THE SIZE OF SPANNING DISKS FOR POLYGONAL CURVES

JOEL HASS, JACK SNOEYINK, AND WILLIAM P. THURSTON

Abstract. For each integer \( n \geq 0 \), there is a closed, unknotted, polygonal curve \( K_n \) in \( \mathbb{R}^3 \) having less than \( 10n + 9 \) edges, with the property that any Piecewise-Linear triangulated disk spanning the curve contains at least \( 2^{n-1} \) triangles.

1. Introduction.

Let \( K \) be a closed polygonal curve in \( \mathbb{R}^3 \) consisting of \( n \) line segments. Assume that \( K \) is unknotted, so that it is the boundary of an embedded disk in \( \mathbb{R}^3 \). This paper considers the question: How many triangles are needed to triangulate a Piecewise-Linear (PL) spanning disk of \( K \)? The main result, Theorem 1 below, exhibits a family of unknotted polygons with \( n \) edges, \( n \to \infty \), such that the minimal number of triangles needed in any triangulated spanning disk grows exponentially with \( n \). More specifically, we construct a sequence of unknotted simple closed curves \( K_n \) in \( \mathbb{R}^3 \) having the following properties for each \( n \geq 0 \).

- The curve \( K_n \) is an unknotted polygon with at most \( 10n + 9 \) edges.
- Any PL embedding of a triangulated disk into \( \mathbb{R}^3 \) with boundary \( K_n \) contains at least \( 2^{n-1} \) triangular faces.

The polygons \( K_1 \) and \( K_3 \) are pictured in Figure 1.

The existence of these curves is related to the complexity of certain topological algorithms. Algorithms to test knot triviality by a search for embedded PL spanning disks are searching for disks that can be exponentially more complicated than their boundary curves. Algorithms of this type include those described in [5], [3], [6], [9], [4]. Some approaches to problems in computational group theory, such as the word problem, are also based on a search for a spanning disk, and may face similar difficulties.

The lower bound given in our examples can be compared with the following upper bound: the results of [9] and [10] show that every unknotted polygon with at most \( n \) edges in \( \mathbb{R}^3 \) bounds a PL embedded triangulated disk which has at most \( Cn^2 \) triangles, where \( C > 1 \) is a constant independent of \( n \). The exponent \( n^2 \) comes from the requirement that the polygon be embedded in the 1-skeleton of a triangulated 3-manifold. A triangulation that contains an \( n \)-edge polygon and using \( O(n^2) \) tetrahedra always exists, and this bound cannot always be improved, see Avis and El Gindy [1]. On the other hand, [1] also shows that a set of points

---

*Date*: March 30, 2022.

1991 *Mathematics Subject Classification*. Primary 57M25; Secondary 53A10.

*Key words and phrases.* knot theory, triangulations, combinatorial complexity, computational topology.

Partially supported by NSF grant DMS-0072348 and the Institute for Advanced Study.

Partially supported by grants from NSERC.

Partially supported by NSF grant DMS-0072540.
in general position, i.e. one for which no four points lie on a plane and no three on a line, can be triangulated using $O(n)$ simplices. It seems plausible that one could obtain an improved upper bound of $C^n$ triangles for a PL spanning disk of a polygon whose vertices are in general position.

A result similar to the one proved in this paper was announced in [14], but the geometric analysis suggested there seems difficult to establish rigorously. We consider here a different set of polygonal curves $K_n$ than those used in [14], and to establish their properties use topological arguments based on ideas from the classification of diffeomorphisms of surfaces [15] and from Morse theory [12].

Although the main result concerns PL curves and surfaces, in some parts of the proof it becomes convenient to work with smooth surfaces and smooth mappings. This allows use of basic results from smooth Morse theory. The arguments could be carried out entirely in the PL context, at the expense of using less well known versions of Morse theory. Passing between the PL and smooth settings is achieved by approximating PL maps by smooth maps.

2. Construction of $K_n$

We now describe how to construct the unknotted curves $K_n$. The curve $K_0$ is contained in the $xz$-plane, see Figure 3. The construction of $K_n$ begins with the PL 4-braid $\alpha$ depicted in Figure 2, where $\alpha = \sigma_1\sigma_2^{-1}$ in terms of the standard generators $\sigma_1, \sigma_2, \sigma_3$ of the braid group on four strands, see [2]. This braid consists of four arcs running between the planes $\{z = 1\}$ and $\{z = 0\}$, along each of which $z$ is monotonically decreasing. The planes $\{z = 0\}, \{z = 1\}$ each intersect $\alpha$ at four points. We arrange these points along the $x$-axis and label them by $p_1 = (-2, 0), p_2 = (-1, 0), p_3 = (1, 0), p_4 = (2, 0)$. In this labeling we only consider the $xy$-coordinates.

A diffeomorphism $\varphi$ of the 4-punctured plane $\mathbb{R}^2 \setminus \{p_1, p_2, p_3, p_4\}$ is associated to the braid $\alpha$. This diffeomorphism is induced by taking the punctured plane at level $z = 1$ and sliding it down the braid to level $z = 0$. Its action on the plane in indicated in Figure 3. The action is the identity outside a disk of radius three around the origin.
THE SIZE OF SPANNING DISKS FOR POLYGONAL CURVES

Figure 2. The braid $\alpha$. The four arcs trace out the motion of $p_1, p_2, p_3, p_4$ under a continuous family of plane homeomorphisms beginning with the identity at $z = 1$ and ending with $\varphi$ at $z = 0$. All the arcs lie in the $xz$-plane except the one running from $p_1$ to $p_2$, which has two vertices off this plane.

Figure 3. The diffeomorphism $\varphi$ associated to $\alpha$. The diffeomorphism is the identity outside a disk containing $p_1, p_2, p_3, p_4$. The locations of the marked points remains fixed, $p'_2 = p_1, p'_3 = p_2$.

The curve $K_n$ is formed from an iterated braid

$$\beta_n = \alpha^{-n} \circ \alpha^n,$$

running between the planes $\{z = n\}, \{z = -n\}$. Between each pair of planes $\{z = k\}$ and $\{z = k + 1\}$, $K_n$ consists of a single copy of $\alpha$ for $0 \leq k \leq n - 1$ and a single copy of $\alpha^{-1}$ for $-n \leq k \leq -1$. In the braid group, $\beta_n$ is equivalent to the trivial 4-braid, which consists of four parallel vertical segments. The construction of $K_n$ is completed by appropriately connecting together the four strands at the upper and lower ends to form a closed curve, as shown in Figure 4. Above the plane $\{z = n\}$ we add a pair of line segments from $p_1$ at height $z = n + 2$ to each of $p_1$ and $p_2$ at height $z = n$, and from $p_3$ at height $z = n + 1$ to each of $p_3$ and $p_4$ at height $z = n$. Similarly, below the plane $\{z = -n\}$ we add a pair of line segments from $p_2$ at height $z = -n - 1$ to each of $p_2$ and $p_3$ at height $z = -n$, and from $p_2$ at height $z = -n - 2$ to each of $p_1$ and $p_4$ at height $z = -n$. Because the braids $\alpha^n$ and $\alpha^{-n}$ cancel in the braid group, it is clear that $K_n$ is unknotted.

Our main result is the following:

Theorem 1. For each $n \geq 0$,

1. $K_n$ is unknotted.
Figure 4. A closer look at the unknot $K_1$. All segments lie in the $xz$-plane except three segments from $p_1$ to $p_3$ between $z = 1$ and $z = 0$ and three segments from $p_3$ to $p_1$ between $z = 0$ and $z = -1$. These have $y$ coordinates above and below the $xz$-plane, as indicated by the + and − symbols.

2. $K_n$ contains at most $10n + 9$ edges.
3. Any piecewise-smooth embedded disk spanning $K_n$ intersects the $y$-axis in at least $2^{n-1}$ points.
4. Any embedded PL triangulated disk $D_n$ bounded by $K_n$ contains at least $2^{n-1}$ triangles.

The condition (3) that the disk intersects a line many times implies condition (4), that it contains many triangles, since each triangle can intersect a line transversely at most once.

We prove Theorem 1 in §3. We first construct a standard spanning disk for $K_n$, which we call $F_n$. Figure 5 shows $F_0$ and $F_1$. To understand the behavior of $F_n$ and other disks spanning $K_n$, we prove some facts about diffeomorphisms and “train tracks”. These are applied in §4 to count the intersections of the $y$-axis and $F_n$. In §5 we use Morse Theory to show that any other spanning disk is at least as complicated, along the $y$-axis, as the standard disk.

3. Construction of a standard spanning disk

In this section we describe how to construct for each $K_n$ a particular smooth spanning disk $F_n$. This standard disk intersects each plane $\{z = c\}$, $-n - 1 < c < n + 1$ in two arcs, which are embedded and disjoint. At $z = \pm n$ the arcs lie along the $x$-axis, joining $p_1, p_2$ and $p_3, p_4$ respectively. For $n \geq 2$ these arcs are shown in Figure 6 at heights $z = n, z = n - 1$ and $z = n - 2$. In Figure 6 the four arcs appear in the three pictures in order (1,2,3,4), (2,3,1,4) and (3,1,2,4), read left to right.
Above the plane \( \{z = n\} \) the standard disk consists of two triangles in the \( xz \)-plane, one with a base along the segment from \( p_1 \) to \( p_2 \) and one with a base along the segment from \( p_3 \) to \( p_4 \). Below \( z = -n \) it is bounded by a six-sided polygon in the \( xz \)-plane meeting \( \{z = n\} \) along two segments, one running from \( p_1 \) to \( p_2 \) and one from \( p_3 \) to \( p_4 \). Between \( \{z = n\} \) and \( \{z = -n\} \) the standard disk twists so that its boundary follows \( K_n \), as made precise below.

The arcs in the first disk of Figure 5 are taken by \( \varphi \) to the arcs in the second disk in the Figure, and those in turn are taken by \( \varphi \) to the arcs in the rightmost disk. The arcs of the braid indicate the motion of the disk in the process of sliding from \( z = 1 \) to \( z = 0 \). A composition of a counterclockwise half-twist interchanging the first two punctures, followed by a clockwise half-twist interchanging the second two punctures gives \( \varphi \). Corresponding to the braid \( \alpha^n \) is the iterate \( \varphi^n \) of \( \varphi \).

We now give a precise description of the construction of \( F_n \), based on \( \varphi \). Begin with a planar polygonal curve bounding a disk \( L_n \) in the \( xz \)-plane, formed as follows: Take vertical segments from \((-2,0, -n)\) to \((-2,0, n)\), \((-1,0, -n)\) to \((-1,0, n)\), \((1,0,-n)\) to \((1,0, n)\) and \((2,0, -n)\) to \((2,0, n)\). At the top, add a line segment from \((-2,0, n+2)\) to each of \((-2,0, n)\), \((-1,0, n)\) and from \((1,0, n+1)\) to each of \((1,0, n)\), \((2,0, n)\). At the lower end, add a line segment connecting \((-2,0, -n-2)\) to each of \((-2,0, -n)\), \((2,0, -n)\) and \((-1,0, -n-1)\) to each of \((-1,0, -n)\), \((1,0, -n)\), as shown in Figure 6.

The standard disk \( F_n \) is the image of this planar disk \( L_n \) under a diffeomorphism \( J_n : \mathbb{R}^3 \to \mathbb{R}^3 \), that preserves the \( z \)-coordinates of points, and carries the boundary of \( L_n \) to \( K_n \). The diffeomorphism \( \varphi \) is isotopic to the identity map on the plane. So there is a continuous family of diffeomorphisms of the plane \( \varphi_t \), \( 0 \leq t \leq 1 \),
A planar disk $L_n$, lying between \( \{ z = n + 2 \} \) and \( \{ z = -n - 2 \} \), is twisted to form $F_n$.

with $\varphi_1 = \text{identity}$ and $\varphi_0 = \varphi$. Define a diffeomorphism $G : \mathbb{R}^2 \times [0, 1] \to \mathbb{R}^2 \times [0, 1]$ by $G(x, y, t) = (\varphi_t(x, y), t)$. Then $G$ carries the vertical line segments in $\mathbb{R}^2 \times [0, 1]$ with $xy$-coordinates $(-2, 0), (-1, 0), (1, 0), (2, 0)$ to the braid $\alpha$. Also define a diffeomorphism $H : \mathbb{R}^2 \times [0, 1] \to \mathbb{R}^2 \times [0, 1]$, by $H(x, y, t) = (\varphi(x, y), t)$. This extends $\varphi$ to $\mathbb{R}^2 \times [0, 1]$ as a product. The diffeomorphism $J_n$ is defined to be the identity for $z \geq n$ and $z \leq -n$. For $0 \leq k \leq t \leq k + 1 \leq n$, $J_n(x, y, t) = G \circ H^{n-k-1}(x, y, t-k) + (0, 0, k)$, and for $-n \leq k-1 \leq t \leq k \leq 0$, $J_n(x, y, t) = G \circ H^{n+k-1}(x, y, k-t) + (0, 0, 2t-k)$.

A Morse function $f$ on a smooth closed manifold has a finite number of critical points $\{ c_i \}$ with distinct values under $f$. A Morse function on a manifold with boundary is a Morse function when restricted to both the boundary and the interior of the manifold. Critical points are either interior critical points or boundary critical points. A Morse function on a disk has at least two critical points, one maximum and one minimum, and if there are exactly two critical points then both must occur on the boundary, since there is a maximum and minimum value for the restriction to the boundary. The construction of $F_n$ gives $z$ as a Morse function on $F_n$ that has four critical points. Two are maxima, at $z = n + 2$ and $z = n + 1$, one is a minimum, at $z = -n - 2$, and one is a saddle point, at $z = -n - 1$. All four critical points lie on the boundary of $F_n$.

4. An invariant train track for $\varphi$

To understand the iterates of $\varphi$, we use an associated combinatorial object called an invariant train track. The theory of train tracks is described in [15]; we need here only elementary ideas from this theory. A train track is a 3-valent graph that is embedded on a surface. The edges, called tracks, are embedded smoothly and the three tangent directions at the vertices, called switches, lining up to give a $C^1$-embedding of the union of any pair of edges meeting at a vertex. Train tracks have fibered neighborhoods, closed neighborhoods filled by fibers. Fibers are intervals transverse to the edges, much like the tracks of a monorail, and there is a projection map of the fibered neighborhood to the train track. A curve is carried by a train track if it can be isotoped into the fibered neighborhood so that it is transverse to the fibers. Such a curve is roughly parallel to the tracks, but may run many times over each track. The curve is determined up to isotopy by a set of weights. These are non-negative integers assigned to each track, giving the number of times the curve runs over that edge, in either direction. At each switch there is a switching condition: the weight assigned to the one "incoming" track is the sum
of the weights of the two “outgoing” tracks. The weights for any two tracks near a vertex determine the weight for the third. An example is shown in Figure 8.

Figure 8. A train track $T$, showing its fibered neighborhood, and a curve that it carries with weights $a = 1$ and $b = 1$. Other weights are determined by the switching conditions.

A curve $C$ carried by a train track can be projected onto the train track, meaning that the embedding of the curve can be composed with the projection of each fiber in the fibered neighborhood to the base point of that fiber on the train track. Each track is given a weight by the projection, corresponding to the number of pre-images in $C$ of a point in the interior of the track. The curve $C$ can be reconstructed from these weights, by taking a number of copies of each track given by the weights and joining them together near the switches. There is a unique way to join that gives an embedded curve. The resulting simple closed curve is unique up to isotopy.

As with curves, a train track $T'$ is carried by another train track $T$ if $T'$ can be isotoped into a fibered neighborhood of $T$ so that its vertices are carried to vertices and so that the tracks of $T'$ are transverse to the fibers of the fibered neighborhood of $T$. We can then project $T'$ into $T$ by mapping each fiber to its base point on $T$. If $T'$ carries weights on its branches, then these can be summed to give weights on the branches of $T$ to which it projects, as in Figure 9.

A train track is said to be invariant under a diffeomorphism $\varphi$ of a surface if its image $\varphi(T)$ is carried by $T$.

For later application, we replace the level planes $\{z = c\}$ of the height function $z$ with the level sets of a different function $f_n : \mathbb{R}^3 \to \mathbb{R}$, that agrees with $z$ in a large ball around the origin, a ball that contains the disks we will be considering. Thus in subsequent arguments we will be able to view either $f_n$ or $z$ interchangeably as the Morse function we are using. The level sets of $f_n$ are a family of spheres rather than planes $\{z = c\}$. To construct $f_n$, we first choose a large constant $R_n > 0$ such that a ball of radius $R_n$ centered at the origin contains $F_n$ in its interior. For each $t$ with $-R_n < t < R_n$, define $\Sigma_t$ to be the 2-sphere obtained by taking the disk $\{(x, y, z) : x^2 + y^2 \leq R_n, \ z = t\}$ and capping it to form a convex 2-sphere enclosing the point $(0, 0, -2R_n)$. Figure 10 shows some of these spheres. The spheres are the level sets of a function $f_n : \mathbb{R}^3 \to [-2R_n, \infty)$.

The restriction of $f_n$ to the disks we will consider agrees with $z$, and $R_n$ will be chosen large enough so that the level sets of $f_n$ look identical to flat planes in a ball containing the disks. Note that we can use different functions $f_n$ for different values of $n$, if necessary, to ensure that our choice of $R_n$ is sufficiently large. The diffeomorphism of the 4-punctured plane $\varphi$, which was the identity outside of a
Figure 9. A train track $T'$ is projected into another train track $T'$ that carries it. The weights of the tracks on $T'$ that project to a given track on $T$ are summed to give weights on that track.

Figure 10. Level sets of the function $f_n$ are spheres, but in a ball containing $K_n$ the level sets are the same as those of the $z$ coordinate.

There is an invariant train track $T$ for $\varphi$, depicted in Figure 1, and also shown with a fibered neighborhood in Figure 8. An assignment of weights to all the tracks of $T$ is completely determined by assigning two weights $a$ and $b$ on the two indicated tracks, as in Figure 8. The non-negative integers $a$ and $b$ are arbitrary, but all other weights are determined by the switching conditions. Each choice of $a, b$ gives rise to a unique simple closed curve carried by $T$, and we refer to $a, b$ as the weights with which this curve is carried by $T$. 
Figure 11. An invariant train track $T$ for $\varphi$. Here $\sigma_1$ and $\sigma_2^{-1}$ denote diffeomorphisms of the four punctured disks associated to the braid group elements $\sigma_1$ and $\sigma_2^{-1}$. The composition of these two diffeomorphisms gives $\varphi$, showing that $T$ is invariant under $\varphi$.

To understand the iterates of $\varphi$ we study the image of $T$ under $\varphi$. The image $\varphi(T)$ can be isotoped so that vertices of $\varphi(T)$ go to vertices of $T$ and tracks of $\varphi(T)$ are transverse to the fibers of the fibered neighborhood of $T$, as indicated in Figure 11.

**Lemma 2.** The train track $T$ is invariant under the homeomorphism $\varphi$. A curve carried by $T$ with weights $a, b$ is mapped by $\varphi$ to a curve carried by $T$ with weights $a + b$ and $a + 2b$.

**Proof:** The image of $T$ under $\varphi$ can be isotoped into the fibered neighborhood of $T$ as shown in Figure 11. The tracks with initial weights $a$ and $b$ have projected onto them tracks with total weight $a + b$ and $a + 2b$ respectively. A curve carried by $T$ with weights $a, b$ is similarly carried to a curve carried with weights $a + b, a + 2b$. $\square$
So $T$ is an invariant train track for $\varphi$, and a curve $C$ carried by $T$ with weights $a$ and $b$, has image $\varphi(C)$ which is also carried by $T$, but with weights $a + b$ and $a + 2b$. When $\varphi$ is iterated, the weights on these two tracks grow according to a Fibonacci sequence:

$$\{(a, b), (a + b, a + 2b), (2a + 3b, 3a + 5b), (5a + 8b, 8a + 11b), \ldots\}.$$ 

**Lemma 3.** A curve carried by $T$ with weights $a_0 \geq 0$ and $b_0 \geq a_0$ is mapped by the diffeomorphism $\varphi^n$ to a curve carried by $T$ with weights $a_n$ and $b_n$, satisfying $a_n \geq 2^n a_0$ and $b_n \geq 2^n b_0$.

**Proof:** Under the action of $\varphi$ the weight $a$ corresponding to a curve $C$ is transformed to the weight $a + b \geq 2a$ corresponding to $\varphi(C)$ and the weight $b$ to $a + 2b \geq 2b$. The result follows by iterating $n$ times.

Let $B$ denote the simple closed curve on a 4-punctured sphere $S$ that separates the points $p_1, p_2$ from $p_3, p_4$, as shown in Figure 12. We analyze the number of intersections between $B$ and a curve $C$ in the 4-punctured sphere $S$ that is carried by $T$ with weights $a, b$. We show there is no isotopy of $C$ in the 4-punctured sphere which can reduce the number of intersections with $B$ below $2a + 2b$.

**Lemma 4.** A curve $C$ in $S$ that is carried by the train track $T$ with weights $a$ and $b$ intersects $B$ in at least $2a + 2b$ points.

**Proof:** In a surface containing two intersecting simple closed curves, a 2-gon is a disk on the surface whose boundary consists of an arc from each of the curves and whose interior is disjoint from each of them. It is shown in [11, Lemma 3.1, pp. 108] that if two simple closed curves on a surface have more intersections than the minimal possible number in their isotopy class, then each contains an arc such that the two arcs together bound a 2-gon on the surface.

Let $C$ be a curve lying in the fibered neighborhood of $T$, transverse to the fibers and carried with weights $a$ and $b$. It follows from the above that if $C$ can be isotoped in $S$ to have fewer than $2a + 2b$ points of intersection with $B$, then there exists an arc $\beta$ contained in $B$ and an arc $\gamma$ contained in $C$ that together bound a 2-gon in $S$, whose interior is disjoint from $B \cup C$. We will show that there is no such 2-gon between $C$ and $B$, and hence that $C$ can not be isotoped to reduce the number of its intersections with $B$.

The arc $\gamma$ lies on $C$ and so lies in the fibered neighborhood of $T$, and is transverse to the fibers. Moreover $\gamma$ intersects $B$ only at its two endpoints, and therefore lies

![Figure 12. The curve $B$ intersects a curve carried by the train track $T$ in at least $2a + 2b$ points.](image)
either to the right or to the left of \( B \) on \( S \), where “left” refers to the side containing \( p_1, p_2 \) and “right” to the side containing \( p_3, p_4 \).

An arc carried by \( T \) with interior to the right of \( B \) runs once around the third puncture and, together with \( \beta \), must separate the third and fourth punctures. Similarly an arc carried by \( T \) with interior on the left of \( B \) runs once around either the first or second punctures before returning to \( B \), and together with \( \beta \) separates the first and second punctures. In either case such an arc is not homotopic to an arc in \( B \), (rel boundary), and therefore cannot cobound a disk with an arc \( \beta \) contained in \( B \). So \( \beta \cup \gamma \) cannot cobound a 2-gon, and it follows that the number of intersections of \( B \) and \( C \) cannot be reduced. \( \square \)

**Corollary 5.** Let \( p_1, p_2, p_3, p_4 \) denote four distinct marked points on a 2-sphere and let \( B \) denote a simple closed curve separating \( p_1, p_2 \) from \( p_3, p_4 \). Let \( \delta \) be the simple closed curve that is the boundary of a neighborhood of an arc joining \( p_1 \) to \( p_2 \) in the complement of \( B \). Then \( \varphi^n(\delta) \) intersects \( B \) in at least \( 2^n \) points.

**Proof:** While \( \delta \) is not carried by \( T \), its image \( \varphi(\delta) \) is carried by \( T \) with weights \( a = 0, b = 1 \), and \( \varphi^2(\delta) \) is carried with weights \( a = 1, b = 2 \). Lemma \( \square \) can be applied to \( \varphi^2(\delta) \) and its iterates, so \( \varphi^n(\delta) = \varphi^{n-2}\varphi^2(\delta) \) is carried with weights at least \( a = 2^{n-2}, b = 2^{n-1} \). By Lemma \( \square \) the curve \( B \) intersects a curve carried by the train track with weights \( a, b \) in at least \( 2a + 2b \) points. Since \( 2a + 2b \geq 2, b \geq 2^n \), the result follows. \( \square \)

**Figure 13.** The curves \( \gamma, \delta \) and \( \varphi(\delta) \) on a sphere with four marked points removed.

**Corollary 6.** An arc \( \gamma \) joining \( p_1 \) to \( p_2 \) in the complement of \( B \) has image under \( \varphi^n \) that intersects the closed curve \( B \) on the 4-punctured 2-sphere \( S \) in at least \( 2^{n-1} \) points.

**Proof:** The simple closed curve \( \delta \) is isotopic to the boundary of a regular neighborhood of \( \gamma \) and \( \varphi^n(\delta) \) is isotopic to the boundary of a regular neighborhood of \( \varphi^n(\gamma) \). For any arc in \( S \) that intersects \( B \) transversely, the boundary of a sufficiently thin neighborhood of the arc intersects \( B \) in twice the number of points that the arc intersects \( B \). If \( \varphi^n(\gamma) \) intersected \( B \) in fewer than \( 2^{n-1} \) points then we can form a thin neighborhood of \( \varphi^n(\gamma) \) whose boundary intersects \( B \) in less than \( 2^n \) points, contradicting Corollary \( \square \). \( \square \)
5. Combinatorial Complexity of Spanning Disks for $K_n$

In this section we show that any PL spanning disk for $K_n$ contains exponentially many triangles, proving the main result.

**Proof of Theorem 1.** Let $n$ be any fixed positive integer. The assertion (1) that $K_n$ is unknotted follows from its construction as the composition of a braid and its inverse.

The curve $K_n$ can be constructed with straight segments as follows: Four segments above $\{z = n\}$ and four below $\{z = -n\}$ cap off the braid. Between $\{z = n\}$ and $\{z = -n\}$, a single line segment forms the entire fourth strand, and the first three strands are formed from $2n$ copies of the first three strands of $\alpha$. Each copy of $\alpha$ requires five segments for the first three strands. The total number of segments needed is no more than $10n + 9$, which is assertion (2).

We prove assertion (3) in three steps. Recall that for each fixed $n$, $K_n$ bounds a smooth disk $F_n$ in $\mathbb{R}^3$ that we call the standard disk, and that $B_0 = \Sigma_0 \cap \{x = 0\}$ is the closed curve obtained by intersecting $\Sigma_0$ with the plane $\{x = 0\}$. We first show that $B_0$ intersects $F_n$ in at least $2^{n-1}$ points. We then consider an arbitrary smooth spanning disk $E_n$, and show that the number of intersections of $E_n$ with $B_0$ is at least as large as that of $F_n$ with $B_0$. In the third step, we approximate an arbitrary PL disk $D_n$ by a smooth disk to obtain the same conclusion in the PL setting.

The standard disk $F_n$ is swept out by arcs joining points of $K_n$ in the level sets $\{z = \text{constant}\}$, as they descend from $\{z > n + 2\}$ to $\{z < n - 2\}$. One arc appears below $\{z = n + 2\}$, the second below $\{z = n + 1\}$. The arcs join together to form a single arc at $\{z = -n - 1\}$, and this in turn disappears below $\{z = n - 2\}$. The height function given by the restriction of the $z$-coordinate to $F_n$ defines a Morse function on $F_n$, and this Morse function has no critical points in the interior of $F_n$.

The arc $\gamma = \gamma_n$ in $F_n \cap \{z = n\}$ connects $p_1$ and $p_2$, as shown in Figure 13. Denote by $\gamma_t$ the arc in $F_n \cap \{z = t\}$ that is in the same component of $F_n \cap \{z \geq t\}$ as $\gamma$. For each integer $k$ with $0 < k \leq n$, as we slide $\gamma_k$ down one unit along $F_n$, $\gamma_k$ is deformed along $F_n$ to $\gamma_{k-1} = \varphi(\gamma_k)$. So as $t$ decreases to 0, the arc $\gamma_n$ is slid along $F_n$ to an arc $\gamma_0$ that is the image of $n$ iterations of $\varphi$.

Let $B_0$ denote the closed curve along which the level set $\Sigma_0 = f_n^{-1}(0)$ intersects the $yz$-plane. Then $B_0$ separates the four points of intersection of $\Sigma_0$ and $K_n$ into pairs, $p_1, p_2$ and $p_3, p_4$. The standard disk $F_n$ intersects $B_0$ in at least $2^{n-1}$ points by Corollary 3.

Our goal is to show that an arbitrary PL disk bounded by $K_n$ intersects $B_0$ in at least as many points as does $F_n$. Before considering PL disks, we first consider a smooth spanning disk $E_n$. In this setting we will apply some basic results from the Morse Theory of smooth functions on surfaces; see [12] for an exposition of smooth Morse Theory. We will then shift back to the PL setting. The height function $z$, or the function $f_n$ that agrees with it on the region we are studying, will serve as the Morse functions. Let $E_n$ denote an arbitrary disk spanning $K_n$ such that

1. $E_n$ has smoothly embedded interior.
2. The height function $z$ restricted to $E_n$ is a Morse function.

We now show, using Morse theory, that the surface $E_n$ intersects the closed curve $B_0$ in at least as many points as does the “standard disk” $F_n$. Choose a value of $R_n$ large enough so that $F_n$ and $E_n$ both lie in the interior of the ball of radius
$R_n$, and as before form the Morse function $f_n$ whose level sets are spheres $\Sigma_t$ for $t > -2R_n$. The intersection of $E_n$ with the spheres $\Sigma_t$ at non-critical levels is contained in $\Sigma_t \cap \{ z = t \}$. As $t$ decreases from $\infty$ to $-2R_n$, the sphere $\Sigma_t$ begins to intersect $K_n = \partial E_n$, when $t = n + 2$. As $t$ decreases there are first one, then two arcs in $\Sigma_t \cap E_n$, along with a (possibly empty) collection of simple closed curves. For $n + 1 < t < n + 2$, $\Sigma_t \cap E_n$ consists of a single arc $\beta_t$, along with a possibly empty collection of simple closed curves. For $-n < t < n$, $\Sigma_t \cap E_n$ contains two arcs connecting the four points of $\Sigma_t \cap K_n$. As $t$ decreases from $n + 1$, $\beta_t$ is continuously deformed as long as $E_n$ is transverse to $\Sigma_t$. As long as the transversality continues to hold, let $\beta_t$ denote the arc that is in the same component of $E_n \cap \{ z \geq t \}$ as $\beta$.

As long as passing through the critical level $\{ t = \pm \}$ does not change which pairs of points on $K_n$ are connected by the pairs of arcs, it is possible to extend the definition of $\beta_t$ to one of the two arcs below the critical point. We define $\beta_{c-\epsilon}$ to be the arc connecting the same pair of points as $\beta_{c+\epsilon}$. In these cases even the isotopy class of the arc is preserved, though the curve $\beta_t$ does not change continuously when $t$ passes through the critical level $c$.

For the values $-n < t < n$, the level sets of $f_n$ are transverse to $K_n$, so any critical points lie in the interior of the disk $E_n$. There are three types of changes in $\beta_t$ that can occur when descending from $\Sigma_{c+\epsilon}$ to $\Sigma_{c-\epsilon}$, as indicated in Figure 14.

1. Moving past a saddle critical point connects $\beta_{c+\epsilon}$ to a simple closed curve of $\Sigma_{c+\epsilon} \cap E_n$ to form $\beta_{c-\epsilon}$.
2. Moving past a saddle critical point connects $\beta_{c+\epsilon}$ to itself to form $\beta_{c-\epsilon}$ together with a simple closed curve.
3. Moving past a saddle critical point connects $\beta_{c+\epsilon}$ to the second arc of $\Sigma_{c+\epsilon} \cap E_n$. No arc $\beta_{c-\epsilon}$ is defined.

The first two types of moves are inverses, since reversing the direction of a type (1) move gives a type (2) move and vice-versa. The level surface in which $\beta_{c+\epsilon}$ lies is either a 2-punctured sphere, if $c \geq n + 1$ or $c \leq -n - 1$, and is a 4-punctured sphere otherwise. In a 2-punctured sphere there is a unique isotopy class of arcs connecting the two punctures, so the isotopy class of $\beta_t$ in $\Sigma_t$ is unchanged when passing through the critical point. In a 4-punctured sphere there are many isotopy classes of arcs connecting two of the punctures, but as we pass a saddle critical point of type (1) or type (2), the curve $\beta_t$ remains in the complement of the second arc of intersection. The complement of an arc in a sphere is homeomorphic to an (open) disk, and in a disk there is a unique isotopy class of arcs connecting any two points. So the isotopy class of $\beta_t$ in $\Sigma_t$ remains unchanged by a saddle move in these cases.

The third type of critical point does change the isotopy class of the arc, since it changes the boundary points connected by the arc. The following lemma asserts that this can occur at most once.

**Lemma 7.** Suppose that $f_n : D \to \mathbb{R}$ is a Morse function on a topological disk $D$ that restricts to a Morse function on $\partial D$. Suppose also that $f_n|_{\partial D}$, the restriction of $f_n$ to $\partial D$, has at most four critical points on $\partial D$. Then $f_n$ can have at most one interior critical point of type (3) that is a saddle connecting distinct arcs in the level set of $f_n$.

**Proof:** Suppose there is an interior critical point with critical value $c$ that is a saddle connecting distinct arcs in the level set $f_n = c + \epsilon$. The four arcs leaving the saddle point hit the boundary of $D$ at four distinct points. These arcs divide
Figure 14. Three types of changes can occur in the level sets $E_n \cap \{z = t\}$ as they pass through a critical point.

$D$ into four quadrants, which meet $\partial D$ in four arcs. Each of the boundary arcs has its two endpoints on the level set $f_n = c$, and is non-constant on these boundary arcs. So each contains at least one maximum or minimum of $f_n|_{\partial D}$. Suppose there were a second saddle critical point on a level set $\{f_n = c'\}$. Since critical points of Morse functions have distinct values, we have that $c' \neq c$. The four arcs emerging from the second saddle are therefore disjoint from the first critical level set, and contained in one of the previously defined quadrants. See Figure 15.

The intersection of this quadrant with $\partial D$ has four points on which $f_n$ takes the value $c'$, and therefore $f_n|_{\partial D}$ has at least three critical points in this quadrant. It follows that $f_n|_{\partial D}$ has at least six critical points, contradicting the hypothesis. So only one saddle of type (3) can occur.

Since $E_n$ is a topological disk whose boundary has four critical points for the height function, at most one type (3) critical point can occur. Assume first that a type (3) critical point does not occur for $t > 0$. On the standard disk, $\gamma_0$ is obtained from $\gamma_n$ by a continuous deformation involving no critical points, while $\beta_0$ is obtained from $\beta_n$ by a process that may include passing through critical points of types (1) and (2), but none of type (3). Therefore the isotopy class of $\beta_0$ in the 4-punctured sphere is the same as that of $\beta_{n+2-\epsilon}$. Since $\beta_t$ is isotopic to $\gamma_t$ for $t$ close to $n + 2$, we conclude that $\beta_0$ is isotopic to $\gamma_0$. By Lemma 4, $\beta_0$ intersects $B_0$ in at least as many points as $\gamma_0$.

Now consider the case where a type (3) critical point does occur for some $t > 0$. By Lemma 5, there are no type (3) critical points for $t < 0$. In this case we repeat the previous argument, but using the function $-f_n$ rather than $f_n$. Note that as $z$ increases from $\{z = k\}$ to $\{z = k + 1\}$ for $k$ an integer with $-n \leq k \leq -1$, the level sets of $F_n$ are again transformed by an application of $\varphi$. We replace $\gamma$ with the
Figure 15. The existence of two type (3) saddle critical points on a disk, at levels \( f_n^{-1}(c) \) and \( f_n^{-1}(c') \), imply that the boundary of the disk has at least six critical points.

\[ \text{arc } \gamma' \text{ in } E_n \cap \Sigma_{-n} \text{ which joins } p_2 \text{ to } p_3, \text{ and } \delta \text{ with } \delta', \text{ the boundary of a regular neighborhood of } \gamma'_n. \]

Figure 16. The curves \( \gamma' \), \( \delta' \) and \( \varphi(\delta') \).

The curve \( \delta' \) is carried with weights \( a = 1, b = 0 \) by \( T \) and \( \varphi(\delta') \) is carried with weights \( a = 1, b = 1 \). By Lemma 6, \( \varphi^n(\delta') \) is carried by \( T \) with weights \( a \geq 2^{n-1} \) and \( b \geq 2^{n-1} \). By Lemma 7 a curve isotopic to \( \varphi^n(\delta') \) intersects \( B_0 \) in at least \( 2a + 2b = 2^n + 1 \) points. Then \( \varphi^n(\gamma') \) intersects \( B_0 \) in at least half as many, or \( 2^n \) points. As \( t \) increases from \(-n\) to 0, the arc \( \gamma'_n \) in the component of \( E_n \cap \{ z \leq t \} \) containing \( \gamma' \), is carried to an arc in \( \Sigma_0 \) isotopic to \( \varphi^n(\gamma'_n) \). So \( E_n \) again must intersect \( B_0 \) in at least \( 2^n \) points. In every case the curve \( B_0 \) intersects \( E_n \) in at least \( 2^n \) points.

Now consider an arbitrary PL disk \( D_n \) with boundary \( K_n \). After an arbitrarily small isometry of \( \mathbb{R}^3 \), we can arrange that \( D_n \) intersects the \( y \)-axis transversely in a finite number of points. The disk \( D_n \) can be approximated by a disk \( E_n \), with smoothly embedded interior, that coincides with \( D_n \) in a neighborhood of each intersection point with the \( y \)-axis, and that remains disjoint from other points of the \( y \)-axis. We choose \( R_n \) larger, if necessary, so that in the passage from the Morse function \( z \) to the Morse function \( f_n \), the three surfaces \( D_n, E_n \) and \( F_n \) that we consider only intersect the flat parts of the spheres \( S_t \) which form the level sets of \( f_n \). So the intersection of \( B_0 \) with \( D_n, E_n \) and \( F_n \) is the same as that of the
y-axis with these surfaces, each intersection being in the interior of a triangular face of $D_n$.

But we have shown that $E_n$ intersects $B_0$ in at least $2^n - 1$ points, and it therefore follows that the PL disk $D_n$ also intersects $B_0$ in at least $2^n - 1$ points.

Since a triangle transversely intersects a line in at most one point, and $B_0$ agrees with the $y$-axis in a ball containing $D_n$, $E_n$ and $F_n$, this implies that $D_n$ contains at least $2^n - 1$ triangles, and Theorem 1 is proved. \(\square\)

Remarks:

1. If we allow spanning disks that self-intersect, then the number of triangles required to span $K_n$ grows only linearly with $n$. If a spanning surface of arbitrary genus is allowed, it can be shown that the number of triangles required to span $K_n$ grows at most quadratically in $n$ [7].

2. The number of Reidemeister moves required to transform any unknotted curve constructed with $n$ polygonal edges into a single triangle has an exponential upper bound derived in [6]. For the particular $K_n$ constructed here, the number of Reidemeister moves required to transform the projection of $K_n$ to a projection with no crossings grows only linearly with $n$.

3. The argument establishes somewhat better estimates than claimed above. If we embed $K_n$ into $\mathbb{R}^3$ in a more efficient way, then for large $n$, $K_n$ has at most $6n$ segments and the number of triangles contained in any disk spanning $K_n$ grows faster than a constant times $\phi^{2n}$, where $\phi$ is the golden ratio.

6. Acknowledgments

This paper was completed while the first author was visiting the Institute for Advanced Study. The authors are grateful to J. Lagarias and the referee for helpful suggestions on the exposition.

References

[1] D. Avis and H. El-Gindy, Triangulating point sets in space, Discrete Comput. Geom. 2 (1987), no. 2, 99–111.
[2] J. Birman, Braids, Knots and the Mapping Class Group, Princeton University Lecture Notes.
[3] J. Birman and M.D. Hirsch, Recognizing the unknot, Geom. Topol. 2 (1998), 175–220 (electronic).
[4] S. Galatolo, On a problem in effective knot theory, Atti Accad. Naz. Lincei Cl. Sci. Fis. Mat. Natur. Rend. Lincei (9) Mat. Appl. 9 (1998), no. 4, 299–306 (1999).
[5] W. Haken, Theorie der Normalflachen, ein Isotopie Kriterium für ein Kreis, Acta Math., 105 (1961), 245–375.
[6] J. Hass and J. C. Lagarias, The number of Reidemeister moves needed for unknotting, J. Amer. Math. Soc. 14 (2001), 399-428.
[7] J. Hass and J. C. Lagarias Combinatorial isoperimetric inequalities in $\mathbb{R}^d$, In preparation.
[8] J. Hass, J. C. Lagarias and N. Pippenger, The computational complexity of knot and link problems, preliminary report, Proc. 38th Annual Symposium on Foundations of Computer Science, (1997) 172-181.
[9] J. Hass, J. C. Lagarias and N. Pippenger, The computational complexity of knot and link problems, J. ACM 46 (1999), no. 2, 185–211.
[10] J. Hass, J. C. Lagarias and W.P. Thurston, Area Inequalities for Embedded Disks Bounding Unknotted Curves, in preparation.
[11] J. Hass and G.P. Scott, Intersections of curves on surfaces, Israel J. Math. 51 (1985), no. 1-2, 90–120.
[12] J. Milnor, Morse theory, Annals of Mathematics Studies, No. 51 Princeton University Press, Princeton, N.J. 1963.
[13] R.C. Penner and J.L. Harer, *Combinatorics of train tracks*, Annals of Mathematics Studies, 125. Princeton University Press, Princeton, NJ, 1992.

[14] J. Snoeyink, A trivial knot whose spanning disks have exponential size, Proc. 6th ACM Conf. Computational Geometry 1990, 139–147.

[15] W. P. Thurston, *Three-dimensional geometry and topology* Princeton University Lecture Notes, 1978.

[16] F. Waldhausen, The word problem in fundamental groups of sufficiently large irreducible 3-manifolds, Annals of Math. (2) 88 1968 272–280.

[17] D. J. A. Welsh, *Complexity: Knots, Colourings and Counting*, Cambridge University Press: Cambridge 1993.

[18] D. J. A. Welsh, Knots and braids: some algorithmic questions, in: *Graph Structure Theory* (Seattle, WA 1991), Contemporary Math. Vol. 147, AMS: Providence 1993, pp. 109–123.

Department of Mathematics, University of California, Davis California 95616
E-mail address: hass@math.ucdavis.edu

Department of Computer Science, University of North Carolina, Chapel Hill, NC 27514
E-mail address: snoeyink@cs.unc.edu

Department of Mathematics, University of California, Davis California 95616
E-mail address: wpt@math.ucdavis.edu