From Curves to Words and Back Again: Geometric Computation of Minimum-Area Homotopy

Hsien-Chih Chang[1], Brittany Terese Fasy[2,3], Bradley McCoy[3], David L. Millman, and Carola Wenk[4]

[1] Department of Computer Science, Dartmouth College, Hanover, USA.
[2] Department of Mathematical Sciences, Montana State University, Bozeman, USA.
[3] School of Computing, Montana State University, Bozeman, USA.
[4] Department of Computer Science, Tulane University, New Orleans, USA.

Abstract. Let γ be a generic closed curve in the plane. Samuel Blank, in his 1967 Ph.D. thesis, determined if γ is self-overlapping by *geometrically* constructing a combinatorial word from γ. More recently, Zipei Nie, in an unpublished manuscript, computed the minimum homotopy area of γ by constructing a combinatorial word *algebraically*. We provide a unified framework for working with both words and determine the settings under which Blank's word and Nie's word are equivalent. Using this equivalence, we give a new geometric proof for the correctness of Nie's algorithm. Unlike previous work, our proof is constructive which allows us to naturally compute the actual homotopy that realizes the minimum area. Furthermore, we contribute to the theory of self-overlapping curves by providing the first polynomial-time algorithm to compute a self-overlapping decomposition of any closed curve γ with minimum area.

1 Introduction

A *closed curve* in the plane is a continuous map γ from the circle $\mathbb{S}^1$ to the plane $\mathbb{R}^2$. In the plane, any closed curve is homotopic to a point. A homotopy that sweeps out the minimum possible area is a *minimum homotopy*. Chambers and Wang [4] introduced the minimum homotopy area between two simple homotopic curves with common endpoints as a way to measure the similarity between the two curves. They suggest that homotopy area is more robust against noise than another popular similarity measure on curves called the *Fréchet distance*. However, their algorithm requires that each curve be simple, which is restrictive.

Fasy, Karakoç, and Wenk [12] proved that the problem of finding the minimum homotopy area is easy on a closed curve that is the boundary of an immersed disk. Such curves are called *self-overlapping* [10,15,18,23,24,26]. They also established a tight connection between minimum-area homotopy and self-overlapping curves by showing that any generic closed curve can be decomposed at some vertices into self-overlapping subcurves such that the combined homotopy from the subcurves is minimum. This structural result gives an exponential-time algorithm for the minimum homotopy area problem by testing each decomposition in a brute-force manner.

Nie, in an unpublished manuscript [19], described a polynomial-time algorithm to determine the minimum homotopy area of any closed curve in the plane. Nie's algorithm borrows tools from geometric group theory by representing the curve as a word in the fundamental group $\pi_1(\gamma)$, and connects minimum homotopy area to the *cancellation norms* [2,3,21] of the word, which can be computed using a dynamic program. However, the algorithm does not naturally compute an associated *minimum-area homotopy*.

Alternatively, one can interpret the words from the dynamic program *geometrically* as crossing sequences by traversing any subcurve cyclicly and recording the crossings along with their directions with a collection of nicely-drawn *cables* from each face to a point at infinity. Such geometric representation is known as the Blank words [1,22]. In fact, the first application of these combinatorial words given by Blank is an algorithm that determines if a curve is self-overlapping. Blank words are geometric in nature and thus the associated objects are polynomial in size. When attempting to interpret Nie's dynamic program from the geometric view, one encounters the question of how to extend Blank's definition of cables to *subcurves*, where the cables inherited from the original curve are no longer positioned well with respect to the subcurves. To our knowledge, no geometric interpretation of the dynamic program is known.

1.1 Our Contributions

We first show that Blank and Nie's word constructions are, in fact, equivalent under the right assumptions (Section 3). Next, we extend the definition of Blank's word to subcurves and arbitrary cable drawings (Section 4.1), and interpret the dynamic program by Nie geometrically (Section 4.2). Using the self-overlapping decomposition theorem by Fasy, Karakoç, and Wenk [12] we provide a correctness proof to the algorithm. Finally, we conclude with a new result that a minimum-area self-overlapping decomposition can be found in polynomial time. We emphasize that extending Blank words to allow arbitrary cables is in no way straightforward. In fact, many assumptions on the cables have to be made in order to connect self-overlapping curves and minimum-area homotopy; handling arbitrary cable systems, as seen in the dynamic program, requires further tools from geometric topology like *Dehn twists*.

2 Background

In this section, we introduce concepts and definitions that are used throughout the paper. We assume the readers are familiar with the basic terminology for curves and surfaces.

2.1 Curves and Graphs

A *closed curve* in the plane is a continuous map $\gamma : \mathbb{S}^1 \to \mathbb{R}^2$, and a *path* in the plane is a continuous map $\zeta : [0,1] \to \mathbb{R}^2$. A path ζ is *closed* when $\zeta(0) = \zeta(1)$. In this work, we are presented with a *generic* curve; that is, one where there are a finite number of self-intersections, each of which is transverse and no three strands cross at the same point. See Figure 1 for an example. The image of a generic closed curve is naturally associated with a four-regular plane graph. The self-intersection points of a curve are *vertices*, the paths between vertices are *edges*, and the connected components of the complement of the curve are *faces*. Given a curve, choose an arbitrary starting point $\gamma(0) = \gamma(1)$ and orientation for γ.

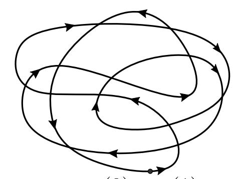


Fig. 1: A generic plane curve induces a four-regular graph.

The *dual graph* γ^* is another (multi-)graph, whose vertices represent the faces of γ, and two vertices in γ^* are joined by an edge if there is an edge between the two corresponding faces in γ. The dual graph is another plane graph with an inherited embedding from γ.

Let T be a spanning tree of γ. Let E denote the set of edges in γ, the tree T partitions E into two subsets, T and $T^* := E \setminus T$. The edges in T^* define a spanning tree of γ^* called the *cotree*. The partition of the edges (T, T^*) is called the *tree-cotree pair*.

We call a rooted spanning cotree T^* of γ^* a *breadth-first search tree* (BFS-tree) if it can be generated from a breadth-first search rooted at the vertex in γ^* corresponding to the unbounded face in γ. Each bounded face f of γ is a vertex in a breadth-first search tree T^*, we associate f with the unique edge incident to f^* in the direction of the root. Thus, there is a correspondence between edges of T^* and faces of γ.

2.2 Homotopy and Isotopy

A *homotopy* between two closed curves γ_1 and γ_2 that share a point p_0 is a continuous map $H\colon [0,1] \times \mathbb{S}^1 \to \mathbb{R}^2$ such that $H(0,\cdot) = \gamma_1$, $H(1,\cdot) = \gamma_2$, and $H(s,0) = p_0 = H(s,1)$. We define a homotopy between two paths similarly, where the two endpoints are fixed throughout the continuous morph. Notice that homotopy between two closed curves as *closed curves* and the homotopy between them as *closed paths* with an identical starting points are different. A homotopy between two injective paths ζ_1 and ζ_2 is an *isotopy* if every intermediate path $H(s,\cdot)$ is injective for all s. The notion of isotopy naturally extends to a collection of paths.

We can think of γ as a topological space and consider the *fundamental group* $\pi_1(\gamma)$. Elements of the fundamental group are called *words*, whose letters correspond to equivalence classes of homotopic closed paths in γ. The fundamental group of γ is a free group with basis consisting of the classes corresponding to the cotree edges of any tree-cotree pair of γ.

Let H be a homotopy between curves γ_1 and γ_2. Let $\#H^{-1}(x)\colon \mathbb{R}^2 \to \mathbb{Z}$ be the function that assigns to each $x \in \mathbb{R}^2$ the number of times the intermediate curves H sweep over x. The homotopy area of H is

$$\mathrm{Area}(H) := \int_{\mathbb{R}^2} \#H^{-1}(x)\, dx.$$

The minimum area homotopy between γ_1 and γ_2 is the infimum of the homotopy area over all homotopies between between γ_1 and γ_2. We denote this by $\mathrm{Area}_H(\gamma_1, \gamma_2) := \inf_H \mathrm{Area}(H)$. When γ_2 is the constant curve at a specific point p_0 on γ_1, define $\mathrm{Area}_H(\gamma) := \mathrm{Area}_H(\gamma, p_0)$. See Figure 2 for an example of a homotopy.

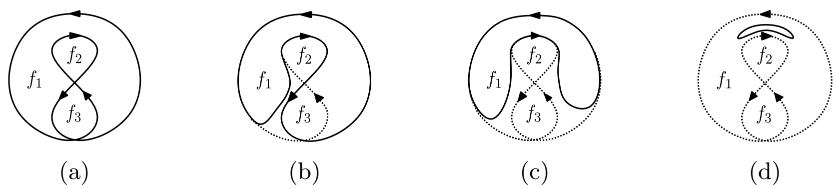


Fig. 2: (a) A generic closed curve in the plane. (b) We see a homotopy that sweeps over the face f_3. (c) The homotopy sweeps f_3 again. (d) The homotopy avoids sweeping over the face f_2. This is a minimum area homotopy for the curve, the area is $\mathrm{Area}(f_1) + 2 \cdot \mathrm{Area}(f_3)$.

For each $x \in \mathbb{R}\backslash\gamma$, the *winding number* of γ at x, denoted as $\mathrm{wind}(x, \gamma)$, is the number of times γ "wraps around" x, with a *positive* sign if it is counterclockwise, and *negative* sign otherwise. The winding number is a constant on each face. The *winding area* of γ is defined to be the integral

$$\mathrm{Area}_W(\gamma) := \int_{\mathbb{R}^2} |\mathrm{wind}(x, \gamma)|\, dx = \sum_{\text{face } f} |\mathrm{wind}(f, \gamma)| \cdot \mathrm{Area}(f).$$

The *depth* of a face f is the minimal number of edges crossed by a path from f to the exterior face. The depth is a constant on each face. We say the depth of a curve is equal the maximum depth over all faces. We define the *depth area* to be

$$\mathrm{Area}_D(\gamma) := \int_{\mathbb{R}^2} \mathrm{depth}(x, \gamma)\, dx = \sum_{\text{face } f} \mathrm{depth}(f) \cdot \mathrm{Area}(f).$$

Chambers and Wang [4] showed that the winding area gives a lower bound for the minimum homotopy area. On the other hand, there is always a homotopy with area $\mathrm{Area}_D(\gamma)$; one such homotopy can be constructed by smoothing the curve at each vertex into simple depth cycles [5], then contracting each simple cycle. Therefore we have

$$\mathrm{Area}_W(\gamma) \leq \mathrm{Area}_H(\gamma) \leq \mathrm{Area}_D(\gamma). \tag{1}$$

2.3 Self-Overlapping Curves

A generic curve γ is *self-overlapping* if there exists an immersion of the two disk $F : \mathbb{D}^2 \to \mathbb{R}^2$ such that $\gamma = F|_{\partial D^2}$. We say a map F *extends* γ. The image $F(\mathbb{D}^2)$ is the *interior* of γ. There are several equivalent ways to define self-overlapping curves [10,24,23,15,18]. Properties of self-overlapping curves are well-studied [9]; in particular, any self-overlapping curve has rotation number 1, where the *rotation number* of a curve γ is the winding number of the derivative γ' about the origin [26]. Also, the minimum homotopy area of any self-overlapping curve is equal to its winding area: $\mathrm{Area}_W(\gamma) = \mathrm{Area}_H(\gamma)$ [12].

The study of self-overlapping curves traces back to Whitney [26] and Titus [24]. Polynomial-time algorithms for determining if a curve is self-overlapping have been given [1,23], as well as NP-hardness result for extensions to surfaces and higher-dimensional spaces [7].

For any curve, the *intersection sequence*[5] $[\gamma]_V$ is a cyclic sequence of vertices $[v_0, v_1, \ldots, v_{n-1}]$ with $v_n = v_0$, where each v_i is an intersection point of γ. Each vertex appears exactly twice in γ_V. Two vertices x and y are *linked* if the two appearances of x and y in γ_V alternate in cyclic order: $\ldots x \ldots y \ldots x \ldots y \ldots$.

A pair of symbols of the same vertex x induces two natural subcurves generated by *smoothing* the vertex x; see Figure 3 for an example. (In this work, every smoothing is done in the way that respects the orientation and splits the curve into two subcurves.) A *vertex pairing* is a collection of pairwise unlinked vertex pairs in $[\gamma]_V$.

A *self-overlapping decomposition* Γ of γ is a vertex pairing such that the induced subcurves are self-overlapping; see Figure 3b and Figure 3d for examples. The subcurves that result from a vertex pairing are not necessary self-overlapping; see Figure 3c. For a self-overlapping decomposition Γ of γ, denote the set of induced subcurves by $\{\gamma_i\}_{i=1}^{\ell}$. Since each γ_i is self-overlapping, the minimum homotopy area is equal to its winding area. We define the *area of self-overlapping decomposition* to be

$$\mathrm{Area}_\Gamma(\gamma) \coloneqq \sum_{i=1}^{\ell} \mathrm{Area}_W(\gamma_i) = \sum_{i=1}^{\ell} \mathrm{Area}_H(\gamma_i).$$

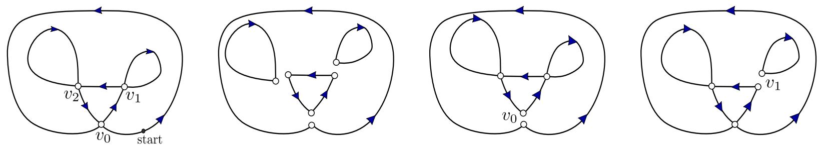
Fasy, Karakoç, and Wenk [12,14] proved the following structural theorem.

Theorem 1 (Self-Overlapping Decomposition [12, Theorem 20]). *Any curve γ has a self-overlapping decomposition whose area is minimum over all null-homotopies of γ.*

3 From Curves to Words

In order to work with plane curves, one must choose a *representation*. An important class of representations for plane curves are the various *combinatorial words*.

[5] also known as the unsigned Gauss code [5,13]



(a) (b) (c) (d)

Fig. 3: (a) Curve γ with intersection sequence $\gamma_V = [v_0, v_1, v_1, v_2, v_2, v_0]$. (b) All vertices are paired. (c) One of the subcurves is not self-overlapping. (d) Both subcurves are self-overlapping.

One example is the *Gauss code* [13]. Determining whether a Gauss code corresponds to an actual plane curve is one of the earliest computational topology questions [8].

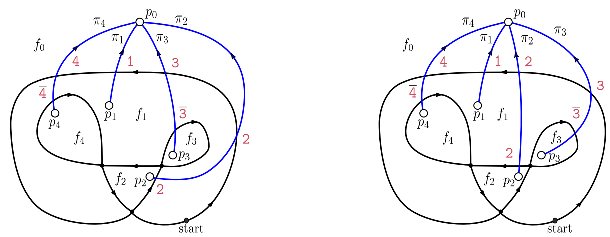
A plane curve (and its homotopic equivalents) can also be viewed as a word in the fundamental group $\pi_1(\gamma)$ of γ [1,22,19]. If we put a point p_i in each bounded face f_i, the curve γ is generated by the unique generators of each $\mathbb{R}^2 - \{p_i\}$. Nie [19] represents curves as words in the fundamental group to find the minimum area swept out by contracting a curve to a point. If the curve lies in a plane with punctures, one can define the *crossing sequence* of the curve with respect to a *system of arcs*, cutting the plane open into a simply-connected region. Blank [1] represents curves using a crossing sequences to determine if a curve is self-overlapping. While Blank constructed the words *geometrically* by drawing arcs and Nie defined the words *algebraically*, the dual view between the system of arcs and fundamental group suggests that the resemblance between Blank and Nie's constructions is not a coincidence.

In this section, we describe the construction by Blank; then, we interpret Blank's construction as a way of choosing the basis for the fundamental group under further restriction [22]. We prove that the Blank word is indeed unique when the restriction is enforced, providing clarification to Blank's original definition. We give a complete description of Nie's word construction and prove that Nie's word and Blank's word are equivalent.

3.1 Blank's Word Construction

We now describe Blank's word construction [1, page 5]. Let γ be a generic closed curve in the plane, pick a point in the unbounded face of γ, call it the *basepoint* p_0. From each bounded face f_i, pick a *representative point* p_i. Now connect each p_i to p_0 by a simple path in such a way that no two paths intersect each other. We call the collection of such simple paths a *cable system*, denoted as Π, and each individual path π_i from p_i to p_0 as a *cable*.

Orient each π_i from p_i to p_0. Now traverse γ from an arbitrary *starting point* of γ and construct a cyclic word by writing down the indices of γ crossing the cables π_i in the order they appear on γ; each index i has a *positive* sign if we cross π_i from right to left and a *negative* sign if from left to right. We denote negative crossing with an overline $\overline{\mathtt{i}}$. We call the resulting combinatorial word over the faces a *Blank word* of γ with respect to $\varPi$, denoted as $[\gamma]_B(\varPi)$. Figure 4 provides an example of Blank's construction.



(a) (b)

Fig. 4: (a) A curve γ with labeled faces and edges, $\varPi_a$ is drawn in blue. The Blank word of γ corresponding to $\varPi_a$ is $[\gamma]_B(\varPi_a) = [\mathtt{2314 2}\overline{\mathtt{3}}\overline{\mathtt{4}}]$. (b) The same curve with a different choice of cables $\varPi_b$. The corresponding Blank word is $[\gamma]_B(\varPi_b) = [\mathtt{3214}\overline{\mathtt{3}}\mathtt{2}\overline{\mathtt{4}}]$.

A word w is *reduced* if there are no two consecutive symbols in w that are identical and with opposite signs. We can enforce every Blank word to be reduced by imposing the following *shortest path assumption*: each cable has a minimum number of intersections with γ among all paths from p_i to p_0. A simple proof [1,6] shows that if $\varPi$ satisfies the shortest path assumption, the corresponding Blank word with respect to $\varPi$ is reduced. However, the choice of the cable system, and how it affects the constructed Blank word, was never explicitly discussed in the original work (presumably because for the purpose of detecting self-overlapping curves, any cable system satisfying the shortest path assumption works). In general, reduced Blank words constructed from different cable systems for the same curve are not identical, see Figure 4a and Figure 4b for an example. In this paper, we show that if the two cable systems have the same *cable ordering*–the (cyclic) order of cables around point p_0 in the unbounded face–then their corresponding (reduced) Blank words are the same, under proper assumptions on the cable system.

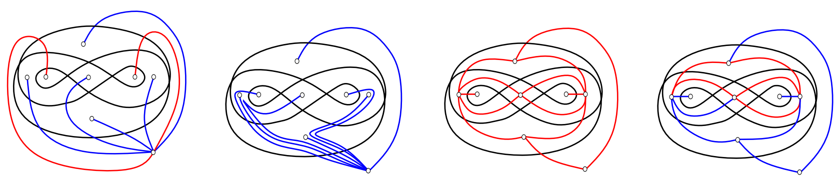
Our first observation is that the Blank words are invariant under cable isotopy; therefore the cable system can be specified up to isotopy.

Lemma 1 (Isotopy Invariance).
The reduced Blank word is invariant under cable isotopy.

Proof. Let γ be a curve. Discretize the isotopy of the cables and consider all the possible *homotopy moves* [5] performed on γ and the cables involving up to two strands from γ and a cable, because isotopy disallows the crossing of two cables. No 1↔0 move—the move that creates/destroys a self-loop—is possible as cables do not self-intersect. Any 2↔0 move which creates/destroys a bigon is in between a cable and a strand from γ, which means the two intersections must have opposite signs, and therefore the reduced Blank word does not change. Any 3→3 move which moves a strand across another intersection does not change the signs of the intersections, so while the order of strands crossing the cable changes, the order of cables crossed by γ remains the same. Thus the reduced Blank word stays the same.

We remark that we can perform an isotopy so that the Blank words are reduced even when the cables are not necessary shortest paths. In the rest of the paper, we sometimes assume Blank words to be reduced based on the context.

Manage the Cable Systems Next, we show that Blank words are well-defined once we fix the choice of basepoint p_0 and the cyclic cable ordering around p_0, as long as the cables are drawn in a reasonable way. Fix a tree-cotree pair (T, T^*) of γ, where the root of the cotree is on p_0. We say that a cable system Π is *managed* with respect to the cotree T^* if each path π_i has to be a path on T^* from the root p_0 to the leaf p_i. Given such a collection of cotree paths, one can slightly perturb them to ensure that all paths are simple and disjoint.[6] See Figure 5 for examples. Not every cable system can be managed with respect to T^*, see Figure 6 for an example.



(a) (b) (c) (d)

Fig. 5: (a) A cable system Π_1 on γ that is not managed. The red cables do not follow existing paths to the exterior face. (b) A managed cable system Π_2 on γ. (c) The dual γ^* in red. (d) The spanning tree T^* in γ^* generated by the managed cable system Π_2.

We now show that if two managed cable systems satisfying shortest path assumption with identical cable ordering around p_0, their corresponding Blank words are the same. Note that managed cable systems require a fixed tree-cotree pair. We emphasize that the shortest path assumption is necessary; one can construct two (not necessarily shortest) cable systems having the same cable ordering but different corresponding reduced Blank words (see Figure 7a and Figure 7b).

Lemma 2 (Blank Word is Unique). *Given a curve γ, if the basepoint p_0 and the cable ordering of a managed cable system Π satisfying the shortest path assumption is fixed, then the Blank word of γ is unique.*

[6] In other words, the cables are *weakly-simple* [25].

Proof. We will argue that once the basepoint p_0 and the order of cables in Π around p_0 is fixed, all the drawings of Π respecting the cable ordering lead to the same Blank word. Because the cable system is managed, the tree-cotree pair of γ are fixed and we can safely contract the primal tree and treat the graph as a collection of nested loops. If the path passes through a vertex of γ, it counts as two crossings. We prove that all the cables to the loops at certain depth have a fixed ordering by induction on the depth. This is sufficient as any two cable systems with the same cables and ordering on every loop of the same depth must be isotopic, thus by Lemma 1 their Blank words are identical. Because of the shortest path assumption, there is only one way to draw the cables to the depth-1 contours.

For any ℓ, imagine all cables of depth at least ℓ are currently drawn from p_0 to the depth-ℓ loops, where on each loop the collection of the cables are precisely those faced contained within the loop, and the cable ordering on the loops is fixed. Due to the shortest path assumption, every cable of depth ℓ has to terminate at their corresponding face. There is at most one unique way to extend each cable of depth ℓ to its representative point in the face while keeping all depth-ℓ cables disjoint and simple, up to isotopy of the cables. (If no such drawing exist, this particular cable ordering is not realizable as a cable system.) By Lemma 1, isotopy does not change the order the curve γ passing through these depth-ℓ cables. Now we partition the cables to faces of depth greater than ℓ based on children loops that contains the corresponding faces. There is one unique way to extend each cable of depth greater than ℓ to the depth-$(\ell+1)$ loops up to isotopy. Again by Lemma 1, isotopy does not change the order the curve γ passing through the cables of depth more than ℓ. By induction, the Blank word of γ is unique.

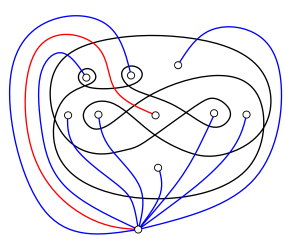
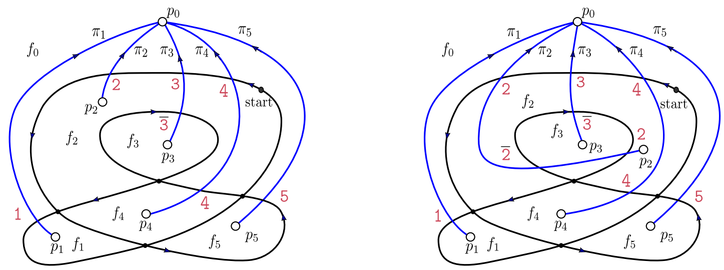


Fig. 6: An example of a cable system with a cable, in red, that cannot be managed.

Therefore, given any plane curve γ, the Blank word is well-defined (if exists), independent of the cable system after specifying a cyclic permutation of all the bounded faces of γ.

3.2 The Nie Word Construction

In an unpublished manuscript [19,20], Nie described how to compute the minimum homotopy area between any two planar closed curves using the language of



(a) (b)

Fig. 7: (a) A curve with labeled faces and edges. A shortest path cable system $\mathit{\Pi}_a$ is drawn in blue. From the indicated start position, the Blank Word is $[\gamma]_B(\mathit{\Pi}_a) =$ $[\mathtt{43254\overline{3}1}]$. (b) The same curve with a cable system $\mathit{\Pi}_b$ that does not fulfill the shortest path assumption. The Blank word is $[\gamma]_B(\mathit{\Pi}_b) = [\mathtt{43254\overline{2}\overline{3}321}]$.

geometric group theory. Nie constructed a combinatorial word representing the planar closed curve, followed by performing dynamic programming on the word based on a structure called "foldings" (see Section 4.1). But first, let us describe the word construction.

Choose a point p_i for each bounded face f_i of γ; denote the collection of points as P. Consider the punctured plane $X \coloneqq \mathbb{R}^2 \setminus P$ and its fundamental group $\pi_1(X)$. Choose a set of generators Σ for $\pi_1(X)$, where each x_i in Σ represents the generator of $\pi_1(\mathbb{R}^2 \setminus \{p_i\}) \cong \mathbb{Z}$.

Now the fundamental group $\pi_1(X)$ is a free group over such generators, and the curve γ can be represented as a word over generators of $\pi_1(X)$. However, there is more than one way to map each generator of $\pi_1(\mathbb{R}^2 \setminus \{p_i\})$ into $\pi_1(X)$, due to the fact that in order for $\pi_1(X)$ to be a group, one has to choose an endpoint x_0 and turn each closed curve in $\pi_1(X)$ into a closed path connecting to x_0. Nie never specified the choice of the connecting path because his algebraic formulation always gives the same answer under any mapping of the generators.

Nie's construction can also be interpreted combinatorially [20]. Again consider the curve γ as a four-regular plane graph. Pick a tree-cotree pair (T, T^*) of γ such that T^* is a BFS-tree; naturally the tree T is contractible. For the sake of illustration, contract T into a single point t; now each cotree edges is a single closed path at t, enclosing at least one point in P. For our purpose of proving word equivalence, there are two natural sets of generators for $X \coloneqq \mathbb{R}^2 \setminus P$:

– set of all cotree edges, and
– set of all face boundaries; i.e. sequences of cotree edges around each face containing p_i.

We now describe the change-of-basis between the two sets of generators in graph-theoretic terms. Traverse γ from some arbitrary starting point and orient each edge of γ accordingly. Now, for each face f_i, define the *boundary operator* ∂ by mapping face f_i to the signed cyclic sequence of edges around face f_i, where each edge is signed positively if it is oriented counter-clockwise and negatively otherwise.

Now, write the curve γ as a cyclic word over the cotree edges T^* by traversing γ, ignoring all tree edges in T. We perform the following procedure inductively on the cotree T^* to construct another cyclic word, this time as an element in the free group over the faces of γ. Starting from the leaves f of T^*, rewrite each edge e bounding the face f (that is, the dual of the unique edge connecting f to its parent in T^*) as a singleton word based on the index of f, with positive sign if edge e is oriented counter-clockwise, or with negative sign otherwise. Next, for any internal node f of T^*, the boundary ∂f consists of a sequence of (1) tree edges, (2) cotree edges to children of f in T^* denoted as $e_1, e_2, \ldots, e_r$, and (3) (a unique) cotree edge to parent of f denoted as e_f:

$$\partial f = [e_f e_1 e_2 \ldots e_r].$$

We can now inductively rewrite each child cotree edge e_i as a free word w_i over the faces (and ignore all tree edges). We emphasize that each word for the child cotree edge constructed inductively is a free word, not a cyclic word. Choose a particular but arbitrary way to break the cyclic sequence of faces and rewrite the equation:

$$e_f = \bar{w}_r \cdot \cdots \cdot \bar{w}_{j+1} \cdot (\bar{w}_j)' \cdot \partial f \cdot (\bar{w}_j)'' \cdot \bar{w}_{j-1} \cdot \cdots \cdot \bar{w}_1,$$

where $\bar{w}_j = (\bar{w}_j)'(\bar{w}_j)''$ is a particular way of breaking the face word $\bar{w}_j$ into two. This gives us a free word over the faces for edge e_f, and thus by induction we have rewritten γ as a free word over the faces. Finally, we can turn the free word back into a cyclic word, by observing that the cyclic permutation of the constructed free word over the faces does not affect the element we are getting in $\pi_1(X)$ (but as a side effect of choosing the basepoint p_0 of γ).

We call the resulting signed sequence of faces the *Nie word* and denoted as $[\gamma]_N(\Sigma)$, where Σ is the choices we made when breaking up the cyclic word at each cotree edge, referred to as a *cycle flattening*. Notice that the definition of $[\gamma]_N$ depends on how we choose to break the cyclic edge sequences, and thus is not well-defined without specifying the choices.

3.3 Word Equivalence

Now we are ready to prove that the two words, one defined geometrically and the other algebraically, are in fact equivalent.

Theorem 2 (Word Equivalence). *Let γ be any plane curve. For a Nie word $[\gamma]_N(\Sigma)$ with a fixed cycle flattening Σ, there is a managed cable system Π such that the Blank word $[\gamma]_B(\Pi)$ is equal to $[\gamma]_N(\Sigma)$. Conversely, any managed cable system Π induces a cycle flattening Σ such that $[\gamma]_B(\Pi)$ and $[\gamma]_N(\Sigma)$ are equal.*

Proof. First, fix a tree-cotree pair (T, T^*) for γ such that T^* is a BFS-tree. Orient the edges of the cotree T^* so that it is rooted at some fixed basepoint p_0. We prove the following statement by induction on the nodes of T^* from leaves to the root, which implies the theorem:

The Blank subword corresponding to any cotree edge e is the same as the Nie subword corresponding to e.

To prove the statement, we will construct the cables in Π gradually from each face to p_0, at each step stopping at the cotree edge e in T^*. Let f be an arbitrary non-root node in T^*, and edge e be the unique edge from f to its parent in T^*. If f is a leaf, e is the only edge in ∂f that is not in tree T. This means, when we write ∂f using edges not in T, we have $\partial f = \pm e$, with positive sign if e is oriented counter-clockwise and negative sign otherwise. We draw the cable from the representative point in face f to e; there is only one possible way to draw the cable up to isotopy.

If f is not a leaf, let $e_1, \ldots, e_r$ be other non-tree edges on ∂f besides e in counter-clockwise order around ∂f, *flipping their orientation defined by traversing γ if necessary.* By induction hypothesis, the Blank subword of e_i is the same as its Nie subword; denote the Blank (or Nie) subword of e_i as w_i. This suggests that as we traverse e_i, the cables in Π seem is exactly equal to w_i. Now we need to draw the cable π_f from the representative point of f to edge e. By construction of the Nie subword and the given cycle flattening Σ, the Blank subword on e must be of the form

$$\bar{w}_r \cdot \cdots \cdot \bar{w}_{j+1} \cdot (\bar{w}_j)' \cdot f \cdot (\bar{w}_j)'' \cdot \bar{w}_{j-1} \cdot \cdots \cdot \bar{w}_1,$$

where $\bar{w}_j = (\bar{w}_j)'(\bar{w}_j)''$ is a particular way of breaking the face word $\bar{w}_j$ into two. (See Figure 8.) Because of the shortest path assumption, the collection of symbols inside each w_i corresponds to exactly the faces contained within the region formed by cotree edge e_i and the primal tree T. Thus, we extend all the cables intersecting the edges $e_1, \ldots, e_r$ to e, and create the representative point of face f within the subregion of f bounded by the last cable in $\bar{w}_r \cdot \cdots \cdot \bar{w}_{j+1} \cdot (\bar{w}_j)'$ and the first cable in $(\bar{w}_j)'' \cdot \bar{w}_{j-1} \cdot \cdots \cdot \bar{w}_1$. This way we can draw the cable π_f so that the corresponding Blank subword is equal to the Nie subword. By induction, we have $[\gamma]_B(\Pi) = [\gamma]_N(\Sigma)$ for the constructed cable system Π, which is managed and satisfies the shortest path assumption by the choice of T^* being a BFS-tree.

From the above construction we can recover a cycle flattening from a given managed cable system satisfying the shortest path property, thus the converse holds as well.

Figure 9 gives an example demonstrating the one-to-one correspondence, for four different cable systems and cycle flattenings, on the same curve and tree-cotree pair. One consequence coming from the equivalence between two words and Lemma 2 is that Nie word is uniquely determined after knowing the subwords corresponding to cotree edges incident to the unbounded face. This is not obviously from the definition of Nie word itself.

With this equivalence in hand, for the remainder of the paper we refer to a Nie word or a Blank word of a curve γ as the *word*, denoted as $[\gamma]$ by dropping the subscripts. Keep in mind, however, that the formal equivalence holds only when the cable system is managed.

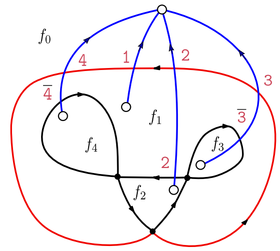


Fig. 8: As we traverse the red edge γ_r intersects $3, 2$, then 1—the cable corresponding to f_1—then 4. As we traverse the boundary of f_1 we traverse the red edge, followed by $\overline{4}$ and $\overline{23}$. This choice of cable system Π corresponds to cycle flattening at f_1 as $\overline{23}3214\overline{4}$ by writing $f_1 = \overline{e_2}e_3e_1e_4$, or equivalently, $\overline{23}3214\overline{4} = \overline{e_2}e_3e_1e_4$. The Blank subword on γ_r, with respect to the cycle flattening, is $e_1 = \overline{\overline{3}2} \cdot \overline{23}3214\overline{4} \cdot \overline{\overline{4}} = 3214$, as expected.

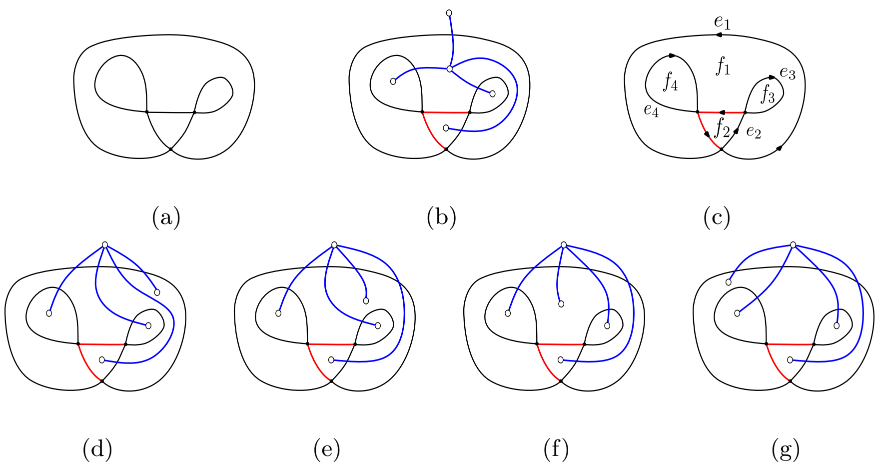


Fig. 9: (a) A curve γ. (b) A spanning tree in red and cotree in blue. (c) A labeling of the coedges and faces. We have four ways to break the cyclic face sequence for e_1, represented using the cable system Π: (d) $e_1 = \partial f_1 e_2 \overline{e_3 e_4}$; (e) $e_1 = e_2 \partial f_1 \overline{e_3 e_4}$; (f) $e_1 = e_2 \overline{e_3} \partial f_1 \overline{e_4}$; and (g) $e_1 = e_2 \overline{e_3 e_4} \partial f_1$.

4 Foldings and Self-Overlapping Decompositions

In this section, we give a geometric proof of the correctness to Nie's dynamic program. To do so, we show that the minimum homotopy area of a curve can be computed from its Blank word using an algebraic quantity of the word called the *cancellation norm*, which is independent of the drawing of the cables. We then show a minimum-area self-overlapping decomposition can be found in polynomial time.

4.1 The Cancellation Norm and Blank Cuts

Given a (cyclic) word w, a *pairing* is a letter and its inverse $(\mathtt{f}, \bar{\mathtt{f}})$ in w. Two letter pairings, $(\mathtt{f_1}, \bar{\mathtt{f}}_\mathtt{1})$ and $(\mathtt{f_2}, \bar{\mathtt{f}}_\mathtt{2})$, are *linked* in a word if the letter pairs occur in alternating order in the word, $[\cdots \mathtt{f_1} \cdots \mathtt{f_2} \cdots \bar{\mathtt{f}}_\mathtt{1} \cdots \bar{\mathtt{f}}_\mathtt{2} \cdots]$. A *folding* of a word is a set of letter pairings such that no two pairings in the set are linked. For example, in the word $[\bar{\mathtt{2}}\mathtt{3154}\mathtt{6}\bar{\mathtt{5}}\mathtt{4}\bar{\mathtt{6}}\mathtt{2}\bar{\mathtt{3}}]$ the set $\{(\mathtt{5}, \bar{\mathtt{5}}), (\bar{\mathtt{3}}, \mathtt{3})\}$ is a folding while $\{(\mathtt{5}, \bar{\mathtt{5}}), (\mathtt{6}, \bar{\mathtt{6}})\}$ is not.

The cancellation norm is defined in terms of pairings. The norm also applies in the more general setting where every letter has an associated nonnegative weight. A letter is *unpaired* in a folding if it does not participate in any pairing of the folding. For a word of length m, computing the cancellation norm takes $\mathcal{O}(m^3)$ time and $\mathcal{O}(m^2)$ space [2,21]. Recently, a more efficient algorithm for computing the cancellation norm appears in Bringmann *et al.* [3]; this algorithm uses fast matrix multiplications and runs in $\mathcal{O}(m^{2.8603})$ time.

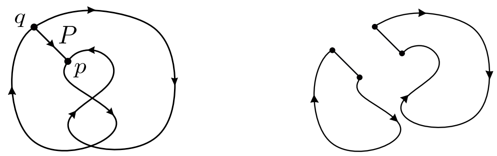
The *weighted cancellation norm* of a word w is defined to be the minimum sum of weights of all the unpaired letters in w across all foldings of w [2,21]. If w is a word where each letter $\mathtt{f_i}$ corresponds to a face f_i of a curve, we define the weight of $\mathtt{f_i}$ to be $\mathrm{Area}(f_i)$. The *area of a folding* is the sum of weights of all the unpaired symbols in a folding. The weighted cancellation norm becomes $\|w\| \coloneqq \min_{\mathcal{F}} \sum_i \mathrm{Area}(f_i)$ where $\mathcal{F}$ is the set of all foldings of w and i ranges over all unpaired letter in w.

A dynamic program, similar to the one for matrix chain multiplication, is applied on the word. Let $w = f_1 f_2 \cdots f_\ell$ where $\ell \geq 2$. Assume we have computed the cancellation norm of all subwords with length less than ℓ. Let $w' = f_1 f_2 \cdots f_{\ell-1}$. If f_ℓ is not the inverse of f_i for $1 \leq i \leq \ell - 1$, then f_ℓ is unpaired and $||w|| = ||w'|| + Area(f_\ell)$. Otherwise, f_ℓ participates in a folding and there exits at least one k where $1 \leq k \leq \ell - 1$ and $f_k = f_\ell^{-1}$. Let $w_1 = f_1 \cdots f_{k-1}$ and $w_2 = f_{k+1} \cdots f_{\ell-1}$. Then, we find the k that minimizes $||w_1||+||w_2||$. We have

$$||w|| = \min\{||w'|| + Area(f_\ell), \min_k\{||w_1|| + ||w_2||\}\}$$

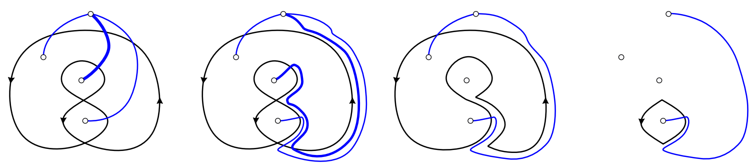
Nie shows that the weighted cancellation norm whose weights correspond to face areas is equal to the minimum homotopy area using the triangle inequality and geometric group theory. Our proof that follows is more geometric and leads to a natural homotopy that achieves the minimum area.

We now show how to interpret the cancellation norm geometrically. Let $(\mathtt{f}, \bar{\mathtt{f}})$ be a face pairing in a folding of the word $[\gamma]_B(\mathit{\Pi})$ for some cable system $\mathit{\Pi}$. Denote the cable in $\mathit{\Pi}$ ending at face f as π_f. Cable π_f intersects γ at two points corresponding to the pairing $(\mathtt{f}, \bar{\mathtt{f}})$, which we denote as p and q respectively. Let π'_f be the simple subpath of π_f so that $\pi'_f(0) = q$ and $\pi'_f(1) = p$. We call π'_f a *Blank cut* [1,10,17] (see Figure 10). Any face pairing defines a Blank cut, and the result of a Blank cut produces two curves each with fewer faces than the original curve: namely, γ_1 which is the restriction of γ from q to p following by the reverse of path π'_f, and γ_2 which is the restriction of γ from p to q followed by path π'_f.



(a) (b)

Fig. 10: (a) A curve with labeled path P. (b) The two induced subcurves from cutting along P.



(a) (b) (c) (d)

Fig. 11: (a) A curve with cables. (b) Isotopy the cables to not partially cut any faces. (c) One subcurve resulting from cutting along the middle cable. The curve is weakly simple and there are two cables in this face. (d) The other subcurve.

In order to not partially cut any face, we require all Blank cuts to occur along the boundary of the face being cut. When cutting face f_i along path π_j, we reroute all cables crossing the interior of f_i, including π_j but excluding π_i, along the boundary of f_i through an *isotopy*, so that no cables intersect π_i. Lemma 1 ensures that the reduced Blank word remains unchanged. See Figure 11 for an example. Notice that different cables crossing f_i might be routed around different sides of f_i in order to avoid intersecting cable π_i and puncture p_i. This way, we ensure the face areas of the subcurves are in one-to-one correspondence with the symbols in the subwords induced by a folding.

Using the concept of Blank cut we can determine if a curve is self-overlapping. A subword σ of w is *positive* if $\sigma = \mathtt{f_1 f_2 \ldots f_k}$, where each letter $\mathtt{f_i}$ is positive. A pairing $(\mathtt{f}, \bar{\mathtt{f}})$ is *positive* if one of the two subwords of the (cyclic) word w in between the two symbols $\mathtt{f}, \bar{\mathtt{f}}$ is positive; in other words, $w = [\mathtt{f} p \bar{\mathtt{f}} w']$ for some positive word p and some word w'. A folding of w is called a *positive folding*[7] if all pairings in w are positive, and the word constructed by replacing each positive pairing (including the positive word in-between) $\mathtt{f} p \bar{\mathtt{f}}$ in the folding with the empty string is still positive. Words that have positive foldings are called *positively foldable*. Blank established the characterization of self-overlapping curves through Blank cuts.

[7] Blank called these pairings *groupings*

Theorem 3 (Self-Overlapping Detection [1]). *Curve γ is self-overlapping if and only if γ has rotation number 1 and $[\gamma]_B(\Pi)$ is positively foldable for any shortest Π.*

However, we face a difficulty when interpreting Nie's dynamic program geometrically. In our proof we have to work with *subcurves* (and their extensions) of the original curve and the induced cable system. For example, after a Blank cut or a vertex decomposition, there might be multiple cables connecting to the same face creating multiple punctures per face, and cables might not be managed or follow shortest paths to the unbounded face (see Figure 11c and Figure 15b). In other words, the subword corresponding to a subcurve with respect to the induced cable system might not be a regular Blank word (remember that Blank word is only well-defined when the cable system is managed, all cables are shortest paths, and the cable ordering is fixed; see Section 2). To remedy this, we tame the cable system first by rerouting them into another cable system that is managed and satisfies the shortest path assumption, then merging all the cables ending at each face. We show that while such operations change the Blank word of the curve, the cancellation norm of the curve and the positive foldability does not change. We summarize the property needed below.

Lemma 3 (Cable Independence). *Let γ be any curve with two cable systems Π and Π' such that the weights of the cables in Π ending at any fixed face sum up to the ones of Π'. Then any folding F of $[\gamma](\Pi)$ can be turned into another folding F' of $[\gamma](\Pi')$, such that the area of the two foldings are identical. As a corollary, the minimum area of foldings (the cancellation norm) of $[\gamma](\Pi)$ and the existence of a positive folding of $[\gamma](\Pi)$ are independent of the choice of Π.*

Next, we prove that for each folding there is a homotopy with equal area.

Lemma 4 (Folding to Homotopy). *Let γ be a curve and Π be a managed cable system satisfying the shortest path assumption, and let F be a folding of $[\gamma](\Pi)$. There exists a null-homotopy of γ with area equal to the area of F.*

Proof. We induct on the number of pairings in F. Consider a folding F with k pairings, and let $(\mathtt{f}, \bar{\mathtt{f}})$ be a pairing in F that does not contain any other pairing in between. A Blank cut along this pairing decomposes the curve into two subcurves, at least one of which does not contain any pairings. Call this subcurve γ_1. We will contract γ_1 to the cut using a homotopy with area equal to the sum of area of all the unpaired letters in the subword of γ_1; such homotopy exist by the base case of induction. However, after we contract γ_1, the cable system on the remaining curve might no longer be managed or shortest, and there might be multiple cables in some faces. Therefore we strengthen the inductive hypothesis by assuming that for any subword of $[\gamma](\Pi)$ and its corresponding subcurve γ' of γ with the induced (unmanaged multi-)cable system, and a folding F' with less k pairings on the subword, there exists a null-homotopy of γ' with area equal to the area of the folding F', and the destination of the null-homotopy can be anywhere on γ'.

Let w' be the subword of $[\gamma](\Pi)$ by removing the subword between the pairing $(\mathtt{f}, \bar{\mathtt{f}})$, folding F' on subword w' be constructed from F by removing the pairing $(\mathtt{f}, \bar{\mathtt{f}})$. Let γ' denote the curve that results from contracting γ_1 to the cut and let Π' denote the cable system induced on γ' by Π. By the (strengthened) induction hypothesis γ' has a null-homotopy with area equal to the one of F'. Combining with the null-homotopy of γ_1 and we are done.

For the base case when F is the empty pairing, if the cables are not managed, construct a managed cable system satisfying the shortest path assumption Π^* for γ and apply Lemma 3; the sum of weights of all letters in the new Blank word remains unchanged. To construct the null-homotopy, decompose the curve into depth cycles by performing a smoothing at each vertex [5], to obtain a null-homotopy with area equal to the depth area. Since each cable in Π^* follows a shortest path to the unbounded face, the number of times an unsigned letter appears in the word is equal to the depth of the face, and thus the homotopy area is equal to the area of the folding. In fact, we can choose any point on the curve to be the destination of the null-homotopy.

What we are left with is to prove Lemma 3: Given a curve γ and a cable system Π, if we have a folding on the face word $[\gamma](\Pi)$, there is an equivalent folding of the new word $[\gamma](\Pi')$ that has the same area if we choose a different cable system. As a result, the cancellation norm and the positive foldability of a word are independent to the cable system chosen. Notice that it is sufficient to assume Π' to be a managed cable system with shortest path assumptions and single cable to each face.

Any two cable systems can always be connected by a sequence of isotopy and order switching between two adjacent cables, followed by a redrawing of the cables induced by a homeomorphism of the plane fixing the punctures $\{p_i\}$, which fixes the cable ordering and the disjointness between cables but the isotopy classes of the cables change. We emphasize that either operation will change the Blank word, but we can always find a folding of the new word that preserves the area.

Next, we show that switching two adjacent cables preserves the area and positivity of the folding

Lemma 5 (Switch Invariance). *Given curve γ, a cable system Π, and a folding F on $[\gamma](\Pi)$. Switching the order of two adjacent cables produces another folding on the new word with equivalent area. Furthermore, the new folding is positive if and only if F is.*

Proof. Consider two cables π_f and π_g in Π that are adjacent in the rotation system. Assume π_g is in the clockwise direction of π_f in the rotation system and we are trying to move π_f across π_g. Let $w := [\gamma]_B(\Pi)$ denote the Blank word before the order of the cables are switched and w' the word after the cables are switched; by Lemma 1 the words are well-defined up to isotopy classes of π_f and π_g. First, we argue that, without loss of generality, we can assume the drawing of π_g follows π_f in an ε-neighborhood before continuing towards puncture p_g, by drawing π_g still counter-clockwise to and ε-close to π_f from p_0

to p_f and back to p_0, followed by the original drawing of π_g. The new drawing is disjoint from the rest of the cables and is isotopic to the original drawing. This way we can ensure any instance of symbol $\mathtt{f}$ in w is followed immediately by a $\mathtt{g}$, and any instance of $\bar{\mathtt{f}}$ has a $\bar{\mathtt{g}}$ proceeding in the word. (Notice we cannot necessarily say the same about $\mathtt{g}$ because cable π_g continues after reaching p_f.)

Let F be a folding of w, we construct a folding F' of w' with equal area. We separate into two cases. If no instance of $(\mathtt{f}, \bar{\mathtt{f}})$ is in F, then all pairings in F can be paired in W' and we can set $F' := F$. Now we assume there is an instance of $(\mathtt{f}, \bar{\mathtt{f}})$ in F. We further split into subcases. Look at the two $\mathtt{g}/\bar{\mathtt{g}}$ symbols adjacent to the $(\mathtt{f}, \bar{\mathtt{f}})$ pair in F. If the two $\mathtt{g}$ symbols is a pairing in F, then we can again set $F' := F$. If exactly one of the $\mathtt{g}/\bar{\mathtt{g}}$ adjacent to the $(\mathtt{f}, \bar{\mathtt{f}})$ pair is the paired with another $\bar{\mathtt{g}}$ or $\mathtt{g}$ that is not adjacent to the $(\mathtt{f}, \bar{\mathtt{f}})$ pair in w, then switching the cables naïvely results in a linked pair,

$$w = [\ldots \mathtt{f}\mathtt{g} \ldots \mathtt{g} \ldots \bar{\mathtt{g}}\bar{\mathtt{f}} \ldots] \longrightarrow w' = [\ldots \mathtt{g}\mathtt{f} \ldots \mathtt{g} \ldots \bar{\mathtt{f}}\bar{\mathtt{g}} \ldots].$$

Instead, after the cables are switched, we pair the $\mathtt{g}$ and $\bar{\mathtt{g}}$ that appear next to the $(\mathtt{f}, \bar{\mathtt{f}})$ pair together in F':

$$w' = [\ldots \mathtt{g}\mathtt{f} \ldots \mathtt{g} \ldots \bar{\mathtt{f}}\bar{\mathtt{g}} \ldots].$$

The new F' is a folding without linked pairs and does not change the area because the same number and types of symbols are paired. Finally, if both $\mathtt{g}$ and $\bar{\mathtt{g}}$ adjacent to the $(\mathtt{f}, \bar{\mathtt{f}})$ pair are paired with letters not adjacent to $\bar{\mathtt{f}}$ and $\mathtt{f}$ in F, we pair $\mathtt{g}$ next to the $\mathtt{f}$ with the $\bar{\mathtt{g}}$ next to $\bar{\mathtt{f}}$ and the $\mathtt{g}$ not adjacent to the $\mathtt{f}$ with the $\bar{\mathtt{g}}$ not adjacent to $\bar{\mathtt{f}}$ in F':

$$w = [\ldots \mathtt{f}\mathtt{g} \ldots \bar{\mathtt{g}} \ldots \mathtt{g} \ldots \bar{\mathtt{g}}\bar{\mathtt{f}} \ldots] \longrightarrow w' = [\ldots \mathtt{g}\mathtt{f} \ldots \bar{\mathtt{g}} \ldots \mathtt{g} \ldots \bar{\mathtt{f}}\bar{\mathtt{g}} \ldots].$$

Again, F' remains a folding without linked pairs and has the same area as F.

Next, we show that the cancellation norm does not depend on the cables having the shortest path property. While the face word is unique up to isotopy of the cables based on Lemma 1, two cable systems with identical ordering around p_0 may not be isotopic to each other. The theory of mapping class groups [11,16] provides tools to convert between all possible isotopy classes of cable systems with same cable ordering.

Let $S_{g,n}$ denote a surface with genus g and n punctures. Let *Diffeo*$^+(S_{g,n})$ denote the group of all orientation-preserving diffeomorphisms from $S_{g,n} \to S_{g,n}$. Define an equivalence relation $\sim$ on $\mathrm{Diffeo}^+(S_{g,n})$, where $\phi \sim \psi$ if there exists an isotopy between ϕ and ψ. The *mapping class group* of $S_{g,n}$ is the group $\mathrm{MCG}(S_{g,n}) = \mathrm{Diffeo}^+(S_{g,n})/\sim$. A simple closed curve or a puncture-to-puncture arc λ on a surface is *nonseparating* if the surface remains connected after cutting along λ. A simple closed curve (or arc) in a surface is *essential* if it is not homotopic to a point and not homotopic to a puncture.

Let λ be a simple closed curve in a surface. We now describe a particular mapping class in $\mathrm{MCG}(S_{g,n})$, a *Dehn twist* — about λ. LetA_λ be an arbitrarily thin annulus homeomorphic to the product of $\mathbb{S}^1$ and the unit interval I; i.e. $A_\lambda \cong \mathbb{S}^1 \times I$ with $\lambda \cong \mathbb{S}^1 \times \{0\}$. The twist is the map $T_\lambda : A_\lambda \to A_\lambda$ where $T_\lambda(\theta, t) = (\theta + 2\pi t, t)$ [11]. Intuitively, we can think of T_λ as acting on any path that crosses A_λ by fixing one boundary component of the annulus and rotating the other boundary component one full revolution, dragging all masses within A_λ along. See Figure 12 for an example. Dehn twists are well-defined as element in $\mathrm{MCG}(S_{g,n})$ and depend only on the isotopy class of λ [16].

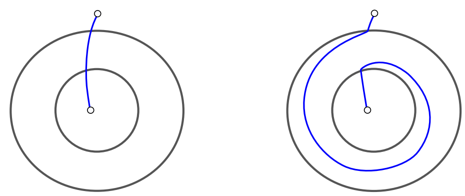


Fig. 12: An illustration of a Dehn twist.

Let D_n be the closed disk $\mathbb{D}^2$ with n punctures. The *pure mapping class group* of D_n, $\mathrm{PMCG}(D_n)$, is the subgroup of $\mathrm{MCG}(D_n)$ that fixes all punctures. The group $\mathrm{PMCG}(D_n)$ is generated by a finite number of closed curves.

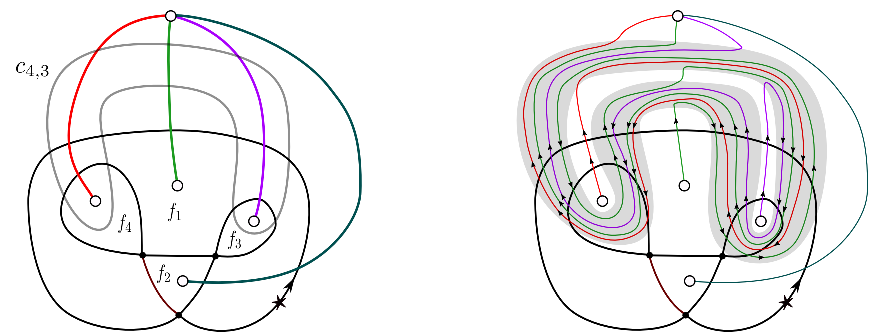
Theorem 4 (Generators [11, §9.3]). $\mathrm{PMCG}(D_n)$ *is generated by Dehn twists about the set of simple closed curves that surround exactly two punctures.*

Thus, given any cable system we can perform Dehn twists about these generators to obtain another cable system with the same cable ordering such that all cables are shortest paths. We choose a disk in the plane that contains the curve γ, the punctures, and the cables. (One subtlety is that even when the two cable systems have the same *cyclic* ordering, the choice of the disk may produce different *linear* cable ordering when we decide the location to take the containing disk. This can be easily resolved by redrawing some of the cables using isotopy across the infinity.) We now show that performing Dehn twists about suitable simple closed curves does not change the cancellation norm of the word.

Lemma 6 (Twist Invariance). *Let γ be a curve and Π be a cable system, and let F be a folding on the corresponding face word $[\gamma](\Pi)$. Dehn twists about the cables in Π produce another folding on the new word with equal area.*

Proof. Let Π be a cable system on a closed curve γ. Given two punctures p_i and p_j, let $c_{i,j}$ denote the closed curve that contains only p_i and p_j and follows the cables π_i and π_j to the exterior face, then connect. See Figure 13a for an illustration. There may be cables between π_i and π_j in the rotation order about p_0. We bundle all such cables together and denote the corresponding face word B. By Theorem 4, the closed curves $c_{i,j}$s generate the pure mapping class group. To show that twisting about $c_{i,j}$ does not change the norm, let w be the word generated by the cables before twisting about $c_{i,j}$ and let w' be the word generated by the cables after the twist. We will show that $||w'|| \leq ||w||$ and $||w|| \leq ||w'||$.

Instances of $\mathtt{f_i}$ and $\bar{\mathtt{f}}_\mathtt{i}$ in w become $(\bar{\mathtt{f}}_\mathtt{i}\bar{B}\bar{\mathtt{f}}_\mathtt{j}B)\mathtt{f_i}(\bar{B}\mathtt{f_j}B\mathtt{f_i})$ and $(\bar{\mathtt{f}}_\mathtt{i}\bar{B}\bar{\mathtt{f}}_\mathtt{j}B)\bar{\mathtt{f}}_\mathtt{i}(\bar{B}\mathtt{f_j}B\mathtt{f_i})$ in w'. Instances of $\mathtt{f_j}$ and $\bar{\mathtt{f}}_\mathtt{j}$ in w become $(B\bar{\mathtt{f}}_\mathtt{i}\bar{B}\bar{\mathtt{f}}_\mathtt{j})\mathtt{f_j}(\mathtt{f_j}B\mathtt{f_i}\bar{B})$ and $(B\bar{\mathtt{f}}_\mathtt{i}\bar{B}\bar{\mathtt{f}}_\mathtt{j})\bar{\mathtt{f}}_\mathtt{j}(\mathtt{f_j}B\mathtt{f_i}\bar{B})$ in w'. See Figure 13 for an example.



(a) Before the twist.

(b) After the twist.

Fig. 13: (a) An example of the Dehn twist. The curve is shown in black. Starting from the star and following the indicated orientation, the Blank word is $w = [\mathtt{2314 2\overline{34}}]$. In the instance we have $i = 4$ and $j = 3$; the curve $c_{4,3}$ is shown in gray. The word B is $\mathtt{1}$. (b) After twisting about an annulus formed by fattening $c_{4,3}$ the new word is $w' = [\mathtt{21\overline{413}3314\overline{1}1\overline{413}14\overline{1}31421\overline{4133}314\overline{1413}1\overline{41}314}]$.

We now construct a folding F' of w' with equal area. All pairs in F are in F'. All letters added by the twist are then paired. If $\mathtt{f_i}$ is in w and unpaired in F then we pair the conjugate letters on each side of $\mathtt{f_i}$ in w'. If $\mathtt{f_i}$ is in w and paired in F then we pair the added letters to the right of $\mathtt{f_i}$ with the added letters to the left of the paired $\bar{\mathtt{f}}_\mathtt{i}$ in w'; similarly, we pair the added letters to the left of $\mathtt{f_i}$ with the added letters to the right of $\bar{\mathtt{f}}_\mathtt{i}$. This can be done because the added words are the same mirror pair around symbol $\mathtt{f_i}$ and $\bar{\mathtt{f}}_\mathtt{i}$.

$$w' = [\ldots(\bar{\mathtt{f}}_\mathtt{i}\bar{B}\bar{\mathtt{f}}_\mathtt{j}B)\overbrace{\mathtt{f_i}(\bar{B}\mathtt{f_j}B\mathtt{f_i})\ldots(\bar{\mathtt{f}}_\mathtt{i}\bar{B}\bar{\mathtt{f}}_\mathtt{j}B)\bar{\mathtt{f}}_\mathtt{i}}(\bar{B}\mathtt{f_j}B\mathtt{f_i})\ldots]$$

where the overbracket indicates the pairing $(\mathtt{f_i}, \bar{\mathtt{f}}_\mathtt{i})$. Thus, a folding of w' of equal area exist. Because the cancellation norm is computed by taking minimum over all foldings, we have $||w'|| \leq ||w||$.

On the other hand, given a folding F' of w' we construct a folding F of w with area at most the area of F'. If an element of w is unpaired in F' it is also unpaired in F. There are three types of pairings in F'. If both letters in the pair are in w: in which case we add this pair to F. If neither letter in the pair is in w, in which case we do nothing to F. The third type of pairing contains a letter of w paired with a letter in that only presents in w' but not in w; call this pair $(\mathtt{f}^0, \bar{\mathtt{f}}^1)$ where $\mathtt{f}^0$ is in w and $\bar{\mathtt{f}}^1$ is in $w' \backslash w$. We will show that there is a unique unpaired letter $\mathtt{f}^k \in w' \backslash w$ that is left unpaired in F, or there is another $\bar{\mathtt{f}}^k \in w$ that we can pair $\mathtt{f}^0$ with in F.

We construct a sequence of letters in w' that are all $\mathtt{f}$ and $\bar{\mathtt{f}}$s and terminates with either an unpaired letter in $w' \backslash w$ or a paired letter in w. Let P denote the function that maps a letter to its paired letter in F', or the identity otherwise. The map C maps a letter of $w' \backslash w$ to its conjugate inverse in the mirror pair added by the Dehn twist if the conjugate inverse has not yet been visited, or the identity otherwise. Note that both P and C are injective, and they only map a symbol $\mathtt{f}$ to its inverse $\bar{\mathtt{f}}$ and vice versa. Beginning with the pairing $(\mathtt{f}^0, \bar{\mathtt{f}}^1)$, we alternate in applying the functions P and C: for any integer i, let $\mathtt{f}^{2i} := C(\mathtt{f}^{2i-1})$ and $\mathtt{f}^{2i+1} := P(\mathtt{f}^{2i})$. Since the word is finite, this unique sequence terminates with either an unpaired letter $\mathtt{f}^k$ of $w' \backslash w$ or an element $\bar{\mathtt{f}}^k$ in w.

- If the sequence of letters ends with an unpaired letter in $w' \backslash w$, this letter uniquely corresponds to $\mathtt{f}^0$ by following pairings and corresponding inverses in F'. For example, let F' be the folding indicated by the overbrackets, with the underbrackets denoting corresponding conjugate inverse letters,

$$[\mathtt{21}\overline{\mathtt{413}}\mathtt{3314}\overline{\mathtt{1}}\mathtt{1}\overline{\mathtt{413}}\mathtt{14}\overline{\mathtt{1}}\mathtt{31421}\overline{\mathtt{4133}}\mathtt{314}\overline{\mathtt{1413}}\mathtt{1}\overline{\mathtt{41}}\mathtt{314}].$$

  The unique unpaired letter corresponding to $\mathtt{3}$ is $\mathtt{3}$ found by considering the second corresponding inverse letter.
- If the sequence of letters ends with a letter $\mathtt{f}^k$ in w, we pair $\mathtt{f}^0$ and $\bar{\mathtt{f}}^k$ in F. This does not create any linked pairs in F since $\mathtt{f}^0$ and $\bar{\mathtt{f}}^k$ participate in pairings in F', they are not linked by any other pairing in F'. For example, let F' be the folding indicated by the overbrackets, with the underbrackets denoting corresponding conjugate inverse letters,

$$[\mathtt{21}\overline{\mathtt{413}}\mathtt{3314}\overline{\mathtt{1}}\mathtt{1}\overline{\mathtt{413}}\mathtt{14}\overline{\mathtt{1}}\mathtt{31421}\overline{\mathtt{4133}}\mathtt{314}\overline{\mathtt{1413}}\mathtt{1}\overline{\mathtt{41}}\mathtt{314}].$$

  The pairing $(\mathtt{3}, \bar{\mathtt{3}})$ is then added to F.

The sum of the areas of the unpaired letters of F is at most the sum of the areas of the unpaired letters of F' and thus $||w|| \leq ||w'||$. This proves the lemma.

We show the invariance for positive foldability

Lemma 7. *Being positively foldable does not depend on Dehn twists.*

Proof. Let w be a word and let w' be the word after a Dehn twist about the simple closed curve $c_{i,j}$ as in the proof of Lemma 6. Suppose w is positively foldable with folding F. Then, the folding F' described in Lemma 6 is a positive folding of w'.

On the other hand, suppose w' has a positive folding F'. Then, each element $\bar{\mathtt{f}}$ of w' is paired with an element $\mathtt{f}$. If both $\bar{\mathtt{f}}$ and $\mathtt{f}$ are in w then we pair them. If neither $\bar{\mathtt{f}}$ nor $\mathtt{f}$ are in w then we ignore them. If one of $\bar{\mathtt{f}}$ and $\mathtt{f}$ are in w and the other is added to w' by the twist, then, since letters are added to w' in pairs, there is another element of w paired with an element of $w' \backslash w$. We can pair the two element in w and ignore the two in $w' \backslash w$ because the pairings in F' are not linked. The resulting pairings give a positive folding of w.

Together, Lemma 5 and Lemma 6 imply the cancellation norm does not depend on the isotopy class of the cables. Moreover, Lemma 5 and Lemma 7 imply the positive foldability of the word does not depend on the isotopy class of the cables. One last subtlety: we show that we can handle multiple cables per face.

Handling multiple cables per face. When a face f in a subcurve contains multiple cables and punctures, we treat the punctures as distinct when applying Lemma 5 and Lemma 6. Now, once all the cables follow the cotree shortest paths, consider all cables ending in an arbitrary face f. If they are separated by other cables in the cyclic ordering, we use Lemma 5 and isotopy to move all the cables not ending in f out of the way until they no longer intersect the face. This makes sure that all the cables ending in f are gathered together in the cyclic ordering. Because they follow the same unique cotree path to the exterior, we can merge them into a single cable, where the corresponding symbols in the word are merged into a single symbol with the combined weight. Thus, the subwords induced by a face pairing correspond to the subcurves generated by the Blank cut.

4.2 Compute Min-Area Homotopy from Self-Overlapping Decomp.

A self-overlapping decomposition is a vertex decomposition where each subcurve is self-overlapping [12]. By Theorem 1, there exists a self-overlapping decomposition and an associated homotopy whose area is equal to the minimum homotopy area of the original curve.

In order to relate vertex decompositions and face decompositions, we define a word that includes both the faces and vertices. Given any curve γ and cable system Π, traverse γ and record both self-crossings and (signed) cable intersections; we call the resulting sequence of vertices and faces the *combined word* $[[\gamma]](\Pi)$. See Figure 14 for an example.

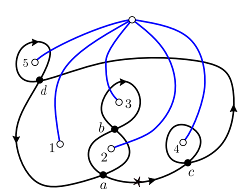


Fig. 14: A curve with combined word [c4c4231d$\bar{5}$da2b$\bar{3}$ba].

We now show that every self-overlapping decomposition (with respect to the vertex word of γ) determines a folding (of the face word of γ) using the combined word.

Theorem 5 (S-O Decomp. to Folding). *Given a self-overlapping decomposition Γ and a cable system Π of γ, there exists a folding F of $[\gamma](\Pi)$ whose area is* $\mathrm{Area}_\Gamma(\gamma)$.

Proof. Begin with the combined word $[[\gamma]](\Pi)$. Decompose $[[\gamma]](\Pi)$ at the vertices given by the self-overlapping decomposition. Let $\Gamma = \{\gamma_1, \gamma_2, \ldots, \gamma_s\}$ be the self-overlapping subcurves and $[[\gamma]](\Pi)_i$ be the corresponding subwords of $[[\gamma]](\Pi)$. If we remove the vertex symbols and turn each $[[\gamma]](\Pi)_i$ into a face word $[\gamma_i]'$, such word may not correspond to Blank words of the subcurves; indeed, when decomposing γ into subcurves by Γ, the subcurve along with the relevant cables may contain multiple cables per face and cables might not be managed or follow shortest paths. See Figure 15 for an example. However, we can first tame the cable system by choosing a new managed cable system Π^* where the cables follow shortest paths and has one cable per face (as in Section 3.1). Lemma 3 ensures that the cancellation norm and positive foldability of the subcurve remain unchanged. Denote the new face word of γ_i with respect to Π^* as $[\gamma_i] = [\gamma_i](\Pi^*)$.

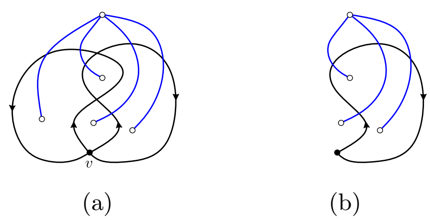


Fig. 15: We decompose the curve in (a) at vertex v into self-overlapping subcurves, the cable system on the induced subcurve in (b) has more than one marked point in a face and cables do not follow shortest paths.

Since each γ_i is a self-overlapping subcurve in Γ, we can find a positive folding F_i of $[\gamma_i]$ by Theorem 3, and the minimum homotopy area of γ_i is equal to the area of folding F_i. Now Lemma 3 implies that the subword $[\gamma_i]'$ from the original combined word also has a positive folding F_i' whose area is equal to the minimum homotopy area of γ_i. By combining all foldings F_i' of each face subword $[\gamma_i]'$, we create a folding F for $[\gamma](\Pi)$ (no pairings between different F_i's can be linked). The area of folding F is equal to the sum of areas of foldings F_i', which in turns is equal to $\sum_i \mathrm{Area}_H(\gamma_i)$, that is, the homotopy area of self-overlapping decomposition $\mathrm{Area}_\Gamma(\gamma)$. This proves the theorem.

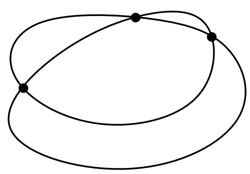
Corollary 1 (Geometric Correctness). *The dynamic programming algorithm computes the minimum-area homotopy for any curve γ.*

Proof. By Theorem 1, there exists a self-overlapping decomposition with minimum homotopy area. By Theorem 5, some folding achieves a minimum area. Using Lemma 4, the minimum-area folding produces a minimum-area homotopy.

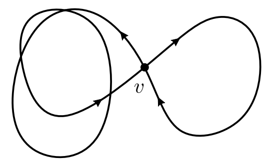
4.3 Min-Area Self-Overlapping Decomposition in Polynomial Time

Finally, we show how to construct a self-overlapping decomposition from a maximal folding with equal area. Before we begin the proof, we include definitions that will help describe the types of curves we encounter. A curve γ is a *k-stack* if it has rotation number k, all bounded faces have positive winding numbers, and $\mathrm{Area}_H(\gamma) = \mathrm{Area}_W(\gamma)$ (see Figure 16a for any example).

Any k-stack has a vertex decomposition into k self-overlapping curves [9, Theorem 2.15].[8] A curve is a *$-k$-stack* if its reversal is a $+k$-stack. We called a curve γ a *stack* if γ is a $\pm k$-stack. A curve is *good* [9] if the depth of each face is equal to the absolute value of the winding number; any good curve must have $\text{Area}_D(\gamma) = \text{Area}_H(\gamma) = \text{Area}_W(\gamma)$ by Equation 1. Good curves are almost stacks, except that some faces might not have positive winding numbers. A vertex of γ is *sign-changing* if the four incident faces have winding numbers $[1, 0, -1, 0]$ in cyclic order (see Figure 16b for an example). If we smooth a sign-changing vertex of a good curve, the two induced subcurves remain good. We can always decompose a good curve into a collection of stacks at all the sign-changing vertices [9, Theorem 5.7].



(a) A 2-stack.



(b) Sign-changing vertex.

Fig. 16: (a) A k-stack where $k = 2$. Smoothing at any of the three vertices gives a self-overlapping decomposition. (b) The vertex v is a sign-changing vertex.

With these definitions in hand, we now show how to construct a self-overlapping decomposition from a maximal folding with equal area.

Theorem 6 (Folding to S.O.D.). *Let γ be a curve and Π be a cable system. Given a maximal folding F of $[\gamma](\Pi)$, there is a self-overlapping decomposition of γ whose area is equal the area induced by the folding F.*

Proof. Let F be any maximal folding of $[\gamma](\Pi)$. Without loss of generality we can assume that the cable system does not cut through the interior of any face by rerouting the cables similar to Section 4.1. By Lemma 1 the Blank word remains unchanged.

Let γ_g be an arbitrary subcurve generated after Blank cutting along the pairings in F. The curve γ_g is must be good: otherwise, there is a face with winding number not equal to its depth, and thus γ_g would cross the corresponding cable from left to right and from right to left. Therefore, we can introduce an extra pair into the folding and F remains unlinked; thus F would not be maximal. Decomposing at all the sign-changing vertices [9, Theorem 5.7] turns γ_g into a collection of stacks, each of which can be further decomposed into collection of self-overlapping curves [9, Theorem 2.15].

The area of the folding F is equal to the number of unpaired faces in $[\gamma](\Pi)$, which is also equal to the sum of depth area of each good subcurve. Since each subcurve γ_g of folding is good, the depth area of γ_g is equal to its winding

[8] k-stacks are called *interior-boundaries* by Titus [24] and the terminology was used in various previous work [17,12,9]. The name *k-stack* is well-justified as such curve is the boundary of a stack of k disks [9].

area. Any additional decomposition of γ_g into self-overlapping curves respects the additivity of winding areas. Thus, the area of folding F is equal to the area induced by our chosen self-overlapping decomposition.

The above theorem implies a polynomial-time algorithm to compute a self-overlapping decomposition with minimum area.

Corollary 2 (Polynomial Optimal Self-Overlapping Decomposition). *Let γ be a curve. A self-overlapping decomposition of γ with area equal to minimum homotopy area of γ can be found in polynomial time.*

Proof. Apply the dynamic programming algorithm to compute the minimum-area folding F for $[\gamma](\Pi)$ with respect to some cable system Π. By Theorem 5 the area of F is equal to the minimum homotopy area of γ, and so does the corresponding self-overlapping decomposition given by Theorem 6.

Acknowledgements Brittany Terese Fasy and Bradley McCoy are supported by NSF grant DMS 1664858 and CCF 2046730. Carola Wenk is supported by NSF grant CCF 2107434 .

References

1. Samuel Joel Blank. *Extending Immersions and Regular Homotopies in Codimension 1*. PhD thesis, Brandeis University, May 1967.
2. Michael Brandenbursky, Światosław Gal, Jarek Kędra, and Michałl Marcinkowski. The cancelation norm and the geometry of bi-invariant word metrics. *Glasg. Math. J.*, page 153–176, 2015.
3. Karl Bringmann, Fabrizio Grandoni, Barna Saha, and Virginia Vassilevska Williams. Truly subcubic algorithms for language edit distance and RNA folding via fast bounded-difference min-plus product. *SIAM J. Comput.*, 48(2):481–512, 2019.
4. Erin Chambers and Yusu Wang. Measuring similarity between curves on 2-manifolds via homotopy area. *29th ACM Symp. Comput. Geom.*, pages 425–434, 2013.
5. Hsien-Chih Chang and Jeff Erickson. Untangling planar curves. *Discrete Comput. Geom.*, 58:889, 2017.
6. Dennis Frisch. Extending immersions into the sphere. 2010. URL: http://arxiv.org/abs/1012.4923.
7. David Eppstein and Elena Mumford. Self-overlapping curves revisited. *20th ACM-SIAM Symp. Discrete Algorithms*, pages 160–169, 2009.
8. Jeff Erickson. One-dimensional computational topology lecture notes. Lecture 7, 2020. URL: https://mediaspace.illinois.edu/channel/CS+598+JGE+âĂŤ%C2%A0Fall+2020/177766461/.
9. Parker Evans. On Self-Overlapping Curves, Interior Boundaries, and Minimum Area Homotopies. Bachelor's thesis, Tulane University, 2018.
10. Parker Evans, Brittany Terese Fasy, and Carola Wenk. Combinatorial properties of self-overlapping curves and interior boundaries. *36th ACM Symp. on Comput. Geom.*, 2020.
11. Benson Farb and Dan Margalit. *A Primer on Mapping Class Groups*. Princeton University Press, 2011.
12. Brittany Terese Fasy, Selçuk Karakoç, and Carola Wenk. On minimum area homotopies of normal curves in the plane. 2017. URL: http://arxiv.org/abs/1707.02251.
13. Carl Friedrich Gauss. Nachlass. I. Zur geometria situs. *Werke*, vol. 8, 271–281, 1900.
14. Selçuk Karakoç. *On Minimum Homotopy Areas*. PhD thesis, Tulane University, 2017.
15. Yijing Li and Jernej Barbič. Immersion of self-intersecting solids and surfaces. *ACM Trans. on Graph.*, 45:1–14, 2018.
16. Bruno Martelli. An introduction to geometric topology, 2016. arXiv:arXiv:1610.02592.
17. Morris L. Marx. Extensions of normal immersions of S^1 into R. *Trans. Amer. Math. Soc.*, 187:309–326, 1974.
18. Uddipan Mukherjee. Self-overlapping curves: Analysis and applications. *Comput.-Aided Des.*, 40:227–232, 2014.
19. Zipei Nie. On the minimum area of null homotopies of curves traced twice. 2014. URL: http://arxiv.org/abs/1412.0101.
20. Zipei Nie. Private correspondence. 2016.
21. Ruth Nussinov and Ann B. Jacobson. Fast algorithm for predicting the secondary structure of single-stranded RNA. *Proc. Natl. Acad. Sci. USA*, 77(11):6309–6313, 1980.

22. Valentin Poénaru. Extension des immersions en codimension 1 (d'après Samuel Blank). *Séminaire N. Bourbaki (1966–1968)*, 10:473–505, 1968.
23. Peter Shor and Christopher Van Wyk. Detecting and decomposing self-overlapping curves. *Comput. Geom.: Theory and Applications*, 2:31–50, 1992.
24. Charles J. Titus. The combinatorial topology of analytic functions on the boundary of a disk. *Acta Math.*, pages 106(1–2):45–64, 1961.
25. Godfried Toussaint. On separating two simple polygons by a single translation. *Discrete Comput. Geom.*, 4(3):265–278, June 1989.
26. Hassler Whitney. On regular closed curves in the plane. *Compos. Math.*, 4:276–284, 1937.