What Segments Are the Best in Representing Contours?

Vladik Kreinovich and Chris Quintana
Department of Computer Science
University of Texas at El Paso
El Paso, TX 79968, USA
vladik@utep.edu

Abstract

A pixel-by-pixel (*iconic*) computer representation of an image takes too much memory, demands lots of time to process, and is difficult to translate into terms that are understandable for human users. Therefore, several image compression methods have been proposed and efficiently used. The natural idea is to represent only the contours that separate different regions of the scene. At first these contours were mainly approximated by a sequence of straight-line segments. In this case, for smooth contours a few segments are sufficient to represent the contours; the more curved it is, the more segments we need and the less compression we get as a result. To handle curved edges, circle arcs are now used in addition to straight lines.

Since the choice of segments essentially influences the quality of the compression, it is reasonable to analyze what types of segments are the best. In the present paper we formulate the problem of choosing the segment type as a mathematical optimization problem and solve it under different optimality criteria. As a result, we get a list of segment types that are optimal under these criteria. This list includes both the segments that were empirically proven to be efficient for some problems, and some new types of segments that may be worth trying.

1 Introduction to the Problem

Why is contour segmentation necessary? A pixel-by-pixel (*iconic*) computer representation of a map takes too much memory, demands lots of time to process, and is difficult to translate into terms that are understandable for human users [2, 4, 5]. Therefore, several image compression methods have been proposed and efficiently used. The natural idea is to represent only the contours that separate different regions of the scene.

Storing the contour pixel-by-pixel would also take too much memory, and therefore too much time to process. Therefore, in order to represent a contour it is reasonable to choose a class of standard segments (e.g, straight-line segments), approximate a contour by a sequence of segments, and represent this contour by the parameters that characterize these segments.

Such representations are successfully used in many areas:

- for automated navigation [5];
- in robotic vision; for example, for the analysis of an image that contains industrial parts that have to be assembled (or processed in some way) [2, 4, 10, 11];
- for automated analysis of line drawings in documents, especially the analysis of flow charts [2, 18].

What segments are used now? The simplest approximation for curves are chains of straight line segments, and such approximations have been intensively used and studied in image processing [1, 5, 7, 17, 19].

In this case, a few segments are sufficient to represent the smooth contours; the more curved it is, the more segments we need and the less compression we get as a result. To handle curved edges, it was proposed in [2] and [11] to use circle arcs in addition to straight lines. This addition essentially improved the quality of the compression.

Formulation of the problem. Since the choice of segments essentially influences the quality of the compression, we can gain much by choosing new types of segments. Therefore, it is necessary to analyze what types of segments are the best to choose (in some reasonable sense).

What we are planning to do. In the present paper, we formulate the problem of choosing a segment type as a mathematical optimization problem, and solve it under different optimality criteria.

As a result, we get a list of segment types that are optimal under these criteria. This list includes both the segments that were empirically proven to be efficient for some problems (that is, line segments, circle arcs, etc), and some new types of segments (like hyperbolic curves) that may be worth trying.

2 Motivations of the Proposed Mathematical Definitions

How to represent segments? How can we represent a contour in mathematical terms? From a mathematical viewpoint, a contour is a curve, i.e., when we follow a contour, both the x and y coordinates change. In principle, we can represent a contour by describing how, for example, the y coordinate depends on x, i.e., by a function $y = f(x)$. But in many applications it is important to be able to rotate the image: e.g., in robotic vision, when the robot moves, the image that it sees rotates; in automated navigation it is desirable to have a map in which the current course of the ship is vertical; in this case, when a ship turns, the map has to be rotated. Rotating a map in mathematical terms means that we change our coordinate frame from x, y to the new coordinates x' and y':

$$x' = x\cos(\phi) + y\sin(\phi)$$

$$y' = -x\sin(\phi) + y\cos(\phi)$$

It is difficult to transform the dependency $y(x)$ into the dependency of y' on x'.

Therefore, instead of describing the explicit dependency $f = y(x)$, it is more reasonable to represent a curve by an implicit function, i.e., by an equation $F(x, y) = 0$. This equation describes a curve just as the equation $x^2 + y^2 = 1$ describes a circle in analytical geometry. For such implicit representations it is very easy to change the representation if we make a rotation: it is sufficient to substitute into the expression for $F(x, y)$ the expressions for x and y in terms of x' and y':

$$x = x'\cos(\phi) - y'\sin(\phi)$$

$$y = x'\sin(\phi) + y'\cos(\phi)$$

The segment $F(x, y) = 0$ must fit different curves, therefore we must have an expression for $F(x, y)$ with one or several parameters, so that by adjusting the values of these parameters we would be able to get the best fit for each segment. The more parameters we allow, the better the approximation, but, on the other hand, the more parameters we need. So we must somehow fix the number m of parameters that we will use.

The most commonly used m-parametric families are obtained as follows: we fix m smooth functions $f_i(x, y)$ (their set is called a *basis* because they are the basis for our approximations), and then use arbitrary functions $f(x, y) = \sum_{i=1}^{m} C_i f_i(x, y)$, where C_i are the adjustable parameters. The functions that are obtained for different values of C_i form a *family* that is m-parametric (or m-dimensional) in the sense that in order to choose a function from that family, it is necessary to give the values of m parameters.

We'll consider only functions $f_i(x, y)$ that are maximally smooth, namely, analytical (i.e., can be expanded into Taylor series). Let's give two examples.

Example 1. In particular, if we take $m = 3$, $f_1(x, y) = 1$, $f_2(x, y) = x$ and $f_3(x, y) = y$, we get all the curves of the type $C_1 + C_2x + C_3y = 0$, i.e., *all possible straight lines.*

Example 2. If we take $m = 4$ and in addition to the above functions $f_i(x, y)$ for $i = 1, 2, 3$ take $f_4(x, y) = x^2 + y^2$, we obtain all possible curves of the type $C_1 + C_2x + C_3y + C_4(x^2 + y^2) = 0$. The geometric meaning of this expression is easy to get: If $C_4 = 0$, we get straight lines. If $C_4 \neq 0$, we can divide both sides of the equation $F(x, y) = 0$ by C_4, thus reducing it to the equation $x^2 + y^2 + \tilde{C}_2x + \tilde{C}_3y + \tilde{C}_1$, where $\tilde{C}_i = C_i/C_4$. If we apply to this equation the transformation that helps to solve ordinary quadratic equations (namely, $x^2 + ax + b = (x + (a/2))^2 + (b - (a/2)^2))$, we come to the conclusion that the original equation $F(x, y) = 0$ is equivalent to the following equation: $(x - x_0)^2 + (y - y_0)^2 = Z$, where $x_0 = -1/2\ \tilde{C}_2$, $y_0 = -1/2\ \tilde{C}_3$ and $Z = C_2^2 + C_3^2 - C_1$.

If $Z < 0$, then this equation has no solutions at all, because its left-hand side is always non-negative. If $Z = 0$, the solution is possible only when $x = x_0$ and $y = y_0$, i.e., the curve consist of only one point (x_0, y_0). If $Z > 0$, then by introducing a new variable $R = \sqrt{Z}$ we can reduce the above equation to the equation $(x - x_0)^2 + (y - y_0)^2 = Z$, that describes a circle with a center in the point (x_0, y_0) and radius R.

Summarizing these two possible cases ($C_4 = 0$ and $C_4 \neq 0$), we come to the following conclusion. If we use segments that are described by the curves $\sum_{i=1}^{4} C_i f_i(x, y)$ with the above- described functions $f_i(x, y)$, then *we approximate a curve by straight-line segments and circle arcs.*

Main problem. When the functions $f_i(x, y)$ are chosen, the problem of choosing C_i is easy to solve: there exist many software programs that implement least square methods and find the values C_i that are the best fit for a curve that is normally described by a sequence of points (x_j, y_j) that densely follow each other. However, as we have already mentioned, the quality of this approximation and the quality of the resulting compression essentially depend on the choice of the basis: for some bases the approximation is much better, for some it is much worse. So the problem is: what basis to choose for approximations?

Why is this problem difficult? We want to find a basis that is in some reasonable sense the best. For example, we may look for a basis, for which an average error is the smallest possible, or for which the running time of the corresponding approximation algorithm is the smallest possible, etc. The trouble is that even for the simplest bases we do not know how to compute any of these possible characteristics. How can we find a basis for which some characteristics if optimal if we cannot compute this characteristic even for a single basis? There does not seem to be a likely answer.

However, we will show that this problem is solvable (and give the solution).

The basic idea of our solution is that we consider all possible optimization criteria on the set of all bases, impose some reasonable invariance demands and from them deduce the precise formulas for the optimal basis. This approach has been applied to various problems in [6, 13, 14, 15, 16].

What family is the best? Among all m-dimensional families of functions, we want to choose the best one. In formalizing what "the best" means we follow the general idea outlined in [13] and applied to various areas of computer science (expert systems in [14], neural networks in [15], and genetic algorithms in [16]); there exists also an application to physical chemistry, see [6].

The criteria to choose may be computational simplicity, minimal average approximation error, or something else. In mathematical optimization problems, numeric criteria are most frequently used, where to every family we assign some value expressing its performance, and choose a family for which this value is maximal. However, it is not necessary to restrict ourselves to such numeric criteria only. For example, if we have several different families that have the same average approximation error E, we can choose between them the one for which the average running time T of an approximation algorithm is the smallest. In this case, the actual criterion that we use to compare two families is not numeric, but more complicated: *a family* Φ_1 *is better than the family* Φ_2 *if and only if either* $E(\Phi_1) < E(\Phi_2)$ *or* $E(\Phi_1) = E(\Phi_2)$ *and* $T(\Phi_1) < T(\Phi_2)$. A criterion can be even more complicated. What a criterion *must* do is to allow us for every pair of families to tell whether the first family is better with respect to this criterion (we'll denote it by $\Phi_1 > \Phi_2$), or the second is better ($\Phi_1 < \Phi_2$) or these families have the same quality in the sense of this criterion (we'll denote it by $\Phi_1 \sim \Phi_2$).

The criterion for choosing the best family must be consistent. Of course, it is necessary to demand that these choices be consistent, e.g., if $\Phi_1 > \Phi_2$ and $\Phi_2 > \Phi_3$ then $\Phi_1 > \Phi_3$.

The criterion must be final. Another natural demand is that this criterion must be *final* in the sense that it must choose a *unique* optimal family (i.e., a family that is better with respect to this criterion than any other family). The reason for this demand is very simple. If a criterion does not choose any family at all, then it is of no use. If several different families are "the best" according to this criterion, then we still have a problem choosing the absolute "best". Therefore, we need some additional criterion for that choice. For example, if several families turn out to have the same average approximation error, we can choose among them a family with minimal computational complexity. So what we actually do in this case is abandon that criterion for which there were several "best" families, and consider a new "composite" criterion instead: Φ_1 is better than Φ_2 according to this new criterion if either it was better according to the old criterion or according to the old criterion they had the same quality, and Φ_1 is better than Φ_2 according to the additional criterion. In other words, if a criterion does not allow us to choose a unique best family, it means that this criterion is not ultimate; we have to modify it until we come to a final criterion that will have that property.

The criterion must be reasonably invariant. As we have already mentioned, in many applications it is desirable to be able to rotate the image, i.e., change the coordinates from $\vec{r} = (x, y)$ to the rotated ones $\vec{r}\,' = (x', y') = U(\vec{r})$, where by U we denoted the coordinate transformation induced by rotation (its explicit expression was given before).

Suppose now that we first fixed some coordinates, compared two different bases $f_i(\vec{r})$ and $\tilde{f}_i(\vec{r})$, and it turned out that the basis $f_i(\vec{r})$ is better (or, to be more precise, that the family $\Phi = \left\{\sum\limits_i C_i f_i(\vec{r})\right\}$ is better than the family $\tilde{\Phi} = \left\{\sum\limits_i C_i \tilde{f}_i(\vec{r})\right\}$). This means the following: suppose that we have a family of curves (x^c_j, y^c_j) (where c is an index that characterizes the curve). Then for these curves in some reasonable average sense the quality of approximation by segments $F(x, y) = 0$ with $F \in \Phi$ is better than the quality of approximation by the segments $F(x,y)=0$ with $F\in\tilde{\Phi}$.

It sounds reasonable to expect that the relative quality of two bases should not change if we rotate the axes. After such a rotation a curve that was initially described by its points $\vec{r}_j^{\,c}=(x_j^c,y_j^c)$ will now be characterized by the values $U(\vec{r}_j^{\,c})$. So we expect that when we apply the same approximation algorithms to the rotated data, the results of approximating by segments from Φ will still be better than the results of applying $\tilde{\Phi}$.

Let us now take into consideration that rotation is a symmetry transformation in the sense that the same approximation problem can be viewed from two different viewpoints: Namely, if we describe this approximation problem in the new coordinates (x',y'), then the description is that we approximate the rotated curve $U(\vec{r}_j^{\,c})$ by segments that have the prescribed type in these new coordinates, i.e., they are described by the formula $F(\vec{r}\,')=0$ with F from an appropriate family. But we could as well consider the same problem in the old coordinates (x,y) that can be determined from the new ones by the inverse transformation U^{-1}: $\vec{r}=U^{-1}(\vec{r}\,')$. In this case the curve is described by its old coordinates $\vec{r}_j$; as for the approximating segments, their expression in the old coordinates can be obtained by substituting the expression $U(\vec{r})$ instead of $\vec{r}\,'$ into the expression, that describes this segment in terms of the new coordinates. So this expression is $F(U(\vec{r}))=0$.

Let's now substitute into the equation $F(U(\vec{r}))=0$ the expression for $F(x,y)$ in terms of the basis functions ($F(x,y)=\sum\limits_i C_i f_i(x,y)$, or $F(\vec{r})=\sum\limits_i C_i f_i(\vec{r})$). We can then conclude that from the viewpoint of the old coordinates we are approximating the same curve with segments that are described by the equations $\sum\limits_i C_i f_i(U(\vec{r}))=0$. These are arbitrary linear combinations of the functions $f_i(U(\vec{r}))$. In other words, from the viewpoint of the old coordinate system we are approximating the same curve, but we are using a different basis, consisting of functions $f_i^U(\vec{r})=f_i(U(\vec{r}))$.

Similar to that approximation of a rotated image by functions from $\tilde{\Phi}$ is equivalent to approximating the original (non-rotated image) by the functions from a basis $\{\tilde{f}_i(U(\vec{r}))\}$. So from the demand that the relative quality of the two bases should not change after the rotation, we can conclude that if a basis $\{f_i(\vec{r})\}$ is better than $\{\tilde{f}_i(\vec{r})\}$, then for every rotation U the basis $\{f_i^U(\vec{r})\}$ must be better than $\{\tilde{f}_i^U(\vec{r})\}$, where $f_i^U(\vec{r})=f_i(U(\vec{r}))$ and $\tilde{f}_i^U(\vec{r})=\tilde{f}_i(U(\vec{r}))$.

Another reasonable demand is related to the possibility to change the origin of the coordinate system. This happens in navigation problems or in robotic vision, if we take our position as an origin. Then, whenever we move, the origin changes and therefore all the coordinates change: a point that initially had coordinates $\vec{r}=(x,y)$ will now have new coordinates $\vec{r}\,'=(x-x_0,y-y_0)$, where x_0 and y_0 are the old coordinates of our current position. This transformation is called *translation* and is usually denoted by T.

It is again reasonable to expect that the relative quality of two bases should not change if we change the origin (i.e., apply an arbitrary translation). If we consider this as a demand that the reasonable preference relation must satisfy, then arguments like the ones we used for rotation allow us then to make the following conclusion: if a basis $\{f_i(\vec{r})\}$ is better than $\{\tilde{f}_i(\vec{r})\}$, then for every translation T the basis $\{f_i^T(\vec{r})\}$ must be better than $\{\tilde{f}_i^T(\vec{r})\}$, where $f_i^T(\vec{r}) = f_i(T(\vec{r}))$ and $\tilde{f}_i^T(\vec{r}) = \tilde{f}_i(T(\vec{r}))$.

The last invariance demand is related to the possibility of using different units to measure coordinates. For example, in navigation we can measure them in miles or in kilometers, and thus get different numerical values for the same point on the map. If (x, y) are the numerical values of the coordinates in miles, then the values of the same coordinates in kilometers will be (cx, cy), where $c = 1.6\ldots$ is the ratio of these two units (number of kilometers per mile).

Suppose now that we first used one unit, compared two different bases $f_i(\vec{r})$ and $\tilde{f}_i(\vec{r})$, and it turned out that $f_i(X)$ is better (or, to be more precise, that the family $\Phi = \left\{\sum\limits_i C_i f_i(\vec{r})\right\}$ is better than the family $\tilde{\Phi} = \left\{\sum\limits_i C_i \tilde{f}_i(\vec{r})\right\}$).

It sounds reasonable to expect that the relative quality of the two bases should not depend on what units we used. So we expect that when we apply the same methods, but to the data in which coordinates are expressed in the new units (in which we have $c\vec{r}_j$ instead of $\vec{r}_j$), the results of applying $f_i(\vec{r})$ will still be better than the results of applying $\tilde{f}_i(\vec{r})$. But again we can view this same approximation process from two different viewpoints: if we consider the numeric values expressed in new units (kilometers), then we approximate the values $\vec{r}_j\,' = c\vec{r}_j$ by the segments $\sum C_i f_i(\vec{r}\,') = 0$. But we could as well consider the same approximation problem in old units (in this case, in miles). Then we have a problem of approximating the points $\vec{r}_j$ by the segments, whose equations are $\sum C_i f_i(c\vec{r}) = 0$. This is equivalent to using new basis functions $f_i^c(\vec{r})$, defined as $f_i^c(\vec{r}) = f_i(c\vec{r})$, to the numerical values of coordinates in the old units. So, just like in the two previous cases, we conclude that if a basis $\{f_i(\vec{r})\}$ is better than $\{\tilde{f}_i(\vec{r})\}$, then the basis $\{f_i^c(\vec{r})\}$ must be better than $\{\tilde{f}_i^c(\vec{r})\}$, where $f_i^c(\vec{r}) = f_i(c\vec{r})$ and $\tilde{f}_i^c(\vec{r}) = \tilde{f}_i(c\vec{r})$. This must be true for every c, because we can use not only miles and kilometers, but other units as well.

Now we are ready for the formal definitions.

3 Definitions and the Main Result

Definitions. Assume that an integer m is fixed. By a *basis* we mean a set of m analytical linearly independent functions $f_i(x, y), i = 1, 2, \ldots, m$ (from 2-dimensional space R^2 to the set R of real numbers). By an *m-dimensional family* of functions we mean the set of all functions of the type $f(x, y) = \sum\limits_{i=1}^{m} C_i f_i(x, y)$ for some basis $\{f_i(x, y)\}$, where C_i are arbitrary constants. The set of all m-dimensional families will be denoted by S_m.

Comment. "Linearly independent" means that all these linear combinations $\sum\limits_i C_i f_i(x, y)$ are different. If the functions $f_i(x, y)$ are not linearly independent, then one of them can be expressed as a linear combination of the others, and so the set of all their linear combinations can be obtained by using a subset of $< m$ functions. From the well known algebraic fact that every linear space has a basis, we conclude that for any set of functions $f_i(x, y)$, the set of all linear combinations $\sum_i C_i f_i(x, y)$ either forms an m-dimensional family, or it forms a l-dimensional family for some $l < m$.

A pair of relations $(<, \sim)$ on a set is called *consistent* if it satisfies the following conditions:

(1) if $a < b$ and $b < c$ then $a < c$;

(2) $a \sim a$;

(3) if $a \sim b$ then $b \sim a$;

(4) if $a \sim b$ and $b \sim c$ then $a \sim c$;

(5) if $a < b$ and $b \sim c$ then $a < c$;

(6) if $a \sim b$ and $b < c$ then $a < c$;

(7) if $a < b$ then $b < a$ or $a \sim b$ are impossible.

Assume that a set A is given. Its elements will be called *alternatives*. By an *optimality criterion* we mean a consistent pair $(<, \sim)$ of relations on the set A of all alternatives. If $a > b$, we say that a is *better* than b; if $a \sim b$, we say that the alternatives a and b are *equivalent* with respect to this criterion. We say that an alternative a is *optimal* (or *best*) with respect to a criterion $(<, \sim)$ if for every other alternative b either $b < a$ or $a \sim b$.

We say that a criterion is *final* if there exists an optimal alternative, and this optimal alternative is unique.

Comment. In the present paper we consider optimality criteria on the set S_m of all m-dimensional families.

Assume that a rotation U is given. By the *result of applying* U to a function $f(\vec{r})$ we mean a function $\tilde{f}(\vec{r}) = f(U(\vec{r}))$. By the *result of applying* U to a family Φ we mean the set of the functions that are obtained from $f \in \Phi$ by applying U. In other words, if a family Φ is obtained from a basis $f_i(\vec{r})$, then the result of applying U to Φ corresponds to a basis $f_i(U(\vec{r}))$. This result will be denoted by $U(\Phi)$. We say that an optimality criterion on S_m is *rotation-invariant* if for every two families Φ and $\tilde{\Phi}$ and for every rotation U, the following two conditions are true:

i) if Φ is better than $\tilde{\Phi}$ in the sense of this criterion (i.e., $\tilde{\Phi} < \Phi$), then

$$U(\tilde{\Phi}) < U(\Phi).$$

ii) if Φ is equivalent to $\tilde{\Phi}$ in the sense of this criterion (i.e., $\Phi \sim \tilde{\Phi}$), then

$$U(\Phi) \sim U(\tilde{\Phi}).$$

Assume that a translation T is given (i.e., a vector $\vec{r}_0$). By the *result of applying* T to a function $f(\vec{r})$ we mean a function $\tilde{f}(\vec{r}) = f(T(\vec{r}))$, where $T(\vec{r}) = \vec{r}+\vec{r}_0$. By the *result of applying* T to a family Φ we mean the set of the functions that are obtained from $f \in \Phi$ by applying T. In other words, if a family Φ is obtained from a basis $f_i(\vec{r})$, then the result of applying T to Φ corresponds to a basis $f_i(T(\vec{r}))$. This result will be denoted by $T(\Phi)$. We say that an optimality criterion on S_m is *translation-invariant* if for every two families Φ and $\tilde{\Phi}$ and for every translation T, the following two conditions are true:

i′) if Φ is better than $\tilde{\Phi}$ in the sense of this criterion (i.e., $\tilde{\Phi} < \Phi$), then

$$T(\tilde{\Phi}) < T(\Phi).$$

ii′) if Φ is equivalent to $\tilde{\Phi}$ in the sense of this criterion (i.e., $\Phi \sim \tilde{\Phi}$), then

$$T(\Phi) \sim T(\tilde{\Phi}).$$

By a *result of a unit change* in a function $f(\vec{r})$ to a unit that is $c > 0$ times smaller we mean a function $\tilde{f}(\vec{r}) = f(c\vec{r})$. By the *result of a unit change* in a family Φ by $c > 0$ we mean the set of all the functions, that are obtained by this unit change from $f \in \Phi$. This result will be denoted by $c\Phi$. So if Φ was generated by the basis $f_i(\vec{r})$, the family $c\Phi$ is generated by the basis $f_i(c\vec{r})$. We say that an optimality criterion on S_m is *unit-invariant* if for every two families Φ and $\tilde{\Phi}$ and for every number $c > 0$ the following two conditions are true:

i″) if Φ is better than $\tilde{\Phi}$ in the sense of this criterion (i.e., $\tilde{\Phi} < \Phi$), then

$$c\tilde{\Phi} < c\Phi.$$

ii″) if Φ is equivalent to $\tilde{\Phi}$ in the sense of this criterion (i.e., $\Phi \sim \tilde{\Phi}$), then

$$c\Phi \sim c\tilde{\Phi}.$$

Comment. As we have already remarked, the demands that the optimality criterion is final, rotation-, translation- and unit-invariant are quite reasonable. The only problem with them is that at first glance they may seem rather weak. However, they are not, as the following Theorem shows:

Theorem 1. *If an m-dimensional family Φ is optimal in the sense of some optimality criterion that is final, rotation-, translation- and unit-invariant, then all its elements are polynomials.*

(The proofs are given in Section 4).

Comment. Curves that are described by the equation $F(x, y) = 0$ are called *algebraic curves* in mathematics, so we can reformulate Theorem 1 by saying that every optimal approximation family consists only of algebraic curves.

For small m we can explicitly enumerate the optimal curves. In order to do it let's introduce some definitions.

Definitions. If a function $F(x,y)$ belongs to a family Φ, we say that a curve $F(x,y)=0$ *belongs* to a family Φ; we'll also say that a family Φ *consists* of all the curves $F(x,y)=0$ for all functions $F(x,y)$ from Φ. By a *hyperbola* we mean a curve $y=C/x$ and all the results of its rotation, translation or unit change.

Theorem 2. *Suppose that an m-dimensional family Φ is optimal in the sense of some optimality criterion that is final, rotation-, translation- and unit-invariant. Then:*

- *for $m=1$, Φ consists of all constant functions (so the equation $F(x,y)=0$ does not define any approximation curves);*
- *for $m=2$, there is no such criterion;*
- *for $m=3$, the optimal family consists of all straight lines;*
- *for $m=4$, the optimal family consists of all straight lines and circles;*
- *for $m=5$, the optimal family consists of all straight lines and hyperbolas.*

Comments.

1) This result explains why straight-line approximation is the most often used, and why approximation by straight line segments and circle arcs turned out to be sufficiently good [2, 11].

2) This result also shows that the next family of approximation curves that is worth trying is by segments of hyperbolas. Preliminary experiments with hyperbolas were performed in [9]: he approximated the isolines (lines of equal depth) of the Pacific Ocean and showed that hyperbolas were a reasonably good approximation: namely, when using them one needs 2–3 times less parameters than by using straight line segments and circle arcs. So we get a 2–3 times compression.

3) Theorems 1 and 2 subsume the (weaker) results that appeared in [8] and [12].

4 Proofs

1. Let us first prove that the optimal family Φ_{opt} exists and is *rotation-invariant* in the sense that $\Phi_{opt}=U(\Phi_{opt})$ for an arbitrary rotation U.

Indeed, we assumed that the optimality criterion is final, therefore there exists a unique optimal family Φ_{opt}. Let's now prove that this optimal family is rotation-invariant. The fact that Φ_{opt} is optimal means that for every other Φ, either $\Phi<\Phi_{opt}$ or $\Phi_{opt}\sim\Phi$. If $\Phi_{opt}\sim\Phi$ for some $\Phi\neq\Phi_{opt}$, then from the definition of the optimality criterion we can easily deduce that Φ is also optimal, which contradicts the fact that there is only one optimal family. So for every Φ either $\Phi < \Phi_{opt}$ or $\Phi_{opt} = \Phi$.

Take an arbitrary rotation U and let $\Phi = U(\Phi_{opt})$. If $\Phi = U(\Phi_{opt}) < \Phi_{opt}$, then from the invariance of the optimality criterion (condition *ii*)) we conclude that

$$U^{-1}(U(\Phi_{opt})) = \Phi_{opt} < U^{-1}(\Phi_{opt}),$$

and that conclusion contradicts the choice of Φ_{opt} as the optimal family. So the inequality

$$\Phi = U(\Phi_{opt}) < \Phi_{opt}$$

is impossible, and therefore $\Phi_{opt} = \Phi = U(\Phi_{opt})$, i.e., the optimal family is really rotation-invariant.

2. Similar arguments show that the optimal family is translation-invariant and unit-invariant, i.e., $\Phi_{opt} = T(\Phi_{opt})$ for any translation T and $\Phi_{opt} = c\Phi_{opt}$ for any $c > 0$.

3. Let's use these invariances to prove that all the functions $F(x,y)$ from the optimal family are polynomials. We supposed that all the functions $F(x,y)$ from Φ are analytical, so they can be expanded into a Taylor series:

$$F(x,y) = a + bx + cy + dx^2 + exy + fy^2 + \ldots$$

Every term in this expansion is of the type $cx^k y^l$, where c is a real number and k and l are non-negative integers. The sum $k + l$ is called an *order* of this term (so constant terms are of order 0, linear terms are of order 1, quadratic terms are of order 2, etc). For every function $F(x,y)$ by $F^{(k)}(x,y)$ we'll denote the sum of all terms of order k in its expansion. Then

$$F(x,y) = F^{(0)}(x,y) + F^{(1)}(x,y) + F^{(2)}(x,y) + \ldots,$$

where $F^{(0)}(x,y) = a$, $F^{(1)}(x,y) = bx + cy$, $F^{(2)}(x,y) = dx^2 + exy + fy^2$, etc.

Some terms in this equation can be equal to 0. For example, for a linear function

$$F^{(0)}(x,y) = 0,$$

for quadratic functions $F^{(0)}(x,y) = F^{(1)}(x,y) = 0$ and the first non-zero term is $F^{(2)}(x,y)$, etc. The first term $F^{(k)}(x,y)$ in the above expansion that is different from 0 is called the *main term* of this function.

Comment. The reason why such a definition is often used is that when $x, y \to 0$, then these terms are really the main terms in the sense that all the others are asymptotically smaller.

4. Let us now use the unit-invariance of the optimal family Φ_{opt} to prove that if a function $F(x,y)$ belongs to it, then so does its main term $F^{(k)}(x,y)$. Indeed, due to unit invariance for every $c > 0$ the function $F(c\vec{r}) = F(cx, cy)$ also belongs to Φ_{opt}. Since any family is a set of all linear combinations of some functions $f_i(x,y)$, it is a linear space, i.e., a linear combination of any

two functions from this family also belongs to this family. In particular, since $F(cx, cy)$ belong to Φ_{opt}, we conclude that the function $c^{-k}F(cx, cy)$ belong to Φ_{opt}.

Since the main term in $F(x, y)$ has order k, the expansion for $F(x, y)$ starts with k-th term:

$$F(x, y) = F^{(k)}(x, y) + F^{(k+1)}(x, y) + F^{(k+2)}(x, y) + \ldots$$

Substituting cx and cy instead of x and y, we conclude that

$$F(cx, cy) = F^{(k)}(cx, cy) + F^{(k+1)}(cx, cy) + F^{(k+2)}(cx, cy) + \ldots$$

Every term of order p is a sum of several monomials of order p, i.e., monomials of the type $x^q y^{p-q}$. If we substitute cx and cy instead of x and y into each of these monomials, we get $c^q x^q c^{p-q} y^{p-q}$. If we combine together the powers of c, we conclude that this monomial turns into $c^p x^q y^{p-q}$. So the result of substituting cx and cy into each monomial of $F^{(p)}(x, y)$ is equivalent to multiplying this monomial by c^p. Therefore, the same is true for the whole term, i.e., $F^{(p)}(cx, cy) = c^p F^{(p)}(x, y)$. Substituting this formula into the above expression for $F(cx, cy)$, we conclude that

$$F(cx, cy) = c^k F^{(k)}(x, y) + c^{k+1} F^{(k+1)}(x, y) + c^{k+2} F^{(k+2)}(x, y) + \ldots$$

After multiplying both sides of this equation by c^{-k} we conclude that

$$c^{-k} F(cx, cy) = F^{(k)}(x, y) + cF^{(k+1)}(x, y) + c^2 F^{(k+2)}(x, y) + \ldots$$

We have already proved that the left-hand side of this equality belongs to Φ_{opt}.

The linear space Φ_{opt} is finite-dimensional, and therefore it is closed (i.e., it contains a limit of every sequence of elements from it). If we are tending c to 0, then all the terms in the right-hand side converge to 0, except the first one $F^{(k)}(x, y)$. So the function $F^{(k)}(x, y)$ equals to a limit of the functions $c^{-k}F(cx, cy)$ from Φ_{opt}, and therefore this function also belongs to the optimal family Φ_{opt}. Thus we proved that the main term of any function from an optimal family belongs to this same optimal family.

5. Let us now prove that if a function $F(x, y)$ belongs to an optimal family, then all its terms $F^{(p)}(x, y)$ (and not only its main term) belong to this same family.

We will prove this consequently (i.e., by induction) for $p = k, k+1, k+2, \ldots$. For k it is already true. Suppose now that we have already proved that $F^{(p)}(x, y) \in \Phi_{opt}$ for $p = k, k+1, k+2, \ldots, q$ and let's prove this statement for $p = q+1$. If $F^{(q+1)}(x, y)$ is identically 0, then this is trivially true, because Φ_{opt} is a linear space, and therefore it has to contain 0. So it is sufficient to consider the case when $F^{(q+1)}(x, y) \neq 0$.

In this case, by definition of the terms $F^{(p)}(x, y)$ we have the following expansion:

$$F(x, y) = F^{(k)}(x, y) + F^{(k+1)}(x, y) + \ldots + F^{(q)}(x, y) + F^{(q+1)}(x, y) + \ldots$$

We assumed that $F(x,y)$ belongs to Φ_{opt}, and we proved that $F^{(p)}(x,y)$ belongs to Φ_{opt} for $p = k, k+1, \ldots, q$. If we move all the terms, which we have already proven belong to Φ_{opt}, to the left-hand side, we get the following equation:

$$F(x,y) - F^{(k)}(x,y) - F^{(k+1)}(x,y) - \ldots - F^{(q)}(x,y) = F^{(q+1)}(x,y) + \ldots$$

Since we consider the case when $F^{(q+1)}(x,y) \neq 0$, this means that $F^{(q+1)}(x,y)$ is the main term of the function in the left-hand side. But Φ_{opt} is a linear space, therefore it contains a linear combination of any functions from it. In particular, it contains a function

$$F(x,y) - F^{(k)}(x,y) - F^{(k+1)}(x,y) - \ldots - F^{(q)}(x,y).$$

So according to statement 4. of this proof the main term of this function also belongs to Φ_{opt}. But we have already mentioned that this main term is $F^{(q+1)}(x,y)$. So $F^{(q+1)}(x,y)$ also belongs to Φ_{opt}. The inductive step is proven, and so is this statement 5.

6. Now we are ready to prove Theorem 1, i.e., to prove that all functions from the optimal family Φ_{opt} are polynomials. Indeed, suppose that some function $F(x,y)$ from this family is not polynomial. This means that its Taylor expansion has infinitely many terms, and, therefore infinitely many terms $F^{(k)}(x,y)$ in its expansion are different from 0. According to 5. all these terms belong to Φ_{opt}. But since they are all polynomials of different orders, they are all linearly independent. So Φ_{opt} contains infinitely many linearly independent functions, which contradicts to our assumption that Φ_{opt} is finite-dimensional, and therefore it can contain at most m linearly independent functions. This contradiction proves that a Taylor expansion of $F(x,y)$ cannot contain infinitely many terms, so it contains only finitely many terms and is therefore a polynomial. Theorem 1 is proved.

7. In order to prove Theorem 2, let us first prove that if $F(x,y)$ belongs to Φ_{opt}, then so do its partial derivatives $\partial F(x,y)/\partial x$ and $\partial F(x,y)/\partial y$.

Indeed, since Φ_{opt} is translation-invariant, it contains the function $F(x+a,y)$ that is obtained from $F(x,y)$ by a translation by $\vec{r}_0 = (a,0)$. Since Φ_{opt} is a linear space and it contains both $F(x,y)$ and $F(x+a,y)$, it must also contain its difference $F(x+a,y) - F(x,y)$ and, moreover, the linear combination $(F(x+a,y) - F(x,y))/a$. We have already mentioned that since Φ_{opt} is finite-dimensional, it is closed, i.e., contains the limit of any convergent sequence of its elements. The sequence $(F(x+a,y) - F(x,y))/a$ evidently converges: to the partial derivative $\partial F(x,y)/\partial x$, so we conclude that the partial derivative also belongs to Φ_{opt}.

For the partial derivative with respect to y, the proof is similar.

8. Let's now prove that if Φ_{opt} contains at least one term of order $k > 0$, then it must contain non-zero terms of orders $k-1, k-2, \ldots, 0$.

It is sufficient to show that it contains a non-zero term of order $k-1$, all the other cases can be then proved by mathematical induction. Indeed, suppose that $F(x,y)$ is of order k. Then either $\partial F(x,y)/\partial x$ or $\partial F(x,y)/\partial y$ are different from 0 (because else F is identically constant and cannot therefore be of order $k>0$). In both cases this partial derivative is of order $k-1$ and according to 7. it belongs to Φ_{opt}. The statement is proven.

9. From 8. we conclude that an optimal m-dimensional family can contain at most terms of order $m-1$. Indeed, with any term of order k it contains $k+1$ linearly independent terms of different orders $k, k-1, k-2, \ldots, 2, 1, 0$, and in an m-dimensional space there can be at most m linearly independent elements. So $k+1 \leq m$, hence $k \leq m-1$.

10. For $m=1$ we can already conclude that Φ_{opt} contains only a constant: in this case this inequality turns into $k \leq 0$. So all the functions $F(x,y)$ can contain only terms of order 0, i.e., they are all constants.

11. Let us now consider the case $2 \leq m \leq 3$. According to 9., it must contain only terms of order ≤ 2. If it contains only terms of order 0, then all the functions $F(x,y)$ from this family are constants, and therefore the dimension m of this family is $m=1$. So since $m \geq 2$, the family Φ_{opt} must contain at least one function of order >0. If it contains a function of order 2, then according to 8. it must also contain a linear function. So in all the cases Φ_{opt} contains a linear function $F(x,y)=bx+cy$. Applying the same statement 8. once again, we conclude that Φ_{opt} must also contain a 0-th order function (i.e., a constant).

Now we can use rotation-invariance of the optimal family, i.e., the fact that the optimal family contains a function $F(U(x,y))$ together with any $F(x,y)$ for an arbitrary rotation U. In particular, if we apply a rotation U on 90^o (for which $U(x,y)=(y,-x)$) to a function $F(x,y)=bx+cy$, we conclude that $F(U(x,y))=by-cx$ also belongs to Φ_{opt}.

Let's show that these two functions $bx+cy$ and $by-cx$ are not linearly independent. Indeed, if $c=0$ or $b=0$, this is evident. Let us consider the case when $b \neq 0$ and $c \neq 0$. Then if they are linearly independent, this means that $bx+cy=\lambda(by-cx)$ for some λ and for all x, y. If two linear functions coincide, then its coefficients coincide, i.e., $b=-\lambda c$ and $c=\lambda b$, and, multiplying these equalities, that $bc=-\lambda^2 bc$. Since $b \neq 0$ and $c \neq 0$, we conclude that $\lambda^2=-1$, but for real numbers λ it is impossible.

So these two functions are linearly independent elements of the 2-dimensional space of all linear function $\{C_1x+C_2y\}$. So their linear combinations form a 2-dimensional subspace of a 2-dimensional space, and therefore coincide with the whole space. So the optimal family Φ_{opt} contains all linear functions.

We have already shown that it contains all constants, so its dimension is at least 3. Therefore $m=2$ is impossible. If $m=3$, then we have already 3 linearly independent elements in our space: 1, x and y. Every three linearly independent elements of a 3-dimensional space form its basis, so all the elements $F(x,y)$ of Φ_{opt} are arbitrary linear functions $a+bx+cy$. Therefore all the curves of the type $F(x,y)=0$ are straight lines.

12. Now let us consider the case $m=4$. In this case, Φ_{opt} cannot contain only linear functions, because in this case its dimension will be equal to 3. So

it must contain at least one function of order ≥ 2. If it contains a function of order 3 or more, then it must contain also a function of order 2, of order 1, etc. From the fact that it contains a function of order 1 we conclude (as in 11.) that it contains all linear functions. So in this case we have at least five linearly independent functions: 1, x, y, a function of order 2, and a function of order 3. But the whole dimension of the space Φ_{opt} equals to 4, so there cannot be 5 linearly independent functions in it. This contradiction proves that Φ_{opt} cannot (for $m = 4$) contain any function of order > 2.

So it must contain a function of order 2. Any term of order 2 can be represented as

$$F(x, y) = dx^2 + exy + fy^2$$

for some real numbers d, e, f, i.e., as a quadratic form. It is known in linear algebra that by an appropriate rotation U (namely, a rotation to a basis consisting of eigen vectors) any quadratic form can be represented in a diagonal form $F(x, y) = C_1(x')^2 + C_2(y')^2$, where $(x', y') = U(x, y)$. Since the optimal family is rotation-invariant, we can conclude that a function $\tilde{F}(x, y) = F(U^{-1}(x, y)) = C_1x^2 + C_2y^2$ also belongs to Φ_{opt}.

Let us show that $C_1 = C_2$. Indeed, if $C_1 \neq C_2$ and $C_1 \neq -C_2$, then by applying to this function rotation-invariance with the same rotation $\tilde{U}(x, y) = (y, -x)$ as in statement 11, we conclude that Φ_{opt} must contain a function $\tilde{F}(\tilde{U}(x, y)) = C_2x^2 + C_1y^2$, that is linearly independent with $\tilde{F}(x, y)$. So we get at least 5 linearly independent functions in Φ_{opt}: 1, x, y and two of these. It is impossible because the total dimension m of Φ_{opt} is 4.

If $C_2 = -C_1$, then since Φ_{opt} is a linear space, it must contain together with $\tilde{F}(x, y)$ a function $\tilde{F}(x, y)/C_1 = x^2 - y^2$. Applying a 45^o rotation U, we conclude that Φ_{opt} must also contain a function xy. So again we have two linearly independent quadratic functions: $x^2 - y^2$ and xy, and the total dimension is at least 5, which contradicts to $m = 4$.

So in all the cases except $C_1 = C_2$ we get a contradiction. Therefore $C_1 = C_2$, and we conclude that Φ_{opt} contains a function $C_1(x^2 + y^2)$. Since Φ_{opt} is a linear space, it must also contain a function that is obtained from that one by dividing by C_1, i.e., $x^2 + y^2$.

So a 4-dimensional space Φ_{opt} contains 4 linearly independent functions $1, x, y, x^2 + y^2$, hence it coincides with the set of their linear combinations. We have already shown that the correspondent curves are straight lines and circles. For $m = 4$, the statement is proved.

13. Let us finally consider the case $m = 5$. Let us first show that in this case Φ_{opt} also cannot contain terms of order 3 or more. Indeed, in this case out of 5 possible linearly independent functions at least one must be of order > 2, so the set of all polynomials of order ≤ 2 is of dimension ≤ 4. We have already proven (in 12.) that in this case, all second order terms are proportional to $x^2 + y^2$. If there is a term of order 3 or more, there must be a non-trivial term of order 3 (according to 8.). Its partial derivatives must also belong to Φ_{opt} (due to 7.)

and, since they are of second order, they must be proportional to $x^2 + y^2$. So

$$\frac{\partial F(x,y)}{\partial x} = a(x^2 + y^2)$$

and

$$\frac{\partial F(x,y)}{\partial y} = b(x^2 + y^2)$$

for some constants a and b that are not both equal to 0 (because else $F(x,y)$ would be a constant). If we differentiate the left-hand side of the first equality with respect to y and the second one with respect to x, we get the same results. So the same must be true if we differentiate the right-hand sides. Therefore, we conclude that $2ay = 2bx$ for all x, y. This is possible only if $a = b = 0$, and, as we have already remarked, at least one of the values a, b must be non-zero. This contradiction shows that third order terms in Φ_{opt} are (for $m = 5$) impossible, and therefore Φ_{opt} must contain only second-order terms.

If all these terms are proportional to $x^2 + y^2$, then the total dimension is 4. So there must be at least one quadratic term $F(x,y)$ in Φ_{opt}, that is not proportional to $x^2 + y^2$. Like in 12., we conclude that the family Φ_{opt} must contain a function $C_1x^2 + C_2y^2$ with $C_1 \neq C_2$. Again like in 12., we can apply the 90^o-rotation and conclude that the family also contains a function $C_2x^2 + C_1y^2$. Since Φ_{opt} is a linear space, it must contain a difference of these functions $(C_1 - C_2)(x^2 - y^2)$ and, since $C_1 \neq C_2$ (and thence $C_1 - C_2 \neq 0$) their linear combination $x^2 - y^2$. Again like in 12. we can now apply a 45^o-rotation and thus prove that a function xy belongs to Φ_{opt}.

So we already have 5 linearly independent functions $1, x, y, x^2 - y^2, xy$ in a 5-dimensional linear space Φ_{opt}. Therefore this space Φ_{opt} coincides with the set of all possible linear combinations of these functions.

It can be then easily checked that the corresponding curves are either straight lines or hyperbolas.

References

[1] D. H. Ballard, "Strip trees, a hierarchical representation for curves," *Communications of the ACM*, 1981, Vol. 24, No. 5, pp. 310–321.

[2] N. Bartneck, "A general data structure for image analysis based on a description of connected components", *Computing*, 1989, Vol. 42, pp. 17–34.

[3] R. Bellman, *Introduction to Matrix Analysis.* McGraw-Hill, N. Y., 1970.

[4] H. Bunke, *Modellgestuerte Bildanalyse.* B. G. Teubner, Stuttgart, 1985.

[5] E. Davis, *Representing and Acquiring Geographic Knowledge.* Pitman, London; Morgan Kaufmann, Los Altos, CA, 1986.

[6] M. Davis, A. Ybarra, V. Kreinovich and C. Quintana. *Patterns of molecular aggregation in mixtures: what mathematical models fit them best?* Technical Report, University of Texas at El Paso, Computer Science Department, 1992.

[7] H. Freeman, "Computer processing of line-drawn images", *Computing Surveys*, 1974, Vol. 6, No. 1, pp. 57–97.

[8] R. I. Freidzon, V. Kreinovich, et al. *A knowledge-based navigation system*, Technical Report, Soviet Ministry of Marine Industry, Soviet National Bureau of Cartography and Geodesy and Soviet Ministry of Defence, Leningrad, 1989 (in Russian).

[9] A. I. Gerasimov, *Applications of projective (piecewise-linear) transformations.* Master's Thesis, Leningrad Polytechnic Institute, Leningrad, 1989 (in Russian).

[10] K. Grebner, "Model based analysis of industrial scenes", *Proceedings of the Conference on Computer Vision and Pattern Recognition*, Miami, 1986, pp. 28–33.

[11] U. Kirchner, "Problem adapted modeling for industrial scene analysis", *Proceedings of the Conference on Computer Vision and Pattern Recognition*, San Diego, CA, 1989.

[12] V. Kreinovich, *Sea cartography: optimal data representation.* Center for New Informational Technology Informatika, Technical Report, Leningrad, 1989.

[13] V. Kreinovich, V. "Group-theoretic approach to intractable problems", Lecture Notes in Computer Science, Springer-Verlag, Berlin, Vol. 417, 1990, pp. 112–121.

[14] V. Kreinovich and S. Kumar, "Optimal choice of &- and ∨-operations for expert values", *Proceedings of the 3rd University of New Brunswick Artificial Intelligence Workshop*, Fredericton, New Brunswick, Canada, 1990, pp. 169–178.

[15] V. Kreinovich and C. Quintana. "Neural networks: what non-linearity to choose", *Proceedings of the 4th University of New Brunswick Artificial Intelligence Workshop*, Fredericton, New Brunswick, 1991, pp. 627–637.

[16] V. Kreinovich, C. Quintana and O. Fuentes. *Genetic algorithms: what fitness scaling is optimal?* Technical Report, University of Texas at El Paso, Computer Science Department, 1992.

[17] U. Montanari, "A note on minimal length polygon approximation", *Communications of the ACM*, 1970, Vol. 13, pp. 41–47.

[18] H. Niemann, *Pattern Analysis.* Springer-Verlag, Berlin, 1981.

[19] T. Pavlidis, "Curve fitting as a pattern recognition problem", *Proceedings of the 6th International Joint Conference on Pattern Recognition*, Munich, 1982, Vol. 2, p. 853.