

A Dido Problem as modernized by Fejes Tóth

Alan Siegel¹

COURANT INSTITUTE OF MATHEMATICAL SCIENCES
NEW YORK UNIVERSITY
New York

Abstract

Let a finite number of line segments be located in the plane. Let C be a circle that surrounds the segments. Define the region enclosed by these segments to be those points that cannot be connected to C by a continuous curve, unless the curve intersects some segment. We show that the area of the enclosed region is maximal precisely when the arrangement of segments defines a simple polygon that satisfies a fundamental isoperimetric inequality, and thereby answer the most basic of the modern day Dido-type questions posed by Fejes Tóth.

1 Introduction

Area optimization problems have a long and remarkably enduring history. Their study can be traced as far back as Zenodorus (second century, B.C.), and pragmatic applications can even be found in the Aeneid [16], which recounts the legend of Queen Dido, and her founding of Carthage. According to Virgil, Queen Dido purchased the right to as much land as she could demark with the skin of an ox, and exercised the option by splitting the hide into extremely thin strips of leather and tying them together. She then used this string to construct a giant semicircle that, when combined with the natural boundary imposed by the sea, turned out to encompass far more area than the seller could have ever imagined.

An elementary presentation of contemporary area optimization problems and their mathematical history can be found in Kazarinoff's very accessible text [8]. In brief, the ancient Greeks knew the Isoperimetric Inequality, which states that among all plane figures with a given perimeter, the circle encompasses the greatest area. Similarly, they understood that among all n -gons with a given perimeter, the regular n -gon encloses the greatest area. Pappus discusses these results [13], and attributes them to Zenodorus, whose original writings on the subject have been lost to antiquity. In the mid-nineteenth century, Steiner realized that the Greeks had actually failed to prove these bounds. Pappus, it seems, had just assumed that the very reasonable pictures he drew captured all of the cases, and that the reshaping transformations he specified could always be applied, somehow, to take any polygonal figure into a better one where the result could be established more directly. The difficulty is that while he was right about the reshaping (as is readily understood with 2000+ years of hindsight), he did not explain how to sequence Zenodorus's ideas in a provably correct manner.

In 1841, Steiner endeavored to give a rigorous proof of the Isoperimetric Inequality by publishing the first of five proposed proofs, which have all come to be recognized as insightful but incomplete. The first rigorous proofs, which used concepts in analysis and the calculus of variations, were given by F. Edler and by Weierstrass. As of the mid 1960's, the question of finding a simple geometric proof was, according to the literature, widely believed to be open (cf. [8, 4, 14]). However, there have been some not so simple proofs of Eastern European origin that lie within elementary geometry. See, for example, [17]. Among the calculus proofs, perhaps the shortest and most elegant is due to Peter Lax [9]. A very elegant proof that uses continuous (but elementary) deformations of convex regions is presented in [1]. A survey of more sophisticated isoperimetric inequalities including formulations for higher dimensions can be found in [10].

¹This research was supported in part by NSF grant CCR-9503793

2 A Dido-type problem

The modern formulations of Dido-type problems date from the mid 1960's, when Fejes Tóth wondered, among other things, what Queen Dido would have done if her strips of hide had hardened into rigid straight line segments. To be more precise, G. Hajós framed this specific version, and questions of this nature have become known as Dido-type problems (cf. [6, 11, 5]). It is to this most basic version as formulated by Hajós that we now turn our attention.

Let E be a collection of finite line segments located in R^2 . A formal definition of the area enclosed by E is as follows. Let $\mathcal{U}$ be the union of the bounded components of $R^2 \setminus E$. Let $\mathcal{W}$ be the closure of the point set defined by $\mathcal{U}$.

- Define the area of $\mathcal{W}$ to be the area enclosed by E , and let this value be denoted by $\text{Area}(E)$.

Informally, the problem posed by Fejes Tóth and Hajós is to prove that the area enclosed by an arbitrary configuration of segments E cannot exceed the largest possible area enclosed by a polygon whose segments are no longer than their counterparts in E . Now, it is fair to suggest that Fejes Tóth and Hajós understood that the statement had to be true, since it is a matter of common sense. They were really asking how mathematics can deal with the lack of organization posed by a collection of segments that can be arranged in arbitrary ways. What do we do if the segments cannot be directed to define a closed oriented path? How do we prove that such chaotic collections can be rearranged as polygons with greater area? At issue are the very many irregular cases that are obviously poor choices for maximizing the area, and which would have been dismissed without comment by the ancient geometers some 2000 years ago.

In 1973, Fejes Tóth published an incorrect solution to a restricted version of the problem [7]. Upon noticing the error, he communicated the problem to János Pach, who gave a correct solution for the restricted case [11]. Although the basic problem was posed over 30 years ago, no solutions have ever appeared in the literature, and it has been listed among the open questions in geometry [5, p30-32]. In 1990, however, Gábor Kertész discovered a solution, and it was accepted as his Masters Thesis at Eötvös University. According to knowledgeable sources, the proof is very complex and has yet to appear in journal form [12].

We give a proof that is simple, fairly direct and, we believe, insightful. Part of the proof's simplicity comes from the use of divide and conquer to reduce the complexity imposed by a morass of arbitrarily placed segments. The proof also uses such intuitive notions as the direct mapping of area, which is reminiscent of ancient geometric algebra. In addition, we will formulate area coverage problems in terms of sufficiency conditions about sums of angles, which is, perhaps, more representative of some developments in modern geometry. Lastly, we exploit the realization that that certain generic exceptions are remarkably innocuous. Yet, for reasons that are not evident, all of these notions seem to need fairly strong quantification for the problem at hand.

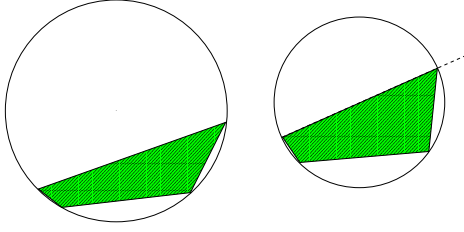
We will soon specify this so-called Dido-type problem precisely, but defer doing so in order to present some preliminary definitions and background geometry that will simplify the subsequent exposition.

- Let the term *multiset* denote collections of elements where different members can be equal.

Suppose S is a multiset of n nonnegative numbers. It is well known how to define the polygons that have side lengths comprising the elements in S , and which have an area that is as large as possible. The solution is to create a polygon with the correct edge lengths that is inscribed in a circle of suitable radius.

To see when this construction makes sense, consider the following procedure. Create a path of segments with lengths defined by the elements in S . Now remove, from the path, some segment that has the longest length, and let this segment length be s . Let the new path of $n - 1$ segments have its vertices placed along a circle of huge radius, and consider how the path is forced to curl up as the radius is decreased. Evidently, the distance between the endpoints begins as the sum of the $n - 1$ segment lengths when the circle radius is infinite, and decreases continuously as the radius diminishes. If s equals this initial sum, the area of the resulting polygon is zero, and the circle is a straight line with infinite radius. If s is less, we can reduce the radius until the distance between the endpoints equals s , and then close the path by connecting its endpoints with the removed segment. This construction gives a polygon that has the maximum area among all n -gons with side lengths comprising the elements of S . For completeness, we note that the construction would fail if we attempt to diminish the diameter of the circle to a value that is smaller than some edge in the chain of

$n - 1$ segments, since such a segment could not lie on the circle. But since all such lengths are bounded by s , no such reduction will occur in the construction.



As is well known, there is a general case where the area can be increased by reducing s . Suppose that the previous procedure creates a polygon that is inscribed within a semicircle of the bounding circumcircle. In this case, it is better to continue reducing the radius and stop the reduction exactly when the path fits tightly within a semicircle, and close the path with a diameter of this final circle as shown in the accompanying figure. This procedure should also be applied if

s exceeds the sum of the other values in S . In these cases, the Reflection Principle shows that this modified construction gives the largest area among all polygons with side lengths that are bounded by the elements in S .

In particular, let $\hat{S}$ comprise the multiset of lengths with s removed. Consider the problem of constructing the n -gon with the greatest area where $n - 1$ of its sides have lengths that comprise the elements of $\hat{S}$, and the last side is built from a subset of an infinite line ℓ . The solution is to configure the edges as a path D that is tightly inscribed in a minimum radius semicircle. Indeed, if some other arrangement R gave a greater area, then we can reflect it across ℓ and adjoin the two copies. The result would be a polygon with side lengths comprising the elements of the multiset $\hat{S} \cup \hat{S}$. But one of the standard isoperimetric inequalities for polygons states that among all polygons with a prescribed multiset of side lengths, those with the greatest area are precisely the ones that can be inscribed in a circle. In this case, D as adjoined with its reflected duplicate is such a maximum area polygon. Hence the supposition that R yields a greater area than D contradicts this isoperimetric inequality.

- Let the term *Dido exception* denote the instances of this Dido-type problem where the longest edge should be replaced by a shorter segment that acts as a diameter of the resulting circumcircle.

Of course, if the original construction produces a polygon that is not inscribed within some semicircle, then it has the greatest possible area, and there is no Dido exception. These area maximization results are all standard; see, for example, any of [8, 11, 14].

- Given a multiset S , let $Poly(S)$ be a simple polygon that has side lengths bounded from above by the numbers in S , and has the maximum possible area.

- Let $C(S)$ be the circumcircle of $Poly(S)$, o_S its center, and $Radius(S)$ its radius.
- If $S = \{s_1, s_2, \dots, s_n\}$, let $e(s_i)$ refer to the edge of $Poly(S)$ that corresponds to s_i .
- Let $\Delta(s_i; S)$ be the triangle with base edge $e(s_i)$ and opposing vertex o_S .

Of course, we can permute these triangles any way we like to get all of the equivalent polygons that have the maximum area.

- Let $alt(s_i; S)$ be the altitude from the base $e(s_i)$ of $\Delta(s_i; S)$ to o_S .
- Let $DArea(S)$ be the area of $Poly(S)$, which includes the reduction of the longest edge whenever necessary to get a polygon of greater area.

Given these definitions, we can now formalize the problem posed by Fejes Tóth and Hajós.

Theorem 3. Let $\hat{S}$ be a collection of edges placed in the plane R^2 . Let S be the multiset whose elements are the lengths of the edges in $\hat{S}$. Then

$$Area(\hat{S}) \leq DArea(S).$$

The proof relies on turning the problem about arbitrary edge collections into a well structured and specific inequality where calculus can be applied with ease. Accordingly, we first seek to quantify $DArea(S)$ as a function of the defining edge lengths in S .

Lemma 1. Let $S = \{s_1, s_2, \dots, s_n\}$ be a multiset of nonnegative values, and $A = DArea(S)$. Then

$$\frac{\partial A}{\partial s_i} = alt(s_i; S).$$

Proof: Let $r = \text{Radius}(S)$, so that the circle $C(S)$ has radius r and center o_S . Let $a_i = \text{alt}(s_i, S)$, and define θ_i to be the angle at vertex o_S of $\triangle(s_i; S)$. By definition, $|e(s_i)| = 2r \sin \frac{\theta_i}{2}$, and $a_i = r \cos \frac{\theta_i}{2}$.

Let us suppose, for the moment, that we do not have a Dido exception so that $s_i = |e(s_i)|$ for all i .

Then

$$\begin{aligned} dA &= \sum_{i=1}^n d\left(\frac{a_i s_i}{2}\right) \\ &= \frac{1}{2} \sum_{i=1}^n a_i ds_i + \frac{1}{2} \sum_{i=1}^n s_i da_i \end{aligned} \quad (1)$$

Now,

$$da_i = \cos \frac{\theta_i}{2} dr - \frac{1}{2} r \sin \frac{\theta_i}{2} d\theta_i, \quad (2)$$

and

$$ds_i = 2 \sin \frac{\theta_i}{2} dr + r \cos \theta_i d\theta_i. \quad (3)$$

So

$$\begin{aligned} \frac{1}{2} s_i da_i &= r \sin \frac{\theta_i}{2} \cos \frac{\theta_i}{2} dr - \frac{1}{2} r^2 \sin^2 \frac{\theta_i}{2} d\theta_i \\ &= r \sin \frac{\theta_i}{2} \cos \frac{\theta_i}{2} dr - \frac{1}{2} r^2 (1 - \cos^2 \frac{\theta_i}{2}) d\theta_i \\ &= \frac{1}{2} r \cos \frac{\theta_i}{2} (2 \sin \frac{\theta_i}{2} dr + r \cos \theta_i d\theta_i) - \frac{1}{2} r^2 d\theta_i. \end{aligned}$$

Substituting from (3) gives

$$\frac{1}{2} s_i da_i = \frac{1}{2} a_i ds_i - \frac{1}{2} r^2 d\theta_i. \quad (4)$$

Substituting from (4) in (1) gives:

$$dA = \sum_{i=1}^n a_i ds_i - \frac{r^2}{2} \sum_{i=1}^n d\theta_i. \quad (5)$$

But $\sum_{i=1}^n \theta_i = 2\pi$, and hence $\sum_{i=1}^n d\theta_i = 0$. Since $dA \equiv \sum_{i=1}^n \frac{\partial A}{\partial s_i} ds_i$, the result now follows when there is no Dido exception.

We now need to observe that $\frac{\partial A}{\partial s_i}$ is continuously differentiable right through any Dido exception, which can occur in just two ways. If s_i were to increase so that θ_i became π , then subsequent increases in s_i would yield $\frac{\partial A}{\partial s_i} = 0$, which gives a smooth continuation of A as a function of s_i . If, on the other hand, a decrease of s_i increased some θ_k , for $k \neq i$ to π , then the system of equations would change. In particular, we would have $\theta_k = \pi$ and $a_k = 0$ as s_i continues to decrease. The net consequence is that all of the equations would still hold, and $\frac{\partial A}{\partial s_i}$ will be precisely $r \cos \frac{\theta_i}{2}$ on either side of this transition, and the area therefore remains continuously differentiable across the Dido exception. ■

It should be noted that a Dido exception does cause a discontinuity in $\frac{\partial r}{\partial s_i}$, but this does not affect the area. It is also worth remarking that the altitude, in Lemma 1, is actually a signed quantity. In particular, if an angle θ_i exceeds π , then a_i will be negative, as will the contribution of $\text{Area}(\triangle(s_i; S))$ to A . But for our present purposes, this issue is irrelevant, since the construction for Dido exceptions prohibits such an occurrence.

Proposition 1. Let $\triangle ABC$ be an isosceles triangle with base $\overline{BC}$. Let D be on the segment $\overline{BC}$, and let $\triangle EBD$ and $\triangle FDC$ be isosceles with respective bases $\overline{BD}$ and $\overline{DC}$. Suppose that the three triangles have equal altitudes with respect to the baseline $\overline{BC}$. Let $\angle BAC = \theta$, $\angle BED = \alpha$, and $\angle DFC = \beta$.

Then $\alpha + \beta \leq 4 \arctan(\frac{1}{2} \tan \frac{\theta}{2})$, and equality holds if and only if D is the midpoint of $\overline{BC}$.

Proof: Straightforward. ■

Proposition 2. Let $g(\theta) = 4 \arctan(\frac{1}{2} \tan \frac{\theta}{2})$, for $0 \leq \theta \leq \pi$. Then

1) g is strictly convex.

2) $g(\gamma x) < \gamma g(x)$ for $0 < \gamma < 1$ and $0 < x \leq \pi$.

Proof: 1) Differentiating gives $g'(\theta) = \frac{1}{4 + \frac{3}{4} \cos^2 \frac{\theta}{2}}$, which implies that $g'' > 0$. Thus g is strictly convex.

2) Let x be any fixed value that satisfies $0 < x \leq \pi$. Since $g(0) = 0$ and g is strictly convex, the curve

$g(\gamma x)$ lies under the line $y = \gamma g(x)$ for $0 < \gamma < 1$. ■

Lemma 2. For $0 \leq \theta \leq \pi$,

$$\frac{2\pi - \theta}{\theta} 4 \arctan\left(\frac{1}{2} \tan \frac{\theta}{2}\right) \leq 2\pi,$$

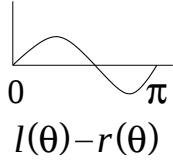
and equality holds if and only if $\theta = 0$ or $\theta = \pi$.

Proof: It is evident that equality is achieved in the limit for $\theta = 0, \pi$. We need to show that

$$4 \arctan\left(\frac{1}{2} \tan \frac{\theta}{2}\right) < \frac{\theta}{1 - \theta/2\pi}$$

for $0 < \theta < \pi$. Let $L(\theta) = 4 \arctan(\frac{1}{2} \tan \frac{\theta}{2})$, and $R(\theta) = \frac{\theta}{1 - \theta/2\pi}$. To see that the $L(\theta) < R(\theta)$ for $0 < \theta < \pi$, we exploit the fact that equality is achieved at the endpoints. It suffices to show that if $L'(\theta) > R'(\theta)$, then $L'(\hat{\theta}) > R'(\hat{\theta})$, for all $\hat{\theta}$: $\theta < \hat{\theta} < \pi$. Indeed, this last inequality would imply that since $L(0) - R(0) = 0$, this difference must be initially decreasing and hence less than 0 for all $\theta < \theta_0$, for some fixed θ_0 , at which point the difference starts to increase, and continues rising to ultimately reach $L(\pi) - R(\pi) = 0$.

Differentiating gives $L' = \frac{1}{\cos^2 \frac{\theta}{2} + \frac{1}{4} \sin^2 \frac{\theta}{2}}$ and $R' = \frac{1}{(1 - \frac{\theta}{2\pi})^2}$. We need only show that $L'(\theta) - R'(\theta)$ has one zero in $(0, \pi)$, and that the transition is from negative to positive. Alternatively, we can prove that $\frac{1}{L'(\theta)} - \frac{1}{R'(\theta)}$ has one zero in $(0, \pi)$, and that the transition is from positive to negative. Accordingly, let $l(\theta) = \frac{1}{L'(\theta)} = \cos^2 \frac{\theta}{2} + \frac{1}{4} \sin^2 \frac{\theta}{2}$, and $r(\theta) = \frac{1}{R'(\theta)} = (1 - \frac{\theta}{2\pi})^2$.



Now, $l(\theta) - r(\theta)$ is zero for $\theta = 0, \pi$. To see that $l(\theta) - r(\theta)$ has only one zero in $(0, \pi)$ and is initially increasing, we can differentiate $l - r$ and see if $l'(\theta) - r'(\theta)$ is initially positive and has only two zeros in $(0, \pi)$. The two zeros will occur because $l(\theta) - r(\theta)$ must be zero at three points in $[0, \pi]$. If $l - r$ were zero at additional locations, $l' - r'$ would have additional zeros as well.

Differentiating gives $l'(\theta) = -\frac{3}{8} \sin \theta$, and $r'(\theta) = -\frac{1}{\pi}(1 - \frac{\theta}{2\pi})$. Since $-\frac{3}{8} \sin \theta$ is strictly convex on $[0, \pi]$, and $r'(\theta)$ defines a straight line, the two curves can intersect at most twice. Finally, we note that $l'(0) - r'(0) = \frac{1}{\pi} > 0$, and the desired conclusion now follows. ■

Proposition 3. Let $S = \{s_1, s_2, \dots, s_n\}$, and $T = \{t_1, t_2, \dots, t_n\}$ be multisets of nonnegative elements where $s_i \leq t_i$, for $i = 1, 2, \dots, n$. Then

$$\text{Radius}(S) \leq \text{Radius}(T),$$

and equality holds if and only if the largest value in S is at least the length of the longest edge in $\text{Poly}(T)$, and all of the other values in S equal the corresponding values in T .

Proof: Let $C(T)$ be a circle with the inscribed figure $\text{Poly}(T)$. Let t_n be a maximal element in T . By swapping the value assigned to a maximal sized element in S with the assignment to s_n , we can assume that s_n is maximal. It is easy to see that this change preserves the inequalities $s_i \leq t_i$ for all i .

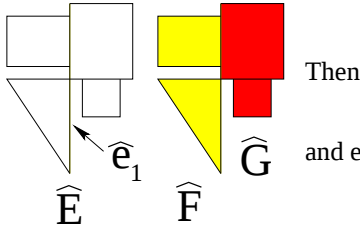
Let a chain of segments with lengths $s_1, s_2, \dots, s_{n-1}$ be inscribed in $C(T)$. Since these segments are no longer than their counterparts in $\text{Poly}(T)$, the chain must span a radial arc of angle α that is no larger than that spanned by the corresponding $n - 1$ segments of $\text{Poly}(T)$. Evidently, the chain will span the same radial arc if and only if $t_i = s_i$ for $i = 1, 2, \dots, n - 1$. In this case, the $\text{Radius}(S)$ will equal $\text{Radius}(T)$ if and only if s_n is at least as large as the longest edge in $\text{Poly}(T)$, which is bounded by t_n .

Now suppose that α is less than the angle spanned by the corresponding edges of $\text{Poly}(T)$. If $\alpha \geq \pi$, then the gap between the endpoints of the chain must exceed t_n , and $\text{Radius}(S)$ must therefore be less than $\text{Radius}(T)$. On the other hand, if $\alpha < \pi$, then $\text{Radius}(S)$ must be less than $\text{Radius}(T)$ because either s_n causes a Dido exception or s_n is less than the length of the gap between the endpoints of the chain as inscribed in $C(T)$. ■

The next inequality enables divide and conquer to be used in an area conserving (or decreasing) way. While the theorem is just a statement about the multisets of numbers E , F , and G , it will be applied as a statement about special collections $\hat{E}$, $\hat{F}$, and $\hat{G}$ of located segments and the areas they enclose.

Theorem 3.1. Let $E = \{e_i\}_{i=1}^n$, $F = \{f_i\}_{i=1}^n$ and $G = \{g_i\}_{i=1}^n$ be multisets with nonnegative entries. Suppose that

- 1) $f_1 \leq e_1 \geq g_1$; (Intuitively, segment $\hat{e}_1$ can be shared by subregions $\hat{F}$ and $\hat{G}$ of the region $\hat{E}$.)
- 2) $e_i = f_i + g_i$, for $i > 1$; (Intuitively, any other segment might be partitioned by $\hat{e}_1$.)
- 3) $e_1 \geq e_j$ for all j where $f_j \cdot g_j > 0$. (Intuitively, $\hat{e}_1$ is longer than any segment that crosses $\hat{e}_1$.)



Then

$$\text{DArea}(E) \geq \text{DArea}(F) + \text{DArea}(G),$$

and equality can only hold if $\text{DArea}(F) \cdot \text{DArea}(G) = 0$.

Proof: The proof will be by contradiction.

Suppose that $\text{DArea}(F) + \text{DArea}(G) - \text{DArea}(E) > 0$. Let us seek to maximize this expression subject to the constraints that e_1 is held constant and all other values can be diminished but not increased. In view of Lemma 1, we can assume that $e_1 = f_1 = g_1$. Formally, let $X = \{x_1, x_2, \dots, x_n\}$, $Y = \{y_1, y_2, \dots, y_n\}$, $Z = \{z_1, z_2, \dots, z_n\}$, and suppose that $\text{DArea}(Y) + \text{DArea}(Z) - \text{DArea}(X)$ is maximal for X , Y , and Z constrained as follows: $x_1 = y_1 = z_1 = e_1$, and for $i = 2, 3, \dots, n$: $0 \leq x_i \leq e_i$, $0 \leq y_i$, $0 \leq z_i$, $x_i = y_i + z_i$.

This problem is just a matter of calculus. Since the function $\text{DArea}(Y) + \text{DArea}(Z) - \text{DArea}(X)$ is continuous on the compact domain for $X \times Y \times Z$, the maximum value must be achieved somewhere, so let (X, Y, Z) be a point where this maximum occurs.

Let $\text{Poly}(X)$, $\text{Poly}(Y)$, and $\text{Poly}(Z)$, have the respective circumcenters o_X , o_Y , and o_Z . Let the radii of the circumcircles be $r_X = \text{Radius}(X)$, $r_Y = \text{Radius}(Y)$, and $r_Z = \text{Radius}(Z)$. Let segments connect each vertex of $\text{Poly}(X)$, $\text{Poly}(Y)$, and $\text{Poly}(Z)$ with, respectively, o_X , o_Y , and o_Z , so that each of the inscribed polygons is triangulated with the polygon's circumcenter serving as a common apex for all of its triangles.

For convenience, let x_i denote the edge $e(x_i)$ as well as a numerical element in X , and let the definitions for y_i and z_i be extended in an analogous manner.

Let, for each i , $\triangle(x_i; X)$ have an angle of θ_i at vertex o_X ; let $\triangle(y_i; Y)$ and $\triangle(z_i; Z)$ have the respective angles α_i , and β_i at o_Y and o_Z .

Let X be fixed. By the properties of calculus, it follows that for $i > 1$, either y_i is extremal (i.e., $y_i = 0$ or $y_i = x_i$) or $\frac{\partial \text{DArea}(Y)}{\partial y_i} + \frac{\partial \text{DArea}(Z)}{\partial y_i} = 0$. By Lemma 1, the equation $\frac{\partial \text{DArea}(Y)}{\partial y_i} + \frac{\partial \text{DArea}(Z)}{\partial y_i} = 0$ says that the local altitude from the edge y_i to o_Y and from edge z_i to o_Z are equal: $\text{alt}(y_i; Y) = \text{alt}(z_i; Z)$, since $z_i = x_i - y_i$. Thus, we view each edge x_i , for $i > 1$ as either whole and belonging to either Y or Z , or as partitioned among Y and Z , in which case the two pieces y_i and z_i have equal altitudes as defined above. As the remarks that conclude the proof of Lemma 1 point out, the occurrence of a Dido exception does not limit the scope this observation.

On the other hand, if an edge x_j is not partitioned, then $\text{alt}(x_j; X) > \text{alt}(y_j; Y)$, if $y_j = x_j$, or $\text{alt}(x_j; X) > \text{alt}(z_j; Z)$, if $z_j = x_j$, since $r_X > r_Y$ and $r_X > r_Z$. In this case, we can reduce x_j (while maintaining equality between x_j and its counterpart in Y or Z), whence $\text{DArea}(Y) + \text{DArea}(Z) - \text{DArea}(X)$ will increase, since the partial derivative of the sum with respect to x_j will be negative. While it is possible that x_j and its counterpart might be Dido exceptions, the values can still be reduced, and x_j would be the first to cease being an exception, which would occur when $x_j = 2r_X$. So even here the observation holds: at the (phantom) maximization point, all X elements other than x_1 must be partitioned between the Y and Z multisets, so that $y_i z_i \neq 0$, for all i .

Consequently, $\text{alt}(y_i, Y) = \text{alt}(z_i, Z)$, for $i > 1$.

By the same reasoning, we can conclude that $\text{alt}(y_j, Y) \geq \text{alt}(x_j, X)$, for $j > 1$: otherwise fix z_j ; $\frac{\partial \text{DArea}(Y)}{\partial x_j} + \frac{\partial \text{DArea}(Z)}{\partial x_j} - \frac{\partial \text{DArea}(X)}{\partial x_j}$ will be negative, and diminishing x_j will cause $\text{DArea}(Y) + \text{DArea}(Z) - \text{DArea}(X)$ to increase.

At this point, the assumptions of Theorem 2 can be seen to ensure that $|e(x_1)| \geq |e(x_j)|$ for all j , since all edges of X are partitioned, the size of edge x_1 did not diminish from e_1 (unless it became a Dido exception) and none of the other lengths increased from their corresponding values in E .

We can now use Proposition 2 and Lemma 2 to complete the contradiction. Evidently, $\sum \theta_i = \sum_i \alpha_i = \sum_i \beta_i = 2\pi$. By construction, $\alpha_1 \leq \pi$, and likewise $\beta_1 \leq \pi$. We will show that $\sum_{j>1} (\alpha_j + \beta_j) < 2\pi$, which gives a contradiction.

Let $g(\theta) = 4 \arctan(\frac{1}{2} \tan \frac{\theta}{2})$. Proposition 1 shows that

$$\sum_{j>1} (\alpha_j + \beta_j) \leq \sum_{j>1} g(\theta_j).$$

Since $\theta_1 \geq \theta_i$ for all i , the convexity established in Proposition 2 ensures that

$$g(\theta_j) = g(\frac{\theta_j}{\theta_1} \theta_1) \leq \frac{\theta_j}{\theta_1} g(\theta_1),$$

whence

$$\sum_{j>1} (\alpha_j + \beta_j) \leq \sum_{j>1} g(\theta_j) \leq \frac{g(\theta_1)}{\theta_1} \sum_{j>1} \theta_j = \frac{g(\theta_1)(2\pi - \theta_1)}{\theta_1},$$

and hence

$$\sum_{j>1} (\alpha_j + \beta_j) \leq 2\pi$$

by Lemma 2. Moreover, Lemma 2 shows that $\sum_{j>1} (\alpha_j + \beta_j) = 2\pi$ only if all angles θ_i are zero or π , and we conclude that X must be composed of two edges of length $2r_X$, and any number of edges of zero length. Consequently, the partitioning must give rise to comparably trivial polygons $\text{Poly}(Y)$ and $\text{Poly}(Z)$, which each have no area.

The supposition that $\text{DArea}(F) + \text{DArea}(G) - \text{DArea}(E) > 0$ has given a contradiction. $\blacksquare$

Evidently, the insufficiency in the sum of the angles α_i and β_i is a consequence of the assumption that the altitudes associated with Y and Z exceed those associated with X . In reality, those altitudes are, overall, smaller, and the areas for Y and Z are not as great, in total, as that for X .

At this point, Theorem 1 is an easy exercise.

Theorem 1. Let $\hat{S}$ be a collection of edges placed in the plane R^2 . Let S be the multiset corresponding to the lengths of the edges in $\hat{S}$. Then

$$\text{Area}(\hat{S}) \leq \text{DArea}(S).$$

Proof: Let $\mathcal{W}$ be the region enclosed by $\hat{S}$. Let $\hat{S}$ be optimized so that each edge in $\hat{S}$ cannot be reduced in length without changing the area of $\hat{S}$. It follows that every edge in $\hat{S}$ has some neighborhood of each endpoint on the boundary of $\mathcal{W}$.

- Define an edge e in $\hat{S}$ to be *normal* if one entire side of e is exposed to the exterior of $\mathcal{W}$, and the other entire side is exposed to the interior. An edge is *abnormal* if it is not normal.

Evidently, the collection $\hat{S}$ of optimized edges defines a polygon if and only if every edge is normal.

The edge lengths of $\hat{S}$ will comprise the multiset S . Let the longest abnormal edge $e \in \hat{S}$ be used to partition $\hat{S}$ into two arrangements $\hat{T}$ and $\hat{U}$. The associated multisets T and U will meet the criteria of the Theorem 2. Formally, both $\hat{T}$ and $\hat{U}$ will consist of an abnormal splitting edge e and one or more connected point sets of $\cup_{\hat{s} \in \hat{S}} \hat{s} \setminus e$, where the removal of e splits each edge crossing e into two disjoint pieces. While this splitting might create more than two connected components, we can group the components arbitrarily

into two sets $\hat{T}$ and $\hat{U}$, which meet the requirements of Theorem 1. Thus, Theorem 2 gives an (implicit) partitioning of $\text{DArea}(S)$ into $\text{DArea}(T)$ and $\text{DArea}(U)$ with some area left over, and the process can be repeated recursively for $\hat{T}$ and $\hat{U}$. Since $\hat{T}$ and $\hat{U}$ both define fewer subregions than does the edge set $\hat{S}$, the refined collections of edges will eventually define just one subregion each, and have no abnormal edges. Consequently, the standard Isoperimetric Theorem for polygons with a fixed multiset of edge lengths can be applied to show that each terminal subregion will have an area that is bounded by its allotted fraction of $\text{DArea}(S)$. ■

A more formal proof of Theorem 1 would be by contradiction, and would suppose that $\hat{S}$ is a counterexample to Theorem 1 where $R^2 \setminus \hat{S}$ has as small a number of components as possible. The recursive formulation was offered to illustrate the intuition behind the divide and conquer proof technique.

3 Conjectures and open problems

Suppose that F is a finite collection of segments.

- Let $\text{Dido}(F) = \text{DArea}(F^+)$, where F^+ is the multiset of lengths $\{|f|\}_{f \in F} \cup \{\sum_{f \in F} |f|\}$.

Thus, $\text{Dido}(F)$ is the area of $\text{ConvHull}(G)$, where G is an open polygonal path that is tightly inscribed in a semicircle, and is composed of segments that are congruent to those in F .

In [2], A. and K. Bezdek posed, and gave a partial proof of the following Dido-type conjecture.

Let F comprise a finite collection of located segments in the plane, and suppose that F is connected as a point-set.

Then

$$\text{Area}(\text{ConvHull}(F)) \leq \text{Dido}(F).$$

This bound was fully established and generalized in [15]. We now offer a somewhat stronger conjecture plus extensions to include the Dido-type problem of Fejes Tóth and Hajós.

Conjecture 1. Let F be a finite collection of segments located in the plane. Let $\text{Umbra}(F)$ be the set of points x such that every line through x intersects some segment in F . Then

$$\text{Area}(\text{Umbra}(F)) \leq \text{Dido}(F).$$

Conjecture 2. Let F be a finite collection of segments located in the plane. Let $\text{Encloak}(F)$ be the set of points x such that every line ℓ through x intersects at least two segments in F . Then

$$\text{Area}(\text{Encloak}(F)) \leq \text{DArea}(F).$$

Other questions. It should be evident that additional problems can be formulated by demanding higher intersection counts. These problems can also be extended to higher dimensions. For example, the formulations of Böröczky, Bárány, Makai Jr. and Pach [3] might be adapted as follows. Suppose a region F is enclosed by a collection P of polytopes that each lie in some hyperplane of R^n . Then we should expect that $\text{Vol}(F)$ is bounded by largest volume of any polytope whose facets (i.e. $(n - 1)$ -dimensional faces) have surface areas that are bounded by corresponding areas of the elements in P , and likewise for $\text{Vol}(\text{Encloak}(P))$.

References

- [1] Benson, R.V. Euclidean Geometry and Convexity, McGraw-Hill, Inc., 1966.
- [2] Bezdek, A., Bezdek, K. (1989) On a discrete Dido-type question. *Elem. Math.* 4: 92–100.
- [3] Böröczky, K., Bárány, I., Makai Jr., E., and Pach, J., Maximal volume enclosed by plates and proof of the chessboard conjecture, *Discrete Math.* 60 (1986), 101–120.

- [4] Courant, Richard, and Herbert Robbins. What is Mathematics? Oxford University Press, 1960.
- [5] Croft, H.T., K.J. Falconer and R.K. Guy, Unsolved Problems in Geometry, Springer-Verlag New York, Inc., 1991.
- [6] Fejes Tóth, L. Über das Didosche Problem, *Elem. Math.* 23 (1968), 97–101.
- [7] Fejes Tóth, L. Research Problem No. 6. *Period. Math. Hungar* 4 (1973), 231–232.
- [8] Kazarinoff, N.D., Geometric Inequalities. Random House, 1961.
- [9] Lax, P.D. A Short Path to the Shortest Path, *American Mathematical Monthly*, 158–159, Feb., 1995.
- [10] Osserman, R. The isoperimetric inequality, *Bulletin of the American Mathematical Society*, 84 (1978), 1182–1238.
- [11] Pach, J. On an isoperimetric problem. *Studia Sci. Math. Hungar.* 13 (1978), no. 1-2, 43–45.
- [12] Pach, J. Personal communication.
- [13] Pappus, of Alexandria. La collection mathématique; oeuvre traduite pour la première fois du grec en français par Paul Ver Eecke. Paris, Bruges, Desclee, de Brouwer et cie, 1933.
- [14] Rademacher, H. and O. Toeplitz. The Enjoyment of Mathematics, Princeton University Press, 1994.
- [15] Siegel, A. Some Dido-type Inequalities. *Elem. Math.* 56 (2001), 1–4.
- [16] Virgil (A.K.A. Publius Vergilius Maro). Aeneid I, 365–368.
- [17] Yaglom, I.M., and V.G. Boltanskiĭ. Convex Figures, translated by P.J. Kelly and L.F. Walton, Holt, Rinehart and Winton, 1961.