# The distance trisector curve


Tetsuo Asano*
School of Information Science, JAIST
1-1 Asahidai, Nomi, Ishikawa
923-1292 Japan

Jiří Matoušek**
Department of Applied Mathematics and
Institute of Theoretical Computer Science (ITI)
Charles University
Malostranské nám. 25, 118 00 Praha 1
Czech Republic

Takeshi Tokuyama‡
Graduate School of Information Sciences
Tohoku University
Aramaki Aza Aoba, Aoba-ku, Sendai
980-8579 Japan


Rev. 7/X/2006 JM


**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 $f\colon \mathbb{R} \to \mathbb{R}$, $C_2$ 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 $\varepsilon$ 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 *zone diagram* and which we investigate in a companion paper.



*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.

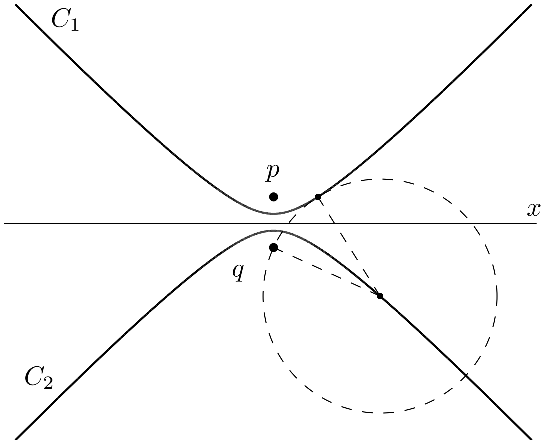


Figure 1: The distance trisector curves

## 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. The bisector (usually called perpendicular bisector) is a fundamental tool in geometry.

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); see Fig. 2. We note that $C_2$ is the bisector of $C_1$ and $C_3$ by symmetry, and that iterating this construction (for $k = 7$, say) doesn't work. 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 $f\colon \mathbb{R} \to \mathbb{R}$ is a convex continuous function.*

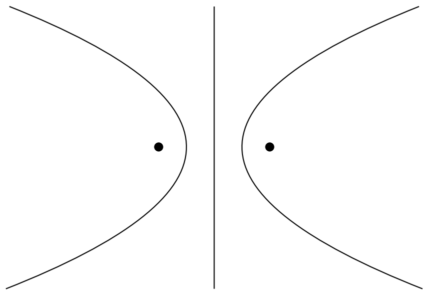

Figure 2: Three equidistant curves separating two points

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 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 (Properties and 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 $\varepsilon$ in time polynomial in $\log \frac{1+|x|}{\varepsilon}$. (The time is measured in the standard Turing machine model of computation; we count the number of bit operations. We assume that $x$ is accessed via an oracle that returns the first $n$ significant digits of $x$ in time polynomial in $n$.)*

Part (i) shows that the trisector curve is “nice” in some sense (which seems intuitively very plausible). The methods of the proof are also used in part (ii). In a nutshell, on a very small neighborhood of 0, we approximate $f$ by a power series truncated to constantly many terms, and we use a functional equation to extend the approximation to the whole of the real axis.

**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.

**Zone diagrams.** 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.

The Voronoi diagram is one of the most popular structures in computational geometry. It is frequently used as a mathematical model to represent a pattern created by a competitive growth process where many bodies grow simultaneously to form a geometric structure together, such as the cell structure of a biological tissue, a crystal-lattice structure, a geographic/geological pattern, an economic/political regional equilibrium, or a gravity/electromagnetic field.

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. However, geometric structures are sometimes observed in the nature in which the union of the cells has a nonempty complement region (called the *neutral zone*). We can regard such a structure as a result of growth process in which the growth terminates before the cell boundaries meet each other, and the termination is due to some non-contact action of other regions. The *zone diagram*, which we investigate in the companion paper [1], is a way of modeling such a structure.

The idea can be explained by a story on equilibrium in the "age of wars". There are $n$ mutually hostile kingdoms. The $i$th kingdom has a castle at a given location $\mathbf{p}_i$ and a territory $R_i$ around it. The $n$ territories are separated by a no-man's land, the neutral zone. If the territory $R_i$ is attacked from another

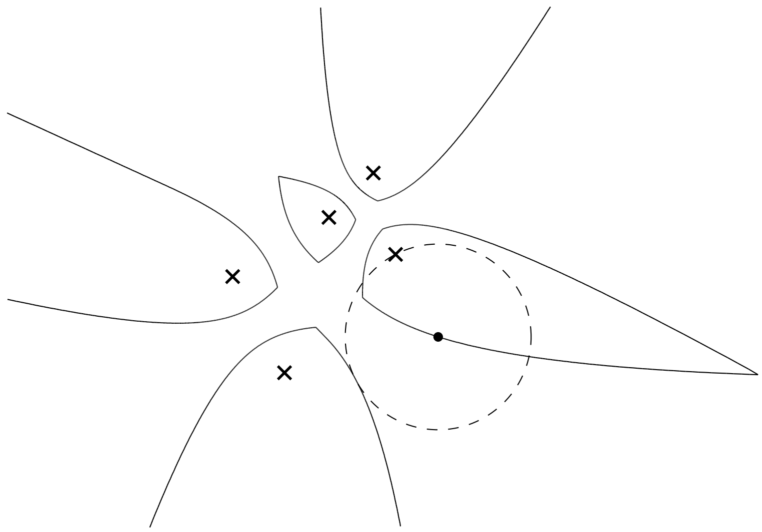

Figure 3: A zone diagram

kingdom, an army departs from the castle $\mathbf{p}_i$ to intercept the attack. The interception succeeds if and only if the defending army arrives at the attacking point on the border of $R_i$ sooner than the enemy. However, the attacker can secretly move his troops inside his territory, and the defense army can start from its castle only when the attacker leaves his territory. The zone diagram is an equilibrium configuration of the territories, such that every kingdom can guard the territory and no kingdom can grow without risk of invasion by other kingdoms. Mathematically speaking, the distance of each point $x$ on the border of the territory $R_i$ to the capital $\mathbf{p}_i$ equals the distance of $x$ to the union of the other $R_j$, $j \neq i$. Fig. 3 is an illustration with five castles (the castles are marked by crosses, and the equidistance property is indicated for one boundary point).

If there are only two castles $\mathbf{p}$ 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 zone diagram 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 zone diagrams.

## 2 Existence and uniqueness

In this section we prove 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.

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

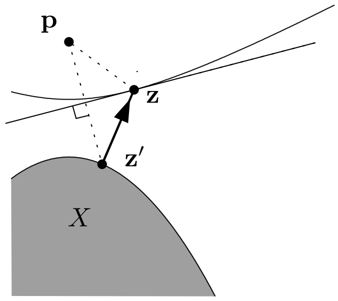


Figure 4: Illustration to Lemma 3(v).

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* $f : \mathbb{R} \to [0, \infty)$*.*

(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. 4.*

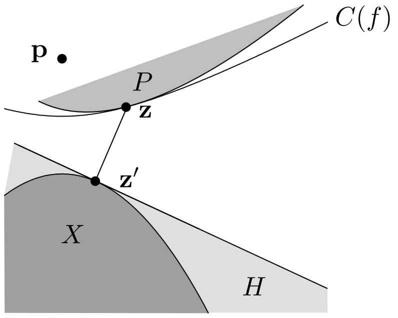


Figure 5: Existence of the derivative of the bisector.

**Proof.** For (i) we note that $\mathrm{dom}(\mathbf{p}, X) = \bigcap_{\mathbf{x}\in X} \mathrm{dom}(\mathbf{p}, \{\mathbf{x}\})$, and the right-hand side is an intersection of (closed) halfplanes.

Part (ii) is clear from the definition.

As for (iii), it is immediate that the boundary of $\mathrm{dom}(\mathbf{a}, X)$ is contained in $\mathrm{bisect}(\mathbf{a}, X)$. Next, let $\mathbf{z} \in \mathrm{bisect}(\mathbf{a}, X) \subseteq \mathrm{dom}(\mathbf{a}, X)$. Let $\mathbf{x}$ be a point of the closure of $X$ nearest to $\mathbf{z}$. Then $\mathrm{dom}(\mathbf{a}, X) \subseteq \mathrm{dom}(\mathbf{a}, \{\mathbf{x}\})$, and the latter is a halfplane (since $\mathbf{a} \neq \mathbf{x}$ by the assumption) having $\mathbf{z}$ on the boundary. Hence $\mathbf{z}$ is on the boundary of $\mathrm{dom}(\mathbf{a}, X)$ as well.

In (iv), by antimonotonicity we have $\mathrm{dom}(\mathbf{p}, X) \subseteq \mathrm{dom}(\mathbf{p}, \{\mathbf{q}\})$, and the latter is the upper halfplane $U$. On the other hand, we have $\mathrm{dom}(\mathbf{p}, X) \supseteq \mathrm{dom}(\mathbf{p}, L)$, and the latter is the region $P$ above the parabola that is the bisector curve of $\mathbf{p}$ and $L$. Hence $P \subseteq \mathrm{dom}(\mathbf{p}, X) \subseteq U$, and since $\mathrm{dom}(\mathbf{p}, X)$ is convex, each vertical line intersects it in a ray directed upwards. This shows that the boundary is a graph of a convex function defined on $\mathbb{R}$.

As for (v), let $\mathbf{z} = (x, f(x)) \in \mathrm{bisect}(\mathbf{p}, X)$. Since $f$ is convex, its graph has at least one supporting line at $\mathbf{z}$, and it suffices to show that the supporting line is unique. Let $\mathbf{z}'$ be a point of $X$ that is nearest to $\mathbf{z}$; see Fig. 5. We have $d(\mathbf{z}, \mathbf{p}) = d(\mathbf{z}, \mathbf{z}')$. The segment $\mathbf{z}\mathbf{z}'$ doesn't intersect $X$ except for at $\mathbf{z}'$, and hence it can be non-strictly separated from the convex set $X$ by a line. That is, there exists a halfplane $H$ having $\mathbf{z}'$ on the boundary, with $\mathbf{z} \notin H$, and with $X \subseteq H$. By antimonotonicity we have $\mathrm{dom}(\mathbf{p}, X) \supseteq \mathrm{dom}(\mathbf{p}, H)$, and the right-hand side is the region inside a parabola passing through $\mathbf{z}$. Hence $\mathrm{dom}(\mathbf{p}, X)$ has at most one supporting line at $\mathbf{z}$, and so does its boundary, which is the graph of $f$.

In (vi), if the line through $\mathbf{z}'$ perpendicular to $\mathbf{z}'\mathbf{z}$ were not a supporting line of $X$, then a small neighborhood of $\mathbf{z}'$ would contain points of $X$ closer to $\mathbf{z}$ than $\mathbf{z}'$. Further, $\mathbf{z}$ lies on the perpendicular bisector of $\mathbf{p}\mathbf{z}'$ since $d(\mathbf{z}, \mathbf{p}) = d(\mathbf{z}, \mathbf{z}')$, and if the tangent at $\mathbf{z}$ were not perpendicular to $\mathbf{p}\mathbf{z}'$, then a small neighborhood of $\mathbf{z}$ would contain a point $\mathbf{x} \in \mathrm{bisect}(\mathbf{p}, X)$ with $d(\mathbf{x}, \mathbf{p}) > d(\mathbf{x}, \mathbf{z}') \geq d(\mathbf{x}, X)$, a contradiction. $\square$

**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:

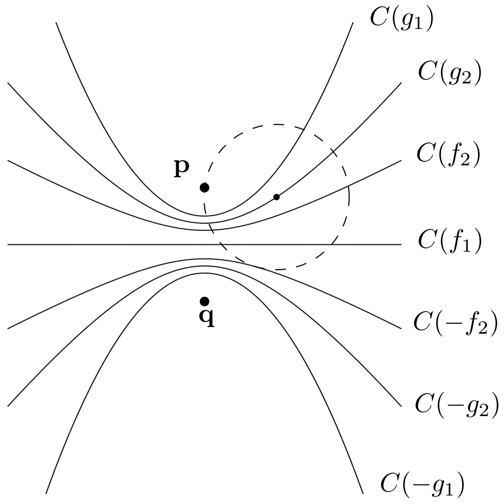


Figure 6: The functions $f_i$ and $g_i$.

(1) $f_1 \equiv 0$,

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

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

So we start with $f_1 \equiv 0$ and iterate the bisector operator, with the $f_i$ being $(2i-1)$st iteration and $g_i$ the $(2i)$th iteration. The first few steps of the construction are illustrated in Fig. 6.

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 $C(f) = \text{bisect}(\mathbf{p}, C(-g))$ and $C(g) = \text{bisect}(\mathbf{p}, C(-f))$.

The following proposition is the technical core of the proof.

**Proposition 4** *$f = g$.*

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

[1] Another very natural proof idea is to define a suitable metric on a suitable space of convex curves such that the operator $C(f) \mapsto \text{bisect}(\mathbf{p}, C(-f))$ is a contraction. Then Banach's theorem would immediately yield existence and uniqueness of a fixed point (which is a distance

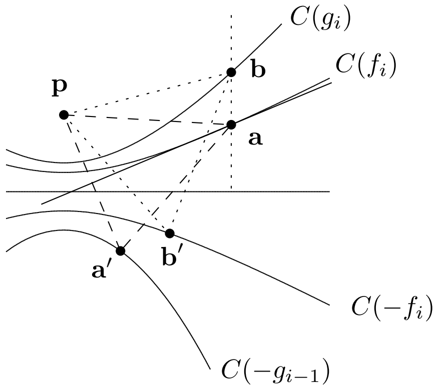


Figure 7: Proving $f_i' \leq g_i'$.

**Properties of the $f_i$ and $g_i$.** From convexity we get that the derivatives $f_i'$ and $g_i'$ are nondecreasing functions. The function $f_1 \equiv 0$ is even, i.e. the graph is symmetric about the $y$-axis, and since the operator bisect$(\mathbf{p}, .)$ preserves this symmetry, all the $f_i$ and $g_i$ are even: $f_i(-x) = f_i(x)$ and $g_i(-x) = g_i(x)$.

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

**Proof.** By induction on $i$ we prove that $f_i' \leq g_i'$ and $f_{i+1}' \leq g_i'$ (here and in the rest of this proof, the inequalities are meant to hold on $[0, \infty)$). Then we will get that $g_i - f_i$ is nondecreasing on $[0, \infty)$, and hence the limit $g - f$ is nondecreasing there, too.

We have $f_1(x) = 0$ and $g_1(x) = (x^2+1)/2$, and so indeed $f_1' \leq g_1'$ on $[0, \infty)$; this is the basis of the induction.

In the induction step, from $f_{i-1}' \leq g_{i-1}'$ we derive $f_i' \leq g_{i-1}'$, and from this we further derive $f_i' \leq g_i'$. We show only the second derivation, since the first one is almost identical.

We thus assume $f_i' \leq g_{i-1}'$, and for an arbitrary $x_0 > 0$ we want to verify $f_i'(x_0) \leq g_i'(x_0)$. Let $\mathbf{a} = (x_0, f_i(x_0))$, and let $\mathbf{a}' = (x_1, -g_{i-1}(x_1))$ be the point of $C(-g_{i-1})$ nearest to $\mathbf{a}$. Similarly, $\mathbf{b} = (x_0, g_i(x))$, and $\mathbf{b}' = (x_2, -f_i(x_2))$ is the point of $C(-f_i)$ nearest to $\mathbf{b}$; see Fig. 7.

By Lemma 3(vi), $\mathbf{a}'\mathbf{a}$ is the outward normal of $C(-g_{i-1})$ at $\mathbf{a}'$, and its slope is thus $1/g_{i-1}'(x_1)$. Hence $g_{i-1}'(x_1)$ has the same sign as $x_0 - x_1$, and this implies $0 \leq x_1 \leq x_0$. Analogously, the slope of $\mathbf{b}'\mathbf{b}$ equals $1/f_i'(x_2)$ and we have $0 \leq x_2 \leq x_0$.

By Lemma 3(vi) again, the tangent to $C(f_i)$ at $\mathbf{a}$ is the perpendicular bisector of $\mathbf{pa}'$, and the tangent of $C(g_i)$ at $\mathbf{b}$ is the perpendicular bisector of $\mathbf{pb}'$

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 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.

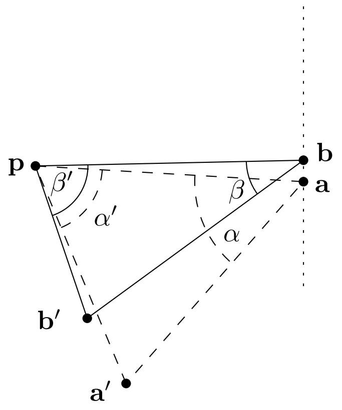


Figure 8: Proving $f_i' \geq g_i'$ continued.

So for checking $f_i'(x_0) \leq g_i'(x_0)$, it suffices to verify that the line $\mathbf{pa}'$ decreases more steeply than the line $\mathbf{pb}'$, or in other words, that $\mathbf{b}'$ is above the line $\mathbf{pa}'$.

We assume the contrary (see Fig. 8) and we are going to derive a contradiction. Assuming $\mathbf{b}'$ below the line $\mathbf{pa}'$, we have $\beta' > \alpha'$, where $\alpha'$ and $\beta'$ are the angles at $\mathbf{p}$. Since the triangles $\mathbf{paa}'$ and $\mathbf{pbb}'$ are isosceles, for the angles at $\mathbf{a}$ and at $\mathbf{b}$ we have $\alpha = \pi - 2\alpha'$ and $\beta = \pi - 2\beta'$. The angular difference between the directions of the segments $\mathbf{pb}$ and $\mathbf{pa}$ is at most $\beta' - \alpha'$, and the difference in angular directions of $\mathbf{b}'\mathbf{b}$ and $\mathbf{a}'\mathbf{a}$ is at most $\beta' - \alpha' + \beta - \alpha = \beta' - \alpha' + (\pi - 2\beta') - (\pi - 2\alpha') = \alpha' - \beta' < 0$. Hence the slope of $\mathbf{b}'\mathbf{b}$ is *smaller* than the slope of $\mathbf{a}'\mathbf{a}$. From this we get that the segment $\mathbf{b}'\mathbf{b}$ intersects the segment $\mathbf{pa}'$, and in particular, $\mathbf{b}'$ lies left of $\mathbf{a}'$; that is, $x_2 < x_1$. By the induction hypothesis $f_i' \leq g_{i-1}'$ and monotonicity of $f_i'$, we then get $f_i'(x_2) \leq f_i'(x_1) \leq g_{i-1}'(x_1)$. This means that the slope of $\mathbf{b}'\mathbf{b}$, which is an outer normal of $C(-f_i)$ at $\mathbf{b}'$, should be *larger* than the slope of $\mathbf{a}'\mathbf{a}$, which is an outer normal of $C(-g_{i-1})$ at $\mathbf{a}'$. But we have derived that the slope of $\mathbf{b}'\mathbf{b}$ is smaller than that of $\mathbf{a}'\mathbf{a}$, and this is the desired contradiction proving the lemma. $\square$

**Proof of Proposition 4.** First let us choose $x_0 > 0$ with $g(x_0) \leq 1$. 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. 9.

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 $f(x_0) < g(x_0) \leq 1$, we get $d(\mathbf{p}, \mathbf{a}) > d(\mathbf{p}, \mathbf{b})$. Now we consider a point $\bar{\mathbf{a}}$ such that $\mathbf{b}'\bar{\mathbf{a}}\mathbf{ab}$ 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 $s = \sup\{x \geq 0 : f(x) = g(x)\} \geq x_0$. 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 $x_1' \leq s$; see Fig. 10. This is possible by a continuity argument, since the outer normal of $C(-f)$ at $\mathbf{z} = (s, -f(s))$ either intersects

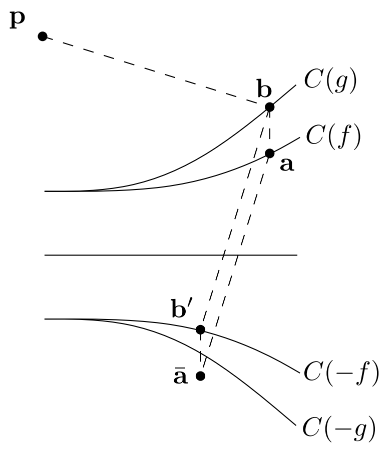


Figure 9: Proving $f(x_0) = g(x_0)$.

$C(g)$ 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 $\mathbf{bb}'$ with $C(f)$. We have $\mathbf{a} \neq \mathbf{b}$ 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. □

# 3 More properties

Let $f\colon \mathbb{R} \to [0, \infty)$ be the convex function whose existence and uniqueness is guaranteed by Theorem 1. Moreover, the proof also implies the following result:

**Proposition 6** *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 $C(f) = V_a \cap \operatorname{bisect}(\mathbf{p}, C(-f))$ (where $C(f) = \{(x, f(x)) : x \in (-a, a)\}$).*

Indeed, it suffices to note that in the proof, knowing $f_i$ on $(-a, a)$ determines $g_i$ on $(-a, a)$, which in turn determines $f_{i+1}$ on $(-a, a)$, and so all arguments can be restricted to $V_a$ instead of the plane.

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 $x \geq 0$ we have $0 \leq t(x) \leq x$, 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 $C(f) = \operatorname{bisect}(\mathbf{p}, C(-f))$, Lemma 3(v) shows that $f'(x)$ exists for every $x \in \mathbb{R}$.

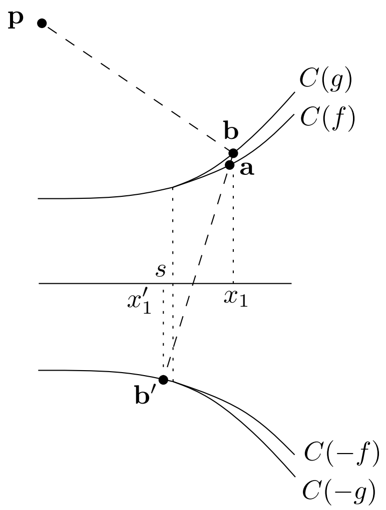


Figure 10: Extending $f = g$ to $\mathbb{R}$.

**Proposition 7** *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 5, show that $t_{\max} \approx 5.648708769021159$ and $\lim_{x\to\infty} f'(x) \approx 1.083629958775032$.

**Proof of Proposition 7.** First we note that $t$ is continuous. Indeed, if $C$ is a closed convex set, then the mapping assigning to a point $\mathbf{x}$ its nearest point in $C$ (the *metric projection*) is well-known to be continuous, and $t(x)$ is the first coordinate of the point of the convex set bounded by $C(-f)$ that is nearest to $(x, f(x))$. Hence the range is an interval.

The injectivity of $t$ follows immediately from Lemma 3(vi), since if $\mathbf{z} = (t(x), -f(t(x))) \in C(-f)$, then the bisector of $\mathbf{zp}$ is a tangent to $C(f)$ at $(x, f(x))$, and thus $x$ is determined uniquely by $t(x)$.

Next, we prove $\sup_{x\in\mathbb{R}} t(x) < \infty$; for contradiction, we suppose the contrary, which means that each point of $C(-f)$ is the nearest point to some point of $C(f)$.

Then no line through $\mathbf{p}$ has two intersections with $C(-f)$, for if it did, at points $\mathbf{z}_1$ and $\mathbf{z}_2$, then both the bisector of $\mathbf{pz}_1$ and the bisector of $\mathbf{pz}_2$ would be tangents of $C(f)$ (Lemma 3(vi)), but $C(f)$ cannot have two distinct parallel tangents.

Now we consider the line $\ell$ through $\mathbf{p}$ with slope $-1$, with equation $y = 1-x$. Let us suppose that $\ell$ intersects the open region below $C(-f)$. Then there is a line $\tilde{\ell}$ through $\mathbf{p}$ with slope $\alpha > -1$ intersecting $C(-f)$, and the bisector of the intersection point $\mathbf{x}$ and of $\mathbf{p}$ is a tangent of $C(f)$ with slope strictly larger than 1. But then $C(-f)$ has a tangent with slope strictly smaller than $-1$, and this implies that $C(-f)$ has to intersect $\ell$ twice, which is a contradiction.

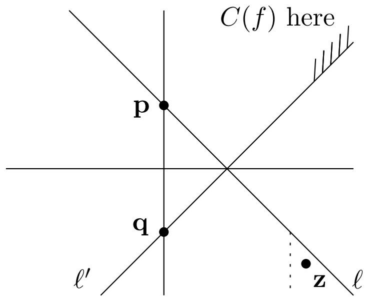


Figure 11: Illustration to the proof of Proposition 7.

Thus, $\ell$ either misses $C(-f)$ or it is tangent to it, and so $C(-f)$ lies in the halfplane below $\ell$. By symmetry, $C(f)$ lies above (non-strictly) the reflection of $\ell$ about the $x$-axis, which we denote by $\ell'$; see Fig. 11. We let $\mathbf{z} \in C(-f)$ be a point with $x$-coordinate larger than 2. Then the bisector of $\mathbf{pz}$ has slope smaller than 1, it avoids the region lying above $\ell'$ and above the $x$-axis, and thus it cannot be tangent to $C(f)$—a contradiction. $\square$

# 4 Power series expansions

## 4.1 Equations

Up until now, we have been arguing geometrically. Now we set up two equations satisfied by $f$ and $t$, and we will work with them mostly analytically.

**Lemma 8** *The following equations are satisfied for every $x \in \mathbb{R}$:*

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

$$t(x) - x + (f(x) + f(t(x))f'(t(x)) = 0, \tag{2}$$

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

**Proof.** The first equation just says that the point $(x, -f(x))$ has 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

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

and this yields (2). $\square$

The system (1), (2) may look like an innocent system of differential equations, but what seems to make it different from differential equations usually encountered in textbooks is the occurrence of $f$, $t$, *and* the composition $f \circ t$.

Next, we check that (1), (2), and some extra conditions determine $f$ and $t$ uniquely.

**Lemma 9** *Let $x_0 > 0$ and suppose that $\tilde{f}\colon(-x_0, x_0) \to (0, \infty)$ and $\tilde{t}\colon(-x_0, x_0) \to (-x_0, x_0)$ are functions such that $\tilde{f}$ is convex, $\tilde{f}'$ exists on $(-x_0, x_0)$, and (1) and (2) are satisfied by $\tilde{f}$ and $\tilde{t}$ for all $x \in (-x_0, x_0)$. Then $\tilde{f} = f$ and $\tilde{t} = t$ on $(-x_0, x_0)$.*

**Proof.** By Proposition 6, it suffices to check that each point of $C(\tilde{f})$ has the same distance to $\mathbf{p}$ and to $C(-\tilde{f})$. Let us fix a point $\mathbf{z} = (x, \tilde{f}(x))$, $x \in (-x_0, x_0)$. By (1), $\mathbf{z}$ has equal distance to $\mathbf{p}$ and to $(\tilde{t}(x), -\tilde{f}(\tilde{t}(x))$. Since $\tilde{f}$ is convex, the distance $d(\mathbf{z}, (y, -\tilde{f}(y))$, considered as a function of $y \in (-x_0, x_0)$, has a unique minimum $y^*$ and it is strictly decreasing on $(-x_0, y^*)$ and strictly increasing on $(y^*, x_0)$. Thus, it may have zero derivative only at $y^*$, and it follows from (2) that $y^* = \tilde{t}(x)$. $\square$

### 4.2 Power series for $f$ and $t$ near the origin

**Lemma 10** *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 $x$.*

**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 the plan of the proof. We do not claim at this moment that $X$, $Y$, and $\lambda$ as above necessarily exist; the existence becomes clear only at the end of the proof. We first investigate what $X, Y, \lambda$ would have to look like if they existed. More precisely, we reformulate equations (1) and (2) in terms of $X, Y, \lambda$, 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, by Lemma 9. We now begin with executing the plan.

The equations (1) and (2) rewritten in terms of $X$ and $Y$ become

$$(X(z) - X(\lambda z))^2 + (Y(z) + Y(\lambda z))^2 - X(z)^2 - (Y(z) - 1)^2 = 0 \tag{3}$$

and

$$\frac{1}{2} \cdot \frac{\partial}{\partial w} \left( (X(w) - X(z))^2 + (Y(w) + Y(z))^2 \right) \Big|_{w=\lambda z}$$

$$= (X(\lambda z) - X(z))X'(\lambda z) + (Y(\lambda z) + Y(z))Y'(\lambda z) = 0 \tag{4}$$

for all $z \in (0, z_0)$.

We write $X$ and $Y$ as power series (for the moment we think of them as formal power series):

$$X(z) = \sum_{i=0}^{\infty} p_i z^i, \quad Y(z) = \sum_{i=0}^{\infty} q_i z^i.$$

We have $X(0) = 0$ and $Y(0) = \frac{1}{3}$, so

$$p_0 = 0, q_0 = \frac{1}{3}.$$

The constant term in (3) is 0, while in (4) it is $\frac{2}{3}q_1$, hence

$$q_1 = 0.$$

Next, we set

$$p_1 = 1.$$

This is not a great loss of generality, since if some functions $X(z)$ and $Y(z)$ satisfy the equations (3), (4) on $(0, z_0)$, then for any scaling factor $\alpha > 0$ the functions $\bar{X}(z) = X(\alpha z)$ and $\bar{Y}(z) = Y(\alpha z)$ satisfy them on $(0, z_0/\alpha)$. Hence by setting $p_1 = 1$ we have excluded only the case $p_1 = 0$, which doesn't lead to a solution anyway.

With the setting so far, the coefficient of $z$ in (3) is 0, while using the coefficient of $z^2$ in (3) and the coefficient of $z$ in (4), we obtain

$$\lambda - 1 + \frac{4}{3}\lambda q_2 = 0, \quad \lambda^2 - 2\lambda + \frac{4}{3}(\lambda^2 + 2)q_2 = 0.$$

These two equations, with $\lambda > 0$, are consistent only for

$$\lambda = \sqrt{3} - 1,$$

in which case they yield

$$q_2 = \frac{3}{8}(\sqrt{3} - 1).$$

Now we prove by induction that all coefficients $p_k$ and $q_k$ are uniquely determined by (3) and (4).

We observe that the lowest power of $z$ in (3) whose coefficient may involve $q_k$ is $z^k$, and in (4) it is $z^{k-1}$. For $p_k$ the lowest power in (4) is $z^{k+1}$ (since $p_0 = 0$), and in (4) it is $z^k$.

Let us suppose that $p_0, \ldots, p_{k-2}$ and $q_0, \ldots, q_{k-1}$ have already been computed in such a way that the coefficients of $z^0, \ldots, z^{k-1}$ in (3) and the coefficients of $z^0, \ldots, z^{k-2}$ in (4), which are fully determined by $p_0, \ldots, p_{k-2}$ and $q_0, \ldots, q_{k-1}$, are all 0. We have done this for $k = 3$.

We prove that there is exactly one choice for $p_{k-1}$ and $q_k$ that makes the coefficient of $z^k$ in (3) and the coefficient of $z^{k-1}$ in (4) equal to 0.

We calculate that the coefficient of $z^k$ in (3) has the form

$$2(\lambda^k - \lambda^{k-1} - \lambda)p_{k-1} + \frac{4}{3}(\lambda^k + 2)q_k + R_k,$$

where $R_k$ depends only on $p_0, \ldots, p_{k-2}$ and $q_0, \ldots, q_{k-1}$. Similarly, for the coefficient of $z^{k-1}$ in (4) we get

$$(k\lambda^{k-1} - (k-1)\lambda^{k-2} - 1)p_{k-1} + \frac{4}{3}k\lambda^{k-1}q_k + S_k,$$

again with $S_k$ depending on $p_0, \ldots, p_{k-2}$ and $q_0, \ldots, q_{k-1}$. We thus have the following system of two linear equations with $p_{k-1}$ and $q_k$ as unknowns:

$$\begin{aligned} A_k p_{k-1} + B_k q_k &= -R_k \\ C_k p_{k-1} + D_k q_k &= -S_k, \end{aligned}$$

with $A_k = 2(\lambda^k - \lambda^{k-1} - \lambda)$, $B_k = \frac{4}{3}(\lambda^k + 2)$, $C_k = (k\lambda^{k-1} - (k-1)\lambda^{k-2} - 1)$, and $D_k = \frac{4}{3}k\lambda^{k-1}$. The determinant $d_k$ of the matrix of this system is

$$d_k = A_k D_k - B_k C_k = \frac{4}{3}\left(2 + (2k-2)\lambda^{k-2} - 2k\lambda^{k-1} - (k-1)\lambda^k - \lambda^{2k-2}\right).$$

It is easy to verify $d_k \neq 0$ for all $k \geq 3$, and hence $p_{k-1}$ and $q_k$ are uniquely determined as claimed.

In particular, we compute $p_2 = q_3 = 0$, which we will soon use.

So far we have shown that (3) and (4) with the initial conditions have a solution in the ring of formal power series. Next, we check that these series have a nonzero radius of convergence. For a sufficiently large constant $M$ we prove by induction that

$$|p_{k-1}| \leq M^{k-3}, |q_k| \leq M^{k-3} \text{ for all } k \geq 3.$$

This holds for $k = 3$ since $p_2 = q_3 = 0$. Let $k \geq 4$. Since the absolute values of the coefficients $A_k, B_k, C_k, D_k$ in the considered linear system are bounded above by a constant, and the determinant $d_k$ is bounded below by a positive constant ($d_4 \approx 1.41532$ is the smallest), Cramer's rule yields $|p_{k-1}|, |q_k| \leq c_1 \cdot (R_k + S_k)$ for a suitable constant $c_1$. It suffices to estimate $R_k + S_k$ by $c_2 M^{k-4}$ for some constant $c_2$ (then $M = c_1 c_2$ will do).

Let us look at $S_k$, for example, and consider the term in $S_k$ that comes from the product $X(z)X'(\lambda z)$ in (4).That is, we are interested in estimating the coefficient of $z^{k-1}$ in $X(z)X'(\lambda z)$, excluding terms that involve $p_{k-1}$, i.e., the expression

$$\sum_{i=2}^{k-2} p_i \cdot (k-i)\lambda^{k-i-1} p_{k-i}.$$

We estimate the absolute value, using the inductive hypothesis $|p_i| \leq M^{i-2}$, by

$$\sum_{i=2}^{k-2} M^{i-2}(k-i)M^{k-i-2}\lambda^{k-i-1} = M^{k-4}\sum_{j=0}^{k-4}(j+2)\lambda^{j+1} = O(M^{k-4}).$$

A similar calculation works for all other terms in $S_k$ and $R_k$; an important fact is that if we expand (3) and (4) into a sum of products like $X(z)X'(\lambda z)$, the one considered above, each of the surviving products has argument $\lambda z$ in at least one of the terms (note that $X(z)^2$ and $Y(z)^2$ cancel out in (3)). This allows us to estimate by $O(M^{k-4})$ for each of the products.

We have proved that (3) and (4) are satisfied by analytic functions on $(0, z_0)$ for some $z_0 > 0$. We note that $X(z) = z + O(z^2)$ maps $(0, z_0)$ bijectively to some interval $(0, x_0)$, provided that $z_0$ is sufficiently small, and hence $\tilde{f}(x) = Y(X^{-1}(x))$ and $\tilde{t}(x) = X(\lambda X^{-1}(x))$ define analytic solutions to (1), (2) in some neighborhood of 0. Calculating the first few coefficients yields

$$\tilde{f}(x) = \frac{1}{3} + \frac{3}{8}(\sqrt{3} - 1)x^2 + O(x^4)$$

and

$$\tilde{t}(x) = (\sqrt{3} - 1)x + O(x^3).$$

The assumptions of Lemma 9 hold in a small neighborhood of 0 (in particular, $\tilde{f}$ is convex because $\tilde{f}''(x) = \frac{3}{4}(\sqrt{3}-1)+O(x^2) > 0$ for a sufficiently small $x$), and so the power series indeed define $f$ and $t$ around 0. Lemma 10 is proved. □

### 4.3 Extending to all of $\mathbb{R}$

We have shown that $f(x)$ and $t(x)$ are analytic 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 11** *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 $\Phi$ and $\Psi$ 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).

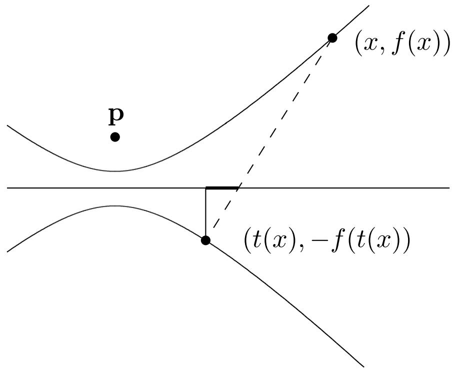


Figure 12: Bounding $t(x)/x$.

**Proof of Lemma 11.** By Lemma 3(vi), the tangent of $C(f)$ at $(x, f(x))$ is the perpendicular bisector of $\mathbf{p}$ and $(t(x), -f(t(x))$, and this yields

$$f'(x) = \frac{t(x)}{1 + f(t(x))}. \tag{5}$$

(This can also be derived directly from (1) and (2), by taking the derivative of (1) with respect to $x$ and solving the resulting equation plus (2) for $f'(x)$. This way, however, we would need to assume the existence of $t'(x)$.)

We substitute for $f'(t(x))$ into (2) using (5). which yields

$$t(x) - x + \frac{t(t(x)) \cdot (f(x) + f(t(x)))}{1 + f(t(t(x)))} = 0.$$

From this equation and equation (1) we can express $x$ and $f(x)$ in terms of $t(x)$, $t(t(x))$, $f(t(x))$, and $f(t(t(x)))$, and we arrive at the statement of the lemma. $\square$

**Lemma 12** *There exists a constant $\beta < 1$ such that for every $x > 0$ we have $t(x) \leq \beta x$.*

**Proof.** For $x \leq x_0$, with some $x_0$ sufficiently small, we have $t(x) \leq 0.9x$ because of the Taylor expansion $t(x) = (\sqrt{3} - 1)x + O(x^3)$ derived in the proof of Lemma 10.

Let us fix such an $x_0$ and let us set $\delta = f'(x_0/2)$; we have $\delta > 0$ (e.g., using the Taylor expansion of $f$, or the fact that $f$ cannot be constant in any neighborhood of 0). We note that $f(x) \geq \delta x$ for all $x > 0$; indeed, for $x \leq x_0$ we can use $f(x) \geq f(0) = \frac{1}{3}$, while for larger $x$ we have $f(x) \geq f(x_0) + f'(x_0)(x - x_0) \geq \frac{1}{3} + \delta(x - x_0) \geq \delta x$ using the monotonicity of $f'$.

Now we let $x \geq x_0$ and we want to bound $x - t(x)$ from below. If $t(x) \leq x_0/2$, then we are done, so we assume $t(x) \geq x_0/2$. We estimate $x - t(x)$ by the length of the thick segment in Fig. 12, which is $f(t(x))f'(t(x)) \geq \delta t(x) f'(t(x)) \geq \delta^2 t(x)$. This leads to $t(x) \leq x/(1 + \delta^2)$. $\square$

**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 12. 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 12, $t$ and $f$ are analytic on a neighborhood of $y_0 = t(x_0)$, as well as on a neighborhood of $t(y_0)$, and $\Phi$ 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). $\square$

## 5 An algorithm for evaluating $f$

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 $\varepsilon$ (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 10 ($k$ is constant, but the time depends on the required precision, which in turn will depend on $z$ and $\varepsilon$).[2]

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

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 11, 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

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

[2]Actually, here we don't need the parameterization as in the proof of Lemma 10, since recurrences for the $a_k$ and $b_k$ can be set up directly using (1),(2) and solved numerically. The advantage of the recurrences for $p_k$ and $q_k$ in the proof of Lemma 10 is that they are simpler and thus make the proof of uniqueness and the estimates manageable.

By Lemma 11 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 $x_{i-1} > t_{\max}$, the recursive formulas also yield some value for $x_i$, but of course, it is no longer $t^{-1}(x_{i-1})$.)

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.

Later, using some of the fact derived in the error analysis, we will estimate the number of halving steps by $O(\log(z/\varepsilon))$. Before we go into the error analysis, we need to discuss another issue. The computation of $i_{\text{last}}$ and $x_{\text{last}}$ involves comparisons of $x_i$ with $t_{\max}$. We will not discuss how $t_{\max}$ can be computed; rather, we will see that there is an elegant way of comparing $x_i$ with $t_{\max}$, even though we don't know $t_{\max}$ explicitly.

**Lemma 13** *Let $x > 0$. Then we have $x < t_{\max}$ if and only if*

$$\frac{1+f(x)}{x} > \frac{t(x)}{1+f(t(x))}. \tag{6}$$

*In other words, $x < t_{\max}$ is equivalent to $Q(x, t(x), f(x), f(t(x))) > 0$, where $Q(x_1, x_2, y_1, y_2) = (1+y_1)(1+y_2) - x_1x_2$ is the denominator in $\Phi$ and $\Psi$.*

**Proof.** Let us consider the vertical ray emanating from the point $\mathbf{q} = (-1, 0)$ and let us start turning it clockwise around $\mathbf{q}$. First it intersects the graph of the smooth strictly convex function $f$ at a single point, then for some time we

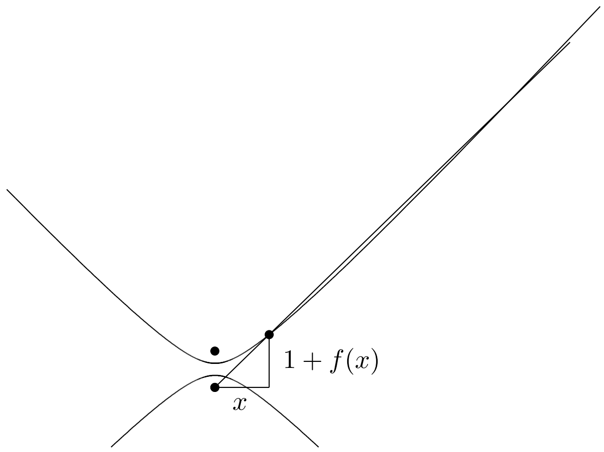


Figure 13: A ray with two intersections.

have two intersections as in Fig. 13, then we reach a ray $\rho_{\max}$ with a single point of tangency, and after that, there are no intersections anymore.

If $\mathbf{z} = (x, f(x))$ is an intersection point of a ray $\rho$ originating from $\mathbf{q}$ with $C(f)$, the slope of $\rho$ is $(1 + f(x))/x$, the left-hand side of (6). If $\mathbf{z}$ is the first intersection of the ray with $C(f)$ (looking from $\mathbf{q}$), then the slope of $\rho$ is greater or equal to $f'(x)$, the slope of the tangent at $x$, while the opposite inequality holds if $\mathbf{z}$ is the second intersection. By (5), we have $f'(x) = t(x)/(1+f(t(x)))$, which is the right-hand side of (6). We can thus see that as we increase $x$ from 0 to $\infty$, the direction of inequality between the left and right-hand sides of (6) changes exactly once, at the point of tangency of the ray $\rho_{\max}$.

It remains to see that this tangency point has the $x$-coordinate equal to $t_{\max}$. As we know, the normal of the graph of $f$ at $t_{\max}$ becomes an asymptote of the graph of $-f$, and hence its slope is $\lim_{x\to\infty} -f'(x)$. Using (5), this equals $\lim_{x\to\infty} -t(x)/(1+f(t(x))) = -t_{\max}/(1+f(t_{\max}))$. At the same time, the normal of the graph of $f$ at $t_{\max}$ has slope $-1/f'(t_{\max}) = -(1+f(t(t_{\max})))/t(t_{\max})$, and so equality in (6) indeed holds for $x = t_{\max}$. □

**Error analysis.** The considerations so far assumed that $x_0(s) = t(s)$, $y_0(s) = f(x_0(s))$, and $y_1(s) = f(s)$ can be evaluated exactly, which is not the case: We can really compute only approximations $\tilde{x}_0$, $\tilde{y}_0$, and $\tilde{y}_1$ to $x_0, y_0, y_1$, respectively, with errors $\varepsilon_0 = |x_0 - \tilde{x}_0|$, $\varepsilon'_0 = |y_0 - \tilde{y}_0|$, $\varepsilon'_1 = |y_1 - \tilde{y}_1|$ bounded by $C_0\delta^{k+1}$, with some (explicit) constant $C_0$. Then, even if we ignore rounding errors, we compute only the approximate quantities $\tilde{x}_i = \Phi(\tilde{x}_{i-1}, \tilde{x}_{i-2}, \tilde{y}_{i-1}, \tilde{y}_{i-2})$, $\tilde{y}_i = \Psi(\tilde{x}_{i-1}, \tilde{x}_{i-2}, \tilde{y}_{i-1}, \tilde{y}_{i-2})$ (where $\tilde{x}_1 = x_1$), and we have the errors

$$\varepsilon_i = |x_i - \tilde{x}_i|,\ \varepsilon'_i = |y_i - \tilde{y}_i|.$$

We also define $\bar{\varepsilon}_i = \max(\varepsilon_i, \varepsilon'_i, \varepsilon_{i-1}, \varepsilon'_{i-1})$.

We want to estimate $\bar{\varepsilon}_i$ using $\bar{\varepsilon}_{i-1}$. If the point $(\tilde{x}_{i-1}, \tilde{x}_{i-2}, \tilde{y}_{i-1}, \tilde{y}_{i-2})$ is at distance at most $\bar{\varepsilon}_{i-1}$ from $(x_{i-1}, x_{i-2}, y_{i-1}, y_{i-2})$ (in the maximum norm), then by the Mean Value Theorem, $|x_i - \tilde{x}_i|$ is bounded by $4\bar{\varepsilon}_{i-1}$ times the maximum of the absolute values of the first partial derivatives of $\Phi$ over points $(\xi_1, \xi_2, \eta_1, \eta_2)$ at distance at most $\bar{\varepsilon}_{i-1}$ from $(x_{i-1}, x_{i-2}, y_{i-1}, y_{i-2})$. Such a maximum is estimated in the following lemma:

**Lemma 14** *There exist constants $A_1, A_2, k_1, k_2$ with the following properties. For a real number $x > 0$, let us set*

$$\mathbf{u}(x) = \Big(t(x), t(t(x)), f(t(x)), f(t(t(x)))\Big) \in \mathbb{R}^4.$$

*If $\mathbf{v} \in \mathbb{R}^4$ is at distance at most $A_1(1+x)^{-k_1}$ from the point $\mathbf{u}(x)$, then all first partial derivatives of $\Phi$ at $\mathbf{v}$ and all first partial derivatives of $\Psi$ at $\mathbf{v}$ are bounded by $A_2(1+x)^{k_2}$ in absolute value. In particular, for $x < t_{\max}$, the partial derivatives are bounded by a constant $A_3$.*

**Proof.** It would not be difficult to analyze the partial derivatives explicitly and get explicit numerical bounds, but we will use a simple general argument instead.

Each partial derivative of either $\Phi(x_1, x_2, y_1, y_2)$ or $\Psi(x_1, x_2, y_1, y_2)$ has the form $P(x_1, x_2, y_1, y_2)/Q(x_1, x_2, y_1, y_2)^2$, where $P$ is some polynomial and

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

is the denominator of $\Phi$ and $\Psi$. Since $P(\mathbf{v})$ is obviously bounded by a polynomial function of $1 + x$ for $\mathbf{v}$ at distance at most $1 + x$ from $\mathbf{u}(x)$, say (using $t(x) \le x$ and $f(x) \le x^2$, for example), it suffices to bound $1/Q(\mathbf{v})$.

We have

$$x = \Phi(\mathbf{u}(x)) = t(x) + \frac{t(t(x))\left(t(x)^2 + (1 + f(t(x)))^2\right)}{Q(\mathbf{u}(x))},$$

and so

$$Q(\mathbf{u}(x)) \ge \frac{t(t(x))\left(t(x)^2 + (1 + f(t(x)))^2\right)}{x} \ge \frac{t(t(x))}{x}.$$

For $x$ small, we have $t(t(x)) \approx (\sqrt{3} - 1)^2 x$, so $t(t(x))/x = \Omega(1)$, while for $x$ not so small, $t(t(x))$ is bounded below by a constant and hence $t(t(x))/x = \Omega(x^{-1})$. Thus $Q(\mathbf{u}(x)) = \Omega((1 + x)^{-1})$ always.

By a simple argument using an obvious boundedness of first partial derivatives of $Q(x_1, x_2, y_1, y_2)$ on a small neighborhood of $\mathbf{u}(x)$, we see that $Q(\mathbf{v})$ remains of order $\Omega((1 + x)^{-1})$ if $\mathbf{v}$ is at most $A_1(1 + x)^{-k_1}$ away from $\mathbf{u}(x)$, for suitable $A_1$ and $k_1$. This implies the lemma. $\square$

Let us return to the considerations before the lemma just proved. We consider $s \in [\delta, 2\delta)$ fixed for a moment, $x_i, y_i$ and $\tilde{x}_i, \tilde{y}_i$ are the terms of the corresponding exact and approximate sequences, respectively, and $\bar{\varepsilon}_i$ is the maximum error in the $i$th terms. Let us put $B = x_{\mathrm{last}}(s)$. We have $\bar{\varepsilon}_1 \le C_0 \delta^{k+1}$ and, assuming that $\bar{\varepsilon}_{i-1} \le A_1(1 + B)^{-k_1}$, Lemma 14 and induction yield $\bar{\varepsilon}_i \le 4A_3 \bar{\varepsilon}_{i-1}$ for $i < i_{\mathrm{last}}(s)$, while

$$\bar{\varepsilon}_{i_{\mathrm{last}}} \le 4A_2(1 + B)^{k_2} \bar{\varepsilon}_{i_{\mathrm{last}}-1} \le 4A_2(1 + B)^{k_2}(4A_3)^{i_{\mathrm{last}}-1} C_0 \delta^{k+1}.$$

Now $x_{i_{\rm last}-1} < t_{\max}$, and since $t^{-1}(x) \ge x/\beta$ by Lemma 12, we have $i_{\rm last} \le 1 + \log_{1/\beta}(t_{\max}/\delta) \le C_2 \log \frac{1}{\delta}$ for some constant $C_2$. Therefore, $(4A_2)^{i_{\rm last}} \le C_3\delta^{-k_3}$ with suitable constants $C_3$ and $k_3$. This finally yields

$$\bar\varepsilon_{i_{\rm last}} \le C_4 \delta^{k+1-k_3}(1+B)^{k_2}.$$

Thus, we can set $k := k_3$, say, and we see that the error $\bar\varepsilon_{i_{\rm last}}$ in the final step can be brought below $\varepsilon$ by choosing $\delta$ as a suitable polynomial function of $\varepsilon$ and $1/(1+B)$.

It is tempting to substitute $B = z$ into the just derived bound and say that the errors of the computation are under control, but there is a potential difficulty. Even though in the final computation we need to reach only to $z$, we may encounter much larger values of $x_{\rm last}$ during the binary search. But fortunately, we don't really care about the exact value of $x_{\rm last}$ once we know that it is much bigger than $z$; the latter information is sufficient to drive the search algorithm. We don't even need to compute $x_{\rm last}$: If its magnitude is enormous, we can detect this, since the denominator $Q$ in the recursive formula must be very close to 0. The error analysis can easily be modified so that it guarantees that the cases $x_{\rm last} < 10z$ and $x_{\rm last} > 100z$, say, are never confused, while the required $\delta$ is bounded polynomially in $\varepsilon$ and $1/(1+z)$.

There is a similar issue in comparing $x_i$ with $t_{\max}$; if they are very close, replacing $x_i$ with $\tilde x_i$ may change the result of the comparison even if $|x_i - \tilde x_i|$ is very small. Here the following strategy works: When $\tilde x_i(s_{\rm low})$ in the algorithm turns out to be extremely close to $t_{\max}$, we pretend that it is actually larger than $t_{\max}$, while for $\tilde x_i(s_{\rm high})$ extremely close to $t_{\max}$ we pretend that it is actually smaller. We omit further details of these considerations.

**Bounding the running time.** In the error analysis above we have shown that each of the computed sequences $(\tilde x_0, \tilde x_1, \ldots)$ and $(\tilde y_0, \tilde y_1, \ldots)$ has $O(\log \frac{1}{\delta}) = O(\log \frac{1+z}{\varepsilon})$ terms. It is also sufficient to make the computations with $O(\log \frac{1+z}{\varepsilon})$-bit numbers.

It remains to estimate the number of steps of the binary search. The error analysis above implies that the preimage of the interval $(z-\varepsilon, z]$ under $t^{-i}$ ($i$-times iterated inverse function to $t$) has length at least

$$\eta := \frac{\varepsilon}{C_5\delta^{-k_3}(1+z)^{k_2}}.$$

Since the binary search starts with an interval of length at most $\delta$, after $\log_2(\delta/\eta) = O(\log \frac{1+z}{\varepsilon})$ halving steps we must hit the interval of length $\eta$ and thus find a suitable starting point $s$.

Altogether the algorithm makes $O((\log \frac{1+z}{\varepsilon})^2)$ arithmetic operations with $O(\log \frac{1+z}{\varepsilon})$-bit numbers. This concludes the proof of Theorem 2(ii). $\square$

# 6 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 $\log \frac{1+z}{\varepsilon}$ (in the RAM model, say).


### 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 consultation and a very helpful hint (suggesting the passage from (1) and (2) to (5)). We also thank Tomáš Kaiser for stimulating discussions. Finally, we thank an anonymous referee for careful reading and, in particular, for pointing out a mistake in the proof of Proposition 7.


## References


[1] T. Asano, J. Matoušek, T. Tokuyama. Zone diagrams: existence, uniqueness, and algorithmic challenge. Submitted. Extended abstract to appear in *Proc. ACM-SIAM Symposium on Discrete Algorithms* 2007.

[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.