A Family of Tangent Continuous Cubic Algebraic Splines

MARCO PALUSZNY
Universidad Central de Venezuela
and
RICHARD R. PATTERSON
Indiana University and Purdue University at Indianapolis

We present an algorithm for creating tangent continuous splines from segments of algebraic cubic curves. The curves used are cubic ovals, and thus are guaranteed convex. Each segment is given by an equation which has five coefficients, thus four degrees of freedom available for shape control. We describe shape handles that work via the coefficients to control the curve. Each segment can be chosen to interpolate one more point and slope and has two additional fullness parameters to control the shape. This family of curves naturally contains conic splines as a subfamily.

Categories and Subject Descriptors: I.3.5 [**Computer Graphics**]: Computational Geometry and Object Modeling—*curve, surface, solid and object representations*; J.6 [**Computer Applications**]: Computer-Aided Engineering—*computer-aided design*

General Terms: Algorithms, Design

Additional Key Words and Phrases: Cubic ovals, cubic splines, implicit curves, interpolation, piecewise algebraic curves

1. INTRODUCTION

In this paper we study splines created from segments of algebraic curves. These are not the usual parametric splines in which each component function x, y depends on a parameter t: $\langle x, y \rangle = \langle f(t), g(t) \rangle$. In an algebraic spline each segment is given implicitly as a segment of the graph of an implicit equation $f(x, y) = 0$. The subject was initiated by T. W. Sederberg in [21] and [22].

The main advantage of parametric curves is that they are easy to graph and to control using the methods of B-Splines and Bézier curves [7]. The

Authors' addresses: M. Paluszny, Escuela de Física y Matemáticas, Facultad de Ciencias, Universidad Central de Venezuela, Apartado 47809, Los Chaguaramos, Caracas 1041-A, Venezuela; R. R. Patterson, Department of Mathematical Sciences, Indiana University-Purdue University at Indianapolis, 402 N. Blackford St., Indianapolis, IN 46202-3216.

main advantage of implicit curves is the ease of checking whether a given point lies on, to the right of, or to the left of the curve. Implicit curves of a given degree greater than two also have more degrees of freedom for shape control.

There are two areas in which much more work is needed in order to make implicit curves as useful as parametric curves are: graphing methods and shape control. Graphing methods so far are relatively slow, but are becoming more sophisticated to provide greater accuracy. Methods studied so far include walking along the curve pixel by pixel [4], stepping along the tangent vector and making corrections, taking into account singular points [2], parametrizing the curve by elliptic functions (elliptic curves only) [16], finding the intersections of the curve with a triangulation of space [6], and finding local rational approximations [3]. The curves used in this paper are so predictable they can be graphed by the most naive method.

Some contributions to the study of shape control of implicit curves are made by Sederberg [22] in the original paper on this subject. Li, Hoschek, and Hartmann study a case in [12] which has one degree of freedom. Sederberg [20] studies the problem of finding intersections of implicit curves, and Garrity and Warren [8, 25] compare several definitions of geometric continuity for them. More attention has been given to implicit surfaces [1, 5, 9–14, 19, 21, 23, 24].

This paper is a contribution to the study of shape control of implicit cubics. The general cubic has ten coefficients, but after endpoint positions and tangents are imposed has six coefficients and five degrees of freedom left. A complete geometric analysis of these five degrees of freedom is the ultimate goal, but in this paper we restrict to a subfamily with four degrees of freedom. The curves in this subfamily are arcs of cubic ovals. The shape handles discussed in this paper are (primarily) interpolation conditions. In a subsequent paper we will discuss curvature conditions.

The individual segments of our splines inherit from cubic ovals the properties of convexity and nonsingularity. Thus any inflection points or double points in a spline must be explicitly put in by the user. In addition these curves and the splines created from them have the properties

—local control,
—variation diminishing,
—convex hull,
—affine invariance, and
—quadratic precision.

These curves are a direct generalization of conic splines, which form a natural subfamily. Unlike many spline constructions, there are no knot considerations and no global systems of equations to solve.

We begin in Section 2 by reviewing the notation we use for barycentric coordinates. In Section 3 we review the classification of algebraic cubic curves. Sederberg's construction is reviewed in Section 4. In Section 5 we define the family of curves we use and prove that they are segments of cubic ovals. The shape handles used for individual segments are described in Sections 6 and 7. In Section 8 we describe the spline construction.

2. BARYCENTRIC COORDINATES

Barycentric coordinates in the affine plane are defined with respect to a triangle. If the vertices P_0, P_1, and P_2 have Euclidean coordinates $P_0(x_0, y_0)$, $P_1(x_1, y_1)$, and $P_2(x_2, y_2)$, the Euclidean coordinates (x, y) of any point P can be expressed

$$(x, y) = s(x_0, y_0) + t(x_1, y_1) + u(x_2, y_2)$$

with $s + t + u = 1$. Thus P has barycentric coordinates $P(s, t, u)$. The barycentric coordinates of the vertices themselves are $P_0(1, 0, 0)$, $P_1(0, 1, 0)$, and $P_2(0, 0, 1)$.

Passing between barycentric and Euclidean coordinates is done using a matrix

$$(x, y, 1) = (s, t, u)\begin{pmatrix} x_0 & y_0 & 1 \\ x_1 & y_1 & 1 \\ x_2 & y_2 & 1 \end{pmatrix}$$

and its inverse. Using these substitutions we can write an equation in either Euclidean or barycentric coordinates.

Passing from barycentric coordinates to general homogeneous coordinates based on the triangle and with its barycenter $P(1/3, 1/3, 1/3)$ as unit point is done by sending (s, t, u) to its equivalence class $[s, t, u]$, where (s, t, u) is equivalent to (ks, kt, ku) for $k \neq 0$. Passing from homogeneous to barycentric coordinates is possible for those points not on the line at infinity $s + t + u = 0$ by choosing the representative of the class $[s, t, u]$ which satisfies $s + t + u = 1$. See [17] for more about general homogeneous coordinates.

3. CLASSIFICATION OF CUBICS

In this section we review the classification of real algebraic cubic curves. When its equation can be factored, a cubic is *reducible* and is either a line and a conic or three lines. Otherwise the curve is *irreducible*, in which case it is either *singular* or *nonsingular*. A singular cubic has exactly one double point which is either a *crunode*, an *acnode*, or a *cusp* [15]. A nonsingular cubic has either one circuit or two as a curve in the real projective plane. For the curves with two circuits, one is the *oval*, which separates the projective plane into a component which is topologically a disk and one which is a Mobius band, and the other is the *odd circuit*, which does not separate the projective plane; see Figure 1. A cubic with one circuit has only the odd circuit.

In order to determine from the coefficients of the cubic which type is present, we can use a discriminant Δ defined in [18, pp. 191–192].

THEOREM 3.1. *A homogeneous cubic F is irreducible and nonsingular if and only if $\Delta(F) \neq 0$. Moreover F has one or two circuits as $\Delta(F) > 0$ or $\Delta(F) < 0$, respectively.*

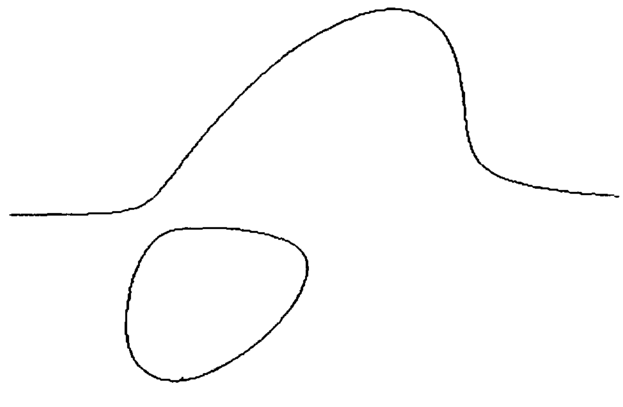


Fig. 1. A two-circuit cubic.

The polynomial Δ has degree 12 in the 10 coefficients of the general cubic and has over 10,000 terms. However in the case of the cubic we shall be using,

$$F(s,t,u) = as^2u + bsu^2 - cst^2 - dt^2u + estu - ft^3,$$

four of the coefficients are zero and the formula for Δ reduces to

$$\begin{aligned}\Delta = a^2b^2(&27a^2b^2f^4 - abe^3f^3 - 36a^2bdef^3 - 36ab^2cef^3 + ade^4f^2\\ &+bce^4f^2 + 8a^2d^2e^2f^2 + 46abcde^2f^2 + 8b^2c^2e^2f^2 + 16a^3d^3f^2\\ &-24a^2bcd^2f^2 - 24ab^2c^2df^2 + 16b^3c^3f^2 - cde^5f - 8acd^2e^3f\\ &-8bc^2de^3f - 16a^2cd^3ef + 64abc^2d^2ef - 16b^2c^3def - c^2d^2e^4\\ &-8ac^2d^3e^2 - 8bc^3d^2e^2 - 16a^2c^2d^4 + 32abc^3d^3 - 16b^2c^4d^2),\end{aligned}$$

up to a positive scalar.

4. REVIEW OF SEDERBERG'S CONSTRUCTION

Sederberg [21, 22] has proposed the following idea which we review here for the case of cubics. It begins with the Bézier construction of triangular surface patches. Given a reference triangle $P_0P_1P_2$, some specific points X_{ijk} in the triangle are defined by

$$X_{ijk} = \frac{i}{3}P_0 + \frac{j}{3}P_1 + \frac{k}{3}P_2,$$

for $i, j, k \geq 0$ and $i + j + k = 3$ as in Figure 2. If a z-coordinate P_{ijk} is assigned to each point X_{ijk}, the corresponding points (X_{ijk}, P_{ijk}) are the control points of a triangular Bézier patch; the graph of

$$z = F(s,t,u) = \sum \binom{3}{i,j,k} P_{ijk} s^i t^j u^k.$$

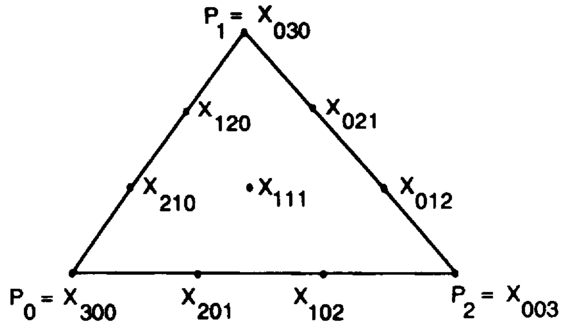


Fig. 2. The points X_{ijk} in the triangle.

The intersection of this surface with the plane $z = 0$ is a plane curve $F(s, t, u) = 0$, defined implicitly in terms of the barycentric coordinates of the triangle. Now there are exactly as many points X_{ijk} as there are coefficients in a polynomial of degree 3. So the P_{ijk} can be reinterpreted both as the coefficients of an implicit equation and as weights at the points X_{ijk}. In this way every implicit cubic equation, when rewritten in barycentric coordinates with respect to some triangle, can be thought of as defining the curve of intersection of a Bézier surface with the plane.

Using this interpretation of an implicit equation, Sederberg described several geometric properties of the curve in terms of the coefficients. He showed that if the weight at a vertex of the triangle is zero then the curve passes through the vertex. Furthermore if an adjoining weight on one edge is also zero the curve is tangent at the vertex to that side of the triangle.[1] In Figure 3 the points X_{ijk} are labeled with the corresponding coefficients P_{ijk}. Only that portion of the curve which lies inside the triangle is graphed.

5. CUBIC OVALS

Suppose we are given two points P_0 and P_2 and lines through them which intersect at a point P_1. We are interested in finding convex cubic arcs which lie within the triangle $P_0P_1P_2$, join P_0 to P_2 and are tangent at P_0 and P_2 to the two given lines. If we use this triangle to define barycentric coordinates as in Section 2, we know from the results of Sederberg that for a cubic satisfying the interpolation conditions the coefficients of s^3, s^2t, tu^2, and u^3 are zero. The equation of such a curve therefore has the form

$$F(s,t,u) = as^2u + bsu^2 - cst^2 - dt^2u + estu - ft^3 = 0.$$

The notation has been simplified to avoid the triple subscripts since a good deal of algebra must be done with these coefficients. Figure 4 illustrates how these coefficients and the barycentric coordinates correspond to points in the triangle.

[1]This tangency property may be lost if the vertex happens to be a multiple point; see Figure 9b.

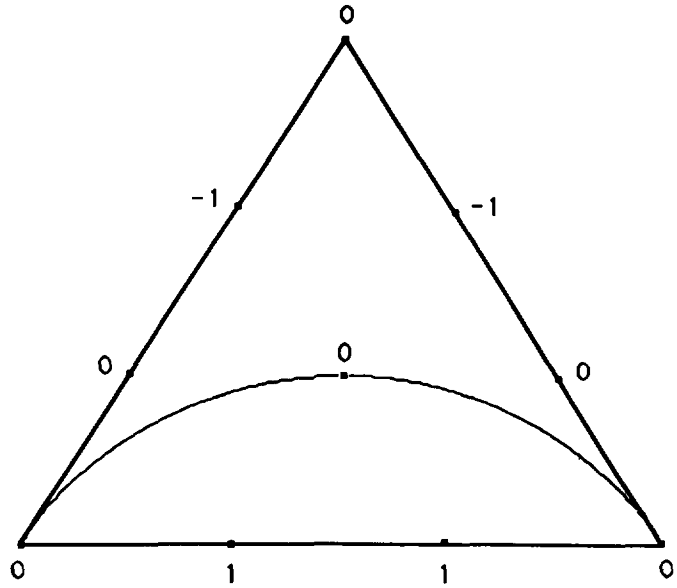


Fig. 3. Graph of $s^2u + su^2 - st^2 - t^2u = 0$.

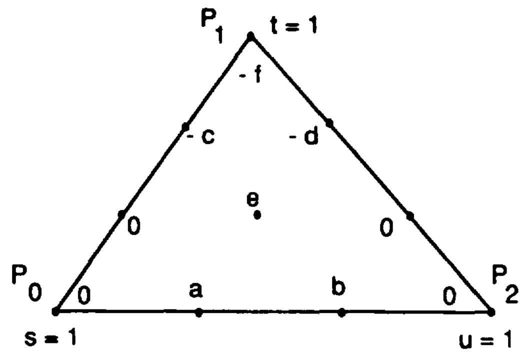


Fig. 4. Notation for the triangle.

We now begin to impose additional conditions on the coefficients so that the curves in the family include a convex arc within the triangle from P_0 to P_2, tangent to the sides. The first condition is that a and b must have the same sign. This is because the line $t = 0$ intersects the curve in P_0, P_2 and the point $P(s, 0, u)$ whose coordinates satisfy $as + bu = 0$. If a and b have opposite signs, P lies between P_0 and P_2. To avoid this we now impose the condition $a, b > 0$.

Next we impose the condition $f = 0$. This causes the curve to pass through P_1; more is said about this at the end of this section.

We can see by example that we have not yet imposed enough conditions to keep the curve inside the triangle. Figure 5 shows the curve when $a = b = 1$,

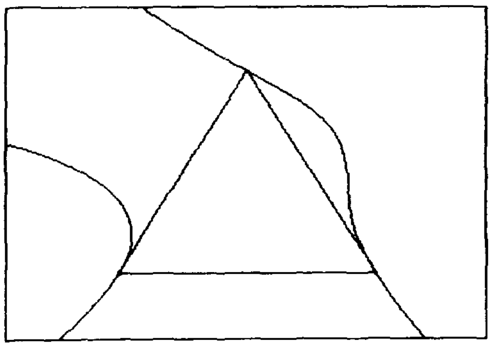


Fig. 5. A curve that lies outside the triangle.

$c = -1$, $d = -1/2$, and $e = 0$. This and similar examples suggest that we might need $c, d > 0$. The next two results combine to show that in this case we have a useful subfamily of curves.

THEOREM 5.1. *If $a, b, c, d \geq 0$, the cubic*

$$F(s, t, u) = as^2u + bsu^2 - cst^2 - dt^2u + estu$$

is an irreducible and nonsingular two-circuit cubic except in the following cases:

—*If $a = 0$ (respectively $b = 0$), the curve is singular with a double point at P_0 (respectively P_2);*
—*If $c = 0$ (respectively $d = 0$), the curve is reducible with a factor u (respectively s);*
—*If $ad = bc$ and $e = 0$, the curve is reducible*

$$F(s, t, u) = (as + bu)\left(su - \frac{c}{a}t^2\right).$$

PROOF. If $a = 0$, all three first partial derivatives of F have the root $s = 1$, $t = u = 0$, so that P_0 is a double point. Similarly if $b = 0$, P_2 is a double point. The conditions involving c and d are easy to see. Because $f = 0$, the discriminant Δ reduces to

$$\Delta = -a^2b^2c^2d^2\left(e^4 + 8(ad + bc)e^2 + 16(ad - bc)^2\right).$$

If $a, b, c, d > 0$, Δ is negative and the curve therefore has two circuits unless the factor involving e is zero; in other words

$$e^2 = -4(ad + bc) \pm 8\sqrt{abcd}\,.$$

The minus sign is impossible and

$$-4(ad + bc) + 8\sqrt{abcd} \geq 0$$

if and only if

$$(ad - bc)^2 \leq 0.$$

Thus the factor involving e is zero if and only if $ad - bc = 0$ and $e = 0$, in which case F is reducible as indicated. □

Now that we know the curve normally has two circuits we can ask which one passes through the points P_0 and P_2.

THEOREM 5.2. *If $a, b, c, d > 0$, the graph of $as^2u + bsu^2 - cst^2 - dt^2u + estu = 0$ includes an arc of a cubic oval which lies inside the triangle, joins P_0 to P_2, and is tangent to the sides of the triangle. If $ad - bc = 0$ and $e = 0$, the curve reduces to a conic arc.*

PROOF. The proof entails looking for silhouette points. According to Salmon [18], Articles 167 and 200, from a point on a one-circuit curve, two lines can be drawn which are tangent to the curve at other points. These points of tangency are silhouette points as seen from the given point. From a point on the odd circuit of a two-circuit curve, four such lines can be drawn—two to points on the odd circuit and two to points on the oval. From a point on the oval, none can be drawn.

We use this idea to test whether P_0 lies on the oval, first in the case $e = 0$. Treat $P_0(1, 0, 0)$ as the origin and search for lines through P_0 which are tangent to the curve. Into $as^2u + bsu^2 - cst^2 - dt^2u = 0$, $a, b > 0$, substitute $u = mt$ and $s = 1$.[2] We get

$$t[-dmt^2 + (bm^2 - c)t + am].$$

The factor t corresponds to the intersection at P_0. The quadratic factor has a double root if its discriminant $G(m)$ equals 0:

$$G(m) = b^2m^4 + (4ad - 2bc)m^2 + c^2 = 0.$$

Thus the question is reduced to whether G has 0 or 4 real roots, since we have already determined that the curve has two circuits. In other words, P_0 lies on the oval if and only if the minimum of G is positive. From

$$G'(m) = 4b^2m^3 + 4(2ad - bc)m$$

we see the critical points occur when $m = 0$ and

$$m^2 = \frac{bc - 2ad}{b^2}.$$

The value of G at 0 is positive if $c \neq 0$. When $bc - 2ad < 0$, 0 is the only point that needs to be considered. So the dotted area in Figure 6 above the line $bc - 2ad = 0$ in c, d-parameter space contains coefficient values for which P_0 lies on the oval. In addition, when $bc - 2ad \geq 0$ we have those

[2] No points are missed by ignoring the line $t = 0$ through P_0 and P_2. It already crosses the curve twice. It can not be tangent at another point without violating Bezout's theorem, according to which a line meets a cubic in at most three points.

Letting $s = 1$ means we are thinking about homogeneous coordinates rather than affine coordinates for a silhouette point which does not lie on the line $s = 0$. No such point could lie on the line $s = 0$ since $s = 0$ already meets the curve once at P_1 and twice at P_2.

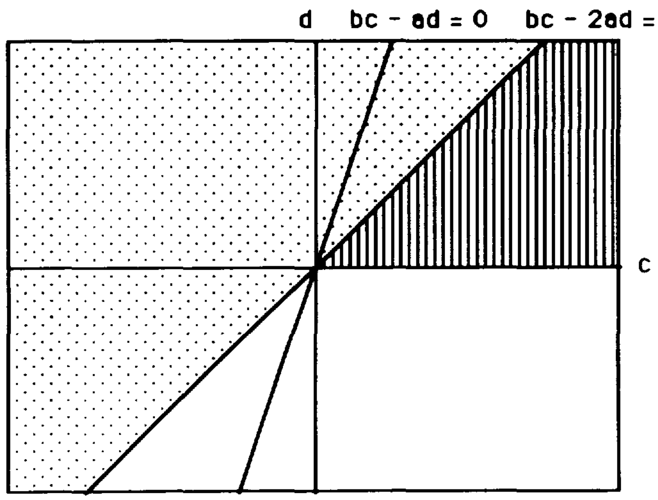


Fig. 6. The region for which P_0 lies on the oval.

points for which the value of G at $\pm m$ is positive. Substituting for m^2 we find

$$b^2G = 4ad(bc - ad).$$

Thus since $d > 0$, G is positive if and only if $bc - ad > 0$, which adds the region that is vertically striped.

Repeating the argument with P_2 we find the region in c, d-parameter space that corresponds to having the point P_2 on the oval. The intersection of the two regions is the first quadrant.

Because P_0 and P_2 lie on the oval when $e = 0$, this remains true by continuity for all e. □

Finally, we can explain the reason for choosing $f = 0$; this causes the odd circuit of the curve to pass through P_1. With the oval tangent at P_0 and P_2, each of the lines $s = 0$ and $u = 0$ has three known intersections with the cubic. Because of Bezout's theorem the odd circuit has no other intersections with these lines and is thus under control safely out of the way.

6. THE CURVE CONSTRUCTION

In this section we describe the parameters used as shape handles to control the curve via the coefficients. The shape handles we examine here are additional interpolation conditions; curvature conditions will be described in a sequel.

We first select a point $B_0(s_0, t_0, u_0)$ that lies inside the triangle and is to be interpolated. Second we specify the tangent line at B_0. One way to describe it is in terms of the point $Q(s_1, 0, u_1)$ where the tangent line at B_0 meets the line P_0P_2 (it is possible for Q to lie at infinity). A more convenient parameter for theoretical purposes is the cross ratio R of either the four points Q, B_0,

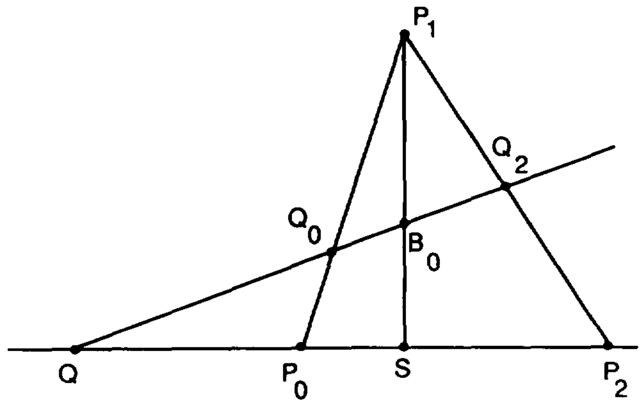


Fig. 7. Defining the cross ratio R.

Q_0, and Q_2 shown in Figure 7, or equivalently the four points Q, S, P_0, and P_2, where $S(s_0/(s_0+u_0), 0, u_0/(s_0+u_0))$ is the projection of B_0. We obtain

$$R = R(Q, S, P_0, P_2) = \frac{\begin{vmatrix} s_1 & u_1 \\ 1 & 0 \end{vmatrix} \begin{vmatrix} \dfrac{s_0}{s_0+u_0} & \dfrac{u_0}{s_0+u_0} \\ 0 & 1 \end{vmatrix}}{\begin{vmatrix} s_1 & u_1 \\ 0 & 1 \end{vmatrix} \begin{vmatrix} \dfrac{s_0}{s_0+u_0} & \dfrac{u_0}{s_0+u_0} \\ 1 & 0 \end{vmatrix}} = \frac{s_0 u_1}{u_0 s_1}.$$

For Q between P_0 and P_2, R is positive, and for Q outside this interval R is negative. (Another parameter m, more intuitive for a designer, is presented in Section 8.) The remaining two degrees of freedom are controlled by a pair (β_1, β_2), each of which can vary from 0 to 1.

THEOREM 6.1. *Given a point $B_0(s_0, t_0, u_0)$ inside the triangle $P_0P_1P_2$ and a point $Q(s_1, 0, u_1)$ on the line $t = 0$, there exists a cubic in the family $as^2u + bsu^2 - cst^2 - dt^2u + estu = 0$, $a, b, c, d > 0$ which passes through B_0 tangent to the line B_0Q if and only if Q is outside the interval $[P_0, P_2]$; $s_1 < 0$ or $s_1 > 1$. A solution corresponds to a choice of (β_1, β_2) in the unit square, with coefficients determined as follows:*

$$a := -R\beta_2 t_0^2 u_0,\ b := \beta_1 s_0 t_0^2,$$
$$c := (1-\beta_1)s_0 u_0^2,\ d := -R(1-\beta_2)s_0^2 u_0,\ e := (1 - R - 2\beta_1 + 2R\beta_2)s_0 t_0 u_0.$$

The proof of Theorem 6.1 is postponed to the Appendix. For the rest of this section we study properties of this curve construction. In the next section we examine the effect of the β's on the shape of the curve.

If B_0 is given, because R must be negative the parameter space is the infinitely long bar in Figure 8. In general each point in this bar determines a different oval arc that satisfies the contact conditions at P_0, P_2 and passes

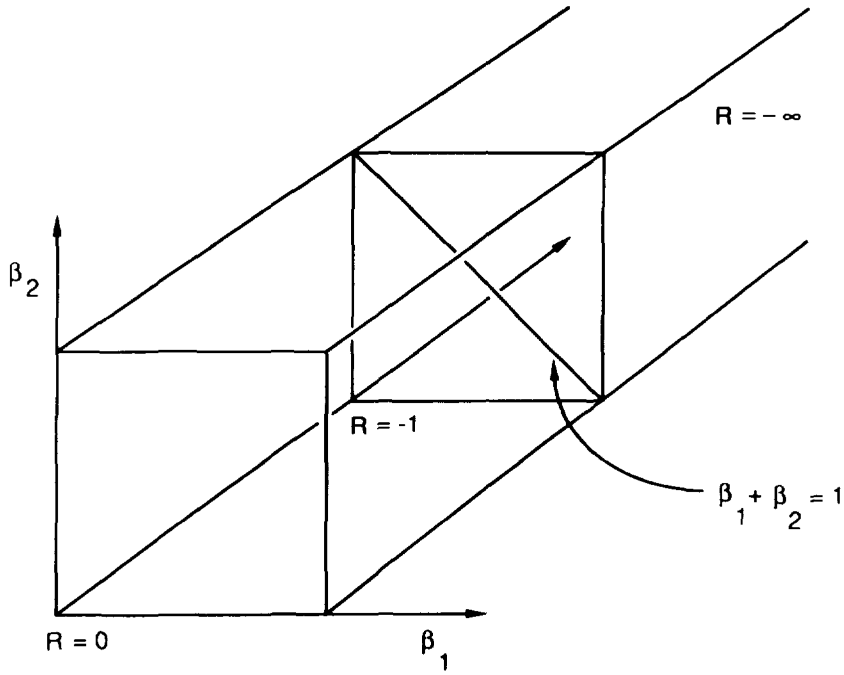


Fig. 8. The parameter domain.

through B_0. However certain points determine the reducible curves which occur when $ad = bc$ and $e = 0$. When the curve is reducible, the quadratic factor describes a conic which satisfies the contact conditions and passes through B_0. Here is an equivalent way to express the reducibility condition in terms of the parameters β_1, β_2, and R.

PROPOSITION 6.2. *The curve is reducible if and only if* $R = -1$ *and* $\beta_1 + \beta_2 = 1$; *see Figure* 8. *The arc of the reducible curve is independent of* β_1 *and* β_2 *so long as* $\beta_1 + \beta_2 = 1$.

PROOF. From $ad - bc = 0$ and $e = 0$, we deduce

$$R^2\beta_2(1 - \beta_2) = \beta_1(1 - \beta_1)$$

and

$$(1 - 2\beta_1) = R(1 - 2\beta_2). \tag{1}$$

Eliminating R we obtain

$$(\beta_1 - \beta_2)(\beta_1 + \beta_2 - 1) = 0. \tag{2}$$

Since R is negative we obtain from (1) that $\beta_1 \geq 1/2$, $\beta_2 \leq 1/2$ or the reverse. In either case from (2), $\beta_1 + \beta_2 = 1$. Then from (1), $R = -1$. Conversely if $R = -1$ and $\beta_1 + \beta_2 = 1$ we can evaluate and find $ad - bc = 0$ and $e = 0$.

When $R = -1$ and $\beta_1 + \beta_2 = 1$

$$a = \frac{t_0^2}{s_0 u_0} c \quad \text{and} \quad b = \frac{t_0^2}{s_0 u_0} d.$$

Letting

$$\gamma = \frac{t_0^2}{s_0 u_0}$$

we have the specific reduction

$$as^2u + bsu^2 - cst^2 - dt^2u = (cs + du)(\gamma su - t^2).$$

As γ is independent of c and d, any change to c or d or to β_1 or β_2 goes into a change of the linear factor, not the quadratic. □

In consequence our curves contain the subfamily of conics. Assuming that we graph our curves by tracing from P_0 or P_2, in the reducible case the conic arc is what is drawn.

In fact this construction has quadratic precision. Suppose we have a conic arc from P_0 to P_2 within the triangle, with P_1 the intersection of the tangent lines and with $B_0(s_0, t_0, u_0)$ any interior point of the arc. If we then use the value of R determined by the tangent line at B_0 and any β_1, β_2 such that $\beta_1 + \beta_2 = 1$, the construction recovers the original conic arc. Because of the endpoint and tangency conditions on the conic, the coefficients of s^2, st, tu, and u^2 are zero. We can then normalize the equation to the form $\gamma su - t^2 = 0$. Because B_0 lies on the curve, $\gamma = t_0^2/s_0 u_0$. The tangent line to the conic at B_0 has equation $\gamma u_0 s - 2t_0 t + \gamma s_0 u = 0$. This line meets the line P_0P_2 at the point $Q(s_1, 0, u_1)$, where

$$s_1 = \frac{s_0}{s_0 - u_0} \quad \text{and} \quad u_1 = \frac{-u_0}{s_0 - u_0}.$$

We find $R = -1$. For any β_1, β_2 such that $\beta_1 + \beta_2 = 1$ the curve is reducible and we recover the original conic.

7. THE EFFECTS OF THE β's

In this section we investigate the behavior of the curves $F = 0$ when B_0 and the slope at B_0 are fixed and the parameters β_1 and β_2 are varied.

The general formula for F is

$$\begin{aligned} F(s,t,u) = {} & -R\beta_2 t_0^2 u_0 s^2 u + \beta_1 s_0 t_0^2 su^2 - (1 - \beta_1) s_0 u_0^2 st^2 \\ & + R(1 - \beta_2) s_0^2 u_0 t^2 u \\ & + (1 - R - 2\beta_1 + 2R\beta_2) s_0 t_0 u_0 stu. \end{aligned}$$

This can be factored

$$(1 - \beta_1)(1 - \beta_2)F_{00} + (1 - \beta_1)\beta_2 F_{01} + \beta_1(1 - \beta_2)F_{10} + \beta_1\beta_2 F_{11}. \quad (3)$$

The formulas for the individual F_{ij} are

$$\begin{aligned}
F_{00} &= s_0 u_0 t[-u_0 st + Rs_0 tu + (1-R)t_0 su],\\
F_{01} &= u_0 s\left[-Rt_0^2 su - s_0 u_0 t^2 + (1+R)s_0 t_0 tu\right],\\
F_{10} &= s_0 u\left[t_0^2 su + Rs_0 u_0 t^2 - (1+R)t_0 u_0 st\right],\\
F_{11} &= t_0 su[-Rt_0 u_0 s + (R-1)s_0 u_0 t + s_0 t_0 u].
\end{aligned}$$

These four extremal curves are all reducible. The three lines composing F_{11} are the sides of the triangle and the line $B_0 Q$. Figure 9 displays the graphs of these curves and eight other extreme curves which correspond to values of β_1 and β_2 on the edges of the unit square. The curves all interpolate the barycenter $B_0(1/3, 1/3, 1/3)$ with $R = -1/3$.

Because the coefficient of $\beta_1 \beta_2$ in (3) is zero, we have the identity

$$F_{00} - F_{01} - F_{10} + F_{11} = 0.$$

By evaluating at $(1/2, 0, 1/2)$ we see that F is positive below the graph of $F = 0$ and negative above. (The meaning of "below," of course, is relative to the current orientation of the triangle.) Similarly the quadratic factors of F_{00}, F_{01}, and F_{10} and all three linear factors of F_{11} are positive at $(1/2, 0, 1/2)$. We can use this to show that every oval that corresponds to a point (α_1, α_2) in the shaded region in Figure 10 lies above the oval that corresponds to the corner point (β_1, β_2).

PROPOSITION 7.1. *The graph of $F_{\alpha_1, \alpha_2} = 0$ is above the graph of $F_{\beta_1, \beta_2} = 0$ (except at B_0) if $\alpha_1 \geq \beta_1$, $\alpha_2 > \beta_2$ or if $\alpha_1 > \beta_1$, $\alpha_2 \geq \beta_2$.*

PROOF. Consider $F_{\beta_1, \beta_2} = 0$ and $F_{\beta_1+\epsilon, \beta_2} = 0$ for positive ϵ. We can write $F_{\beta_1+\epsilon, \beta_2}$ as

$$F_{\beta_1, \beta_2} + \epsilon\beta_1(F_{00} - F_{01} - F_{10} + F_{11}) + \epsilon(F_{10} - F_{00}) = F_{\beta_1, \beta_2} + \epsilon(F_{10} - F_{00}).$$

Now

$$F_{10} - F_{00} = s_0 s(t_0 u - u_0 t)^2$$

is positive at any point P other than B_0 on the graph of $F_{\beta_1, \beta_2} = 0$ inside the triangle. So $F_{\beta_1 + \epsilon, \beta_2}(P)$ is positive except at B_0, where it is zero. It follows that the graph of $F_{\beta_1+\epsilon, \beta_2} = 0$ lies above the graph of $F_{\beta_1, \beta_2} = 0$.

A similar argument works for vertical segments. The rest of the proof follows by transitivity. □

The behavior of the curves on the diagonal $\beta_1 = \beta_2$ is simple; as the β's are increased the curve swells into the corners. See Figure 11.

The behavior transverse to this diagonal is much more subtle. To explain this we consider the problem of interpolating another point U. The region in which we can choose U is bounded by the graphs of $F_{00} = 0$ and $F_{11} = 0$ and is divided by the graphs of $F_{01} = 0$ and $F_{10} = 0$; see Figure 12. If U is chosen above both of the graphs of $F_{01} = 0$ and $F_{10} = 0$ there are reducible curves $F_{\beta_1^r, 1}$ and F_{1, β_2^r} that pass through U. Every curve which corresponds to a

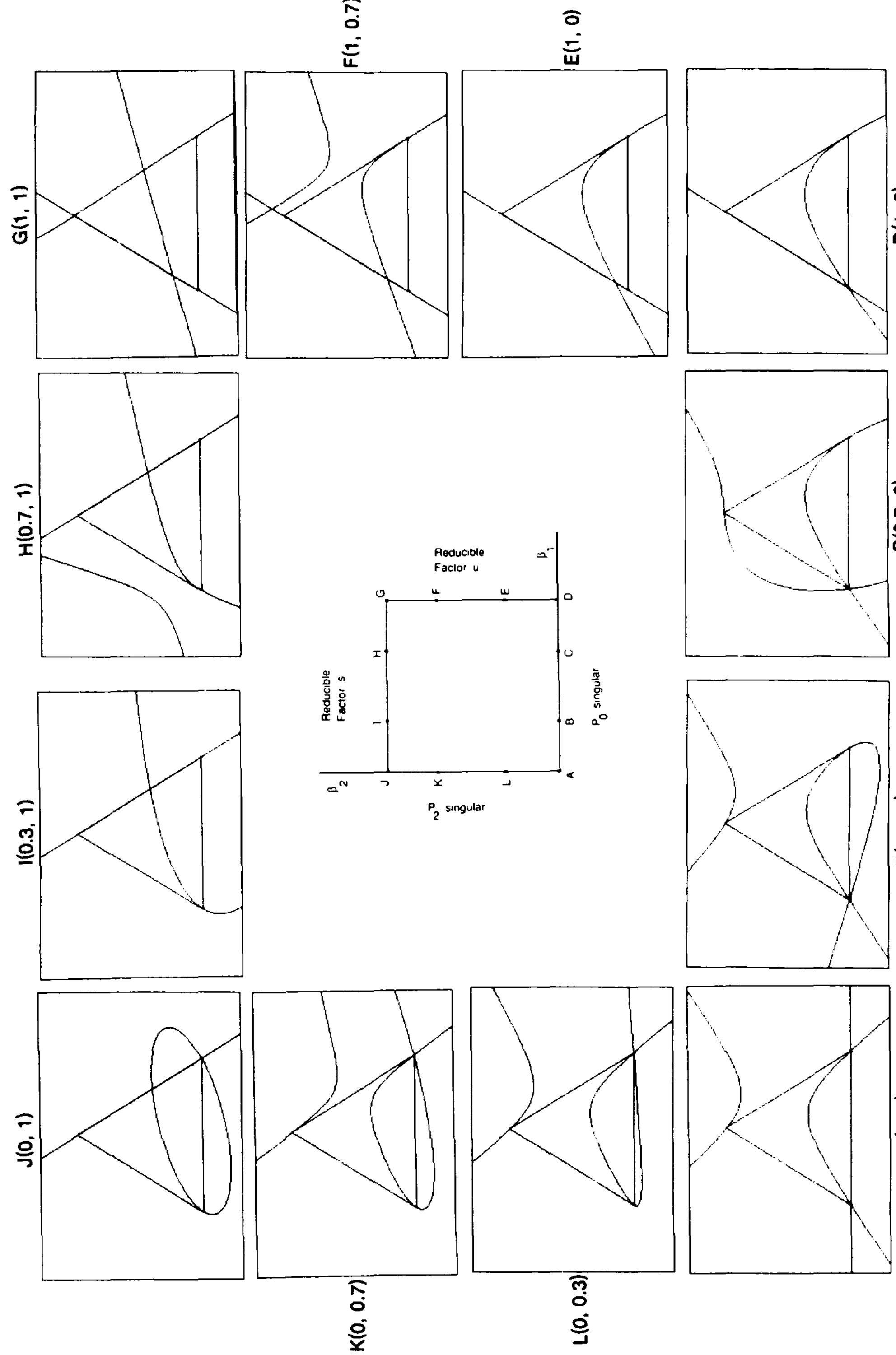


Fig. 9. Twelve extremal curves (A)–(L).

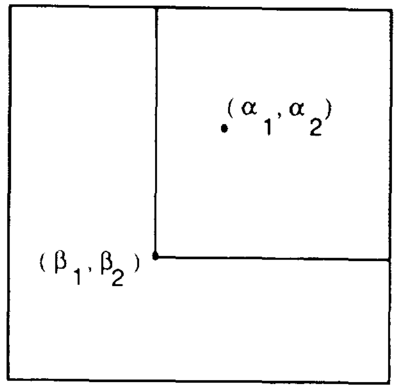


Fig. 10. Parameters of the curves that lie above F_{β_1, β_2}.

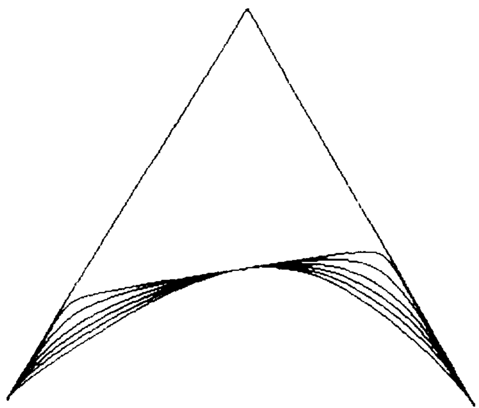


Fig. 11. Curves through $B_0(1/3, 1/3, 1/3)$, $m = 0.3$ with $\beta_1 = \beta_2 = .1, .3, .5, .7, .9, .99$.

point on the segment between $(\beta_1^r, 1)$ and $(1, \beta_2^r)$ determines an oval that passes through U. Figure 13 illustrates this for a particular point U. More generally the motion along a line of positive slope in the (β_1, β_2) control square, say towards $(\beta_1^r, 1)$, corresponds to the swelling of the oval towards the conic arc given by $(\beta_1^r, 1)$. Figures 9h and 9i display such arcs.

The point U could alternatively be chosen in the region corresponding to a segment between $(\beta_1^r, 1)$ and $(\beta_1^s, 0)$ (a reducible curve and a singular curve), $(0, \beta_2^s)$ and $(1, \beta_2^r)$ (a singular curve and a reducible curve), or $(0, \beta_2^s)$ and $(\beta_1^s, 0)$ (two singular curves); see Figure 14. Each of these segments must have negative slope; otherwise the existence of the intersection point U would contradict Proposition 7.1.

8. TANGENT CONTINUOUS SPLINES

To create a tangent continuous spline the user chooses a locally convex control polygon that we label as in Figure 15. The control polygon has endpoints P_0 and P_{2n} (which may be the same point), corner points with odd subscripts and junction points with even subscripts. For tangent continuity the junction points must be constrained to lie on the segments between

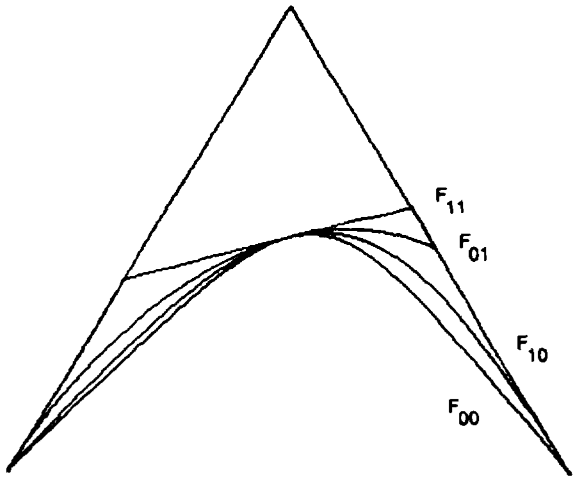


Fig. 12. The region in which another point U can be chosen; $B_0(1/4, 1/2, 1/4)$, $m = 0.3$.

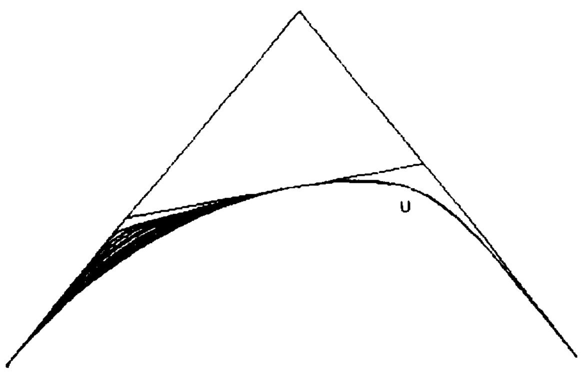


Fig. 13. A selection of curves through U; $B_0(1/4, 1/2, 1/4)$, $m = 0.3$.

corner points. The control polygon defines a sequence of triangles. Let Δ_i be the triangle with vertices P_{2i}, P_{2i+1}, and P_{2i+2}. The user selects an interpolation point B_i in each Δ_i, a tangent line at each B_i, and settings for $(\beta_1[i], \beta_2[i])$ in each Δ_i, $0 \le \beta_1[i], \beta_2[i] \le 1$.

Instead of using the parameter R to describe the tangent line at B_i, an alternative that is more intuitive is an affine version of slope. In Δ_0 the parameter $m_0 \in (-1, 1)$ is defined by

$$m_0 = \frac{1}{s_1 - u_1}, \quad \text{thus } (s_1, u_1) = \left(\frac{m_0 + 1}{2m_0}, \frac{m_0 - 1}{2m_0} \right).$$

This transformation sets up a one-to-one correspondence between points Q that lie on the line P_0P_2 but are outside the interval $[P_0, P_2]$ and the

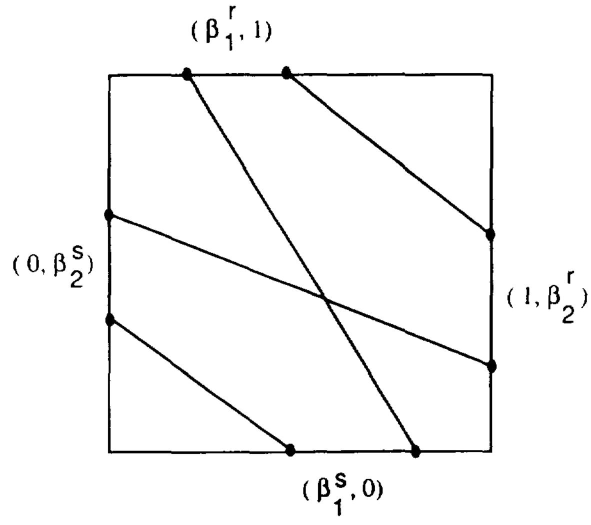


Fig. 14. Segments that correspond to pencils of ovals through various points U.

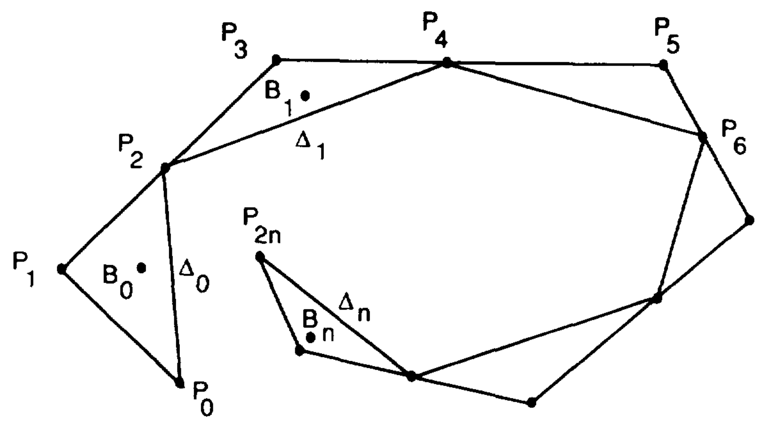


Fig. 15. Control polygon.

interval $(-1, 1)$, with $m_0 = 0$ corresponding to the point at infinity. The value $m_0 = -1$ corresponds to the line through B_0 and P_2, $m_0 = 0$ corresponds to the line parallel to the line P_0P_2, and $m_0 = 1$ corresponds to the line through P_0.

The relationship between R and m_0 is given by

$$R = \frac{s_0(m_0 - 1)}{u_0(m_0 + 1)} \text{ and } m_0 = \frac{s_0 + u_0 R}{s_0 - u_0 R}.$$

When m_0 corresponds to R, $-m_0$ corresponds to $1/R$.

Conic splines are created if all the $\beta_1[i] = \beta_2[i] = 0.5$ and each m_i defaults to $(s_0 - u_0)/(s_0 + u_0)$, where s_0, t_0, and u_0 are the coordinates relative to Δ_i of B_i. These conic splines are modified by moving the interpolation points B_i rather than by the usual way of changing weights at the corners, but this is equivalent. Further modification can then be made in any given segment by varying m_i, $\beta_1[i]$ or $\beta_2[i]$, thus moving to cubic segments.

The remaining properties of these splines listed in the Introduction are all fairly obvious except perhaps for affine invariance. This holds because all the constructions are affine.

Figure 16a and 16b are conic splines. In 16a, the curves pass through the barycenters but in 16b, B_1 has coordinates $(1/6, 2/3, 1/6)$. In 16c and 16d barycenters are used again. In 16c all the $m_i = 0$ and all the β's $= 0.9$. In 16d $m_1 = 0.5$ and all the β's $= 0.5$.

APPENDIX—PROOF OF THEOREM 6.1

The condition that the curve should pass through B_0 gives the first requirement

$$as_0^2 u_0 + bs_0 u_0^2 + es_0 t_0 u_0 = cs_0 t_0^2 + dt_0^2 u_0.$$

The condition that the tangent line should pass through $Q(s_1, 0, u_1)$ gives the second requirement

$$as_0(2u_0 s_1 + s_0 u_1) + bu_0(u_0 s_1 + 2s_0 u_1) + et_0(u_0 s_1 + s_0 u_1) = ct_0^2 s_1 + dt_0^2 u_1.$$

The two requirements can be written together in matrix form

$$(as_0, bu_0, et_0)\begin{pmatrix} 2u_0 s_1 + s_0 u_1 & s_0 u_0 \\ u_0 s_1 + 2s_0 u_1 & s_0 u_0 \\ u_0 s_1 + s_0 u_1 & s_0 u_0 \end{pmatrix} = (ct_0^2, dt_0^2)\begin{pmatrix} s_1 & s_0 \\ u_1 & u_0 \end{pmatrix}$$

or

$$(as_0, bu_0, et_0)\frac{1}{t_0^2(u_0 s_1 - s_0 u_1)}\begin{pmatrix} 2u_0^2 s_1 & -s_0^2 u^1 - s_0 u_0 u_1 \\ u_0^2 s_1 + s_0 u_0 u_1 & -2s_0^2 u_1 \\ u_0^2 s_1 & -s_0^2 u_1 \end{pmatrix} = (c, d).$$

Introducing the notation $R = s_0 u_1/u_0 s_1$, we shorten the above formula to

$$(as_0, bu_0, et_0)\frac{1}{t_0^2(1 - R)}\begin{pmatrix} 2u_0 & -s_0(1 + R) \\ u_0(1 + R) & -2s_0 R \\ u_0 & -s_0 R \end{pmatrix} = (c, d).$$

We are seeking solutions with both (a, b) and (c, d) in the first quadrant.

One condition may be imposed on the five coefficients. We select temporarily the condition

$$as_0 + bu_0 = 1$$

in addition to the requirement $a, b > 0$. With this normalization condition the point (as_0, bu_0, et_0) can be written $as_0(1, 0, et_0) + bu_0(0, 1, et_0)$. Let $w_1(e)$

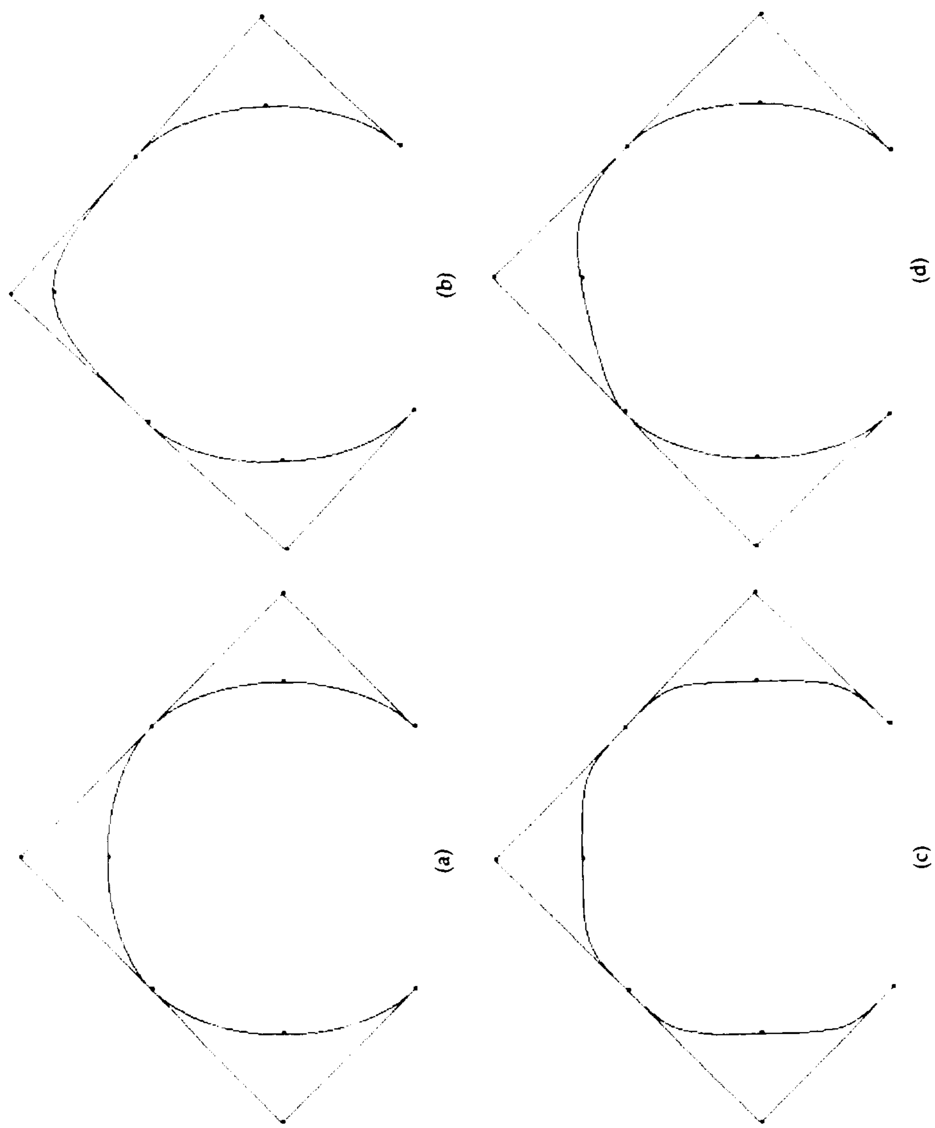


Fig. 16. (a) Conic spline through barycenters, (b) Conic spline through $B_1(1/6, 2/3, 1/6)$, (c) Cubic spline through barycenters, all $m_i = 0$, all β's = 0.9, (d) Cubic spline with $m_1 = 0.5$, all β's = 0.5.

and $w_2(e)$ be the transforms using the matrix above of $(1, 0, et_0)$ and $(0, 1, et_0)$, respectively. Then the point (as_0, bu_0, et_0) transforms to $(c, d) = as_0 w_1(e) + bu_0 w_2(e)$. We can express $w_1(e)$ and $w_2(e)$ in the following way:

$$w_1(e) = \frac{1}{t_0^2(1-R)}[(u_0, -s_0) + (1 + et_0)(u_0, -s_0 R)]$$

and

$$w_2(e) = \frac{1}{t_0^2(1-R)}[R(u_0, -s_0) + (1 + et_0)(u_0, -s_0 R)].$$

Because s_0 and u_0 are positive, if R is negative, the point

$$V_1 = \frac{1}{t_0^2(1-R)}(u_0, -s_0)$$

is in the fourth quadrant,

$$V_2 = \frac{R}{t_0^2(1-R)}(u_0, -s_0)$$

is in the second quadrant, and

$$W = \frac{1}{t_0^2(1-R)}(u_0, -Rs_0)$$

is in the first quadrant of the c, d plane; see Figure 17. As e is allowed to vary, the points $w_1(e) = V_1 + (1 + et_0)W$ and $w_2(e) = V_2 + (1 + et_0)W$ slide along parallel lines l_b and l_a. The point $as_0 w_1(e) + bu_0 w_2(e)$ lies on a segment which runs between the parallel lines. Because of the requirement $c,d > 0$, the domain of variability of (c, d) is the shaded region in Figure 17 where the band meets the first quadrant.

If R if positive, the corresponding constructions for the cases $0 < R < 1$ and $R > 1$ lead to bands that do not intersect the first quadrant; see Figure 18. Hence there is no solution to the interpolation problem if Q lies in $[P_0, P_2]$.

We can now obtain expressions in the case R is negative for e, a, and b in terms of c and d chosen in the domain of variability. The equation of the line through $w_1(e)$ and $w_2(e)$ is independent of R:

$$s_0 t_0^2 c + t_0^2 u_0 d = (1 + et_0) s_0 u_0.$$

Hence

$$e = \frac{1}{s_0 t_0 u_0}\left[s_0 t_0^2 c + t_0^2 u_0 d - s_0 u_0\right]. \tag{4}$$

Using $bu_0 = 1 - as_0$, we can solve

$$as_0 w_1(e) + bu_0 w_2(e) = (c, d)$$

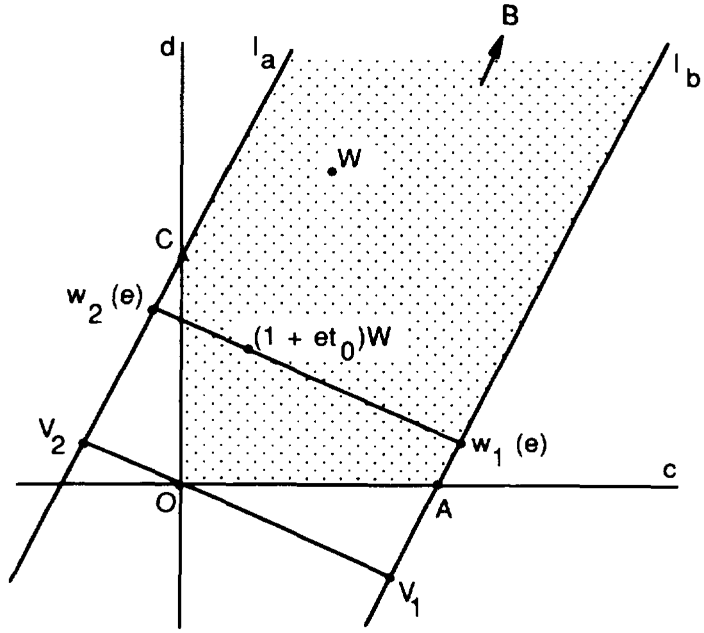


Fig. 17. The domain of variability.

for a. Equating either the first or second component we obtain

$$a = \frac{1}{s_0^2 u_0 (R - 1)} \left[s_0 t_0^2 Rc + t_0^2 u_0 d + s_0 u_0 R \right]. \tag{5}$$

Similarly

$$b = \frac{1}{s_0 u_0^2 (1 - R)} \left[s_0 t_0^2 Rc + t_0^2 u_0 d + s_0 u_0 \right]. \tag{6}$$

The line l_a that passes through the points $w_2(e)$ and V_2 has the equation

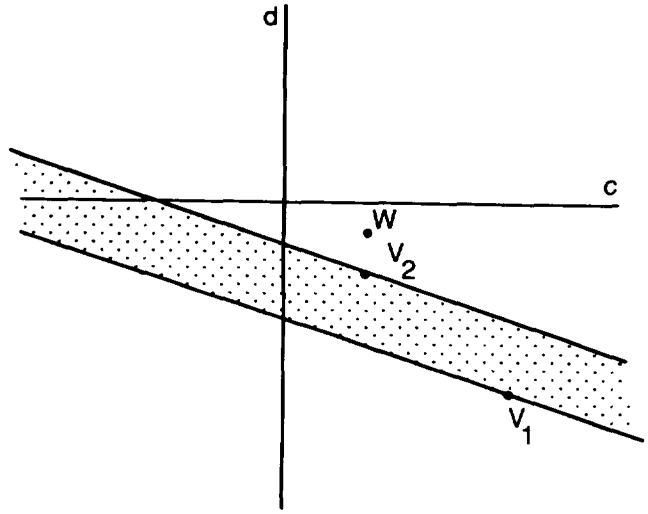
$$s_0 t_0^2 Rc + t_0^2 u_0 d + s_0 u_0 R = 0.$$

The requirement $a > 0$ is equivalent to the condition that (c, d) lie below the line l_a. Similarly the line l_b has equation

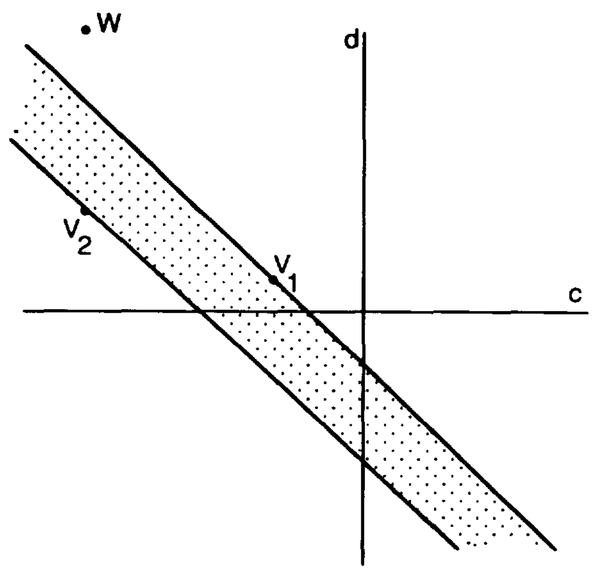
$$s_0 t_0^2 Rc + t_0^2 u_0 d + s_0 u_0 R = 0$$

and the requirement $b > 0$ is equivalent to the condition that (c, d) lie above the line l_b. To summarize, a point (c, d) chosen from the domain of variability determines the other coefficients a, b, and e of a curve which satisfies the interpolation conditions. The four sides of the domain of variability correspond to parameter values of the reducible or singular curves which occur when one or more of a, b, c, or d is equal to 0.

Because the domain of variability is a quadrilateral in the projective plane, we can apply a projective transformation to transform it to a square. This not



(a)



(b)

Fig. 18. The bands in case (a) $0 < R < 1$ (b) $R > 1$.

only makes it easier to describe the parameter domain, it turns out to simplify the formulas.

Using the equations for l_a and l_b, we can find the homogeneous coordinates of the vertices of the domain of variability (see Figure 17): $O[0, 0, 1]$, $A[-u_0, 0, t_0^2 R]$, $B[u_0, -s_0 R, 0]$, and $C[0, -s_0 R, t_0^2]$. The homogeneous coordinates of the unit square are $O[0, 0, 1]$, $U_1[1, 0, 1]$, $U[1, 1, 1]$, and $U_2[0, 1, 1]$.

Using the methods of [17], we find the matrix M of the projective transformation which sends U_1 to C, U to O, U_2 to A, and O to B:

$$M = \begin{pmatrix} -u_0 & 0 & t_0^2 \\ 0 & s_0 R & -t_0^2 R \\ u_0 & -s_0 R & 0 \end{pmatrix}.$$

The inverse of M is (projectively equivalent to)

$$\begin{pmatrix} s_0 t_0^2 R^2 & s_0 t_0^2 R & s_0 t_0^2 R \\ t_0^2 u_0 R & t_0^2 u_0 & t_0^2 u_0 R \\ s_0 u_0 R & s_0 u_0 R & s_0 u_0 R \end{pmatrix}.$$

A point $[\beta_1, \beta_2, 1]$ in the square transforms to $[u_0(1 - \beta_1), s_0 R(\beta_2 - 1), t_0^2(\beta_1 - R\beta_2)]$. From this we obtain formulas for c and d in terms of β_1 and β_2:

$$c = \frac{u_0(1 - \beta_1)}{t_0^2(\beta_1 - R\beta_2)} \text{ and } d = \frac{s_0 R(\beta_2 - 1)}{t_0^2(\beta_1 - R\beta_2)}.$$

From equations (4), (5), and (6) we then obtain

$$e = \frac{1 - R - 2\beta_1 + 2R\beta_2}{t_0(\beta_1 - R\beta_2)}, \; a = \frac{-R\beta_2}{s_0(\beta_1 - R\beta_2)}, \text{ and } b = \frac{\beta_1}{u_0(\beta_1 - R\beta_2)}.$$

Because these all have the common denominator $\beta_1 - R\beta_2$, we can cancel it. (Note we no longer have the normalization condition $as_0 + bu_0 = 1$ after this cancellation.) Then if we multiply by $s_0 t_0^2 u_0$ we obtain the formulas stated in the theorem. □

ACKNOWLEDGMENTS

The authors wish to thank Thomas Berry for, among other things, informing them about the cubic invariants in [18].

REFERENCES

1. Bajaj, C. L. Surface fitting using implicit algebraic surface patches. In *Topics in Surface Modeling*, H. Hagen, Ed., SIAM, 1992.
2. Bajaj, C. L., Hoffmann, C. M., Hopcroft, J. E., and Lynch, R. E. Tracing surface intersections. *Comput. Aided Geom. Des. 5* (1988), 285–307.
3. Bajaj, C. L., and Xu, G. Piecewise rational approximation of real algebraic curves. Comput. Sci. Tech. Rep. CAPO 91-19, Purdue Univ., 1991.
4. Chandler, R. E. A tracking algorithm for implicitly defined curves. *IEEE Comput. Graph. Appl. 8* (1988), 83–89.
5. Dahmen, W. Smooth piecewise quadratic surfaces. In *Mathematical Methods in Computer Aided Geometric Design*, G. Farin, Ed., Academic Press, Boston, 1989.
6. Dobkin, D. P., Levy, S. V. F., Thurston, W. P., and Wilks, A. R. Contour tracing by piecewise linear approximations. *ACM Trans. Graph. 9*, (1990), 389–423.
7. Farin, G. *Curves and Surfaces for Computer Aided Geometric Design: A Practical Guide*. Academic Press, Boston, 1988.

8. GARRITY, T., AND WARREN, J. Geometric continuity. *Comput. Aided Geom. Des. 8* (1991), 51–65.
9. HOFFMANN, C. M., AND HOPCROFT, J. E. Automatic surface generation in computer aided design. *Visual Comput. 1* (1985), 95–100.
10. HOFFMANN, C. M., AND HOPCROFT, J. E. The potential method for blending surfaces and curves. *Geometric Modeling: Algorithms and New Trends*, G. Farin, Ed., SIAM, Philadelphia, 1987.
11. HOSCHEK, J., AND HARTMANN, E. Functional splines for interpolation, approximation, blending of curves, surfaces and solids. In *Topics in Surface Modeling*, H. Hagen, Ed., SIAM, 1992.
12. LI, J., HOSCHEK, J., AND HARTMANN, E. G^{n-1}-functional splines for interpolation and approximation of curves, surfaces and solids. *Comput. Aided Geom. Des. 7* (1990), 209–220.
13. LODHA, S., AND WARREN, J. Bézier representation for cubic surface patches. *Comput. Aided Des.* To appear.
14. LODHA, S., AND WARREN, J. Bézier representation for quadric surface patches. *Comput. Aided Des. 22* (1990), 574–579.
15. PATTERSON, R. R. Parametric cubics as algebraic curves. *Comput. Aided Geom. Des. 5* (1988), 139–159.
16. PATTERSON, R. R. Parametrizing and graphing nonsingular cubic curves. *Comput. Aided Des. 20* (1988), 615–623.
17. PENNA, M. A., AND PATTERSON, R. R. *Projective Geometry and its Applications to Computer Graphics*. Prentice-Hall, Englewood Cliffs, NJ, 1985.
18. SALMON, G. *Higher Plane Curves*. Hodges, Foster and Figgis, Dublin, 1879.
19. SEDERBERG, T. W. Algebraic geometry for surfaces and solid modeling. *Geometric Modeling: Algorithms and New Trends*, G. Farin, Ed., SIAM, Philadelphia, 1987.
20. SEDERBERG, T. W. Algorithm for algebraic curve intersection. *Comput. Aided Des. 21* (1989), 547–554.
21. SEDERBERG, T. W. Piecewise algebraic surface patches. *Comput. Aided Geom. Des. 2* (1985), 53–59.
22. SEDERBERG, T. W. Planar piecewise algebraic curves. *Comput. Aided Geom. Des. 1* (1984), 241–255.
23. WARREN, J. Blending algebraic surfaces. *ACM Trans. Graph. 8* (1989), 263–278.
24. WARREN, J. Free form blending: A technique for creating piecewise implicit surfaces. In *Topics in Surface Modeling*, H. Hagen, Ed., SIAM, 1992.
25. WARREN, J. Several notions of geometric continuity for implicit plane curves. Monografías de la Academia de Ciencias de Zaragoza, Comunicaciones presentadas en el Instituto de Estudios Avanzados de la OTAN sobre "Computación de Curvas y Superficies", Puerto de la Cruz (Tenerife, España, 1989) 2 (1990).

Received January 1991; revised May 1992; accepted September 1992

Editor: Anthony DeRose