The distance trisector curve

Tetsuo Asano*
School of Information Science
JAIST
1-1 Asahidai, Nomi, Ishikawa,
923-1292 Japan
t-asano@jaist.ac.jp

Jiří Matoušek**
Department of Applied Mathematics and
Institute of Theoretical Computer Science,
Charles University
Malostranské nám. 25,
118 00 Praha 1
Czech Republic
matousek@kam.mff.cuni.cz

Takeshi Tokuyama‡
GSIS
Tohoku University
Aramaki Aza Aoba, Aoba-ku,
Sendai, 980-8579 Japan
tokuyama@dais.is.tohoku.ac.jp

ABSTRACT

Given points $\mathbf{p}$ and $\mathbf{q}$ in the plane, we are interested in separating them by two curves C_1 and C_2 such that every point of C_1 has equal distance to $\mathbf{p}$ and to C_2, and every point of C_2 has equal distance to C_1 and to $\mathbf{q}$. We show by elementary geometric means that such C_1 and C_2 exist and are unique. Moreover, for $\mathbf{p} = (0, 1)$ and $\mathbf{q} = (0, -1)$, C_1 is the graph of a function $f : \mathbb{R} \to \mathbb{R}$, C_2 is the graph of $-f$, and f is convex and analytic (i.e., given by a convergent power series at a neighborhood of every point). We conjecture that f is not expressible by elementary functions and, in particular, not algebraic. We provide an algorithm that, given $x \in \mathbb{R}$ and $\varepsilon > 0$, computes an approximation to $f(x)$ with error at most ε in time polynomial in $\log \frac{1+|x|}{\varepsilon}$.

The separation of two points by two "trisector" curves considered here is a special (two-point) case of a new kind of Voronoi diagram, which we call the *Voronoi diagram with neutral zone* and which we investigate in a companion paper.

1. INTRODUCTION

The two curves C_1 and C_2 in Figure 1 have the following property: Every point of C_2 has the same distance to the point $\mathbf{q} = (0, -1)$ and to C_1 (as is indicated for one point of C_2 in the drawing), and similarly, every point of C_1 has equal distance to $\mathbf{p} = (0, 1)$ and to C_1. Some preliminary results about such curves have been reported in [2].

We call such C_1 and C_2 *distance trisector curves* of $\mathbf{p}$ and $\mathbf{q}$. This notion is motivated by a routing problem on a printed circuit board layout raised by Dr. Hiroshi Murata from Kitakyusyu University (personal communication to T. Asano, 2002): Given two points $\mathbf{p}$ and $\mathbf{q}$ in the plane, we want to draw k "equally spaced curves" $C_1, C_2, \ldots, C_k$ separating them. A natural interpretation of this requirement is this: C_i should be a bisector of C_{i-1} and C_{i+1}, where $C_0 = \{\mathbf{p}\}$ and $C_{k+1} = \{\mathbf{q}\}$. That is, C_i is the set of points with equal distance to C_{i-1} and C_{i+1}, $i = 1, 2, \ldots, k$.

For $k = 1$, C_1 is the *bisector* of $\mathbf{p}$ and $\mathbf{q}$, i.e., the line perpendicular to the segment $\mathbf{pq}$ and going through its midpoint. For $k = 3$, we can take the bisector of $\mathbf{p}$ and $\mathbf{q}$ for C_2, and C_1 and C_3 are parabolas (bisectors of a point and a line). The cases $k = 1$ and $k = 3$ are the only ones where the existence of such curves is obvious, and even in the $k = 3$ case, the uniqueness of the solution is not immediate.

Main results. In this paper we consider the case $k = 2$ (distance trisector curves). By elementary geometric arguments we prove the following:

THEOREM 1 (EXISTENCE AND UNIQUENESS). *There exists exactly one pair of curves (C_1, C_2) that are distance trisector curves of the points $\mathbf{p} = (0, 1)$ and $\mathbf{q} = (0, -1)$. They are the graphs of f and $-f$, respectively, where $f : \mathbb{R} \to \mathbb{R}$ is a convex continuous function.*

Computational geometry usually works with lines, circles, quadrics, or bounded-degree algebraic curves. These curves are considered to be "known": Operations such as locating a query point with respect to them, say above/below, or intersecting them with other such curves, are assumed to be doable in constant time, and implementations are available for the most common cases.

Only upon encountering the distance trisector curve did we realize that it is not so clear what one means by "knowing" a curve. For example, it is one thing to be able to plot the curve, and another thing to be able to decide a point-location query. We suspect that the *exact* point location query for the trisector curve might be undecidable in the Real RAM model, since we conjecture the curve to be

*The part of this research by T.A. was partially supported by the Ministry of Education, Science, Sports and Culture, Grant-in-Aid for Scientific Research on Priority Areas and Scientific Research (B).

**Parts of this research by J.M. were done during visits to the Japanese Advanced Institute for Science and Technology (JAIST) and to the ETH Zürich; the support of these institutions is gratefully acknowledged.

‡The part of this research by T.T. was partially supported by the Ministry of Education, Science, Sports and Culture, Grant-in-Aid for Scientific Research on Priority Areas.

STOC'06, July 22-24, 2006, Seatle, Washington.

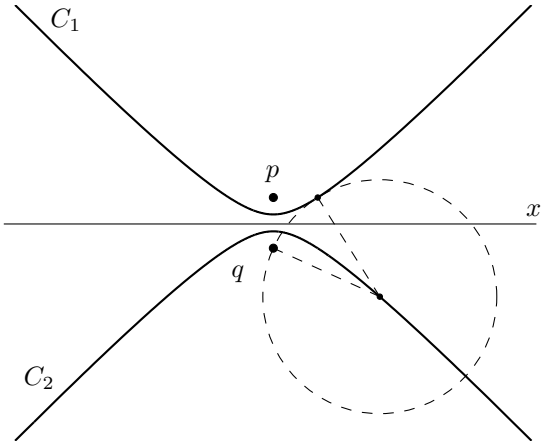

Figure 1: The distance trisector curves

highly transcendental. Yet it turns out that the curve can be approximated efficiently; essentially, it can be evaluated at any point to n digits in time polynomial in n.

THEOREM 2 (APPROXIMATE EVALUATION).

(i) The function f as in Theorem 1 is analytic. That is, for every x_0 there is a neighborhood on which it can be expressed by a convergent power series in $x - x_0$.

(ii) For every $x \in \mathbb{R}$ and every $\varepsilon > 0$, the value of $f(x)$ can be computed with accuracy ε in time polynomial in $\log \frac{1+|x|}{\varepsilon}$. (We assume that x is accessed via an oracle that returns the first n significant digits of x in time polynomial in n.)

Discussion. We consider the definition of the distance trisector curve very natural, and we were surprised to find no traces of it in the literature (so far; we will be very grateful for any pointers or tips). Before starting this research, we had a vague general feeling that all "natural" curves had been discovered and thoroughly investigated, if not by Newton, Euler, or the Bernoullis, then in the 19th century at the latest. However, curves commonly mentioned in the literature (see, for example, the "Famous Curves Index" [4]) have a (simple) algebraic equation, or at least they can be expressed using exponential and trigonometric functions. Moreover, geometrically they are usually defined in terms of other, previously defined objects (as caustic curves, evolutes, involutes, pedal curves, inverse curves, etc.). If the initial objects are curves with equations expressible by elementary functions, then the listed constructions do not leave the realm of such curves either.

In contrast, the definition of the distance trisector curve is self-referential; the curve can be regarded as a fixed point of a certain operator acting globally on curves. Moreover, the definition involves distances of points to the curve being defined, and so, expressed formally, it is not a first-order predicate (roughly speaking, it is not sufficient to talk about finitely many points at a time in the definition).

We conjecture that the distance trisector curve is not algebraic, and actually, that it cannot be expressed by elementary functions. Such a result would resemble the famous results, going back to Liouville, on the impossibility of expressing certain primitive functions, such as $\int e^{x^2} \mathrm{d}x$, in terms of elementary functions (see, e.g., [6]). However, the techniques used there do not seem immediately applicable to our problem, and probably one should begin with the more modest goal of proving the curve to be transcendental.

Voronoi diagrams with neutral zones. Another direction of generalizing the distance trisector curve, besides the problem of k equidistant curves, is an apparently new and interesting variation on the classical notion of Voronoi diagram. There are several generalizations and variations of Voronoi diagrams, and their geometric properties and computational complexities are widely studied; see, e.g., [3, 7]. A common feature of these variations is that they define *partitions* of space into regions (*Voronoi cells*), each of which is the dominating region of an input point or object. The *Voronoi diagram with neutral zone*, which we investigate in the companion paper [1], can be regarded as a model of a growth process where the growth from each site terminates before the boundaries meet, and the termination is due to some long-distance action.

A Voronoi diagram with neutral zone for 5 points (marked by crosses) is shown in Fig. 2. The region of a site $\mathbf{p}$ consists of points that are closer to $\mathbf{p}$ than to the union of the regions of the remaining sites.

If there are only two sites $\mathbf{p}$ and $\mathbf{q}$, the borders of the regions are exactly the distance trisector curves of $\mathbf{p}$ and $\mathbf{q}$. In this respect, the distance trisector curves play a role somewhat analogous to the role of perpendicular bisectors (lines) in ordinary Voronoi diagrams. However, while all regions in an ordinary Voronoi diagram are bounded by segments of the bisectors (line segments), the regions in the Voronoi diagram with neutral zone are in general *not* bounded by segments of the distance trisector curves. Still, it is clear that understanding the distance trisector curves is a necessary prerequisite for studying Voronoi diagrams with neutral zone.

2. EXISTENCE AND UNIQUENESS

In this section we sketch the proof of Theorem 1. We begin with preliminaries, we formally define the bisector of a point and a set and the closely related concept of dominance region, and we prove some simple properties.

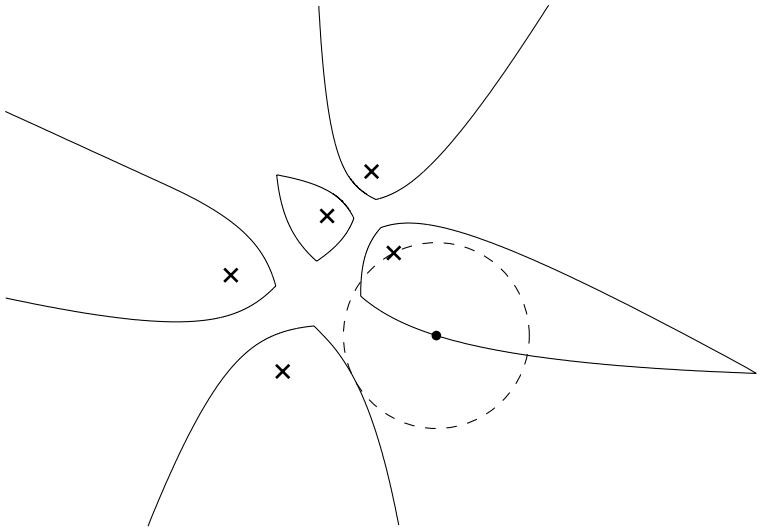

Figure 2: A Voronoi diagram with neutral zones

For a function $f\colon \mathbb{R} \to \mathbb{R}$ we let $C(f) = \{(x, f(x)) : x \in \mathbb{R}\} \subset \mathbb{R}^2$ denote the graph of f.The inequality $f \le g$ between functions means $f(x) \le g(x)$ for all $x \in \mathbb{R}$.

Dominance region and bisector. For a point $\mathbf{a}$ and a set $X \subseteq \mathbb{R}^2$ we define the *dominance region* of $\mathbf{a}$ with respect to X as

$$\mathrm{dom}(\mathbf{a}, X) = \{\mathbf{z} \in \mathbb{R}^2 : d(\mathbf{z}, \mathbf{a}) \le d(\mathbf{z}, X)\},$$

where $d(\cdot,\cdot)$ denotes the Euclidean distance and $d(\mathbf{z}, X) = \inf_{\mathbf{x}\in X} d(\mathbf{z}, \mathbf{x})$. The *bisector* of $\mathbf{a}$ and X is

$$\mathrm{bisect}(\mathbf{a}, X) = \{\mathbf{z} \in \mathbb{R}^2 : d(\mathbf{z}, \mathbf{a}) = d(\mathbf{z}, X)\}.$$

LEMMA 3 (PROPERTIES OF BISECTORS).

(i) $\mathrm{dom}(\mathbf{a}, X)$ *is a closed convex set for every* $\mathbf{a}$ *and every* X.

(ii) (Antimonotonicity) The operator $\mathrm{dom}(.,.)$ *is antimonotone with respect to the second argument; that is, if* $X \subseteq X'$, *then* $\mathrm{dom}(\mathbf{a}, X) \supseteq \mathrm{dom}(\mathbf{a}, X')$.

(iii) If $\mathbf{a}$ *doesn't lie in the closure of* X, *then the bisector* $\mathrm{bisect}(\mathbf{a}, X)$ *equals the boundary of* $\mathrm{dom}(\mathbf{a}, X)$.

(iv) Let $\mathbf{p} = (0, 1)$ *and suppose that* X *is contained in the lower halfplane* $L = \{(x, y) : y \le 0\}$ *and contains the point* $\mathbf{q} = (0, -1)$. *Then* $\mathrm{bisect}(\mathbf{p}, X)$ *is contained in the upper halfplane and it intersects every vertical line exactly once; thus, it is the graph of a convex function* $f\colon \mathbb{R} \to [0, \infty)$.

(v) If $\mathbf{p}$, X, *and* f *are as in (iv) and, moreover,* X *is a closed convex set, then the derivative* $f'(x)$ *exists for all* $x \in \mathbb{R}$.

(vi) If $\mathbf{p}$ *and* X *are as in (v), and* $\mathbf{z}$ *is a point of* $\mathrm{bisect}(\mathbf{p}, X)$, *then there exists a unique point* $\mathbf{z}' \in X$ *nearest to* $\mathbf{z}$, *the segment* $\mathbf{z}'\mathbf{z}$ *is an outer normal of* X *at* $\mathbf{z}'$ *(that is, it is perpendicular to some supporting line of* X *at* $\mathbf{z}'$*), and the (unique) tangent of* $\mathrm{bisect}(\mathbf{p}, X)$ *at* $\mathbf{z}$ *is the perpendicular bisector of the points* $\mathbf{p}$ *and* $\mathbf{z}'$; *see Fig. 3.*

Proof. Easy and omitted. □

Outline of the proof of Theorem 1. We define two infinite sequences $(f_1, f_2, f_3, \ldots)$ and $(g_1, g_2, g_3, \ldots)$ of convex functions $\mathbb{R} \to \mathbb{R}$ as follows:

(1) $f_1 \equiv 0$,

(2) $C(g_i) = \mathrm{bisect}(\mathbf{p}, C(-f_i))$, where $\mathbf{p} = (0, 1)$, $i = 1, 2, \ldots$, and

(3) $C(f_{i+1}) = \mathrm{bisect}(\mathbf{p}, C(-g_i))$, $i = 1, 2, \ldots$.

By Lemma 3 the functions f_i and g_i are well-defined, convex, differentiable, and nonnegative. Antimonotonicity yields $f_1 \le f_2 \le f_3 \le \cdots \le g_3 \le g_2 \le g_1$. The sequence $(f_1, f_2, \ldots)$ is nondecreasing and bounded from above (by g_1, say), and so it converges to a (pointwise) limit f, which is finite and convex, and therefore continuous (the convergence is uniform on every bounded interval, but we don't need this). Similarly, the g_i converge to a convex continuous function g, and we have $f \le g$. It is easily seen that $C(f) = \mathrm{bisect}(\mathbf{p}, C(-g))$ and $C(g) = \mathrm{bisect}(\mathbf{p}, C(-f))$.

The following proposition is the technical core of the proof.

PROPOSITION 4. $f = g$.

Once we prove this, we get $C(f) = \mathrm{bisect}(\mathbf{p}, C(-f))$, and thus $C(f)$ is a distance trisector curve. Moreover, supposing that another function $h\colon \mathbb{R} \to [0, \infty)$ satisfies $C(h) = \mathrm{bisect}(\mathbf{p}, C(-h))$, we start with the inequality $f_1 \le h$, and by repeatedly applying $\mathrm{bisect}(\mathbf{p}, .)$ to both sides we get $f_i \le h \le g_i$ for all i. Therefore, $f = h = g$, and the uniqueness follows.[1]

The proof of Proposition 4 relies on the following lemma.

[1]Another very natural proof idea is to define a suitable metric on a suitable space of convex curves such that the operator $C(f) \mapsto \mathrm{bisect}(\mathbf{p}, C(-f))$ is a contraction. Then Banach's theorem would immediately yield existence and uniqueness of a fixed point (which is a distance trisector curve), and we would get some other consequences, such as bounding the convergence of $g_i - f_i$ to 0 by a geometric series. It turned out, though, that some natural metrics do

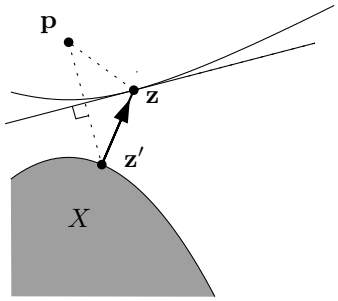

Figure 3: Illustration to Lemma 3(v).

LEMMA 5. *The difference $g-f$ is nondecreasing on $[0,\infty)$.*

Sketch of proof. It suffices to prove $f_i' \le g_i'$ on $[0,\infty)$ for all i. We proceed by induction: from $f_{i-1}' \le g_{i-1}'$ we derive $f_i' \le g_{i-1}'$, and from this we further derive $f_i' \le g_i'$. We omit the proof in this extended abstract.

Proof of Proposition 4. First let us choose $x_0 > 0$ with $g(x_0) \le 1$. We show that $f(x_0) = g(x_0)$; since $g - f$ is nondecreasing, we then have $f = g$ on $[0, x_0]$. For contradiction, let us assume that $\mathbf{a} = (x_0, f(x_0))$ and $\mathbf{b} = (x_0, g(x_0))$ are different points; see Fig. 4left.

Let $\mathbf{b}' = (x_0', -f(x_0'))$ be the point of $C(-f)$ nearest to $\mathbf{b}$. The segment $\mathbf{b}'\mathbf{b}$ has a positive slope, and thus $x_0' < x_0$. We have $d(\mathbf{p},\mathbf{b}) = d(\mathbf{b},\mathbf{b}')$, and since $f(x_0) < g(x_0) \le 1$, we get $d(\mathbf{p},\mathbf{a}) > d(\mathbf{p},\mathbf{b})$. Now we consider a point $\bar{\mathbf{a}}$ such that $\mathbf{b}'\bar{\mathbf{a}}\mathbf{a}\mathbf{b}$ is a parallelogram. Since $g - f$ is nondecreasing, the segment $\mathbf{a}\bar{\mathbf{a}}$ intersects the curve $C(-g)$. Thus $d(\mathbf{a}, C(-g)) < d(\mathbf{a},\bar{a}) = d(\mathbf{b},\mathbf{b}') = d(\mathbf{p},\mathbf{b}) < d(\mathbf{p},\mathbf{a})$, contradicting to $\mathbf{a}$ lying on $C(f) = \mathrm{bisect}(\mathbf{p}, C(-g))$.

We have shown $f = g$ on $[0, x_0]$. Let us now put $s = \sup\{x \ge 0 : f(x) = g(x)\} \ge x_0$. Assuming $s < \infty$, we derive a contradiction. Let us choose a point $\mathbf{b} = (x_1, g(x_1))$ on $C(g)$ such that $x_1 > s$ but the point $\mathbf{b}' = (x_1', -f(x_1'))$ of $C(-f)$ nearest to $\mathbf{b}$ satisfies $x_1' \le s$; see Fig. 4 right. This is possible by a continuity argument, since the outer normal of $C(-f)$ at $\mathbf{z} = (s, -f(s))$ either intersects $C(g)$ right of the vertical line $x = s$, or it misses $C(g)$ altogether, and as we move $\mathbf{z}$ left along $C(-f)$, after some time the outer normal intersects $C(g)$ at $(s, g(s))$.

Having $\mathbf{b}$ and $\mathbf{b}'$ as above, we let $\mathbf{a}$ be the intersection of the segment $\mathbf{b}\mathbf{b}'$ with $C(f)$. We have $\mathbf{a} \ne \mathbf{b}$ since $f(x_1) < g(x_1)$. But since $\mathbf{b}' \in C(-f)$ and $\mathbf{b}'\mathbf{a}$ is normal to $C(-f)$, $\mathbf{b}'$ should be the point of $C(-f)$ nearest to $\mathbf{a}$. We should have both $d(\mathbf{p},\mathbf{b}) = d(\mathbf{b}',\mathbf{b})$ and $d(\mathbf{p},\mathbf{a}) = d(\mathbf{b}',\mathbf{a})$, but this is impossible, because the ray $\mathbf{b}'\mathbf{b}$ contains only one point equidistant to $\mathbf{b}'$ and $\mathbf{p}$. This concludes the proof of Proposition 4, as well as of Theorem 1. □

More properties. We note that the above proof also implies the following result: *For every $a > 0$, the distance trisector curves of $\mathbf{p}$ and $\mathbf{q}$ on the vertical strip $V_a = (-a, a)\times\mathbb{R}$ are uniquely determined; that is, there exists exactly one function $f\colon (-a,a) \to [0,\infty)$ with $C(f) = V_a \cap \mathrm{bisect}(\mathbf{p}, C(-f))$ (where $C(f) = \{(x, f(x)) : x \in (-a,a)\}$).*

We also know that for every $x \in \mathbb{R}$ there exists a unique point of $C(-f)$ nearest to $(x, f(x))$. Let $t(x)$ denote the x-coordinate of this point. For $x \ge 0$ we have $0 \le t(x) \le x$, and $t(-x) = -t(x)$ since f is even. In particular, $t(0) = 0$, and from this we can also see that $f(0) = \frac{1}{3}$.

Since $C(f) = \mathrm{bisect}(\mathbf{p}, C(-f))$, Lemma 3(v) shows that $f'(x)$ exists for every $x \in \mathbb{R}$.

The proof of the following proposition is omitted:

PROPOSITION 6. *The function t is injective (distinct points have distinct images), and it maps $[0,\infty)$ onto the interval $[0, t_{\max})$, where $t_{\max} = \sup\{t(x) : x \in \mathbb{R}\} < \infty$.*

Remark. Numerical computations, using the methods of Section 4, show that $t_{\max} \approx 5.648708769021159$ and $\lim_{x\to\infty} f'(x) \approx 1.083629958775032$.

3. POWER SERIES EXPANSIONS

LEMMA 7. *The following equations are satisfied for every $x \in \mathbb{R}$:*

$$(t(x)-x)^2+(f(t(x))+f(x))^2-x^2-(f(x)-1)^2 = 0, \quad \textit{and} \quad (1)$$

$$t(x) - x + (f(x) + f(t(x))f'(t(x)) = 0, \qquad (2)$$

where $f'(t(x))$ is the derivative of f evaluated at $t(x)$.

Proof. The first equation just says that the point $(x, -f(x))$ equal distances to $\mathbf{p}$ and to $(t(x), f(t(x))$.

For a fixed x, the point $(t(x), -f(t(x)))$ minimizes the squared distance of $(x, f(x))$ to $(t, -f(t))$ among all t. Hence

$$\frac{\partial}{\partial t}\left((t-x)^2 + (f(t)+f(x))^2\right)\Big|_{t=t(x)} = 0,$$

and this yields (2). □

It is easy to check that (1) and (2) determine f and t uniquely. More precisely, if $\tilde{f}$ and $\tilde{t}$ are defined on $(-x_0, x_0)$, $-x_0 < \tilde{t}(x) < x_0$ for all $x \in (-x_0, x_0)$, and $\tilde{f}$ is convex and differentiable, then $\tilde{f} = f$ and $\tilde{t} = t$ on $(-x_0, x_0)$.

LEMMA 8. *There exists $x_0 > 0$ such that on $(-x_0, x_0)$, f and t can be represented as sums of convergent power series in x.*

not work. After the proof presented in this section was finished, and after some experimentation, the second author has found a metric that does yield a proof via Banach's theorem, but formally verifying that we indeed obtain a contraction looks quite complicated at present. So for now we decided to stick to the original proof.

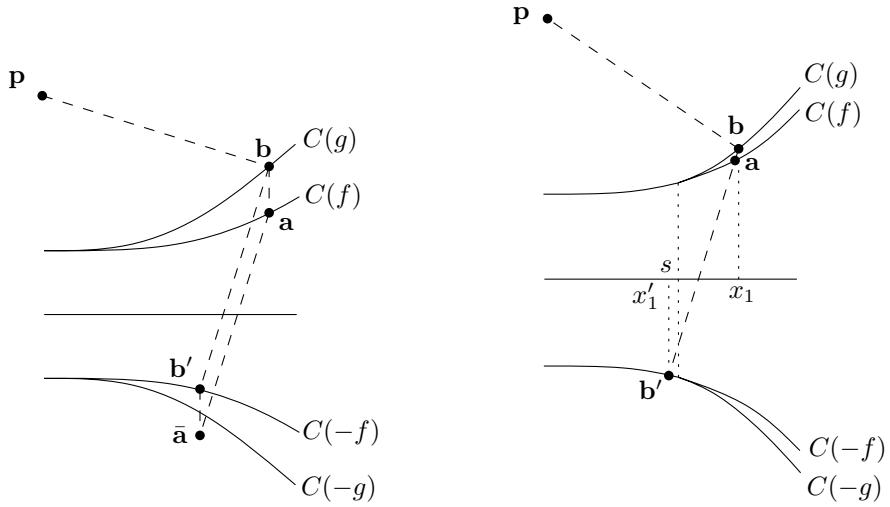

Figure 4: Proving $f = g$.

Sketch of proof. We use the following ingenious parameterization, which was suggested to us by Christian Blatter and which, in a different context, goes back at least to an 1884 paper of Kœnigs (see, e.g., [5], Theorem 8.2). We introduce a new variable z (time) and we look for a real number $\lambda \in (0,1)$ and functions $X(z)$ and $Y(z)$ on some interval $[0, z_0)$ such that for all $z \in [0, z_0)$, if

$$x = X(z),$$

then

$$f(x) = Y(z),\ t(x) = X(\lambda z),\ \text{and}\ f(t(x)) = Y(\lambda z).$$

Here is an outline of the proof. We do not claim at this moment that X, Y, and λ as above necessarily exist; the existence becomes clear only at the end of the proof. We first investigate what X, Y, λ would have to look like if they existed. More precisely, we reformulate equations (1) and (2) in terms of X, Y, λ, and assuming that X and Y are given by power series, we arrive at recurrences for the coefficients of these power series. Next, we verify that these recurrences force $\lambda = \sqrt{3} - 1$ and that they determine all coefficients uniquely. Simple estimates of the coefficients show that the resulting power series converge in some neighborhood of 0. Then the analytic functions $\tilde{X}$ and $\tilde{Y}$ defined by them determine functions $\tilde{f}$ and $\tilde{t}$ on some interval $(-x_0, x_0)$ that satisfy (1) and (2), and hence they equal f and t, respectively. We omit further details. □

We have shown that $f(x)$ and $t(x)$ are given by power series on some neighborhood of 0. Now we are going to extend this neighborhood iteratively.

The next lemma provides functional equations for f and t.

LEMMA 9. *For every $x \in \mathbb{R}$ we have*

$$\begin{aligned} x &= \Phi\Big(t(x), t(t(x)), f(t(x)), f(t(t(x)))\Big) \text{ and} \\ f(x) &= \Psi\Big(t(x), t(t(x)), f(t(x)), f(t(t(x)))\Big), \end{aligned}$$

where Φ and Ψ are the following rational functions:

$$\begin{aligned} \Phi(x_1, x_2, y_1, y_2) &= x_1 + \frac{x_2\left(x_1^2 + (1+y_1)^2\right)}{2Q(x_1, x_2, y_1, y_2)}, \\ \Psi(x_1, x_2, y_1, y_2) &= \frac{2x_1x_2y_1 + (1+y_2)(1+x_1^2-y_1^2)}{2Q(x_1, x_2, y_1, y_2)}, \end{aligned}$$

with $Q(x_1, x_2, y_1, y_2) = (1+y_1)(1+y_2) - x_1x_2$.

Thus if, for some a, we know $t(a)$, $f(a)$, and $f(t(a))$, we can easily calculate $b = t^{-1}(a)$ and $f(b)$, provided that $t^{-1}(a)$ exists (which is equivalent to $|a| < t_{\max}$). This will be one of the main ingredients of the algorithm for evaluating f. The lemma also shows that the inverse function t^{-1} can be expressed using t and f; namely, $t^{-1}(y) = \Phi(y, t(y), f(y), f(t(y)))$. This will be used in the proof of Theorem 2(i).

The proof of the lemma is a simple calculation based on (1) and (2) which we omit.

LEMMA 10. *There exists a constant $\beta < 1$ such that for every $x > 0$ we have $t(x) \le \beta x$.*

The proof is simple and is omitted.

Proof of Theorem 2(i). Suppose that we have already proved that f and t are analytic on $[0, a)$ for some $a > 0$, and let $x_0 \in [a, a/\beta)$, where $\beta < 1$ is as in Lemma 10. On a neighborhood of x_0 we have $x = F(t(x))$, where $F(y) := \Phi(y, t(y), f(y), f(t(y)))$. By the assumption and by Lemma 10, t and f are analytic on a neighborhood of $y_0 = t(x_0)$, as well as on a neighborhood of $t(y_0)$, and Φ is a rational function, and hence F is analytic on a neighborhood of y_0. (One might worry that $F(y)$ might become the indeterminate expression $\frac{0}{0}$ for some y, but the numerator $x_2(x_1^2 + (1+y_1)^2)$ in $\Phi(x_1, x_2, y_1, y_2)$ is obviously nonzero whenever $x_2 > 0$.)

Hence, on a neighborhood of x_0, t is the inverse function to F, and thus analytic. Then $f(x) = G(t(x))$, with $G(y) = \Psi(y, t(y), f(y), f(t(y)))$, is analytic there as well. This proves Theorem 2(i). □

4. EVALUTAION ALGORITHM

Given $z \in \mathbb{R}$, which for notational convenience we always assume to be positive, and $\varepsilon > 0$, we want to compute $f(z)$ with error at most ε (compared to the statement of Theorem 2, we have renamed x to z, so that we can use x as a variable). The idea is as follows.

We choose a sufficiently small $\delta = \delta(z, \varepsilon)$. For $x \le 2\delta$ we can evaluate $t(x)$ and $f(x)$ with high precision using the power series expansions

$$f(x) = \sum_{i=0}^{k} a_i x^i + O\left(x^{k+1}\right), \quad t(x) = \sum_{i=0}^{k} b_i x^i + O\left(x^{k+1}\right).$$

Here k is a suitable constant chosen once and for all. The required a_i and b_i can be computed in polynomial time to any desired precision using the approach of Lemma 8.

Let us suppose, for the moment, that we can evaluate t and f *exactly* on $[0, 2\delta)$.

What do we do if $z > 2\delta$? The idea is to start with a suitable $s \in [\delta, 2\delta)$, compute $f(s)$, $t(s)$, and $f(t(s))$, and "step up" all the way to z using the functional equations from Lemma 9, which, as we recall, allow us to calculate $t^{-1}(x)$ and $f(t^{-1}(x))$ from the knowledge of $f(x)$, $t(x)$, and $f(t(x))$, and both x and $t(x)$ are smaller than $t^{-1}(x)$ (at least by a constant factor $\beta < 1$).

Thus, for a starting point $s \in [\delta, 2\delta)$ we define the sequences $(x_0, x_1, x_2, \ldots)$ and $(y_0, y_1, y_2, \ldots)$, depending on s, by

$$\begin{array}{lcllcl} x_1 &=& s, & & y_1 &=& f(x_1), \\ x_0 &=& t(x_1), & & y_0 &=& f(x_0), \\ x_i &=& \Phi(x_{i-1}, x_{i-2}, y_{i-1}, y_{i-2}), & & & & \\ y_i &=& \Psi(x_{i-1}, x_{i-2}, y_{i-1}, y_{i-2}), & i \ge 2. & & & \end{array}$$

By Lemma 9 and by induction we find that $x_i = t^{-1}(x_{i-1})$ and $y_i = f(x_i)$ *provided* that $x_j < t_{\max} = \sup_{x \in \mathbb{R}} t(x)$ for all $j \le i - 1$.

If we are extremely lucky and pick the starting s so that z appears as one of the terms x_i, we have calculated $f(z) = f(x_i) = y_i$ in this way. But typically we do not hit z with any of the x_i. So we are going to adjust s using a binary search strategy, so that eventually some x_i approaches z sufficiently closely. To describe the binary search, we first introduce some notation.

Let $i_{\text{last}} = i_{\text{last}}(s)$ be the maximum i such that $x_0, x_1, \ldots, x_{i-1} < t_{\max}$, and let $x_{\text{last}} = x_{\text{last}}(s) = x_{i_{\text{last}}}$.

Let us say that s *reaches* a point $\bar{x}$ if there exists $i \le i_{\text{last}}$ with $x_i = \bar{x}$. We thus want to find a starting point s that reaches some point very close to z.

We start the search by setting $s := \delta$, and we compute which points this s reaches.

First, let us assume that there is $i_0 < i_{\text{last}}(\delta)$ with $x_{i_0}(\delta) \le z < x_{i_0+1}(\delta)$. Then we initialize $s_{\text{low}} := \delta$ and $s_{\text{high}} := x_2(\delta) = t^{-1}(\delta) < 2\delta$ (note that $x_{i_0}(s_{\text{high}}) = x_{i_0+1}(s_{\text{low}}) > z$), and we repeatedly halve the current interval $[s_{\text{low}}, s_{\text{high}})$ to find an s with $z - \varepsilon < x_{i_0}(s) \le z$. The invariant in this search is $x_{i_0}(s_{\text{low}}) \le z < x_{i_0}(s_{\text{high}})$.

It remains to deal with the case where $x_{\text{last}}(\delta) < z$. We fix $i_0 = i_{\text{last}}(\delta)$ and we again set $s_{\text{low}} := \delta$ and $s_{\text{high}} := x_2(\delta) = t^{-1}(\delta)$ and search by interval halving. This time the invariant is $i_{\text{last}}(s_{\text{low}}) = i_0$, $x_{\text{last}}(s_{\text{low}}) \le z$, and $i_{\text{last}}(s_{\text{high}}) < i_0$. In each halving step we set $s_{\text{mid}} := (s_{\text{low}} + s_{\text{high}})/2$. If $i_{\text{last}}(s_{\text{mid}}) < i_0$, then we set $s_{\text{high}} := s_{\text{mid}}$ and continue with the next halving. If $i_{\text{last}}(s_{\text{mid}}) = i_0$ and $x_{\text{last}}(s_{\text{mid}}) \le z$, then we set $s_{\text{low}} := s_{\text{mid}}$, and we continue. Finally, if $i_{\text{last}}(s_{\text{mid}}) = i_0$ and $x_{\text{last}}(s_{\text{mid}}) > z$, we set $s_{\text{high}} := s_{\text{mid}}$, and we now have a situation as in the previous paragraph, with $x_{i_0}(s_{\text{low}}) \le z < x_{i_0}(s_{\text{high}})$, and we continue as described there.

The computation of i_{last} and x_{last} involves comparisons of x_i with $t_{\max}$. There is an elegant way of comparing a given number with $t_{\max}$, even though we don't know $t_{\max}$ explicitly. Namely, we have $x < t_{\max}$ if and only if $\frac{1+f(x)}{x} > \frac{t(x)}{1+f(t(x))}$ (proof omitted).

So far we have assumed that f and t can be computed exactly on $(-2\delta, 2\delta)$, which is not the case. One needs to analyze the propagation of the errors in the algorithm, which involves bounding some partial derivatives. This can be done without much calculation; only some rather general properties of Φ and Ψ are used. Similar reasoning also leads to bounding the number of binary search steps by $O(\log \frac{1+|z|}{\varepsilon})$, and the overall running time by $O((\log \frac{1+|z|}{\varepsilon})^2)$. We omit this part here.

5. CONCLUSION

Here we outline possible directions for further work.

One obvious question is a generalization to k equidistant curves separating two points; we haven't touched it at all.

We have shown the existence and uniqueness of the distance trisector curve by elementary geometric arguments. It would be nice to obtain a simpler and more conceptual proof, say based on Banach's theorem on fixed points of a contractive map, or on existence theorems for differential equations.

A possibly quite challenging problem is to find more about the nature of the distance trisector curve. Is it algebraic, can it be expressed by elementary functions, or as a solution to an ordinary differential equation (or even PDE) with coefficients expressible by elementary functions?

As for the algorithm for evaluating $f(x)$, can one eliminate the binary search used in our approach? A related open problem is to find an algorithm with running time linear or near-linear in $\log \frac{1+z}{\varepsilon}$.

Acknowledgment. We would like to thank Christian Blatter for a suggestion that led to a substantial simplification of the proof and to obtaining a stronger result, and to Michael Struwe for a very helpful hint. We also thank Tomáš Kaiser for stimulating discussions.

6. REFERENCES

[1] T. Asano, J. Matoušek, T. Tokuyama. Voronoi Diagrams with Neutral Zone, *in preparation.*

[2] T. Asano and T. Tokuyama. Drawing Equally-Spaced Curves between Two Points, *Proc. Fall Conference on Computational Geometry,* Boston, Massachusetts, November 2004, pages 24–25.

[3] F. Aurenhammer. Voronoi Diagrams – A Survey of a Fundamental Geometric Data Structure, *ACM Computing Surveys* 23,3(1991) 345–405.

[4] *Famous Curves Index*, `http://www-history.mcs.st-andrews.ac.uk/history/Curves/Curves.html`, as of June 2005.

[5] J. Milnor. *Dynamics in one complex variable. Introductory lectures.* Vieweg, Wiesbaden 1999.

[6] M. Rosenlicht. Integration in finite terms. *American Mathematical Monthly* 79(1972), 963–972.

[7] A. Okabe, B. Boots, K. Sugihara. *Spatial Tessellations, Concepts and Applications of Voronoi Diagrams*, John Wiley & Sons, New York, NY 1992.