On zero curves of bivariate polynomials

Michael S. Floater
Departamento de Matemática Aplicada, Universidad de Zaragoza, Spain

January 1995, revised November 1995

Abstract. Two conditions on the signs of the coefficients of a bivariate polynomial which ensure that the zero set is a single curve are derived. The first condition demands that all but one of the coefficients have the same sign. The second requires that the signs are 'split' by any straight line. Both properties are demonstrated by generalizing the set of isoparametric lines to a certain two-parameter family of curves. Equivalent curves are found for power, tensor-product Bernstein, exponential and triangular Bernstein polynomials. The second property will allow greater freedom when using algebraic curves for geometric modelling.

§1. Introduction

With the emergence of algebraic curves and surfaces in geometric modelling [2,4,6,11,12, 21,22] it is important to be able to predict how many connected components the zero set of a multivariate function has in terms of its coefficients. It would be especially useful to find a condition which ensures that the zero set is a single curve or surface.

For univariate polynomials Descartes' Rule of Signs [24] bounds the number of zeros by the number of sign changes in the sequence of coefficients. For Bernstein polynomials Descartes' rule usually goes under the name 'variation-diminishing' [13]. Further information can be gleaned from the Budan-Fourier theorem, used for example in [16], and from Sturm's sequences [24] which can be used to get an exact count of the zeros and to separate them.

For bivariate polynomials no such comprehensive theories exist and it is known from numerical experiments that the zeros can take on complex configurations. Karlin [18] has pointed out the lack of a satisfactory concept of total-positivity for multivariate functions and total positivity has been used to derive many variation-diminishing properties.

There are situations where the zeros are predictable. For example, if the control net $\hat{f}$ of a Bernstein-Bézier triangle f is convex, then so is f itself [9]. This would imply that the zeros of f were either non-existent or formed those parts of some convex curve which intersected the triangle. If $\hat{f}$ is monotonically increasing in some direction $\mathbf{d}$, then f is also monotonically increasing in the same direction [17]. So in this case the zero set, if non-empty, must consist of an open curve, transversal to $\mathbf{d}$. To a more limited extent, similar properties could be derived for the zeros of tensor-product Bernstein polynomials using convexity and monotonicity [10,14,17].

However, we would prefer to have simpler conditions than these, conditions which are in terms only of the signs of the coefficients, similar to the univariate case.

In this paper two simple conditions are presented which, barring degenerate cases, ensure that the zero set has one or at most one connected component, that component usually being a smooth curve. Typically, f will have the form $f(x,y) = \sum_i \sum_j a_{i,j}\phi_i(x)\phi_j(y)$, where the ϕ_i are the basis functions, the $a_{i,j}$ are the coefficients, and f will be defined on some subset Ω of $\mathbb{R}^2$. We decompose Ω into the three sets

$$f_- = f^{-1}((-\infty,0)), \qquad f_0 = f^{-1}(\{0\}), \qquad f_+ = f^{-1}((0,\infty)). \tag{1}$$

The following two properties of f will be derived for certain bases.

(P1) If at most one of the coefficients of f is negative, then the set f_- on which f is negative is either empty or simply connected.

(P2) If the matrix of coefficients $a_{i,j}$, regarded as forming a uniform grid, can be divided by any straight line, such that those coefficients on one side of the line are negative and those on the other side are positive, then the zero set f_0 of f is a single analytic curve. In short, if the signs of the coefficients are split by a straight line, then f_0 is a single curve.

By considering $-f$, it is clear from (P1), that also if at most one of the coefficients of f is positive, then the set f_+ is either empty or simply connected. In (P1), if all the coefficients are non-zero, one can deduce that f_0 can only be empty, a point, or a single non-intersecting curve. Moreover if the negative coefficient is in the interior of the grid, this curve will be closed. Otherwise it will start and finish on the boundary of the domain (when the domain is infinite in some direction, this needs to be interpreted accordingly). We regard (P1) as mainly of theoretical interest but it could lead to a more general property which bounds the number of connected components of f_- in terms of the number, and possibly position, of the negative coefficients.

The special case of (P2) in which the straight line is horizontal or vertical is certainly well-known in geometric modelling. A similar special case for Bernstein-Bézier triangles, when the line is parallel to one of the sides, was proved by Bajaj and Xu [5] and used to develop a scheme for modelling with piecewise algebraic curves, called A-splines [6]. It was also generalized by Bajaj, Chen and Xu [3,4] in two ways to Bernstein polynomials on tetrahedra. However (P2) in general provides more freedom in the choice of the signs of the coefficients while still ensuring that there is only one curve in the domain of interest.

Properties (P1) and (P2) are derived from two things: (i) a certain two-parameter family of curves in the parameter domain which generalizes the set of isoparametric lines, and (ii) Descartes' Rule of Signs for *generalized* univariate polynomials, polynomials having non-integer exponents. Indeed we use Descartes' rule for generalized univariate polynomials in order to deduce something about *ordinary* multivariate polynomials. The curves (i) may be an important tool in the study of zeros of multivariate polynomials.

Equivalent families of curves and properties (P1) and (P2) are found for power, exponential, and Bernstein tensor-product polynomials and for triangular Bernstein-Bézier triangles.

This paper is organized as follows. Section 2 covers Descartes' Rule of Signs for generalized polynomials and some simple consequences. In Sections 3 and 4 the two-parameter family of curves is defined and used to prove (P1) and (P2) for bivariate polynomials with power bases. Corresponding curves and properties are then developed for each of the bases: exponential, tensor-product Bernstein, and triangular Bernstein-Bézier in Sections 5,6,7. We finish in Section 8 with some remarks about what can be deduced when f has several negative coefficients, similar to (P1).

§2. Preliminaries

The subsequent development of properties of zeros of bivariate polynomials will require two special cases of Descartes' Rule of Signs for generalized univariate polynomials. Let $g : (0, \infty) \to \mathbb{R}$ be the generalized polynomial defined by $g(x) = \sum_{k=1}^{r} g_k x^{p_k}$, where $p_1 < p_2 < \cdots < p_r$ is any increasing sequence of real numbers.

Generalized Descartes' Rule of Signs. *The number of positive roots of the equation* $g(x) = 0$ *does not exceed the number of changes of signs in the sequence* $g_1, g_2, \dots, g_r$.

The proof, using properties of the Wronskian of the basis functions x^{p_k} can be found in [7]. We now derive from the Rule of Signs two almost self-evident consequences which will constitute the building blocks for the variation-diminishing properties derived later for bivariate polynomials.

Lemma 1. *Suppose that for some* $q \in I\!R$, *the coefficients of* g *are such that either* $g_k \leq 0$ *for* $k < q$ *and* $g_k \geq 0$ *for* $k > q$ *or* $g_k \geq 0$ *for* $k < q$ *and* $g_k \leq 0$ *for* $k > q$. *Suppose further that at least one coefficient is positive and at least one negative. Then* g *has exactly one zero,* $x_0 > 0$, *say and* $g'(x_0) \neq 0$.

This is simply Descartes' Rule of Signs when the number of sign changes is exactly one. The second lemma concerns the case when the coefficients have two sign changes but limited to when they are consecutive. We define g_-, g_0, g_+ analogously to f_-, f_0, f_+.

Lemma 2. *Suppose that at most one of the coefficients* g_k *is negative. Then if the set* g_- *is non-empty, it is connected, i.e. it is an open interval.*

Proof. We know that g has at most two roots. If there are none or one, g_- is empty or connected, respectively. Otherwise there are two roots. Then $g > 0$ for small enough $x > 0$ and for large enough x, and $g < 0$ between the roots. $\triangleleft$

We shall generalize the two properties above in a certain way to (ordinary) bivariate polynomials. Initially we will concentrate on the power basis functions. Let $\Omega = (0, \infty) \times (0, \infty)$ and define $f : \Omega \to I\!R$ as

$$f(x, y) = \sum_{i=0}^{m} \sum_{j=0}^{n} a_{i,j} x^i y^j. \tag{2}$$

§3. One negative (or positive) coefficient

Consider the isoparametric line corresponding to

$$y = c, \tag{3}$$

where c is some positive constant. If we define $g : (0, \infty) \to I\!R$ as f evaluated along the curve, i.e. $g(x) = f(x, c)$ then

$$g(x) = \sum_{i=0}^{m} \sum_{j=0}^{n} a_{i,j} x^i c^j = \sum_{i=0}^{m} g_i x^i, \qquad g_i = \sum_{j=0}^{n} a_{i,j} c^j.$$

Now suppose at most one coefficient $a_{i,j}$ is negative. Then, since $c > 0$, it is clear that at most one coefficient g_i is negative. Then we may apply Descartes' Rule of Signs (Lemma 2) to g, which in this case is an ordinary polynomial, and deduce that the set g_- is either empty or an open interval. Thus the set f_- intersects the line $\{y = c\}$ in at most one interval. Clearly a similar property is valid for isoparametric lines defined by

$$x = c, \tag{4}$$

and we summarize these two properties as a lemma.

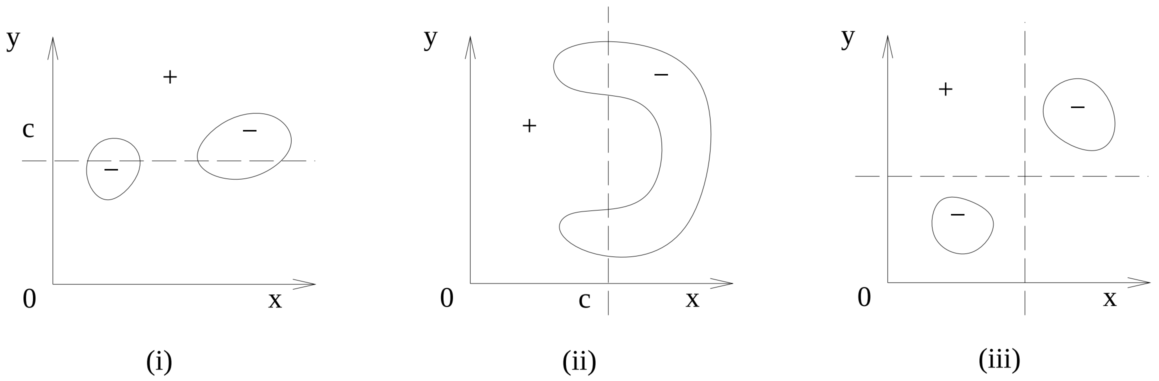

Fig. 1. Impossible configurations of the components of f_-.

Lemma 3. *Suppose only one of the coefficients $a_{i,j}$ is negative. Then the intersection set of f_- with any isoparametric line is either empty or connected (an open line segment).*

Lemma 3 means that if at most one coefficient $a_{i,j}$ is negative, the zeros of f cannot take on the configurations of (i) and (ii) in Figure 1. However it is not in itself sufficient to prevent the situation in (iii) occurring. It will be shown that the latter situation is impossible by extending the family of isoparametric curves sufficiently that through any two points in Ω, there is a curve having the property that its intersection with f_- is connected. This will imply that the set f_- is path-wise connected. Consider the curves

$$y = y(x) = \alpha x^\beta, \qquad \alpha, \beta \in \mathbb{R}, \qquad \alpha > 0, \beta \neq 0. \tag{5}$$

Lemma 4. *Suppose only one of the coefficients $a_{i,j}$ is negative. Then the intersection set of f_- with any curve of the form $y = \alpha x^\beta$, where $\alpha > 0$, $\beta \neq 0$, is either empty or connected.*

Proof. If one defines $g(x) = f(x, \alpha x^\beta)$, one finds that

$$g(x) = \sum_{i=0}^{m} \sum_{j=0}^{n} a_{i,j} \alpha^j x^{i+\beta j}.$$

Then collect common terms (there may be many if β is 1, -1, 2, or $-\frac{1}{2}$, for example, but none if β is transcendental) and get

$$g(x) = \sum_{k=1}^{r} g_k x^{p_k},$$

for some increasing sequence of powers $p_1 < p_2 < \cdots < p_r$, and $r \leq (m+1)(n+1)$. Now since each $a_{i,j}$ contributes to only one g_k, only one, at most, of the g_k can be negative. Then we may apply Descartes' Rule of Signs to the basis $\{x^{p_k}\}$ (Lemma 2), and deduce that g_- is either empty or connected, i.e. either empty or an open interval. It immediately follows that the intersection of f_- with the curve $y = \alpha x^\beta$, is either empty or connected. $\triangleleft$

Applying the whole family of curves defined by (3–5) we deduce the following, namely (P1).

Proposition 5. *Suppose only one coefficient $a_{i,j}$ is negative. Then the set f_- is either empty or consists of one simply connected component.*

Proof. Suppose that $f_- \neq \emptyset$. We first show that f_- is connected. Let $\mathbf{x}_1 = (x_1, y_1)$ and $\mathbf{x}_2 = (x_2, y_2)$ be points in f_-. So $f(\mathbf{x}_1) < 0$ and $f(\mathbf{x}_2) < 0$. First of all, if $y_1 = y_2 = c$, for some c then consider the isoparametric line $\{y = c\}$. By Lemma 3, $f(x, y) < 0$ for all $x \in (x_0, x_3)$ for some x_0, x_3, with $0 \leq x_0 < x_1 < x_2 \leq x_3 \leq \infty$. Consequently the straight line segment

$$(1 - \lambda)\mathbf{x}_1 + \lambda\mathbf{x}_2, \qquad \lambda \in [0, 1],$$

is a path joining $\mathbf{x}_1$ and $\mathbf{x}_2$ lying inside f_-. Similarly, if $x_1 = x_2$, the vertical straight line segment joining the points is a path inside f_-.

Now suppose $x_1 \neq x_2$ and $y_1 \neq y_2$. One can find uniquely α and β such that the curve $y = \alpha x^\beta$ interpolates $\mathbf{x}_1$ and $\mathbf{x}_2$. Indeed, solving

$$y_1 = \alpha x_1^\beta, \qquad y_2 = \alpha x_2^\beta,$$

one finds the unique solution

$$\alpha = \exp\left(\frac{\ln x_2 \ln y_1 - \ln x_1 \ln y_2}{\ln x_2 - \ln x_1}\right) > 0, \qquad \beta = \frac{\ln y_2 - \ln y_1}{\ln x_2 - \ln x_1} \neq 0.$$

Note that the sign of β is the same as that of $(x_2 - x_1)(y_2 - y_1)$. By Lemma 4, and the fact that $g(x_1) < 0$, $g(x_2) < 0$, we must have that $g(x) < 0$ for all $x \in (x_1, x_2)$. Therefore $(x, y(x))$ for $x \in [x_1, x_2]$ is a path lying entirely inside f_-, joining $\mathbf{x}_1$ to $\mathbf{x}_2$. Since the points $\mathbf{x}_1$ and $\mathbf{x}_2$ were chosen arbitrarily, it follows that f_- is connected [23].

In order to show that f_- is simply connected, we suppose f_- to be non-empty and having a hole, in the sense of the Jordan Curve Theorem [1]. Then there exists a simple closed curve $\Gamma \subset f_-$ and at least one point $\mathbf{x}_1 \notin f_-$ lying inside Γ. Then consider the isoparametric line $\{y = y_1\}$. This line must cross Γ at least twice. Moreover, it must cross it at least once at a point (x_0, y_1), $x_0 < x_1$ and at least once at a point (x_2, y_1), $x_2 > x_1$. Therefore $f(x_0, y_1) < 0$, $f(x_1, y_1) \geq 0$, $f(x_2, y_1) < 0$, which contradicts Lemma 3. $\triangleleft$

Observations.

(a) If in Proposition 5 the negative coefficient is an internal one ($1 \leq i \leq m-1$, $1 \leq j \leq n-1$) and if all remaining coefficients are positive then f_0 is one of the following:
  (i) empty,
  (ii) a single point,
  (iii) a simple closed curve.

(b) If many of the coefficients are zero, f_0 can have two connected components. For example the zeros in Ω of $f(x, y) = x - 3xy + 2xy^2$ are the lines $y = 1$ and $y = 2$. The set f_- can also be arbitrarily close to a curve of the form $y = \alpha x^k$. For example, $f(x, y) = (y - x^2)^2 - \epsilon x^2 y = y^2 - (2 + \epsilon)x^2 y + x^4$ has one negative coefficient, and as $\epsilon \longrightarrow 0$, f_- converges towards the curve $y = x^2$. In the limit it is empty.

To see (a), suppose first that f_- is empty. Then if f_0 contains two points they can be interpolated by one of the curves (3–5) and since f cannot be identically zero along it, it must be negative between the two points. This contradicts the fact that f_- is empty. Alternatively, suppose that f_- is non-empty. Then by hypothesis it is bounded away from the boundary of Ω and we claim that f_0 consists of the boundary of f_-. Indeed, if f_0 contained, in addition, some point lying outside the closure of f_- then we could join it to any point on the boundary of f_- with one of the curves (3–5). Now f would be negative along this curve between the two points, contradicting the fact that f is non-negative at all points outside f_-.

§4. Splitting the signs of the coefficients by straight lines

Having presented the class of curves (3–5), we now use them to give a second property of 'variation-diminishing' type. Suppose now that for some $k \in 1, \dots, m-1$, that either the condition

$$\begin{aligned} a_{i,j} &\le 0, \qquad \text{for } i < k, \\ a_{i,j} &\ge 0, \qquad \text{for } i > k, \end{aligned} \tag{6}$$

or the condition

$$\begin{aligned} a_{i,j} &\ge 0, \qquad \text{for } i < k, \\ a_{i,j} &\le 0, \qquad \text{for } i > k, \end{aligned} \tag{7}$$

is placed on the coefficients. We will further suppose, to avoid trivialities that at least one of the coefficients is negative and at least one is positive. Note that the signs of the coefficients $a_{k,j}$ are arbitrary. Consider the isoparametric line (3). It was found before that if $g(x) = f(x, c)$, then

$$g(x) = \sum_{i=0}^{m} g_i x^i, \qquad \text{where} \qquad g_i = \sum_{j=0}^{n} a_{i,j} c^j.$$

So from (6), it follows that $g_i \le 0$ for $i < k$ and $g_i \ge 0$ for $i > k$. And the opposite is true for (7). Thus there is one sign change in $\{g_i\}$ and Lemma 1 shows that g has exactly one zero. Since it depends on c, we call it $x(c)$. This defines a function $x : (0, \infty) \to \mathbb{R}$. Since the curves (3) cover the whole domain $(0, \infty) \times (0, \infty)$, the only zeros of f are of the form $(x(c), c)$.

In fact f_0 is an analytic curve. We know that $g'(x(c)) \ne 0$, i.e. $\frac{\partial}{\partial x} f(x(c), c) \ne 0$ for each $c > 0$. Moreover f is an analytic function. So by the Implicit Function Theorem [20] the zeros of f in a neighbourhood of $(x(c), c)$ form an analytic curve. Since this curve must coincide with $(x(c), c)$ it follows that the whole curve $(x(c), c)$ is analytic. The condition is easy to define diagrammatically. Figure 2 shows an example of (6). If we arrange the coefficients in a rectangular grid, the property (6) says that all coefficients on one side of some vertical line have one sign or are zero while all coefficients on the other side have the other sign or are zero. If the line passes exactly through one column of coefficients, those signs are arbitrary.

By an analogous argument, one can show that f_0 is an analytic curve if there exists an $l \in \{1, \dots, n-1\}$ for which

$$\begin{aligned} a_{i,j} &\le 0, \qquad \text{for } j < l, \\ a_{i,j} &\ge 0, \qquad \text{for } j > l, \end{aligned} \tag{8}$$

or

$$\begin{aligned} a_{i,j} &\ge 0, \qquad \text{for } j < l, \\ a_{i,j} &\le 0, \qquad \text{for } j > l. \end{aligned} \tag{9}$$

We summarize as follows.

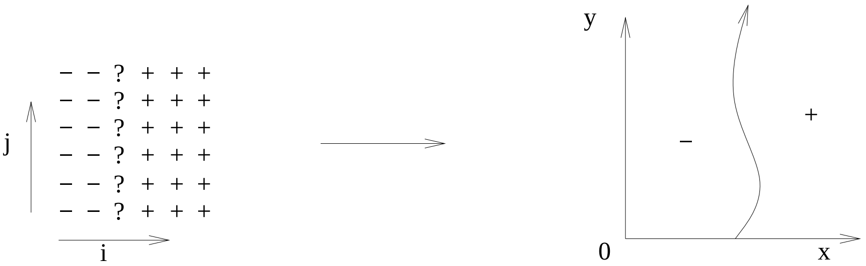

Fig. 2. Splitting the signs of the coefficients with a vertical line.

Proposition 6. *If either property (6) or (7) holds for some k, then f_0 is an open analytic curve transversal to the direction of the x-axis. If either (8) or (9) holds for some l, then f_0 is an open analytic curve transversal to the direction of the y-axis.*

By applying the curves (5), we may generalize Proposition 6 to (P2), a property involving any straight line which splits the signs of the coefficients of f; see Figure 3.

Proposition 7. *Suppose $\exists\beta \in I\!R$, $\beta \neq 0$ and $\gamma \in I\!R$ for which either*

$$\begin{aligned} a_{i,j} &\leq 0, \qquad \text{for } i+\beta j < \gamma, \\ a_{i,j} &\geq 0, \qquad \text{for } i+\beta j > \gamma, \end{aligned} \tag{10}$$

or

$$\begin{aligned} a_{i,j} &\geq 0, \qquad \text{for } i+\beta j < \gamma, \\ a_{i,j} &\leq 0, \qquad \text{for } i+\beta j > \gamma. \end{aligned} \tag{11}$$

Suppose further that there is at least one negative and one positive coefficient. Then f_0 consists of one analytic curve of the form $(x(\alpha), \alpha x^\beta(\alpha))$, parametrized by α, and is transversal at every point to the curve of the form $(x, \alpha x^\beta)$ which passes through that point.

Proof. For any $\alpha > 0$, define the curve $\mathbf{p}_\alpha(x) = (x, \alpha x^\beta)$ and consider $g_\alpha(x) = f(\mathbf{p}_\alpha(x))$. Then

$$g_\alpha(x) = \sum_{i=0}^{m}\sum_{j=0}^{n} a_{i,j}\alpha^j x^{i+\beta j} = \sum_{k=1}^{r} g_k x^{p_k},$$

for some increasing sequence $p_1 < p_2 < \cdots < p_r$, and $r \leq (m+1)(n+1)$. Now by hypothesis, either $g_k \leq 0$ for $p_k < \gamma$ and $g_k \geq 0$ for $p_k > \gamma$ or $g_k \geq 0$ for $p_k < \gamma$ and $g_k \leq 0$ for $p_k > \gamma$. In either case, Lemma 1 applies and one finds that g_α has exactly one zero $x(\alpha)$ say. Also $g'_\alpha(x(\alpha)) \neq 0$ and so, by the chain rule,

$$\nabla f \,.\, \mathbf{p}'_\alpha \neq 0,$$

at $x = x(\alpha)$. Then the Implicit Function Theorem applies and shows that the curve $\mathbf{q}(\alpha) = (x(\alpha), \alpha x^\beta(\alpha))$ is analytic and crosses each $\mathbf{p}_\alpha$ transversally. By varying α, the curves $\mathbf{p}_\alpha$ cover the whole domain and so the curve $\mathbf{q}$ contains all the zeros of f. $\triangleleft$

Figure 3 shows examples of f satisfying the hypothesis of Proposition 7. In all cases in Figure 2 and 3, f_0 is a single curve, dividing Ω into two connected subsets f_- and f_+. The splitting of signs is reminiscent of Newton's diagram [19] used in the study of double points of algebraic curves. In a Newton diagram all coefficients to one side of a line are zero.

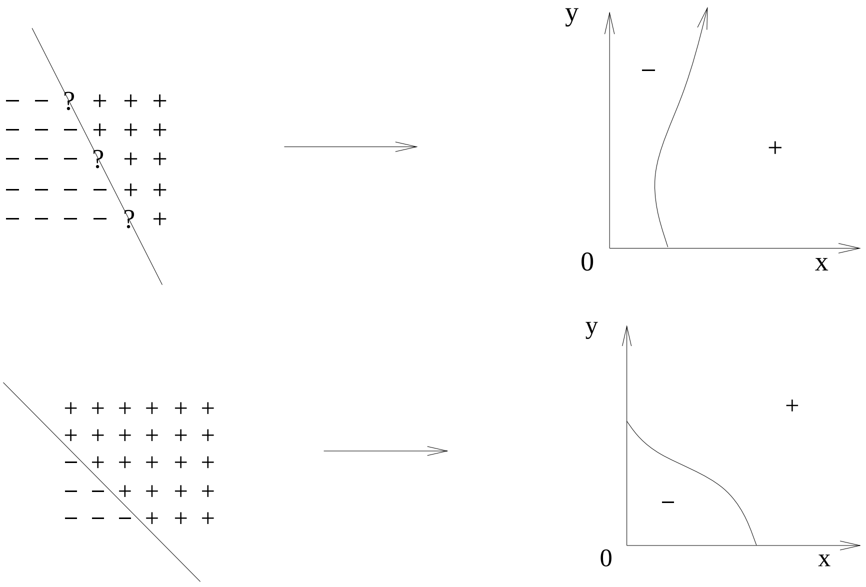

Fig. 3. Splitting the signs of the coefficients with straight lines.

§5. Exponential polynomials

The log functions encountered in the proof of Proposition 5 suggest that it might be a good idea to make the transformation of variables $x = e^X$, $y = e^Y$. Substituting these into (2) and letting $\tilde{f}(X, Y) = f(x, y)$, we find that

$$\tilde{f}(X, Y) = \sum_{i=0}^{m} \sum_{j=0}^{n} a_{i,j} e^{iX+jY}, \tag{12}$$

an exponential polynomial. It can easily be proved that $\tilde{f}$, as a function defined on the whole of $\mathbb{R}^2$ satisfies (P1) and (P2). They can be derived in an analogous way by noticing that the curves (3–5) are transformed into

$$Y = \tilde{c}, \qquad X = \tilde{c}, \qquad \text{and} \qquad Y = \beta X + \lambda, \quad \beta \neq 0,$$

where $\tilde{c} = \ln(c)$ and $\lambda = \ln(\alpha)$. These are now the set of all straight lines in $\mathbb{R}^2$. However it is simpler to notice that the topologies of $\tilde{f}_-$, $\tilde{f}_0$, $\tilde{f}_+$ are the same as the topologies of f_-, f_0, f_+. The sets $\tilde{f}_-$, $\tilde{f}_0$, $\tilde{f}_+$ decompose the domain of $\tilde{f}$ which is $\mathbb{R}^2$. Indeed, let $\phi : \Omega \to \mathbb{R}^2$ be the transformation $\phi(x, y) = (\ln(x), \ln(y))$. Then $(X, Y) = \phi(x, y)$. It follows that $\tilde{f}_- = \phi(f_-)$, and $\tilde{f}_0 = \phi(f_0)$. Since ϕ is an isomorphism [23], this establishes (P1) and (P2).

There is a third way of proving (P1) which uses convexity and is more in the spirit of the proof of Descartes' Rule of Signs. Indeed, if $a_{k,l}$ is the negative coefficient, we can define

$$\tilde{h}(X, Y) = e^{-kX-lY} \tilde{f}(X, Y) = \sum_{i=0}^{m} \sum_{j=0}^{n} a_{i,j} e^{(i-k)X+(j-l)Y}$$

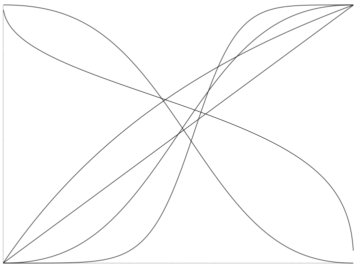

Fig. 4. Some of the curves in the tensor-product Bernstein case.

and we find, that for any λ_1, λ_2 in $\mathbb{R}$,

$$\begin{aligned}
\lambda_1^2\tilde{h}_{XX}(X,Y) &+ 2\lambda_1\lambda_2\tilde{h}_{XY}(X,Y) + \lambda_1^2\tilde{h}_{YY}(X,Y) \\
&= \sum_{i=0}^{m}\sum_{j=0}^{n}(\lambda_1(i-k)+\lambda_2(j-l))^2 a_{i,j}e^{(i-k)X+(j-l)Y} \\
&= \sum_{i=0}^{m}\sum_{\substack{j=0 \\ (i,j)\neq(k,l)}}^{n}(\lambda_1(i-k)+\lambda_2(j-l))^2 a_{i,j}e^{(i-k)X+(j-l)Y} \geq 0.
\end{aligned}$$

This means that $\tilde{h}$ is convex and it follows that $\tilde{h}_-$ is either empty or convex. Since $\tilde{f}_- = \tilde{h}_-$, $\tilde{f}_-$ is also either empty or convex, and in particular simply connected.

§6. Bernstein polynomials

Again we can transform variables in order to derive (P1) and (P2) for tensor-product Bernstein polynomials. Let $x = X/(1-X)$ and $y = Y/(1-Y)$ and $F(X,Y) = (1-X)^m(1-Y)^n f(x,y)$. Then

$$F(X,Y) = \sum_{i=0}^{m}\sum_{j=0}^{n} b_{i,j}\binom{m}{i}X^i(1-X)^{m-i}\binom{n}{j}Y^j(1-Y)^{n-j},$$

where $b_{i,j} = a_{i,j}/\binom{m}{i}/\binom{n}{j}$. If $\phi : (0,\infty)\times(0,\infty) \to (0,1)\times(0,1)$ is the transformation $\phi(x,y) = (x/(1+x), y/(1+y))$ then $F_- = \phi(f_-)$. So (P1) holds also for F, where $a_{i,j}$ is replaced by $b_{i,j}$. Similarly (P2) follows in a similar way.

Notice that the family of curves (3–5) is transformed to

$$Y = C, \qquad X = C, \qquad \text{and} \qquad Y = \frac{\alpha X^\beta}{(1-X)^\beta + \alpha X^\beta},$$

where $C = c/(1+c)$. The third type of curve is monotonically increasing in X when $\beta > 0$ and decreasing when $\beta < 0$. Some of these curves are depicted in Figure 4.

§7. Bernstein polynomials on triangles

In this section we derive (P1) and (P2) for Bernstein-Bézier triangles. Here (P2) just has to be interpreted a little differently; the grid of coefficients $a_{i,j}$ is now triangular rather than rectangular.

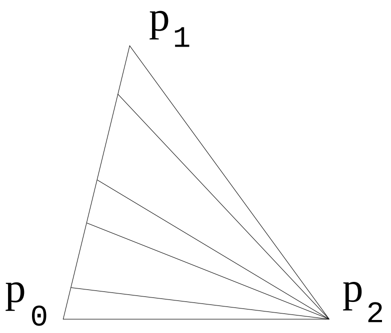

Fig. 5. Isoparametric lines radiating from $\mathbf{p}_2$.

It is straightforward to follow the same arguments used previously once we have found the two-parameter family of curves. Let $\mathbf{p}_0$, $\mathbf{p}_1$, $\mathbf{p}_2$ be any three affinely independent points in $\mathbb{R}^2$ and let Ω be the (open) interior of the triangle having these three points as vertices. We may define a polynomial $f : \Omega \to \mathbb{R}$ of degree n as

$$f(\mathbf{x}) = \sum_{|\mathbf{i}|=n} a_{\mathbf{i}} \binom{n}{\mathbf{i}} \tau_0^{i_0} \tau_1^{i_1} \tau_2^{i_2}, \qquad \binom{n}{\mathbf{i}} = \frac{n!}{i_0! i_1! i_2!},$$

where the τ_i are the barycentric coordinates of $\mathbf{x}$ with respect to $\triangle \mathbf{p}_0\mathbf{p}_1\mathbf{p}_2$, i.e. $\tau_0+\tau_1+\tau_2 = 1$ and $\mathbf{x} = \tau_0\mathbf{p}_0 + \tau_1\mathbf{p}_1 + \tau_2\mathbf{p}_2$. Any point in Ω has strictly positive barycentric coordinates.

The lines in the triangle which play the same role as isoparametric lines for tensor-product polynomials are given by the equations

$$\tau_0/\tau_1 = c, \qquad \tau_1/\tau_2 = c, \qquad \tau_2/\tau_0 = c. \tag{13}$$

Various examples of the first type of line are shown in Figure 5. If we parametrize the first type, for example, as

$$\Gamma(t) = (1-t)\left(\frac{c}{(1+c)}\mathbf{p}_0 + \frac{1}{(1+c)}\mathbf{p}_1\right) + t\mathbf{p}_2,$$

then $f(\Gamma(t))$ becomes a polynomial of degree n in t, in Bernstein form [13]:

$$g(t) = f(\Gamma(t)) = \sum_{|\mathbf{i}|=n} a_{\mathbf{i}} \binom{n}{\mathbf{i}} \left(\frac{c}{1+c}\right)^{i_0} \left(\frac{1}{1+c}\right)^{i_1} t^{i_2}(1-t)^{i_0+i_1} = \sum_{i_2=0}^{n} g_{i_2} \binom{n}{i_2} t^{i_2}(1-t)^{n-i_2}.$$

Thus if, for example, only one of the coefficients $a_{\mathbf{i}}$ is negative, then at most one of the g_{i_2} is negative and then the intersection of f_- with the line Γ is either empty or a single line segment. Also if, for example, $a_{\mathbf{i}} \leq 0$ for $i_2 < q$ while $a_{\mathbf{i}} \geq 0$ for $i_2 > q$, for any $q \in (0, n)$, then g has exactly one zero (provided there is at least one negative and one positive coefficient). Varying $c \in (0, \infty)$, and considering all such lines Γ, we deduce that the zero set f_0 of f, is a single analytic curve passing from side $\mathbf{p}_0\mathbf{p}_2$ to side $\mathbf{p}_1\mathbf{p}_2$ as depicted in Figure 6. This latter property was proved by Bajaj and Xu [5]. It was used in [15] to determine precisely

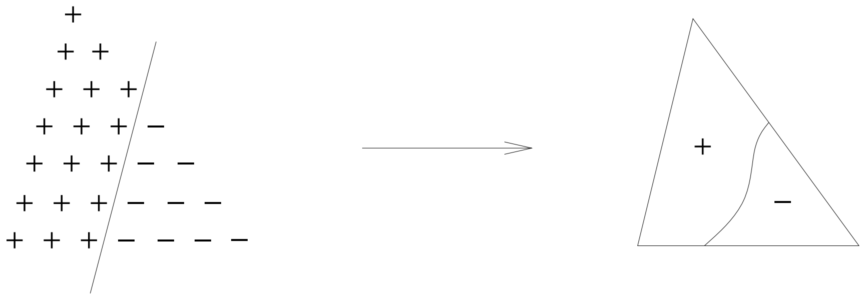

Fig. 6. f_0 is a single curve.

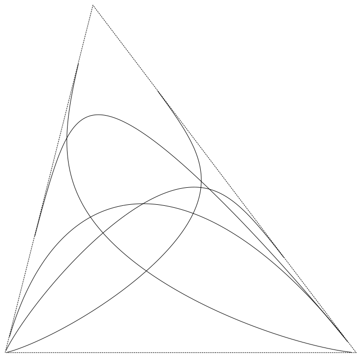

Fig. 7. Some of the curves in the triangular Bernstein case.

when the implicit form of a certain rational cubic Bézier curve in a triangle has no more zeros curves than the cubic curve itself.

In order to demonstrate (P1) and (P2) it is necessary to generalize Γ to a richer family of curves. These are easiest to define when they are divided into three categories. We introduce, the curve $\Gamma : [0,1] \to \mathbb{R}^2$ defined as

$$\Gamma(t) = \frac{(1-t)\mathbf{q}_0 + \alpha t^\beta (1-t)^{1-\beta}\mathbf{q}_1 + t\mathbf{q}_2}{(1-t) + \alpha t^\beta (1-t)^{1-\beta} + t}.$$

The points $(\mathbf{q}_0, \mathbf{q}_1, \mathbf{q}_2)$ are any of the three permutations $(\mathbf{p}_0, \mathbf{p}_1, \mathbf{p}_2)$, $(\mathbf{p}_1, \mathbf{p}_2, \mathbf{p}_0)$, $(\mathbf{p}_2, \mathbf{p}_0, \mathbf{p}_1)$. Also $\alpha > 0$ and $\beta \in (0,1)$. The curve Γ begins at $\mathbf{q}_0$ and ends at $\mathbf{q}_2$. Examples of Γ are displayed in Figure 7.

Now let

$$W(t) = (1-t) + \alpha t^\beta (1-t)^{1-\beta} + t,$$

and consider the case where $\mathbf{q}_i = \mathbf{p}_i$. Substituting Γ into f one finds that

$$\begin{aligned} f(\Gamma(t)) &= \frac{1}{W^n(t)} \sum_{|\mathbf{i}|=n} a_\mathbf{i} \binom{n}{\mathbf{i}} \alpha^{i_1} t^{i_2+\beta i_1} (1-t)^{(1-\beta)i_1+i_0} \\ &= \frac{1}{W^n(t)} \sum_{|\mathbf{i}|=n} a_\mathbf{i} \binom{n}{\mathbf{i}} \alpha^{i_1} t^{i_2+\beta i_1} (1-t)^{n-(i_2+\beta i_1)}. \end{aligned}$$

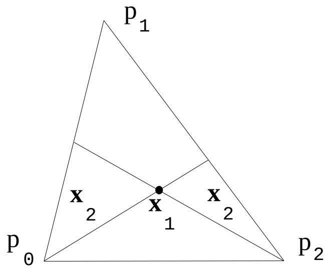

Fig. 8. One of the three cases of the position of $\mathbf{x}_2$ relative to $\mathbf{x}_1$.

Further, making the transformation of variables $t = x/(1+x)$, we find that

$$f(\Gamma(t)) = \frac{1}{W^n(t)} \frac{1}{(1+x)^n} \sum_{|\mathbf{i}|=n} a_{\mathbf{i}} \binom{n}{\mathbf{i}} \alpha^{i_1} x^{i_2+\beta i_1}.$$

This is a positive multiple of a generalized polynomial which plays the same role as g in Sections 3 and 4. Explicitly, if only one of the coefficients $a_{\mathbf{i}}$ is negative then g_- is empty of connected where $g(t) = f(\Gamma(t))$, $t \in (0,1)$. So in order to derive (P1) it remains to show that through any two points in the triangle, there is a curve of the form Γ interpolating them. Indeed, let $\mathbf{x}_1$ and $\mathbf{x}_2$ be points in the triangle and let ϕ_0, ϕ_1, ϕ_2, be the barycentric coordinates of $\mathbf{x}_1$ and ψ_0, ψ_1, ψ_2, be the barycentric coordinates of $\mathbf{x}_2$. If either

$$\frac{\phi_0}{\psi_0} < \frac{\phi_1}{\psi_1} < \frac{\phi_2}{\psi_2} \qquad \text{or} \qquad \frac{\phi_0}{\psi_0} > \frac{\phi_1}{\psi_1} > \frac{\phi_2}{\psi_2}, \tag{14}$$

then we set $\mathbf{q}_i = \mathbf{p}_i$. Using standard facts about barycentric coordinates (14) can be seen to be equivalent to $\mathbf{x}_2$ lying in one of the subtriangles shown in Figure 8. The other four cases are handled by the other two permutations of the indices (i_0, i_1, i_2). Using the implicit form of Γ:

$$\tau_1 - \alpha \tau_0^{1-\beta} \tau_2^{\beta} = 0,$$

we solve for α and β:

$$\beta = \ln\left(\frac{\phi_1 \psi_0}{\psi_1 \phi_0}\right) \Big/ \ln\left(\frac{\phi_2 \psi_0}{\psi_2 \phi_0}\right),$$

and

$$\alpha = \phi_1 \phi_0^{(\beta-1)} \phi_2^{-\beta},$$

with $\alpha > 0$ and from (14), $0 < \beta < 1$. By forming the ratio of ϕ_0 to ϕ_2 we find that $\mathbf{x}_1 = \Gamma(t_1)$ and $\mathbf{x}_2 = \Gamma(t_2)$ where

$$t_1 = \phi_2/(\phi_0 + \phi_2),$$

and

$$t_2 = \psi_2/(\psi_0 + \psi_2).$$

This establishes (P1); see Figure 9.

We have already shown (P2) in the case when the straight line is parallel to one of the sides of the triangle. We considered the case when the straight line is $i_2 = q$, crossing the

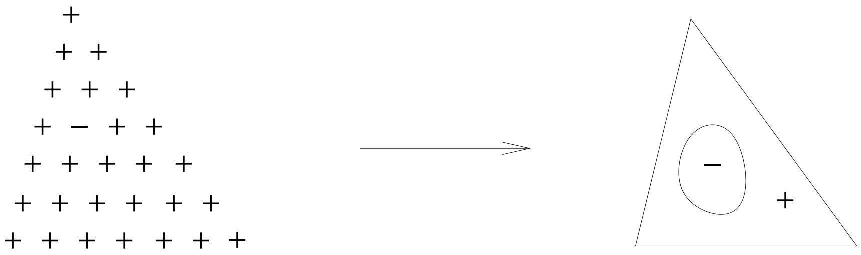

Fig. 9. f_- is empty or has one simply connected component.

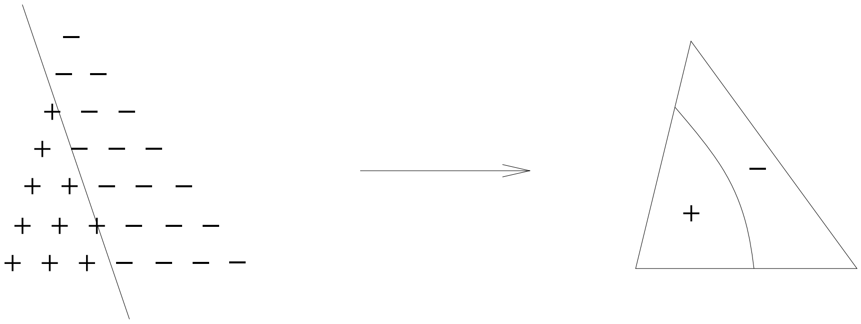

Fig. 10. f_0 is a single curve.

sides $\mathbf{p}_0\mathbf{p}_2$ and $\mathbf{p}_1\mathbf{p}_2$. In order to show that (P2) generalizes to any straight line, we split all possible lines into three categories and treat only one. Consider any line whose equation is $i_2 + \beta i_1 = \gamma$, for any $\beta \in (0,1)$. Such a line has an unsigned direction lying in a double cone, bounded by the unsigned directions $\mathbf{p}_1 - \mathbf{p}_0$ and $\mathbf{p}_1 - \mathbf{p}_2$. In this case we consider Γ with $\mathbf{q}_i = \mathbf{p}_i$. Straight lines whose unsigned directions lie in the remaining two double cones, can be treated by the remaining two permutations of indices. Now, considering the form of $f(\Gamma(t))$ it merely remains to apply Lemma 2 to deduce that if $a_{\mathbf{i}} \leq 0$ for $i_2 + \beta i_1 < \gamma$ and $a_{\mathbf{i}} \geq 0$ for $i_2 + \beta i_1 > \gamma$ then f_0 consists of a single curve as in Figure 10. This establishes (P2).

Finally it may be interesting to note that Γ can be any conic section which passes through two of the vertices tangentially to the sides. Let $\beta = \frac{1}{2}$, and reparametrize Γ with respect to $s \in [0,1]$, where $s^2/(1-s)^2 = t/(1-t)$. Then multiplying numerator and denominator of $\Gamma(t)$ by $(1-s)^2/(1-t)$, one obtains

$$\Gamma(t) = \frac{(1-s)^2\mathbf{q}_0 + \alpha s(1-s)\mathbf{q}_1 + s^2\mathbf{q}_2}{(1-s)^2 + \alpha s(1-s) + s^2},$$

a rational quadratic Bézier curve. The sign of $\alpha - 2$ determines the type of conic [8]. For each α, f has at most one zero along this conic when the signs of the coefficients of f are split by any line parallel to the line passing through $\mathbf{q}_1$ and $(\mathbf{q}_0 + \mathbf{q}_2)/2$. For example the equation for the line passing through $\mathbf{p}_1$ and $(\mathbf{p}_0 + \mathbf{p}_2)/2$ can be written as $i_0 = i_2$. Since $i_1 = 1 - i_0 - i_2$ this is equivalent to the equation $i_2 + \frac{1}{2}i_1 = \frac{1}{2}$.

§8. Final Remarks

Can one bound the number of connected components of f_- in terms of the number of negative coefficients of f? This is unclear but numerical examples show that when two of the coefficients of f are negative, f_- can have three connected components. However, using yet again the family of curves (3–5), we can say the following about (2).

Proposition 8. *If at most two of the coefficients $a_{i,j}$ are negative, then every connected component of f_- is simply connected.*

Proof. Suppose, to obtain a contradiction, that f_- contains a multiply connected component. Then there exists a simple closed curve $\Gamma \subset f_-$ and a point $\mathbf{x}_1 = (x_1, y_1) \notin f_-$ inside Γ. It may be assumed that there are exactly two negative coefficients. First, suppose that they are a_{i,j_1}, a_{i,j_2}. Then along any isoparametric line $\{y = c\}$, $g(x) = f(x,c)$ has the property that g_- is connected since a_{i,j_1}, a_{i,j_2} contribute to the same coefficient of g, namely g_i. But when $c = y_1$, the line $y = c$ must cross Γ on both sides of $x = x_1$ which is a contradiction.

A similar argument eliminates the case when the two negative coefficients have their second index in common. The remaining case is when the two negative coefficients are a_{i_1,j_1} and a_{i_2,j_2} with $i_1 \neq i_2$, $j_1 \neq j_2$. Now define

$$\beta = (i_1 - i_2)/(j_2 - j_1), \qquad \alpha = y_1/x_1^\beta,$$

and consider the curve $y = \alpha x^\beta$. It clearly passes through $\mathbf{x}_1$ and must cross Γ on either side. If $g(x) = f(x, \alpha x^\beta)$, then g has at most one negative coefficient since a_{i_1,j_1} and a_{i_2,j_2} contribute to the same basis function x^{p_k} where $p_k = i_1 + \beta j_1 = i_2 + \beta j_2$. Lemma 2 shows that g_- is therefore connected and this is again a contradiction. $\lhd$

It is also possible to obtain further statements which exclude multiply connected components. For example, if exactly three coefficients are negative then at most one connected component of f_- can be multiply connected. Also if f_- does have such a component then no other component can lie inside one of its holes. However such statements seem to have less and less value as one increases the number of negative coefficients.

It would be more interesting if it could be shown that when f has two negative coefficients which are *adjacent* then f_- had at most two connected components (numerical examples can be constructed for which there are two components in this case). Adjacency would need to made concrete. For example the two negative coefficients might be $a_{i,j}$ and $a_{i+1,j}$.

Finally, are there families of curves like those considered in this paper for tensor-product splines [8]?

§9. References

1. Ahlfors L. V., *Complex analysis*, New York, McGraw-Hill, 1953.
2. Bajaj C., Some applications of constructive real algebraic geometry, in *Algebraic geometry and its applications, C. Bajaj (ed.)*, Springer-Verlag, New York, 1994, 303–405.
3. Bajaj C., J. Chen, G. Xu, Modeling with cubic A-patches, Comp. Sci. Report, CAPO-93-02, Purdue University, 1993.
4. Bajaj C., J. Chen, G. Xu, Free form surface design with A-patches, in *Proc. of Graphics Interface 94, Canadian Information Processing Society*, Vancouver, Canada, 1994.
5. Bajaj C., G. Xu, A-splines: local interpolation and approximation using C^k-continuous piecewise real algebraic curves, Comp. Sci. Report CAPO-92-44, Purdue University, 1992.

6. Bajaj C., G. Xu, Data fitting with cubic A-splines, *Proc. of Computer Graphics International, CGI94*, Melbourne, Australia, 1994.
7. Berezin I. S., N. P. Zhidkov, *Computing methods, Vol II*, Pergamon Press, Oxford, 1965.
8. de Boor C., *A Practical Guide to Splines*, Springer-Verlag, New York, 1978.
9. Chang G., P. J. Davis, The convexity of Bernstein polynomials over triangles, J. Approx. Theory **40** (1984), 11–28.
10. Dahmen W., Convexity and Bernstein-Bézier polynomials, *Curves and Surfaces, P. J. Laurent, A. Le Méhauté, and L. L. Schumaker (eds.), Academic Press, Boston* (1991), 107–134.
11. Dahmen W., Smooth piecewise quadric surfaces, in *Mathematical Methods in Computer-Aided Geom. Design , T. Lyche & L. L. Schumaker (eds.), Academic Press, Boston* (1989), 181–193.
12. Dahmen W., T. Thamm-Schaar, Cubicoids: modelling and visualization, Computer-Aided Geom. Design **10** (1993), 89–108.
13. Farin G., *Curves and surfaces for computer aided geometric design*, Academic Press, San Diego, 1988.
14. Floater M. S., A weak condition for the convexity of tensor-product Bézier and B-spline surfaces, Advances in Comp. Math. **2** (1994), 67–80.
15. Floater M. S., Rational cubic implicitization, in *Mathematical Methods for Curves and Surfaces, M. Dæhlen, T. Lyche and L. Schumaker (eds.), Vanderbilt University Press, Nashville* (1996), 151–159.
16. Goodman T. N. T., Properties of β-splines, J. of Approx. Theory **44** (1985), 132–153.
17. Goodman T. N. T., Shape preserving representations, in *Mathematical Methods in Computer-Aided Geom. Design , T. Lyche & L. L. Schumaker (eds.), Academic Press, Boston* (1989), 333–351.
18. Karlin S., *Total positivity*, Stanford University Press, Stanford, 1968.
19. Primrose E. J. F., *Plane algebraic curves*, Macmillan, London, 1955.
20. Rudin W., *Principles of mathematical analysis*, McGraw-Hill, New York, 1976.
21. Sederberg T. W., Planar piecewise algebraic curves, Computer-Aided Geom. Design **1** (1984), 241–255.
22. Sederberg T. W., Piecewise algebraic patches, Computer-Aided Geom. Design **2** (1985), 53–59.
23. Simmons G. F., *Introduction to topology and modern analysis*, McGraw-Hill, Tokyo, 1963.
24. Uspensky J. V., *Theory of equations*, McGraw-Hill, New York, 1948.