-1}[i, i+1]$ as desired. Figure 22 depicts the situation. □

Claim 2. *If c_k has a $\bar{1}$ turn at index $k+1$, with $0 < k < i-1$, then there is an index r with $0 < r < k$ such that*

- *$c_k[r, r-1] = c_k[k, k+1]$,*
- *the arc $c_k[r, r-1]$ lies to the right[6] of $c_k[k, k+1]$,*
- *if $r \neq 1$, then c_k has a 1 turn at $r-1$.*

Proof. Refer to Figure 22 for the claim and its proof. A $\bar{1}$ turn can only occur at the

[6] More formally, the arc occurrence $[r, r-1]$ lies to the right of $[k, k+1]$ in the restriction of Π to $[0 \xrightarrow{k+1}]$.

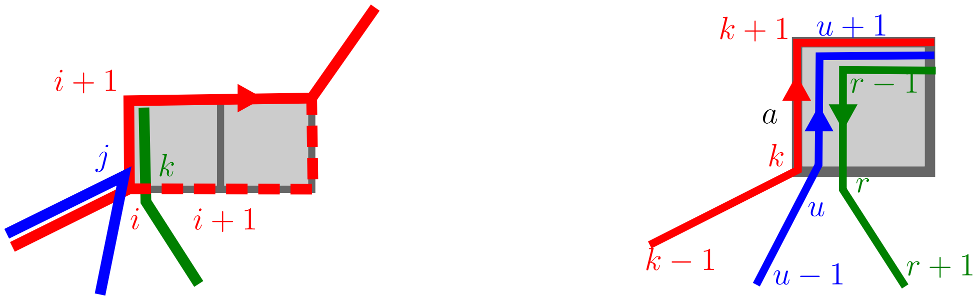

Figure 22: Left, illustration for the proof of claim 1 when $[i, i+1]$ was switched just before its insertion. c_i and c_{i-1} appear respectively in plain and doted lines. Right, illustration for claim 2.

destination of an arc that has been switched. It follows that the arc $c_k[k, k+1]$ must have been switched just before its insertion, thus witnessing the existence of a triggering arc a of the form $c_k[u, u-1]$ or $c_k[u, u+1]$, with $0 < u < k$, that lies parallel to and to the right of $c_k[k, k+1]$ (with respect to Π). In the case $a = c_k[u, u-1]$, we may set $r = u$. Otherwise, $a = c_k[u, u+1]$ and recalling that the restriction of Π to $[0 \xrightarrow{i}]$, hence to $[0 \xrightarrow{k+2}]$, has no crossing[7], it must be that $c_k[u \xrightarrow{2}]$ also lies parallel to and to the right of $c_k[k \xrightarrow{2}]$. So c_i has a $\bar{1}$ turn at index $u+1$, and we are back to the hypothesis of the claim, decreasing the value of k to u. We can repeat the same arguments inductively, each time decreasing the value of k in the claim. Since $k \geq 1$, the process must stop, implying that we have reached the former case. It remains to note that whenever $r > 1$ the double point $(k+1, r-1)$ cannot be a crossing by the choice of i, so that c_k makes a 1 turn at $r-1$. □

Recall that i is the smallest index such that the restriction of Π to $[0 \xrightarrow{i+1}]$ has a crossing. We denote by $c' = c_{\ell-1}$ the geodesic homotopic to c resulting from all the switches in the course of the algorithm execution. Let j be the minimal index such that (i, j) is a crossing of Π. We consider two lifts $\tilde{d}_i$ and $\tilde{d}_j$ of c' in the Poincaré disk such that $\tilde{d}_i(i) = \tilde{d}_j(j)$. We first suppose $i < \ell$.

Claim 3. *The piece of lift $\tilde{d}_i[i \xrightarrow{+\infty}]$ has no crossing with $\tilde{d}_j$. (Crossings are defined with respect to the lift of Π in the Poincaré disk.)*

Proof. For the sake of a contradiction, suppose that $\tilde{d}_i[i \xrightarrow{+\infty}]$ and $\tilde{d}_j$ cross. Let s be the smallest positive integer such that $\tilde{d}_i$ and $\tilde{d}_j$ crosses at $\tilde{d}_i(i+s)$. We thus have a bigon of the form $([i \xrightarrow{s}], [j \xrightarrow{\varepsilon s}])$ for some $\varepsilon \in \{-1, 1\}$. Moreover, this bigon has no intermediate crossings by the choice of s. Let Δ be a disk diagram for this bigon, oriented consistently with the system of quads. Δ must start with a staircase part by Claim 1. In particular, the turn t between $c'[j, j+\varepsilon]$ and $c'[i, i+1]$ should be ± 1. The unzip algorithm may have run across four possible situations at step i.

1. Either $[i, i+1]$ was switchable just before its insertion but was not switched,

[7] When $i = \ell$, a crossing of the form $(0, j)$ is not considered as a crossing of the restriction of Π to $[0 \xrightarrow{\ell}]$.

2. or it was switchable and switched,
3. or it was not switchable because of an inappropriate turn sequence,
4. or it was not switchable because the part to be switched contains $[0, 1]$.

In the first situation, we know by Remark 3 that $c'(i)$ is followed by a turn sequence of the form 2^*1. Hence, t is exactly 1; but this contradicts the fact that $[i, i+1]$ was not switched though switchable. Indeed, $[j, j+\varepsilon]$ should have triggered the switch. The third situation together with Remark 3 also lead to a contradiction as the inappropriate turn sequence prevents $[i \overset{s}{\to}]$ from being part of any staircase. Thanks to Claim 1, the second situation equally prevents $[i \overset{s}{\to}]$ from being part of any staircase. It remains to consider the fourth situation. We first suppose that $\varepsilon = -1$, i.e. that the diagram $\varDelta$ corresponds to the bigon $([i \overset{s}{\to}], [j \overset{-s}{\to}])$. By Lemma 32, we have $s < \ell$ and the fourth situation implies that the $[i \overset{s}{\to}]$ side of the bigon contains $[0, 1]$. These two properties imply that $0 < i + s - \ell < i$. It ensues that the $[j \overset{-s}{\to}]$ side contains index i, for otherwise $(i + s, j - s)$ would be a crossing occurring before step i. However, the occurrence of i on both sides of the bigon again contradicts Lemma 32. Hence, it must be that $\varepsilon = 1$, i.e. that $\varDelta$ is bounded by two forward paths. Since the $[i \overset{s}{\to}]$ side of $\varDelta$ contains $[0, 1]$, the $[j \overset{s}{\to}]$ side must lie to the left of $\varDelta$ and thus contain $\bar{1}$ turns. Let $k+1$ be the first index of a $\bar{1}$ turn along this side. Note that $j \leq k < i - 1$ because i is followed by a turn sequence of the form 2^*1. Claim 2 ensures the existence of a smaller index $r < k$ such that $c_k[k, k+1] = c_k[r, r-1]$ with $c_k[r, r-1]$ to the right of $c_k[k, k+1]$. See Figure 23. If $r > 1$, Claim 2 further states that c_k has a 1 turn at index $r - 1$. This is also true when $r = 1$ and $\varDelta$ starts with a vertical step: situation (4) implies a 1 turn at index 0 in this case. It follows in both cases that $c_i[r \to]$ is stuck inside the first staircase step of $\varDelta$: it cannot turn left as this would create a bracket; it cannot turn right as this would imply a crossing lexicographically smaller than (i, j); and it can neither go straight as it would also imply a crossing lexicographically smaller than (i, j). To see this we remark that the indices in $[r \overset{k+1-j}{\longrightarrow}]$ are each smaller than j for otherwise it would be followed by a 1 turn, inducing a bracket. Similar arguments holds in the case where $r = 1$ and $\varDelta$ starts with a horizontal step. We conclude that situation (4) also leads to a contradiction and that $\tilde{d}_i[i \overset{+\infty}{\to}]$ and $\tilde{d}_j$ indeed have no crossing. □

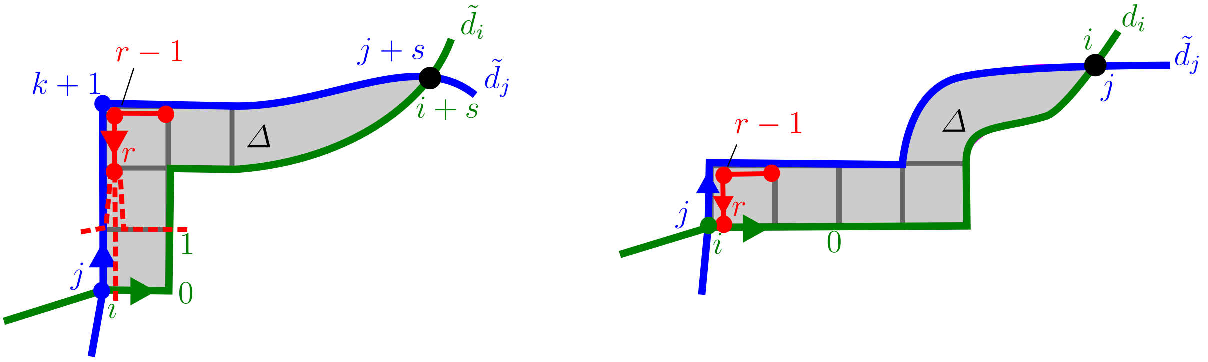

Figure 23: $\varDelta$ may start with a vertical step (left) or a horizontal step (right).

We next consider the smallest positive integer r such that $\tilde{d}_i$ and $\tilde{d}_j$ crosses at $\tilde{d}_i(i-r)$. If no such r exists, then by the above Claim 3, $\tilde{d}_j$ and $\tilde{d}_i$ have a unique intersection point and we may conclude that their limit points alternate. We can thus assume the existence of a bigon $([i-r \overset{r}{\to}], [j-\varepsilon r \overset{\varepsilon r}{\to}])$ with $r > 0$ and $\varepsilon \in \{-1, 1\}$, and without intermediate crossings. We first examine the case $\varepsilon = -1$ where the bigon $([i-r \overset{r}{\to}], [j+r \overset{-r}{\to}])$ has oppositely oriented index paths. We must have $r \geq i-j$ for otherwise the tip $(i-r, j+r)$ of the bigon would define a crossing occurring before step i, contradicting the choice of i. Hence, $[i-r \overset{r}{\to}]$ contains j. This is however impossible by Lemma 32. We now look at the case of two forwards index paths ($\varepsilon = 1$). We must have $r \geq j$ for otherwise the tip $(i-r, j-r)$ of the bigon would define a crossing occurring before step i, again contradicting the choice of i. It follows that $[j-r \overset{r}{\to}]$ contains $[0,1]$ and that the forward branches $\tilde{d}_i[i-j \overset{+\infty}{\to}]$ and $\tilde{d}_j[0 \overset{+\infty}{\to}]$ have a unique intersection point. The bigon labels a disk diagram Δ composed of paths and staircases as described in Theorem 10.

- If $[0,1]$ and $\alpha := [i-j, i-j+1]$ label the same arc of a path part in Δ there are two possibilities: either it holds initially that $c[0,1] = c(\alpha)$ or α was switched in the course of the algorithm. In the former case, we know by the preprocessing phase and Lemma 35 that the left-to-right order of $[0,1]$ and α is coherent with its extension in the backward direction. This implies that ultimately in the backward direction $\tilde{d}_i$ lies on the same side of $\tilde{d}_j$ as does α. In the latter case, as described in the end of the proof of Lemma 36, we know by the insertion procedure that at least one of the limit points of $\tilde{d}_i$ lies on the same side of $\tilde{d}_j$ as does α. Since (i,j) is the only crossing in the forward direction, we conclude in both cases that $\tilde{d}_j$ and $\tilde{d}_i$ have alternating limit points.

- Otherwise, $[0,1]$ and α label two distinct arcs, say a_0 and a_{i-j}, of a staircase part σ of Δ. Let $[i-v \overset{v-u}{\to}]$ and $[j-v \overset{v-u}{\to}]$, $0 \leq u < v \leq r$, be the index paths corresponding to the sides σ_L and σ_R of σ as pictured in Figure 24. By Remark 4, $[j-v \overset{v-u}{\to}]$ must label the right side σ_R while $[i-v \overset{v-u}{\to}]$ cannot contain $[0,1]$. It follows that $v < i$, whence $[i-v \overset{v-u}{\to}] \subset [1 \overset{i-1}{\to}]$. If a_{i-j} belongs to a horizontal part of σ_L, the first vertex in this part has a $\bar{1}$ turn and we let m be the index of this vertex. Otherwise, a_{i-j} belongs to a vertical part whose last vertex has a $\bar{1}$ turn. It follows that the vertex of index $i-j+1$ had a $\bar{1}$ turn at step $i-j$ of the algorithm. We set $m = i-j+1$ in this case. See Figure 24. In both cases, c_{m-1} had a $\bar{1}$ turn at index m and by Claim 2, there is an arc occurrence $[x, x-1]$, with $x < m-2$, that lies to the right of a_{i-j}. We view Δ as a subset of the Poincaré disk so that Δ_L and Δ_R can be seen as portions of $\tilde{d}_i$ and $\tilde{d}_j$ respectively. Let $\tilde{q}$ be the lift of c' that extends the above occurrence $[x, x-1]$ in $\mathbb{D}$. We denote by $q_+ := \tilde{q}[x \overset{+\infty}{\to}]$ and $q_- := \tilde{q}[x \overset{-\infty}{\to}]$ the portion of $\tilde{q}$ respectively after and before its vertex with index x. See Figure 25.

  We claim that q_+ cannot cross $\tilde{d}_i$. Otherwise, we would get a pair of homotopic paths, one piece of q_+ starting at index x and one piece of $\tilde{d}_i$ ending at index $m-1$, with opposite orientations. By Lemma 32, the q_+ piece would not contain index $m-1$ and would thus have length at most $m-1-x < m-1$. In turn, this would imply that the $\tilde{d}_i$ piece would only contain vertices with indices in $[1, m]$. The crossing of q_+ and $\tilde{d}_i$ would thus occur before step i, a contradiction.

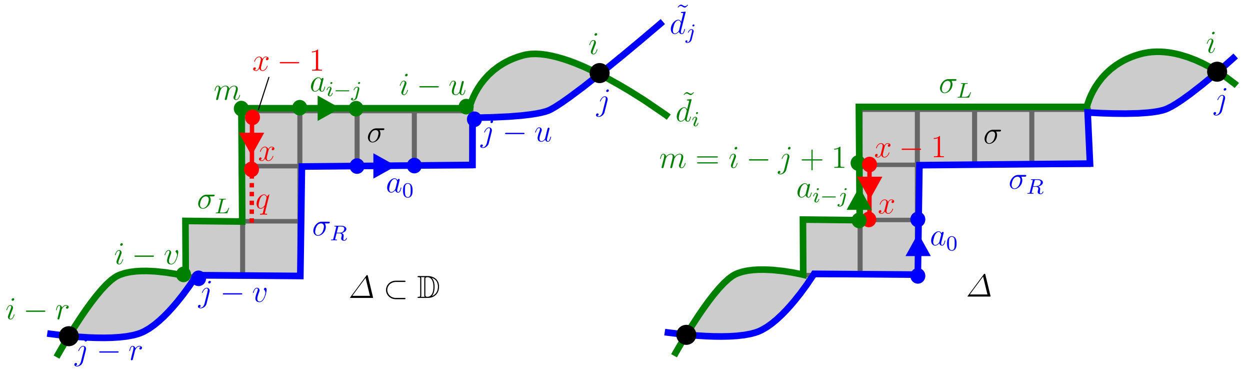

Figure 24: a_{i-j} may belong to a horizontal (left figure) or vertical (right figure) part of σ_L.

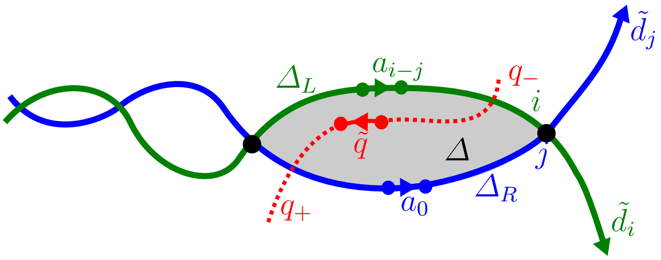

Figure 25: q_+ cannot cross $\tilde{d}_i$, q_- cannot cross $\tilde{d}_j$, and $\tilde{d}_j$ cannot cross $\tilde{d}_i[i \xrightarrow{+\infty}]$.

We next claim that q_- cannot cross $\tilde{d}_j$. To see this, first note that q_+ crosses $\tilde{d}_j$ before a_0 (i.e., at a point with index in $[j-r \xrightarrow{r-j}]$) by the previous claim. If q_- crossed $\tilde{d}_j$, say at $q_-(s) = \tilde{d}_j(t)$, then we would have a bigon between $\tilde{q}$ and $\tilde{d}_j$ whose $\tilde{d}_j$ side contains a_0. Since no crossing occurs before step i, we would have either $s \geq i$ or $t \geq i$. We infer in both cases that the index path of the $\tilde{q}$ side of the bigon contains index 0. This would contradict Lemma 32.

It follows from the above claims that the two lifts $\tilde{q}$ and $\tilde{d}_i$ of c' have alternating limit points.

It remains to consider the case $i = \ell$ where the first crossing appears after the last arc insertion. Since all the arcs have been inserted without introducing crossings it means that the crossings of the computed perturbation Π have the form $(0, j)$. We first claim that each bigon of Π must have its two sides oriented the same way. Otherwise, the tips of the bigon have the form $(0, j)$ and $(j, 0)$ for some index $j \neq 0$, in contradiction with Lemma 32. Let $\tilde{d}_\ell$ and $\tilde{d}_j$ be two intersecting lifts of c' in the Poincaré disk. If $\tilde{d}_\ell$ and $\tilde{d}_j$ intersect only once, then we are done as they must have alternating limit points. We now suppose that $\tilde{d}_\ell$ and $\tilde{d}_j$ intersect at least twice and we consider the bigon Δ between their *last two* intersections in the forward direction. By Lemma 34, the length of Δ is smaller than ℓ. Let $(\ell - j, 0)$ and $(0, j)$ be the tips of Δ, so that $\tilde{d}_j[0 \xrightarrow{j}] \sim \tilde{d}_\ell[\ell - j \xrightarrow{j}]$.

- If $[0, 1]$ and $\alpha = [\ell - j, \ell - j + 1]$ label the same arc $\tilde{d}_j[0, 1] = \tilde{d}_\ell(\alpha)$ then, by the insertion procedure, one of the limit points of $\tilde{d}_\ell$ is on the same side of $\tilde{d}_j$ as α. The same argument as in the general case $i < \ell$ allows us to conclude that $\tilde{d}_\ell$ and $\tilde{d}_j$ have alternating limit points.

- We finally suppose that $[0,1]$ and α label distinct arcs, say a_0 and $a_{\ell-j}$, of a staircase part σ of Δ. As in the general case (i.e., $i < \ell$), a_0 must see σ on its left, while $a_{\ell-j}$ must see σ on its right. Hence, the $\tilde{d}_j$ side of σ is canonical while the side along $\tilde{d}_\ell$ is not and α must have been switched. By Claim 2, there is an arc occurrence β to the right of $\tilde{d}_\ell[\ell-j, \ell-j+1](\alpha)$ with the opposite orientation. Let $\tilde{q}$ be the lift of c' that extends β. As in the general case $i < \ell$, we can show that the part of $\tilde{q}$ after β cannot cross $\tilde{d}_\ell$, while the part before β cannot cross $\tilde{d}_j$. Using that $(0,j)$ is the last crossing along $\tilde{d}_\ell$ and $\tilde{d}_j$, we equally conclude that $\tilde{d}_\ell$ and $\tilde{q}$ have alternating limit points. See Figure 25 with $i = \ell$.

□

Proof of Theorem 3. Let c be a combinatorial curve of length ℓ on a combinatorial surface of size n. We compute its canonical form in $O(n+\ell)$ time and check in linear time that c is primitive. In the negative, we conclude that either c is contractible, hence reduced to a vertex, or that c has no embedding by Proposition 26. In the affirmative, we apply the unzip algorithm to compute a perturbation Π of some geodesic c' homotopic to c. According to Proposition 38, we have $i(c) = 0$ if and only if Π has no crossings. This is easily verified in $O(n+\ell)$ time by checking for each vertex v of the system of quads that the set of paired arc occurrences with v as middle vertex form a well-parenthesized sequence with respect to the local ordering $\prec_v$ induced by Π. We conclude the proof thanks to Lemma 36. □

10 Concluding remarks

The existence of a singular bigon claimed in Theorem 30 relies on Theorem 4.2 of Hass and Scott [HS85]. As noted by the authors themselves this result is “surprisingly difficult to prove”. Except for this result and the recourse to some hyperbolic geometry in the general strategy of Section 3 our combinatorial framework allows us to provide simple algorithms and to give simple proofs of results whose known demonstrations are rather involved. Concerning Proposition 28, the existence of a perturbation without bigon could be achieved in our combinatorial viewpoint by showing that if a perturbation has bigons, then one of them can be swapped to reduce the number of crossings. (The example in Figure 1 shows that such a bigon need not be singular.) What is more, if those swappable bigons could be found easily this would provide an algorithm to compute a minimally crossing perturbation of two curves by iteratively swapping bigons as in Section 8 for the case of a single curve. However, we were unable to show the existence or even an appropriate definition of a swappable bigon. Note that in the analogous approaches using Reidemeister-like moves by de Graff and Schrijver [dGS97] or by Paterson [Pat02], the number of moves required to reach a minimal configuration is unknown[8]. Comparatively, the number of bigon swaps would be just half the excess crossing of a given perturbation. We would thus obtain a

[8] In a very recent work, Chang et al. [CEL$^+$18] were able to show that a *single* curve with n self-intersections on an orientable surface could be put in minimal configuration by applying $O(n^4)$ Reidemeister-like moves. In our framework a combinatorial perturbation of a closed walk of length n may have $\Omega(n^2)$ self-intersections, so that their approach leads to $O(n^8)$ moves to reach a minimal configuration.

polynomial time algorithm for computing a minimally crossing perturbation of two curves (there is an exponential time algorithm by a result of Neumann-Coto [NC01, Prop. 2.2]).

Although the geometric intersection number of a combinatorial curve of length ℓ may be $\Omega(\ell^2)$, it is not clear that the complexity in Theorem 1 is optimal. In particular, it would be interesting to see if the unzip algorithm of Section 9.2 yields minimally crossing curves even with curves that are not homotopic to simple curves, thus improving Theorem 2. This would yield a quasi-linear algorithm for computing the geometric intersection number based on the counting of inversions in a permutation [BLRvW84]. It is also tempting to check whether the unzip algorithm applies to compute the geometric intersection number of two curves rather than a single curve. Another intriguing question related to the computation of minimally crossing perturbations comes from the fact that they are not unique. Given a combinatorial perturbation we can construct a continuous realization as described in the proof of Lemma 27. Say that two realizations are in the **same configuration** if there is an ambient isotopy of the surface where they live that brings one realization to the other. It was shown by Neumann-Coto [NC01] that every minimally crossing perturbation is in the configuration of shortest geodesics for some Riemaniann metric μ, but Hass and Scott [HS99] gave counterexamples to the fact that we could always choose μ to be hyperbolic. We conclude with the following open question: *Is there an algorithm to construct or detect combinatorial perturbations that have a realization in the configuration of geodesics for some hyperbolic metric?*

References

[AAET17] Hugo A Akitaya, Greg Aloupis, Jeff Erickson, and Csaba D Tóth. Recognizing weakly simple polygons. *Discrete & Computational Geometry*, 58(4):785–821, 2017.

[Are15] Chris Arettines. A combinatorial algorithm for visualizing representatives with minimal self-intersection. *Journal of Knot Theory and Its Ramifications*, 24(11):1–17, 2015.

[BLRvW84] Jon Bentley, Charles Leiserson, Ronald Rivest, and Christopher van Wyk. Problem 84-2 (leo guibas editor). *Journal of Algorithms*, 5(4):592–594, 1984.

[BS84] Joan S. Birman and Caroline Series. An algorithm for simple curves on surfaces. *J. London Math. Soc.*, 29(2):331–342, 1984.

[Bus92] Peter Buser. *Geometry and Spectra of Compact Riemann Surfaces*. Birkhäuser, 1992.

[CdVL05] Éric Colin de Verdière and Francis Lazarus. Optimal System of Loops on an Orientable Surface. *Discrete & Computational Geometry*, 33(3):507 – 534, March 2005.

[CEL⁺18] Hsien-Chih Chang, Jeff Erickson, David Letscher, Arnaud De Mesmay, Saul Schleimer, Eric Sedgwick, Dylan Thurston, and Stephan Tillmann.

Tightening curves on surfaces via local moves. In *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 121–135. Society for Industrial and Applied Mathematics, 2018.

[CEX15] Hsien-Chih Chang, Jeff Erickson, and Chao Xu. Detecting weakly simple polygons. In *Proc. 26th Ann. ACM-SIAM Symp. Discrete Algorithms*, pages 1655–1670, 2015.

[Cha14] Moira Chas. Self-intersection numbers of length-equivalent curves on surfaces. *Experimental Mathematics*, 23(3):271–276, 2014.

[Chi69] David R.J. Chillingworth. Simple closed curves on surfaces. *Bulletin of the London Mathematical Society*, 1(3):310–314, 1969.

[Chi71] David R.J. Chillingworth. An algorithm for families of disjoint simple closed curves on surfaces. *Bulletin of the London Mathematical Society*, 3(1):23–26, 1971.

[Chi72] David R.J. Chillingworth. Winding numbers on surfaces. II. *Mathematische Annalen*, 199(3):131–153, 1972.

[CL87] Marshall Cohen and Martin Lustig. Paths of geodesics and geometric intersection numbers: I. In *Combinatorial group theory and topology*, volume 111 of *Ann. of Math. Stud.*, pages 479–500. Princeton Univ. Press, 1987.

[CL12] Moira Chas and Steven P Lalley. Self-intersections in combinatorial topology: statistical structure. *Inventiones mathematicae*, 188(2):429–463, 2012.

[DG99] Tamal K. Dey and Sumanta Guha. Transforming Curves on Surfaces. *Journal of Computer and System Sciences*, 58(2):297–325, 1999.

[dGS97] Maurits de Graaf and Alexander Schrijver. Making curves minimally crossing by Reidemeister moves. *Journal of Combinatorial Theory, Series B*, 70(1):134–156, 1997.

[DLH10] Pierre De La Harpe. Topologie, théorie des groupes et problèmes de décision: célébration d'un article de max dehn de 1910. *Gazette des mathématiciens*, 125:41–75, 2010.

[EW13] Jeff Erickson and Kim Whittelsey. Transforming curves on surfaces redux. In *Proc. 24th Ann. ACM-SIAM Symp. Discrete Algorithms*, 2013.

[FM12] Benson Farb and Dan Margalit. *A primer on mapping class groups*. Princeton university Press, 2012.

[GKZ05] Daciberg L. Gonçalves, Elena Kudryavtseva, and Heiner Zieschang. An algorithm for minimal number of (self-)intersection points of curves on surfaces. In *Proceedings of the Seminar on Vector and Tensor Analysis*, volume 26, pages 139–167, 2005.

[GS90] Steve M. Gersten and Hamish B. Short. Small cancellation theory and automatic groups. *Inventiones mathematicae*, 102:305–334, 1990.

[HS85] Joel Hass and Peter Scott. Intersections of curves on surfaces. *Israel Journal of Mathematics*, 51(1-2):90–120, 1985.

[HS94] Joel Hass and Peter Scott. Shortening curves on surfaces. *Topology*, 33(1):25–43, 1994.

[HS99] Joel Hass and Peter Scott. Configurations of curves and geodesics on surfaces. *Geometry and Topology Monographs*, 2:201–213, 1999.

[KMP77] Donald E. Knuth, James H. Morris, Jr, and Vaughan R. Pratt. Fast pattern matching in strings. *SIAM journal on computing*, 6(2):323–350, 1977.

[Lot97] M. Lothaire. *Combinatorics on words*. Cambridge University Press, 1997.

[LR12] Francis Lazarus and Julien Rivaud. On the homotopy test on surfaces. In *Proc. 53rd Ann. IEEE Symp. Foundations of Computer Science*, pages 440–449, 2012.

[Lus87] Martin Lustig. Paths of geodesics and geometric intersection numbers: II. In *Combinatorial group theory and topology*, volume 111 of *Ann. of Math. Stud.*, pages 501–543. Princeton Univ. Press, 1987.

[Mir08] Maryam Mirzakhani. Growth of the number of simple closed geodesics on hyperbolic surfaces. *Annals of Mathematics*, pages 97–125, 2008.

[Mir16] Maryam Mirzakhani. Counting mapping class group orbits on hyperbolic surfaces. Preprint arxiv:1601.03342, January 2016.

[MT01] Bojan Mohar and Carsten Thomassen. *Graphs on Surfaces*. Studies in the Mathematical Sciences. Johns Hopkins University Press, 2001.

[NC01] Max Neumann-Coto. A characterization of shortest geodesics on surfaces. *Algebraic and Geometric Topology*, 1:349–368, 2001.

[Pat02] J.M. Paterson. A combinatorial algorithm for immersed loops in surfaces. *Topology and its Applications*, 123(2):205–234, 2002.

[Poi04] Henri Poincaré. Cinquième complément à l'analysis situs. *Rendiconti del Circolo Matematico di Palermo*, 18(1):45 –110, 1904.

[Rei62] Bruce L. Reinhart. Algorithms for jordan curves on compact surfaces. *Annals of Mathematics*, 75(2):209–222, 1962.

[Sap15] Jenya Sapir. Bounds on the number of non-simple closed geodesics on a surface. Preprint arxiv:1505.07171, May 2015.

[SSS08] Marcus Schaefer, Eric Sedgwick, and Daniel Stefankovic. Computing Dehn twists and geometric intersection numbers in polynomial time. In *CCCG*, pages 111–114, 2008.

[Sti93] J. Stillwell. *Classical Topology and Combinatorial Group Theory*. Graduate Texts in Mathematics. Springer, 1993.

[Tur79] Vladimir G. Turaev. Intersections of loops in two-dimensional manifolds. *Mathematics of the USSR-Sbornik*, 35(2):229, 1979.

[Zie65] Heiner Zieschang. Algorithmen für einfache kurven auf flächen. *Mathematica Scandinavica*, 17:17–40, 1965.

[Zie69] Heiner Zieschang. Algorithmen für einfache kurven auf flächen II. *Mathematica scandinavica*, 25:49–58, 1969.

Appendix

A Proof of the direct implication of Proposition 28

Proposition. *If a combinatorial perturbation of one or two primitive curves has excess (self-)crossing then it contains a bigon or a monogon.*

We first introduce some terminology. An **elementary homotopy** on a combinatorial curve c consists in adding or removing a spur, or replacing in c a possibly empty part of a facial walk by its complementary part. For a free elementary homotopy we can also apply a circular shift to the indices of the closed curve c. The equivalence relation generated by (free) homotopies is called combinatorial (free) homotopy.

Let $\rho : G \to S$ be a cellular embedding of G in a topological surface S corresponding to the combinatorial surface we are working with. Every combinatorial curve c can be realized by a continuous curve on S by replacing each arc in c with a continuous arc deduced from the restriction of ρ to the corresponding arc. We denote this continuous realization by $\rho(c)$. It is part of the folklore that (free) combinatorial homotopy coincides with continuous (free) homotopy, meaning that c is (freely) homotopic to d if an only if $\rho(c)$ is (freely) homotopic to $\rho(d)$.

An **elementary move** of a combinatorial perturbation consists either in an elementary homotopy or in an **adjacent transposition**, i.e. in exchanging the left-to-right order of two occurrences in a same arc where one occurrence is next to the right of the other. We further require before performing an elementary homotopy that the perturbation is in **good position**. This means, if the elementary homotopy applies to a nonempty part u of a facial walk of some face, that each arc occurrence in u should be the rightmost element of its arc, i.e. the most interior to the face. When removing a spur, we just require that the two arc occurrences to be removed are adjacent in left-to-right order. See Figure 26. The elementary homotopy does not modify the order of the remaining arc occurrences. When inserting a spur we make the inserted arc occurrences adjacent and when inserting part of a facial walk we insert each arc occurrence after the rightmost element of its arc. Remark that by a sequence of adjacent transpositions we can always enforce a perturbation to be in good position.

Proof of the Proposition. We consider the case of a perturbation Π of two curves c and d. Suppose that Π has p excess crossings. We shall prove the stronger claim that Π has at least $\lceil p/2 \rceil$ bigons whose tips are pairwise distinct. Let Ψ be a perturbation of two curves c' and d' respectively homotopic to c and d such that Ψ has no excess crossings. Consider a sequence of elementary homotopies from (c', d') to (c, d). Following the above remark, we can insert adjacent transpositions between each elementary homotopy in order to obtain a sequence of elementary moves from Ψ to Π. The claim is trivially verified by Ψ. We now check that the claim remains true after each elementary move. If the move is an adjacent transposition we may assume that we exchange an arc occurrence of c with an arc occurrence of d since we only consider intersections between c and d. Without loss of generality we also assume exchanging the forward occurrences $[i, i+1]_c$ and $[j, j+1]_d$. There are three cases to consider.

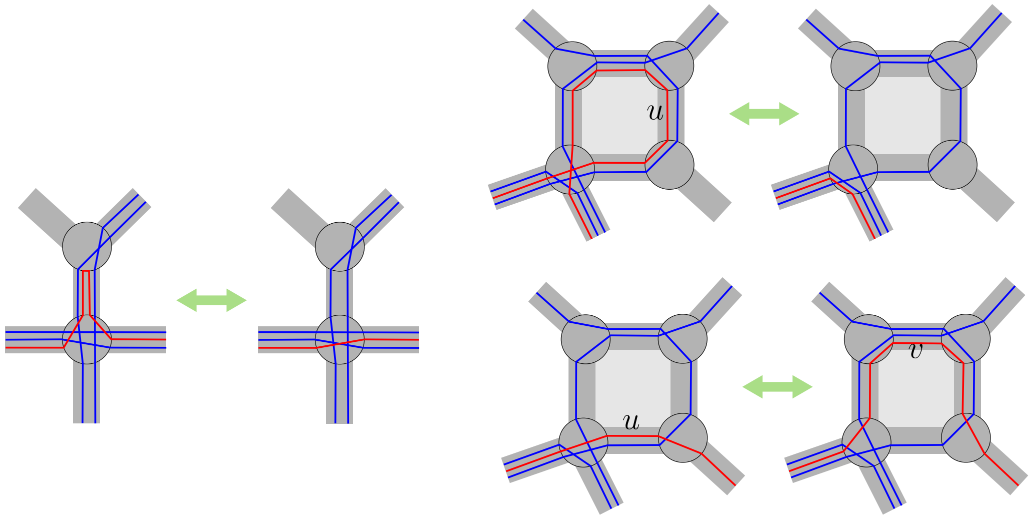

Figure 26: Left, the removal of a spur in good position. Right, two elementary homotopies replacing a subpath u of a facial walk by the complementary subpath v. In the upper right, v is empty. The perturbations being in good positions, crossings may only appear or disappear by pairs.

1. If none of (i, j) and $(i+1, j+1)$ is a crossing then the transposition adds two crossings that we may pair as the tips of a new bigon. We now have $p+2$ excess crossings with at least $\lceil (p+2)/2 \rceil$ bigons as required.

2. If (i, j) is a crossing but $(i+1, j+1)$ is not and if (i, j) is paired to form the tips of a bigon, say $([i \xrightarrow{\ell}]_c, [j \xrightarrow{k}]_d)$, we may just replace this bigon by $([i+1 \xrightarrow{\ell-1}]_c, [j+1 \xrightarrow{k-1}]_d)$ sliding its tip (i, j) to $(i+1, j+1)$. A similar procedure applies when $(i+1, j+1)$ is a crossing but (i, j) is not. In each case the number of excess crossings and bigons is left unchanged.

3. It remains the case where both (i, j) and $(i+1, j+1)$ are crossings. If none of the two is paired to form the tips of a bigon, then the transposition removes two crossings and no pairing, so that the claim remains trivially true. If exactly one of the two is paired or if the two are paired together, we loose one bigon and two crossings after the transposition so that the claim remains true. Otherwise, there are two bigons of the form $([i \xrightarrow{\ell}]_c, [j \xrightarrow{k}]_d)$ and $([i+1 \xrightarrow{\ell'}]_c, [j+1 \xrightarrow{k'}]_d)$ that we can recombine to form the bigon $([i+\ell \xrightarrow{1+\ell'-\ell}]_c, [j+k \xrightarrow{1+k'-k}]_d)$. We again have one less bigon and two less crossings.

We now consider the application of an elementary homotopy as described before the proof. There are again three possibilities.

1. If the homotopy replaces a nonempty subpath u of a facial walk uv^{-1} by the nonempty complementary part v then no crossing may appear or disappear as we assume the perturbation in good position. In particular, no crossing may use an internal vertex of u and u is either entirely included in or excluded from any side of any bigon. We can replace u by v in any bigon side where u occurs to obtain

valid bigons in the new perturbation after the elementary homotopy is applied. The number of excess crossings and bigons is left unchanged.

2. When u is empty in the above replacement, or when inserting a spur, we may only add crossings by pairs forming bigons with one zero-length side and the complement of u or the spur as the other side. A similar analysis as in case (2) of a transposition applies to take care of each pair.
3. When v is empty in the above replacement, or when removing a spur, we may only remove crossings by pairs and a similar analysis as in case (3) of a transposition applies to take care of each pair.

This ends the proof for the case of two curves. A similar proof holds for the existence of a bigon or a monogon in a perturbation with excess self-intersection. This time the excess crossings are either paired to form bigons or left alone as the tips of monogons. □

B Proof of Lemma 29

Lemma. *Swapping the two sides of a singular bigon of a perturbation of a geodesic primitive curve decreases its number of crossings by at least two.*

Proof. Consider a singular bigon $([i \xrightarrow{\ell}], [j \xrightarrow{\ell}])$ of a perturbation Π of a closed primitive canonical curve c. Let Ψ be the perturbation after the bigon has been swapped. We shall partition the set of potential double points and show that the net change of the number of crossings with respect to Π and Ψ is non positive in each part. As the tips of the bigon are not crossings in Ψ, this will prove the lemma. We set for $x \in \mathbb{Z}/|c|\mathbb{Z} \setminus ([i \xrightarrow{\ell}] \cup [j \xrightarrow{\ell}])$ and $0 < p, q < \ell$:

$$\begin{array}{ll}
\mathbb{D}_{p,q} = \{(i+p, j+q), \{(i+q, j+p)\}, & \mathbb{D}_{x,p} = \{(x, i+p), (x, j+p)\}, \\
\mathbb{D}_{p,0} = \{(i, i+p), (i, j+p), (j, i+p), (j, j+p)\}, & \mathbb{D}_{x,0} = \{(x, i), (x, j)\}, \\
\mathbb{D}_{p,\ell} = \{(i+\ell, i+p), (i+\ell, j+p), (j+\ell, i+p), (j+\ell, j+p)\}, & \mathbb{D}_{x,\ell} = \{(x, i+\ell), (x, j+\ell)\}, \\
\mathbb{D}'_{p,q} = \{(i+p, i+q), \{(j+q, j+p)\}, & \mathbb{D}_1 = \{(y, z) \mid y, z \notin ([i \xrightarrow{\ell}] \cup [j \xrightarrow{\ell}])\}, \\
\mathbb{D}_2 = \{(i, i+\ell), (i, j+\ell), (j, j+\ell), (j, i+\ell)\}, & \mathbb{D}_3 = \{(i, j), (i+\ell, j+\ell)\}
\end{array}$$

When $i = j + \ell$ or $j = i + \ell$, the set $\mathbb{D}_2$ consists of three double points only (or zero if $c(i) \neq c(i+\ell)$). Note that we cannot have both equalities as this would imply $j = j + 2\ell$, whence $\ell = |c|/2$ and c would be a square, in contradiction with the hypothesis that c is primitive. For each double point in $\mathbb{D}_1$ the four incident arc occurrences are left in place in Π and Ψ. It ensues that $\mathbb{D}_1$ has the same crossings in Π and Ψ. The double points in $\mathbb{D}_3$ are the tips of the bigon. They are crossings in Π and not in Ψ, whence a net change of -2 crossings. For each of $\mathbb{D}_{x,p}, \mathbb{D}_{p,q}$ or $\mathbb{D}'_{p,q}$, the first double point in their above definition is a crossing in Π (resp. Ψ) if and only if the second one is a crossing in Ψ (resp. Π). Their net change of crossings is thus null. For $\mathbb{D}_{x,0}$, $\mathbb{D}_{x,\ell}$, $\mathbb{D}_{p,0}$ and $\mathbb{D}_{p,\ell}$ there are a few case analysis depending on the relative ordering of the two arc occurrences incident to x (resp. $i+p$, $j+p$) with respect to the crossing (i, j) (resp. $(i+\ell, j+\ell)$. In each case (see Figure 27) the number of crossings cannot increase from Π to Ψ. For $\mathbb{D}_2$ there are three cases according to whether $i + \ell = j$, $j + \ell = i$, or none of the two identifications occurs.

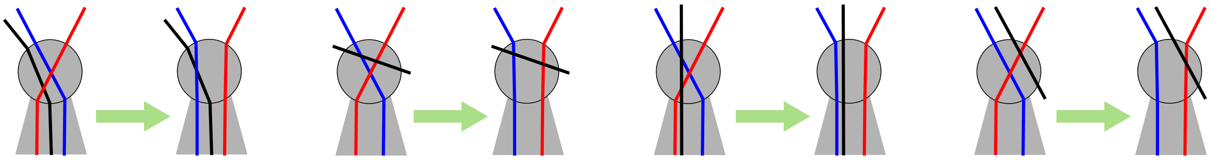

Figure 27: Up to some obvious symmetries the case $\mathbb{D}_{x,0}$, $\mathbb{D}_{x,\ell}$, $\mathbb{D}_{p,0}$ or $\mathbb{D}_{p,\ell}$ has four distinct possible configurations. The blue and red strands represent the crossing (i, j) (resp. $(i+\ell, j+\ell)$) and each arrow links the left configuration before the bigon swap (in Π) with the right configuration after the swap is applied (in Ψ).

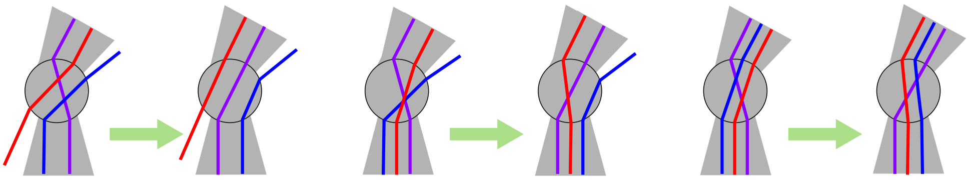

Figure 28: When $i + \ell = j$, the end of $[i \xrightarrow{\ell}]$ overlaps with the beginning of $[j \xrightarrow{\ell}]$ (in the purple strand). The third configuration on the right is actually forbidden by the definition of a singular bigon! It corresponds to the non-singular bigon on Figure 15.

Recall from the paragraph before the lemma that we cannot have both identifications. The first two cases can be treated the same way. See Figure 28. The number of possible configurations is much larger in the last case, i.e. when $i + \ell \neq j$ and $j + \ell \neq i$. We essentially have to shuffle two circular orderings of length four corresponding to the two crossing tips. Without loss of generality we can assume that the path from i to $i + \ell$ is the right side of the bigon. We consider the following arc occurrences:

$$\begin{array}{llll} \alpha = [i, i+1], & \alpha' = [i, i-1], & \beta = [j, j+1], & \beta' = [j, j-1] \\ \gamma = [i+\ell, i+\ell+1], & \gamma' = [i+\ell, i+\ell-1], & \delta = [j+\ell, j+\ell+1], & \delta' = [j+\ell, j+\ell-1] \end{array}$$

Since (i, j) and $(i+\ell, j+\ell)$ are crossings we must see $(\alpha, \beta, \alpha', \beta')$ in this counterclockwise order around the vertex $c(i) = c(i+\ell)$ and similarly for $(\gamma, \delta, \gamma', \delta')$. We denote by S_1 and S_2, respectively, these two circular sequences. We need to consider the effect of the bigon swapping on all the possible shuffles of S_1 and S_2. The restriction of these shuffles to α, β, γ and δ gives the 6 possible shuffles of (α, β) and (γ, δ). Among them the order $(\alpha, \delta, \gamma, \beta)$ cannot occur. Indeed, since the bigon is a thick double path the arcs $c(\alpha)$ and $c(\beta)$ either coincide or form a corner of a quad. This would force $c(\delta)$ and $c(\gamma)$ to lie in a similar configuration. In turn, the constrained order S_2 would also enforce $c(\delta')$ and $c(\delta)$ to coincide or form a corner of a quad in contradiction with the hypothesis that c has no spurs or brackets. Similar arguments show that the orders $(\alpha, \gamma, \beta, \delta)$, $(\alpha, \delta, \beta, \gamma)$, $(\alpha, \gamma, \delta, \beta)$ and $(\alpha, \beta, \delta, \gamma)$ can only occur as factors in the possible shuffles of S_1 and S_2. These 4 orders thus leads to 24 distinct shuffles of S_1 and S_2 by factoring with the 6 shuffles of (α', β') with (γ', δ'). By exchanging the roles of (α, β) and (γ, δ) and by turning clockwise instead of counterclockwise we see that $(\alpha, \delta, \beta, \gamma)$ leads to the same orders as $(\alpha, \gamma, \beta, \delta)$ and a similar correspondence holds for $(\alpha, \gamma, \delta, \beta)$ and $(\alpha, \beta, \delta, \gamma)$. We thus only need to check the 12 configurations depicted on Figure 29. It remains to consider the order $(\alpha, \beta, \gamma, \delta)$. Because of the constrained order S_2, γ' and δ' cannot lie between γ and

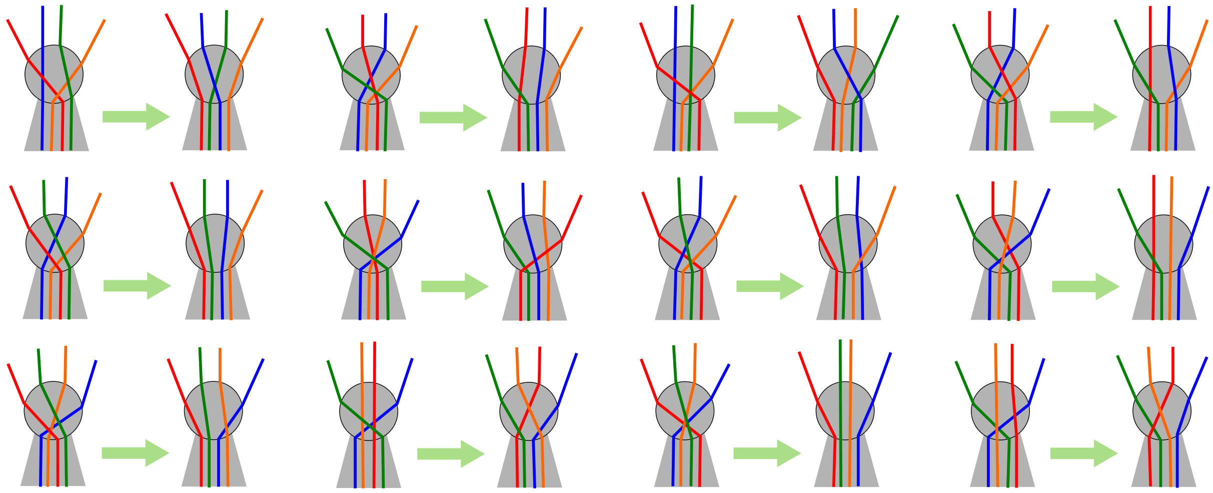

Figure 29: Effect of swapping the singular bigon $([i \overset{\ell}{\to}], [j \overset{\ell}{\to}])$ on $\mathbb{D}_2$ for all the orders including the factor $(\alpha, \gamma, \beta, \delta)$ or $(\alpha, \gamma, \delta, \beta)$. The blue strand represents (α, α'), the red one (β, β'), the orange one (γ, γ') and the green one (δ, δ').

δ. This leaves out $\binom{4}{2} = 6$ possible shuffles of (γ', δ') and $(\alpha, \beta, \gamma, \delta)$:

$$\begin{array}{lll} S'_1 : (\alpha, \beta, \gamma, \delta, \gamma', \delta') & S'_2 : (\alpha, \beta, \delta', \gamma, \delta, \gamma') & S'_3 : (\alpha, \beta, \gamma', \delta', \gamma, \delta) \\ S'_4 : (\alpha, \delta', \beta, \gamma, \delta, \gamma') & S'_5 : (\alpha, \gamma', \beta, \delta', \gamma, \delta) & S'_6 : (\alpha, \gamma', \delta', \beta, \gamma, \delta) \end{array}$$

We finally shuffle each of these orders with (α', β'). Since α' and β' cannot lie between α and β, we obtain $\binom{6}{2} = 15$ possible shuffles when considering either S'_1, S'_2 or S'_3, $\binom{5}{2} = 10$ possible shuffles with S'_4 or S'_5, and $\binom{4}{2} = 6$ possible shuffles with S'_6. By swapping left and right and turning clockwise instead of counterclockwise, we remark that S'_1 and S'_3 lead to the same orders and similarly for S'_4 and S'_5. We thus only need to consider the shuffles of (α', β') and S'_1, S'_2, S'_4 or S'_6 to complete the inspection of all the cases. Those shuffles are represented on Figures 30, 31, 32, 33 respectively.

In each of the configurations, we trivially check that the number of crossings is not increasing. This allows to conclude the lemma when the two index paths of a singular bigon are directed the same way. A similar analysis can be made when their directions are opposite, that is when the considered bigon has the form $([i \overset{\ell}{\to}], [j \overset{-\ell}{\to}])$. □

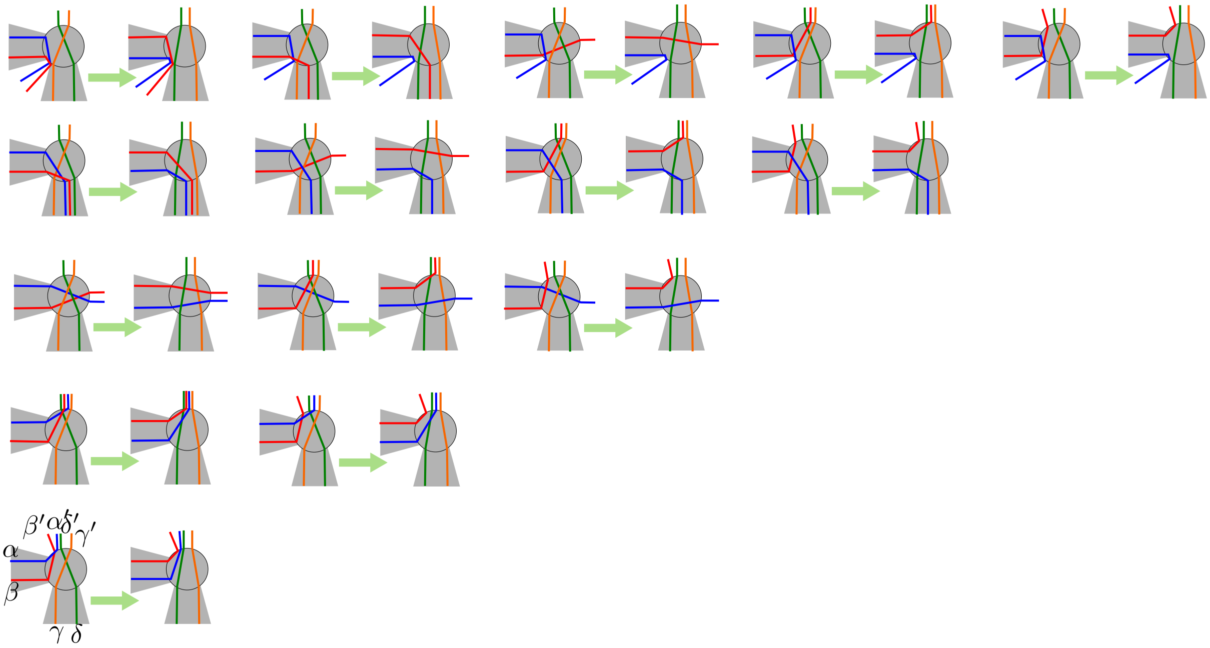

Figure 30: Effect of the bigon swapping on the 15 shuffles of S_1' and (α', β').

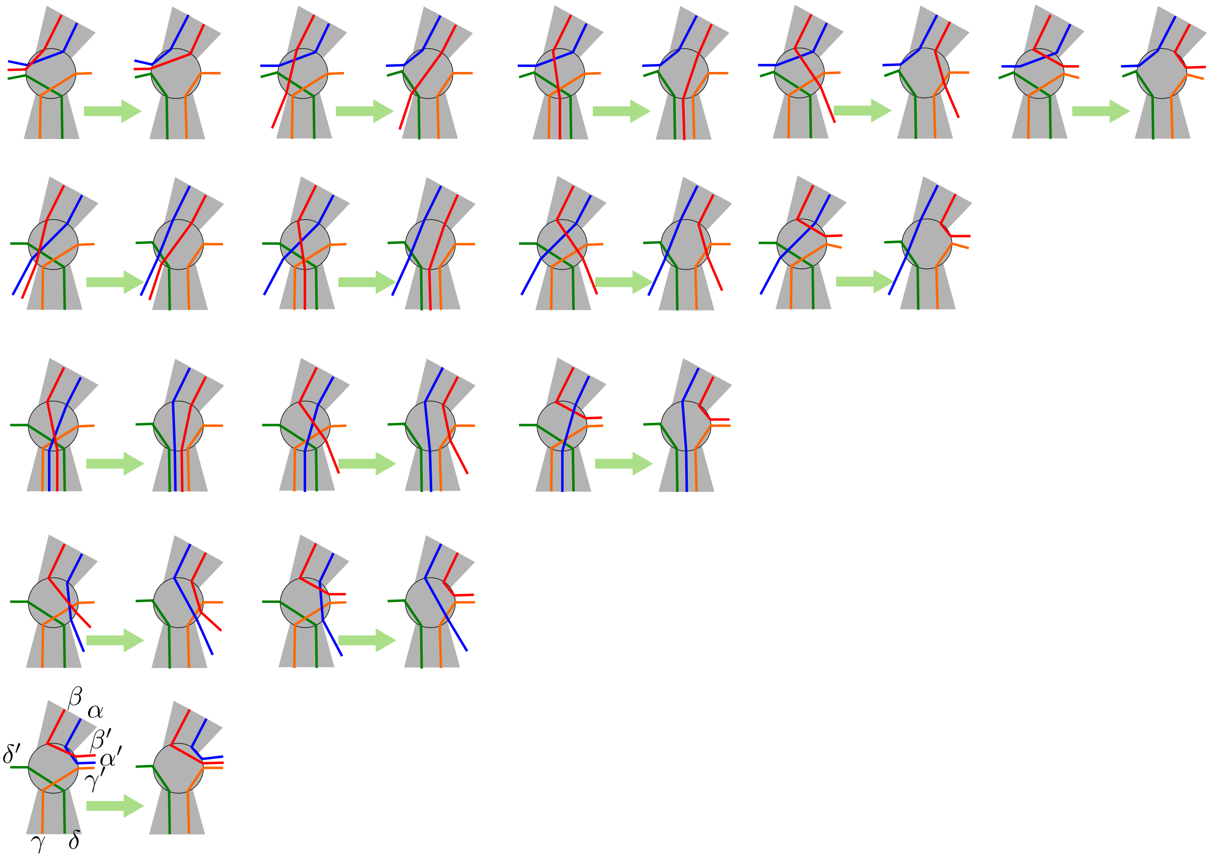

Figure 31: Effect of the bigon swapping on the 15 shuffles of S_2' and (α', β').

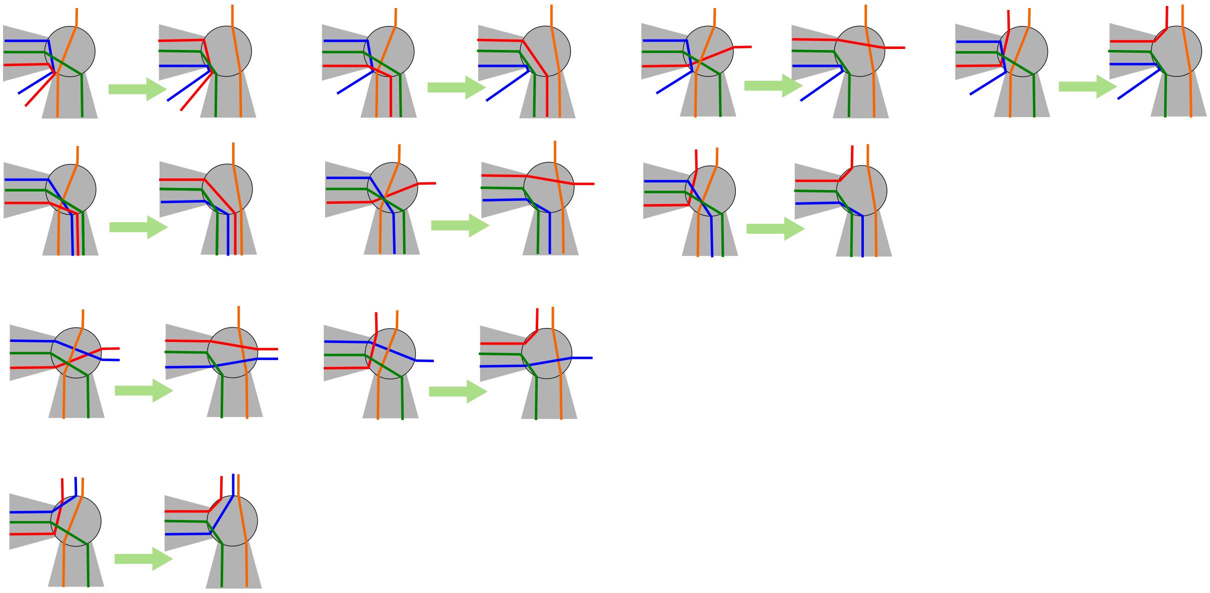

Figure 32: Effect of the bigon swapping on the 10 shuffles of S'_4 and (α', β').

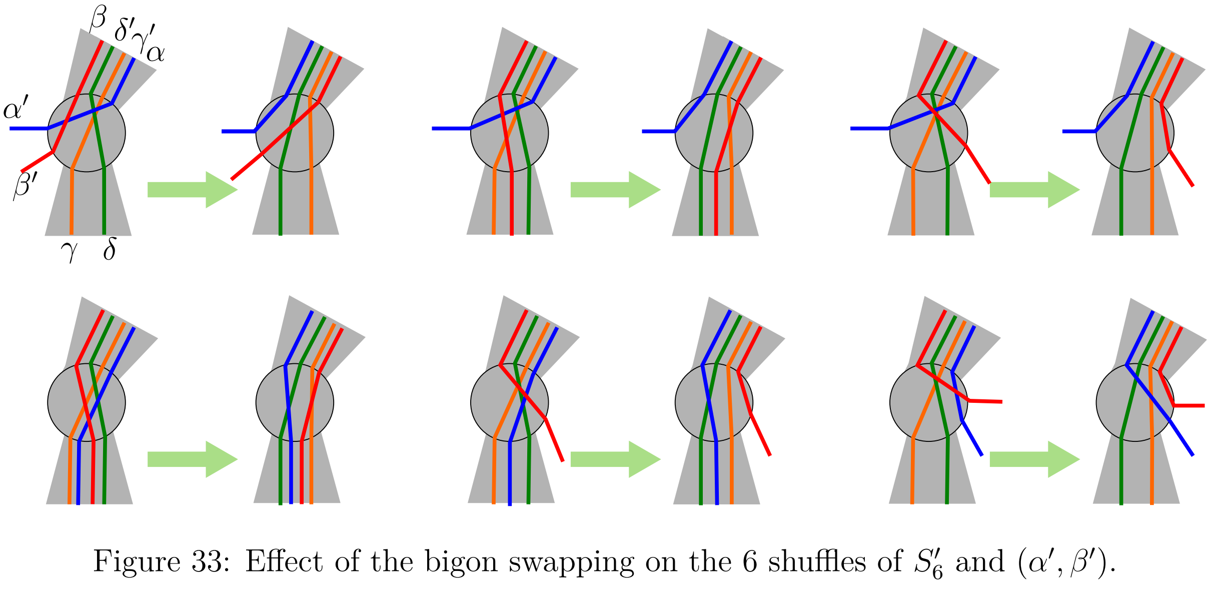

Figure 33: Effect of the bigon swapping on the 6 shuffles of S'_6 and (α', β').