# An Introduction to the Use of Splines in Computer Graphics

*Richard H. Bartels*
*John C. Beatty*

Computer Graphics Laboratory
Department of Computer Science
University of Waterloo
Waterloo, Ontario, Canada N2L 3G1
(519) 888-4534

*Brian A. Barsky*

Berkeley Computer Graphics Laboratory
Computer Science Division
University of California
Berkeley, California 94720 USA
(415) 642-9838

University of Waterloo TR CS-83-09
UC Berkeley TR UCB/CSD 83/136
Revised May 1985

## ABSTRACT

Piecewise linear approximations to curves and surfaces have many disadvantages which representational techniques based on B-splines or Beta-splines alleviate. Unfortunately the existing literature on B-splines in particular is more suitable for the numerical analyst than for the computer scientist. We survey the mathematics and use of these techniques, in the context of computer graphics, for the design and manipulation of freeform curves and surfaces and with an emphasis on developing the reader's intuition.

We begin with a brief discussion of cubic spline interpolation that leads naturally to a treatment of uniform cubic B-splines. These provide a convenient forum in which to introduce many of the ideas needed subsequently.

We then present a development of the general B-splines, of order $k$ over an arbitrary (*i.e.* non-uniform) knot sequence, as a divided difference of the one-sided power function. Starting from the one-sided power function makes the continuity properties of the B-splines intelligible. The divided difference operator is then applied to the one-sided power function in order to symbolically pre-cancel high order terms so as to avoid cancellation error and overflow, resulting in the computationally satisfactory recursive definition of a B-spline.

The use of general B-splines specifically in the context of graphical applications is also discussed to show how the flexibility inherent to them can be used to control the shape of parametric curves and surfaces. Of particular interest is the Oslo algorithm, which follows naturally from our development of the B-splines, and (among other things) can be used to refine the user's control over part of a curve or surface to an arbitrarily small region for the purpose of introducing fine detail.

Finally we introduce the Beta-splines. They provide an intuitive means of controlling the continuity loss which results when multiple knots underlie a B-spline. We concentrate on cubic Beta-splines, and show how replacing parametric continuity with geometric continuity makes available two shape parameters which control the tension in a curve or surface.

© 1985 by the University of Waterloo and the University of California. This work was supported by funds from the Natural Sciences and Engineering Research Council of Canada (Grants A3022, A4076 & G0651), the National Research Council of Canada (Contract O5SU.31155-1-3103), the Province of Ontario, the University of Waterloo, the United States National Science Foundation (Grant ECS-82-04381), and the Defense Advanced Research Projects Agency of the United States (Contract N00039-82-C-0235).

# Contents

# 1. Introduction

The most basic output primitives in every computer graphics library are "LineSegment()" and "Polygon()," or their equivalent. These are, of course, sufficient in the sense that any curved line or surface can be arbitrarily well approximated by straight line segments or planar polygons, but in many contexts that is not enough. Such approximations often require large amounts of data in order to obtain satisfactory smoothness, and are awkward to manipulate. Then too, even with the most sophisticated continuous shading models, polygonal techniques generally result in visually objectionable images: Mach bands are apparent at the borders between adjacent polygons, and there is always a telltale jaggedness to polygonal silhouettes. Hence many modelling systems are augmented by circles, spheres, cylinders, *etc.*, and allow such simple primitives to be combined to form quite complex objects.

There is a substantial class of curves and surfaces, however, that do not display the sort of regularity that makes such modelling convenient. For these, systems using primitives which can themselves be irregularly curved are more natural. Broadly speaking these are based either on the interpolation or on the approximation of points called *control vertices* that are supplied by the user. In either case a curve is defined by piecing together a succession of *curve segments*, and a surface is defined by stitching together a mosaic of *surface patches*; such a *piecewise* approach is taken for reasons of flexibility and generality.

Interpolatory approaches, most commonly based on cubic splines, are perhaps the more easily understood by users, but they are less readily suited to real-time manipulations. We shall see that they lack some of the desirable properties possessed by most approximation techniques, although in any case they can provide a convenient means of generating an initial curve or surface from which one can compute other representations based on approximation.

Early work by Coons [Coons64, Coons67] and Bézier [Bézier70, Bézier77] introduced the use of nonlinear parametric polynomial representations for the segments and patches from which we assemble piecewise curves and surfaces. Although the techniques they introduced are still in use, we will emphasize the B-splines introduced to computer graphics more recently by Riesenfeld [Riesenfeld73, Gordon74] and discussed extensively in [deBoor78] and [Schumaker81] because they best combine flexibility, power and generality. Moreover, they are less widely understood.

Parametric B-spline curves have many advantages. Among them is the ability to control the degree of continuity at the joints between adjacent curve segments, and at the borders between surface patches, independent of the order of the segments or the number of control vertices being approximated. However, the notion of parametric first or second degree continuity at joints does not always correspond to intuition or to a physically desired effect. For piecewise cubic curves and bicubic surfaces these parametric continuity constraints can be replaced by the more meaningful requirements of continuous unit tangent and curvature vectors. Doing so introduces certain constrained discontinuities in the first and second parametric derivatives which are expressed in terms of "bias" and "tension" parameters called $\beta 1$ and $\beta 2$ in [Barsky81, Barsky85]. These provide additional means of locally controlling shape and give rise to Beta-spline curves and surfaces, which we shall also discuss.

Our intent is to survey the spline techniques that seem most useful in computer graphics, developing the theory in such a way as to emphasize the development of intuition and understanding, and to encourage the use of these techniques by computer scientists. Although we are particularly interested in the modelling of three dimensional freeform surfaces, the mathematics and algorithms of interest are best understood in the context of two dimensional curves, which we can then generalize to define surfaces. Because we are interested in being able to design such curves and surfaces as well as in being able to represent them, we will also consider techniques for displaying them.

### 1.1. General References

For a general introduction to computer graphics which mentions splines briefly see [Foley82] or [Newman73]. Foley and van Dam discuss Hermite, cubic Be'zier and cubic B-spline curves and surfaces. They use matrix notation for compactness. Newman and Sproull discuss Be'zier and uniform B-spline curves. In both cases the treatment is short and prescriptive, as befits introductory texts.

A general survey of curve and surface representations may be found in [Rogers76]. A succinct development is generally provided, together with an example or two. The text was not typeset, and as a result the book is often hard to read. Program code is supplied in many cases, but unfortunately in BASIC.

[Faux79] is an excellent and more recent text containing a good treatment of basic analytic and differential geometry and introductions to most of the standard curve and surface techniques, although the treatment of B-splines is again rather summary.

[deBoor78] is an excellent and comprehensive treatment of B-splines from the numerical analyst's point of view. Much of it is also accessible to anyone with a decent mathematical background willing to read carefully. All of the basic algorithms are presented in Fortran, and are available on magnetic tape from International Math and Statistics Library, Inc., in Houston, Texas. [Schumaker81] provides an excellent and extensive treatment of splines in general, but is accessible only to those with an advanced grounding in mathematical analysis.

## 2. Preliminaries

It is usually convenient to represent a two-dimensional curve parametrically as

$$\mathbf{Q}(\bar{u}) = ( X(\bar{u}), Y(\bar{u}) )$$

where $X(\bar{u})$ and $Y(\bar{u})$ are each single-valued functions of the parameter $\bar{u}$ which yield the $x$- and $y$-coordinates, respectively, of a point on the curve in question for any value of $\bar{u}$.

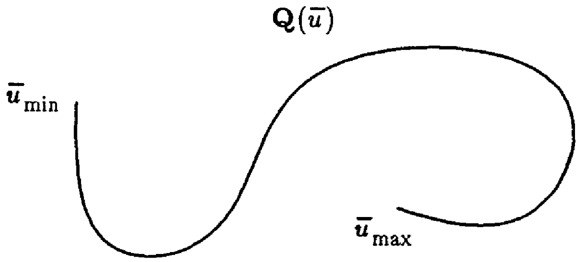


Figure 1. A parametrically defined curve.

Although polynomials are computationally efficient and easy to work with, it is not usually possible to define a satisfactory curve using single polynomials for $X(\bar{u})$ and $Y(\bar{u})$. Instead it is customary to break the curve into some number $m$ of pieces called *segments*, each defined by separate polynomials, and hook the segments together to form a *piecewise polynomial* curve. The parameter $\bar{u}$ then varies between some initial minimum value $\bar{u}_{\text{min}}$ and some final maximum value $\bar{u}_{\text{max}}$ as we move along the curve; the values of $\bar{u}$ that correspond to *joints* between segments are called *knots*. The sequence of knot values is required to be nondecreasing, so that

$$\bar{u}_{\text{min}} = \bar{u}_0 \leq \bar{u}_1 \leq \cdots \leq \bar{u}_m = \bar{u}_{\text{max}} .$$

The sequence of knot values is usually written

$$\bar{u}_0 , \bar{u}_1 , \cdots , \bar{u}_{m-1} , \bar{u}_m$$

and is called either a *knot sequence* or a *knot vector*.

Thus the parametric functions $X(\bar{u})$ and $Y(\bar{u})$ are each composed of $m$ polynomial pieces, the first covering the interval of $\bar{u}$ ranging from $\bar{u}_0$ to $\bar{u}_1$, the second covering values from $\bar{u}_1$ to $\bar{u}_2$, and so on. Usually $X(\bar{u})$ and $Y(\bar{u})$ are required to satisfy some continuity constraints at the joints between successive polynomial segments; if the $0^{\text{th}}$ through $d^{\text{th}}$ derivatives are everywhere continuous (in particular, at the joints), then $X$ and $Y$ are said to be $C^d$ continuous. Sometimes we will assume that the knots are a consecutive sequence of integers, say with $\bar{u}_i = i$; this is called a *uniform knot sequence.*

Also, it will usually be simpler to write $X(\bar{u}-\bar{u}_i)$ rather than $X(\bar{u})$ and we shall generally do so; the reparametrization is easily accomplished by substitution. Thus for the $i^{\text{th}}$ segment, which we shall agree runs between $\bar{u}_i$ and $\bar{u}_{i+1}$, we might write $Y_i(u) = u^2$, so that $u=0$ corresponds to the left end of the segment, rather than the more cumbersome

$$Y_i(\bar{u}) = (\bar{u}-\bar{u}_i)^2 = \bar{u}^2 - 2\bar{u}_i\bar{u} + \bar{u}_i^2 \ .$$

To distinguish between these two conventions, we shall write $Y_i(u)$ when we are parametrizing from the left end of the $i^{\text{th}}$ interval ($u = \bar{u}-\bar{u}_i$), and $Y_i(\bar{u})$ when referring to a single parametrization of the entire curve. Thus

$$Y_i(u) = u^2$$

and

$$Y_i(\bar{u}) = (\bar{u}-\bar{u}_i)^2 = \bar{u}^2 - 2\bar{u}_i\bar{u} + \bar{u}_i^2$$

are equivalent, each being a reparametrization of the other.

There are a variety of ways in which to define a specific curve. They can be broadly classified as being based on "interpolation" or on "approximation". In both cases one begins by specifying a sequence of points called *control vertices*, which we will represent in illustrations by a dot "●" or a "+" sign. In the case of interpolation the curve is required to pass through the control vertices in the order specified:

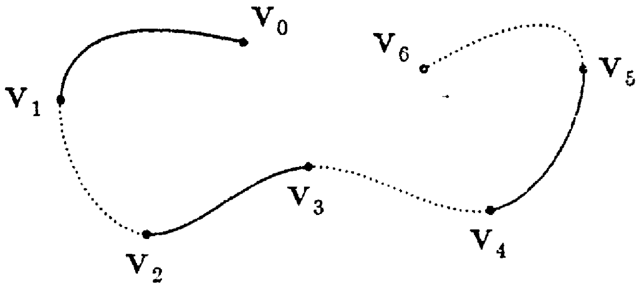


Figure 2. A curve defined by interpolation.

For those techniques based on approximation the curve is required only to pass "near" the control vertices — exactly what "near" means depends on the particular approximation technique used. (See Figure 3.)

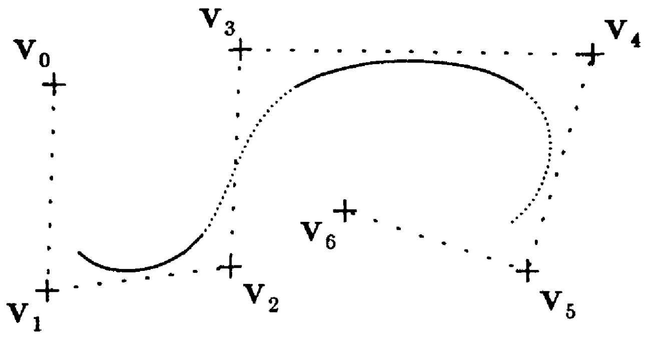


Figure 3. An example of a curve defined by a sequence of *control vertices*, represented here by "+" signs, near which the curve passes. The lightly dotted line connecting the control vertices forms the *control polygon*, and indicates the order in which the control vertices are to be approximated. The solid and heavily dotted curves represent distinct curve *segments*. Each is a single parametric cubic. The point at which two successive segments meet is called a *joint*. The value of the parameter $\bar{u}$ which corresponds to a joint is called a *knot*.

In either case, moving the control vertices alters the curve.

We have connected the control vertices together with lightly dotted lines in Figure 3 to form what is called the *control polygon*. This control polygon indicates the order in which the control vertices are approximated.

# 3. Hermite and Cubic Spline Interpolation

Suppose that we have $m+1$ data points $V_0$, ..., $V_m$ through which we wish to draw a curve such as the following (in which $m=6$).

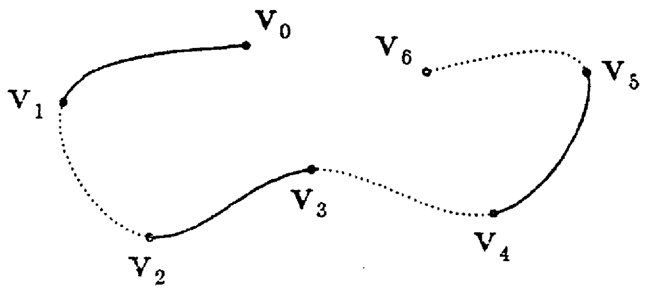


Figure 4. An interpolating cubic spline.

Each successive pair of control vertices is connected by a distinct curve segment. Since each such segment $\mathbf{Q}_i(u)$ is represented parametrically as $(X_i(u),Y_i(u))$, we are really concerned with how the $X_i(u)$ and $Y_i(u)$ are determined by the control vertices

$$\mathbf{V}_i = ( x_i, y_i ) .$$

In general, the $x$-coordinates $X(\overline{u})$ of points on a curve are determined solely by the $x$-coordinates $x_0$, ..., $x_m$ of the control vertices, and similarly for $Y(\overline{u})$. Since both $X(\overline{u})$ and $Y(\overline{u})$ are treated in the same way we shall discuss only $Y(\overline{u})$; indeed, to obtain curves in three dimensions we simply define a $Z(\overline{u})$ as well and let $\mathbf{Q}_i(u)$ be given by $(X_i(u),Y_i(u),Z_i(u))$.

For ease of computation we shall limit ourselves to the use of polynomials in defining $X_i(u)$, $Y_i(u)$ and $Z_i(u)$. Indeed cubic polynomials usually provide sufficient flexibility at reasonable cost. For the above curve, then, $Y(\overline{u})$ is the following.

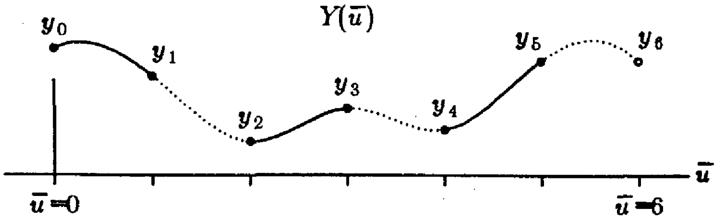


Figure 5. $Y(\bar{u})$ for the curve shown in Figure 4 above. In this example we have rather arbitrarily chosen to use uniform knot spacing, so that the knot sequence is (0,1,2,3,4,5,6).

Each $Y_i(u)$ is a cubic polynomial in the parameter $u$. We know two things in particular about

$$Y_i(u) = a_i + b_i u + c_i u^2 + d_i u^3$$

namely that

$$Y_i(0) = y_i = a_i$$

$$Y_i(1) = y_{i+1} = a_i + b_i + c_i + d_i \ .$$

Notice that, as promised earlier, we have parametrized each $Y_i(u)$ separately so that $u=0$ corresponds to its left end. We have used a uniform knot sequence, so $u=1$ must then correspond to its right end.

Because we have four coefficients to determine, we need two other constraints in order to completely determine a particular $Y_i(u)$. One easy way to do this is to simply pick, arbitrarily, first derivatives $D_i$ of $Y(u)$ at each knot $\bar{u}_i$, so that

$$Y_i^{(1)}(0) = D_i = b_i$$

$$Y_i^{(1)}(1) = D_{i+1} = b_i + 2c_i + 3d_i \ .$$

These four equations can be solved symbolically, once and for all, to yield

$$\begin{aligned} a_i &= y_i \\ b_i &= D_i \\ c_i &= 3(y_{i+1}-y_i) - 2D_i - D_{i+1} \\ d_i &= 2(y_i - y_{i+1}) + D_i + D_{i+1} \ . \end{aligned} \tag{1}$$

Since we use $D_i$ as the derivative at the left end of the $i^{th}$ segment (*i.e.* as $Y_i^{(1)}(0)$) and at the right end of the $(i-1)^{st}$ segment (as $Y_{i-1}^{(1)}(1)$), $Y(u)$ has a continuous first derivative.

This technique is called *Hermite interpolation*. It can be generalized to higher order polynomials. In fact we will subsequently need to perform interpolation by fifth degree polynomials whose first and second derivatives match at the joints between successive segments. The development is entirely analogous to that of cubic Hermite interpolation.

How are the $D_i$ specified? One possibility is to compute the tangents automatically, perhaps by fitting a parabola through $y_{i-1}$, $y_i$, and $y_{i+1}$, and using its derivative at $y_i$ as $D_i$; arbitrary values (such as 0) can be used at the end points [Kochanek82]. Or tangents can be computed as weighted averages of the vector from $y_{i-1}$ to $y_i$ and the vector from $y_i$ to $y_{i+1}$ [Kochanek84]. Or the user may directly specify the derivatives. Since our curves are described parametrically, the user would actually specify an $x$ and a $y$ derivative at each knot, comprising a derivative vector.

It is possible to arrange that successive segments match second as well as first derivatives at joints, using only cubic polynomials. Suppose, as above, that we want to interpolate the $(m+1)$ vertices $V_0$, ..., $V_m$ by such a curve. Each of the $m$ segments $Y_0(u)$, ..., $Y_{m-1}(u)$ is a cubic polynomial determined by four coefficients. Hence we have $4m$ unknown values to determine. At each of the $(m-1)$ interior knots $\bar{u}_1$, ..., $\bar{u}_{m-1}$ (where two segments meet) we have four conditions:

$$\begin{aligned} Y_{i-1}(1) &= y_i \\ Y_i(0) &= y_i \\ Y_{i-1}^{(1)}(1) &= Y_i^{(1)}(0) \\ Y_{i-1}^{(2)}(1) &= Y_i^{(2)}(0) \ . \end{aligned}$$

Since we also require that

$$\begin{aligned} Y_0(0) &= y_0 \\ Y_{m-1}(1) &= y_m \end{aligned}$$

we have a total of $4m-2$ conditions from which to determine our $4m$ unknowns. Thus we need two more conditions. These may be chosen in a variety of ways. A common choice is simply to require that the second derivatives at the endpoints $\bar{u}_0$ and $\bar{u}_m$ both be zero; this yields what is called a *natural cubic spline*. Figure 4 is actually a natural cubic spline.

### 3.1. Practical Considerations - Computing Natural Cubic Splines

We do not need to directly solve $4m$ equations — the problem can be simplified. Notice that a natural cubic spline is actually a special case of Hermite interpolation; we have simply chosen first derivative vectors so as to match second derivatives as well. If we can compute the needed $D_i$, we have already obtained definitions of the $a_i$, $b_i$, $c_i$ and $d_i$ in terms of the $D_i$.

Thus at each internal joint we want to choose $D_i$ so that -

$$Y_{i-1}^{(2)}(1) = Y_i^{(2)}(0)$$

or

$$2c_{i-1} + 6d_{i-1} = 2c_i \ .$$

Substituting in our earlier solutions (1) for $c_{i-1}$, $d_{i-1}$ and $c_i$, we have

$$2\Big[3(y_i - y_{i-1}) - 2D_{i-1} - D_i\Big] + 6\Big[2(y_{i-1} - y_i) + D_{i-1} + D_i\Big]$$
$$= 2\Big[3(y_{i+1} - y_i) - 2D_i - D_{i+1}\Big] \ .$$

Simplifying, and moving the unknowns to the left, we have

$$D_{i-1} + 4D_i + D_{i+1} = 3(y_{i+1} - y_{i-1}) \ . \qquad (2)$$

Since there are $m-1$ internal joints, there are $m-1$ such equations. Requiring that the second derivative at the beginning of the curve be zero implies that

$$2c_0 = 0$$

$$2\Big[3(y_1 - y_0) - 2D_0 - D_1\Big] = 0$$

$$2D_0 + D_1 = 3(y_1 - y_0) \ .$$

Requiring that the second derivative at the end of the curve be zero similarly results in

$$D_{m-1} + 2D_m = 3(y_m - y_{m-1}) \ .$$

We now have $m+1$ equations in $m+1$ unknowns. Representing them in matrix form we have

$$\begin{bmatrix} 2 & 1 & & & & & \\ 1 & 4 & 1 & & & & \\ & 1 & 4 & 1 & & & \\ & & 1 & 4 & 1 & & \\ & & & \cdot & \cdot & \cdot & \cdot \\ & & & & 1 & 4 & 1 \\ & & & & & 1 & 2 \end{bmatrix} \begin{bmatrix} D_0 \\ D_1 \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ D_m \end{bmatrix} = \begin{bmatrix} 3(y_1 - y_0) \\ 3(y_2 - y_0) \\ \cdot \\ \cdot \\ \cdot \\ 3(y_m - y_{m-2}) \\ 3(y_m - y_{m-1}) \end{bmatrix} .$$

Beginning at the top, each row is easily combined with the row below it to yield

$$\begin{bmatrix} 1 & \alpha_0 & & & & \\ & 1 & \alpha_1 & & & \\ & & 1 & \alpha_2 & & \\ & & & \cdot & \cdot & \cdot \\ & & & & \alpha_{m-2} & \\ & & & & 1 & \alpha_{m-1} \\ & & & & & 1 \end{bmatrix} \begin{bmatrix} D_0 \\ D_1 \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ D_m \end{bmatrix} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ \beta_m \end{bmatrix} .$$

("forward elimination"). This directly yields the value of $D_m$, and it is then a simple matter to solve successively for $D_{m-1}$, $D_{m-2}$, ..., $D_1$ and finally $D_0$ ("backward substitution").

The multiplicative factors $s_i$ that accomplish the forward substitution and the corresponding values $\alpha_i$ need only be computed once. The $\beta_i$'s must be computed and the backward substitution performed separately for each coordinate. When a data point $y_j$ is moved the values $\beta_j$, ..., $\beta_m$ must be recomputed and the entire backward substitution again performed.

### 3.2. Other End Conditions For Cubic Interpolating Splines

There are many other ways in which to determine the additional two constraints we need to fully define a $C^2$ continuous interpolating cubic spline. It is generally most useful for these conditions to be applied to the ends of a curve so that they have the most effect there, as for the natural cubic splines, but even so the choice of conditions has some influence over the shape of the entire curve.

For example, instead of fixing the second derivatives at the first and last knot to be zero, we may fix the first derivatives there to be zero.

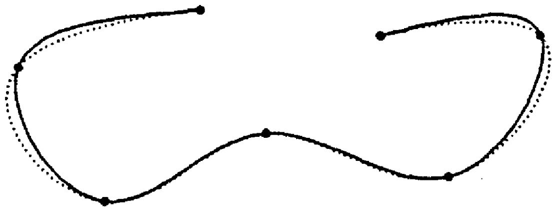


Figure 6. The solid line is a natural cubic interpolating spline; that is, the second parametric derivatives at the ends of the solid curve are zero. For the dotted curve the first derivatives at the ends have been set to zero instead.

Another possibility, which de Boor calls the *not-a-knot* condition [deBoor78], is to require $C^3$ continuity at the second and next-to-last knots $u_1$ and $u_{m-1}$. In effect the first two segments are a single polynomial, as are the last two.

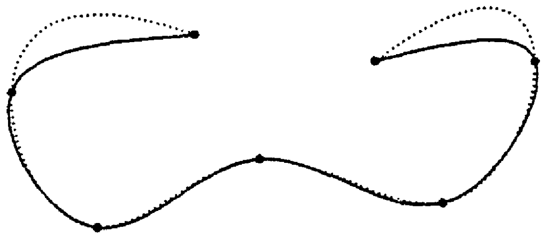


Figure 7. The solid line is a natural cubic interpolating spline. For the dotted curve, $C^3$ continuity has been forced between the first and second segments, and between the last and the next-to-last segments.

Yet another alternative, suggested by Forsythe, Malcolm and Moler [Forsythe77], is to use the third derivatives of the cubic polynomials that interpolate the first and last four control vertices as the (constant) third derivatives of the first and last segments.

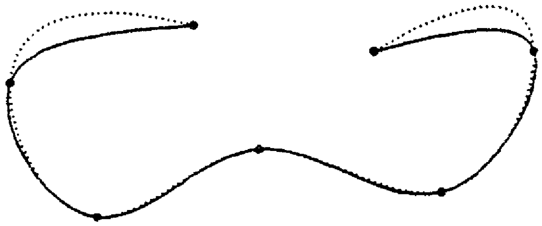


Figure 8. The solid line is a natural cubic interpolating spline. For the dotted curve the third derivative of the polynomial that interpolates the first four control vertices is used as the (constant) third derivative of the first segment, and similarly for the last segment.

Or one might allow the user to explicitly supply any two of the first, second, or third derivative vectors at the ends. In any case, we can construct and solve a set of equations very much as we did for the natural cubic splines. Additional discussion of how this can be done, and algorithms, are given in chapter four of [Forsythe77] and in chapter four of [deBoor78]. For a uniform knot vector, and indeed for any reasonable strictly increasing sequence of knots, these equations are well-conditioned and can be solved easily and accurately.

### 3.3. Knot Spacing

While the end conditions discussed above affect the entire curve, their principal influence is felt at the endpoints. Gross changes to a curve's shape can be made elsewhere without moving the control vertices by varying the knot spacing.

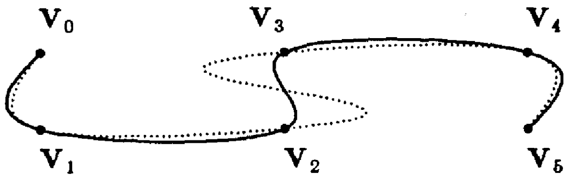


Figure 9. The solid line is a natural cubic interpolating spline in which the knots are spaced a unit apart. Unit knot spacing is used also in the dotted curve except for the parametric interval corresponding to the segment between $V_2$ and $V_3$, for which the knots are spaced four units apart.

With the single exception of Figure 9, we have used a uniform knot sequence in defining the interpolating cubic spline curves discussed above. Thus the knot vector for the solid curve in Figure 9 is

$$0,\ 1,\ 2,\ 3,\ 4,\ 5$$

while the dotted curve interpolates the same data points, but for the knot vector

$$0,\ 1,\ 2,\ 6,\ 7,\ 8\ .$$

Thus knot spacing can be used to control shape; the more difficult question is how that control can be used intuitively.

Uniform knot spacing is one obvious way to define a knot sequence. The Euclidean distance between control vertices is a second natural choice for the length of the parametric interval over which $u$ varies in defining a segment.

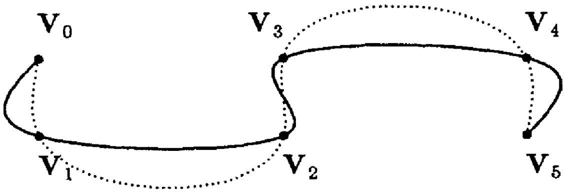


Figure 10. The solid line in the above figure is a natural cubic interpolating spline in which the knots are spaced a unit apart. In the case of the dotted curve, the knots corresponding to two successive control vertices differ in value by the Euclidean distance separating the two control vertices.

### 3.4. Closed Curves

It is also sometimes useful to generate closed curves such as the following.

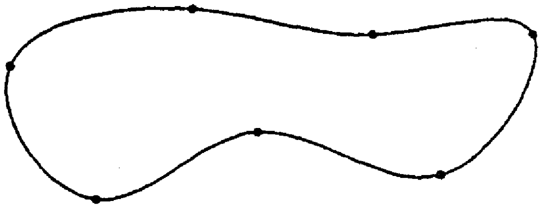


Figure 11. A closed interpolating cubic spline.

In this case equation (2) applies at each of the $m$ vertices, with the caveat that indices must be computed modulo $m+1$. The system of equations that results looks a little different:

$$\begin{bmatrix} 4 & 1 & & & & & 1 \\ 1 & 4 & 1 & & & & \\ & 1 & 4 & 1 & & & \\ & & 1 & 4 & 1 & & \\ & & & \cdot & \cdot & \cdot & \cdot \\ & & & & 1 & 4 & 1 \\ 1 & & & & & 1 & 4 \end{bmatrix} \begin{bmatrix} D_0 \\ D_1 \\ . \\ . \\ . \\ . \\ D_m \end{bmatrix} = \begin{bmatrix} 3(y_1-y_m) \\ 3(y_2-y_0) \\ . \\ . \\ . \\ 3(y_m-y_{m-2}) \\ 3(y_0-y_{m-1}) \end{bmatrix} .$$

Basically one solves this system as one solved for the $D_i$ for an open curve. During forward elimination, however, it is necessary to compute and save nonzero values for entries in the rightmost column and to successively cancel the leftmost nonzero value in the bottom row. The obvious change must also be made to the back substitution process as well.

## 4. A Simple Approximation Technique — Uniform Cubic B-splines

Later we will develop B-spline curves and surfaces in their full generality. In this section we will introduce them by looking at a simpler and particularly useful special class of B-splines called the *uniform cubic B-splines.* As the name implies, we make use of parametric cubic polynomials on a uniform knot sequence.

A particular property of the B-splines is *local control*, by which we mean that altering the position of a single control vertex causes only a part of the curve to change. This makes it possible to modify part of a curve (or surface) without affecting other portions that are already satisfactory, which is often useful in geometric design and modelling. An added benefit of local control is that it minimizes the work required to recompute a curve after a control vertex has been moved since only a small part of the curve has changed.

The way in which local control is obtained is most easily explained by considering first a "piecewise linear" interpolation of the control vertices. Consider the following "piecewise linear curve."

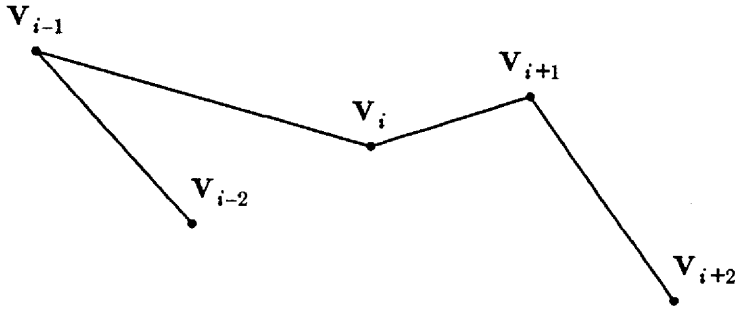


Figure 12. A piecewise linear B-spline.

If we represent the segments of this curve in the obvious way we have

$$\begin{aligned}
\mathbf{Q}_{i-1}(u) &= (\, X_{i-1}(u), Y_{i-1}(u)\,) = (1-u)\mathbf{V}_{i-2} + u\,\mathbf{V}_{i-1} && u = (\bar{u}-\bar{u}_{i-1})/(\bar{u}_i-\bar{u}_{i-1}) \\
\mathbf{Q}_{i}(u) &= (\, X_{i}(u), Y_{i}(u)\,) = (1-u)\mathbf{V}_{i-1} + u\,\mathbf{V}_{i} && u = (\bar{u}-\bar{u}_{i})/(\bar{u}_{i+1}-\bar{u}_{i}) \\
\mathbf{Q}_{i+1}(u) &= (\, X_{i+1}(u), Y_{i+1}(u)\,) = (1-u)\mathbf{V}_{i} + u\,\mathbf{V}_{i+1} && u = (\bar{u}-\bar{u}_{i+1})/(\bar{u}_{i+2}-\bar{u}_{i+1}) \\
\mathbf{Q}_{i+2}(u) &= (\, X_{i+2}(u), Y_{i+2}(u)\,) = (1-u)\mathbf{V}_{i+1} + u\,\mathbf{V}_{i+2} && u = (\bar{u}-\bar{u}_{i+2})/(\bar{u}_{i+3}-\bar{u}_{i+2})
\end{aligned}$$

where

$$\begin{aligned}
X_{i-1}(u) &= (1-u)x_{i-2} + u\,x_{i-1} \\
Y_{i-1}(u) &= (1-u)y_{i-2} + u\,y_{i-1} \\
X_i(u) &= (1-u)x_{i-1} + u\,x_i \\
Y_i(u) &= (1-u)y_{i-1} + u\,y_i \\
X_{i+1}(u) &= (1-u)x_i + u\,x_{i+1} \\
Y_{i+1}(u) &= (1-u)y_i + u\,y_{i+1} \\
X_{i+2}(u) &= (1-u)x_{i+1} + u\,x_{i+2} \\
Y_{i+2}(u) &= (1-u)y_{i+1} + u\,y_{i+2} \ .
\end{aligned}$$

Altering $\mathbf{V}_i$ clearly affects only the two segments $\mathbf{Q}_i(u)$ and $\mathbf{Q}_{i+1}(u)$ adjacent to it: $\mathbf{V}_i$ does not appear in the formulae for any other segments. We can represent our piecewise linear curve so as to isolate the influence of each control vertex and make a generalization to higher order, smoother piecewise polynomial curves straightforward.

If we plot $Y(\bar{u})$ as a function of $\bar{u}$, and represent the contribution of $y_i$ to $Y(\bar{u})$ by a dashed line, we obtain

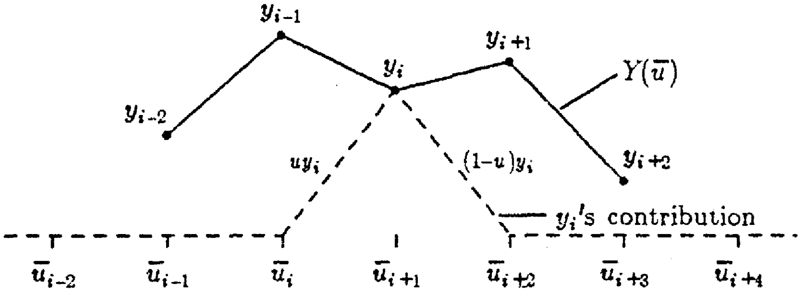


Figure 13. The contribution of $y_i$ to $Y(\bar{u})$. In this example the knots are spaced one unit apart, although that is not essential.

Notice that this contribution is zero both to the left of $\bar{u}_i$ and to the right of $\bar{u}_{i+2}$. Similarly, the contribution of $y_{i-1}$ is

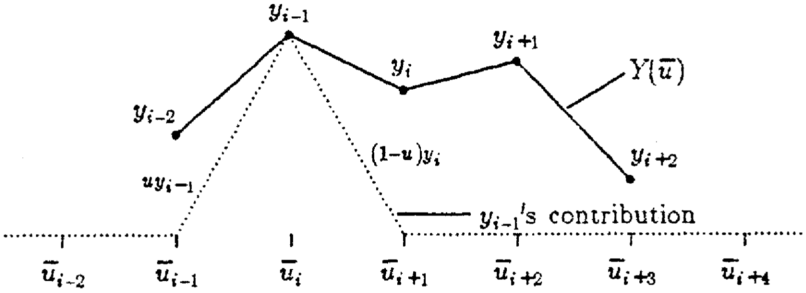


Figure 14. The contribution of $y_{i-1}$ *to* $Y(\bar{u})$.

Plotting these two *hat functions* together gives us a graphical representation of the fact that $Y_i(u) = (1-u)y_{i-1} + u\,y_i$. (For the sake of clarity we will stop extending these hat functions to the left and right by zero when we draw them because these extensions would all be drawn on top of one another).

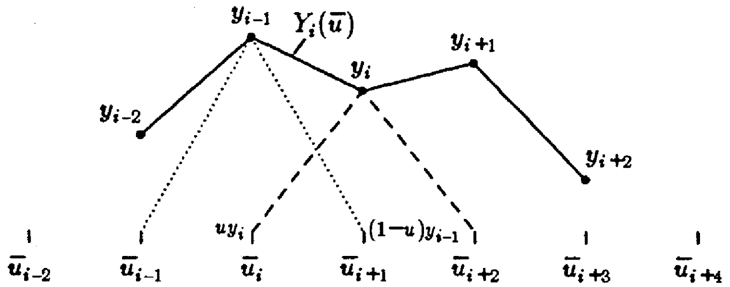


Figure 15. A simultaneous look at the contributions of $y_{i-1}$ and $y_i$ to the curve in general, and to $Y_i(\bar{u})$ in particular.

It is useful to think of $y_{i-1}$ and $y_i$ as each *scaling* a corresponding "unit hat function" (see below) whose maximum height is one, and each of which is a simple translation of the others. It is also useful to think of $y_{i-1}$ and $y_i$ as each *being weighted by* a corresponding unit hat function. As $\bar{u}$ increases from $\bar{u}_{i-1}$, the contribution of $y_{i-1}$ grows from nothing at $\bar{u}=\bar{u}_{i-1}$, peaks at $\bar{u}=\bar{u}_i$, and dies away to nothing again at $\bar{u}=\bar{u}_{i+1}$. A similar thing is true for $y_i$ on the interval $\bar{u}_i \le \bar{u} < \bar{u}_{i+2}$. More profoundly, we have seen that $\mathbf{Q}(\bar{u})$ is entirely determined by $\mathbf{V}_{i-1}$ and $\mathbf{V}_i$ alone in the interval $\bar{u}_i \le \bar{u} < \bar{u}_{i+1}$. In this interval $Y(\bar{u})$ is just a weighted average of $y_{i-1}$ and $y_i$, namely

$$Y_i(\bar{u}) = \left(\frac{\bar{u}_{i+1}-\bar{u}}{\bar{u}_{i+1}-\bar{u}_i}\right) y_{i-1} + \left(\frac{\bar{u}-\bar{u}_i}{\bar{u}_{i+1}-\bar{u}_i}\right) y_i$$

$$= (1-u)y_{i-1} + u\,y_i \quad .$$

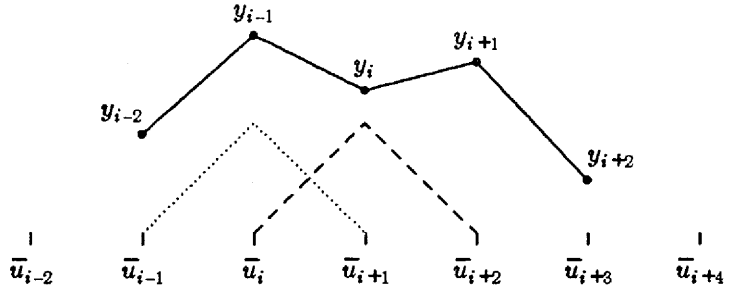


Figure 16. Multiplying the two unit (height one) hat functions shown here by $y_{i-1}$ and $y_i$ yields the scaled hat functions shown in Figure 15.

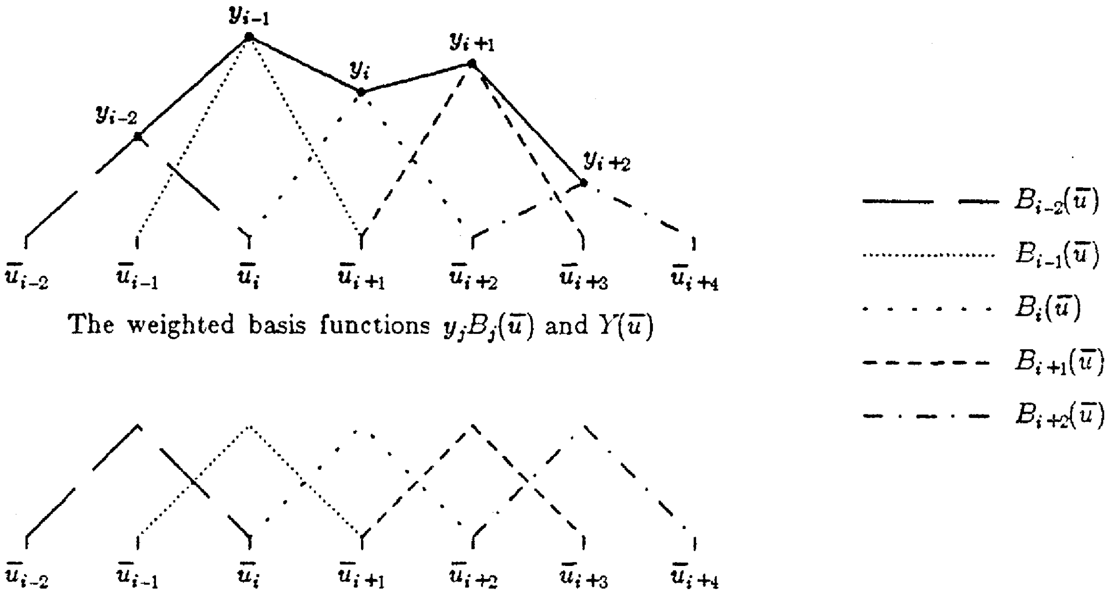
If we call the "dotted" unit hat function $B_{i-1}(\bar{u})$ and the "dashed" unit hat function $B_i(\bar{u})$ (to be compatible with later material we name a hat function after the knot at its left extremity), then the line segment attaching $y_{i-1}$ to $y_i$ may be written as

$$Y_i(\bar{u}) = y_{i-1}B_{i-1}(\bar{u}) + y_i B_i(\bar{u}) \qquad \text{for } \bar{u}_i \le \bar{u} < \bar{u}_{i+1} \tag{3}$$

where

$$B_i(\bar{u}) = \begin{cases} \dfrac{\bar{u}-\bar{u}_i}{\bar{u}_{i+1}-\bar{u}_i} & \bar{u}_i \le \bar{u} < \bar{u}_{i+1} \\ \dfrac{\bar{u}_{i+2}-\bar{u}}{\bar{u}_{i+2}-\bar{u}_{i+1}} & \bar{u}_{i+1} \le \bar{u} < \bar{u}_{i+2} \end{cases} . \tag{4}$$

We can represent the other segments of our piecewise linear curve in the same way; equation (3) is quite general. In the following illustration we show all the hat functions $B_{i-2}(\bar{u})$, $\cdots$, $B_{i+2}(\bar{u})$ that define our example "curve".



The weighted basis functions $y_jB_j(\bar{u})$ and $Y(\bar{u})$

The unweighted basis functions $B_j(\bar{u})$

Figure 17. Representing a piecewise linear curve as a linear combination of hat functions.

The entire curve can now be written as

$$\mathbf{Q}(\bar{u}) = \sum_i \mathbf{V}_i B_i(\bar{u}) = \sum_i \left( x_i B_i(\bar{u}), y_i B_i(\bar{u}) \right) . \tag{5}$$

Depending on the point of view we wish to take, we may speak of (5) as a *linear combination* of the functions $B_i$, or as a *weighted sum* of the control vertices $\mathbf{V}_i$. For any particular $i$, equation (5) simply reduces to equation (3) because all the hat functions except $B_{i-1}(\bar{u})$ and $B_i(\bar{u})$ are zero inside the interval from $\bar{u}_i$ to $\bar{u}_{i+1}$. It should be clear that we can represent any piecewise linear curve in this way, and the unit hat functions $B_i(\bar{u})$ are called *basis functions* for this reason. Also, we may now turn our argument around: any particular vertex $\mathbf{V}_i$ contributes to the curve we are defining only where $B_i(\bar{u})$ is nonzero. Since $B_i(\bar{u})$ is nonzero only over the two successive intervals $[\bar{u}_i,\bar{u}_{i+1})$ and $[\bar{u}_{i+1},\bar{u}_{i+2})$, the actual position of $\mathbf{V}_i$ can only influence the two corresponding segments $\mathbf{Q}_i(\bar{u})$ and $\mathbf{Q}_{i+1}(\bar{u})$ of the curve — local control.

Notice that we have made use of the half-open intervals $[\bar{u}_i,\bar{u}_{i+1})$ and $[\bar{u}_{i+1},\bar{u}_{i+2})$ so that $\mathbf{Q}_i(\bar{u})$ defines the curve for values of $\bar{u}$ up to but not including $\bar{u}_{i+1}$ because the first interval is open at the right, and $\mathbf{Q}_{i+1}(\bar{u})$ then takes over at $\bar{u}_{i+1}$ itself because the second interval is closed at the left.

The hat functions that we have introduced are continuous, although their derivatives usually have jumps at knots (the technical term is $C^0$ continuous). When we use them to weight control vertices and sum them via equation (5) we obtain a curve that is consequently continuous, but whose first derivative vector may be discontinuous at knots: a piecewise linear curve, as we knew from the beginning.

Our real objective, of course, is to define curves like the one below by assembling pieces that are curved rather than straight. Just as in Chapter 3, and for the same reasons, we choose to consider piecewise cubic curves.

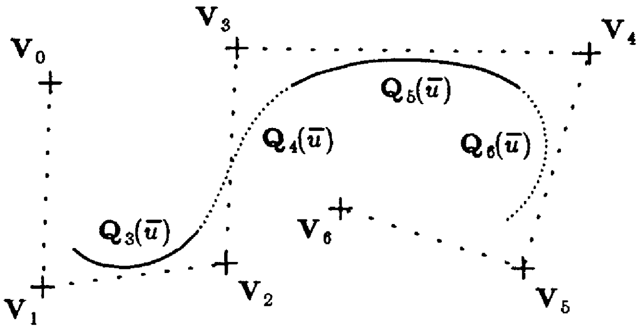


Figure 18. The curve shown is constructed from cubic B-splines over a uniform knot sequence. Notice how it smoothly approximates the indicated vertices, and indeed the control polygon.

The technique we are now developing does not, in general, interpolate the control vertices — that is a special property of the piecewise linear curves we have considered. Instead, each sequence of control vertices defines a curve that "passes near" those vertices. As before we may restrict our attention to a single coordinate such as $Y(\bar{u})$.

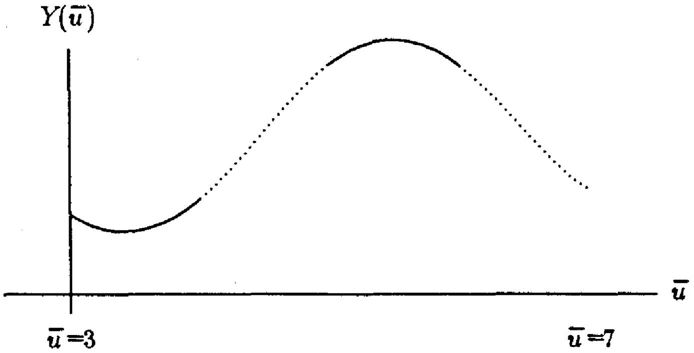


Figure 19. $Y(\bar{u})$ for the curve in Figure 18.

We choose to focus on piecewise cubic polynomial curves assembled from cubic polynomials $X_i(u)$ and $Y_i(u)$ that have positional, first derivative and second derivative continuity ($C^2$ continuity) at the joints between successive segments, so that they satisfy the equations

$$Q_{i-1}(\bar{u}_i) = Q_i(\bar{u}_i) \tag{6}$$

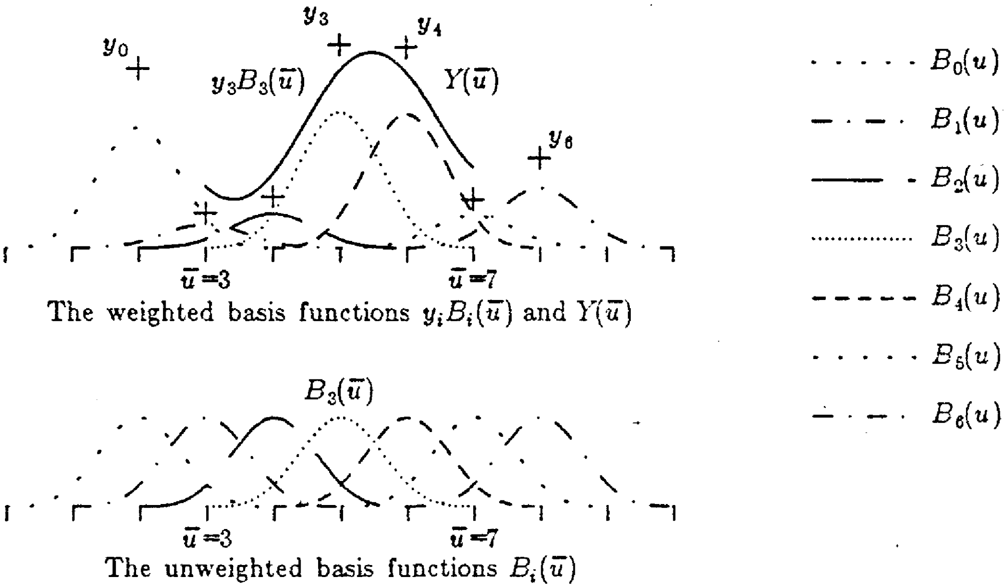
$$Q_{i-1}^{(1)}(\bar{u}_i) = Q_i^{(1)}(\bar{u}_i) \tag{7}$$

$$Q_{i-1}^{(2)}(\bar{u}_i) = Q_i^{(2)}(\bar{u}_i) \ . \tag{8}$$

In particular, this implies that

$$Y_{i-1}(\overline{u}_i) = Y_i(\overline{u}_i)$$
$$Y^{(1)}_{i-1}(\overline{u}_i) = Y^{(1)}_i(\overline{u}_i)$$
$$Y^{(2)}_{i-1}(\overline{u}_i) = Y^{(2)}_i(\overline{u}_i)$$

and similarly for $X(\overline{u})$. We can achieve the desired continuity if the basis functions with which we define $X(\overline{u})$ and $Y(\overline{u})$ are themselves $C^2$ continuous piecewise cubic polynomials with knots at the $\overline{u}_i$, since a linear combination (scaled sum) of such basis functions will also be a $C^2$ continuous piecewise cubic polynomial. Much as for the piecewise linear case, locality can be obtained if all but a small number of the parametric polynomial segments defining a basis function are identically zero. The basis functions we use will be smoother, and it turns out that this means they have to be nonzero on a somewhat wider interval, but the construction is otherwise quite analogous to the linear case we have already considered. For example, $Y(\overline{u})$ for the curve of Figure 18 can be represented in the following way as a sum of scaled $C^2$ continuous piecewise cubic basis functions.



The weighted basis functions $y_iB_i(\overline{u})$ and $Y(\overline{u})$

The unweighted basis functions $B_i(\overline{u})$

Figure 20. The $y$ component of the curve in Figure 18 as a scaled sum of basis functions.

Figure 20 illustrates several conventions. We choose to index control vertices from zero through $m$ (here 6). As we shall see, it requires four basis functions to properly define each cubic curve segment. Hence there are three more basis functions (and three more control vertices) than there are curve segments. Each basis function is nonzero over four parametric intervals. The leftmost basis function extends three additional intervals to the left of the curve, and the rightmost basis function extends three additional intervals to the right. Summarizing: there are $m+1$ control vertices, $m+1$ basis functions, $m-2$ curve segments bounded by $m-1$ knots, and $m-1+3+3 = m+5$ knots altogether. The curve is generated (swept out) as $\overline{u}$ runs from $\overline{u}_3$ to $\overline{u}_{m+1}$.

Let us now see how to actually define these basis functions. Using a little foresight, we suppose each basis function to be nonzero over four successive intervals (which for convenience we assume all have length one),

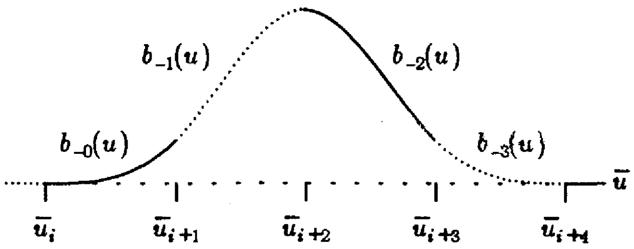


Figure 21. The uniform cubic B-spline $B_i(\bar{u})$ is a cubic $C^2$ basis function centred at $\bar{u}_{i+2}$. It is zero for $\bar{u} \leq \bar{u}_i$ and for $\bar{u} \geq \bar{u}_{i+4}$. The nonzero portion of $B_i(\bar{u})$ is composed of the four polynomial segments $b_{-0}(u)$, $b_{-1}(u)$, $b_{-2}(u)$ and $b_{-3}(u)$.

and ask that within each interval a basis function be defined by a cubic polynomial

$$a_j + b_j u + c_j u^2 + d_j u^3 \,, \qquad i-3 \leq j \leq i$$

having four coefficients. Thus the nonzero portion of our cubic B-spline basis function $B(\bar{u})$ consists (from left to right) of four *basis segments* $b_{-0}(u)$, $b_{-1}(u)$, $b_{-2}(u)$ and $b_{-3}(u)$, and there are sixteen coefficients to determine. By assumption $B_i(\bar{u})$ is identically zero for $\bar{u} \leq \bar{u}_i$ and for $\bar{u} \geq \bar{u}_{i+4}$, so the first and second derivatives $B_i^{(1)}(\bar{u})$ and $B_i^{(2)}(\bar{u})$ are also identically zero outside the interval $(\bar{u}_i, \bar{u}_{i+4})$. The requirement that positions, first derivatives, and second derivatives match at each knot $\bar{u}_j$ then implies that

$$\begin{array}{rclcrclcrcl}
0 &=& b_{-0}(0) & & 0 &=& b_{-0}^{(1)}(0) & & 0 &=& b_{-0}^{(2)}(0) \\
b_{-0}(1) &=& b_{-1}(0) & & b_{-0}^{(1)}(1) &=& b_{-1}^{(1)}(0) & & b_{-0}^{(2)}(1) &=& b_{-1}^{(2)}(0) \\
b_{-1}(1) &=& b_{-2}(0) & & b_{-1}^{(1)}(1) &=& b_{-2}^{(1)}(0) & & b_{-1}^{(2)}(1) &=& b_{-2}^{(2)}(0) \\
b_{-2}(1) &=& b_{-3}(0) & & b_{-2}^{(1)}(1) &=& b_{-3}^{(1)}(0) & & b_{-2}^{(2)}(1) &=& b_{-3}^{(2)}(0) \\
b_{-3}(1) &=& 0 & & b_{-3}^{(1)}(1) &=& 0 & & b_{-3}^{(2)}(1) &=& 0
\end{array}$$

where for simplicity each segment is individually parametrized so that $u = 0$ corresponds to its left endpoint and $u = 1$ corresponds to its right endpoint. These constitute fifteen constraints. We shall see that it is convenient to require that

$$b_{-0}(0) + b_{-1}(0) + b_{-2}(0) + b_{-3}(0) = 1 \ . \qquad (10)$$

Because $b_{-0}(0) = 0$ this simplifies to

$$b_{-1}(0) + b_{-2}(0) + b_{-3}(0) = 1 \ .$$

Because our knots are equally spaced, this amounts to assuming that when we add together an unscaled sequence of basis functions $B_i$, each of which is a copy of $B$ shifted so that its *support* (the parameter values for which it is nonzero) begins at $\bar{u}_i$, the three basis functions $B_{j-3}$, $B_{j-2}$ and $B_{j-1}$ which are nonzero at $\bar{u}_j$ sum to one. Such an assumption is said to be a *normalizing condition* and serves to define the function $B(\bar{u})$ uniquely. Rather miraculously, we shall see in the next section that this normalizing condition will in fact hold at all other values of $\bar{u}$ as well; that is,

$$b_{-0}(u) + b_{-1}(u) + b_{-2}(u) + b_{-3}(u) = 1, \qquad \text{for all } 0 \leq u < 1.$$

(Notice that our hat functions also summed to one.)

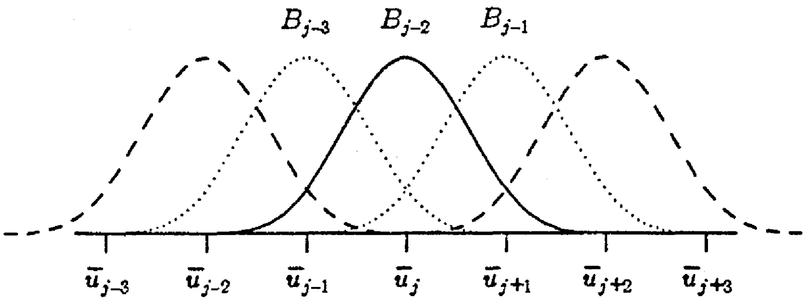


Figure 22. The basis functions which are not 0 at $u_j$.

We now have sixteen equations in sixteen unknowns (that is why we assumed that our basis function had four cubic segments), and we may solve for the coefficients $a_j$, $b_j$, $c_j$ and $d_j$ of the four segments $b_{-0}$, $b_{-1}$, $b_{-2}$, and $b_{-3}$ comprising our basis function $B$. Doing so yields the polynomials

$$\begin{aligned} b_{-0}(u) &= \frac{1}{6}\ u^3 \\ b_{-1}(u) &= \frac{1}{6}\ (\ 1 + 3u + 3u^2 - 3u^3\ ) \\ b_{-2}(u) &= \frac{1}{6}\ (\ 4 - 6u^2 + 3u^3\ ) \\ b_{-3}(u) &= \frac{1}{6}\ (\ 1 - 3u + 3u^2 - u^3\ )\ . \end{aligned} \tag{11}$$

These four segments define the *uniform cubic B-spline;* again, the term *uniform* means that the knots are equally spaced. The "B" is short for "Basis", which is appropriate because they can be used to represent any $C^2$ spline over a uniform knot sequence. Finally, it is easy to directly verify that these segments have the continuity necessary to qualify them as $C^2$ splines. Consider, for example, the joint between $b_{-2}(u)$ and $b_{-3}(u)$. So far as positional continuity is concerned, we have

$$b_{-2}(1) = b_{-3}(0) = \frac{1}{6}$$

Consider the first parametric derivative at their common joint. We have

$$\begin{aligned} b^{(1)}_{-2}(u) &= \frac{1}{6}\ (\ -12u + 9u^2\ ) \\ b^{(1)}_{-3}(u) &= \frac{1}{6}\ (\ -3 + 6u - 3u^2\ ) \end{aligned}$$

and

$$b^{(1)}_{-2}(1) = b^{(1)}_{-3}(0) = -\frac{1}{2}\ .$$

Their second parametric derivatives are given by

$$\begin{aligned} b^{(2)}_{-2}(u) &= (\ -2 + 3u\ ) \\ b^{(2)}_{-3}(u) &= (\ 1 - u\ ) \end{aligned}$$

so that

$$b_{-2}^{(2)}(1) = b_{-3}^{(2)}(0) = 1 \ .$$

However,

$$b_{-2}^{(3)}(u) = 3$$

$$b_{-3}^{(3)}(u) = -1$$

so that their common third parametric derivatives are not equal. Notice that we also have

$$b_{-3}(1) = b_{-3}^{(1)}(1) = b_{-3}^{(2)}(1) = 0 \ .$$

Since the basis function (and consequently all its derivatives) are identically zero to the right of $b_{-3}(1)$, we have positional as well as first and second derivative continuity at the right end of $b_{-3}(u)$ as well.

To determine a curve, we select a set of control vertices $\mathbf{V}_i$ and use them to define the curve

$$\mathbf{Q}(\bar{u}) = \sum_i \mathbf{V}_i B_i(\bar{u}) = \sum_i \left( x_i B_i(\bar{u}), y_i B_i(\bar{u}) \right) \qquad (12)$$

in which each $B_i$ is simply a copy of $B$, shifted so that its support extends from $\bar{u}_i$ to $\bar{u}_{i+4}$, and the coefficients in the summation are given by the control vertices

$$\mathbf{V}_i = ( x_i , y_i ) \ .$$

Notice that because the basis functions are nonzero on only four successive intervals, if $\bar{u}_i \leq \bar{u} < \bar{u}_{i+1}$ then

$$\mathbf{Q}_i(\bar{u}) = \sum_{r=-3}^{r=0} \mathbf{V}_{i+r} B_{i+r}(\bar{u}) \qquad (13)$$

$$= \mathbf{V}_{i-3} B_{i-3}(\bar{u}) + \mathbf{V}_{i-2} B_{i-2}(\bar{u}) + \mathbf{V}_{i-1} B_{i-1}(\bar{u}) + \mathbf{V}_{i-0} B_{i-0}(\bar{u}) \ .$$

If we replace each basis function $B_j(\bar{u})$ by the particular segment which pertains to the interval $[\bar{u}_i , \bar{u}_{i+1})$, then (13) can be written as

$$\mathbf{Q}_i(u) = \sum_{r=-3}^{r=0} \mathbf{V}_{i+r} b_r(u) \qquad (14)$$

$$= \mathbf{V}_{i-3} b_{-3}(u) + \mathbf{V}_{i-2} b_{-2}(u) + \mathbf{V}_{i-1} b_{-1}(u) + \mathbf{V}_{i-0} b_{-0}(u) \ .$$

Notice that the segments of our basis function are numbered from right to left because that is the order in which they appear when summed to form a curve: the leftmost control vertex scales the rightmost basis segment, and so on. Equation (14) also reflects the convenience of parametrizing each basis segment from $u=0$ at its left end; since the basis functions are all translates of one another, this convention allows us to use the same formulas in defining each basis function, and hence in computing each curve segment.

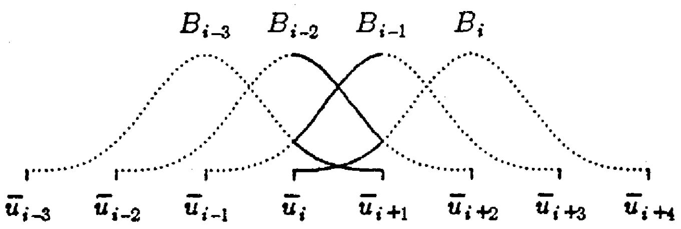


Figure 23. The four uniform cubic B-spline basis functions which are nonzero on the $i^{th}$ interval $[u_i,u_{i+1})$.

### 4.1. The Convex Hull Property

The *convex hull* of a set of control vertices in the plane can be thought of as the region lying inside a rubber band stretched so as to contain the control vertices, and then released so that it "snaps tightly against them." Formally, the convex hull defined by the control vertices $\mathbf{V}_i$ consists of all points that can be written as $\Sigma w_i \mathbf{V}_i$ where $\Sigma w_i = 1$ and $w_i \geq 0$. Thus the line segment joining any two points in a convex hull is also within the convex hull.

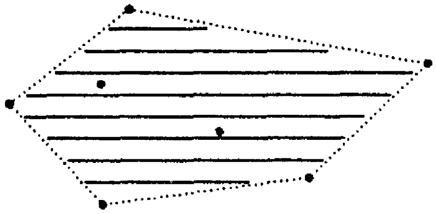


Figure 24. The convex hull of a set of control vertices.

It turns out to be a consequence of the way in which we have constructed the $B_i$ (specifically, a result of their normalization) that the $i^{th}$ segment of a uniform cubic B-spline curve lies within the convex hull of the vertices $\mathbf{V}_{i-3}$, $\mathbf{V}_{i-2}$, $\mathbf{V}_{i-1}$ and $\mathbf{V}_i$, as shown below.

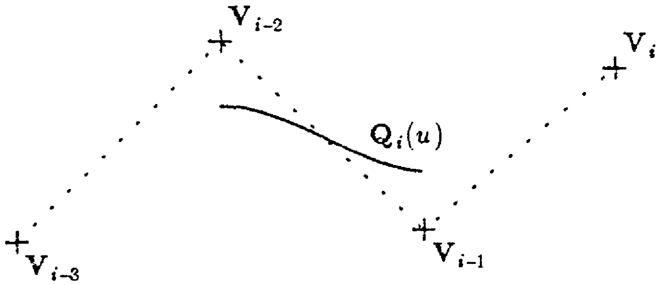


Figure 25. The $i^{th}$ segment lies within the convex hull of $\mathbf{V}_{i-3}$, $\mathbf{V}_{i-2}$, $\mathbf{V}_{i-1}$ and $\mathbf{V}_i$.

Although we only required that the basis functions sum to one at the knots, it is easy to verify directly by summing equations (11) that

$$\sum_{r=-3}^{r=0} b_r(u) = b_{-3}(u) + b_{-2}(u) + b_{-1}(u) + b_{-0}(u) = 1 \qquad (15)$$

for the entire interval $(0 \le u < 1)$. This is important because of the following general proposition: if we weight any set of $m+1$ points $\mathbf{V}_0$, ..., $\mathbf{V}_m$ by nonnegative coefficients $c_0$, ..., $c_m$ which sum to one $(c_0 + c_1 + \cdots + c_m = 1)$, then the resulting point

$$\mathbf{P} = c_0\mathbf{V}_0 + c_1\mathbf{V}_1 + \cdots + c_{m-1}\mathbf{V}_{m-1} + c_m\mathbf{V}_m$$

lies within the convex hull of the $\mathbf{V}_i$. Although we are only interested in such *convex combinations* of four vertices at the moment, we shall go ahead and show that it is true for an arbitrary number of vertices because we shall need the more general result later. An induction proof is straightforward:

First of all, notice that this is trivially true for one vertex. It is almost as trivially true for two vertices, say $\mathbf{V}_0$ and $\mathbf{V}_1$. We have

$$\mathbf{P} = c_0\mathbf{V}_0 + c_1\mathbf{V}_1 \ .$$

Since $c_0 + c_1 = 1$ we may rewrite this as

$$\mathbf{P} = (1-c_1)\mathbf{V}_0 + c_1\mathbf{V}_1$$

which for $0 \le c_1 \le 1$ we recognize as lying on the line segment joining $\mathbf{V}_0$ and $\mathbf{V}_1$. Thus the basis step.

Now suppose for the induction hypothesis that a convex combination of $m$ vertices lies within the convex hull of those vertices, and suppose that we have $m+1$ vertices $\mathbf{V}_0$, $\mathbf{V}_1$, ..., $\mathbf{V}_{m-1}$, $\mathbf{V}_m$ and corresponding nonnegative weights $c_0$, $c_1$, ..., $c_{m-1}$, $c_m$ which sum to one. To make use of our induction hypothesis we form the following convex combination $\mathbf{W}$ of the first $m$ vertices.

$$S = c_0 + c_1 + \cdots + c_{m-2} + c_{m-1} \neq 0 \qquad \text{(else trivial)}$$

$$d_i = \frac{c_i}{S} , \qquad 0 \le i \le m-1$$

$$\mathbf{W} = d_0\mathbf{V}_0 + d_1\mathbf{V}_1 + \cdots + d_{m-2}\mathbf{V}_{m-2} + d_{m-1}\mathbf{V}_{m-1} \ .$$

We have defined the $d_i$ in such a way that they sum to one $(d_0 + d_1 + \cdots + d_{m-1} = 1)$; hence $\mathbf{W}$ lies within the convex hull of the $\mathbf{V}_0$, $\mathbf{V}_1$, ..., $\mathbf{V}_{m-2}$, $\mathbf{V}_{m-1}$. Now consider

$$\mathbf{P} = S\mathbf{W} + (1-S)\mathbf{V}_m \ .$$

Since $\mathbf{V}_m$ is a control vertex and we already know that $\mathbf{W}$ is in the convex hull of the $\mathbf{V}_i$, $\mathbf{P}$ is in the convex hull of the $\mathbf{V}_i$. But by substituting in the definitions of $S$ and of $\mathbf{W}$, and noting that $(1-S)$ is $c_m$, we see that $\mathbf{P}$ is simply

$$c_0\mathbf{V}_0 + c_1\mathbf{V}_1 + \cdots + c_{m-1}\mathbf{V}_{m-1} + c_m\mathbf{V}_m$$

which is the point we wished to establish as being in the convex hull of the control vertices $\mathbf{V}_0$, $\mathbf{V}_1$, ..., $\mathbf{V}_{m-1}$, $\mathbf{V}_m$. Thus the induction step.

It now follows from equation (15) that the $i^{\text{th}}$ segment of a uniform cubic B-spline curve lies within the convex hull of $\mathbf{V}_{i-3}$, $\mathbf{V}_{i-2}$, $\mathbf{V}_{i-1}$ and $\mathbf{V}_i$. Thus if four successive control vertices of such a curve are visible on a display screen, so is the segment they define. An entire curve "follows" the control vertices in the sense that each successive segment lies within the convex hull of the next group of four control vertices; as we go from one segment to the next, the "oldest" is dropped because it no longer contributes to the curve, and a new vertex is picked up.

It also follows from this discussion that we may consider the B-splines as a "parameter dependent, varying convex combination" or "running average" of the control vertices.

### 4.2. Translation Invariance

It is highly desirable that translating all the control vertices by the same amount not change the shape of the curve defined. Like the convex hull property, this is an easy consequence of equation (15).

Suppose that we translate the control vertices by $\mathbf{t}=(dx,dy)$. Let $\mathbf{Q}(\bar{u})$ be the curve defined by the control vertices $\mathbf{V}_i$, and let $\mathbf{Q}_t(\bar{u})$ be the curve defined by the control vertices $\mathbf{V}_i+\mathbf{t}$. From (12) we have

$$\mathbf{Q}_t(\bar{u}) = \sum_i (\mathbf{V}_i + \mathbf{t})B_i(\bar{u}) = \sum_i \mathbf{V}_i B_i(\bar{u}) + \mathbf{t}\sum_i B_i(\bar{u}) \ .$$

From (15), then, we have

$$\mathbf{Q}_t(\bar{u}) = \sum_i \mathbf{V}_i B_i(\bar{u}) + \mathbf{t} = \mathbf{Q}(\bar{u}) + \mathbf{t} \ .$$

Thus we may either translate the control vertices and then compute the curve they define, or compute the curve first and then translate the points lying on it — the result is the same.

### 4.3. Rotation and Scaling Invariance

It is also important that we be able to rotate a curve without changing its shape.

Suppose that we rotate the control points by some angle $\theta$. Let $\mathbf{R}$ be the matrix accomplishing this rotation. Again $\mathbf{Q}(\bar{u})$ is the curve defined by the control points $\mathbf{V}_i$, and let $\mathbf{Q}_r(\bar{u})$ be the curve defined by the control points $\mathbf{R}\cdot\mathbf{V}_i$. From (12) we have

$$\mathbf{Q}_r(\bar{u}) = \sum_i (\mathbf{R}\cdot\mathbf{V}_i)B_i(\bar{u}) \ .$$

Since for any matrix $\mathbf{M}$ and vectors $\mathbf{a}$ and $\mathbf{b}$, $\mathbf{M}\cdot\mathbf{a}+\mathbf{M}\cdot\mathbf{b} = \mathbf{M}\cdot(\mathbf{a}+\mathbf{b})$, we have

$$\mathbf{Q}_r(\bar{u}) = \mathbf{R}\cdot\sum_i \mathbf{V}_i B_i(\bar{u}) = \mathbf{R}\cdot\mathbf{Q}(\bar{u}) \ .$$

Thus we may either rotate the control points and then compute the curve they define, or compute the curve first and then rotate the points lying on it — the result is the same.

Since scaling can be represented as a matrix operation, a similar argument establishes that the shape of a cubic B-spline curve is not affected by scaling the control vertices — the same curve is obtained if we scale points on the curve instead.

### 4.4. The Perspective Transformation

We note in passing that the perspective transformation does not preserve the shape of B-spline surfaces. That is, the surface obtained by computing points on a surface and then applying the perspective transformation is not identical to the surface obtained by applying the perspective transformation to the control vertices and then computing points on the surface defined by the transformed control vertices. In fact, the perspective transformation of a cubic curve or surface is not necessarily expressible as a cubic. It is instead a rational polynomial, namely the quotient of a cubic polynomial and a linear polynomial in $z$ which represents the depth of the surface, and it is easy to construct examples of these which are relatively prime.

## 4.5. End Conditions for Curves

Let's consider the beginning of a uniform cubic B-spline curve.

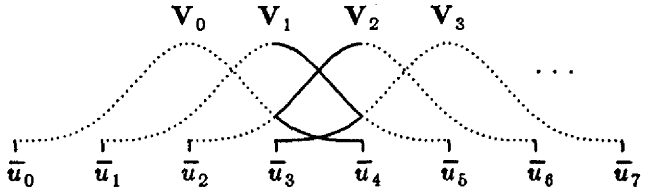


Figure 26. The four basis functions which define the first curve segment.

It is only in the fourth interval $[\bar{u}_3,\bar{u}_4)$ that we have four vertices with which to properly define a curve segment using equation (14). We might, of course, choose to define segments in the first three intervals by simply eliminating from equation (14) those terms for which we do not have a control vertex. This is not prudent, however, since the resulting curve segments will not necessarily lie within the convex hull of the control vertices defining them. Consider the first interval $[u_0,u_1)$: the corresponding curve segment will necessarily begin at $\mathbf{0}$ since only the one control vertex $\mathbf{V}_0$ is available and at the left end of the interval its corresponding basis function has the value 0.

Similarly, we can continue plotting a curve only so long as we have four control vertices with which to scale B-splines. The last knot is $\bar{u}_{m+4}$, but the last interval on which the curve can be defined is $[\bar{u}_m,\bar{u}_{m+1})$.

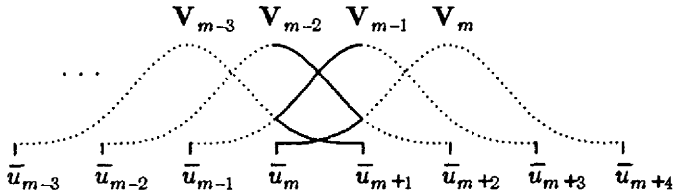


Figure 27. The four basis functions which define the last curve segment.

Thus for a uniform cubic B-spline curve we have three fewer segments than we have control vertices.

Since we often want the beginning or ending of a curve to have some particular property, the behaviour of uniform cubic B-spline curves at their end points is of interest. This subject is discussed in [Barsky82], from which the following presentation is drawn.

### 4.5.1. Curvature

One of the properties in which we are interested is curvature: whether, or how much, a curve "bends" at some point. Curvature is defined quantitatively in the following way.

At a given point $\mathbf{P}$ on a parametrically defined curve $\mathbf{Q}(\bar{u})$, the circle which has the same first and second derivative vectors as the curve is called the *osculating circle*. The centre and radius of this circle are called the *centre of curvature* $c(\bar{u})$ and the *radius of curvature* $\rho(\bar{u})$, respectively, at this point; the *curvature* $\kappa(\bar{u})$ at this point is the reciprocal, $1/\rho(\bar{u})$, of the radius of curvature. Thus if the osculating circle has a large radius, the curvature is small, as our intuition tells us. The *curvature vector* $\mathbf{K}(\bar{u})$ has a magnitude equal to the curvature and points from $\mathbf{P}$ towards the centre of curvature.

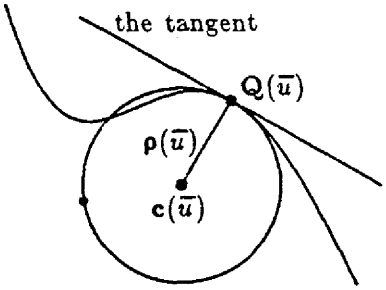


Figure 28. The osculating circle.

With the use of a bit of differential geometry it is possible to show (see [Barsky81] [Barsky85], or [Faux79, pp 99-101]) that the vector

$$\frac{\mathbf{Q}^{(1)}(\bar{u}) \times \mathbf{Q}^{(2)}(\bar{u})}{|\,\mathbf{Q}^{(1)}(\bar{u})\,|^3}$$

has magnitude equal to the curvature. However, this vector is perpendicular to the plane containing the osculating circle (the *osculating plane*). An additional cross-product with

$$\frac{\mathbf{Q}^{(1)}(\bar{u})}{|\mathbf{Q}^{(1)}(\bar{u})|} \tag{16}$$

results in a vector of the same length lying in the osculating plane, which is the curvature vector:

$$\mathbf{K}(\bar{u}) = \frac{(\mathbf{Q}^{(1)}(\bar{u}) \times \mathbf{Q}^{(2)}(\bar{u})) \times \mathbf{Q}^{(1)}(\bar{u})}{|\,\mathbf{Q}^{(1)}(\bar{u})\,|^4} \ . \tag{17}$$

From (17) it follows that:

- if the second derivative vector is zero, then the curvature is zero;
- if the first and second derivative vectors are nonzero but linearly dependent (collinear), then the curvature is zero;
- if the first and second derivative vectors are linearly independent (not collinear), then the curvature is nonzero.

### 4.5.2. No End Conditions

If we simply evaluate equation (14) at $u=0$ for the vertices $\mathbf{V}_0$, $\mathbf{V}_1$, $\mathbf{V}_2$, and $\mathbf{V}_3$, and at $u=1$ for the vertices $\mathbf{V}_{m-3}$, $\mathbf{V}_{m-2}$, $\mathbf{V}_{m-1}$, and $\mathbf{V}_m$, using (11) to define the basis functions, we find that the curve begins at

$$\mathbf{P}_s = \mathbf{Q}_3(0) = \frac{1}{6}\,(\,\mathbf{V}_0 + 4\mathbf{V}_1 + \mathbf{V}_2\,) \tag{18}$$

and ends at

$$\mathbf{P}_e = \mathbf{Q}_m(1) = \frac{1}{6}\,(\,\mathbf{V}_{m-2} + 4\mathbf{V}_{m-1} + \mathbf{V}_m\,) \ .$$

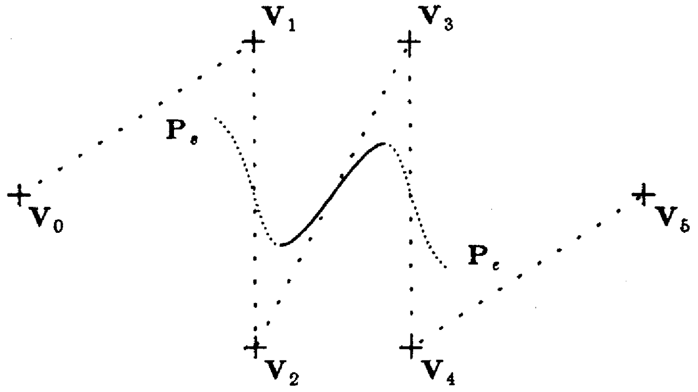


Figure 29. A 3-segment B-spline curve with no end conditions.

There are two natural ways in which to better control, among other things, the positions of $\mathbf{P}_s$ and $\mathbf{P}_e$. The first is to simply extend the vertex sequence $\mathbf{V}_0, \mathbf{V}_1, \ldots, \mathbf{V}_{m-1}, \mathbf{V}_m$ by repeating the end vertices $\mathbf{V}_0$ and $\mathbf{V}_m$ some number of times; this technique is said to make use of ***multiple vertices***. A second technique is to compute additional *phantom vertices* $\mathbf{V}_{-1}$ and $\mathbf{V}_{m+1}$ at either end (instead of having the user specify them), extending the curve by two segments so that $\mathbf{P}_s$ and $\mathbf{P}_e$ satisfy some condition.

### 4.5.3. Double Vertices

Suppose that we *double* the first and last vertices. That is, the user specifies the sequence of $m+1$ vertices $\mathbf{V}_0, \mathbf{V}_1, \ldots, \mathbf{V}_{m-1}, \mathbf{V}_m$, but we actually compute a curve of $m$ segments from the sequence of $m+3$ vertices $\mathbf{V}_0, \mathbf{V}_0, \mathbf{V}_1, \ldots, \mathbf{V}_{m-1}, \mathbf{V}_m, \mathbf{V}_m$. By adding a vertex to each end of the curve, we add an additional segment to each end as well. The new segments have the form

$$\mathbf{Q}_2(u) = \mathbf{V}_0[b_{-3}(u)+b_{-2}(u)] + \mathbf{V}_1 b_{-1}(u) + \mathbf{V}_2 b_{-0}(u) \tag{19}$$

$$\mathbf{Q}_{m+1}(u) = \mathbf{V}_{m-2} b_{-3}(u) + \mathbf{V}_{m-1} b_{-2}(u) + \mathbf{V}_m [b_{-1}(u)+b_{-0}(u)] \ . \tag{20}$$

If we evaluate these at $u=0$ and $u=1$, respectively, to obtain the first and last points on the curve (or substitute $\mathbf{V}_0$ for $\mathbf{V}_1$ and $\mathbf{V}_1$ for $\mathbf{V}_2$ in (18)) we find that

$$\mathbf{P}_s = \mathbf{Q}_2(0) = \frac{1}{6}(5\mathbf{V}_0 + \mathbf{V}_1) = (1-\tfrac{1}{6})\mathbf{V}_0 + \tfrac{1}{6}\mathbf{V}_1$$

$$\mathbf{P}_e = \mathbf{Q}_{m+1}(1) = \frac{1}{6}(\mathbf{V}_{m-1} + 5\mathbf{V}_m) = \tfrac{1}{6}\mathbf{V}_{m-1} + (1-\tfrac{1}{6})\mathbf{V}_m \ .$$

Thus the curve begins at a point $\mathbf{P}_s$ which is one-sixth of the way from $\mathbf{V}_0$ to $\mathbf{V}_1$ and ends at a point $\mathbf{P}_e$ which is one-sixth of the way from $\mathbf{V}_m$ to $\mathbf{V}_{m-1}$. Differentiating (19) and (20) and evaluating at $u=0$ and $u=1$, we find the first derivative vectors at $\mathbf{P}_s$ and $\mathbf{P}_e$ to be

$$\mathbf{Q}_2^{(1)}(0) = \frac{1}{2}(\mathbf{V}_1 - \mathbf{V}_0)$$

$$\mathbf{Q}_{m+1}^{(1)}(1) = \frac{1}{2}(\mathbf{V}_m - \mathbf{V}_{m-1})$$

Thus the curve is tangent at its endpoints to the first and last line segments of the control polygon.

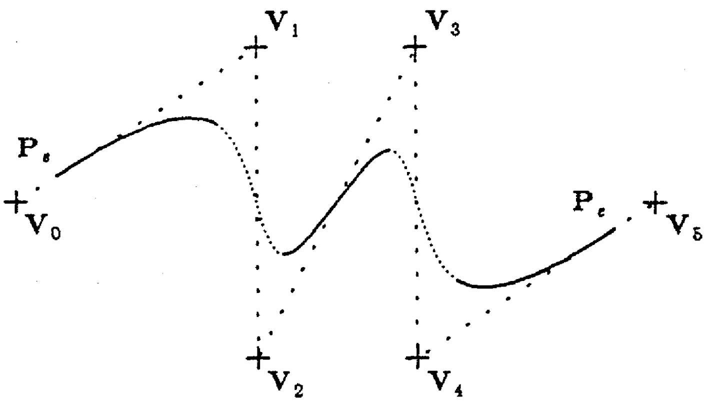


Figure 30. $V_0$ and $V_5$ are double vertices.

If we compute the second derivative vectors at $\mathbf{P}_s$ and $\mathbf{P}_e$ as well we find that they are collinear with the tangent vectors, so that the curvature at $\mathbf{P}_s$ and $\mathbf{P}_e$ is 0. However, it is not necessary to verify this directly, as it follows from the consideration of triple vertices below.

### 4.5.4. Triple Vertices

Suppose instead that we now *triple* the first and last vertices, so that the curve is computed from the $m+5$ vertices $\mathbf{V}_0, \mathbf{V}_0, \mathbf{V}_0, \mathbf{V}_1, \ldots, \mathbf{V}_{m-1}, \mathbf{V}_m, \mathbf{V}_m, \mathbf{V}_m$. This adds two additional segments

$$\begin{aligned} \mathbf{Q}_1(u) &= \mathbf{V}_0[b_{-3}(u)+b_{-2}(u)+b_{-1}(u)] + \mathbf{V}_1 b_{-0}(u) && (21)\\ \mathbf{Q}_2(u) &= \mathbf{V}_0[b_{-3}(u)+b_{-2}(u)] + \mathbf{V}_1 b_{-1}(u) + \mathbf{V}_2 b_{-0}(u) \end{aligned}$$

to the beginning of the curve and two additional segments

$$\begin{aligned} \mathbf{Q}_{m+1}(u) &= \mathbf{V}_{m-2} b_{-3}(u) + \mathbf{V}_{m-1} b_{-2}(u) + \mathbf{V}_m[b_{-1}(u)+b_{-0}(u)] \\ \mathbf{Q}_{m+2}(u) &= \mathbf{V}_{m-1} b_{-3}(u) + \mathbf{V}_m[b_{-2}(u)+b_{-1}(u)+b_{-0}(u)] && (22) \end{aligned}$$

to the end of the curve. If we now substitute in equations (11) and evaluate $\mathbf{Q}_2(0)$ and $\mathbf{Q}_{m+3}(1)$ we find that

$$\begin{aligned} \mathbf{P}_s &= \mathbf{Q}_1(0) = \mathbf{V}_0 \\ \mathbf{P}_e &= \mathbf{Q}_{m+2}(1) = \mathbf{V}_m \ . \end{aligned}$$

That is, the curve interpolates the first and last control vertices.

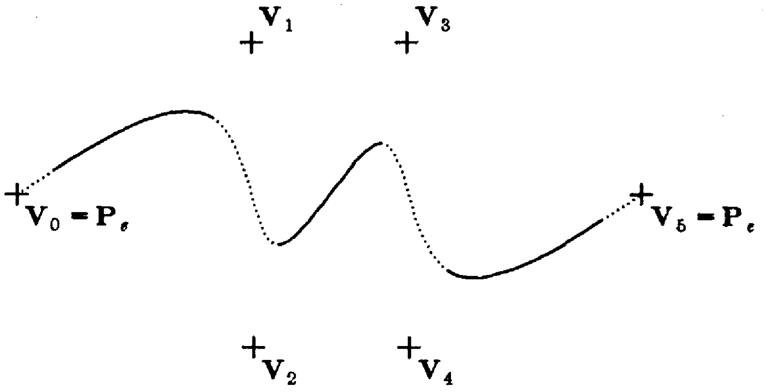


Figure 31. $V_0$ and $V_m$ are triple vertices, and are interpolated. The control polygon has been omitted so that the curve can be seen to reach $V_0$ and $V_m$.

Moreover, the first and last segments of the curve are now short straight line segments. We can verify this easily by simplifying (21) and (22). The equation which results for the first segment is

$$Q_1(u) = \left[1 - \frac{u^3}{6}\right] V_0 + \left[\frac{u^3}{6}\right] V_1$$

or

$$Q_1(s) = (1-s)V_0 + sV_1$$

for $s = u^3/6$, which we recognize as the equation of a line. The last segment of the curve is, analogously,

$$Q_{m+2}(u) = \left[\frac{1-u^3}{6}\right] V_{m-1} + \left[1 - \frac{1-u^3}{6}\right] V_m$$

or

$$Q_{m+2}(t) = t V_{m-1} + (1-t) V_m$$

for $t = (1-u^3)/6$.

The second and penultimate segments $Q_2(u)$ and $Q_{m+1}(u)$ begin and end, respectively, with a double vertex, and so exhibit the behaviour described for double vertices. Thus $Q_2(0)$ lies on the line segment from $V_0$ to $V_1$ and the curvature of $Q_2(u)$ is zero at that point, since it has the same first and second derivatives there as $Q_1(u)$, which is a straight line. By the same argument the curvature at $Q_{m+1}(1)$ is zero.

### 4.5.5. Multiple Interior Vertices

The analysis of double and triple vertices is equally applicable on the interior of a B-spline curve. Triple interior vertices are particularly interesting. So long as the triple vertex and the vertices immediately preceding and succeeding it fail to be collinear, the left and right derivative vectors at the triple vertex also fail to be collinear; the curve is said to be have a *cusp*.

At first sight this may seem to contradict the fact that the curve is $C^2$ continuous. As we will see in the chapter on continuity, the first derivative vector is **0** at the joint and is continuous there. A cusp results because the derivative vectors just to the left and right of the joint point in different directions.

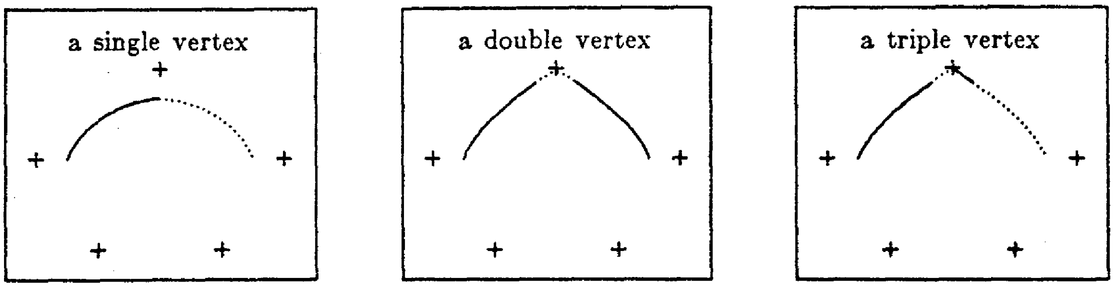


Figure 32. Forming a cusp by tripling a vertex. The double control vertex is not interpolated, while the triple vertex is.

### 4.5.6. Collinear Vertices

It is also useful to know that the segment defined by four collinear control vertices $\mathbf{V}_{i-3}$, $\mathbf{V}_{i-2}$, $\mathbf{V}_{i-1}$ and $\mathbf{V}_i$ is a straight line.

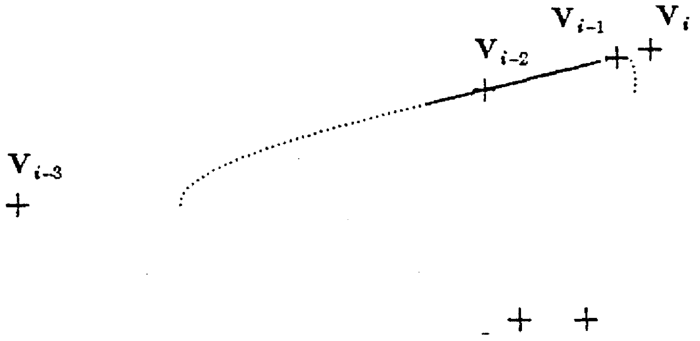


Figure 33. Four collinear control vertices produce a straight line segment.

This follows easily from the convex hull property.

### 4.5.7. Phantom Vertices: Position Specification

The essential idea behind all the *phantom vertex* techniques is to introduce two additional vertices $\mathbf{V}_{-1}$ and $\mathbf{V}_{m+1}$, thus defining two additional segments $\mathbf{Q}_2(u)$ and $\mathbf{Q}_{m+1}(u)$. The positions of $\mathbf{V}_{-1}$ and $\mathbf{V}_{m+1}$ are obtained by solving some constraint equations expressed in terms of $\mathbf{Q}_2(0)$ and $\mathbf{Q}_{m+1}(1)$ for $\mathbf{V}_{-1}$ and $\mathbf{V}_{m+1}$. For instance, we may allow the user to supply additional points $\mathbf{P}_s$ and $\mathbf{P}_e$ at which the curve is to begin and end, respectively, and then solve the equations

$$\mathbf{Q}_2(0) = \mathbf{P}_s = \frac{1}{6}\,(\,\mathbf{V}_{-1} + 4\mathbf{V}_0 + \mathbf{V}_1\,)$$

$$\mathbf{Q}_{m+1}(1) = \mathbf{P}_e = \frac{1}{6}\,(\,\mathbf{V}_{m-1} + 4\mathbf{V}_m + \mathbf{V}_{m+1}\,)$$

for

$$\mathbf{V}_{-1} = 6\mathbf{P}_s - 4\mathbf{V}_0 - \mathbf{V}_1 \tag{23}$$

$$V_{m+1} = 6P_e - 4V_m - V_{m-1} \ .$$

The curvature at $P_s$ and $P_e$ is analyzed by computing the first and second derivative vectors at these two points:

$$Q_2^{(1)}(0) = \frac{V_1 - V_{-1}}{2} = V_1 + 2V_0 - 3P_s \tag{24}$$

$$Q_{m+1}^{(1)}(1) = \frac{V_{m+1} - V_{m-1}}{2} = 3P_e - 2V_m - V_{m-1} \tag{25}$$

$$Q_2^{(2)}(0) = V_{-1} - 2V_0 + V_1 = 6(P_s - V_0) \tag{26}$$

$$Q_{m+1}^{(2)}(1) = V_{m-1} - 2V_m + V_{m+1} = 6(P_e - V_m) \ . \tag{27}$$

Since $V_1$ appears in (24) but not in (26), $Q_2^{(1)}(0)$ is not a scalar multiple of $Q_2^{(2)}(0)$, so that the first and second derivative vectors are linearly independent. As we saw earlier, this is sufficient to conclude that the curvature at $P_s$ is nonzero. Similarly, $V_m$ appears in (25) but not in (27), so that the curvature at $P_e$ is also nonzero.

### 4.5.8. Phantom Vertices: End Vertex Interpolation

This is really a special case of the position specification described above. Instead of supplying new end points, we ask that phantom vertices $V_{-1}$ and $V_{m+1}$ be found that cause the curve to interpolate $V_0$ and $V_m$. Substituting $V_0$ for $P_s$ and $V_m$ for $P_m$ in (23) yields the following equations for the phantom vertices $V_{-1}$ and $V_{m+1}$.

$$V_{-1} = 2V_0 - V_1$$

$$V_{m+2} = 2V_m - V_{m-1} \ .$$

For this special case the derivative vectors given in equations (24)-(27) become

$$Q_2^{(1)}(0) = V_1 - V_0$$

$$Q_{m+1}^{(1)}(1) = V_m - V_{m-1}$$

$$Q_2^{(2)}(0) = 0$$

$$Q_{m+1}^{(2)}(1) = 0 \ .$$

Thus for end vertex interpolation by means of phantom vertices the curve is tangent to the control polygon with zero curvature at its endpoints. This case is distinct from the end vertex interpolation resulting from triple end vertices since it does not usually result in straight line segments for the first and last curve segments.

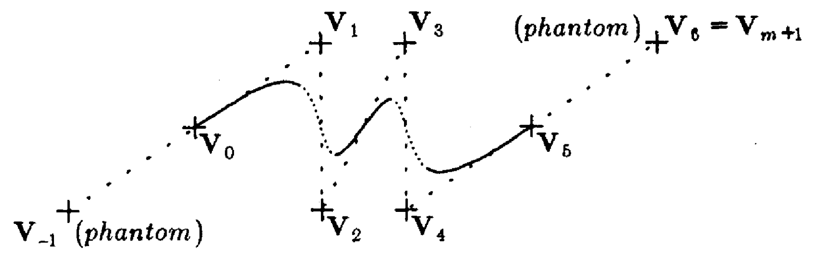


Figure 34. End vertex interpolation via phantom vertices

### 4.5.9. Phantom Vertices: Fixing Derivative Vectors

It is also possible to compute phantom vertices that give the curve selected first or second derivative vectors at its initial and final points. The actual initial and final positions of the curve are fixed as a result of specifying a derivative vector; since they do not coincide with any particularly meaningful positions (such as a control vertex), we will not discuss them. Further details may be found in [Barsky82].

### 4.5.10. End Conditions: Closed Curves

The curves we have discussed so far are *open* curves, which is to say that the two endpoints do not, in general, coincide. A $C^2$ continuous *closed* curve whose endpoints do meet, and which is $C^2$ continuous there as well, is obtained if the first three control vertices are identical to the last three, since if we use the $m+4$ vertex sequence $\mathbf{V}_0, \mathbf{V}_1, \mathbf{V}_2, \ldots, \mathbf{V}_{m-1}, \mathbf{V}_m, \mathbf{V}_0, \mathbf{V}_1, \mathbf{V}_2$ to define $m+1$ segments,

$$\mathbf{P}_s = \frac{1}{6}(\mathbf{V}_0 + 4\mathbf{V}_1 + \mathbf{V}_2) \qquad (18)$$

and ends at

$$\mathbf{P}_e = \frac{1}{6}(\mathbf{V}_{m+1} + 4\mathbf{V}_{m+2} + \mathbf{V}_{m+3}) = \frac{1}{6}(\mathbf{V}_0 + 4\mathbf{V}_1 + \mathbf{V}_2)$$

so that the curve is continuous. To see that the curve is, in fact, $C^2$ continuous, notice that the last curve segment defined by this vertex sequence is determined by $\mathbf{V}_m$, $\mathbf{V}_0$, $\mathbf{V}_1$ and $\mathbf{V}_2$; if we think of the vertex sequence as wrapping around on itself circularly, with $V_m$ followed by $\mathbf{V}_0$, then the following segment (with which it would join $C^2$ continuously) would be determined by $\mathbf{V}_0$, $\mathbf{V}_1$, $\mathbf{V}_2$ and $\mathbf{V}_3$. But this is simply the first segment of the curve we have defined, so it is clear that the head and tail of the curve join with first and second derivative continuity.

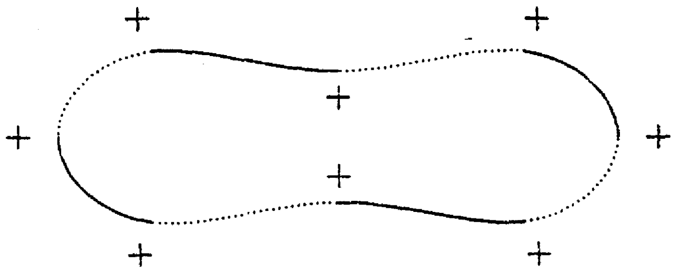


Figure 35. A *closed* uniform cubic B-spline curve.

### 4.6. Uniform Bicubic B-Spline Surfaces

The formation of uniform bicubic B-spline surfaces is a natural and straightforward generalization of the uniform cubic B-spline curves. We want to form our surface as a scaled sum of basis functions, as in (12), but now $X$, $Y$ and $Z$ must be functions of two independent parameters:

$$\begin{aligned}\mathbf{Q}(\bar{u},\bar{v}) &= \sum_{i,j} \mathbf{V}_{i,j} B_{i,j}(\bar{u},\bar{v}) \qquad (28)\\ &= \sum_{i,j} \left( x_{i,j}B_{i,j}(\bar{u},\bar{v}),\ y_{i,j}B_{i,j}(\bar{u},\bar{v}),\ z_{i,j}B_{i,j}(\bar{u},\bar{v}) \right) .\end{aligned}$$

For scale factors we again use the $x$-, $y$- and $z$-coordinates of what is now a two-dimensional array of control vertices $\mathbf{V}_{i,j}$ called the *control mesh* or *control graph* near which the surface is to pass. (See

Figure 36 below.) To obtain locality we would like the new basis functions $B_{i,j}(\bar{u},\bar{v})$ to be nonzero only for a small range of $\bar{u}$ and $\bar{v}$. An easy way to arrange this is to let $B_{i,j}(\bar{u},\bar{v})=B_i(\bar{u})B_j(\bar{v})$, where $B_i(\bar{u})$ and $B_j(\bar{v})$ are simply the univariate B-splines defined by (11). Since each is nonzero only over four successive intervals, if $\bar{u}_{i-1}\le\bar{u}\le\bar{u}_i$ and $v_{j-1}\le\bar{v}\le v_j$ we can rewrite (28) as

$$\mathbf{Q}(\bar{u},\bar{v}) = \sum_{r=-3}^{0}\sum_{s=-3}^{0} \mathbf{V}_{i+r,j+s}B_{i+r}(\bar{u})B_{j+s}(\bar{v}) \ . \tag{29}$$

This is simply the tensor or Cartesian product of two univariate B-spline curve segments. If we rewrite (29) in terms of basis segments instead of basis functions and adopt the convention that the portion of $\mathbf{Q}(\bar{u},\bar{v})$ defined by this set of values for $\bar{u}$ and $\bar{v}$ is denoted by $\mathbf{Q}_{i,j}(u,v)$, then we can write

$$\mathbf{Q}_{i,j}(u,v) = \sum_{r=-3}^{0}\sum_{s=-3}^{0} \mathbf{V}_{i+r,j+s}b_r(u)b_s(v) \tag{30}$$

so that $\mathbf{Q}_{i,j}(u,v)$, the $i,j^{th}$ *patch*, is completely determined by sixteen control vertices. Thus the four by four array

$$\begin{array}{cccc}
\mathbf{V}_{0,3} & \mathbf{V}_{1,3} & \mathbf{V}_{2,3} & \mathbf{V}_{3,3} \\
\mathbf{V}_{0,2} & \mathbf{V}_{1,2} & \mathbf{V}_{2,2} & \mathbf{V}_{3,2} \\
\mathbf{V}_{0,1} & \mathbf{V}_{1,1} & \mathbf{V}_{2,1} & \mathbf{V}_{3,1} \\
\mathbf{V}_{0,0} & \mathbf{V}_{1,0} & \mathbf{V}_{2,0} & \mathbf{V}_{3,0}
\end{array}$$

of control vertices can be used to define the single patch

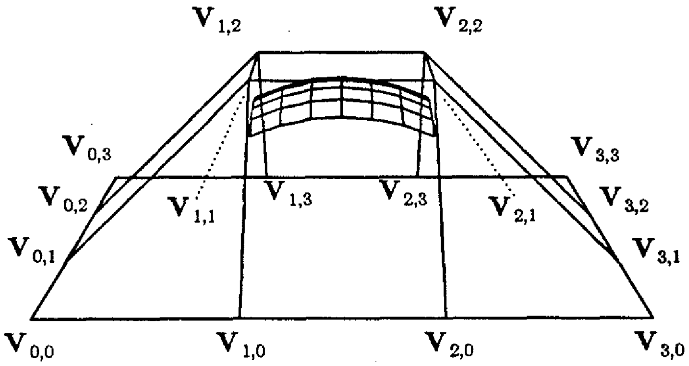


Figure 36. A B-spline surface, consisting of a single patch, with its control graph. The patch is rendered here by drawing seven lines of constant $\bar{u}$ which are equally spaced in $\bar{v}$, and seven lines of constant $\bar{v}$ which are equally spaced in $\bar{u}$. Adjacent vertices in the control graph are connected by straight line segments. Notice how the patch lies close to the central four control vertices.

The separability of $B_{i,j}(\bar{u},\bar{v})$ into $B_i(\bar{u})$ and $B_j(\bar{v})$ can be useful. For example, we can expand (30) as

$$
\begin{aligned}
\mathbf{Q}_{i,j}(u,v) = & \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (31)\\
& [\,\mathbf{V}_{i-3,j}\ b_{-3}(u) + \mathbf{V}_{i-2,j}\ b_{-2}(u) + \mathbf{V}_{i-1,j}\ b_{-1}(u) + \mathbf{V}_{i,j}\ b_{-0}(u)]\ b_{-0}(v) \quad + \\
& [\,\mathbf{V}_{i-3,j-1} b_{-3}(u) + \mathbf{V}_{i-2,j-1} b_{-2}(u) + \mathbf{V}_{i-1,j-1} b_{-1}(u) + \mathbf{V}_{i,j-1} b_{-0}(u)]\ b_{-1}(v) \quad + \\
& [\,\mathbf{V}_{i-3,j-2} b_{-3}(u) + \mathbf{V}_{i-2,j-2} b_{-2}(u) + \mathbf{V}_{i-1,j-2} b_{-1}(u) + \mathbf{V}_{i,j-2} b_{-0}(u)]\ b_{-2}(v) \quad + \\
& [\,\mathbf{V}_{i-3,j-3} b_{-3}(u) + \mathbf{V}_{i-2,j-3} b_{-2}(u) + \mathbf{V}_{i-1,j-3} b_{-1}(u) + \mathbf{V}_{i,j-3} b_{-0}(u)]\ b_{-3}(v) \ .
\end{aligned}
$$

From this it is clear that if we fix $u$ at some arbitrary value between 0 and 1 then we can write (31) as

$$\mathbf{Q}_{i,j,u}(v) = \mathbf{W}_0 b_{-3}(v) + \mathbf{W}_1 b_{-2}(v) + \mathbf{W}_2 b_{-1}(v) + \mathbf{W}_3 b_{-0}(v)$$

where the appearance of $u$ in the subscript indicates that its value has been fixed, and

$$
\begin{aligned}
\mathbf{W}_3 &= \mathbf{V}_{i-3,j}\ b_{-3}(u) + \mathbf{V}_{i-2,j}\ b_{-2}(u) + \mathbf{V}_{i-1,j}\ b_{-1}(u) + \mathbf{V}_{i,j}\ b_{-0}(u)\\
\mathbf{W}_2 &= \mathbf{V}_{i-3,j-1}\ b_{-3}(u) + \mathbf{V}_{i-2,j-1}\ b_{-2}(u) + \mathbf{V}_{i-1,j-1}\ b_{-1}(u) + \mathbf{V}_{i,j-1}\ b_{-0}(u)\\
\mathbf{W}_1 &= \mathbf{V}_{i-3,j-2}\ b_{-3}(u) + \mathbf{V}_{i-2,j-2}\ b_{-2}(u) + \mathbf{V}_{i-1,j-2}\ b_{-1}(u) + \mathbf{V}_{i,j-2}\ b_{-0}(u)\\
\mathbf{W}_0 &= \mathbf{V}_{i-3,j-3}\ b_{-3}(u) + \mathbf{V}_{i-2,j-3}\ b_{-2}(u) + \mathbf{V}_{i-1,j-3}\ b_{-1}(u) + \mathbf{V}_{i,j-3}\ b_{-0}(u)
\end{aligned}
$$

Thus $\mathbf{Q}_{i,j,u}(v)$ is simply the uniform cubic B-spline curve segment defined by the "control vertices" $\mathbf{W}_0$, $\mathbf{W}_1$, $\mathbf{W}_2$ and $\mathbf{W}_3$. It is not hard to see that the curve segment $\mathbf{Q}_{i,j+1,u}(v)$, in the next patch "up", is given by

$$\mathbf{Q}_{i,j+1,u}(v) = \mathbf{W}_1 b_{-3}(v) + \mathbf{W}_2 b_{-2}(v) + \mathbf{W}_3 b_{-1}(v) + \mathbf{W}_4 b_{-0}(v)$$

where

$$\mathbf{W}_4 = \mathbf{V}_{i-3,j+1} b_{-3}(u) + \mathbf{V}_{i-2,j+1} b_{-2}(u) + \mathbf{V}_{i-1,j+1} b_{-1}(u) + \mathbf{V}_{i,j+1} b_{-0}(u) \ .$$

This is simply the second segment in a uniform cubic B-spline curve defined by the "control vertices" $\mathbf{W}_1$, $\mathbf{W}_2$, $\mathbf{W}_3$ and $\mathbf{W}_4$. It follows immediately that this curve is $C^2$ continuous. Since a completely analogous argument can be made with respect to $u$ by factoring the $b_r(u)$ out of (30) instead of the $b_s(v)$, the uniform cubic B-spline surface we have defined is $C^2$ continuous along lines of constant $u$ and $v$. It follows from elementary calculus that the uniform cubic B-spline surfaces are therefore $C^2$ continuous in every direction.

### 4.7. Boundary Conditions for Surfaces

Just as we are interested in specifying end conditions to control the way in which curves terminate, so are we interested in specifying boundary conditions to control the behaviour on the periphery of a surface.

Recall from equation (30) and Figure 36 that it requires sixteen control vertices to define a single patch. Adding an additional "column" of four control vertices would enable us to define an additional horizontally adjacent patch; adding $m$ additional columns defines $m$ additional patches.

$$\begin{array}{ccccc}
\mathbf{V}_{0,3} & \mathbf{V}_{1,3} & \mathbf{V}_{2,3} & \mathbf{V}_{3,3} & \mathbf{V}_{4,3} \\
\mathbf{V}_{0,2} & \mathbf{V}_{1,2} & \mathbf{V}_{2,2} & \mathbf{V}_{3,2} & \mathbf{V}_{4,2} \\
\mathbf{V}_{0,1} & \mathbf{V}_{1,1} & \mathbf{V}_{2,1} & \mathbf{V}_{3,1} & \mathbf{V}_{4,2} \\
\mathbf{V}_{0,0} & \mathbf{V}_{1,0} & \mathbf{V}_{2,0} & \mathbf{V}_{3,0} & \mathbf{V}_{4,2}
\end{array}$$

A four row by five column array of control vertices, defining two adjacent patches. Columns one, two, three and four together define one patch. Columns two, three, four and five together define the second.

If there are $m+1$ such columns in the control mesh, each with four vertices, then we can generate $m-2$ patches.

Similarly, adding a "row" of four vertices enables us to add an adjacent vertical patch; adding $n$ additional rows of four vertices each adds $n$ additional patches, stacked vertically. A total of $n-2$ such patches result if we have $n+1$ rows.

In general, then, there are three fewer rows and columns of patches than there are rows and columns of control vertices. Hence an $(m+1)\times(n+1)$ array of control vertices defines $(m-2)\times(n-2)$ patches. We can, in a natural way, define additional boundary patches either by repeating boundary vertices or by defining phantom vertices, much as we did for curves.

### 4.7.1. Multiple Vertices

It is easiest to describe and illustrate this process if we start with an array of sixteen control vertices

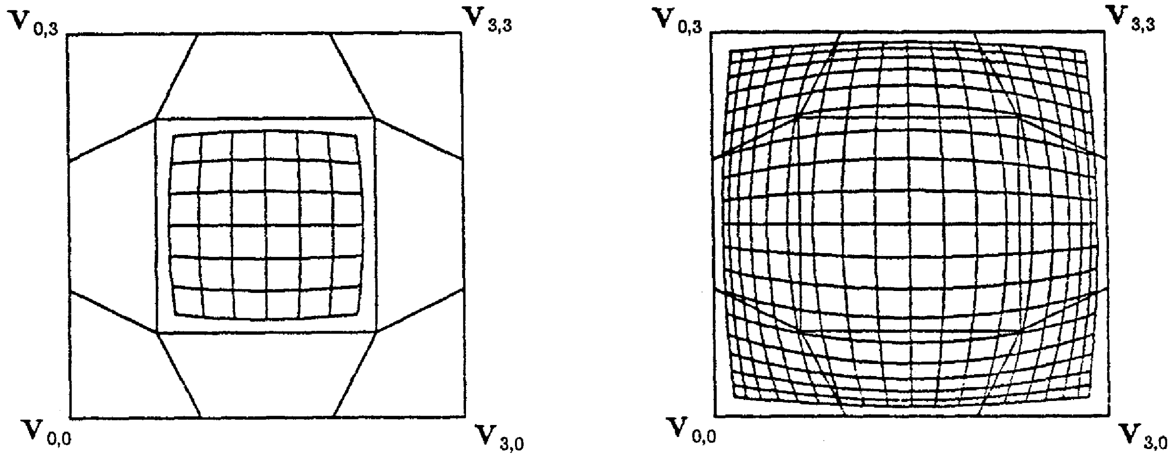
$$\begin{array}{cccc}
\mathbf{V}_{0,3} & \mathbf{V}_{1,3} & \mathbf{V}_{2,3} & \mathbf{V}_{3,3} \\
\mathbf{V}_{0,2} & \mathbf{V}_{1,2} & \mathbf{V}_{2,2} & \mathbf{V}_{3,2} \\
\mathbf{V}_{0,1} & \mathbf{V}_{1,1} & \mathbf{V}_{2,1} & \mathbf{V}_{3,1} \\
\mathbf{V}_{0,0} & \mathbf{V}_{1,0} & \mathbf{V}_{2,0} & \mathbf{V}_{3,0}
\end{array}$$

defining a single patch which is physically proximate to the four central vertices $\mathbf{V}_{1,1}$, $\mathbf{V}_{2,1}$, $\mathbf{V}_{1,2}$ and $\mathbf{V}_{2,2}$ (as in Figure 36). If we double the boundary vertices we obtain the control mesh

$$\begin{array}{cccccc}
\mathbf{V}_{0,3} & \mathbf{V}_{0,3} & \mathbf{V}_{1,3} & \mathbf{V}_{2,3} & \mathbf{V}_{3,3} & \mathbf{V}_{3,3} \\
\mathbf{V}_{0,3} & \mathbf{V}_{0,3} & \mathbf{V}_{1,3} & \mathbf{V}_{2,3} & \mathbf{V}_{3,3} & \mathbf{V}_{3,3} \\
\mathbf{V}_{0,2} & \mathbf{V}_{0,2} & \mathbf{V}_{1,2} & \mathbf{V}_{2,2} & \mathbf{V}_{3,2} & \mathbf{V}_{3,2} \\
\mathbf{V}_{0,1} & \mathbf{V}_{0,1} & \mathbf{V}_{1,1} & \mathbf{V}_{2,1} & \mathbf{V}_{3,1} & \mathbf{V}_{3,1} \\
\mathbf{V}_{0,0} & \mathbf{V}_{0,0} & \mathbf{V}_{1,0} & \mathbf{V}_{2,0} & \mathbf{V}_{3,0} & \mathbf{V}_{3,0} \\
\mathbf{V}_{0,0} & \mathbf{V}_{0,0} & \mathbf{V}_{1,0} & \mathbf{V}_{2,0} & \mathbf{V}_{3,0} & \mathbf{V}_{3,0}
\end{array}$$

which adds a "strip" of boundary patches around the periphery of the original single-patch surface.

Notice that the corner vertices $\mathbf{V}_{0,3}$, $\mathbf{V}_{3,3}$, $\mathbf{V}_{0,0}$ and $\mathbf{V}_{3,0}$ are actually replicated three times (once horizontally, once vertically, and once diagonally to define a new corner vertex), while all other boundary vertices are replicated once (doubled). The surface on the right in the following figure, which is shown from above, is the result.



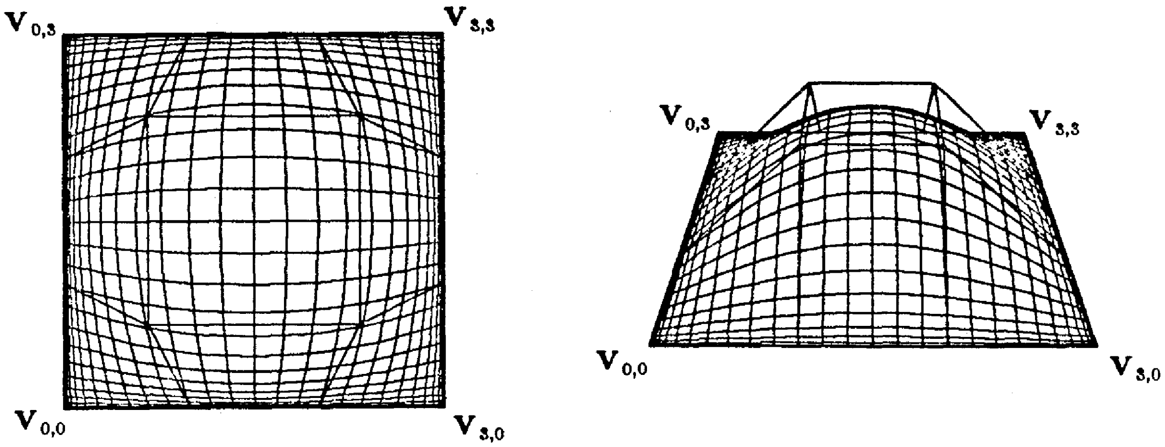
No boundary conditions. Double boundary vertices.

Figure 37. The single patch shown on the left has been rendered by drawing 7 equally spaced lines of constant $\bar{u}$ and 7 equally spaced lines of constant $\bar{v}$. An example of double boundary vertices is shown on the right. This surface, consisting of 9 patches, has been rendered by drawing 19 equally-spaced lines of constant $\bar{u}$ and 19 equally spaced lines of constant $\bar{v}$. Both surfaces are shown from above.

Suppose that we now triple the boundary vertices, so as to define two additional strips of boundary patches. For our example surface this yields the control mesh

$$\begin{array}{cccccccc}
\mathbf{V}_{0,3} & \mathbf{V}_{0,3} & \mathbf{V}_{0,3} & \mathbf{V}_{1,3} & \mathbf{V}_{2,3} & \mathbf{V}_{3,3} & \mathbf{V}_{3,3} & \mathbf{V}_{3,3} \\
\mathbf{V}_{0,3} & \mathbf{V}_{0,3} & \mathbf{V}_{0,3} & \mathbf{V}_{1,3} & \mathbf{V}_{2,3} & \mathbf{V}_{3,3} & \mathbf{V}_{3,3} & \mathbf{V}_{3.3} \\
\mathbf{V}_{0,3} & \mathbf{V}_{0,3} & \mathbf{V}_{0,3} & \mathbf{V}_{1,3} & \mathbf{V}_{2,3} & \mathbf{V}_{3,3} & \mathbf{V}_{3,3} & \mathbf{V}_{3,3} \\
\mathbf{V}_{0.2} & \mathbf{V}_{0,2} & \mathbf{V}_{0.2} & \mathbf{V}_{1,2} & \mathbf{V}_{2,2} & \mathbf{V}_{3,2} & \mathbf{V}_{3,2} & \mathbf{V}_{3,2} \\
\mathbf{V}_{0,1} & \mathbf{V}_{0,1} & \mathbf{V}_{0,1} & \mathbf{V}_{1,1} & \mathbf{V}_{2,1} & \mathbf{V}_{3,1} & \mathbf{V}_{3,1} & \mathbf{V}_{3,1} \\
\mathbf{V}_{0,0} & \mathbf{V}_{0,0} & \mathbf{V}_{0,0} & \mathbf{V}_{1,0} & \mathbf{V}_{2,0} & \mathbf{V}_{3,0} & \mathbf{V}_{3,0} & \mathbf{V}_{3,0} \\
\mathbf{V}_{0,0} & \mathbf{V}_{0,0} & \mathbf{V}_{0,0} & \mathbf{V}_{1,0} & \mathbf{V}_{2,0} & \mathbf{V}_{3,0} & \mathbf{V}_{3,0} & \mathbf{V}_{3,0} \\
\mathbf{V}_{0,0} & \mathbf{V}_{0,0} & \mathbf{V}_{0,0} & \mathbf{V}_{1,0} & \mathbf{V}_{2,0} & \mathbf{V}_{3,0} & \mathbf{V}_{3,0} & \mathbf{V}_{3,0}
\end{array}$$

and defines the following surface.



Triple boundary vertices.

The same surface the front.

Figure 38. Triple boundary vertices. This surface, consisting of 25 patches, has been rendered by drawing 31 equally-spaced lines of constant $\bar{u}$ and 31 equally spaced lines of constant $\bar{v}$.

Doubling and tripling the boundary vertices adds patches which bring the surface closer to the periphery of the control graph. Indeed, for this control graph, tripling the boundary vertices causes the boundary of the surface to interpolate the line segments joining the peripheral boundary vertices. To see why this is so, it is necessary to know that this is a special case; the "bottom" four vertices in the control graph are collinear, as are the "top" four vertices in the control graph, the leftmost four vertices, and the rightmost four vertices. Let us pick an arbitrary four by four sub-array of the control graph that defines a boundary patch of the "tripled" surface, say

$$\begin{array}{cccc}
\mathbf{V}_{0,3} & \mathbf{V}_{0,3} & \mathbf{V}_{0,3} & \mathbf{V}_{1,3} \\
\mathbf{V}_{0,2} & \mathbf{V}_{0,2} & \mathbf{V}_{0,2} & \mathbf{V}_{1,2} \\
\mathbf{V}_{0,1} & \mathbf{V}_{0,1} & \mathbf{V}_{0,1} & \mathbf{V}_{1,1} \\
\mathbf{V}_{0,0} & \mathbf{V}_{0,0} & \mathbf{V}_{0,0} & \mathbf{V}_{1,0}
\end{array}$$

and substitute these into equation (31) to obtain the patch they define. We have

$$\begin{aligned}
\mathbf{Q}_{1,3}(u,v) \; = & \\
& [\mathbf{V}_{0,3}b_{-3}(u) + \mathbf{V}_{0,3}b_{-2}(u) + \mathbf{V}_{0,3}b_{-1}(u) + \mathbf{V}_{1,3}b_{-0}(u)]\; b_{-0}(v) \quad + \\
& [\mathbf{V}_{0,2}b_{-3}(u) + \mathbf{V}_{0,2}b_{-2}(u) + \mathbf{V}_{0,2}b_{-1}(u) + \mathbf{V}_{1,2}b_{-0}(u)]\; b_{-1}(v) \quad + \\
& [\mathbf{V}_{0,1}b_{-3}(u) + \mathbf{V}_{0,1}b_{-2}(u) + \mathbf{V}_{0,1}b_{-1}(u) + \mathbf{V}_{1,1}b_{-0}(u)]\; b_{-2}(v) \quad + \\
& [\mathbf{V}_{0,0}b_{-3}(u) + \mathbf{V}_{0,0}b_{-2}(u) + \mathbf{V}_{0,0}b_{-1}(u) + \mathbf{V}_{1,0}b_{-0}(u)]\; b_{-3}(v) \; .
\end{aligned} \tag{32}$$

Once again we fix $u$ at some arbitrary value between 0 and 1 and write (32) as

$$\mathbf{Q}_{1,3,u}(v) \;=\; \mathbf{W}_0 b_{-3}(v) + \mathbf{W}_1 b_{-2}(v) + \mathbf{W}_2 b_{-1}(v) + \mathbf{W}_3 b_{-0}(v)$$

where

$$\begin{array}{lllllllll}
\mathbf{W}_3 & = & \mathbf{V}_{0,3}\; b_{-3}(u) & + & \mathbf{V}_{0,3}\; b_{-2}(u) & + & \mathbf{V}_{0,3}\; b_{-1}(u) & + & \mathbf{V}_{1,3}\; b_{-0}(u) \\
\mathbf{W}_2 & = & \mathbf{V}_{0,2}\; b_{-3}(u) & + & \mathbf{V}_{0,2}\; b_{-2}(u) & + & \mathbf{V}_{0,2}\; b_{-1}(u) & + & \mathbf{V}_{1,2}\; b_{-0}(u) \\
\mathbf{W}_1 & = & \mathbf{V}_{0,1}\; b_{-3}(u) & + & \mathbf{V}_{0,1}\; b_{-2}(u) & + & \mathbf{V}_{0,1}\; b_{-1}(u) & + & \mathbf{V}_{1,1}\; b_{-0}(u) \\
\mathbf{W}_0 & = & \mathbf{V}_{0,0}\; b_{-3}(u) & + & \mathbf{V}_{0,0}\; b_{-2}(u) & + & \mathbf{V}_{0,0}\; b_{-1}(u) & + & \mathbf{V}_{1,0}\; b_{-0}(u) \quad .
\end{array}$$

However, these four vertices $\mathbf{W}_3$, $\mathbf{W}_2$, $\mathbf{W}_1$ and $\mathbf{W}_0$ are each points on a uniform cubic B-spline curve segment in which the first three vertices are identical. We already know that such a segment lies on the straight line joining the two distinct control vertices involved, and interpolates the triple vertex. Hence at $u=0$

$$\mathbf{Q}_{1,3,0}(v) \;=\; \mathbf{V}_{0,0} b_{-3}(v) + \mathbf{V}_{0,1} b_{-2}(v) + \mathbf{V}_{0,2} b_{-1}(v) + \mathbf{V}_{0,3} b_{-0}(v) \;. \tag{33}$$

But the four control vertices appearing in this equation are exactly the leftmost four vertices in our control graph, which are collinear, and we therefore know that they define a segment of the straight line through $\mathbf{V}_{0,0}$, $\mathbf{V}_{0,1}$, $\mathbf{V}_{0,2}$ and $\mathbf{V}_{0,3}$.

A similar argument establishes that, for this particular surface, all the other boundary curves are straight line segments. Furthermore, since the control vertices along each boundary are collinear, the segments along each boundary are also collinear.

Consider a corner patch, such as the one defined by

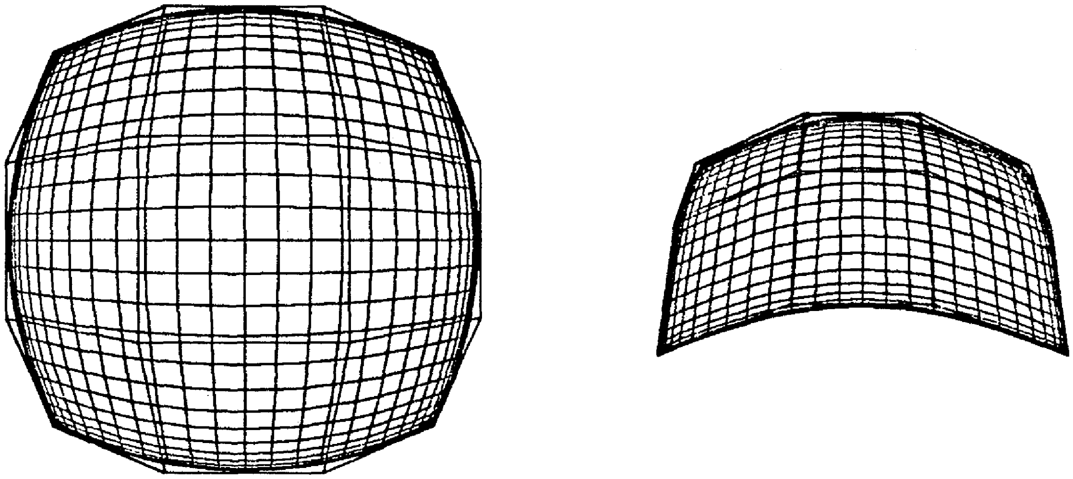
$$\begin{array}{cccc}
\mathbf{V}_{0,1} & \mathbf{V}_{0,1} & \mathbf{V}_{0,1} & \mathbf{V}_{1,1} \\
\mathbf{V}_{0,0} & \mathbf{V}_{0,0} & \mathbf{V}_{0,0} & \mathbf{V}_{1,0} \\
\mathbf{V}_{0,0} & \mathbf{V}_{0,0} & \mathbf{V}_{0,0} & \mathbf{V}_{1,0} \\
\mathbf{V}_{0,0} & \mathbf{V}_{0,0} & \mathbf{V}_{0,0} & \mathbf{V}_{1,0}
\end{array}$$

The left boundary curve of this patch is

$$\mathbf{Q}_{1,1,0}(v) \;=\; \mathbf{V}_{0,0} b_{-3}(v) + \mathbf{V}_{0,0} b_{-2}(v) + \mathbf{V}_{0,0} b_{-1}(v) + \mathbf{V}_{0,1} b_{-0}(v)$$

which, for $v=0$, will interpolate the corner vertex $\mathbf{V}_{0,0}$. Thus the boundary of this surface consists of four straight line segments which join the four corner control vertices.

Unfortunately this behaviour is not very general. We have only to arrange that the boundary vertices not be collinear in order to destroy it, as we illustrate in Figure 39,



From above.

From the front.

Figure 39. This is also an example of a 25 patch surface produced from sixteen control vertices by tripling the boundary vertices. Unlike the previous illustrations, however, these boundary vertices are not coplanar. As a result the surface boundary does not interpolate the boundary of the control polygon.

although the boundary of the surface does closely approximate the periphery of the control graph. It is not hard to see why the failure occurs; if $\mathbf{V}_{0,3}$, $\mathbf{V}_{0,2}$, $\mathbf{V}_{0,1}$ and $\mathbf{V}_{0,0}$ are not collinear then we still have $\mathbf{W}_3 = \mathbf{V}_{0,3}$, $\mathbf{W}_2 = \mathbf{V}_{0,2}$, $\mathbf{W}_1 = \mathbf{V}_{0,1}$ and $\mathbf{W}_0 = \mathbf{V}_{0,0}$, but now (33) simply defines an arbitrary uniform cubic B-spline curve segment. The most that we can conclude is that the four corner vertices of the control graph will be interpolated by the four corners of the surface, since a similar analysis of the four corner patches yields a boundary curve in which the corner vertex is tripled.

Some additional analysis of these boundary conditions appears in [Barsky82]. One can also define phantom vertices by specifying derivative vectors at the boundaries. However, this is probably too cumbersome to be useful. It may occasionally be useful to define phantom vertices which yield zero curvature around the periphery of a patch; again, the details may be found in [Barsky82]

### 4.7.2. Periodic Surfaces

We can "glue together" opposite edges of a surface, in much the same way that we produced closed curves, by simply wrapping the control graph around on itself. By way of example, let's construct something resembling a cylinder. Suppose that $\bar{u}$ is to increase as we move circularly around the cylinder, and the $\bar{v}$ is to increase as we move down the length of the cylinder. We can get a fairly good approximation to a circle by making a closed curve out of four control vertices laid out in a square, using the control vertex sequence

$$\mathbf{V}_0, \mathbf{V}_1, \mathbf{V}_2, \mathbf{V}_3, \mathbf{V}_0, \mathbf{V}_1, \mathbf{V}_2$$

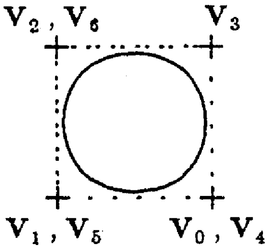


Figure 40. The closed curve defined by four control vertices arranged in a square is approximately a circle. A better approximation can be obtained by distributing more control vertices in a regular circular pattern.

To get a cylinder we simply translate these four control vertices at right angles to the plane in which they lie, making several copies of them as we go. The resulting surface looks like this.

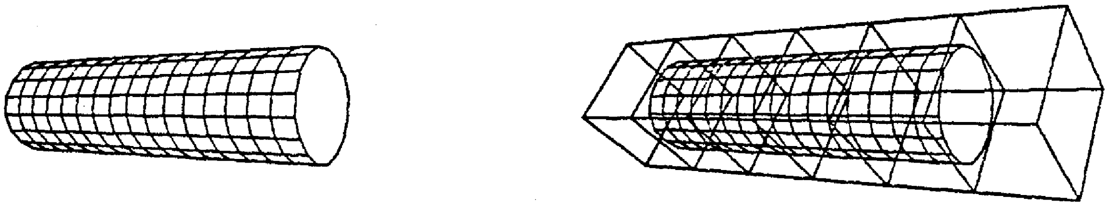

Without the control graph. With the entire control graph.

Figure 41. "Extruding" the control vertices of Figure 40 produces a control graph yielding a reasonable approximation to a cylinder.

Interesting effects can be achieved by pulling some of the "wrapped" points apart. If we take the middle square of control vertices in Figure 41 and pull the two vertices $\mathbf{V}_0$ and $\mathbf{V}_1$ at one end away from the two vertices $\mathbf{V}_5$ and $\mathbf{V}_6$ at the other end to yield

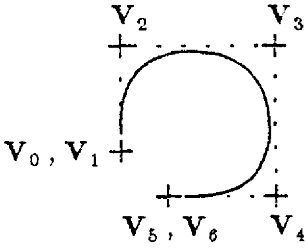


Figure 42.

without moving the other control vertices, we obtain the following surface.

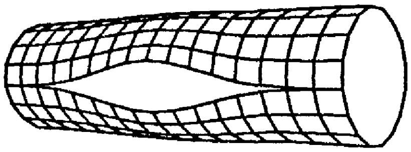


Figure 43. If we take the control graph shown in Figure 41 and separate the midmost plane of control vertices as in Figure 42 so that by themselves they no longer define a closed curve, we open a hole in the cylinder.

We can also "glue" the ends of the control vertex array together at the same time we are wrapping the sides together. If we take the same square of four control vertices, but now revolve it circularly in space at some distance from the origin, we can define a torus.

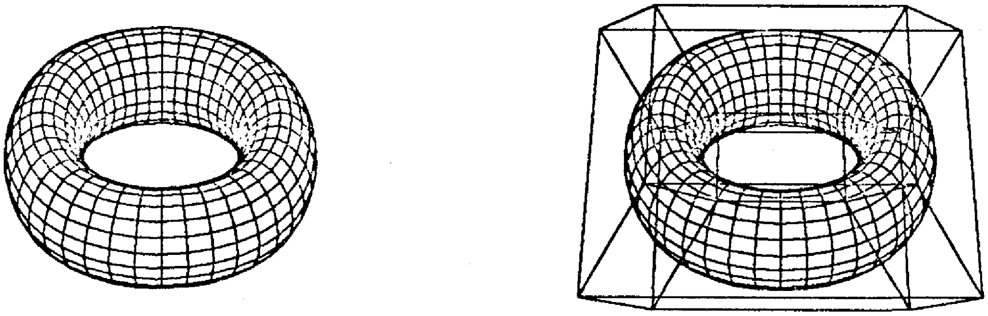


Figure 44. This surface is defined by treating the control vertex array periodically in both $\bar{u}$ and $\bar{v}$.

# 5. Interlude

Thus far our treatment of splines has focused on the uniform cubic B-splines. This has enabled us to introduce most of the concepts in which we are interested in a fairly simple setting which minimized the complexity of the mathematics involved. Moreover, the uniform cubic B-splines are of substantial interest in their own right since they can be efficiently computed and suffice for many applications.

Nevertheless, there are a number of ways in which we might seek to generalize them to obtain greater flexibility and power:

- we might use polynomials of some other degree;
- we might use an irregularly spaced (*nonuniform*) knot sequence;
- we might wish to impose something other than $C^2$ continuity at the joints between successive curve segments.

In principle we could accomplish all of these by simply repeating the development of Section 4, defining the appropriate constraint equations and solving them for the coefficients of the basis segments involved. In fact this might be satisfactory for some especially interesting special cases (such as uniform quintic B-splines). It would be terribly cumbersome, however, to carry this out every time the user wished to try the effect of a different order spline, or of changing the knot spacing. Fortunately we can do better.

It is a remarkable fact, resulting principally from the work of Isaac Schoenberg, Carl de Boor and Maurice Cox [deBoor72, deBoor78, Cox71], that there exists a single unified algorithm by which all three of these generalizations of the uniform cubic B-splines can be accomplished.

Our next objective, then, is to develop a general treatment of B-splines of arbitrary order $k$, defined over irregularly spaced knot sequences, and with any of $C^0$ through $C^{k-2}$ continuity at the joints between segments (or the borders between surface patches). Not surprisingly, this development will be more involved than was our treatment of the uniform cubic B-splines. It does not, however, require anything more than a careful consideration of easily understood properties of polynomials, and so is readily accessible to the careful reader. It is quite beautiful as well.

# 6. Splines in a More General Setting

### 6.1. Preliminaries

By way of introduction we will begin with a few examples, explaining much of the vocabulary that we will need later. Most of what we will do in the following theoretical sections parallels the developments in [Schumaker81], in [deBoor78], and finally in [Cohen80], though with a greater emphasis on intuition and at a much lower level of rigour and formality.

We will carry out our theoretical discussions purely in terms of the variable $\bar{u}$, never reparametrizing the separate segment polynomials into [0,1]. For example, we will now represent the uniform cubic B-spline of (11) on the knot sequence 0,1,2,3,4 by

$$B_0(\bar{u}) = \begin{cases} \quad\quad 0 & -\infty < \bar{u} < 0 \\ b_{-0}(\bar{u}) = \frac{1}{6}\bar{u}^3 & 0 \le \bar{u} < 1 \\ b_{-1}(\bar{u}) = -\frac{1}{6}(3\bar{u}^3-12\bar{u}^2+12\bar{u}-4) & 1 \le \bar{u} < 2 \\ b_{-2}(\bar{u}) = \frac{1}{6}(3\bar{u}^3-24\bar{u}^2+60\bar{u}-44) & 2 \le \bar{u} < 3 \\ b_{-3}(\bar{u}) = -\frac{1}{6}(\bar{u}^3-12\bar{u}^2+48\bar{u}-64) & 3 \le \bar{u} < 4 \\ \quad\quad 0 & 4 \le \bar{u} < +\infty \ . \end{cases}$$

The splines we have discussed so far have usually involved cubic polynomials, so for variety we will use quadratics in the next few paragraphs.

A typical segment polynomial of degree two has the form

$$p(\bar{u}) = c_0 + c_1\bar{u} + c_2\bar{u}^2 \ .$$

If we are to use quadratics, or any other polynomials, to concoct splines, we must carry out, explicitly or implicitly, the sort of construction that we used to form the uniform B-splines. That means, having selected some knot $\bar{u}=\bar{u}_i$ and having decided that

$$p_{left}(\bar{u}) = c_{00} + c_{01}\bar{u} + c_{02}\bar{u}^2$$

and

$$p_{right}(\bar{u}) = c_{10} + c_{11}\bar{u} + c_{12}\bar{u}^2$$

will meet at $\bar{u}_i$ with a certain continuity, we are left with the problem of imposing conditions on the coefficients

$$c_{00}\,,\ c_{01}\,,\ c_{02}\,,\ c_{10}\,,\ c_{11}\,,\ \text{and } c_{12}$$

so that the meeting takes place as desired. One possibility, of course, is to request $C^2$ continuity:

$$\begin{aligned}
p_{left}^{(0)}(\bar{u}_i) &= p_{right}^{(0)}(\bar{u}_i)\\
p_{left}^{(1)}(\bar{u}_i) &= p_{right}^{(1)}(\bar{u}_i)\\
p_{left}^{(2)}(\bar{u}_i) &= p_{right}^{(2)}(\bar{u}_i)\ .
\end{aligned}$$

This can only be accomplished by imposing the conditions

$$\begin{aligned}
c_{00} + c_{01}\bar{u}_i + c_{02}\bar{u}_i^2 &= c_{10} + c_{11}\bar{u}_i + c_{12}\bar{u}_i^2\\
c_{01} + 2c_{02}\bar{u}_i &= c_{11} + 2c_{12}\bar{u}_i\\
2c_{02} &= 2c_{12}\ .
\end{aligned}$$

This clearly results in

$$p_{left}(\bar{u}) \equiv p_{right}(\bar{u})\ ,$$

which is a completely uninteresting outcome; the knot $\bar{u}_i$ might as well not exist. The remaining possibilities are $C^1$ continuity,

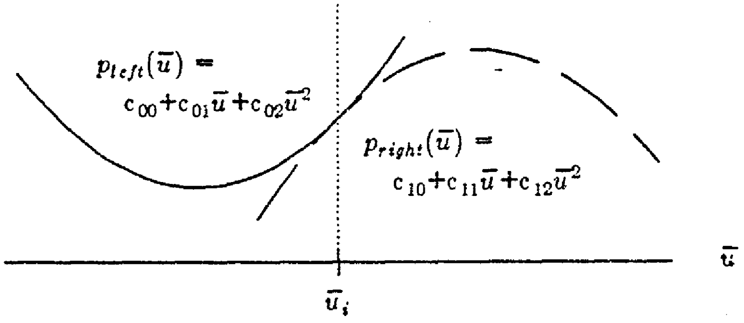

$p_{left}(\bar{u}) =$
$c_{00}+c_{01}\bar{u}+c_{02}\bar{u}^2$
$p_{right}(\bar{u}) =$
$c_{10}+c_{11}\bar{u}+c_{12}\bar{u}^2$
$\bar{u}$
$\bar{u}_i$

Figure 45. $C^1$ or first derivative continuity at $\bar{u}_i$.

$C^0$ continuity,

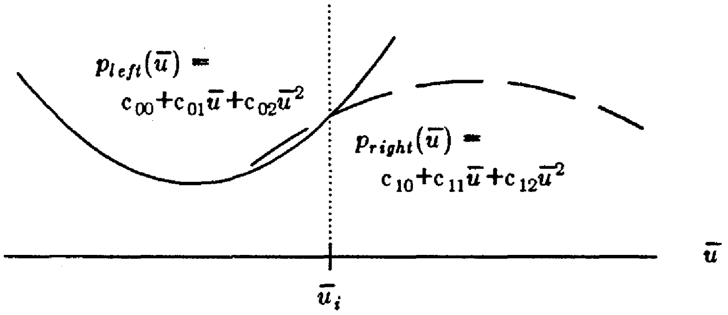

$p_{left}(\bar{u}) =$
$c_{00}+c_{01}\bar{u}+c_{02}\bar{u}^2$
$p_{right}(\bar{u}) =$
$c_{10}+c_{11}\bar{u}+c_{12}\bar{u}^2$
$\bar{u}$
$\bar{u}_i$

Figure 46. $C^0$ or positional continuity at $\bar{u}_i$, with a failure of first derivative continuity.

and even the case of no continuity at all (which we will refer to as $C^{-1}$ continuity).

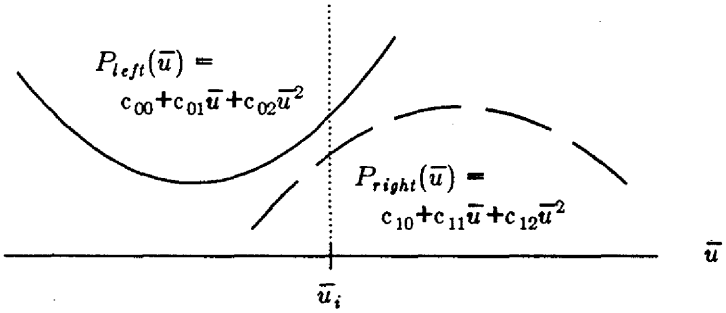

$P_{left}(\bar{u}) =$
$c_{00}+c_{01}\bar{u}+c_{02}\bar{u}^2$
$P_{right}(\bar{u}) =$
$c_{10}+c_{11}\bar{u}+c_{12}\bar{u}^2$
$\bar{u}$
$\bar{u}_i$

Figure 47. A discontinuity at $\bar{u}_i$. $p_{left}(\bar{u})$ and $p_{right}(\bar{u})$ are said to be $C^{-1}$ continuous at $\bar{u}_i$.

What becomes clear is that, in imposing the conditions that force $p_{right}$ to join $p_{left}$ with some degree of continuity, the very fact that three coefficients are involved limits what is possible and what is interesting. In the cubic case there were four coefficients, so $C^2$ continuity was both interesting and attainable, and we should expect that for polynomials having $k$ coefficients, i.e. having highest power at most $k-1$, only continuities $C^{-1}$ through $C^{k-2}$ are both interesting and attainable. In particular, imposing $C^{k-1}$ continuity is uninteresting because it forces $p_{left}(\bar{u})$ and $p_{right}(\bar{u})$ to be identical, resulting not only in $C^{k-1}$ continuity, but also $C^k$, $C^{k+1}$, . . . , $C^\infty$ continuity.

To understand splines properly, it will be necessary to review a few facts about polynomials first.

### 6.2. Polynomials

We will focus our attention on $k^{th}$-*order polynomials*. By this we mean polynomials having precisely $k$ coefficients, with zero coefficients allowed. For example, polynomials of order 4 consist of all functions $p(\bar{u})$ of the variable $\bar{u}$ which can be written in the form

$$p(\bar{u}) \;=\; c_0 + c_1\bar{u} + c_2\bar{u}^2 + c_3\bar{u}^3 \;,$$

and this is intended to include the cases in which

$$c_3{=}0 \text{ and/or } c_2{=}0 \text{ and/or } c_1{=}0 \text{ and/or } c_0{=}0 \;\;.$$

As a consequence, $4^{th}$-order polynomials are the polynomials having at most degree 3. This means that the $k^{th}$-order polynomials include all polynomials up to and including those of degree $k-1$.

**Notation:** $\mathbf{P}^k$ stands for *the set of all* $k^{\text{th}}$*-order polynomials*, i.e. all functions of a real variable $\bar{u}$ which can be represented as

$$p(\bar{u}) = \sum_{i=0}^{k-1} c_i \bar{u}^i$$

for any choice of real constants $c_0, \ldots, c_{k-1}$.

### 6.3. Vector Spaces

To be able to treat polynomials and splines more generally, we will draw a great deal from the concept of a vector space. To review: a vector space over the real numbers is any collection of objects for which there are defined operations of

- *vector addition* between any two members of the collection, yielding a member of the collection, and
- *scalar multiplication* between any real number and any member of the collection, yielding a member of the collection.

These operations must satisfy a number of algebraic conditions: Suppose $Z, Y, X, \cdots$ stand for the objects in the collection, suppose $a, b, c, \cdots$ stand for real numbers, and suppose the operations of scalar multiplication and vector addition are denoted in a natural way; i.e. $aX$ and $X+Y$. It is required that the following hold:

- $X+Y = Y+X$ (commutativity of addition);
- $(X+Y)+Z = X+(Y+Z)$ (associativity of addition);
- there is a "zero" vector $\Theta$ having the property that $X+\Theta = X$ for all $X$ (the additive identity element);
- $a(X+Y) = (aX)+(aY)$ (distributivity of scalar multiplication over addition);
- $(a+b)X = (aX)+(bX)$ (distributivity of addition over scalar multiplication);
- $(ab)X = a(bX)$ (associativity of scalar multiplication);
- $1X = X$ for all $X$ (the multiplicative identity element).

The question of what object constitutes the zero vector is of particular interest, because in the answer lies the *de facto* definition of what constitutes the equality (i.e. the equivalence or indistinguishability) of two vectors. This is a remark which will have particular relevance to splines.

The usual example one has in mind for a vector space is 3-space, i.e. the collection of all objects of the form

$$P = (x, y, z)$$

where $x$, $y$, and $z$ are real numbers. Vector addition follows the familiar format:

$$\begin{aligned} P_1 &= (x_1, y_1, z_1) \\ P_2 &= (x_2, y_2, z_2) \\ P_1 + P_2 &= (\, x_1+x_2,\, y_1+y_2,\, z_1+z_2 \,) \ . \end{aligned}$$

Scalar multiplication ("scaling") follows the format

$$P = (x, y, z)$$
$$aP = (ax, ay, az) \ .$$

The zero vector is, of course,

$$\Theta = (0,0,0) = \mathbf{0} \ ,$$

which implies that $P_1 = P_2$ if and only if $x_1 = x_2$, $y_1 = y_2$, and $z_1 = z_2$.

3-space is only one example. The definition of a vector space is general enough and powerful enough to include many different types of objects. For our deliberations it will be important to observe that $k^{\text{th}}$-order polynomials (and later, $k^{\text{th}}$-order splines on a fixed knot sequence) constitute a vector space.

### 6.4. Polynomials as a Vector Space

It is easily seen that the set of all $k^{\text{th}}$-order polynomials

$$X(\bar{u}), Y(\bar{u}), Z(\bar{u}), \ \cdots$$

form a vector space, since the conventional addition of such polynomials

$$\begin{aligned} X(\bar{u}) + Y(\bar{u}) &\equiv (c_0 + \cdots + c_{k-1}\bar{u}^{k-1}) + (d_0 + \cdots + d_{k-1}\bar{u}^{k-1}) \\ &= (c_0 + d_0) + \cdots + (c_{k-1} + d_{k-1})\bar{u}^{k-1} \end{aligned}$$

and the conventional scaling of such polynomials

$$\begin{aligned} aX(\bar{u}) &\equiv a(c_0 + \cdots + c_{k-1}\bar{u}^{k-1}) \\ &= (ac_0) + \cdots + (ac_{k-1})\bar{u}^{k-1} \end{aligned}$$

satisfy all the rules listed for vector spaces. The polynomial corresponding to $\Theta$ is, of course,

$$\Theta(\bar{u}) = 0 + 0\bar{u} + \cdots + 0\bar{u}^{k-1} \ ,$$

which implies that $X(\bar{u}) = Y(\bar{u})$ if and only if $X$ and $Y$ have precisely the same coefficients. Stating this formally:

**Theorem:** For any $k > 0$, $\mathbf{P}^k$ is a vector space with the usual definitions of polynomial addition and of multiplication by a real number playing the roles, respectively, of vector addition and scalar multiplication.

Polynomials can even be written so as to look like ordinary vectors; that is, they can be written as *k-tuples* of numbers. Since a polynomial is completely determined by its "powers of $\bar{u}$ coefficients," for example, we could write

$$X(\bar{u}) \equiv (c_0, c_1, c_2)$$

with the interpretation that

$$(c_0, c_1, c_2) \equiv c_0(1,0,0) + c_1(0,1,0) + c_2(0,0,1)$$

and, of course,

$$
\begin{aligned}
(1,0,0) &\equiv 1\bar{u}^0 + 0\bar{u}^1 + 0\bar{u}^2 = \bar{u}^0 \\
(0,1,0) &\equiv 0\bar{u}^0 + 1\bar{u}^1 + 0\bar{u}^2 = \bar{u}^1 \\
(0,0,1) &\equiv 0\bar{u}^0 + 0\bar{u}^1 + 1\bar{u}^2 = \bar{u}^2 \ .
\end{aligned}
$$

The mechanics of this interpretation of polynomials as vectors requires two things:

- that we have chosen *coordinate system* or, as we will refer to it, a *basis*, and
- that the $k$-tuple of coefficients given as the description of the polynomial is, in fact, a correct representation in terms of the chosen basis.

These points are worth raising because many coordinate systems (bases) are possible in any given vector space. In the case of polynomials, the following illustration is worth considering. It is easily verified (by expanding the quantities in parentheses and collecting the terms together according to powers of $\bar{u}$) that the polynomial given by

$$5(\bar{u}-1)^2 + 4(\bar{u}-2)^2 + 3(\bar{u}-3)^2$$

is exactly the same as the polynomial given by

$$48\bar{u}^0 - 44\bar{u}^1 + 12\bar{u}^2 \ .$$

In the first case the basis

$$(\bar{u}-1)^2 \ , \ (\bar{u}-2)^2 \ , \ (\bar{u}-3)^2$$

and the coefficients

$$5 \ , \ 4 \ , \ 3$$

go together. The polynomial can be expressed as the 3-tuple $(5,4,3)$ in the "$(\bar{u}-1)^2,(\bar{u}-2)^2,(\bar{u}-3)^2$" coordinate system. In the second case the basis

$$\bar{u}^0 \ , \ \bar{u}^1 \ , \ \bar{u}^2$$

and the coefficients

$$48 \ , \ -44 \ , \ 12$$

go together. The polynomial can be expressed as the 3-tuple $(48,-44,12)$ in the "$\bar{u}^0,\bar{u}^1,\bar{u}^2$" coordinate system.

It would clearly be inappropriate to take the 3-tuple (5,4,3) and to interpret it as a mechanism for describing this polynomial in terms of powers of $\bar{u}$, but it is a valid description of the polynomial in terms of another basis. Thus, we should usually expect to say that some $k$-tuple $(c_0,\ldots,c_{k-1})$ represents some vector $P$ of some vector space $\mathbf{V}$ with respect to some basis $B_0,\ldots,B_{k-1}$. General practice is often less precise than this. When a basis is not explicitly mentioned, some "canonical" basis is understood by common agreement. Thus, for polynomials, we generally interpret a vector

$$(c_0, \ . \ . \ . \ , c_{k-1})$$

as a polynomial in terms of the "powers basis" when no basis is explicitly designated, i.e.

$$c_0\bar{u}^0 + \cdots + c_{k-1}\bar{u}^{k-1} \ \ ,$$

just as in our usual notion of $k$-space we use the "unit coordinates"

$$(1,...,0)\ ,\ .\ .\ .\ ,\ (0,...,1)$$

to interpret a list of numbers

$$(c_0,...,c_{k-1})$$

as a vector

$$c_0(1,...,0) + \cdots + c_{k-1}(0,...,1)\ .$$

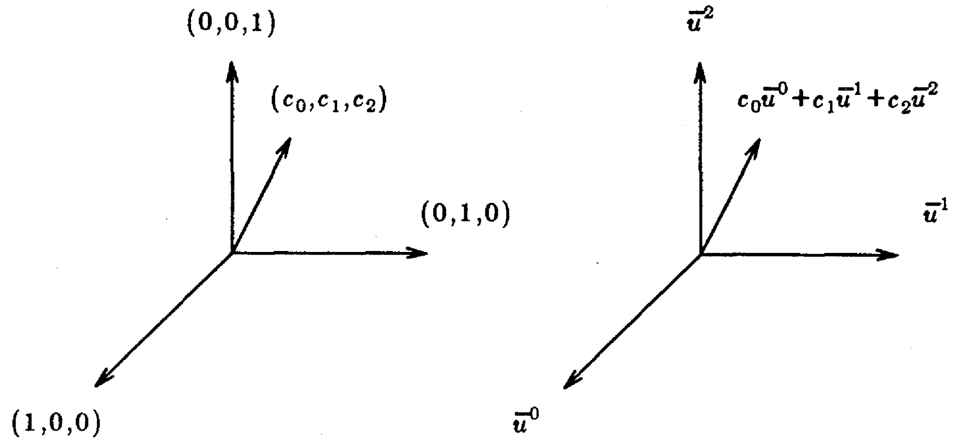


Figure 48. Canonical coordinate systems (bases).

In the following section we expand on the notion of a basis and of the relationships between alternative bases.

## 6.5. Bases and Dimension

To understand what a basis is for a general vector space, it is important to recall the concept of *linear independence*:

Vectors $Z, Y, X, \cdots$ are *linearly independent* if the only scalars $a, b, c, \cdots$ for which

$$aX + bY + \cdots + cZ = \Theta$$

are the trivial ones

$$a = b = \cdots = c = 0\ .$$

A consequence of this is that if some vector $W$ can be represented in terms of linearly independent vectors $X, Y, \ldots, Z$:

$$W = rX + sY + \cdots + tZ$$

then the coefficients in the combination $r, s, \ldots, t$ are uniquely defined.

For an arbitrary $W$ and an arbitrary collection $X, Y, \ldots, Z$ we can't guarantee that $W$ has a representation in terms of the collection, even if the collection is linearly independent. For example $\vec{u}^3$ can't be expressed in terms of $\vec{u}^0$, $\vec{u}^1$, and $\vec{u}^2$ alone. Collections of linearly independent vectors that have the power of representing *every* vector in a space have a special importance.

**Definition:** A *basis* of a vector space is a collection of vectors that is linearly independent and that can express any vector in the space as a linear combination.

Some spaces are so "rich" that they can only be expressed in terms of infinite collections of vectors. Any space of interest to us, however, will be generated by a finite collection of basis vectors. Bases are not unique, but if one finite basis (containing, say, $k$ vectors) for a space exists, then any other basis for that space must also be comprised of exactly $k$ vectors.

**Definition:** The number $k$ (if finite) of the elements in any basis for a vector space, does not depend on the basis. This number is the same for all bases, and it is called the ***dimension*** of the space.

The dimension of 3-space is, of course, 3. The dimension of $\mathbf{P}^k$ is $k$.

As we have already remarked, we will ultimately discover that splines, as well as polynomials, form vector spaces. The B-splines will form our canonical basis. From the point of view of graphics and the construction of objects using splines or polynomials, the dimension of a space may be thought of as indicating the number of "controls" which may be varied to obtain distinct members of the space. The control variables may be regarded as the coefficients of the basis elements

$$a_0 , \ldots , a_m$$

and each is "independent" in the sense that it can be varied by itself to obtain new vectors not obtainable by using any of the other parameters. We will eventually associate the dimension of a spline space with the number of control *vertices* that may be used to construct the spline curves or surfaces in that space.

### 6.6. Change of Basis

The modeling transformations of graphics lead one to confront different bases which define the same space. For example, if 3-space is represented in terms of the basis

$$(1,0,0)\ ,\ (0,1,0)\ ,\ (0,0,1)$$

and this space is subjected to some transformation, $\mathbf{A}$, then the basis undergoes the change

$$(1,0,0)\mathbf{A} \rightarrow (t_{00},t_{01},t_{02})$$
$$(0,1,0)\mathbf{A} \rightarrow (t_{10},t_{11},t_{12})$$
$$(0,0,1)\mathbf{A} \rightarrow (t_{20},t_{21},t_{22})\ .$$

Under what conditions will

$$(t_{00},t_{01},t_{02})\ ,\ (t_{10},t_{11},t_{12})\ ,\ (t_{20},t_{21},t_{22})$$

represent a new coordinate system? The necessary and sufficient condition for these transformed vectors to be a basis is that $\mathbf{A}$ has to be *nonsingular* (that is *invertible*).

More generally, if

$$X_0 , \ldots , X_m$$

is one basis for a vector space and

$$Y_0 , \ldots , Y_m$$

is another, then each vector in the one basis can be expressed as a linear combination of the vectors in the other basis; e.g.

$$\begin{aligned} X_j &= a_{0,j}Y_0 + \cdots + a_{m,j}Y_m \\ &= \sum_{i=0}^{m} a_{i,j}Y_i \end{aligned}$$

for $j=0,1, \cdots ,m$. Conversely, if $Y_0, \ldots , Y_m$ is a basis for a vector space, and if coefficients $a_{i,j}$ are chosen to create linearly independent combinations

$$X_j = a_{0,j}Y_0 + \cdots + a_{m,j}Y_m \quad \text{for } j=0, \ldots ,m$$

then these combinations, $X_j$, will also be a basis. The constants $a_{i,j}$ appearing in the above assertions make up the *change of basis matrix*

$$\mathbf{A} = \begin{bmatrix} a_{00} & \cdots & a_{0m} \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ a_{m0} & \cdots & a_{mm} \end{bmatrix} ;$$

that is,

$$\begin{bmatrix} X_0 ,\ldots, X_m \end{bmatrix} = \begin{bmatrix} Y_0 ,\ldots, Y_m \end{bmatrix} \mathbf{A}$$

Note: A change of basis matrix must be nonsingular.

Notice that normal 3-space could be represented by the basis

$$X_0 = (1,0,0), \; X_1 = (1,1,0), \; X_2 = (1,1,1)$$

as well as by the basis

$$Y_0 = (1,0,0), \; Y_1 = (0,1,0), \; Y_2 = (0,0,1) \; .$$

The 3-tuples representing the $X$'s and $Y$'s are to be interpreted in terms of the canonical 3-space coordinates:

$$\begin{aligned} X_0 &= 1(1,0,0) + 0(0,1,0) + 0(0,0,1) \\ X_1 &= 1(1,0,0) + 1(0,1,0) + 0(0,0,1) \\ X_2 &= 1(1,0,0) + 1(0,1,0) + 1(0,0,1) \end{aligned}$$

and, trivially,

$$\begin{aligned} Y_0 &= 1(1,0,0) + 0(0,1,0) + 0(0,0,1) \\ Y_1 &= 0(1,0,0) + 1(0,1,0) + 0(0,0,1) \end{aligned}$$

$$Y_2 = 0(1,0,0) + 0(0,1,0) + 1(0,0,1) \ .$$

This means that

$$\Big[(1,0,0)\ ,\ (1,1,0)\ ,\ (1,1,1)\Big] = \Big[(1,0,0)\ ,\ (0,1,0)\ ,\ (0,0,1)\Big]\begin{bmatrix}1&0&0\\1&1&0\\1&1&1\end{bmatrix} ,$$

and conversely,

$$\Big[(1,0,0)\ ,\ (0,1,0)\ ,\ (0,0,1)\Big] = \Big[(1,0,0)\ ,\ (1,1,0)\ ,\ (1,1,1)\Big]\begin{bmatrix}1&0&0\\-1&1&0\\0&-1&1\end{bmatrix} .$$

Or, for example in more expanded format, this means that

$$(0,1,0) = (1,1,0)\begin{bmatrix}1&0&0\\-1&1&0\\0&-1&1\end{bmatrix} .$$

For another example, the power basis for the quadratic polynomials

$$\bar{u}^0\ ,\ \bar{u}^1\ ,\ \text{and}\ \bar{u}^2$$

could be replaced by

$$(\bar{u}-t)^0\ ,\ (\bar{u}-t)^1\ ,\ \text{and}\ (\bar{u}-t)^2 \qquad (34)$$

for any fixed value of $t$.

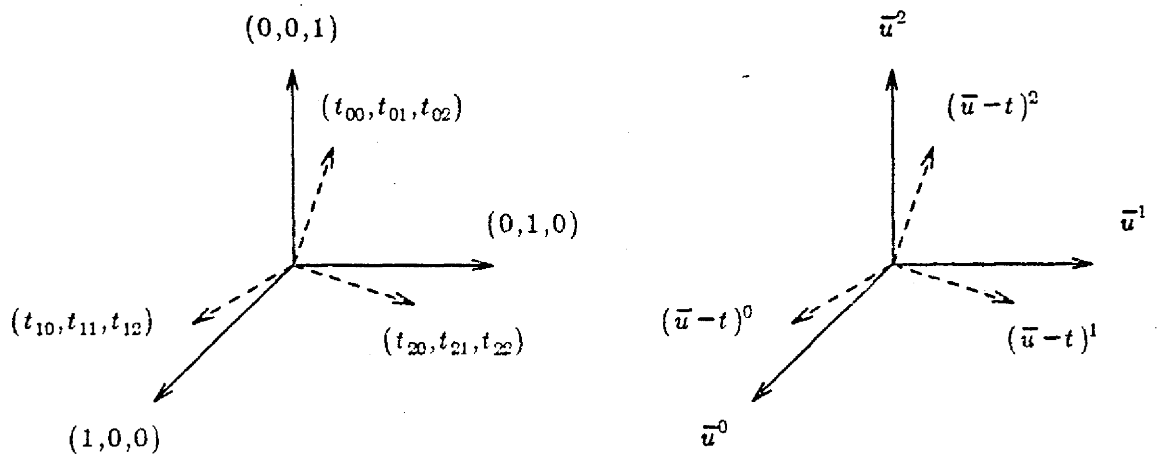


Figure 49. Two bases (coordinate systems) for 3-space or quadratics.

The change of basis matrix for this alternative to the power basis for quadratics is given by

$$\mathbf{A} = \begin{bmatrix}1&-t&t^2\\0&1&-2t\\0&0&1\end{bmatrix} .$$

Since this matrix has a determinant equal to 1, it is nonsingular, and this verifies that (34) is, indeed, a legitimate basis.

Another important example of an alternative polynomial basis is

$$(\bar{u}-\bar{u}_i)^2\ ,\ (\bar{u}-\bar{u}_{i+1})^2\ ,\ \text{and}\ (\bar{u}-\bar{u}_{i+2})^2\ , \tag{35}$$

which will be a legitimate basis when $\bar{u}_i$, $\bar{u}_{i+1}$, and $\bar{u}_{i+2}$ are distinct numbers. We will have many occasions to consider extensions of (34) and (35) to $k^{\text{th}}$-order polynomials in the material which follows.

As a final example, this time for the cubics, the power basis

$$\bar{u}^0\ ,\ \bar{u}^1\ ,\ \bar{u}^2\ ,\ \bar{u}^3$$

could be replaced by the linear combinations

$$\begin{aligned}
-\bar{u}^3+3\bar{u}^2-3\bar{u}+1 &= (1-\bar{u})^3 &&= P_{0,3}\\
3\bar{u}^3-6\bar{u}^2+3\bar{u} &= 3\bar{u}(1-\bar{u})^2 &&= P_{1,3}\\
3\bar{u}^2-3\bar{u}^3 &= 3\bar{u}^2(1-\bar{u}) &&= P_{2,3}\\
\bar{u}^3 &= \bar{u}^3 &&= P_{3,3}
\end{aligned}$$

to obtain an alternative basis for the cubics. The change of basis equations for this example are usually written

$$\begin{bmatrix}\bar{u}^3 & \bar{u}^2 & \bar{u}^1 & \bar{u}^0\end{bmatrix}\begin{bmatrix}1 & -3 & 3 & -1\\ 0 & 3 & -6 & 3\\ 0 & 0 & 3 & -3\\ 0 & 0 & 0 & 1\end{bmatrix} = \begin{bmatrix}P_{3,3} & P_{2,3} & P_{1,3} & P_{0,3}\end{bmatrix}\ .$$

(The members of this new basis for the cubics are called the "Bernstein polynomials" — they are used to define "Be´zier curves," which will be discussed in Chapter 11.)

### 6.7. An Alternative Polynomial Basis

Thus far we have institutionalized the *power basis* for $\mathbf{P}^k$, i.e. the functions

$$\begin{aligned}
q_0(\bar{u}) &= \bar{u}^0\\
q_1(\bar{u}) &= \bar{u}^1\\
&\ \ \vdots\\
q_{k-1}(\bar{u}) &= \bar{u}^{k-1}\ ,
\end{aligned}$$

but we have indicated that other bases are also interesting. If $\bar{u}_0\ ,\ \ldots\ ,\ \bar{u}_{k-1}$ are distinct values for $\bar{u}$, consider functions of the sort mentioned in (35) above:

$$\begin{aligned}
r_0(\bar{u}) &= (\bar{u}-\bar{u}_0)^{k-1}\\
r_1(\bar{u}) &= (\bar{u}-\bar{u}_1)^{k-1}\\
&\ \ \vdots\\
r_{k-1}(\bar{u}) &= (\bar{u}-\bar{u}_{k-1})^{k-1}\ .
\end{aligned} \tag{36}$$

These, too, constitute a basis for $\mathbf{P}^k$. This may be verified by observing that each power of $\bar{u}$,

$$\bar{u}^0, \ldots, \bar{u}^{k-1} \ ,$$

can be expressed in terms of these functions, i.e. constants $a_{i,j}$ can be found such that

$$a_{i,0}r_0(\bar{u}) + \cdots + a_{i,k-1}r_{k-1}(\bar{u}) = q_i(\bar{u}) = \bar{u}^i \quad \text{for } i=0, \ldots, k-1 \ .$$

We will not establish this formally but will merely give as example which suggests the general pattern. Consider $k=3$ and $i=0$. In order to represent $\bar{u}^0$, $a_{0,0}$, $a_{0,1}$, and $a_{0,2}$ must satisfy

$$a_{0,0}(\bar{u}-\bar{u}_0)^2 + a_{0,1}(\bar{u}-\bar{u}_1)^2 + a_{0,2}(\bar{u}-\bar{u}_2)^2 = \bar{u}^0 = 1 \ .$$

Expanding the terms in parentheses and looking at each power of $\bar{u}$ separately on the left-hand side of the equation, it is evident that the $a$'s must satisfy the equalities

$$\begin{aligned}
\bar{u}^2 a_{0,0} &+ \bar{u}^2 a_{0,1} + \bar{u}^2 a_{0,2} = 0\cdot\bar{u}^2 = 0 \\
-2\bar{u}\bar{u}_0 a_{0,0} &- 2\bar{u}\bar{u}_1 a_{0,1} - 2\bar{u}\bar{u}_2 a_{0,2} = 0\cdot\bar{u}^1 = 0 \\
\bar{u}_0^2 a_{0,0} &+ \bar{u}_1^2 a_{0,1} + \bar{u}_2^2 a_{0,2} = 1\cdot\bar{u}^0 = 1 \ .
\end{aligned}$$

This system is uninteresting for $\bar{u}=0$, so we assume that $\bar{u}$ is nonzero, and we divide the first equation by $\bar{u}^2$ and the second by $-2\bar{u}$, which produces

$$\begin{bmatrix} 1 & 1 & 1 \\ \bar{u}_0 & \bar{u}_1 & \bar{u}_2 \\ \bar{u}_0^2 & \bar{u}_1^2 & \bar{u}_2^2 \end{bmatrix} \begin{bmatrix} a_{0,0} \\ a_{0,1} \\ a_{0,2} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \ .$$

The coefficients of the resulting system of linear equations form the well-known *Vandermonde matrix of order* 2

$$\begin{bmatrix} 1 & 1 & 1 \\ \bar{u}_0 & \bar{u}_1 & \bar{u}_2 \\ \bar{u}_0^2 & \bar{u}_1^2 & \bar{u}_2^2 \end{bmatrix}$$

whose determinant is

$$(\bar{u}_0-\bar{u}_1)(\bar{u}_0-\bar{u}_2)(\bar{u}_1-\bar{u}_2) \ .$$

Thus, since $\bar{u}_0$, $\bar{u}_1$, and $\bar{u}_2$ are distinct by assumption, the determinant is not zero; hence, the system is nonsingular and can be solved. Finding constants to represent $\bar{u}$ and $\bar{u}^2$ in terms of the $r$ basis leads to exactly the same system matrix.

For general $k$, the matrix which arises in this exercise is the *Vandermonde matrix of order k*

$$\begin{bmatrix} 1 & 1 & \cdots & 1 \\ \bar{u}_0 & \bar{u}_1 & \cdots & \bar{u}_{k-1} \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \bar{u}_0^{k-1} & \bar{u}_1^{k-1} & \cdots & \bar{u}_{k-1}^{k-1} \end{bmatrix}$$

whose determinant is

$$(\bar{u}_0-\bar{u}_1)(\bar{u}_0-\bar{u}_2)\cdots(\bar{u}_0-\bar{u}_{k-1})(\bar{u}_1-\bar{u}_2)\cdots(\bar{u}_{k-2}-\bar{u}_{k-1})$$
$$= \prod_{i<j}(\bar{u}_i-\bar{u}_j) \ .$$

Suppose the values $\bar{u}_0, \ldots, \bar{u}_{k-1}$ are not distinct. For example, let $\bar{u}_i$ and $\bar{u}_{i+1}$ be brought together, so that

$$\bar{u}_{i-1} < \bar{u}_i = \bar{u}_{i+1} < \bar{u}_{i+2} \ .$$

Then it will no longer do to use $r_0, \ldots, r_{k-1}$ as a basis, since

$$r_i(\bar{u}) = (\bar{u}-\bar{u}_i)^{k-1} = r_{i+1}(\bar{u}) = (\bar{u}-\bar{u}_{i+1})^{k-1}$$

and this destroys linear independence. We will see that

$$(\bar{u}-\bar{u}_i)^{k-1} \quad \text{and} \quad (\bar{u}-\bar{u}_{i+1})^{k-2} = (\bar{u}-\bar{u}_i)^{k-2}$$

can be used in place of

$$(\bar{u}-\bar{u}_i)^{k-1} \quad \text{and} \quad (\bar{u}-\bar{u}_{i+1})^{k-1}$$

to produce an alternative basis. If more knots are brought together, i.e.

$$\bar{u}_{i-1} < \bar{u}_i = \cdots = \bar{u}_{i+j} < \bar{u}_{i+j+1} \ ,$$

then the pattern continues:

$$(\bar{u}-\bar{u}_i)^{k-1} \ , (\bar{u}-\bar{u}_i)^{k-2} \ , \ldots, \ (\bar{u}-\bar{u}_i)^{k-1-j}$$

are used in place of

$$(\bar{u}-\bar{u}_i)^{k-1} \ , (\bar{u}-\bar{u}_{i+1})^{k-1} \ , \ldots, \ (\bar{u}-\bar{u}_{i+j})^{k-1} \ .$$

Of course, we need take no more than the first $k$ of these terms if $j \geq k-1$, since $\mathbf{P}^k$ has dimension $k$.

Note that the alternative basis being taken when some of the $\bar{u}_i$ values are repeated relates to the derivatives of $(\bar{u}-\bar{u}_i)^{k-1}$:

$$\begin{array}{lcc}
r_i^{(0)}(\bar{u}) & = & (\bar{u}-\bar{u}_i)^{k-1} \\
r_i^{(1)}(\bar{u}) & = & (k-1)(\bar{u}-\bar{u}_i)^{k-2} \\
\vdots & \vdots & \vdots \\
r_i^{(j)}(\bar{u}) & = & (k-1)\cdots(k-j)(\bar{u}-\bar{u}_i)^{(k-j-1)} \\
\vdots & \vdots & \vdots \\
r_i^{(k-1)}(\bar{u}) & = & (k-1)\cdots(3)(2)(1)(\bar{u}-\bar{u}_i)^0 = (k-1)!(\bar{u}-\bar{u}_i)^0 \ .
\end{array} \tag{37}$$

It is as if pushing knots together demanded that derivatives be taken as a compensation. This connection between multiple knots and derivatives will resurface later, when we define the $k^{\text{th}}$-order B-splines in Chapter 9.

If all the functions of (37) are used for some fixed $\bar{u}_i = t$, they constitute a basis for the $k^{\text{th}}$-order polynomials. We may scale the function $r_i^{(j)}(\bar{u})$ by the factor

$$\frac{1}{(k-1)\cdots(k-1-j)}$$

if we wish, to leave merely

$$(\bar{u}-\bar{u}_i)^{k-j} \ .$$

The basis (34) mentioned above is an example of this for quadratics.

### 6.8. Subspaces

If any set of $l$ linearly independent vectors is chosen from a space,

$$\mathbf{X} = \{ X_1 , \ldots , X_l \} \ ,$$

and the collection of all vectors formed as linear combinations of this set is considered,

$$\mathbf{W} = \{ W : W = a_1X_1 + \cdots + a_lX_l \} \ ,$$

then $\mathbf{W}$ can be seen to be a vector space of dimension $l$ with the members of $\mathbf{X}$ as a basis. $\mathbf{W}$ may or may not contain all the vectors of the original space — if $l \leq k$, then it is a *subspace* of the original.

A simple example of a subspace is provided by any of the 2-space planes which are imbedded in 3-space and pass through the origin. The $x,y$ plane is a particular instance which is generated by all linear combinations of the collection

$$(1,0,0) \text{ and } (0,1,0) \ .$$

That is, the $x,y$ plane consists of all linear combinations

$$x(1,0,0) + y(0,1,0) = (x,y,0) \ .$$

In the same sense, the linear polynomials constitute a subspace of the quadratics.

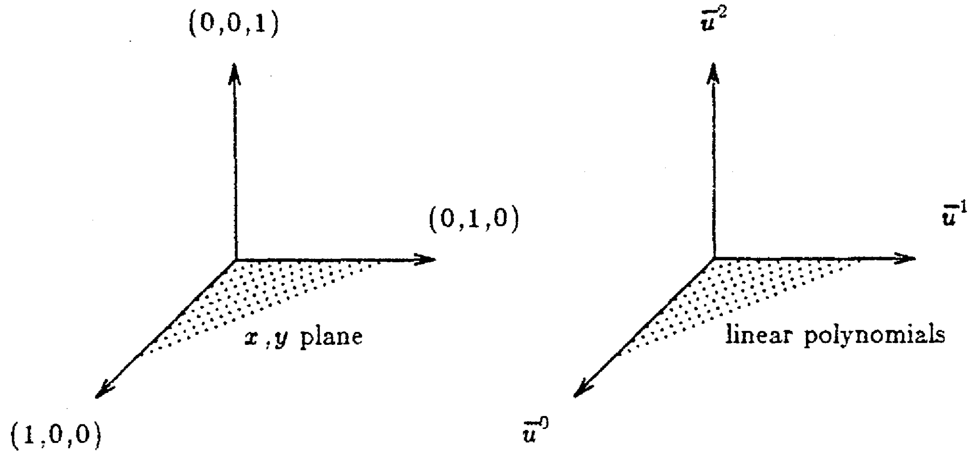


Figure 50. Representative subspaces in 3-space and quadratic polynomials.

More profoundly, we will come to see that the space $\mathbf{P}^k$ is a subspace of the space of $k^{th}$-order splines.

A basis for a subspace need not be chosen as a subset of the basis for a full space, as was the case in the above two examples. The set of all "linear combinations" of

$$(1,2,3)$$

i.e. all scalar multiples of this vector, forms a subspace (in this case a 1-D subspace — a line) of 3-space. However, if $N_0, \ldots, N_n$ form a basis for a vector space, and $B_0, \ldots, B_m$ $(m<n)$ constitutes a basis for one of its subspaces, then we must have the representations

$$B_i = \alpha_{i,0}N_0 + \cdots + \alpha_{i,n}N_n$$

for $i = 0, \ldots, m$ and some (unique) coefficients, $\alpha_{i,0}, \ldots, \alpha_{i,n}$, since each $B_i$ is a member of the space described by the basis $N_0, \ldots, N_n$. That is, the $B$'s may be found from the $N$'s by the use of an $m \times n$ table (matrix) of numbers $\alpha_{i,j}$.

We will show over the course of the next several chapters that the $k^{\text{th}}$-order splines with knots $\{\bar{u}_0, \ldots, \bar{u}_{m+k}\}$ form a vector space, and that the B-splines constitute a basis. The "Oslo algorithm" for B-spline subdivision with which we shall end this theoretical development is simply a method for finding the $\alpha$'s for a given "$N$-basis" spline space and a corresponding "$B$-basis" spline subspace. The $N$-space is generated from the $B$-space by subdividing the range of the parameter into smaller knot intervals.

### 6.9. Knots and Parameter Ranges: Splines as a Vector Space

The order, $k$, of the polynomials with which we deal will force us to change our notion of how many knots we must have and what constitutes the legal range of the parameter among those knots. For cubics $(k=4)$ we required knots

$$\bar{u}_0\ ,\ \bar{u}_1\ ,\ \bar{u}_2\ ,\ \text{and}\ \bar{u}_3$$

to lie "to the left" of the parameter $\bar{u}$ at all times, and knots

$$\bar{u}_{m+1}\ ,\ \bar{u}_{m+2}\ ,\ \bar{u}_{m+3}\ ,\ \text{and}\ \bar{u}_{m+4}$$

to lie "to the right". This meant that, for any legal value of $\bar{u}$, we were on the nonzero domain of four B-splines, precisely the number needed to define any of the cubic segments of the spline curve we sought to construct. The correct numbers for general $k$ turn out to be

$$\bar{u}_0\ ,\ \ldots\ ,\ \bar{u}_{k-1}$$

on the left,

$$\bar{u}_{m+1}\ ,\ \ldots\ ,\ \bar{u}_{m+k}$$

on the right, and

$$\bar{u}_{k-1} \leq \bar{u} < \bar{u}_{m+1}$$

as the legal range for the parameter.

For $k=4$ in our introductory sections on uniform cubic B-splines we constructed curves and surfaces from pieces of cubics which existed as functions for all values of $\bar{u}$

$$-\infty < \bar{u} < +\infty\ .$$

The cubic pieces were, however, not very interesting for

$$\bar{u} < \bar{u}_3\ \text{and}\ \bar{u} \geq \bar{u}_{m+1} \tag{38}$$

since the pieces were quite limited in their character throughout these regions. (Any function which we constructed was zero for $\bar{u}<\bar{u}_0$, was some multiple of $(\bar{u}-\bar{u}_0)^3$ for $\bar{u}_0 \leq \bar{u} < \bar{u}_1$ — the leftmost piece of the leftmost B-spline which we were using — and so on. Similar statements apply to the range $\bar{u} \geq \bar{u}_{m+1}$.)

We took account of such limitations in the ranges (38) by only constructing our curves (or surfaces) for values of the parameter range $[\bar{u}_3,\bar{u}_{m+1})$. The piecewise cubic functions existed everywhere, but we found them useful only on a fixed parameter range. More profoundly, if we had constructed two piecewise $C^2$ cubics

$$s_1(\bar{u}) \text{ and } s_2(\bar{u})$$

which were equal on the range

$$\bar{u}_3 \le \bar{u} < \bar{u}_{m+1}$$

but which differed outside this interval (on the ranges (38)), then it should make no difference whatever to the appearance of $s_1$ or $s_2$ if they are rendered only for $\bar{u}$ in the range $[\bar{u}_3,\bar{u}_{m+1})$. For any practical purpose, $s_1$ and $s_2$ are equivalent.

This has its reflection in general, particularly if we wish to use vector space terminology on splines. Splines will constitute, for us, pieces of $k^{\text{th}}$-order polynomials, each piece being defined on one interval of a collection of intervals, with some differentiability properties satisfied by adjacent pieces. Our notation will be different in the next section, for reasons which we will explain, but for the present this means roughly that a spline will be defined for knots

$$\bar{u}_0,\bar{u}_1,\ \ldots\ ,\bar{u}_{m+k}$$

as a function $s(\bar{u})$ of the form

$$s(\bar{u}) = \begin{cases} p_{-1}(\bar{u}) & \text{defined on} \quad (-\infty,\bar{u}_0) \\ p_0(\bar{u}) & \text{defined on} \quad [\bar{u}_0,\bar{u}_1) \\ p_1(\bar{u}) & \text{defined on} \quad [\bar{u}_1,\bar{u}_2) \\ \quad\cdot & \qquad\cdot \qquad\qquad \cdot \\ \quad\cdot & \qquad\cdot \qquad\qquad \cdot \\ \quad\cdot & \qquad\cdot \qquad\qquad \cdot \\ p_{m+k-1}(\bar{u}) & \text{defined on} \quad [\bar{u}_{m+k-1},\bar{u}_{m+k}) \\ p_{m+k}(\bar{u}) & \text{defined on} \quad (\bar{u}_{m+k},+\infty) \end{cases} \quad ,$$

where $p_i(\bar{u})$ and $p_{i+1}(\bar{u})$ are expected to disagree at $\bar{u}=\bar{u}_{i+1}$ in their "highest $\mu_i$ derivatives", that is, derivatives $k-1,\ \ldots\ ,k-\mu_i$, for some reasonable collection of "continuity-loss indices" $\mu_i$. For the purposes of setting continuity conditions, then, we will associate an index $0<\mu_i\le k$ with each $\bar{u}_i$. For example, a picture might be something like the following:

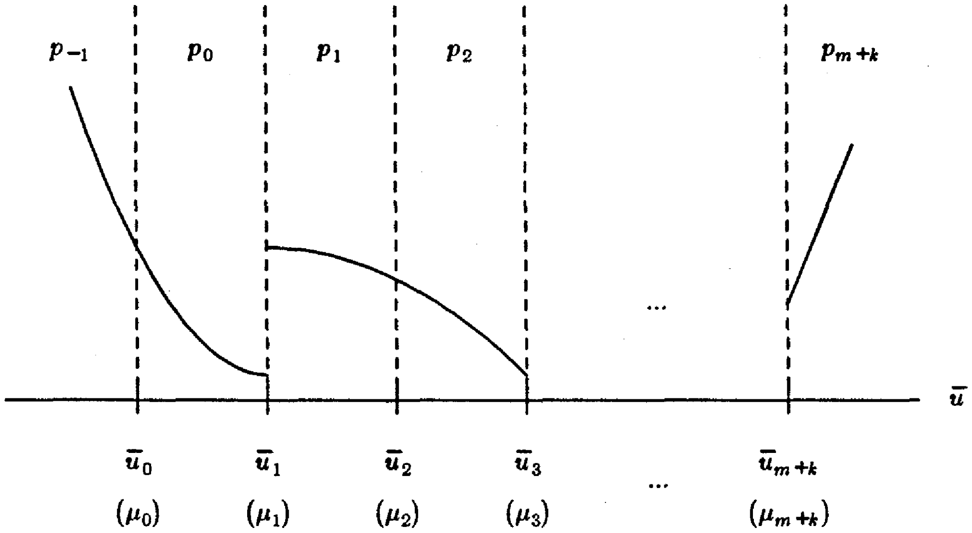


Figure 51. A look at an arbitrary spline.

Vector addition will consist of adding these piecewise polynomials together piece by piece, i.e. if $s_1(\bar{u})$ is defined in terms of pieces $p_{1,i}(\bar{u})$ on a collection of intervals $[\bar{u}_i,\bar{u}_{i+1})$, and another spline $s_2(\bar{u})$ is defined in terms of pieces $p_{2,i}(\bar{u})$ on the same intervals, and with the same continuity properties between the pieces, then we can set

$$s_1(\bar{u}) + s_2(\bar{u}) = p_{1,i}(\bar{u}) + p_{2,i}(\bar{u}) \quad \text{on the interval } [\bar{u}_i,\bar{u}_{i+1})$$

for each $i = 0, \ldots, m+k-1$.

Similarly, the multiplication of a single spline by a scalar can be defined in piece-by-piece fashion. These definitions satisfy the conditions for vector addition and scalar multiplication.

To completely regard these functions as members of a vector space, it will be necessary to define what we mean by equality. We will say that any two splines $s_1(\bar{u})$ and $s_2(\bar{u})$ are equal if they do not differ for any $\bar{u} \in [\bar{u}_{k-1},\bar{u}_{m+1})$. This means that the polynomials

$$p_{-1}(\bar{u})\ , \ldots ,\ p_{k-2}(\bar{u}) \quad \text{and} \quad p_{m+1}(\bar{u})\ , \ldots ,\ p_{m+k}(\bar{u})$$

are of no interest to us when we are comparing two splines. The difference (or identity) of two splines is only a matter to be determined from the polynomials

$$p_{k-1}(\bar{u})\ , \ldots ,\ p_m(\bar{u})\ .$$

Equivalently, as we mentioned in the section above on vector spaces, this is making a statement about what we consider to be the "zero vector"

$$s_1(\bar{u}) - s_2(\bar{u}) = \Theta(\bar{u}) \equiv \text{has value zero for all } \bar{u} \in [\bar{u}_{k-1},\bar{u}_{m+1})\ .$$

The idea that differences are important only if they take place on the legal parameter range, that the concept of a spline space is fundamentally entwined with a restricted parameter range, and that splines are equivalent if they are equal on that range — all this is a reflection of the fact that we kept $\bar{u}$ to the right of $\bar{u}_3$ and to the left of $\bar{u}_{m+1}$ for cubics.

For splines this makes the concept of linear independence a bit more subtle than it was for pure polynomials. The concept of linear independence that is applicable to a spline is that two splines

$$s_1(\overline{u}) \quad \text{and} \quad s_2(\overline{u})$$

are linearly independent if and only if

$$\alpha_1 s_1(\overline{u}) + \alpha_2 s_2(\overline{u}) = 0$$

implies

$$\alpha_1 = \alpha_2 = 0$$

when only the values of $\overline{u}$ on the parameter range of interest are taken into account. As an example, the functions $s_1$ and $s_2$ shown below are linearly independent if we are only being shown a section of them and their parameter range happens to include the interval $[-1,3)$. But viewed as functions whose parameter range is $[0,2)$, they are linearly dependent — $s_1$ is a multiple of $s_2$.

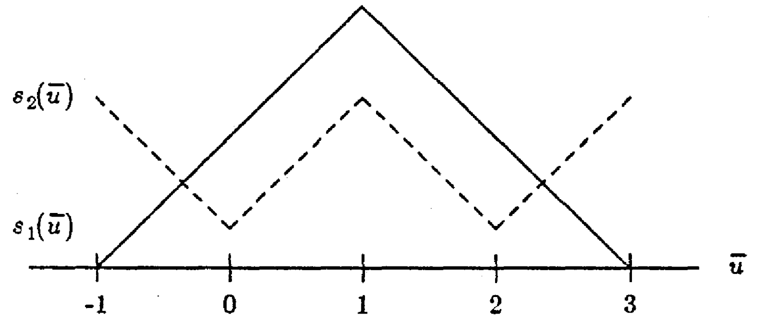


Figure 52. Two $k=2$ splines that are (or are not) linearly independent.

### 6.10. Spline Continuity and Multiple Knots

One of our stated intentions is to relax the requirement that a cubic (4th-order) spline need have $C^2$ continuity across any knot $\overline{u}_i$ or, in general, that a spline involving $k$th-order polynomial segments have $C^{k-2}$ continuity. We suggested above that we associate an index $\mu_i$ with each knot $\overline{u}_i$, indicating what order of continuity is to be imposed at that knot. We could use this index to "count the continuity loss" at the knot $\overline{u}_i$ by the following sort of scheme (remembering that $C^{k-1}$ continuity at a knot implies $C^\infty$ continuity at the knot): let $\mu_i=1$ indicate that $C^{k-1-\mu_i}=C^{k-2}$ continuity is required, let $\mu_i=2$ indicate that $C^{k-1-\mu_i}=C^{k-3}$ continuity is required, and so on through $\mu_i=k$ indicating that no continuity at all, $C^{k-1-\mu_i}=C^{-1}$ continuity, is required at $\overline{u}_i$. This means that we would have to deal with two sequences of numbers to define splines in any generality: the knot sequence

$$\overline{u}_0\,,\ \cdots\ ,\overline{u}_{m+k}$$

and the "index of continuity" sequence

$$\mu_0\,,\ \cdots\ ,\mu_{m+k}\ .$$

This does not turn out to be the best thing to do. The actual situation is more intricate than that, because continuity at a knot turns out to be influenced by the process of moving knots around. For design purposes we might wish to allow the position of knots to be varied, and this will require the option of allowing knots to be pushed together, e.g.

$$\bar{u}_i \rightarrow \bar{u}_{i+1} \ .$$

Suppose that each of these knots is associated with $C^{k-2}$ continuity, and suppose a spline is constructed, for varying choices of $\bar{u}_i$ as it "moves close to" $\bar{u}_{i+1}$; for example:

$$|\bar{u}_i - \bar{u}_{i+1}| = 10^{-1} \ , \ |\bar{u}_i - \bar{u}_{i+1}| = 10^{-2} \ , \ldots , \ |\bar{u}_i - \bar{u}_{i+1}| = 10^{-j} \ .$$

When this is tried computationally, and the $(k-2)^{\text{th}}$ derivative of the spline is studied in sequence, this derivative comes closer and closer to being discontinuous as $\bar{u}_i$ moves closer and closer to $\bar{u}_{i+1}$. The following is an example.

Consider the knots

$$\bar{u}_0 = 0 \ , \ \bar{u}_1 = 1 \ , \ \bar{u}_2 = 1+d \ , \ \text{and} \ \bar{u}_3 = 2+d$$

and construct the quadratic spline $(k=3)$ on the knot intervals

$$[0,1) \ , \ [1,1+d) \ , \ \text{and} \ [1+d,2+d)$$

defined by

$$s(\bar{u}) = \begin{cases} 0 & \bar{u} < 0 \\ \bar{u}^2 & 0 \le \bar{u} < 1 \\ -\dfrac{2\bar{u}(\bar{u}-d-2)+d+2}{d} & 1 \le \bar{u} < 1+d \\ (2+d-\bar{u})^2 & 1+d \le \bar{u} < 2+d \\ 0 & 2+d \le \bar{u} \ . \end{cases}$$

This spline is shown in Figure 53:

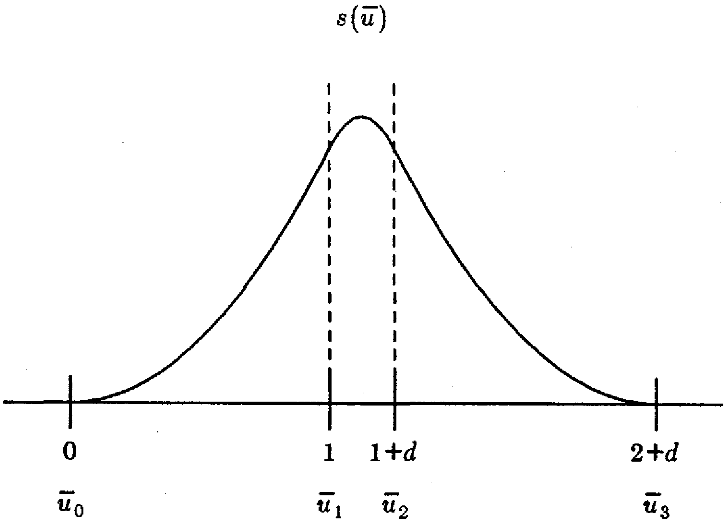


Figure 53. A quadratic B-spline-like spline.

Note that this looks very much like a B-spline — it is zero to the left of $\bar{u}=0$, quadratic in each of the three knot intervals, zero to the right of $\bar{u}=2+d$, positive on $(0,2+d)$, and $C^{3-2}=C^1$ continuous everywhere (i.e. continuous in position and tangent). Its value is 1 at $\bar{u}_1$ and $\bar{u}_2$, and its derivative at these two points is 2 and $-2$, respectively. Both value and derivative at $\bar{u}_1$ and $\bar{u}_2$ are independent of $d$. The value of this spline at the midpoint of its support

$$\bar{u} = 1 + \frac{d}{2}$$

is

$$s(1+\frac{d}{2}) = \frac{d+2}{2} \ .$$

If we take the limit of this function as $d$ goes to zero, the result is clearly

$$s(\bar{u}) = \begin{cases} 0 & \bar{u} < 0 \\ \bar{u}^2 & 0 \leq \bar{u} < 1 \\ (2-u)^2 & 1 \leq \bar{u} < 2 \\ 0 & 2 \leq \bar{u} \end{cases}$$

which looks as follows.

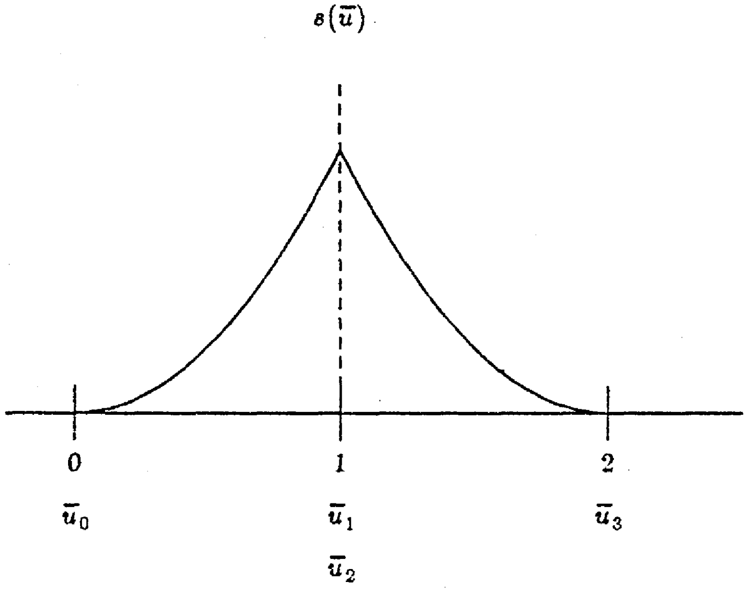


Figure 54. The limiting case as $d$ goes to zero.

When this is done more generally, carefully and mathematically with arbitrary splines, then the resulting continuity at a double knot

$$\bar{u}_i = \bar{u}_{i+1}$$

is always seen to be $C^{k-1-2}=C^{k-3}$, which is just what one would have specified by assigning a $\mu$-index of 2 to the knot value represented by the doublet. If a single, simple knot can be regarded as a point at which there is a single "continuity-loss" for a spline, then this seems to suggest that double knots and double continuity-loss go together. This rule proves true more generally: when more and more knots coalesce together, discontinuities of higher and higher order result at the rate of one $\mu$-index count per multiplicity.

As a further example, suppose $\bar{u}_3$ is pushed together with $\bar{u}_1$ and $\bar{u}_2$ in our above quadratic example.

The resulting "B-spline-like" object will look as follows.

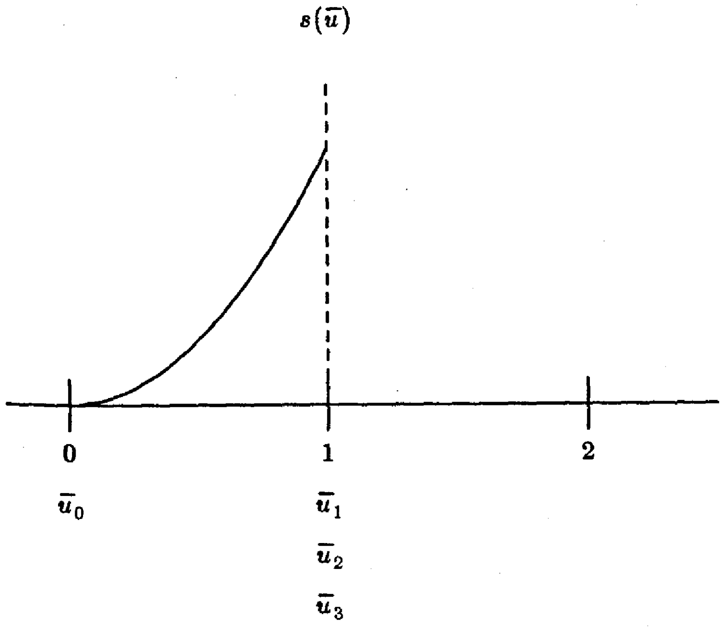


Figure 55. A B-spline-like quadratic with three knots put together. $s(\bar{u})$ is zero for $\bar{u}<0$ and $\bar{u}>1$.

If we move $\bar{u}_0$ into $\bar{u}=1$ to join $\bar{u}_1$, $\bar{u}_2$, and $\bar{u}_3$, then the spline disappears altogether. Since $k=3$ for quadratics and disappearance took place when 4 knots came together, this suggests that we will find it unprofitable in general to push more than $k$ knots together.

More formally, for any $m,k>0$ we will let $\{\bar{u}_i\}_0^{m+k}$ stand for a chosen *sequence* of $m+k+1$ knots

$$\{\bar{u}_i\}_0^{m+k} = \{\ \bar{u}_0\,, \ldots,\ \bar{u}_{m+k}\ \}$$

where

$$\bar{u}_0 \leq \bar{u}_1 \leq \cdots \leq \bar{u}_{m+k}\ .$$

Notice that more than one knot can fall on the same value. Thus we might have

$$\bar{u}_i < \bar{u}_{i+1} = \cdots = \bar{u}_{\lambda-1} = \bar{u}_\lambda = \bar{u}_{\lambda+1} = \cdots = \bar{u}_j < \bar{u}_{j+1}\ ,$$

a situation which we will depict graphically as follows.

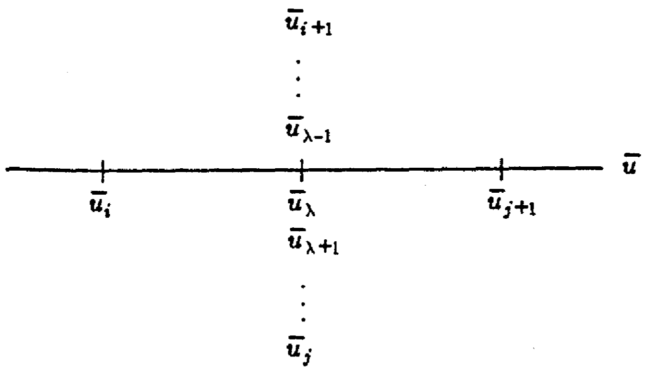


Figure 56. The knots $\bar{u}_{i+1}$ through $\bar{u}_j$ have the same value.

We will say that $\bar{u}_\lambda$ is a knot of *multiplicity* $\mu_\lambda = j-i$. (The same is said of any one of the knots $\bar{u}_{i+1}, \ldots, \bar{u}_j$.) That is, the multiplicity of a knot $\bar{u}_\lambda$ is the count of the knots in the sequence with value equal to $\bar{u}_\lambda$. Having introduced $\mu_\lambda$ as a continuity-loss index, and then argued that continuity loss is to be associated with knot multiplicity, we will use $\mu_\lambda$ to denote the multiplicity of $\bar{u}_\lambda$. The multiplicity count includes $\bar{u}_\lambda$ itself. Notice that $\bar{u}_\lambda$ would be a knot of multiplicity 1, if $\bar{u}_{\lambda-1} < \bar{u}_\lambda < \bar{u}_{\lambda+1}$. The continuity requirement at any knot $\bar{u}_\lambda$ will be $C^{k-1-\mu_\lambda}$, and the number $\mu_\lambda$ will be equal to the number of knots in the knot sequence which are equal to $\bar{u}_\lambda$.

As a simple example, consider the knot sequence

$$\begin{array}{ccccccc} \bar{u}_0 & \bar{u}_1 & \bar{u}_2 & \bar{u}_3 & \bar{u}_4 & \bar{u}_5 & \bar{u}_6 \,. \\ 1 & 3 & 3 & 5 & 5 & 5 & 5 \end{array} \qquad (39)$$

For this sequence

$$\mu_0 = 1,\ \mu_1 = 2,\ \mu_2 = 2,\ \mu_3 = 4,\ \mu_4 = 4,\ \mu_5 = 4, \text{ and } \mu_6 = 4 \ .$$

If we wish to restrict ourselves only to the distinct knots in the sequence, then we may pick a representative knot from each "cluster" of one or more equal knots and list these representatives in their sequence order

$$\bar{u}_{i_0},\ \bar{u}_{i_1}\ , \ldots ,\ \bar{u}_{i_M}\ ,$$

where $M+1$ is the number of distinct values in the knot sequence. As an example, $M=2$ for the knot sequence given in (39), and it could be restricted to

$$\bar{u}_0, \bar{u}_1, \bar{u}_3 \quad (i_0=0, i_1=1, i_2=3)$$

or

$$\bar{u}_0, \bar{u}_2, \bar{u}_5 \quad (i_0=0, i_1=2, i_2=5)$$

or

$$\bar{u}_0, \bar{u}_1, \bar{u}_6 \quad (i_0=0, i_1=1, i_2=6)$$

or any of the other obvious combinations.

A handy way of selecting one representative knot from each cluster, and of specifying a nontrivial interval between adjacent clusters, is to define $\gamma_+(i)$ as the index of the leftmost knot whose value is actually greater than $\bar{u}_i$. For example, in (39) above we would have

$$\gamma_+(0) = 1 \ , \ \gamma_+(1) = 3 \ , \text{ and } \gamma_+(2) = 3 \ .$$

In order to give $\gamma_+$ meaning for knots at the right end of a sequence, we will establish the convention that $+\infty$ is the knot $\bar{u}_{m+k+1}$, which means that, in (39):

$$\gamma_+(3) = 7 \ , \ \gamma_+(4) = 7 \ , \ \gamma_+(5) = 7 \ , \text{ and } \gamma_+(6) = 7 \ .$$

This definition can conveniently be used to choose a representative collection of knots by the following scheme:

$$i_0 = 0$$

and

$$i_j = \gamma_+(i_{j-1}) \quad \text{for } j=1, \ldots, M \ .$$

In (39) above, this would select

$$i_0 = 0 \ , \ i_1 = 1 \ , \text{ and } i_2 = 3 \ .$$

More formally,

**Definition:** For any $i \in \{0, \ldots, m+k\}$ we define

$$\gamma_+(i)$$

to be the smallest index satisfying

$$\bar{u}_{\gamma_+(i)} > \bar{u}_i \ .$$

In order to define $\gamma_+(i)$ for any $i$ satisfying

$$\bar{u}_i = \bar{u}_{m+k} \ ,$$

we let

$$\gamma_+(i) = m+k+1 \text{ if } \bar{u}_i = \bar{u}_{m+k}$$

and we let

$$\bar{u}_{m+k+1} = +\infty \ .$$

Referring to Figure 56:

$$\gamma_+(i) = i+1$$

and

$$\gamma_+(i+1) = \cdots = \gamma_+(\lambda) = \cdots = \gamma_+(j) = j+1 \ .$$

This definition can be "run backwards" by letting $\gamma_-(i)$ be the index of the rightmost knot whose value is strictly less than $\bar{u}_i$. In (39) above, we have

$$\gamma_{-}(1) = 0$$
$$\gamma_{-}(2) = 0$$
$$\gamma_{-}(3) = 2$$
$$\gamma_{-}(4) = 2$$
$$\gamma_{-}(5) = 2$$
$$\gamma_{-}(6) = 2 \quad .$$

In order to give $\gamma_{-}$ meaning for knots at the left end of a sequence, we will establish the convention that $-\infty$ is the knot $\bar{u}_{-1}$, which means that, in (39):

$$\gamma_{-}(0) = -1 \quad .$$

Again more formally,

**Definition:** For any $i \in \{0, \ldots, m+k\}$ we define

$$\gamma_{-}(i)$$

to be the largest index satisfying

$$\bar{u}_{\gamma_{-}(i)} < \bar{u}_i \quad .$$

In order to define $\gamma_{-}(i)$ for any $i$ satisfying

$$\bar{u}_i = \bar{u}_0 \quad .$$

We let

$$\gamma_{-}(i) = -1 \text{ if } \bar{u}_i = \bar{u}_0$$

and we let

$$\bar{u}_{-1} = -\infty \quad .$$

Referring to Figure 56,

$$\gamma_{-}(j+1) = j$$

and

$$\gamma_{-}(j) = \cdots = \gamma_{-}(\lambda) = \cdots = \gamma_{-}(i+1) = i$$

The set of distinct, consecutive values of $\bar{u}$ across which a spline changes from one polynomial into another are often referred to as the *breakpoints* of a spline. In the above example, then, the breakpoints are 1.0, 3.0, and 5.0. The usual practice is to associate the multiplicity of a knot with the unique breakpoint on which it falls. In the above example, then, we may speak of 5.0 as "being associated with (knots of) multiplicity 4."

Also associated with the breakpoints $\bar{u}_i$ are the *breakpoint intervals*,

$$[\bar{u}_i, \bar{u}_{\gamma_+(i)})$$

the half-open intervals over which a spline is merely an ordinary polynomial segment.

There are some subtleties here: the breakpoints are simply *values*. They demarcate the intervals on which a spline is simply a polynomial. The knots, however, are a *sequence*. They might be viewed as "tokens" which are allocated, in order, to the breakpoint positions along the $\bar{u}$ axis — sometimes one to a breakpoint, but sometimes several in a cluster. To list the breakpoints, we may simply select one knot at each breakpoint, thereby choosing a sub-sequence of the knots. That sub-sequence can be chosen whenever and however it suits our convenience. The rules denoted by $\gamma_+$ and $\gamma_-$ are simply two handy ways of determining the sub-sequence.

There is a need for this pedantry. In previous chapters, knots corresponded to breakpoints and there was nothing unusual about the interval between two knots, $[\bar{u}_i, \bar{u}_{i+1})$. If, however, $\bar{u}_i = \bar{u}_{i+1}$, this half-open interval, which is supposed to consist of all values of $\bar{u}$ satisfying $\bar{u}_i \leq \bar{u} < \bar{u}_{i+1}$, is vacuous.

Indeed, the fact that $[\bar{u}_i, \bar{u}_{i+1})$ can be vacuous when $\bar{u}_i$ and $\bar{u}_{i+1}$ are repeated knots leads to a convenient way of designating the breakpoint interval into which a given value of $\bar{u}$ falls. Suppose that $\bar{u}_0 \leq \bar{u} < \bar{u}_m$. Then the phrase

"Let $\delta$ be an index such that $\bar{u}_\delta \leq \bar{u} < \bar{u}_{\delta+1}$"

is a phrase which uniquely specifies $\delta$. This is easiest to see with an example. Consider the knots of (39) and the value of $\bar{u}=4$.

$\delta=0$ is not such that $\bar{u}_\delta = 1 \leq 4 < \bar{u}_{\delta+1} = 3$

$\delta=1$ is not such that $\bar{u}_\delta = 3 \leq 4 < \bar{u}_{\delta+1} = 3$

$\delta=2$ is (!) such that $\bar{u}_\delta = 3 \leq 4 < \bar{u}_{\delta+1} = 5$

$\delta=3$ is not such that $\bar{u}_\delta = 5 \leq 4 < \bar{u}_{\delta+1} = 5$

etc.

Clearly, $\delta=2$, and the breakpoint interval containing $\bar{u}=4$ is

$$[\bar{u}_2, \bar{u}_3) = [3,5) .$$

If knots remained frozen in place, we would have no reason to separate the concepts of "knot" and "breakpoint"; it would suffice to flag each $\bar{u}_i$ with a "continuity-loss" index $\mu_i$, as was suggested above. The need to distinguish between knots and breakpoints arises when knots are moved about, pushed together, and pulled apart.

We close with a formal notation for the set of all splines. In this definition $p_{left}(\bar{u})$ is the *segment polynomial* describing $s(\bar{u})$ in the breakpoint interval

$$[\bar{u}_{\gamma_-(i)}, \bar{u}_\lambda) = [\bar{u}_i, \bar{u}_\lambda) ,$$

and $p_{right}(\bar{u})$ is the segment polynomial describing $s(\bar{u})$ on

$$[\bar{u}_\lambda, \bar{u}_{\gamma_+(\lambda)}) = [\bar{u}_\lambda, \bar{u}_j)$$

the next breakpoint interval to the right Figure 57 below shows this in more detail.

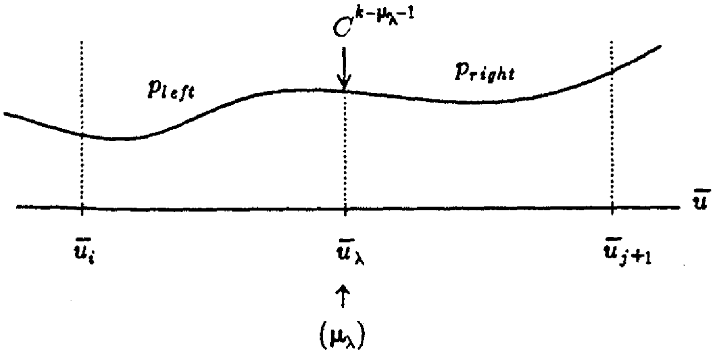


Figure 57. An overview of breakpoint intervals.

The $l^{\text{th}}$ derivative of $p_{left}$ at $\bar{u}_\lambda$ will be denoted by $p_{left}^{(l)}(\bar{u}_\lambda)$, and similarly for $p_{right}$.

**Definition:** Assuming that $k \geq 1$, that $k-1 < m+1$, that $\bar{u}_{k-1} < \bar{u}_{m+1}$, and that $\bar{u}_i \leq \bar{u}_{i+1}$ for all $i = 0, \ldots, m+k-1$, then $S(\mathbf{P}^k, \{\bar{u}_i\}_0^{m+k})$, the set of all $k^{\text{th}}$-order splines on the parameter range $[\bar{u}_{k-1}, \bar{u}_{m+1})$ with the knot sequence $\{\bar{u}_i\}_0^{m+k}$ is the set of all functions $s(\bar{u})$ satisfying:

$$s(\bar{u}) \in \mathbf{P}^k \quad \text{for each interval } [-\infty, \bar{u}_0),\ [\bar{u}_i, \bar{u}_{\gamma_+(i)}),\ \text{and } [\bar{u}_{m+k}) \quad (j = 0, \ldots, M-1)$$

and for any knot $\bar{u}_\lambda$, $\lambda \in \{0, \ldots, m+k\}$, with associated multiplicity $\mu_\lambda$,
if

$$s(\bar{u}) \equiv p_{left}(\bar{u}) \in \mathbf{P}^k \quad \text{on } [\bar{u}_{\gamma_-(\lambda)}, \bar{u}_\lambda)$$

and

$$s(\bar{u}) \equiv p_{right}(\bar{u}) \in \mathbf{P}^k \quad \text{on } [\bar{u}_\lambda, \bar{u}_{\gamma_+(\lambda)})$$

then

$$p_{left}^{(l)}(\bar{u}_\lambda) = p_{right}^{(l)}(\bar{u}_\lambda) \quad \text{for } l = 0, \ldots, k-1-\mu_\lambda \ .$$

No continuity at all is assumed if $k-1-\mu_\lambda < 0$.

Two splines are considered to be *identical* if they are equal for all $\bar{u}$ in the parameter range, even though they may differ outside that range.

It is a trivial observation, but one very useful to make, that any $k^{\text{th}}$-order polynomial is a spline in $S(\mathbf{P}^k, \{\bar{u}_i\}_0^{m+k})$; the converse is not necessarily true. The argument goes something like this:

- On each breakpoint interval a $k^{\text{th}}$-order polynomial is, of course, a $k^{\text{th}}$-order polynomial.
- At each breakpoint $u_\lambda$ a $k^{\text{th}}$-order polynomial is $C^\infty$ differentiable; thus it is certainly $C^{k-1-\mu_\lambda}$ differentiable. There are, obviously, splines in $S(\mathbf{P}^k, \{\bar{u}_i\}_0^{m+k})$ that are not polynomials, so $S(\mathbf{P}^k, \{\bar{u}_i\}_0^{m+k})$ is larger than $\mathbf{P}^k$.

A more profound observation is that the sum of any two functions in $S(\mathbf{P}^k, \{\bar{u}_i\}_0^{m+k})$ is also a function in $S(\mathbf{P}^k, \{\bar{u}_i\}_0^{m+k})$, and that any constant times a function in $S(\mathbf{P}^k, \{\bar{u}_i\}_0^{m+k})$ is also a function in $S(\mathbf{P}^k, \{\bar{u}_i\}_0^{m+k})$. Again, a brief argument:

- On any of the breakpoint intervals we are merely dealing with polynomials. The sum of any two

$k^{th}$-order polynomials is still a $k^{th}$-order polynomial, and a constant times a $k^{th}$-order polynomial is just a $k^{th}$-order polynomial.

- At any point in $[\bar{u}_{k-1}, \bar{u}_{m+1})$ we are dealing with functions having the same fixed differentiability properties. At each point of any breakpoint interval that isn't a knot, all derivatives of any $k^{th}$-order polynomial exist, and so do those of the sums of any two such polynomials or a constant times either one of them. On the other hand, at any knot $\bar{u}_i$, any two members of $\mathbf{S}(\mathbf{P}^k, \{\bar{u}_i\}_0^{m+k})$ will have derivatives of order 0 through $k-\mu_i-1$, hence so must their sum or any constant multiple of either.

That is, in short:

**Theorem:** $\mathbf{S}(\mathbf{P}^k, \{\bar{u}_i\}_0^{m+k})$ is a vector space.

**Theorem:** $\mathbf{P}^k$ is a (proper) subspace of $\mathbf{S}(\mathbf{P}^k, \{\bar{u}_i\}_0^{m+k})$.

As always, $\bar{u}$ is restricted to the range $[\bar{u}_{k-1}, \bar{u}_{m+1})$.

We will establish a few chapters hence that B-splines (suitably generalized) form a basis for this space, but we will arrive at that fact by first discussing a more easily understood basis for constructing $\mathbf{S}(\mathbf{P}^k, \{\bar{u}_i\}_0^{m+k})$.

# 7. The One-Sided Basis

### 7.1. The One-sided Cubic

This chapter will focus on functions of the form $(\bar{u}-t)_+^r$, their properties, and the means by which they are transformed into B-splines. We shall begin with an example.

Suppose that we have a cubic polynomial $p(\bar{u})$, let $t$ be some arbitrary position on the $\bar{u}$ axis, and let $\bar{v}$ indicate the displacement from $t$:

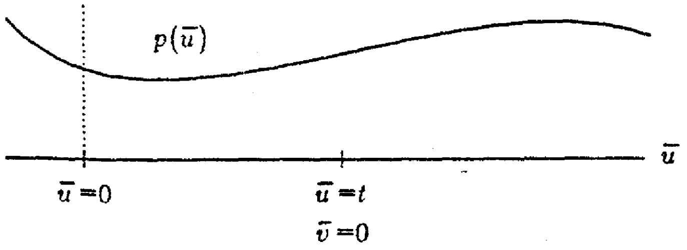


Figure 58.

It is easy to see how to express the value of $p(\bar{u})$ in terms of the displacement $\bar{v}$ from the point $\bar{u}=t$, instead of the displacement $\bar{u}$ from the point $\bar{u}=0$). We know that $\bar{u}=\bar{v}+t$, and all we have to do is substitute $\bar{v}+t$ for $\bar{u}$ in the expression for $p(\bar{u})$. For example, if

$$p(\bar{u}) = 3 - \bar{u} + \bar{u}^2 - \bar{u}^3$$

and $t=1$ then

$$p(\bar{v}) = 2 - 2\bar{v} - 2\bar{v}_2 - \bar{v}^3 \ .$$

Since $\bar{v}=\bar{u}-1$, an alternative representation for $p(\bar{u})$ is now easily seen to be

$$p(\bar{u}) = 2 - 2(\bar{u}-1) - 2(\bar{u}-1)^2 - (\bar{u}-1)^3 \ .$$

Thus, if we are given any polynomial $p(\bar{u})$ and constant $t$, these observations show how easily $p(\bar{u})$ may be given in terms of the powers $(\bar{u}-t)$.

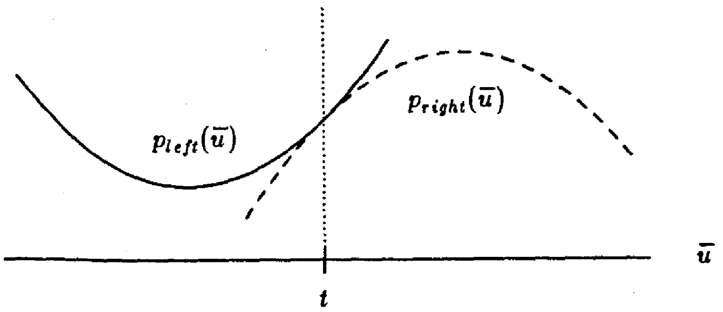


Figure 59. $p_{left}(\bar{u})$ and $p_{right}(\bar{u})$ are cubic polynomials which meed at $\bar{u}=t$ with $C^2$ continuity.

Now suppose that we have two cubic polynomials $p_{left}(\bar{u})$ and $p_{right}(\bar{u})$ which meet with $C^2$ continuity, and no more, at the arbitrary parameter value $\bar{u}=t$ (see Figure 59). Consider the change

$$d(\bar{u}) = p_{right}(\bar{u}) - p_{left}(\bar{u})$$

as we cross the knot $t$. Since $d(\bar{u})$ is clearly a cubic polynomial, we know that $d(\bar{u})$ can be represented in terms of an expansion about $\bar{u}=t$, and so we can write

$$d(\bar{u}) = c_0 + c_1(\bar{u}-1) + c_2(\bar{u}-1)^2 + c_3(\bar{u}-1)^3 \ .$$

What are the coefficients $c_i$? Consider the following:

$$d(\bar{u}) = c_0 + c_1(\bar{u}-t)^1 + c_2(\bar{u}-t)^2 + c_3(\bar{u}-t)^3 \ ,$$

consequently $d(\bar{u}) = c_0$ at $\bar{u} = t$, and

$$\frac{d}{d\bar{u}}d(\bar{u}) = c_1 + 2c_2(\bar{u}-t)^1 + 3c_3(\bar{u}-t)^2 \ ,$$

consequently $\frac{d}{d\bar{u}}d(\bar{u}) = c_1$ at $\bar{u} = t$, and

$$\frac{d^2}{d\bar{u}^2}d(\bar{u}) = 2c_2 + 6c_3(\bar{u}-t)^1 \ ,$$

consequently $\frac{d^2}{d\bar{u}^2}d(\bar{u}) = 2c_2$ at $\bar{u} = t$, and finally,

$$\frac{d^3}{d\bar{u}^3}d(\bar{u}) = 6c_3 \ ,$$

consequently $\frac{d^3}{d\bar{u}^3}d(\bar{u}) = 6c_3$ at $\bar{u} = t$.

The $c_i$ are, essentially, the derivatives of $d(\bar{u})$ at $\bar{u}=t$. But because $p_{left}(\bar{u})$ and $p_{right}(\bar{u})$ are $C^2$ continuous, we know that they have the same value, first derivative and second derivative at $t$; that is,

$$d(t) = 0 \ , \ d'(t) = 0 \ , \text{ and } d''(t) = 0 \ .$$

So

$$c_0 = 0 \ , \ c_1 = 0 \ , \text{ and } c_2 = 0 \ ,$$

and we have

$$d(\bar{u}) = c_3(\bar{u}-t)^3 .$$

In fact, it is clear the value of $c_3$ is determined by the size of the change in the third derivative of the spline at $t$; we have

$$c_3 = \frac{1}{6}\left(p^{(3)}_{right}(t) - p^{(3)}_{left}(t)\right) .$$

It simplifies things a bit if we define a modified version

$$(\bar{u}-t)^3_+$$

(notice the subscript "+" ) of $(\bar{u}-t)^3$ which is zero to the left of $\bar{u}=t$ and acts just like $(\bar{u}-t)^3$ for all $\bar{u} \geq t$. The form of such a function, for $t = \bar{u}_0, \ldots, \bar{u}_8$, is shown below in Figure 60.

Why is all this useful? Notice that $(\bar{u}-t)^3_+$ has the value zero at $\bar{u}=t$. So do the first and second derivatives of $(\bar{u}-t)^3_+$. But the third derivative of $(\bar{u}-t)^3_+$ is discontinuous at $\bar{u}=t$:

| | | | |
|---|---|---|---|
| function value: | $(\bar{u}-t)^3\rvert_{\bar{u}=t}$ | $= (t-t)^3$ | $= 0$ |
| first derivative value: | $3\cdot(\bar{u}-t)^2\rvert_{\bar{u}=t}$ | $= 3\cdot(t-t)^2$ | $= 0$ |
| second derivative value: | $6\cdot(\bar{u}-t)^1\rvert_{\bar{u}=t}$ | $= 6\cdot(t-t)^1$ | $= 0$ |
| third derivative value: | $6\rvert_{\bar{u}=t}$ | | $= 6$ (a constant) . |

Suspiciously, this is exactly the sort of discontinuity displayed by $C^2$ splines in general.

Since $(\bar{u}-t)^3_+$ is zero for $\bar{u}<t$ we can now write

$$p_{left}(\bar{u}) = p_{left}(\bar{u}) + c_3(\bar{u}-t)^3_+$$

and for $\bar{u} \geq t$ we also have

$$p_{right}(\bar{u}) = p_{left}(\bar{u}) + c_3(\bar{u}-t)^3_+ .$$

What we have obtained is a succinct representation, equally valid on either side of $t$, for the spline represented by $p_{left}(\bar{u})$ and $p_{right}(\bar{u})$. All of the foregoing is a special case of Taylor's theorem (in which we have "expanded $d(\bar{u})$ about $\bar{u} = t$").

Suppose that we have the $C^2$ cubic spline shown in Figure 60.

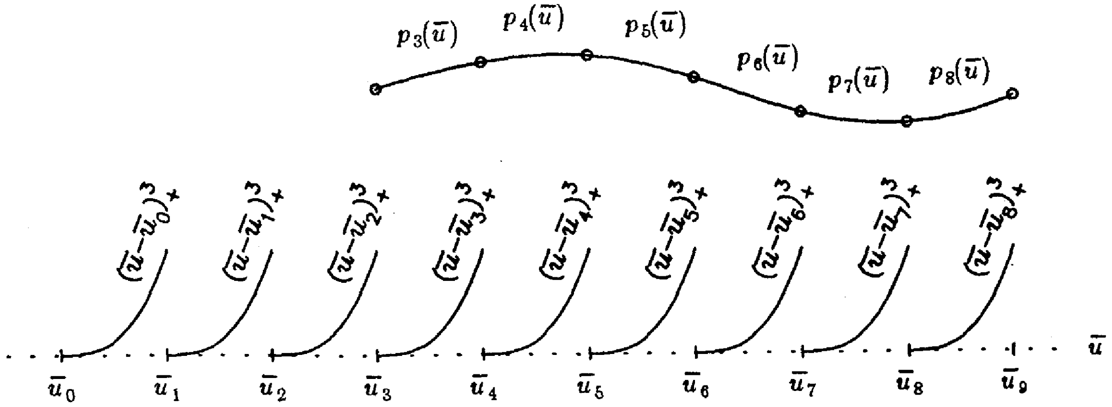

$p_3(\bar{u})$
$p_4(\bar{u})$
$p_5(\bar{u})$
$p_6(\bar{u})$
$p_7(\bar{u})$
$p_8(\bar{u})$
$(\bar{u}-\bar{u}_0)^3_+$
$(\bar{u}-\bar{u}_1)^3_+$
$(\bar{u}-\bar{u}_2)^3_+$
$(\bar{u}-\bar{u}_3)^3_+$
$(\bar{u}-\bar{u}_4)^3_+$
$(\bar{u}-\bar{u}_5)^3_+$
$(\bar{u}-\bar{u}_6)^3_+$
$(\bar{u}-\bar{u}_7)^3_+$
$(\bar{u}-\bar{u}_8)^3_+$
$\bar{u}$
$\bar{u}_0$
$\bar{u}_1$
$\bar{u}_2$
$\bar{u}_3$
$\bar{u}_4$
$\bar{u}_5$
$\bar{u}_6$
$\bar{u}_7$
$\bar{u}_8$
$\bar{u}_9$

Figure 60. A $C^2$ cubic spline expressed with the aid of "one-sided basis functions". Each $p_i(\bar{u})$ is a scaled sum of all the basis functions which depart from zero left of, or at, $\bar{u}_i$.

If we identify $p_3(\bar{u})$ with $p_{left}(\bar{u})$ in the above discussion and $p_4(\bar{u})$ with $p_{right}(\bar{u})$, using $t = \bar{u}_4$, then

$$p_4 = p_3(\bar{u}) + c_{3,4}(\bar{u}-\bar{u}_4)_+^3$$

for some constant $c_{3,4}$. It is also important to note that $p_3(\bar{u})$ can be expressed in terms of $(\bar{u}-\bar{u}_0)_+^3$, $(\bar{u}-\bar{u}_1)_+^3$, $(\bar{u}-\bar{u}_2)_+^3$, and $(\bar{u}-\bar{u}_3)_+^3$:

$$p_3(\bar{u}) = c_{3,0}(\bar{u}-\bar{u}_0)_+^3 + c_{3,1}(\bar{u}-\bar{u}_1)_+^3 + c_{3,2}(\bar{u}-\bar{u}_2)_+^3 + c_{3,3}(\bar{u}-\bar{u}_3)_+^3 \ .$$

This is true simply because

$$\begin{aligned}(\bar{u}-\bar{u}_0)_+^3 &= (\bar{u}-\bar{u}_0)^3 \quad \text{for} \quad \bar{u}\in[\bar{u}_3,\bar{u}_4)\\ (\bar{u}-\bar{u}_1)_+^3 &= (\bar{u}-\bar{u}_1)^3 \quad \text{for} \quad \bar{u}\in[\bar{u}_3,\bar{u}_4)\\ (\bar{u}-\bar{u}_2)_+^3 &= (\bar{u}-\bar{u}_2)^3 \quad \text{for} \quad \bar{u}\in[\bar{u}_3,\bar{u}_4)\\ (\bar{u}-\bar{u}_3)_+^3 &= (\bar{u}-\bar{u}_3)^3 \quad \text{for} \quad \bar{u}\in[\bar{u}_3,\bar{u}_4) \ ,\end{aligned}$$

and $(\bar{u}-\bar{u}_0)^3, \ldots, (\bar{u}-\bar{u}_3)^3$ constitutes a basis for the cubics.

If we shift our attention one interval to the right and regard $p_4(\bar{u})$ as $p_{left}(\bar{u})$, then we see that

$$p_5(\bar{u}) = p_4(\bar{u}) + c_{3,5}(\bar{u}-\bar{u}_5)_+^3 \ .$$

But note that this can also be written as

$$p_5(\bar{u}) = p_3(\bar{u}) + c_{3,4}(\bar{u}-\bar{u}_4)_+^3 + c_{3,5}(\bar{u}-\bar{u}_5)_+^3 \ ,$$

or even as

$$p_5(\bar{u}) = c_{3,0}(\bar{u}-\bar{u}_0)_+^3 + \cdots + c_{3,5}(\bar{u}-\bar{u}_5)_+^3 \ .$$

Continuing in this fashion, we can convince ourselves that the entire cubic spline may be represented as

$$\sum_{i=0}^{8} c_{3,i}\ (\bar{u}-\bar{u}_i)_+^3$$

for some coefficients $c_{3,i}$ that we know how to compute.

Now let us see how to generalize what we have learned to establish an easily understood basis for $\mathbf{S}(\mathbf{P}^k,\{\bar{u}_i\}_0^{m+k})$. We will proceed intuitively; for a rigourous development of this material see [Schumaker81] or [deBoor78].

### 7.2. The General Case

The first segment polynomial of any spline $s(\bar{u})$ that could interest us is the one that is defined on the first breakpoint interval of the parameter range, namely

$$[\bar{u}_{k-1}, \bar{u}_{\gamma+(k-1)}) \ .$$

Let us denote this ($k^{\text{th}}$-order) segment polynomial by $p_{left}(\bar{u})$. If $\bar{u}_0, \ldots, \bar{u}_{k-1}$ are distinct, then $s(\bar{u}) = p_{left}(\bar{u})$ can be represented on this interval as a linear combination of the $k$ basis functions

$$(\bar{u}-\bar{u}_0)^{k-1}, (\bar{u}-\bar{u}_1)^{k-1}, \ldots, (\bar{u}-\bar{u}_{k-1})^{k-1} \ , \tag{40}$$

which is the basis mentioned in (36). If the first $k$ knots are not distinct, then some mixture of the basis described by (36) and (37) will be necessary to describe $p_{left}(\bar{u})$. We will keep things simple, just for the moment, and assume that the basis functions (40) are appropriate, so that

$$s(\bar{u}) = p_{left}(\bar{u}) = \sum_{r=0}^{k-1} c_{k-1,r}(\bar{u}-\bar{u}_r)^{k-1} \text{ on } [\bar{u}_{k-1}, \bar{u}_{\gamma+(k-1)}) \tag{41}$$

for some selection of real coefficients $c_{k-1,0}, c_{k-1,1}, \ldots, c_{k-1,k-1}$. The picture to have in mind for the following discussion is Figure 61:

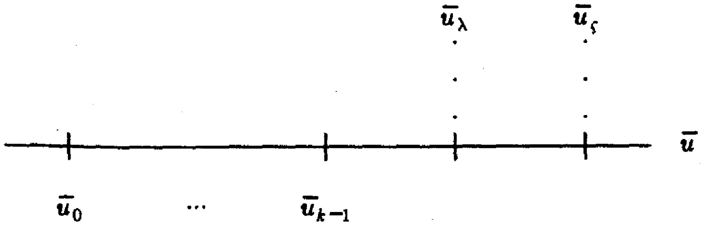


Figure 61. Schema of the knots at the beginning of the parameter range (example for the purpose of discussion).

As $\bar{u}$ crosses the breakpoint at

$$\bar{u}_{\gamma+(k-1)} = \bar{u}_\lambda$$

$s(\bar{u})$ changes from $p_{left}(\bar{u})$, a $k^{\text{th}}$-order polynomial, into $p_{right}(\bar{u})$, also a $k^{\text{th}}$-order polynomial. But $p_{left}(\bar{u})$ and $p_{right}(\bar{u})$ are expected to agree in their first $k-1-\mu_\lambda$ derivatives at $\bar{u}=\bar{u}_\lambda$, because of the multiplicity of the knots associated with this breakpoint. Consequently:

$$\begin{array}{lcl} p_{left}^{(0)}(\bar{u}_\lambda) & = & p_{right}^{(0)}(\bar{u}_\lambda) \\ p_{left}^{(1)}(\bar{u}_\lambda) & = & p_{right}^{(1)}(\bar{u}_\lambda) \\ \vdots & \vdots & \vdots \\ p_{left}^{(k-1-\mu_\lambda)}(\bar{u}_\lambda) & = & p_{right}^{(k-1-\mu_\lambda)}(\bar{u}_\lambda) \ . \end{array} \tag{42}$$

Generalizing the argument of the last section we see that

$$\begin{aligned} d(\bar{u}) &= p_{right}(\bar{u}) - p_{left}(\bar{u}) \\ &= d^{(0)}(\bar{u}_\lambda)(\bar{u}-\bar{u}_\lambda)^0 + d^{(1)}(\bar{u}_\lambda)(\bar{u}-\bar{u}_\lambda)^1 + \cdots + \frac{1}{(k-1)!} d^{(k-1)}(\bar{u}_\lambda)(\bar{u}-\bar{u}_\lambda)^{k-1} \ . \end{aligned}$$

(This representation is also a simple instance of Taylor's Theorem.) We are particularly interested in this expression for $\bar{u}_\lambda \le \bar{u} < \bar{u}_{\gamma+(\lambda)} = \bar{u}_\varsigma$, which is the next breakpoint interval to the right, since for these values

$$s(\bar{u}) = p_{right}(\bar{u}) = p_{left}(\bar{u}) + d(\bar{u}) \ .$$

Now, because of (42) above,

$$d^{(0)}(\bar{u}_\lambda) = \cdots = d^{(k-1-\mu_\lambda)}(\bar{u}_\lambda) = 0 \ .$$

Hence $d(\bar{u})$ can be written more simply as

$$d(\bar{u}) = c_{k-\mu_\lambda,\lambda}(\bar{u}-\bar{u}_\lambda)^{k-\mu_\lambda} + c_{k-\mu_\lambda+1,\lambda}(\bar{u}-\bar{u}_\lambda)^{k-\mu_\lambda+1} + \cdots + c_{k-1,\lambda}(\bar{u}-\bar{u}_\lambda)^{k-1}$$

for some some constants $c_{k-\mu_\lambda,\lambda}, \ldots, c_{k-1,\lambda}$. This suggests that we can write $s(\bar{u})$ on the entire interval $[\bar{u}_{k-1},\bar{u}_\varsigma)$ as

$$s(\bar{u}) = p_{left}(\bar{u}) + d(\bar{u})_+ ,$$

where

$$d(\bar{u})_+ = c_{k-\mu_\lambda,\lambda}(\bar{u}-\bar{u}_\lambda)_+^{k-\mu_\lambda} + c_{k-\mu_\lambda+1,\lambda}(\bar{u}-\bar{u}_\lambda)_+^{k-\mu_\lambda+1} + \cdots + c_{k-1,\lambda}(\bar{u}-\bar{u}_\lambda)_+^{k-1} ,$$

giving us

$$s(\bar{u}) = \begin{cases} p_{left}(\bar{u}) & \bar{u}_{k-1} \le \bar{u} < \bar{u}_\lambda \\ p_{right}(\bar{u}) & \bar{u}_\lambda \le \bar{u} < \bar{u}_\varsigma . \end{cases}$$

(This assumes that $\mu_\lambda \le k$. If $\mu_\lambda > k$, then we take 0 as the lowest exponent in the above, i.e. we use $\min(k,\mu_\lambda)$ in place of $\mu_\lambda$.)

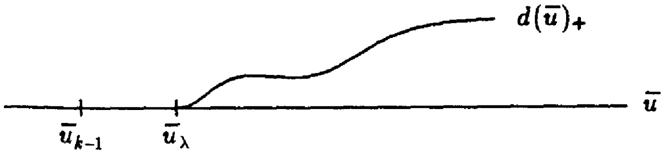


Figure 62. $d(\bar{u})_+$ is the amount which must be "added" to $p_{left}(\bar{u})$ as we cross $\bar{u}_\lambda$ to obtain $p_{right}(\bar{u})$.

This means that for $\bar{u} \in [\bar{u}_{k-1},\bar{u}_{\gamma_+(\lambda)}) = [\bar{u}_{k-1},\bar{u}_\varsigma)$, we have

$$s(\bar{u}) = \sum_{j=0}^{k-1} c_{k-1,j}(\bar{u}-\bar{u}_j)^{k-1} + \sum_{\tau=1}^{\min(k,\mu_\lambda)} c_{k-\tau,\lambda}(\bar{u}-\bar{u}_\lambda)_+^{k-\tau} ,$$

since

$$p_{left}(\bar{u}) = \sum_{j=0}^{k-1} c_{k-1,j}(\bar{u}-\bar{u}_j)^{k-1}$$

for come constants $c_{k-1,0}, \ldots, c_{k-1,k-1}$.

Let us number the breakpoints consecutively:

$$\bar{u}_{i_0} = \bar{u}_0$$

$$\bar{u}_{i_1} = \bar{u}_1$$

$$\vdots$$

$$\bar{u}_{i_{k-1}} = \bar{u}_{k-1}$$

$$\bar{u}_{i_k} = \bar{u}_{\lambda}$$

$$\bar{u}_{i_{k+1}} = \bar{u}_{\varsigma}$$

$$\vdots$$

$$\bar{u}_{i_L} = \bar{u}_{\gamma-(m+1)} \;,$$

where the last breakpoint which needs to be considered is the one associated with the last knot lying strictly within the parameter range, which we have denoted by $\bar{u}_{i_L}$.

The same argument that we made at the breakpoint $\bar{u}=\bar{u}_{\lambda}=\bar{u}_{i_k}$ can be repeated across each successive breakpoint in the parameter range. $s(\bar{u})$ picks up the powers

$$(\bar{u}-\bar{u}_{i_j})^{k-l} \quad \text{for} \quad l=1, \ldots, \min(k,\mu_{i_j})$$

at each breakpoint $\bar{u}_{i_j}$. This means that $s(\bar{u})$ can be represented across the entire parameter range $\bar{u} \in [\bar{u}_{k-1},\bar{u}_{m+1})$ by the formula

$$s(\bar{u}) = \sum_{j=0}^{k-1} c_{k-1,j}(\bar{u}-\bar{u}_{i_j})^{k-1} + \sum_{j=k}^{L} \sum_{r=1}^{\min(k,\mu_{i_j})} c_{k-r,i_j}(\bar{u}-\bar{u}_{i_j})_+^{k-r} \;,$$

where the $c$'s are appropriate coefficients.

Since we are only interested in $s(\bar{u})$ for $\bar{u} \geq \bar{u}_{i_{k-1}}$, we may subscript all of the terms in the first summation by "+", since to the right of $\bar{u}_{i_{k-1}}=\bar{u}_{k-1}$

$$(\bar{u}-\bar{u}_0)^3 = (\bar{u}-\bar{u}_0)_+^3 \,,\; (\bar{u}-\bar{u}_1)^3 = (\bar{u}-\bar{u}_1)_+^3 \,, \ldots ,\, (\bar{u}-\bar{u}_{k-1})^3 = (\bar{u}-\bar{u}_{k-1})_+^3 \;,$$

so

$$s(\bar{u}) = \sum_{j=0}^{k-1} c_{k-1,j}(\bar{u}-\bar{u}_{i_j})_+^{k-1} + \sum_{j=k}^{L} \sum_{r=1}^{\min(k,\mu_{i_j})} c_{k-r,i_j}(\bar{u}-\bar{u}_{i_j})_+^{k-r} \;.$$

Note that the second summation is doubled, with each breakpoint contributing as many terms (successive powers) as there are knots multiply located on that breakpoint. The simplest version of this formula would be the one in which all knots are simple:

$$s(\bar{u}) = \sum_{j=0}^{m} c_{k-1,j}(\bar{u}-\bar{u}_{i_j})_+^{k-1} \;.$$

This is the formula which describes the curve shown in Figure 60, namely

$$s(\bar{u}) = \begin{cases} p_3(\bar{u}) & \text{for} \quad \bar{u} \in [\bar{u}_3, \bar{u}_4) \\ p_4(\bar{u}) & \text{for} \quad \bar{u} \in [\bar{u}_4, \bar{u}_5) \\ p_5(\bar{u}) & \text{for} \quad \bar{u} \in [\bar{u}_5, \bar{u}_6) \\ p_6(\bar{u}) & \text{for} \quad \bar{u} \in [\bar{u}_6, \bar{u}_7) \\ p_7(\bar{u}) & \text{for} \quad \bar{u} \in [\bar{u}_7, \bar{u}_8) \\ p_8(\bar{u}) & \text{for} \quad \bar{u} \in [\bar{u}_8, \bar{u}_9) \end{cases}$$

$$= \sum_{j=0}^{8} c_{3,j} (\bar{u} - \bar{u}_j)_+^3 \ .$$

On the other hand, the formula takes on its most general form if we drop our initial assummption that

$$\bar{u}_0 \ , \ . \ . \ . \ , \ \bar{u}_{k-1}$$

are distinct. These knots, too, could be multiple, giving:

$$s(\bar{u}) = \sum_{j=0}^{L} \sum_{r=1}^{\min(k,\mu_{i_j})} c_{k-r,i_j} (\bar{u} - \bar{u}_{i_j})_+^{k-r} \ .$$

The functions $(\bar{u}-t)_+^r$, for $t=$ any knot and any power $r=0, \ldots, k-1$, are clearly worth studying, since the discussion above leads us to believe that any spline on any collection of knots can be represented as a linear combination of such "one-sided power functions."

### 7.3. One-sided Power Functions

Let us begin with the simplest version of $(\bar{u}-t)_+^r$, namely the one with $r=0$. This is the step function given by

$$(\bar{u} - t)_+^0 \equiv \begin{cases} 0 & \bar{u} < t \\ 1 & \bar{u} \geq t \ . \end{cases} \tag{43}$$

This function is an instance of two first order polynomials (constants) tied together with no continuity whatsoever across the single breakpoint $t$, which also constitutes a knot of multiplicity 1. It is the simplest possible example of a spline: $k=1$, $\bar{u}_0=t$, $\mu_0=1$, and we have $C^{-1}$ continuity across the breakpoint. Moreover, it is just what we need to represent functions like $d(\bar{u})_+$ above, where $d(\bar{u})$ is of order 1; $(\bar{u} - \bar{u}_{i_0})_+^0$ is identically zero to the left of $\bar{u}_{i_0}$, and can be used to represent any constant (polynomial of order 1) to the right of $\bar{u}_{i_0}$. Thus

$$d(\bar{u})_+ = \text{constant} \cdot (\bar{u} - t)_+^0 \ .$$

Figuratively we have

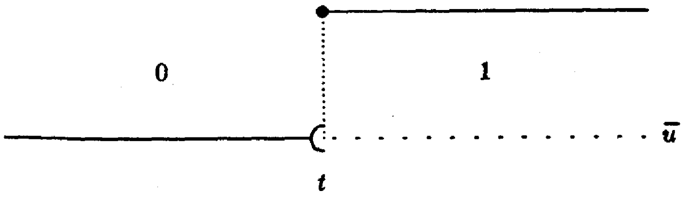


Figure 63. The step function $(\bar{u}-t)_+^0$

The "cup and ball" representation used here indicates visually that $(\bar{u}-t)_+^0$ has the value 0 from $-\infty$ up to, but not including, $\bar{u}=t$ and the value 1 at $\bar{u}=t$ and thereafter to $+\infty$. This is a convention we will use often in subsequent figures.

**Definition:** The $r^{th}$ one-sided power function is given by

$$(\bar{u}-t)_+^r \equiv (\bar{u}-t)_+^0(\bar{u}-t)^r \qquad (44)$$

$$\equiv \begin{cases} 0 & \bar{u} < t \\ (\bar{u}-t)^r & \bar{u} \geq t \end{cases} ,$$

where $r = 0,1,2,...$ and $(\bar{u}-t)_+^0$ is given by (43) above.

It is easily checked that $(\bar{u}-t)_+^r$ is $C^{r-1}$ continuous across any fixed $t$: there will be a match at $\bar{u}=t$ in value and in the first $r-1$ derivatives, while the $r^{\text{th}}$ derivative will be discontinuous at $\bar{u}=t$.

### 7.4. The One-sided Basis

The discontinuities in any spline we will ever construct are exactly like the discontinuity in

$$(\bar{u}-t)_+^0 ,$$

i.e. all our splines will be "open on the right" in any breakpoint interval. This means that they have one characterization up to, but not including, each breakpoint, and they will have another directly at and to the right of that breakpoint. This derives directly from the way in which we have defined $\mathbf{S}(\mathbf{P}^k,\{\bar{u}_i\}_0^{m+k})$, but it also follows from the way in which we will be representing splines using the functions $(\bar{u}-t)_+^r$.

For an arbitrary spline, when all knots have been considered, and all transitions from $p_{left}(\bar{u}) \rightarrow p_{right}(\bar{u})$ dealt with on the parameter interval $\bar{u}_{k-1} \leq \bar{u} < \bar{u}_{m+1}$, we will have made use of the following one-sided power functions in representing the spline:

$$\begin{array}{ll}
(\bar{u}-\bar{u}_{i_0})_+^{k-1}, (\bar{u}-\bar{u}_{i_0})_+^{k-2}, \ldots, (\bar{u}-\bar{u}_{i_0})_+^{k-\min(k,\mu_{i_0})} & \text{for } \bar{u}_{i_0} \\
(\bar{u}-\bar{u}_{i_1})_+^{k-1}, (\bar{u}-\bar{u}_{i_1})_+^{k-2}, \ldots, (\bar{u}-\bar{u}_{i_1})_+^{k-\min(k,\mu_{i_1})} & \text{for } \bar{u}_{i_1} \\
\bullet\ \bullet\ \bullet & \bullet\ \bullet\ \bullet \\
(\bar{u}-\bar{u}_{i_M})_+^{k-1}, (\bar{u}-\bar{u}_{i_M})_+^{k-2}, \ldots, (\bar{u}-\bar{u}_{i_M})_+^{k-\min(k,\mu_{i_M})} & \text{for } \bar{u}_{i_M}\ .
\end{array} \tag{45}$$

The first $k$ of the one-sided power functions in (45) derive from the knots $\bar{u}_0, \ldots, \bar{u}_{k-1}$, and the remaining ones are due to the knots which lie within the legal parameter range. The following points should be kept in mind about equations (45).

- Since $\bar{u} > \bar{u}_{k-1}$, we may write

  $$(\bar{u}-\bar{u}_0)^j, \ldots, (\bar{u}-\bar{u}_{k-1})^j = (\bar{u}-\bar{u}_0)_+^j, \ldots, (\bar{u}-\bar{u}_{k-1})_+^j$$

  for all powers $j$, that is we were able to affix the subscript "+" to each of the first $k$ functions in (45) without changing them on the legal parameter range.
- The first $k$ one-sided power functions are simply

  $$(\bar{u}-\bar{u}_0)_+^{k-1}, \ldots, (\bar{u}-\bar{u}_{k-1})_+^{k-1}$$

  if the $\bar{u}_0, \ldots, \bar{u}_{k-1}$ are distinct.

Schematically we have the following.

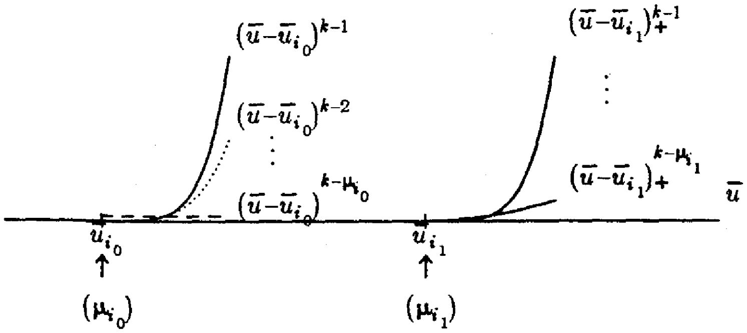

$(\bar{u}-\bar{u}_{i_0})^{k-1}$
$(\bar{u}-\bar{u}_{i_0})^{k-2}$
$(\bar{u}-\bar{u}_{i_0})^{k-\mu_{i_0}}$
$(\bar{u}-\bar{u}_{i_1})_+^{k-1}$
$(\bar{u}-\bar{u}_{i_1})_+^{k-\mu_{i_1}}$
$\bar{u}$
$\bar{u}_{i_0}$
$(\mu_{i_0})$
$\bar{u}_{i_1}$
$(\mu_{i_1})$

Figure 64. It is fairly easy to see how we can build, from these functions, an arbitrary spline having the sort of continuity at knots which we desire. The vertical scale here has been much reduced.

The functions in (45) are linearly independent, and it can easily be seen that each is a member of $\mathbf{S}(\mathbf{P}^k, \{\bar{u}_i\}_0^{m+k})$.

**Theorem:** The functions of (45) above form a basis for $\mathbf{S}(\mathbf{P}^k, \{\bar{u}_i\}_0^{m+k})$.

**Definition:** The functions $(\bar{u}-t)_+^r$, are called the *one-sided* or *truncated power basis* for $\mathbf{S}(\mathbf{P}^k, \{\bar{u}_i\}_0^{m+k})$.

By way of illustrating (45), consider the knot sequence

| $\bar{u}_0$ | $\bar{u}_1$ | $\bar{u}_2$ | $\bar{u}_3$ | $\bar{u}_4$ | $\bar{u}_5$ | $\bar{u}_6$ |
|---|---|---|---|---|---|---|
| 1 | 3 | 3 | 5 | 5 | 5 | 5 |

introduced as (39) in the last chapter. Let us use it to construct quadratic splines ($k=3$). We have

$$\bar{u}_{k-1} = \bar{u}_2 \quad \text{and} \quad \bar{u}_{m+1} = \bar{u}_4 \ .$$

If we choose

$$i_0 = 0 \ , \ i_1 = 1 \ , \ \text{and} \ i_2 = 3$$

then the basis (45) would be

$$(\bar{u}-\bar{u}_0)_+^2 \ , \ (\bar{u}-\bar{u}_1)_+^2 \ , \ (\bar{u}-\bar{u}_1)_+^1 \ , \ \text{and} \ (\bar{u}-\bar{u}_3)_+^2$$

and the space of all quadratic splines on this knot sequence has dimension 4.

To reinforce the concept of a one-sided basis, let's consider the uniform cubic B-spline given in equation (11) on

$$\{\bar{u}_0, \bar{u}_1, \bar{u}_2, \bar{u}_3, \bar{u}_4\} = \{ 0, 1, 2, 3, 4 \} \ .$$

It is, of course, a $C^2$ cubic spline. By a change of variables we have the alternative representation

$$B_0(\bar{u}) = \begin{cases} & 0 & -1 \le \bar{u} < 0 \\ b_{-0}(\bar{u}) = & \frac{1}{6}\bar{u}^3 & 0 \le \bar{u} < 1 \\ b_{-1}(\bar{u}) = & -\frac{1}{6}(3\bar{u}^3 - 12\bar{u}^2 + 12\bar{u} - 4) & 1 \le \bar{u} < 2 \\ b_{-2}(\bar{u}) = & \frac{1}{6}(3\bar{u}^3 - 24\bar{u}^2 + 60\bar{u} - 44) & 2 \le \bar{u} < 3 \\ b_{-3}(\bar{u}) = & -\frac{1}{6}(\bar{u}^3 - 12\bar{u}^2 + 48\bar{u} - 64) & 3 \le \bar{u} < 4 \\ & 0 & 4 \le \bar{u} < 5 \ . \end{cases}$$

As we have seen, the appropriate one-sided basis is

$$(\bar{u}-0)_+^3 , \ (\bar{u}-1)_+^3 , \ (\bar{u}-2)_+^3 , \ (\bar{u}-3)_+^3 , \ \text{and} \ (\bar{u}-4)_+^3 \ .$$

It is straightforward to verify that

$$B_0(\bar{u}) = \frac{1}{6}(\bar{u}-0)_+^3 - \frac{2}{3}(\bar{u}-1)_+^3 + 1(\bar{u}-2)_+^3 - \frac{2}{3}(\bar{u}-3)_+^3 + \frac{1}{6}(\bar{u}-4)_+^3 \ . \tag{47}$$

This is true, since for $0 \le \bar{u} < 1$ all terms in (47) after the equal sign are zero save the first, and (47) becomes

$$B_0(\bar{u}) = \frac{1}{6}(\bar{u}-0)_+^3$$

$$= \frac{1}{6}\overline{u}^3 \ .$$

For the range $1 \leq \overline{u} < 2$:

$$p_{left}(\overline{u}) = b_{-0} \ ,$$

$$p_{right}(\overline{u}) = b_{-1} \ ,$$

the knot at which these segment polynomials join is

$$\overline{u} = 1 \ ,$$

and consequently our previous discussions lead us to believe that

$$b_{-1} = b_{-0} + c(\overline{u}-1)^3$$

for some constant 3. Indeed,

$$c = \frac{b_{-1} - b_{-0}}{(\overline{u}-1)^3}$$

$$= \frac{-\frac{3}{6}\overline{u}^3 + \frac{12}{6}\overline{u}^2 - \frac{12}{6}\overline{u} + \frac{4}{6} - \frac{1}{6}\overline{u}^3}{(\overline{u}-1)^3}$$

$$= -\frac{2}{3} \ .$$

The verification follows this pattern for the remaining breakpoint intervals.

We can, in fact, construct any B-spline we have ever seen, or indeed any spline, as a linear combination of one-sided power functions. Why do we choose to use the B-splines instead of the simpler power functions? Because, for a number of reasons, the one-sided power functions are computationally unsatisfying. Their utility to us is not in constructing splines; for that purpose they suffer from two severe shortcomings: *numerical instability* and *lack of local control*. The utility of the one-sided basis is that it can be easily described and understood. From it we will define the basis we really want to use, namely the B-spline basis.

To elaborate on these shortcomings, consider the above example. The uniform B-spline that we constructed is representative of the curves and surfaces that we encounter in graphics: they do not behave wildly. Indeed, the uniform B-spline goes to zero to the left and to the right; a more randomly-chosen spline can be expected merely to maintain bounded behaviour throughout the region of interest. The one-sided basis, on the other hand, blows up as $\overline{u}$ increases. Hence if the one–sided basis is used to express "reasonable" spline curves and surfaces, the coefficients required to do this can be expected to alternate between large positive and negative values in order to force *numerical cancellation* of the basis function values as $\overline{u}$ increases. Cancellation is computationally undesirable.

A second shortcoming, from the point of view of graphics, is that the one-sided basis functions do not have *local support*; they are all nonzero on at least half the real line. If one represents a curve or surface in the usual way as a scaled sum of basis functions, the lack of local support translates into a lack of *local control*: the adjustment of any scale factor has an influence over the shape of the remainder of the curve. A change in the first scale factor will affect the entire curve. As a result, the adjustment of any scale factor will then give rise to a system of linear equations that must be solved to determine the effect of the adjustment on the curve or surface. The system of equations will be large, involving data from all the control vertices, and the continual need to solve such systems whenever vertices are adjusted is a bottleneck in real-time interactive graphical design. It is precisely because the uniform B-spline goes to zero outside of a closed, bounded interval that it is of such interest to us.

Finally, there is no intuitive relationship between the scale factors weighting the one-sided basis functions and the shape of the curve they define. On the other hand, we have already seen that the scale factors weighting the B-spline representation of a spline curve have a direct physical interpretation as control vertices.

### 7.5. Linear Combinations and Cancellation

The key to constructing a desirable basis from the less desirable (but conceptually simple) one-sided basis is to recognize that cancellation can occur and local support can be achieved analytically, by a symbolic process, before any numerical computation is begun. To this end we will rearrange the one-sided basis functions in the above example by taking linear combinations of them to produce new functions that behave in a much more bounded fashion.

Our game plan is as follows. We will begin with

$$(\bar{u}-0)_+^3 ,$$

which grows cubically for $\bar{u}\geq 0$, and

$$(\bar{u}-1)_+^3 ,$$

which grows cubically for $\bar{u}\geq 1$. By taking an appropriate linear combination

$$c_0(\bar{u}-0)_+^3 + c_1(\bar{u}-1)_+^3$$

of these two functions we can produce a third (combined) function whose $\bar{u}^3$ term is cancelled away for $\bar{u}\geq 1$. The three functions

$$(\bar{u}-0)_+^3$$

$$(\bar{u}-1)_+^3$$

and

$$c_0(\bar{u}-0)_+^3 + c_1(\bar{u}-1)_+^3$$

are linearly dependent, but any two of them will be independent. That is, any one of these three functions can be written as a linear combination of the other two, but no single one of them is merely a multiple of one of the others. Since any two of them can represent the third, any two of them can represent anything that could have been represented by the original two. Since

$$c_0(\bar{u}-0)_+^3 + c_1(\bar{u}-1)_+^3$$

is better behaved than either of the original two functions, i.e. it has no $\bar{u}^3$ term for $\bar{u}\geq 1$ and consequently "grows more slowly" as $\bar{u}\rightarrow +\infty$, we would like to use this combined function as a replacement for one of the original two to obtain a revised basis. Arbitrarily, we will use it to replace

$$(\bar{u}-0)_+^3 .$$

This combination process can be repeated for the pairs

$$(\bar{u}-1)_+^3 \text{ and } (\bar{u}-2)_+^3$$

$$(\bar{u}-2)^3_+ \text{ and } (\bar{u}-3)^3_+$$

and

$$(\bar{u}-3)^3_+ \text{ and } (\bar{u}-4)^3_+$$

to yield functions that can be substituted for

$$(\bar{u}-1)^3_+$$

$$(\bar{u}-2)^3_+$$

and

$$(\bar{u}-3)^3_+ \ .$$

More precisely, we see that

$$(\bar{u}-1)^3_+-(\bar{u}-0)^3_+ = \begin{cases} 0 & \bar{u}<0 \\ -\bar{u}^3 & 0 \le \bar{u} < 1 \\ -3\bar{u}^2+3\bar{u}-1 & 1 \le \bar{u} \ , \end{cases}$$

which grows quadratically for $\bar{u} \ge 1$.

Note that $(\bar{u}-1)^3_+-(\bar{u}-0)^3_+$ goes negative at $\bar{u}=0$, and continues to $-\infty$. From Figure 65 it is clear why this is so: $(\bar{u}-0)^3_+$ is positive between 0 and 1, so $-(\bar{u}-0)^3_+$ is negative between 0 and 1. Since $(\bar{u}-1)^3_+$ is zero in this interval, the entire expression is negative on [0,1).

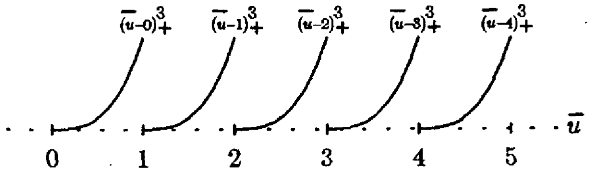


Figure 65. The one-sided basis functions for a uniform $C^2$ cubic spline.

Will this work more generally? Observe that

$$(\bar{u}-\bar{u}_{i+1})^3_+ - (\bar{u}-\bar{u}_i)^3_+ = \begin{cases} 0 & \bar{u} < \bar{u}_i \\ -(\bar{u}-\bar{u}_i)^3 & \bar{u}_i \le \bar{u} < \bar{u}_{i+1} \\ -3\bar{u}^2(\bar{u}_{i+1}-\bar{u}_i) + 3\bar{u}(\bar{u}_{i+1}-\bar{u}_i)(\bar{u}_{i+1}+\bar{u}_i) & \bar{u}_{i+1} \le \bar{u} \ . \\ \quad -(\bar{u}_{i+1}-\bar{u}_i)(\bar{u}^2_{i+1}+\bar{u}_{i+1}\bar{u}_i+\bar{u}^2_i) & \end{cases} \tag{48}$$

We are left, as for uniform knot spacing, with a function that grows only quadratically as $\bar{u} \to \infty$.

### 7.6. Cancellation as a Divided Difference

Our objective will now be to carry this one stage further. We will try to take two such "quadratically-growing" combinations of the one-sided power functions and combine them further so as to cancel away any $\bar{u}^2$ behaviour as $\bar{u}\to\infty$. It was easy to remove the $\bar{u}^3$ terms, for $\bar{u}>\bar{u}_{i+1}$, by combining the two one-sided power functions

$$(\bar{u}-\bar{u}_{i+1})_+^3 = \bar{u}^3 + \cdots$$

and

$$(\bar{u}-\bar{u}_i)_+^3 = \bar{u}^3 + \cdots$$

because their dominating ($\bar{u}^3$) terms had the same, constant coefficient (namely 1). The coefficient of the dominating term (as $\bar{u}\to\infty$) in the rightmost segment of (48), however, is troublesome — it depends on $i$. But notice that the factor $(\bar{u}_{i+1}-\bar{u}_i)$ occurs in every term of the function for $\bar{u}\geq\bar{u}_{i+1}$. We will divide this factor out to simplify matters. (For the time being we will assume that all the knots are distinct.) That is, we will consider the *divided difference* function

$$\frac{(\bar{u}-\bar{u}_{i+1})_+^3-(\bar{u}-\bar{u}_i)_+^3}{\bar{u}_{i+1}-\bar{u}_i} = \begin{cases} 0 & \bar{u}<\bar{u}_i \\ -\dfrac{(\bar{u}-\bar{u}_i)^3}{(\bar{u}_{i+1}-\bar{u}_i)} & \bar{u}_i\leq\bar{u}<\bar{u}_{i+1} \\ -3\bar{u}^2+3\bar{u}(\bar{u}_{i+1}+\bar{u}_i) & \bar{u}_{i+1}\leq\bar{u} \\ \quad -(\bar{u}_{i+1}^2+\bar{u}_{i+1}\bar{u}_i+\bar{u}_i^2) & \end{cases} \tag{49}$$

$$= \left(+\frac{1}{\bar{u}_{i+1}-\bar{u}_i}\right)(\bar{u}-\bar{u}_{i+1})_+^3 + \left(-\frac{1}{\bar{u}_{i+1}-\bar{u}_i}\right)(\bar{u}-\bar{u}_i)_+^3$$

because for $\bar{u}\geq\bar{u}_{i+1}$ it both cancels the cubic term and ensures that the remaining quadratic term will have the constant coefficient 3. This means that we are setting

$$c_0 = \left(-\frac{1}{\bar{u}_{i+1}-\bar{u}_i}\right)$$

and

$$c_1 = \left(+\frac{1}{\bar{u}_{i+1}-\bar{u}_i}\right) .$$

So as to have a short-hand notation for this expression we will write

$$[\bar{u}_i,\bar{u}_{i+1}:t](\bar{u}-t)_+^3 = \frac{(\bar{u}-\bar{u}_{i+1})_+^3-(\bar{u}-\bar{u}_i)_+^3}{\bar{u}_{i+1}-\bar{u}_i} \tag{50}$$

to indicate the operations of (1) selecting the two values $\bar{u}_{i+1}$ and $\bar{u}_i$, (2) substituting them for $t$ in two copies of $(\bar{u}-t)_+^3$, (3) subtracting the results, and (4) dividing by $\bar{u}_{i+1}-\bar{u}_i$. Thus our original divided difference for uniformly spaced knots is

$$[0,1:t](\bar{u}-t)_+^3 = \frac{(\bar{u}-1)_+^3-(\bar{u}-0)_+^3}{1-0}$$

$$= (+1)(\bar{u}-1)_+^3+(-1)(\bar{u}-0)_+^3 \ .$$

We emphasize that this divided difference function (50) is a ***linear combination*** of the one-sided power functions $(\bar{u}-\bar{u}_i)_+^3$ and $(\bar{u}-\bar{u}_{i+1})_+^3$ and that it can be used to *substitute* for either of these one-sided power functions in the collection of basis functions. Because it is a linear combination of $(\bar{u}-\bar{u}_i)_+^3$ and $(\bar{u}-\bar{u}_{i+1})_+^3$, its differentiability properties will be the "union" of those possessed by the two functions individually: it is $C^2$ at $\bar{u}_i$ and $\bar{u}_{i+1}$ and fully differentiable elsewhere.

We choose to let

$$[0,1:t](\bar{u}-t)_+^3$$

(shown in Figure 66) replace

$$(\bar{u}-0)_+^3 \ ,$$

and in doing so we have modified our basis into one which is "nicer" in the sense that its first one-sided member grows only quadratically as $\bar{u} \to \infty$.

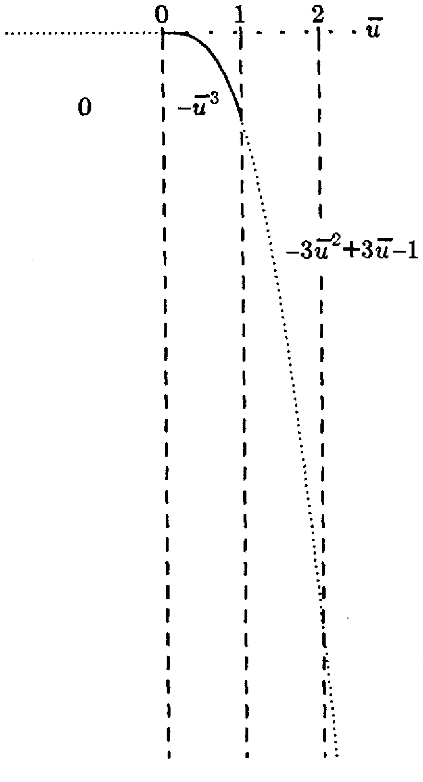


Figure 66. $[0,1:t](\bar{u}-t)_+^3$ This function is quadratic for $\bar{u} \geq 1$.

In a like fashion it may be verified that

$$[1,2:t](\bar{u}-t)_+^3 = \frac{(\bar{u}-2)_+^3-(\bar{u}-1)_+^3}{2-1}$$
$$= \begin{cases} 0 & \bar{u} < 1 \\ -(\bar{u}-1)^3 & 1 \le \bar{u} < 2 \\ -3\bar{u}^2 + 9\bar{u} - 7 & 2 \le \bar{u} \ , \end{cases}$$

that

$$[2,3:t](\bar{u}-t)_+^3 = \frac{(\bar{u}-3)_+^3-(\bar{u}-2)_+^3}{3-2}$$
$$= \begin{cases} 0 & \bar{u} < 2 \\ -(\bar{u}-2)^3 & 2 \le \bar{u} < 3 \\ -3\bar{u}^2 + 15\bar{u} - 19 & 3 \le \bar{u} \ , \end{cases}$$

and that

$$[3,4:t](\bar{u}-t)_+^3 = \frac{(\bar{u}-4)_+^3-(\bar{u}-3)_+^3}{4-3}$$
$$= \begin{cases} 0 & \bar{u} < 3 \\ -(\bar{u}-3)^3 & 3 \le \bar{u} < 4 \\ -3\bar{u}^2 + 21\bar{u} - 37 & 4 \le \bar{u} \ . \end{cases}$$

That is, the trick that worked with respect to the first and second knots will work with respect to the remaining adjacent pairs of knots, and we may use the above three functions to substitute for $(\bar{u}-1)_+^3$, $(\bar{u}-2)_+^3$, and $(\bar{u}-3)_+^3$, respectively.

Thus we can replace some of the "eventually-cubic" functions in our original basis with these "eventually-quadratic" functions. In particular,

$$\begin{array}{lll} [0,1:t](\bar{u}-t)_+^3 & \text{replaces} & (\bar{u}-0)_+^3 \\ [1,2:t](\bar{u}-t)_+^3 & \text{replaces} & (\bar{u}-1)_+^3 \\ [2,3:t](\bar{u}-t)_+^3 & \text{replaces} & (\bar{u}-2)_+^3 \\ [3,4:t](\bar{u}-t)_+^3 & \text{replaces} & (\bar{u}-3)_+^3 \ . \end{array}$$

Figure 67 illustrates this process. The original basis appears on the left in Figure 67, consisting of one-sided power functions which grow as $\bar{u}^3$ for $\bar{u} \rightarrow +\infty$. The new basis appears on the right, and consists of four functions that begin to go negative cubically but eventually grow as $-\bar{u}^2$ for $\bar{u} \rightarrow +\infty$, and one of the original truncated cubics which cannot be replaced because there is no truncated cubic to its right with which it can be differenced.

### 7.7. Cancelling the Quadratic Term — The Second Difference

In the previous section we saw that dividing by the knot spacing $\bar{u}_{i+1}-\bar{u}_i$ set us up to repeat the cancellation process by ensuring that the coefficient of the quadratic term for the rightmost segment was the constant 3. What does the difference

$$[\bar{u}_{i+1},\bar{u}_{i+2}:t](\bar{u}-t)_+^3 - [\bar{u}_i,\bar{u}_{i+1}:t](\bar{u}-t)_+^3$$

look like for sufficiently large $\bar{u}$? Well, the first term is a spline with breakpoints at $\bar{u}_{i+1}$ and $\bar{u}_{i+2}$, while

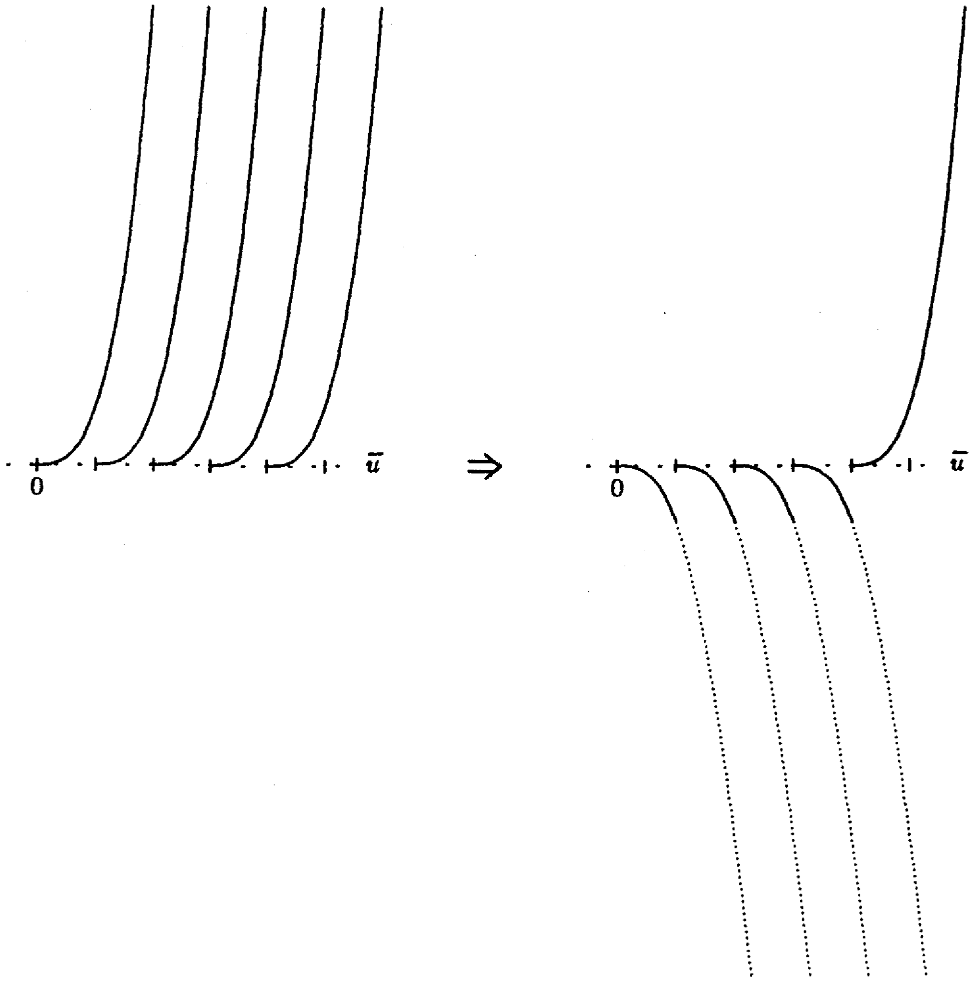


Figure 67. All five of the function shown on the left are cubic. The first four are replaced by four "eventually-quadratic" functions on the right. Each "eventually-quadratic" function consists of three polynomial segments: the first is identically zero, begins at $-\infty$, and is not plotted; the second, drawn as a solid line, is cubic; the third, drawn dotted, is quadratic and continues indefinitely to $-\infty$.

The fifth function on the left cannot be replaced by an "eventually-quadratic" function because there is no one-sided cubic to its right with which it can be differenced. We will deal with this technicality later.

the second term is a spline with breakpoints at $\bar u_i$ and $\bar u_{i+1}$. The difference will therefore have breakponts at $\bar u_i$, $\bar u_{i+1}$ and $\bar u_{i+2}$. Since we are interested in the asymptotic behaviour of this difference as $\bar u \to +\infty$, it is sufficient for our purposes to compute the difference of the rightmost segments for these two terms. From (49) we see that

$$[\bar u_i, \bar u_{i+1}:t](\bar u - t)^3_+ = -3\bar u^2 + 3\bar u(\bar u_{i+1} + \bar u_i) - (\bar u_{i+1}^2 + \bar u_{i+1}\bar u_i + \bar u_i^2) \ ,$$

from which it is clear that

$$[\bar u_{i+1}, \bar u_{i+2}:t](\bar u - t)^3_+ = -3\bar u^2 + 3\bar u(\bar u_{i+2} + \bar u_{i+1}) - (\bar u_{i+2}^2 + \bar u_{i+2}\bar u_{i+1} + \bar u_{i+1}^2) \ .$$

Hence

$$[\bar{u}_{i+1},\bar{u}_{i+2}:t](\bar{u}-t)^3_+ - [\bar{u}_i,\bar{u}_{i+1}:t](\bar{u}-t)^3_+ = 3\bar{u}(\bar{u}_{i+2}-\bar{u}_i) - (\bar{u}_{i+2}-\bar{u}_i)(\bar{u}_{i+2}+\bar{u}_{i+1}+\bar{u}_i) \ .$$

The quadratic term has disappeared, as expected. In order to obtain a constant coefficient for the linear term, it is clear that we need to divide by $\bar{u}_{i+2}-\bar{u}_i$. What we want to compute, then, is

$$\frac{\dfrac{(\bar{u}-\bar{u}_{i+2})^3_+-(\bar{u}-\bar{u}_{i+1})^3_+}{\bar{u}_{i+2}-\bar{u}_{i+1}} - \dfrac{(\bar{u}-\bar{u}_{i+1})^3_+-(\bar{u}-\bar{u}_i)^3_+}{\bar{u}_{i+1}-\bar{u}_i}}{\bar{u}_{i+2}-\bar{u}_i}$$

$$= 3\bar{u} - \bar{u}_{i+2} - \bar{u}_{i+1} - \bar{u}_i \qquad \text{for } \bar{u} \geq \bar{u}_{i+2} \ .$$

Expanding our short-hand notation, we write this as

$$[\bar{u}_i,\bar{u}_{i+1},\bar{u}_{i+2}:t](\bar{u}-t)^3_+ \ .$$

This *second difference* goes positive at $\bar{u}_i$. It is easy to figure out why: we saw earlier that $[\bar{u}_i,\bar{u}_{i+1}:t](\bar{u}-t)^3_+$ goes negative, so $-[\bar{u}_i,\bar{u}_{i+1}:t](\bar{u}-t)^3_+$ goes positive. Since $[\bar{u}_{i+1},\bar{u}_{i+2}:t](\bar{u}-t)^3_+$ is zero between $\bar{u}_i$ and $\bar{u}_{i+1}$ (and $\bar{u}_{i+2}>\bar{u}_i$, so that the denominator is positive), the difference (51) is initially positive. (In fact it remains so.)

Notice that

$$[\bar{u}_i,\bar{u}_{i+1},\bar{u}_{i+2}:t](\bar{u}-t)^3_+ = \frac{[\bar{u}_{i+1},\bar{u}_{i+2}:t](\bar{u}-t)^3_+ - [\bar{u}_i,\bar{u}_{i+1}:t](\bar{u}-t)^3_+}{\bar{u}_{i+2}-\bar{u}_i} \ . \tag{51}$$

This suggests a recursive definition for divided differences — such a definition is to be a subject of the next chapter.

Our second difference, then, is a function which grows only linearly as $\bar{u}\to\infty$. Such functions can be used to replace the functions of (50), which in turn replaced certain of the original one-sided power functions, to yield an even "nicer" basis containing functions which grow only linearly as $\bar{u}\to+\infty$.

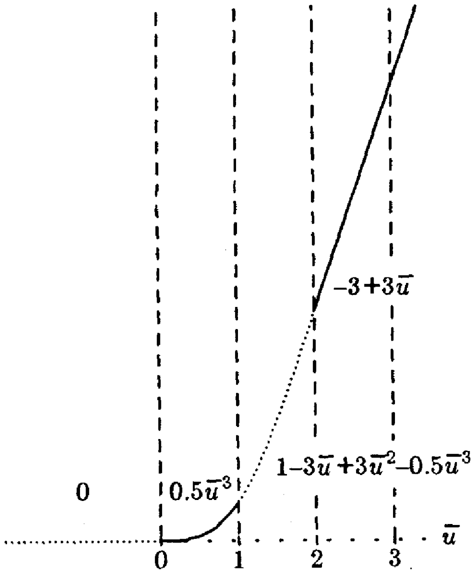


Figure 68. $[0,1,2:t](\bar{u}-t)^3_+$ This function is linear for $\bar{u}\geq 2$.

In the case of our example,

$$[0,1,2{:}t](\bar{u}-t)_+^3$$

$$= \frac{\dfrac{(\bar{u}-2)_+^3-(\bar{u}-1)_+^3}{2-1} - \dfrac{(\bar{u}-1)_+^3-(\bar{u}-0)_+^3}{1-0}}{2-0} = \begin{cases} 0 & \bar{u}<0 \\ \frac{1}{2}(\bar{u})^3 & 0 \le \bar{u} < 1 \\ -\frac{1}{2}(\bar{u}^3-6\bar{u}^2+6\bar{u}-2) & 1 \le \bar{u} < 2 \\ \frac{1}{2}(6\bar{u}-6) & 2 \le \bar{u} \ . \end{cases}$$

As expected, this function is linear for $\bar{u} \ge 2$ (see Figure 68). Now that we know how to cancel the quadratic term, we can do this for each successive pair of eventually-quadratic basis functions:

$$\begin{array}{lll} [0,1,2{:}t](\bar{u}-t)_+^3 & \text{replaces} & [0,1{:}t](\bar{u}-t)_+^3 \\ [1,2,3{:}t](\bar{u}-t)_+^3 & \text{replaces} & [1,2{:}t](\bar{u}-t)_+^3 \\ [2,3,4{:}t](\bar{u}-t)_+^3 & \text{replaces} & [2,3{:}t](\bar{u}-t)_+^3 \ . \end{array}$$

Figure 69 illustrates this process.

## 7.8. Cancelling the Linear Term — The Third Difference

In the preceding section we accomplished the replacement of $(\bar{u}-\bar{u}_i)_+^3$ by

$$[\bar{u}_i,\bar{u}_{i+1},\bar{u}_{i+2}{:}t](\bar{u}-t)_+^3$$
$$= 3\bar{u} - \bar{u}_{i+2} - \bar{u}_{i+1} - \bar{u}_i \qquad \text{for } \bar{u} \ge \bar{u}_{i+2} \ .$$

In the same way we replaced $(\bar{u}-\bar{u}_{i+1})_+^3$ by

$$[\bar{u}_{i+1},\bar{u}_{i+2},\bar{u}_{i+3}{:}t](\bar{u}-t)_+^3$$
$$= 3\bar{u} - \bar{u}_{i+3} - \bar{u}_{i+2} - \bar{u}_{i+1} \qquad \text{for } \bar{u} \ge \bar{u}_{i+3} \ ,$$

and so on. Since

$$[\bar{u}_{i+1},\bar{u}_{i+2},\bar{u}_{i+3}{:}t](\bar{u}-t)_+^3 - [\bar{u}_i,\bar{u}_{i+1},\bar{u}_{i+2}{:}t](\bar{u}-t)_+^3 = (\bar{u}_{i+3}-\bar{u}_i) \ ,$$

we now replace $[\bar{u}_i,\bar{u}_{i+1},\bar{u}_{i+2}{:}t](\bar{u}-t)_+^3$ by the *third difference*

$$\frac{[\bar{u}_{i+1},\bar{u}_{i+2},\bar{u}_{i+3}{:}t](\bar{u}-t)_+^3 - [\bar{u}_i,\bar{u}_{i+1},\bar{u}_{i+2}{:}t](\bar{u}-t)_+^3}{\bar{u}_{i+3}-\bar{u}_i}$$
$$= \frac{3\bar{u} - \bar{u}_{i+3} - \bar{u}_{i+2} - \bar{u}_{i+1} - 3\bar{u} + \bar{u}_{i+2} + \bar{u}_{i+1} + \bar{u}_i}{\bar{u}_{i+3}-\bar{u}_i} \qquad \text{for } \bar{u} \ge \bar{u}_{i+3}$$
$$= -1 \qquad \text{for } \bar{u} \ge \bar{u}_{i+3} \ .$$

which we will denote by

$$[\bar{u}_i,\bar{u}_{i+1},\bar{u}_{i+2},\bar{u}_{i+3}{:}t](\bar{u}-t)_+^3 \ .$$

In a like manner we replace $[\bar{u}_{i+1},\bar{u}_{i+2},\bar{u}_{i+3}{:}t](\bar{u}-t)_+^3$ by

$$[\bar{u}_{i+1},\bar{u}_{i+2},\bar{u}_{i+3},\bar{u}_{i+4}{:}t](\bar{u}-t)_+^3$$
$$= \frac{[\bar{u}_{i+2},\bar{u}_{i+3},\bar{u}_{i+4}{:}t](\bar{u}-t)_+^3 - [\bar{u}_{i+1},\bar{u}_{i+2},\bar{u}_{i+3}{:}t](\bar{u}-t)_+^3}{\bar{u}_{i+4}-\bar{u}_{i+1}}$$

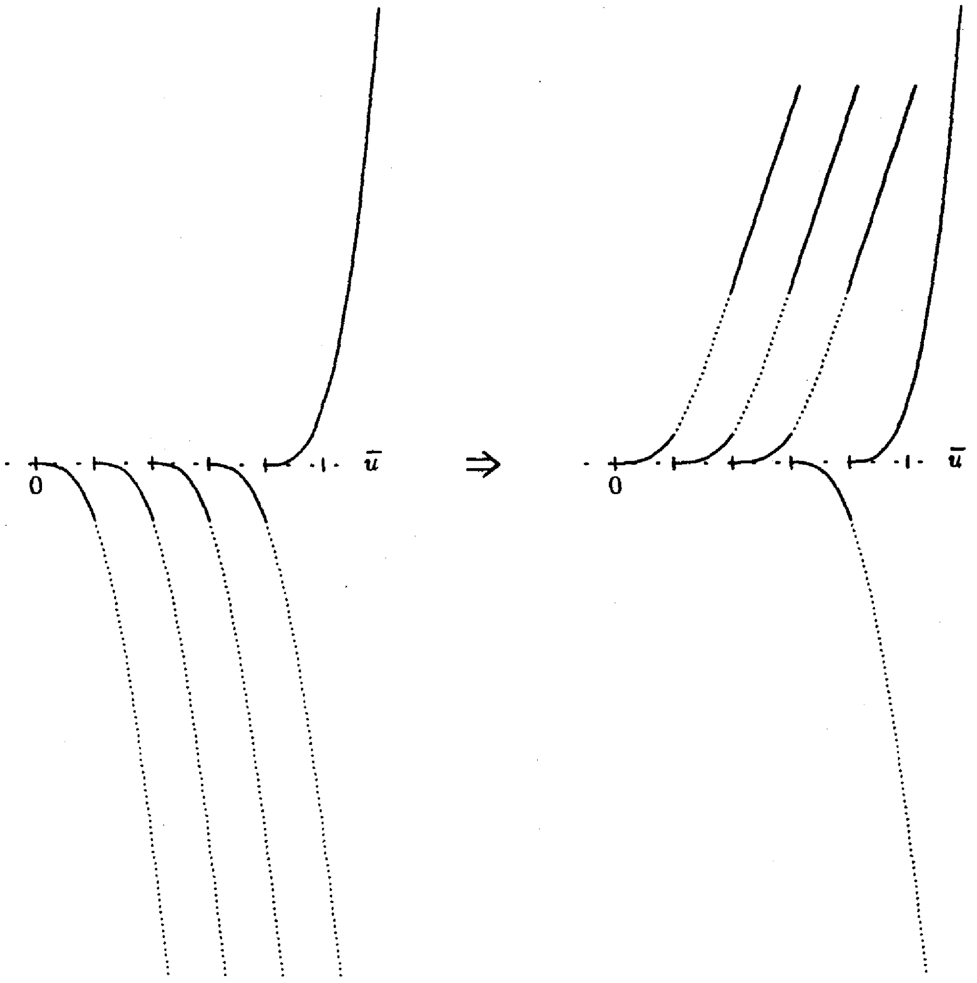


Figure 69. Taking a second divided difference of two "eventually-quadratic" functions allows us to obtain "eventually-linear" functions. Since we only have four "eventually-quadratic" functions on the left, we can only do this three times.

$$= \frac{3\bar{u} - \bar{u}_{i+4} - \bar{u}_{i+3} - \bar{u}_{i+2} - 3\bar{u} + \bar{u}_{i+3} + \bar{u}_{i+2} + \bar{u}_{i+1}}{\bar{u}_{i+4} - \bar{u}_{i+1}} \qquad \text{for } \bar{u} \geq \bar{u}_{i+4}$$

$$= -1 \qquad \text{for } \bar{u} \geq \bar{u}_{i+4} \ ,$$

and so on.

Returning to our example, we have

$$[1,2,3,4:t](\bar{u}-t)_+^3$$

$$= \frac{[1,2,3:t](\bar{u}-t)_+^3 - [0,1,2:t](\bar{u}-t)_+^3}{3-0}$$

$$\frac{\dfrac{\dfrac{(\bar{u}-3)^3_+-(\bar{u}-2)^3_+}{3-2}-\dfrac{(\bar{u}-2)^3_+-(\bar{u}-1)^3_+}{2-1}}{3-1}-\dfrac{\dfrac{(\bar{u}-2)^3_+-(\bar{u}-1)^3_+}{2-1}-\dfrac{(\bar{u}-1)^3_+-(\bar{u}-0)^3_+}{1-0}}{2-0}}{3-0}$$

$$= \begin{cases} 0 & \bar{u}<0 \\ -\frac{1}{6}(\bar{u}^3) & 0\le\bar{u}<1 \\ \frac{1}{6}(2\bar{u}^3-9\bar{u}^2+9\bar{u}-3) & 1\le\bar{u}<2 \\ -\frac{1}{6}(\bar{u}^3-9\bar{u}^2+27\bar{u}-21) & 2\le\bar{u}<3 \\ -1 & 3\le\bar{u}\ . \end{cases}$$

(See Figure 70.) As expected, it is constant for $\bar{u}\ge 3$.

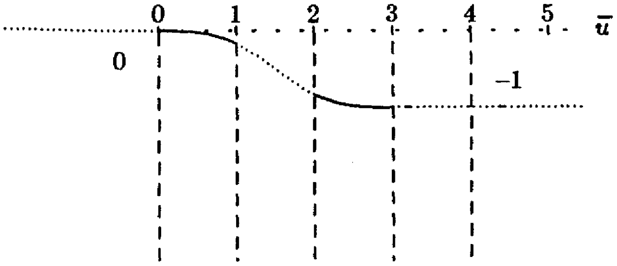


Figure 70. $[0,1,2,3:t](\bar{u}-t)^3_+$ This function is constant for $\bar{u}\ge 3$.

As before, we can now replace successive pairs of eventually-linear functions with eventually-constant functions:

$[0,1,2,3:t](\bar{u}-t)^3_+$ replaces $[0,1,2:t](\bar{u}-t)^3_+$

$[1,2,3,4:t](\bar{u}-t)^3_+$ replaces $[1,2,3:t](\bar{u}-t)^3_+$ .

Figure 71 illustrates the result.

### 7.9. The Uniform Cubic B-spline — A Fourth Difference

The third divided differences are constant, in fact −1, for sufficiently large $\bar{u}$. This avoids the need to cancel large positive values with large negative values. To obtain locality, that is, to obtain functions that return all the way to zero, requires the computation of one more difference. To be consistent with earlier steps we compute

$$\frac{[\bar{u}_{i+1},\bar{u}_{i+2},\bar{u}_{i+3},\bar{u}_{i+4}:t](\bar{u}-t)^3_+ - [\bar{u}_i,\bar{u}_{i+1},\bar{u}_{i+2},\bar{u}_{i+3}:t](\bar{u}-t)^3_+}{\bar{u}_{i+4}-\bar{u}_i} \tag{52}$$

$$= 0 \qquad\qquad \text{for } \bar{u}\ge\bar{u}_{i+4}\ ,$$

although the division by $\bar{u}_{i+4}-\bar{u}_i$ is actually superfluous. We denote this quantity by

$$[\bar{u}_i,\bar{u}_{i+1},\bar{u}_{i+2},\bar{u}_{i+3},\bar{u}_{i+4}:t](\bar{u}-t)^3_+\ . \tag{53}$$

For the uniform knot sequence we have been using as an example, $[0,1,2,3,4:t](\bar{u}-t)^3_+$ is exactly the spline of equation (46) — the uniform cubic B-spline for the knots {0,1,2,3,4} (see Figure 72).

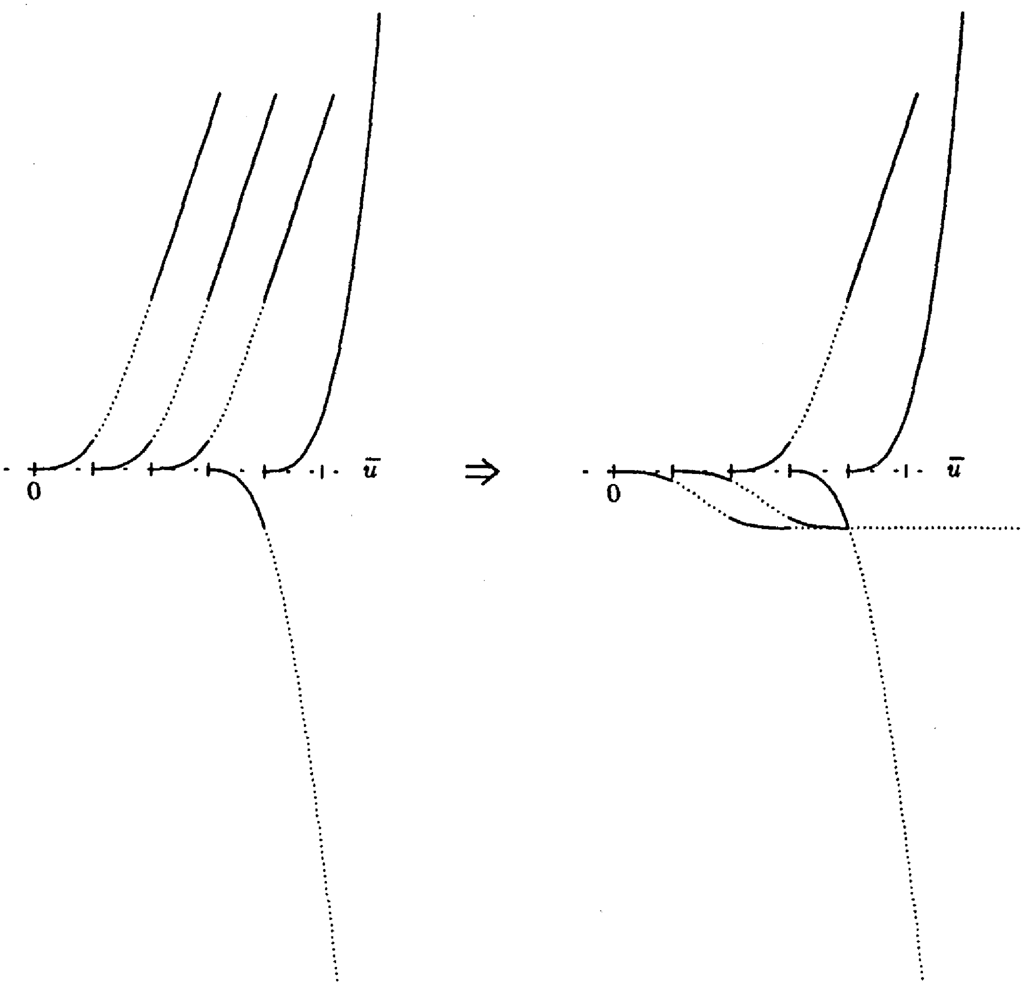


Figure 71. One more difference cancels the coefficient of a linear term to 0, leaving us with a constant function -1 for sufficiently large values of $\bar{u}$. Since we have three eventually-linear functions, we can do this twice to produce two eventually-constant functions.

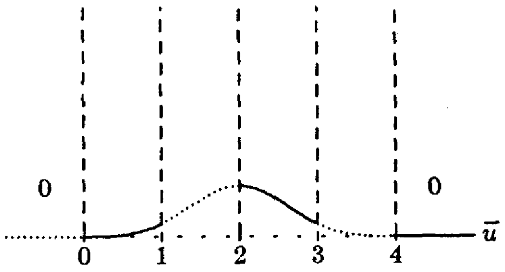


Figure 72. $[0,1,2,3,4{:}t](\bar{u}-t)^3_+$ This is the uniform cubic B-spline we met in Chapter 4.

We can use this eventually-zero function as a substitute for the eventually-constant function (see Figure 73), which substituted for the eventually-linear function, which substituted for the eventually-quadratic function, which substituted for the original cubic function $(\bar{u}-\bar{u}_i)^3_+$. We arrive at this eventually-zero function by combining differences of $(\bar{u}-0)^3_+, \ldots, (\bar{u}-4)^3_+$ so as to cancel, in succession,

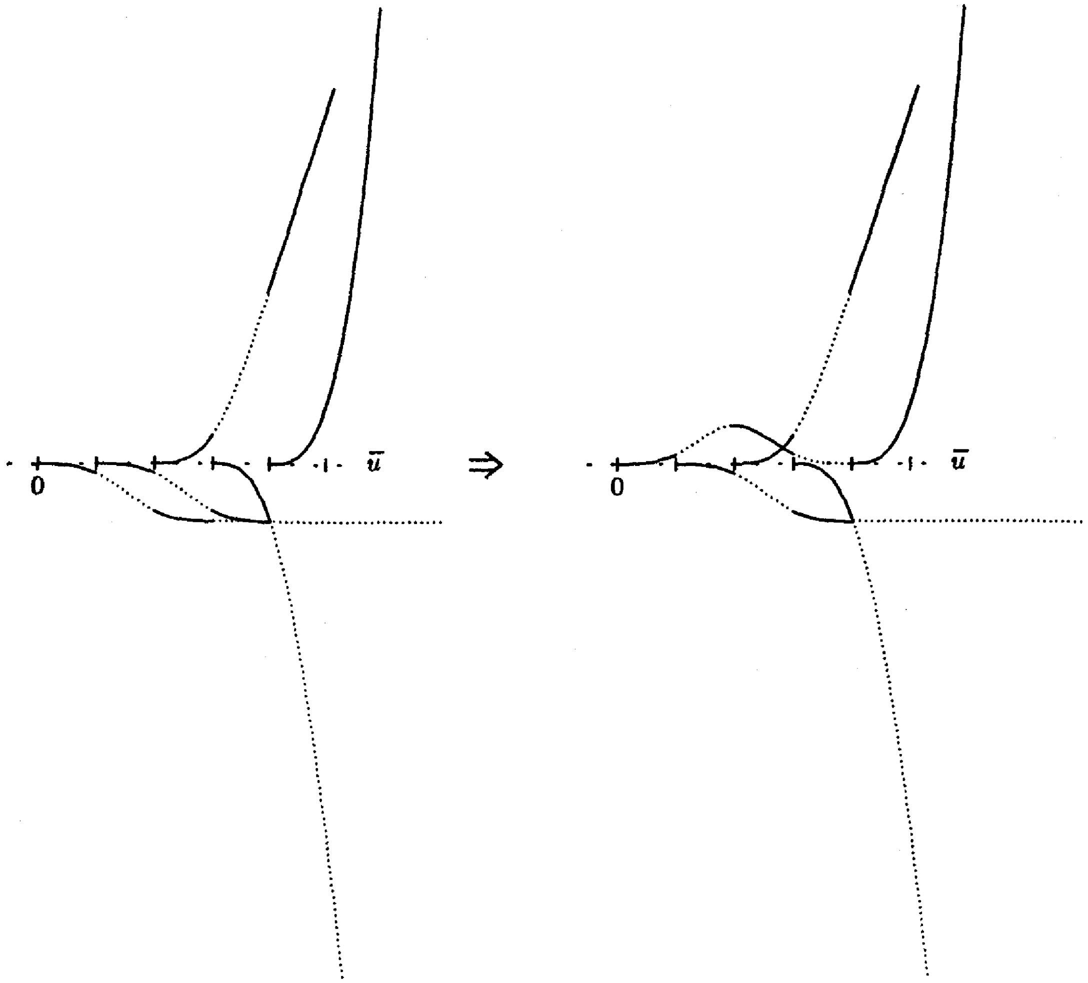


Figure 73. The final differencing step cancels away the constant term for large enough $\bar{u}$, resulting in a piecewise polynomial which is nonzero for only four intervals. The two eventually-constant functions on the left are exactly what we need to produce the single eventually-zero function on the right that has been our objective for the last few pages.

the powers $\bar{u}^3$, $\bar{u}^2$, $\bar{u}^1$, and $\bar{u}^0$.

For future reference we note that because (53) involves an even number of differences it will go positive at $\bar{u}_i$.

We emphasize that for the knot sequences we have been considering, the divided difference notation we have been using is simply shorthand. If we let $f_i = (\bar{u} - \bar{u}_{i+j})_+^3$ then

$[\bar{u}_i,\bar{u}_{i+1},\bar{u}_{i+2},\bar{u}_{i+3},\bar{u}_{i+4}:t](\bar{u}-t)_+^3 =$

$$\frac{\dfrac{\dfrac{\dfrac{f_1-f_0}{\bar u_1-\bar u_0}-\dfrac{f_2-f_1}{\bar u_2-\bar u_1}}{\bar u_2-\bar u_0}-\dfrac{\dfrac{f_2-f_1}{\bar u_2-\bar u_1}-\dfrac{f_3-f_2}{\bar u_3-\bar u_2}}{\bar u_3-\bar u_1}}{\bar u_3-\bar u_0}-\dfrac{\dfrac{\dfrac{f_2-f_1}{\bar u_2-\bar u_1}-\dfrac{f_3-f_2}{\bar u_3-\bar u_2}}{\bar u_3-\bar u_1}-\dfrac{\dfrac{f_3-f_2}{\bar u_3-\bar u_2}-\dfrac{f_4-f_3}{\bar u_4-\bar u_3}}{\bar u_4-\bar u_2}}{\bar u_4-\bar u_1}}{\bar u_4-\bar u_0} .$$

Admittedly this is a rather complicated expression, but we have seen why each of the terms appears. It is also clear from this expression that all we have done is replace $(\bar u-\bar u_i)_+^3$ by a particularly desirable linear combination of $(\bar u-\bar u_i)_+^3$, $(\bar u-\bar u_{i+1})_+^3$, $(\bar u-\bar u_{i+2})_+^3$, $(\bar u-\bar u_{i+3})_+^3$ and $(\bar u-\bar u_{i+4})_+^3$.

Our development has shown us how to obtain the single eventually-zero basis function shown in Figure 73. In actual fact we start with a larger number of one-sided cubics, and replace as many as we can at each step — five truncated cubics would have allowed us to compute two eventually-zero basis functions, and so on. So as to be able to replace all the truncated cubics needed to represent the curve in question, we add four more one-sided cubics, positioned arbitrarily to the right of $\bar u_m$ (see Figure 74).

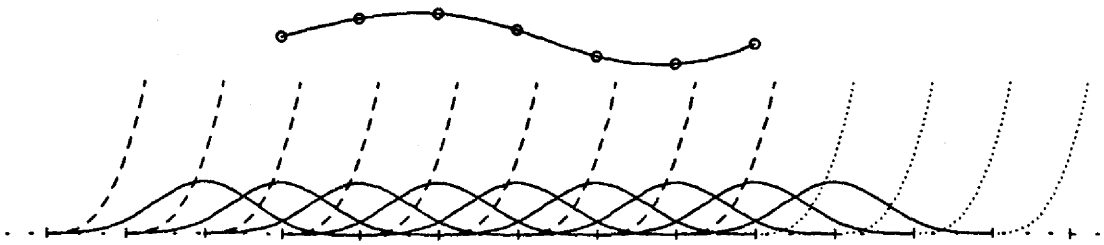


Figure 74. The dotted curves on the right are truncated cubics. They make it possible to difference the rightmost of the original cubics (which are drawn dashed) in order to replace all of the original truncated power functions by B-splines. Notice that there are an equal number of each.

Thus it appears that B-splines can be constructed by differencing one-sided power functions. This proves to be the case, but a little more preparation is needed if we are going to handle completely arbitrary knot sequences: the differencing process we have developed breaks down if two or more knots move together. If we had encountered a situation in which $\bar u_i=\bar u_{i+1}$, we would at some point have divided by zero. The remedy for this difficulty becomes obvious if we watch what happens as $\bar u_i$ and $\bar u_{i+1}$ "move close together," namely:

$$\lim_{\bar u_i\to\bar u_{i+1}} -\frac{(\bar u-\bar u_{i+1})^{k-1}-(\bar u-\bar u_i)^{k-1}}{\bar u_{i+1}-\bar u_i} = \frac{d}{dt}(\bar u-t)_{+_{k-1}}\Big|_{t=u_i} .$$

This suggests that it would be useful to study derivatives of one-sided power functions, and to expand the idea of differencing to include differentiation when multiple knots are encountered.

This, then, is the motivation for the next chapter. In it we will digress briefly from splines to formally introduce the divided difference operation, to study its relationship to differentiation, and to consider the properties of one-sided power functions under differencing and differentiation.

# 8. Divided Differences

In the previous section three important things occurred: we introduced the one-sided basis for the splines $\mathbf{S}(\mathbf{P}^k,\{\bar{u}_i\}_0^{m+k})$, we motivated the consideration of divided differences, and we gave the (correct) impression that uniform B-splines can be represented as divided differences of the one-sided power functions. A logical development would proceed as follows from this background.

- $\mathbf{S}(\mathbf{P}^k,\{\bar{u}_i\}_0^{m+k})$ is a vector space, and we know a set of functions that form a basis for that space.
- Any basis for a vector space can be obtained from suitable linear combinations of the elements of any other basis.
- The divided difference operation, at least in the uniform-knot case, is a mechanism for
  a) constructing linear combinations of functions, and
  b) manufacturing the B-splines.

  We should, therefore, investigate the extent to which, for non-uniform knot sequences, functions like B-splines can be produced from differencing the one-sided power functions.
- We should further investigate the extent to which these difference-manufactured B-splines can be used to generate all splines in $\mathbf{S}(\mathbf{P}^k,\{\bar{u}_i\}_0^{m+k})$.

Such an investigation is our next objective. To begin, however, we must develop a tool kit of results about the one-sided power functions and their interactions with derivative and difference operations.

### 8.1. Differentiation and One-sided Power Functions

We will begin with differentiation. Notice that $(\bar{u}-t)_+^r$ suffers a discontinuity (as a function of $\bar{u}$ for fixed $t$) in the $r^{\text{th}}$ derivative when $\bar{u}=t$, where it is a $C^{r-1}$ function, and that it is a $C^r$ function (at least) everywhere else. In fact, even the discontinuity is not too serious. The discontinuity in any derivative of any spline in $\mathbf{S}(\mathbf{P}^k,\{\bar{u}_i\}_0^{m+k})$ derives ultimately from the behaviour of the function $(\bar{u}-t)_+^0$, and since this function is "open on the right", we can easily verify (and we do just that in the next couple of pages) that a right-handed derivative of $(\bar{u}-t)_+^r$ with respect to $\bar{u}$ does exist at the point $\bar{u}=t$. Moreover, at all other values of $\bar{u}$ the right-handed derivative exists and is equivalent to the standard derivative of the function.

**Definition:** The derivative

$$D_{\bar{u}} f(\bar{u})$$

of any function $f(\bar{u})$ is said to be taken in the *right-handed* sense, if

$$D_{\bar{u}} f(\bar{u}) = \lim_{\substack{\epsilon \to 0 \\ \epsilon > 0}} \frac{f(\bar{u}+\epsilon) - f(\bar{u})}{\epsilon} .$$

Note that the limit is approached from the positive side. Since $\epsilon$ is positive, $\bar{u}+\epsilon$ lies to the right of $\bar{u}$.

**Convention:** The derivatives of the one-sided power function $(\bar{u}-t)_+^r$ with respect to $\bar{u}$ for fixed $t$ are:

for the zero$^{\text{th}}$ derivative,

$$D_{\bar{u}}^{(0)}(\bar{u}-t)_+^r \equiv (\bar{u}-t)_+^r \; ;$$

for the first derivative,

$$D_{\bar{u}}^{(1)}(\bar{u}-t)_+^r \equiv D_{\bar{u}}(\bar{u}-t)_+^r \equiv \lim_{\substack{\epsilon \to 0 \\ \epsilon > 0}} \frac{(\bar{u}+\epsilon-t)_+^r - (\bar{u}-t)_+^r}{\epsilon} \; ;$$

and recursively,

$$D_{\bar{u}}^{(l)}(\bar{u}-t)_+^r \equiv D_{\bar{u}}[D_{\bar{u}}^{(l-1)}(\bar{u}-t)_+^r]$$

for all the succeeding derivatives, $l=2,3, \cdots$ (understood in the right-handed sense).

Consider what this means for $r=0$ and $l=1$. For any chosen $\bar{u}<t$,

$$(\bar{u}+\epsilon-t)_+^0 - (\bar{u}-t)_+^0 = 0-0 = 0$$

for all $\epsilon>0$ small enough. Hence the limit defined above for $\epsilon\to 0$ is equal to zero. On the other hand, for any chosen $\bar{u}\geq t$,

$$(\bar{u}+\epsilon-t)_+^0 - (\bar{u}-t)_+^0 = 1-1 = 0$$

for all $\epsilon>0$. Again, the limit defined above is zero. This means that the one-sided power function $(\bar{u}-t)_+^0$, for variable $\bar{u}$ and fixed $t$, behaves exactly like a constant (a polynomial of order 1) under the application of right-handed differentiation with respect to $\bar{u}$. Since $(\bar{u}-t)_+^0$ is a simple spline of order 1, this is very appealing. Furthermore, for higher orders the product rule for differentiation may be applied, giving

$$
\begin{aligned}
D_{\bar{u}}(\bar{u}-t)_+^r &= D_{\bar{u}}[(\bar{u}-t)_+^0(\bar{u}-t)^r] \\
&= [D_{\bar{u}}(\bar{u}-t)_+^0]\cdot(\bar{u}-t)^r + (\bar{u}-t)_+^0\cdot[D_{\bar{u}}(\bar{u}-t)^r] \\
&= 0\cdot(\bar{u}-t)^r + (\bar{u}-t)_+^0\cdot[r(\bar{u}-t)^{r-1}] \\
&= r(\bar{u}-t)_+^{r-1} \ .
\end{aligned}
\tag{54}
$$

Hence we have the following.

**Theorem:** For all $r,l \geq 0$,

$$
D_{\bar{u}}^{(l)}(\bar{u}-t)_+^r = \begin{cases} \dfrac{r!}{(r-l)!}(\bar{u}-t)_+^{r-l} & \text{for } l \leq r \\ 0 & \text{for } l > r \ . \end{cases}
$$

This is also very appealing. It means, under the agreement that we consider only right-handed derivatives whenever $(\bar{u}-t)_+^r$ is being regarded as a function of $\bar{u}$ for fixed $t$, that $(\bar{u}-t)_+^r$ behaves just like the ordinary polynomial $(\bar{u}-t)^r$.

Notice that $(\bar{u}-t)_+^r$ is also a function of $t$. As such, for fixed $\bar{u}$, it is at least a $C^r$ function of $t$, except when $t=\bar{u}$, and at that critical value of $t$ it has a left-handed derivative with respect to $t$. We can see this best by turning the equations (43) and (44) around to look at them from the "$t$ point of view." Equation (43) becomes

$$
(\bar{u}-t)_+^0 \equiv \begin{cases} 1 & t \leq \bar{u} \\ 0 & t > \bar{u} \end{cases}
$$

and (44) becomes

$$
\begin{aligned}
(\bar{u}-t)_+^r &\equiv (\bar{u}-t)_+^0(\bar{u}-t)^r \\
&\equiv \begin{cases} (\bar{u}-t)^r & t \leq \bar{u} \\ 0 & t > \bar{u} \end{cases}
\end{aligned}
$$

and Figure 63 becomes

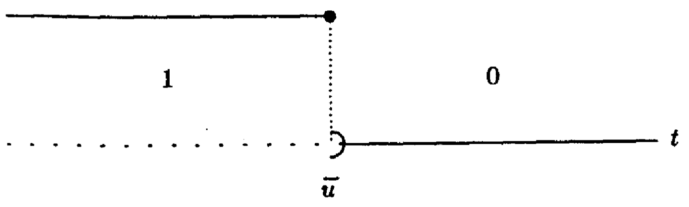


Figure 75.

Somewhat later we will begin dealing with $(\bar{u}-t)_+^r$ as a function of $t$. Observe that, as such, it is "open on the left". Consequently our intuition with respect to $\bar{u}$ can be applied to $t$ as well. For example:

**Definition:** The derivative

$$D_t\, g(t)$$

of any function $g(t)$ is said to be taken in the *left-handed* sense if

$$D_t\, g(t) \equiv \lim_{\substack{\epsilon \to 0 \\ \epsilon > 0}} \frac{g(t-\epsilon) - g(t)}{-\epsilon} \quad ,$$

Note that the limit is approached from the negative side. Since $\epsilon$ is positive, $t-\epsilon$ lies to the left of $t$.

**Convention:** The derivatives of the one-sided power function $(\bar{u}-t)_+^r$ with respect to $t$ for fixed $\bar{u}$ are:

for the zero-th derivative,

$$D_t^{(0)}(\bar{u}-t)_+^r \equiv (\bar{u}-t)_+^r \ ;$$

for the first derivative,

$$D_t^{(1)}(\bar{u}-t)_+^r \equiv D_t(\bar{u}-t)_+^r$$

$$\equiv \lim_{\substack{\epsilon \to 0 \\ \epsilon > 0}} \frac{(\bar{u}-(t-\epsilon))_+^r - (\bar{u}-t)_+^r}{-\epsilon} \quad ;$$

and recursively,

$$D_t^{(l)}(\bar{u}-t)_+^r \equiv D_t[D_t^{l-1}(\bar{u}-t)_+^r]$$

for all the succeeding derivatives, $l=2,3,\cdots$ (understood in the left-handed sense).

Results similar to those of (54) hold for $D_t^{(l)}(\bar{u}-t)_+^r$.

**Theorem:** For all $r,l \geq 0$,

$$D_t^l(\bar{u}-t)_+^r = \begin{cases} (-1)^l \dfrac{r!}{(r-l)!}(\bar{u}-t)_+^{r-l} & \text{for } l \leq r \\[2ex] 0 & \text{for } l > r \ . \end{cases}$$

Divided differences are close relatives of derivatives. In the previous chapter we took the function $(\bar{u}-t)_+^r$, evaluated it at some knot $t=\bar{u}_i$, evaluated it again at some other knot $t=\bar{u}_{i+1}$, and formed the combination

$$\frac{(\bar{u}-\bar{u}_{i+1})_+^r - (\bar{u}-\bar{u}_i)_+^r}{\bar{u}_{i+1}-\bar{u}_i} \ .$$

We can only write this if we assume that $\bar{u}_{i+1}$ and $\bar{u}_i$ are distinct. However, if we should choose to let $\bar{u}_i$ approach and join $\bar{u}_{i+1}$, the above combination would be consistent with the left-handed derivative. That is, letting $\epsilon = u_{i+1} - u_i > 0$).

$$\begin{aligned}
\lim_{\bar{u}_i \to \bar{u}_{i+1}} \frac{(\bar{u}-\bar{u}_{i+1})_+^r - (\bar{u}-\bar{u}_i)_+^r}{\bar{u}_{i+1}-\bar{u}_i} &= \lim_{\bar{u}_i \to \bar{u}_{i+1}} \frac{(\bar{u}-\bar{u}_i)_+^r - (\bar{u}-\bar{u}_{i+1})_+^r}{\bar{u}_i-\bar{u}_{i+1}} \\
&= \lim_{\substack{\epsilon \to 0 \\ \epsilon > 0}} \frac{(\bar{u}-(\bar{u}_{i+1}-\epsilon))_+^r - (\bar{u}-\bar{u}_{i+1})_+^r}{-\epsilon} \\
&= D_t(\bar{u}-t)_+^r \big|_{t=\bar{u}_{i+1}} \ .
\end{aligned}$$

It is this observation that will provide us with a definition for divided differences that includes the case of repeated values, namely, that distinct values are handled by differencing and dividing while repeated values are handled by differentiation.

## 8.2. Divided Differences in a General Setting

Let us work up to a definition for general divided differences gradually, reminding ourselves a little about calculus along the way. To forget about the specific form of the one-sided power functions for a moment, we will frame our discussion in terms of general functions, $f, g, h, \cdots$, of general variables, $x, y, z, \cdots$. We will come back to our specific functions $(\bar{u}-t)_+^r$ in a short while.

Consider any differentiable function. Recall that differentiation is an *operator* which provides a *mapping* of differentiable functions onto other functions; e.g.

$$D_x f(x,y) \equiv g(x,y)$$

and

$$D_y f(x,y) \equiv h(x,y) \ .$$

The "source" function, $f(x,y)$, and the "target" function, $g(x,y)$ or $h(x,y)$, have the same number of variables. In a like fashion we can regard the divided difference to be an operator which provides us with a mapping:

$$[z_1,z_2{:}x]f(x,y) = \frac{f(z_2,y)-f(z_1,y)}{z_2-z_1} \equiv G(z_1,z_2,y)$$

or

$$[z_1,z_2{:}y]f(x,y) = \frac{f(x,z_2)-f(x,z_1)}{z_2-z_1} \equiv H(x,z_1,z_2) \ .$$

These mappings convert the source function, $f$, of two variables into target functions, $G$ or $H$, of three variables. If the appropriate two of these variables are permitted to merge to a common value,

$$z_1 \rightarrow z \leftarrow z_2$$

then we obtain a function of two variables again. In fact,

$$[z,z:x]f(x,y) = \lim_{\substack{z_1\rightarrow z\\ z_2\rightarrow z}} G(z_1,z_2,y) = D_x f(x,y)|_{x=z} = g(z,y)$$

or

$$[z,z:y]f(x,y) = \lim_{\substack{z_1\rightarrow z\\ z_2\rightarrow z}} H(x,z_1,z_2) = D_y f(x,y)|_{y=z} = h(x,z) \ .$$

For higher differences this becomes slightly more complicated, so it will be worthwhile to economize on notation for the purpose of clarity. All except the one variable undergoing the differencing will now be suppressed; i.e. $f(x,y)\equiv f(x)$, if we are differencing with respect to $x$.

Consider

$$[z_1,z_2,z_3:x]f(x) = \frac{[z_2,z_3:x]f(x) - [z_1,z_2:x]f(x)}{z_3 - z_1}$$

$$= \frac{\dfrac{f(z_3)-f(z_2)}{z_3-z_2} - \dfrac{f(z_2)-f(z_1)}{z_2-z_1}}{z_3 - z_1} \ .$$

Using a Taylor series expansion we have

$$\frac{f(z_3)-f(z_2)}{z_3-z_2} = f'(z_2) + \frac{1}{2}(z_3-z_2)f''(z_2) + \mathrm{O}((z_3-z_2)^2)$$

where the expression $\mathrm{O}((z_3-z_2)^2)$ indicates that the remainder of the series will behave like a constant times $(z_3-z_2)^2$ ("will have the same order of behaviour as $(z_3-z_2)^2$") if the values of $z_3$ and $z_2$ approach each other. Similarly

$$-\frac{f(z_2)-f(z_1)}{z_2-z_1} = \frac{f(z_1)-f(z_2)}{z_2-z_1}$$

$$= -f'(z_2) - \frac{1}{2}(z_1-z_2)f''(z_2) + \mathrm{O}((z_1-z_2)^2) \ .$$

Consequently

$$\frac{\dfrac{f(z_3)-f(z_2)}{z_3-z_2} - \dfrac{f(z_2)-f(z_1)}{z_2-z_1}}{z_3 - z_1}$$

$$= \frac{1}{2(z_3-z_1)}[(z_3-z_2) - (z_1-z_2)]f''(z_2) + \frac{\mathrm{O}((z_3-z_2)^2) + \mathrm{O}((z_1-z_2)^2)}{(z_3-z_1)}$$

$$= \frac{1}{2}f''(z_2) + \frac{\mathrm{O}((z_3-z_2)^2) + \mathrm{O}((z_1-z_2)^2)}{(z_3-z_1)} \ .$$

If $z_1,z_2,z_3$ are allowed to approach a common value, $z$, in a reasonable way, then the trailing expression will go to zero. This motivates the following interpretation:

$$\begin{aligned}[z,z,z:x]f(x) &= \lim_{\substack{z_1\to z\\ z_2\to z\\ z_3\to z}}[z_1,z_2,z_3:x]f(x)\\ &= \frac{1}{2}D_x^{(2)}f(x)\big|_{x=z=z_1=z_2=z_3} \ .\end{aligned}$$

And in general, not surprisingly:

$$[z_1, \ldots, z_{1+l}:x]f(x) = \frac{1}{l!}D_x^{(l)}f(x)\big|_{x=z_1=\cdots=z_{1+l}}$$

when $z_1=\cdots=z_{1+l}$.

With these preliminaries, we will give a recursive definition of the *divided difference operator*. The definition begins by regarding the zero$^{\text{th}}$ divided difference as the operation that evaluates a function at a specified value of a variable. This corresponds roughly to the zero$^{\text{th}}$ differentiation operator; i.e.

$$D_x^{(0)}f(x)\big|_{x=z_1} = f(z_1) \ .$$

**Definition:** For any values $z_i \le \cdots \le z_{i+l}$ the $l$-th *divided difference* is given by

$$[z_i:x]f(x) = f(z_i)$$

for $l=0$, and by

$$[z_i, \ldots, z_{i+l}:x]f(x)$$

$$= \begin{cases} \dfrac{[z_{i+1}, \ldots, z_{i+l}:x]f(x) - [z_i, \ldots, z_{i+l-1}:x]f(x)}{z_{i+l}-z_i} & \text{if } z_{i+l} > z_i \\[2ex] \dfrac{1}{l!}D_x^{(l)}f(x)\big|_{x=z_i} & \text{if } z_{i+l} = z_i \end{cases}$$

for $l \ge 1$.

Observe that the notation

$$[z_i:x]\ ,\ [z_i,z_{i+1}:x]\ ,\ [z_i,z_{i+1},z_{i+2}:x]\ , \text{ etc.}$$

is reasonably suggestive for low-order divided differences. However, something like

$$[z_i, \ldots, z_{i+l-1}:x]$$

is often more confusing than helpful. Thus we will sometimes use the more compact form below.

**Notation:** $[z_i, \ldots, z_{i+r}:x] \equiv [z_i(r):x]$ .

The intent of this definition is to express, in shorthand, that $z_i$ followed by the next $r$ elements in the $\{z\}$ sequence define the $r^{\text{th}}$ order divided difference.

Two comments are worth making here, before we proceed with our development:

- Once we move back to the specific case of the one-sided power functions, we will only be interested in the divided differences of $(\bar{u}-t)_+^r$ with respect to $\bar{u}$ or with respect to $t$, rather than the divided differences of general functions. In this case, obviously, the differentiation in the above definition will either be right-handed, for differences with respect to $\bar{u}$, or left-handed, for differences with respect to $t$.
- The definition was motivated by the idea that divided differences could be equated with derivatives when any two or more of the values, $z_i, \ldots, z_{i+l}$, join together. The discussion about this was only motivational, but it can be made rigourous. It is proven in [Schumaker81] that, if $z_{i+r}(\epsilon)$, for $r=0,1,, \ldots, l$, is any sequence of points with $z_{i+r}(\epsilon) \rightarrow z_{i+r}$ as $\epsilon \rightarrow 0$, then it is true for any sufficiently smooth function, $p = p(t)$ (e.g. $(\bar{u}-t)_+^r$), that:

$$\lim_{\epsilon \to 0} [z_i(\epsilon), z_{i+1}(\epsilon), \ldots, z_{i+l}(\epsilon):t]p(t) = [z_i, z_{i+1}, \ldots, z_{i+l}:t]p(t) .$$

In particular, the divided difference over an arbitrary set of points $z_i, \ldots, z_{i+l}$ that contains repetitions is the limit of divided differences over distinct points. Up until now we have thought of the knots $\bar{u}_i$ in $\mathrm{S}(\mathrm{P}^k, \{\bar{u}_i\}_0^{m+k})$ as being fixed, and usually we have regarded them as distinct. This gives us the permission to regard knots as movable and, at times, confluent.

### 8.3. Algebraic and Analytic Properties

We have reminded ourselves that differentiation can be regarded as an operator that maps functions into functions, and we have taken this same view in the case of the divided difference. We close by establishing a few of the algebraic properties of these operators.

Let us begin by recalling that the differentiation operator is a *linear operator*; i.e.

$$D_x\{\alpha f(x)\} \equiv \alpha\{D_x f(x)\}$$

for any scalar $\alpha$, and

$$D_x\{f_1(x)+f_2(x)\} \equiv \{D_x f_1(x)\} + \{D_x f_2(x)\}$$

for any two functions $f_1(x)$ and $f_2(x)$. This holds for right-handed and left-handed differentiation as well as for ordinary, unrestricted differentiation. This means that, for any sum,

$$D_x\{\sum_j \alpha_j f_j(x)\} = \sum_j \alpha_j\{D_x f_j(x)\} .$$

It is just as easily seen that the simple divided difference operation (50) also behaves linearly. For example:

$$[z_1, z_2:x]\{\alpha f(x)\} = \frac{\alpha f(z_2) - \alpha f(z_1)}{z_2 - z_1}$$

$$= \alpha \frac{f(z_2) - f(z_1)}{z_2 - z_1}$$

$$= \alpha\{[z_1,z_2:x]f(x)\}$$

for any scalar $\alpha$. Similarly,

$$[z_1,z_2:x]\{f_1(x)+f_2(x)\} = \{[z_1,z_2:x]f_1(x)\} + \{[z_1,z_2:x]f_2(x)\}$$

for any two functions, $f_1(x)$ and $f_2(x)$. As a result of the definition above and some of our preceding observations, we have the following.

**Theorem:** The divided difference is a linear operator; i.e. for each fixed $l$, if

$$[z_i(l):x]f_j(x)$$

exists for each $j$ in some set of indices $\mathbf{J}$, then

$$[z_i(l):x]\{\sum_{j\in J}\alpha_j f_j(x)\} = \sum_{j\in J}\alpha_j\{[z_i(l):x]f_j(x)\}$$

for any scalars $\alpha_j$, $j\in\mathbf{J}$.

This provides us with a final observation. Notice that the order in which we perform differencing and differentiation can be swapped whenever these operations act on different variables.

**Corollary:**

$$D_x^{(r)}[z_i(l):y]\,f(x,y) = [z_i(l):y]\,D_x^{(r)}f(x,y)$$

# 9. General B-splines

Back in Chapter 7 we found that

$$[0,1,2,3,4:t](\bar{u}-t)_+^3$$

is exactly the uniform B-spline of Chapter 4. More generally, we found that

$$[\bar{u}_0,\bar{u}_1,\bar{u}_2,\bar{u}_3,\bar{u}_4:t](\bar{u}-t)_+^3$$

yielded an eventually-zero function. Of course, any scalar multiple of an eventually-zero function is also eventually-zero, and we shall see that if we use

$$(\bar{u}_4-\bar{u}_0)[\bar{u}_0,\bar{u}_1,\bar{u}_2,\bar{u}_3,\bar{u}_4:t](\bar{u}-t)_+^3 \;=\; (\bar{u}_4-\bar{u}_0)[\bar{u}_0(4):t](\bar{u}-t)_+^3$$

as our cubic B-spline then the curves we define will have the convex hull property. We will denote this function by

$$B_{0,4}(\bar{u}) \;.$$

The notation is intended to remind us of the following.

- $B_{0,4}(\bar{u})$ is a member of $\mathrm{S}(\mathrm{P}^4,\{\bar{u}_i\}_0^{m+k})$; the second subscript indicates its order.
- $B_{0,4}(\bar{u})$ is positive for $\bar{u}$ between $\bar{u}_0$ and $\bar{u}_4$; the two subscripts together indicate its support, i.e. the range of parameter values for which it is nonzero.

For B-splines in general, i.e. with unequally-spaced or multiple knots, these will continue to be important observations.

**Definition:** Assuming that $i \le m$, the *B-spline of order k associated with the knots* $\bar{u}_i, \ldots, \bar{u}_{i+k}$ is given by

$$B_{i,k}(\bar{u}) \;=\; (-1)^k(\bar{u}_{i+k}-\bar{u}_i)[\bar{u}_i(k):t](\bar{u}-t)_+^{k-1} \;.$$

It should be observed that $B_{i,k}(\bar{u})$ is vacuous if $\bar{u}_i = \bar{u}_{i+k}$.

In Chapter 7 we saw that odd divided differences of $(\bar{u}-t)_+^{k-1}$ were negative. The term $(-1)^k$, which is $-1$ for odd values of $k$ and $+1$ for even values of $k$, is therefore introduced so that $B_{i,k}(\bar{u})$ will be positive for all $k$.

Our earlier observations about the cubic B-splines generalize as follows.

- $B_{i,k}(\bar{u})$ is a member of $S(P^k, \{\bar{u}_i\}_0^{m+k})$; it is composed of segment polynomials having order $k$.
- Knot multiplicities greater than $k$ are of no interest in constructing splines, and in their absence we have $\bar{u}_{i+k} > \bar{u}_i$ for every $i$. As we shall see later, it follows that $B_{i,k}(\bar{u}) > 0$ for $\bar{u}$ between $\bar{u}_i$ and $\bar{u}_{i+k}$.
- $B_{i,k}(\bar{u}) = 0$, for $\bar{u} < \bar{u}_i$ and $\bar{u}_{i+k} \leq \bar{u}$. Thus $i,k$ indicates the interval of support. (The value of $B_{i,k}(\bar{u})$ at $\bar{u} = \bar{u}_i$ will depend upon the multiplicity of $\bar{u}_i$ and upon the value of $k$. Notice particularly that $\bar{u}_i$ is *included* and $\bar{u}_{i+k}$ is *excluded*. This is a result of the fact that $B_{i,k}(\bar{u})$ is constructed from the functions $(\bar{u}-t)_+^r$ and is open on the right.)

### 9.1. A Simple Example — Step Function B-splines

Let us recall how we arrived at this definition by examining the B-spline representation of piecewise linear functions. First consider

$$\begin{aligned} B_{i,1}(\bar{u}) &= (-1)(\bar{u}_{i+1}-\bar{u}_i)[\bar{u}_i(1):t](\bar{u}-t)_+^0 \\ &= (-1)(\bar{u}_{i+1}-\bar{u}_i)[\bar{u}_i,\bar{u}_{i+1}:t](\bar{u}-t)_+^0 \ . \end{aligned}$$

Since $k=1$, only multiplicities of 1 are interesting, and we must have $\bar{u}_i < \bar{u}_{i+1}$. The definition of a divided difference tells us that

$$\begin{aligned} B_{i,1}(\bar{u}) &= (-1)(\bar{u}_{i+1}-\bar{u}_i)\frac{(\bar{u}-\bar{u}_{i+1})_+^0 - (\bar{u}-\bar{u}_i)_+^0}{(\bar{u}_{i+1}-\bar{u}_i)} \\ &= (\bar{u}-\bar{u}_i)_+^0 - (\bar{u}-\bar{u}_{i+1})_+^0 \\ &= \begin{cases} 0 & \text{for } \bar{u} < \bar{u}_i \\ 1 & \text{for } \bar{u}_i \leq \bar{u} < \bar{u}_{i+1} \\ 0 & \text{for } \bar{u}_{i+1} \leq \bar{u} \ . \end{cases} \end{aligned}$$

Figure 76 illustrates the differencing process from which this function is constructed as a means of displaying, in this very simple context, the idea of combining two adjacent one-sided functions to obtain a function that dies away to zero.

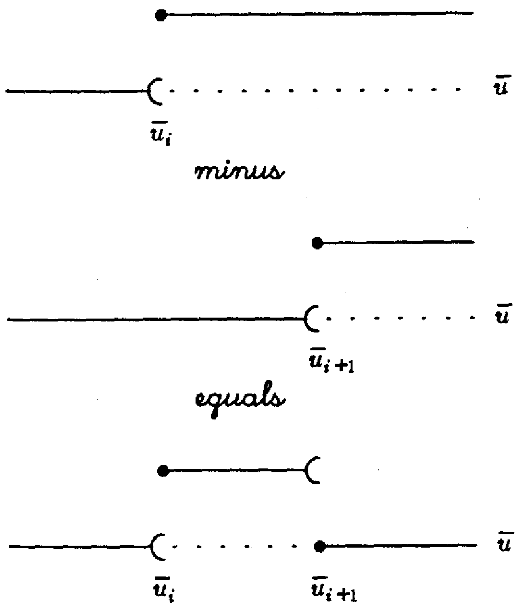


Figure 76. The construction of $B_{i,1}(\bar{u})$ (bottom) from $(\bar{u}-\bar{u}_i)_+^0$ (top) and $(\bar{u}-\bar{u}_{i+1})_+^0$ (middle).

This B-spline is as reasonable a function to associate with piecewise, first-order, $C^{-1}$ polynomials (i.e. step functions) as were the one-sided functions $(\bar{u}-\bar{u}_i)_+^0$ and $(\bar{u}-\bar{u}_{i+1})_+^0$. Let us consider the simple case in which the knots are

$$\begin{array}{cccc} \bar{u}_0 & \bar{u}_1 & \bar{u}_2 & \bar{u}_3 \\ 0 & 2 & 4 & 5 \end{array} ,$$

the legal parameter range is

$$\bar{u}_{k-1} \equiv \bar{u}_0 \leq \bar{u} < \bar{u}_3 \equiv \bar{u}_{m+1} ,$$

and the randomly-chosen step function to be represented (see Figure 77) is:

$$s_1(\bar{u}) = \begin{cases} 0.25 & \bar{u} < 0 \\ 0.50 & 0 \leq \bar{u} < 2 \\ 0.75 & 2 \leq \bar{u} < 4 \\ 1.25 & 4 \leq \bar{u} < 5 \\ 2.00 & 5 \leq \bar{u} . \end{cases}$$

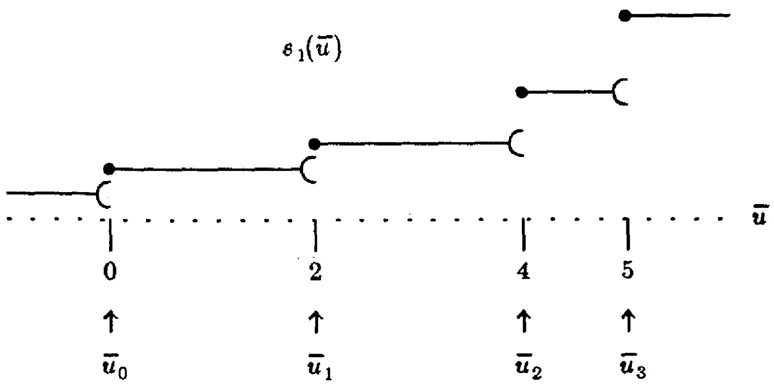


Figure 77. A piecewise constant function.

This step function is a member of the space

$$S(\mathbf{P}^1,\{0,2,4,5\}) \;,$$

for which the appropriate one-sided basis is

$$(\bar{u}-\bar{u}_0)_+^0 \equiv (\bar{u}-0)_+^0 \equiv \begin{cases} 0 & \bar{u} < 0 \\ 1 & \bar{u} \geq 0 \end{cases}$$

$$(\bar{u}-\bar{u}_1)_+^0 \equiv (\bar{u}-2)_+^0 \equiv \begin{cases} 0 & \bar{u} < 2 \\ 1 & \bar{u} \geq 2 \end{cases}$$

$$(\bar{u}-\bar{u}_2)_+^0 \equiv (\bar{u}-4)_+^0 \equiv \begin{cases} 0 & \bar{u} < 4 \\ 1 & \bar{u} \geq 4 \end{cases}$$

(see Figure 78). Note that there is no one-sided power function in the basis which could possibly account for the behavour of the step function outside of $[0,5)$. That region of the $\bar{u}$ axis, however, is outside of the parameter range associated with the spline space that we are considering.

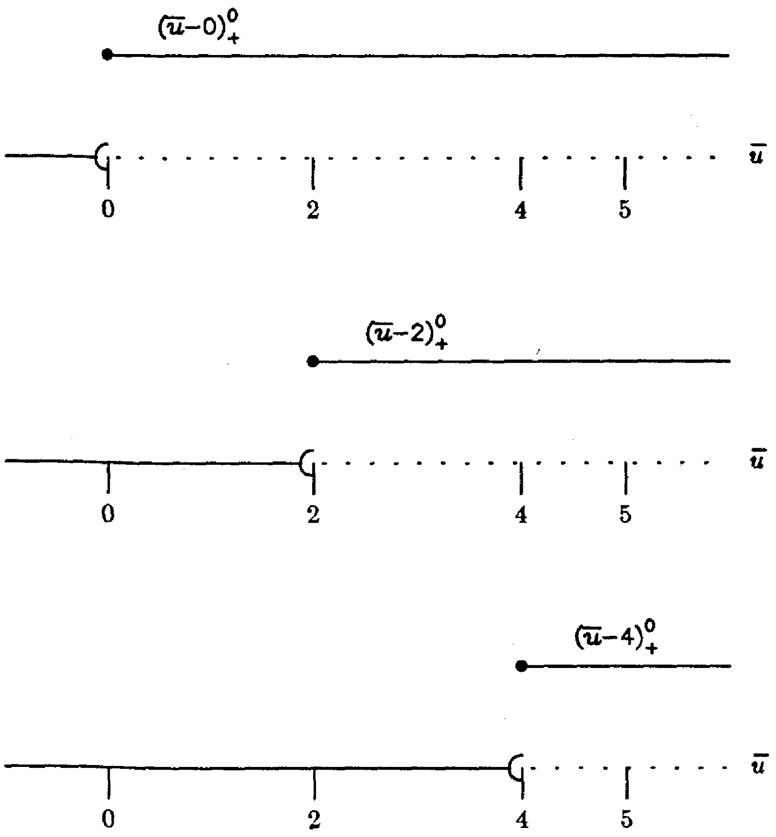


Figure 78. The one-sided basis for $S(P^1,\{0,2,4,5\})$

The one-sided power functions cannot reproduce $s_1(\bar{u})$ on the entire axis, as we have remarked before, but we can reproduce the behaviour of $s_1(\bar{u})$ restricted to the parameter range [0,5) by a function $s_2(\bar{u})$:

$$s_2(\bar{u}) = (0.50)(\bar{u}-0)^0_+ + (0.25)(\bar{u}-2)^0_+ + (0.50)(\bar{u}-4)^0_+ \qquad (55)$$

(see Figure 79).

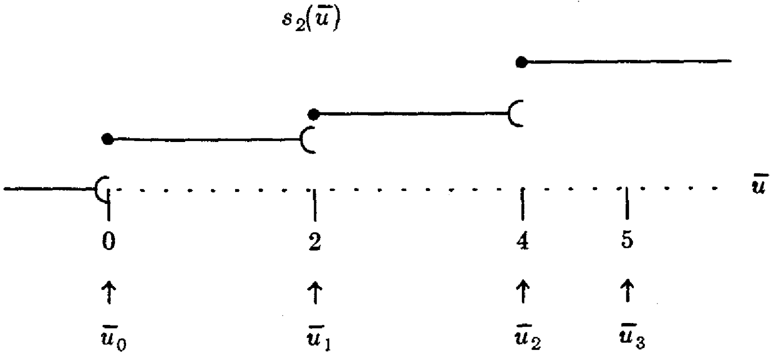


Figure 79. The one-sided basis representation of the function $s_1(\bar{u})$ in Figure 77.

Again, notice that the representation $s_2(\bar{u})$ differs from the given step function $s_1(\bar{u})$ outside

$[\bar{u}_0,\bar{u}_3) = [0,5)$, a fact which is not at all disturbing: recall that members of $\mathbf{S}(\mathbf{P}^1,\{0,2,4,5\})$ are indistinguishable if they only differ outside the legal parameter range.

The obvious divided-differencing process to be considered uses

$$(\bar{u}-0)_+^0 \text{ and } (\bar{u}-2)_+^0$$

to construct the B-spline

$$B_{0,1}(\bar{u}) = (-1)(2-0)[0,2{:}t](\bar{u}-t)_+^0 \equiv \begin{cases} 0 & \bar{u} < 0 \\ 1 & 0 \leq \bar{u} < 2 \\ 0 & 2 \leq \bar{u} \end{cases}$$

which is substituted for $(\bar{u}-0)_+^0$, and uses

$$(\bar{u}-2)_+^0 \text{ and } (\bar{u}-4)_+^0$$

to construct the B-spline

$$B_{1,1}(\bar{u}) = (-1)(4-2)[2,4{:}t](\bar{u}-t)_+^0 \equiv \begin{cases} 0 & \bar{u} < 2 \\ 1 & 2 \leq \bar{u} < 4 \\ 0 & 4 \leq \bar{u} \end{cases}$$

which is substituted for $(\bar{u}-2)_+^0$. Finally, the one-sided power function

$$(\bar{u}-5)_+^0 \ ,$$

which was not needed for the one-sided-basis representation, can be differenced with

$$(\bar{u}-4)_+^0 \ ,$$

to produce the final B-spline:

$$(\bar{u}-\bar{u}_3)_+^0 \equiv (\bar{u}-5)_+^0 = \begin{cases} 0 & \bar{u} < 5 \\ 1 & 5 \leq \bar{u} \end{cases}$$

$$B_{2,1}(\bar{u}) = (-1)(5-4)[4,5{:}t](\bar{u}-t)_+^0 = \begin{cases} 0 & \bar{u} < 4 \\ 1 & 4 \leq \bar{u} < 5 \\ 0 & 5 \leq \bar{u} \ , \end{cases}$$

which substitutes for $(\bar{u}-4)_+^0$. Figure 80 below depicts this basis. Note, as for the one-sided basis, that this (B-spline) basis can not account for the behaviour of the step function to the left of the leftmost knot and to the right of the rightmost knot.

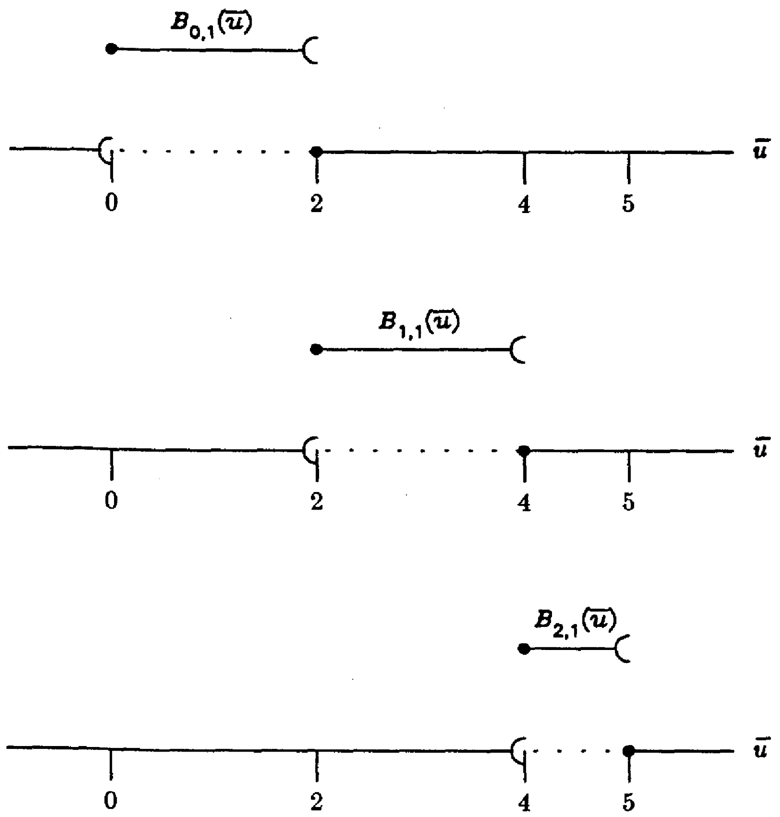


Figure 80. The B-spline basis for $S(P^1,\{0,2,4,5\})$

This allows us to represent $s_1(\bar u)$, restricted to [0,5), by the functions $s_3(\bar u)$:

$$s_3(\bar u) = (0.50)B_{0,1}(\bar u) + (0.75)B_{1,1}(\bar u) + (1.25)B_{2,1}(\bar u)$$

(see Figure 81).

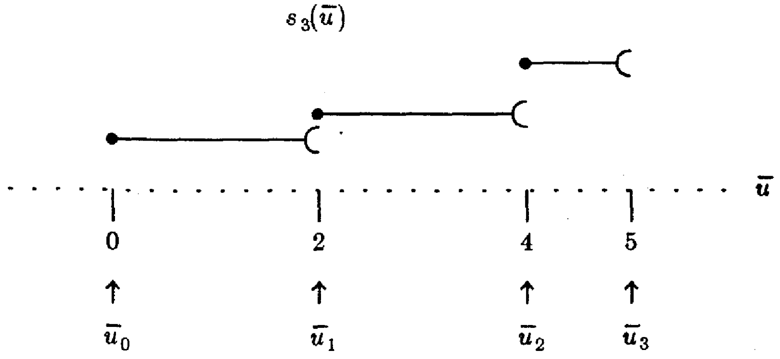


Figure 81. The B-spline representation of the function in Figure 77.

Compare this B-spline representation $s_3(\bar u)$ (shown in Figure 81) with the the original step function $s_1(\bar u)$ (given by (55) and shown in Figure 77), and with its one-sided representation $s_2(\bar u)$ (shown in Figure 79). All three are identical on the parameter range [0,5); consequently they are, by convention, the same spline with respect to the space $S(P^1,\{0,2,4,5\})$. In practice, too, they are indistinguishable, since they should never be rendered for values of the parameter $\bar{u}$ outside $[0,5)$, in the same sense as uniform cubic splines were not to be rendered outside the interval $[\bar{u}_3,\bar{u}_{m+1})$.

### 9.2. Linear B-splines

The reader should be cautious about following us too far without objection. The above example was chosen to have unequally-spaced knots, so it is a bit more general than the uniform case. But it is possible that any inferences we might draw from this example break down in the presence of multiple knots. Let us introduce another knot at $\bar{u}=4$ to see what can be learned. To do this, we will go to order $k=2$, since knots of multiplicity 2 are uninteresting for $k=1$. Furthermore, so that this multiplicity falls strictly within the parameter range and contributes to the "significant" portion of any spline, we will add the knot

$$\bar{u}_5 = 7$$

so that the knot sequence is

| $\bar{u}_0$ | $\bar{u}_1$ | $\bar{u}_2$ | $\bar{u}_3$ | $\bar{u}_4$ | $\bar{u}_5$ . |
|---|---|---|---|---|---|
| 0 | 2 | 4 | 4 | 5 | 7 |

Consider $S(P^2,\{0,2,4,4,5,7\})$. The elements of this space are:

| | |
|---|---|
| linear | for $\bar{u} < 0$ |
| $C^0$ | at $\bar{u} = 0$ |
| linear | for $0 \le \bar{u} < 2$ |
| $C^0$ | at $\bar{u} = 2$ |
| linear | for $2 \le \bar{u} < 4$ |
| $C^{-1}$ | at $\bar{u} = 4$ |
| linear | for $4 \le \bar{u} < 5$ |
| $C^0$ | at $\bar{u} = 5$ |
| linear | for $5 \le \bar{u} < 7$ |
| $C^0$ | at $\bar{u} = 7$ |
| linear | for $7 \le \bar{u}$ . |

A representative spline from this space is shown in Figure 82.

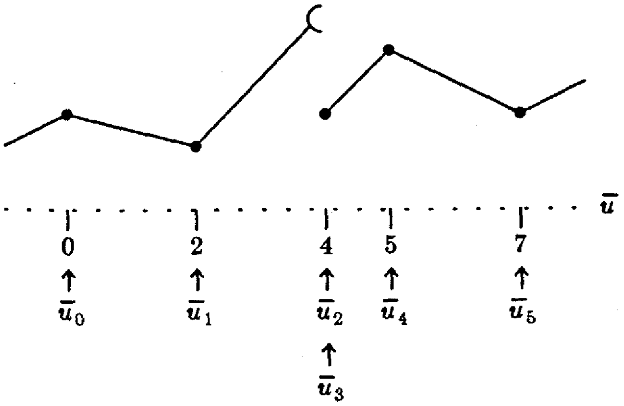


Figure 82. A piecewise linear function.

The legal parameter range of this space is

$$\bar{u}_{k-1} \equiv \bar{u}_1 \leq \bar{u} < \bar{u}_4 \equiv \bar{u}_{m+1} \ .$$

The legal parameter range begins with $\bar{u}_1$ because we need two (linearly independent) functions to define the first segment; the knot $\bar{u}_7$ is added to the right of the legal parameter range so that there will be a one-sided function to difference with $(\bar{u}-\bar{u}_4)_+^1$.

The one-sided basis for this space is

$$(\bar{u}-0)_+^1 \, , \ (\bar{u}-2)_+^1 \, , \ (\bar{u}-4)_+^0 \, , \ (\bar{u}-4)_+^1 \ .$$

The one-sided representation for the spline shown in Figure 82 is shown in Figure 83 below.

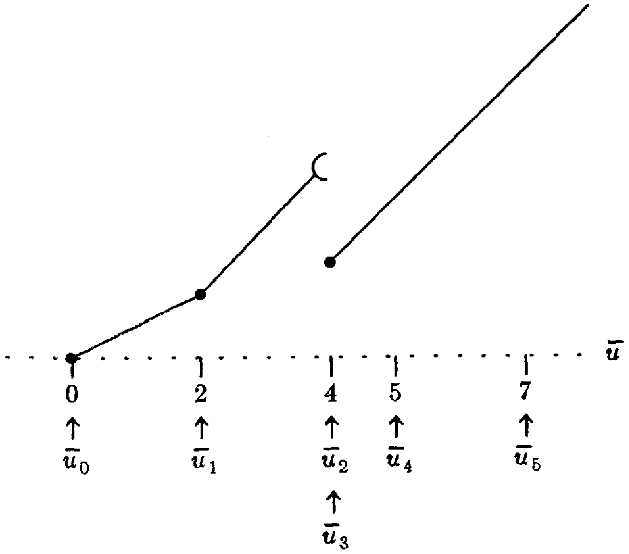


Figure 83.

We use

$$(\bar{u}-0)^1_+,\ (\bar{u}-2)^1_+,\ (\bar{u}-4)^1_+\ ,$$

together with the extra one-sided power functions

$$(\bar{u}-5)^1_+ \text{ and } (\bar{u}-7)^1_+\ ,$$

to construct the B-splines

$$B_{0,2}(\bar{u}) = (-1)^2(\bar{u}_2-\bar{u}_0)[\bar{u}_0,\bar{u}_1,\bar{u}_2{:}t](\bar{u}-t)^1_+$$

$$= (\bar{u}_2-\bar{u}_0)\left(\frac{[\bar{u}_1,\bar{u}_2{:}t](\bar{u}-t)^1_+ - [\bar{u}_0,\bar{u}_1{:}t](\bar{u}-t)^1_+}{\bar{u}_2-\bar{u}_0}\right)$$

$$= (\bar{u}_2-\bar{u}_0)\left(\frac{\dfrac{[\bar{u}_2{:}t](\bar{u}-t)^1_+ - [\bar{u}_1{:}t](\bar{u}-t)^1_+}{\bar{u}_2-\bar{u}_1} - \dfrac{[\bar{u}_1{:}t](\bar{u}-t)^0_+ - [\bar{u}_0{:}t](\bar{u}-t)^1_+}{\bar{u}_1-\bar{u}_0}}{\bar{u}_2-\bar{u}_0}\right)$$

$$= (\bar{u}_2-\bar{u}_0)\left(\frac{\dfrac{(\bar{u}-\bar{u}_2)^1_+ - (\bar{u}-\bar{u}_1)^1_+}{\bar{u}_2-\bar{u}_1} - \dfrac{(\bar{u}-\bar{u}_1)^1_+ - (\bar{u}-\bar{u}_0)^1_+}{\bar{u}_1-\bar{u}_0}}{\bar{u}_2-\bar{u}_0}\right)$$

$$= (4-0)\left(\frac{\dfrac{(\bar{u}-4)^1_+ - (\bar{u}-2)^1_+}{4-2} - \dfrac{(\bar{u}-2)^1_+ - (\bar{u}-0)^1_+}{2-0}}{4-0}\right)$$

$$\equiv \begin{cases} 0 & \bar{u} < 0 \\ \dfrac{\bar{u}-0}{2-0} & 0 \le \bar{u} < 2 \\ \dfrac{2-\bar{u}}{4-2}+1 & 2 \le \bar{u} < 4 \\ 0 & 4 \le \bar{u} \end{cases}$$

$$B_{1,2}(\bar{u}) = (-1)^2(\bar{u}_3-\bar{u}_1)[\bar{u}_1,\bar{u}_2,\bar{u}_3{:}t](\bar{u}-t)^1_+$$

$$= (\bar{u}_3-\bar{u}_1)\left(\frac{[\bar{u}_2,\bar{u}_3{:}t](\bar{u}-t)^1_+ - [\bar{u}_1,\bar{u}_2{:}t](\bar{u}-t)^1_+}{\bar{u}_3-\bar{u}_1}\right)$$

$$= (\bar{u}_3-\bar{u}_1)\left(\frac{D_t^{(1)}(\bar{u}-t)_+^1\big|_{t=\bar{u}_2=\bar{u}_3} - \dfrac{[\bar{u}_2{:}t](\bar{u}-t)_+^1 - [\bar{u}_1{:}t](\bar{u}-t)_+^1}{\bar{u}_2-\bar{u}_1}}{\bar{u}_3-\bar{u}_1}\right)$$

$$= (\bar{u}_3-\bar{u}_1)\left(\frac{-(\bar{u}-\bar{u}_2)_+^0 - \dfrac{(\bar{u}-\bar{u}_2)_+^1 - (\bar{u}-\bar{u}_1)_+^1}{\bar{u}_2-\bar{u}_1}}{\bar{u}_3-\bar{u}_1}\right)$$

$$= (4-2)\left(\frac{-(\bar{u}-4)_+^0 - \dfrac{(\bar{u}-4)_+^1 - (\bar{u}-2)_+^1}{4-2}}{4-2}\right)$$

$$\equiv \begin{cases} 0 & \bar{u} < 2 \\ \dfrac{\bar{u}-2}{4-2} & 2 \le \bar{u} < 4 \\ 0 & 4 \le \bar{u} \end{cases}$$

$$B_{2,2}(\bar{u}) = (-1)^2(\bar{u}_4-\bar{u}_2)[\bar{u}_2,\bar{u}_3,\bar{u}_4{:}t](\bar{u}-t)_+^1$$

$$\equiv \begin{cases} 0 & \bar{u} < 4 \\ \dfrac{4-\bar{u}}{5-4}+1 & 4 \le \bar{u} < 5 \\ 0 & 5 \le \bar{u} \end{cases}$$

$$B_{3,2}(\bar{u}) = (-1)^2(\bar{u}_5-\bar{u}_3)[\bar{u}_3,\bar{u}_4,\bar{u}_5{:}t](\bar{u}-t)_+^1$$

$$\equiv \begin{cases} 0 & \bar{u} < 4 \\ \dfrac{\bar{u}-4}{5-4} & 4 \le \bar{u} < 5 \\ \dfrac{5-\bar{u}}{7-5}+1 & 5 \le \bar{u} < 7 \\ 0 & 7 \le \bar{u} \; . \end{cases}$$

These B-splines are plotted in Figure 84. Pay particular attention to the B-splines $B_{1,2}(\bar{u})$ and $B_{2,2}(\bar{u})$, which contain a discontinuity at the breakpoint $\bar{u}=4$.

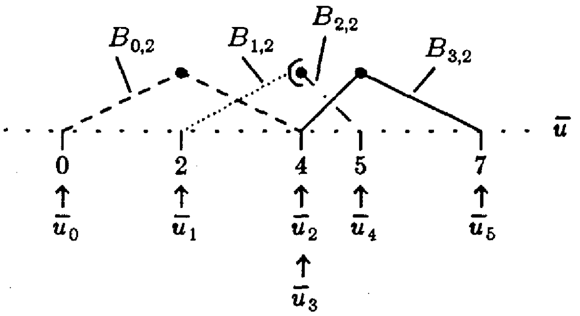


Figure 84. The B-splines of order 2 with which we can represent the piecewise linear curve shown in Figure 82.

Multiplying these B-splines by the appropriate scale factors results in a curve indistinguishable from that of Figure 82 on the legal parameter range [2,5).

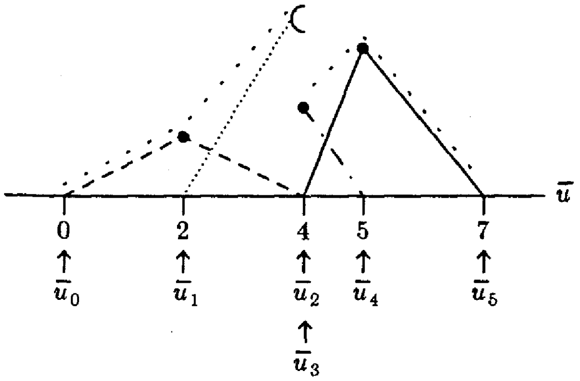


Figure 85. The piecewise linear curve of Figure 82 (shown with a lightly dotted line) represented as a linear combination of 2nd-order B-splines. Notice how the double knot at $\bar{u}=4$ allows us to represent a discontinuity in the curve at this point.

Any spline of the sort shown in Figure 82 can be represented by this collection of B-splines on the parameter range $[\bar{u}_1,\bar{u}_4) = [2,5)$.

## 9.3. General B-spline Bases

The previous example suggests how a pure B-spline basis can be constructed for any spline space $\mathbf{S}(\mathbf{P}^k,\{\bar{u}_i\}_0^{m+k})$.

**Construction:** For any given knot sequence $\{\bar{u}_i\}_0^{m+k}$, let

$$B_{i,k}(\bar{u}) = (-1)^k(\bar{u}_{i+k}-\bar{u}_i)[\bar{u}_i(k):t](\bar{u}-t)_+^{k-1}$$

for $i=0,\ldots,m$.

**Theorem:** The $B_{0,k}(\bar{u}), \ldots, B_{m,k}(\bar{u})$ constructed in this fashion are a basis for $\mathbf{S}(\mathbf{P}^k, \{\bar{u}_i\}_0^{m+k})$.

This is a result due to Curry and Schoenberg, and a proof may be found in [deBoor78]. It implies that the dimension of $\mathbf{S}(\mathbf{P}^k, \{\bar{u}_i\}_0^{m+k})$ is $m+1$, provided no knot has multiplicity greater than $k$, which would result in some of the B-splines being vacuous. (Recall that the legal parameter range is $[\bar{u}_k, \bar{u}_{m+1})$.)

## 9.4. Examples — Quadratic B-splines

The divided-difference formulation could be used directly to evaluate the B-splines (though we will discourage this from the standpoint of numerical accuracy in some subsequent remarks). We will illustrate this process of evaluation with a couple of examples. First, consider $B_{0,3}(\bar{u})$ with $\bar{u}=2$ on the knots

| $\bar{u}_0$ | $\bar{u}_1$ | $\bar{u}_2$ | $\bar{u}_3$ , |
|---|---|---|---|
| 0 | 1 | 3 | 4 |

shown in Figure 86.

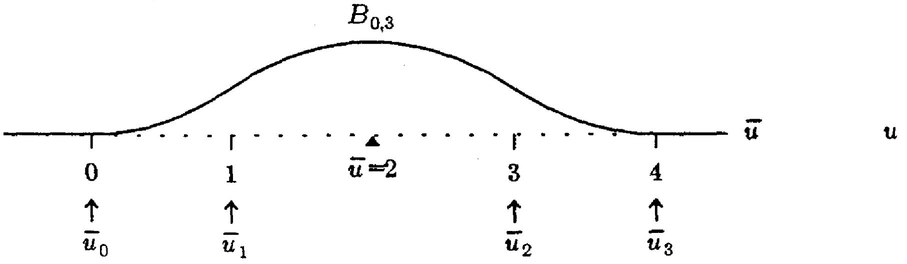


Figure 86. A quadratic B-spline with single knots.

In the table below, beginning at the second column, each entry is the divided difference of the entry to its left in the preceding column and the entry just above that. That is, for each pattern

A
B C

we have

$$C = \frac{B - A}{\bar{u}_{i_1} - \bar{u}_{i_2}}$$

for some appropriate knots $\bar{u}_{i_1}$ and $\bar{u}_{i_2}$.

| $\bar{u}_i$ | $(\bar{u}-\bar{u}_i)^2_+$ | $[*,*]$ | $[*,*,*]$ | $[*,*,*,*]$ |
|---|---|---|---|---|
| 0 | $(2-0)^2_+=4$ | | | |
| 1 | $(2-1)^2_+=1$ | $\frac{1-4}{1-0}=-3$ | | |
| 3 | $(2-3)^2_+=0$ | $\frac{0-1}{3-1}=-\frac{1}{2}$ | $\frac{-\frac{1}{2}+3}{3-0}=\frac{5}{6}$ | |
| 4 | $(2-4)^2_+=0$ | $\frac{0-0}{4-3}=0$ | $\frac{0+\frac{1}{2}}{4-1}=\frac{1}{6}$ | $\frac{\frac{1}{6}-\frac{5}{6}}{4-0}=-\frac{1}{6}$ |

The above differencing process establishes that

$$\begin{aligned} B_{0,3}(2) &= (-1)^3(4-0)(-\frac{1}{6}) \\ &= \frac{2}{3} \end{aligned}$$

for the knot sequence

$$\{\bar{u}_i\}_0^{m+k} = \{0,1,3,4\} \ .$$

A more involved example is given by the computation of $B_{0,3}(2)$ if the knot sequence is

$$\begin{matrix} \bar{u}_0 & \bar{u}_1 & \bar{u}_2 & \bar{u}_3 \ , \\ 0 & 1 & 1 & 3 \end{matrix}$$

shown in Figure 87.

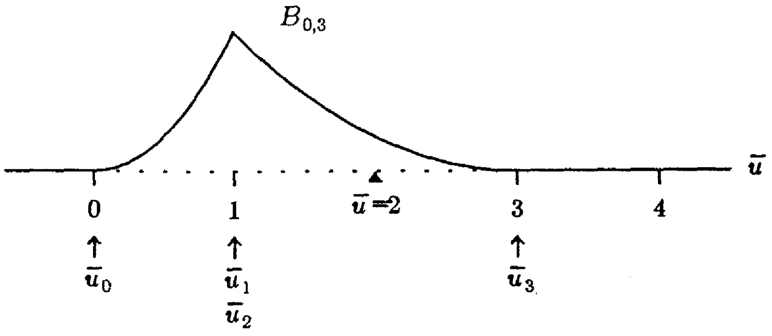


Figure 87. A quadratic B-spline with a double knot.

The table for the divided-difference computation is given below. Note that the repeated knot requires the computation of a derivative.

| $\bar{u}_i$ | $(\bar{u}-\bar{u}_i)^2_+$ | $[*,*]$ | $[*,*,*]$ | $[*,*,*,*]$ |
|---|---|---|---|---|
| 0 | $(2-0)^2_+=4$ | | | |
| 1 | $(2-1)^2_+=1$ | $\frac{1-4}{1-0}=-3$ | | |
| 1 | $(2-1)^2_+=1$ | $D_t(2-t)^2\vert_{t=1}=-2$ | $\frac{-2+3}{1-0}=1$ | |
| 3 | $(2-3)^2_+=0$ | $\frac{0-1}{3-1}=-\frac{1}{2}$ | $\frac{-\frac{1}{2}+2}{3-1}=\frac{3}{4}$ | $\frac{\frac{3}{4}-1}{3-0}=-\frac{1}{12}$ |

This table establishes that

$$B_{0,3}(2) = (-1)^3(3-0)(-\frac{1}{12}) = \frac{1}{4} .$$

This example raises another issue we should consider. Our preliminary discussions introduced the operation of divided differencing in the context of simple knots. We motivated the use of divided differences informally on the grounds that, when

$$\bar{u}_i < \cdots < \bar{u}_{i+k} ,$$

this operation was precisely what was needed to form linear combinations of

$$(\bar{u}-\bar{u}_i)^{k-1}_+ , \ldots , (\bar{u}-\bar{u}_{i+k})^{k-1}_+$$

having compact support. That is, the differencing process is a means of finding coefficients $d_i$, ..., $d_{i+k}$ for which

$$B_{i,k}(\bar{u}) = d_{i+k}(\bar{u}-\bar{u}_{i+k})^{k-1}_+ + \cdots + d_i(\bar{u}-\bar{u}_i)^{k-1}_+$$

had the property that

$$B_{i,k}(\bar{u}) = 0 \text{ for } \bar{u} \geq \bar{u}_{i+k} .$$

This informal explanation does not apply when knots become multiple, yet we proceeded to define the divided difference operator in general, including the case of multiple knots, and then claim, or at least imply, that the functions

$$B_{i,k}(\bar{u}) = (-1)^k(\bar{u}_{i+k}-\bar{u}_i)[\bar{u}_i(k):t](\bar{u}-t)^{k-1}_+$$

would have this property of compact support. We will give empirical evidence below that this is true, and we will prove it formally in a later section. Let us look at the example of the knots $\{0,1,1,3\}$ to see what happens.

The one-sided power functions that are appropriate for this knot sequence are

$$(\bar{u}-0)^2_+ , \ (\bar{u}-1)^2_+ , \ (\bar{u}-1)^1_+ , \text{ and } (\bar{u}-3)^2_+ . \tag{56}$$

Note that there are two powers of $(\bar{u}-1)$ associated with the double knot $\bar{u}=1$, a first power and a square. Recall that $(\bar{u}-1)^2_+$ allows us to alter the third derivative as we cross a knot; in the same way, the truncated power function $(\bar{u}-1)^1_+$ allows us to alter the second derivative as we cross a knot. To develop any spline basis suitable for this knot sequence we must restrict our attention to linear combinations of these functions.

The divided difference table above, for general $\bar{u}$, would be

| $\bar{u}_i$ | $(\bar{u}-\bar{u}_i)^2_+$ | $[*,*]$ | $[*,*,*]$ | $[*,*,*,*]$ |
|---|---|---|---|---|
| 0 | $(\bar{u}-0)^2_+$ | | | |
| 1 | $(\bar{u}-1)^2_+$ | $\dfrac{(\bar{u}-1)^2_+-(\bar{u}-0)^2_+}{1-0}$ | | |
| 1 | $(\bar{u}-1)^2_+$ | $D_t(\bar{u}-t)^2\|_{t=1}=-2(\bar{u}-1)^1_+$ | A | |
| 3 | $(\bar{u}-3)^2_+$ | $\dfrac{(\bar{u}-3)^2_+-(\bar{u}-1)^2_+}{3-1}$ | B | C |

where

$$A = \frac{-2(\bar{u}-1)^1_+ - \dfrac{(\bar{u}-1)^2_+ - (\bar{u}-0)^2_+}{1-0}}{1-0}$$

$$B = \frac{\dfrac{(\bar{u}-3)^2_+ - (\bar{u}-1)^2_+}{3-1} + 2(\bar{u}-1)^1_+}{3-1}$$

and

$$C = \frac{A-B}{3-0} \quad .$$

Clearly, we are still producing linear combinations of the one-sided power functions (56). Indeed, the derivative that arises when a multiple knot is encountered reduces the order of the one-sided function just enough to introduce the appropriate lower-degree continuity.

What is more, the compact support property (locality) also holds. If the expression for C above is written out, and then simplified symbolically under the assumption that $\bar{u}>3$ so that the "+" subscripts are no longer relevant, then all terms in the numerator cancel to zero.

We end this section with a cautionary remark. Observe that both computational tables above, for $B_{0,3}(2)$ with single knots and for a double knot, involve arithmetic with a mixture of positive and negative numbers. This implies that cancellations can take place in floating-point to produce inaccurate results. These inaccuracies will be pronounced in cases where knots are nearly but notexactly multiple. The divided-difference definition of a B-spline is not the recommended formula to use in numeric computations. We will establish more viable methods of computation in a later chapter.

### 9.5. The Visual Effect of Knot Multiplicities — Cubic B-splines

We will end this section with a few more examples of B-splines, chosen now from the more useful cubic case. (Most of the material in this and the following subsection is taken from [Barsky83].) Let us begin with the knots we have been using for the uniform B-spline discussions:

$$\begin{array}{ccccc} \bar{u}_0 & \bar{u}_1 & \bar{u}_2 & \bar{u}_3 & \bar{u}_4 \\ 0 & 1 & 2 & 3 & 4 \end{array} .$$

We will progressively increase the multiplicity of the knot at $\bar{u}=1$ to watch what happens. As a reminder, we give the description of the uniform B-spline on these knots once again:

$$\begin{aligned} B_{0,4}(\bar{u}) &= (-1)^4(\bar{u}_4-\bar{u}_0)[\bar{u}_0(4):t](\bar{u}-t)_+^3 \\ &= (-1)^4(\bar{u}_4-\bar{u}_0)[\bar{u}_0,\bar{u}_1,\bar{u}_2,\bar{u}_3,\bar{u}_4:t](\bar{u}-t)_+^3 \\ &= (4-0)[0,1,2,3,4:t](\bar{u}-t)_+^3 \\ &= \begin{cases} b_{-0}(\bar{u}) & 0 \le \bar{u} < 1 \\ b_{-1}(\bar{u}) & 1 \le \bar{u} < 2 \\ b_{-2}(\bar{u}) & 2 \le \bar{u} < 3 \\ b_{-3}(\bar{u}) & 3 \le \bar{u} < 4 \end{cases} , \end{aligned}$$

where the segment polynomials are given by

$$\begin{aligned} b_{-0}(\bar{u}) &= \frac{\bar{u}^3}{6} \\ b_{-1}(\bar{u}) &= -\frac{3\bar{u}^3-12\bar{u}^2+12\bar{u}-4}{6} \\ b_{-2}(\bar{u}) &= \frac{3\bar{u}^3-24\bar{u}^2+60\bar{u}-44}{6} \\ b_{-3}(\bar{u}) &= -\frac{\bar{u}^3-12\bar{u}^2+48\bar{u}-64}{6} . \end{aligned}$$

Figure 88 shows the graph of $B_{0,4}(\bar{u})$.

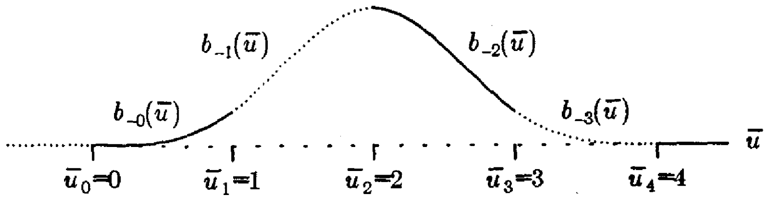


Figure 88. $B_{0,4}(\bar{u})$, a uniform cubic B-spline, with each of the segment polynomials comprising the basis function labeled and distinguished by the alternating use of dotted and solid lines.

If the knot at $\bar{u}=1$ is doubled, then $B_{0,4}(\bar{u})$ looks as follows.

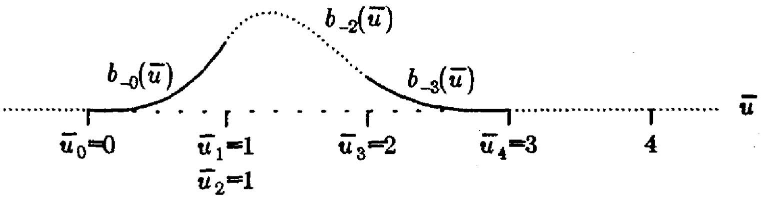


Figure 89. The double knot at $\bar{u}=1$ eliminates second derivative (curvature) continuity there, although first derivative (slope) continuity remains. Notice that the basis function is no longer symmetric.

Notice that the support of this function (the region on which it is different from zero) is still $(\bar{u}_0,\bar{u}_4)$, but this now represents the interval

$$0 < \bar{u} < 3$$

because the knots $\bar{u}_1$ and $\bar{u}_2$ have "moved together." The segment polynomials $b_{-0}(\bar{u})$ and $b_{-2}(\bar{u})$ have only their value and first derivative in common at $\bar{u}=1$; i.e. $B_{0,4}(\bar{u})$ has $C^1$ continuity at the breakpoint $\bar{u}_1$. The segment polynomials are given by

$$b_{-0}(\bar{u}) = \frac{\bar{u}^3}{2}$$

$b_{-1}(\bar{u})$ *is vacuous*

$$b_{-2}(\bar{u}) = \frac{5\bar{u}^3-27\bar{u}^2+45\bar{u}-21}{4}$$

and

$$b_{-3}(\bar{u}) = -\frac{\bar{u}^3-9\bar{u}^2+27\bar{u}-27}{4} .$$

Since there are now only three segments to the B-spline instead of four, we have had to choose a new numbering of the segment polynomials. Our choice reflects the idea that, since the interval between $\bar{u}_1$ and $\bar{u}_2$ has now disappeared, we should dispense with $b_{-1}(\bar{u})$. The first derivatives of the remaining segment polynomials are

$$b_{-0}^{(1)}(\bar{u}) = \frac{3\bar{u}^2}{2}$$

$$b_{-2}^{(1)}(\bar{u}) = \frac{15\bar{u}^2-54\bar{u}+45}{4}$$

and

$$b_{-3}^{(1)}(\bar{u}) = -\frac{3\bar{u}^2-18\bar{u}+27}{4} .$$

Notice that $b_{-0}^{(1)}(1)=b_{-2}^{(1)}(1)=1.5$ and that $b_{-2}^{(1)}(2)=b_{-3}^{(1)}(2)=-0.75$, thus establishing first derivative continuity at $\bar{u}=1$ and $\bar{u}=2$. The second derivatives are

$$b_{-0}^{(2)}(\bar{u}) = 3\bar{u}$$

$$b_{-2}^{(2)}(\bar{u}) = \frac{30\bar{u}-54}{4}$$

and

$$b^{(2)}_{-3}(\bar{u}) = -\frac{6\bar{u}-18}{4} .$$

We see that $b^{(2)}_{-0}(1)=3$ while $b^{(2)}_{-2}(1)=-6$, so that $B_{i,k}(\bar{u})$ has a discontinuous second derivative at $\bar{u}=\bar{u}_1$. Observe however, that $b^{(2)}_{-2}(2)=b^{(2)}_{-3}(2)=1.5$, so $B_{i,k}(\bar{u})$ does have a continuous second derivative at $\bar{u}=\bar{u}_3$. This follows, of course, from the fact that we have not increased the knot multiplicity at $\bar{u}=2$, and the discontinuity that we have introduced at the breakpoint $\bar{u}=\bar{u}_1=\bar{u}_2=1$ has no influence on the other breakpoints.

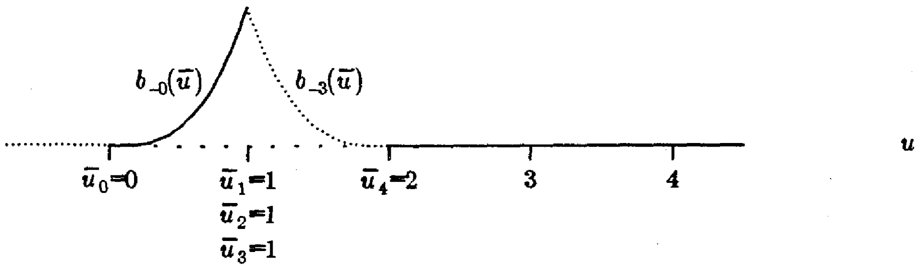


Figure 90. A knot of multiplicity 3, which reduces the cubic B-spline to positional continuity at $\bar{u}=1$.

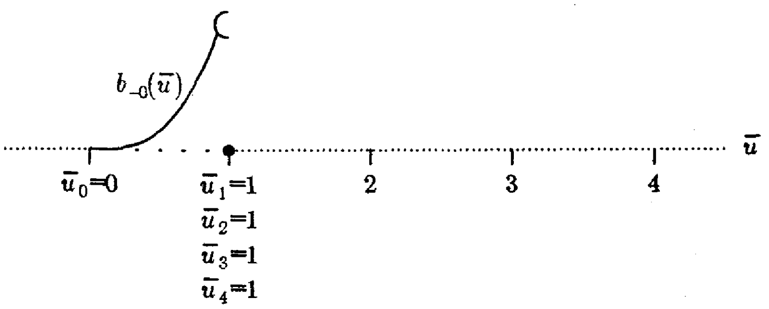


Figure 91. A knot of multiplicity 4, which eliminates even positional continuity at $\bar{u}=1$.

Figure 90 contains a knot of multiplicity three ($C^0$, or positional continuity) and Figure 91 a knot of multiplicity four (no continuity). Notice that in each case the basis function, which is cubic, is nonzero over the span of four knots, namely for $u\in[\bar{u}_0,\bar{u}_4)$. The two cubics that meet at the triple knot in Figure 90 meet only with $C^0$ (that is, $C^{4-1-\mu}$ continuity). Each additional time a knot is repeated, the parametric continuity of the underlying basis functions, and hence the parametric continuity of any curve they might construct, is reduced by one order. The segment polynomials for the triple-knot case are:

$$b_{-0}(\bar{u}) = \bar{u}^3$$

$b_{-1}(\bar{u})$ *is vacuous*

$b_{-2}(\bar{u})$ *is vacuous*

$$b_{-3}(\bar{u}) = -\bar{u}^3+6\bar{u}^2-12\bar{u}+8 .$$

Finally, the segment polynomials for the quadruple-knot case are:

$b_{-0}(\bar{u}) = \bar{u}^3$

$b_{-1}(\bar{u})$ *is vacuous*

$b_{-2}(\bar{u})$ *is vacuous*

$b_{-3}(\bar{u})$ *is vacuous* .

For purposes of comparison, Figure 92 summarizes the various ways in which multiplicities may be distributed among the knots defining a cubic B-spline in which the non-vacuous intervals all have unit length.

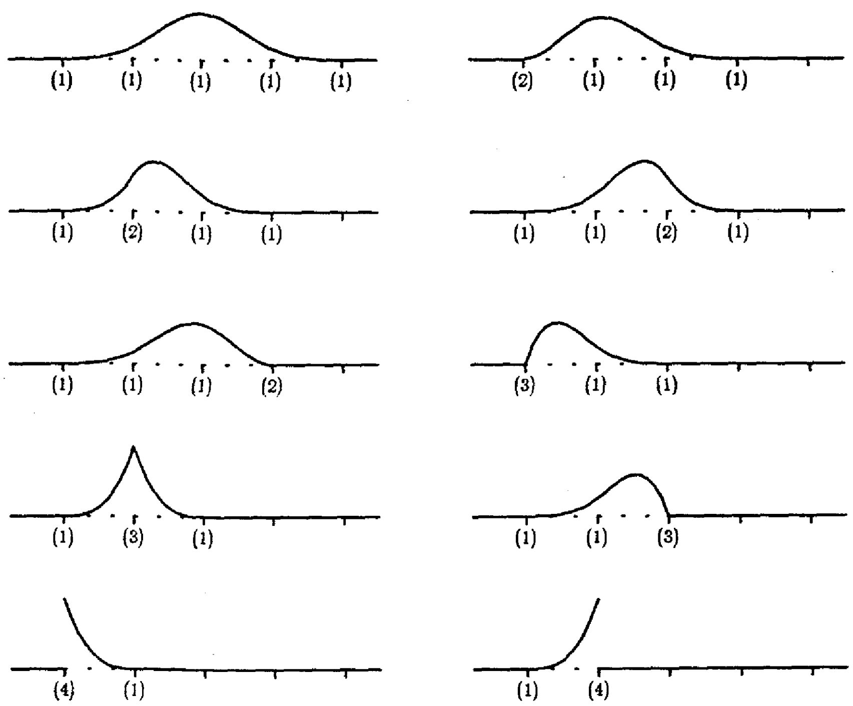


Figure 92. The various ways in which knot multiplicities can be distributed among the knots defining a cubic B-spline. The multiplicities are indicated in parentheses.

The following set of figures illustrates the effect that multiple knots have on the shape of a parametric curve.

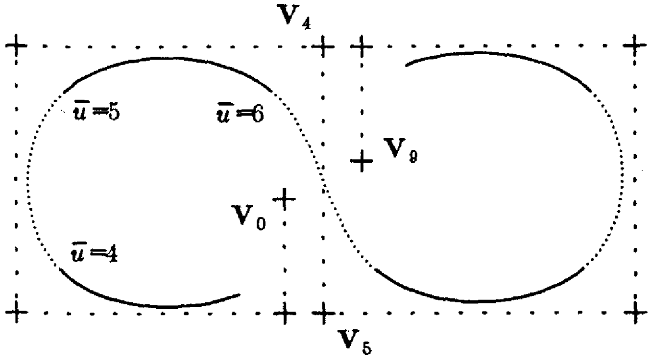


Figure 93. This is a uniform cubic B-spline curve: The knots are equally spaced and of multiplicity 1.

Figure 93 shows a simple curve obtained from ten control vertices using uniform cubic B-splines.

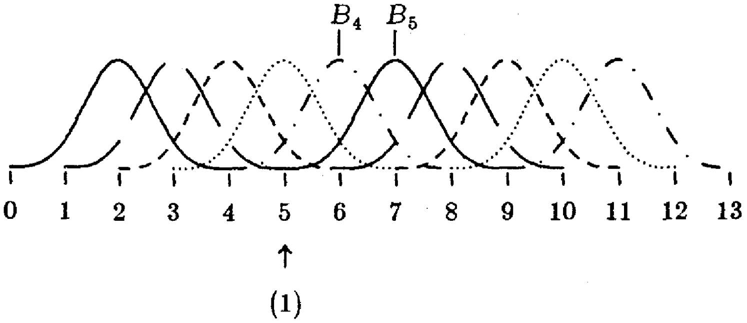


Figure 94. These are the B-splines used in constructing the curve of Figure 93. They are all translates of one another.

We have flagged the B-splines as follows in Figure 94:

$$B_{4,4}(\bar{u}) \to B_4 \quad \text{and} \quad B_{5,4}(\bar{u}) \to B_5 \ .$$

These particular B-splines are distinguished because we will be increasing the multiplicity of the knot at $\bar{u}=5$, and these basis functions are the ones that will show the most effect. The curve will likewise display the largest change in the interval between the control vertices $\mathbf{V}_4$ and $\mathbf{V}_5$.

The space of splines under consideration is

$$\mathbf{S}(\mathbf{P}^4,\{\bar{u}_i\}_0^{13})$$

and the dimension of this space is 10, which is just what we need to manage ten control vertices. The parameter range is $[3,10)$, and as $\bar{u}$ travels from 3 to 10, it passes through the seven intervals

$$[3,4)\ ,\ [4,5)\ ,\ [5,6)\ ,\ [6,7)\ ,\ [7,8)\ ,\ [8,9)\ ,\ \text{and}\ \ [9,10)\ .$$

The portions of the curve generated as $\bar{u}$ runs through these segments are indicated by alternating solid and dotted lines.

Consider now what happens if we double the knot at $\bar{u}=5$. To retain the same dimension, i.e. to use the same control vertices, we have now to use the knot sequence

$$\{\,0,1,2,3,4,5,5,6,7,8,9,10,11,12\,\}$$

so that the parameter range becomes [3,9). This means that $\bar{u}$ will now travel through only the six intervals

[3,4) , [4,5) , [5,6) , [6,7) , [7,8) , and [8,9)

as the curve is traced out. That is, there are six curve segments. Figures 95 and 96 show what happens.

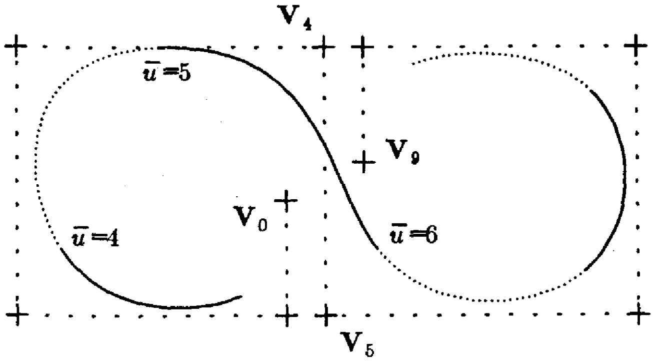


Figure 95. The breakpoints defining this curve are equally spaced, but there is a double knot at $\bar{u}$=5. (See Figure 96.)

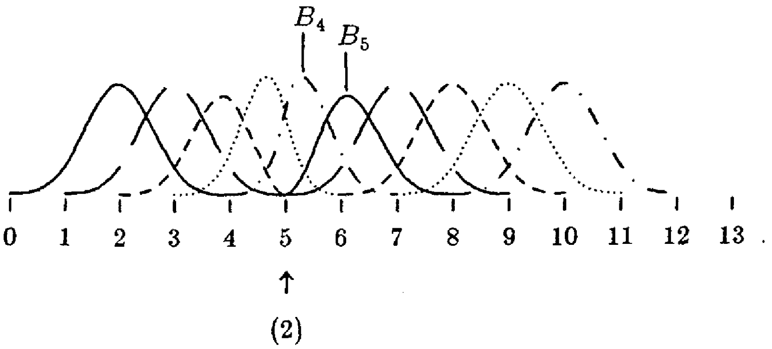


Figure 96. The B-splines used to construct the curve of Figure 96. There is a double knot at $\bar{u}$=5.

When the knot at $\bar{u}$=5 is tripled, the knot sequence becomes

$$\{\,0,1,2,3,4,5,5,5,6,7,8,9,10,11\,\}\ ,$$

the parameter range becomes [3,8), and there are five segments. The fact that some of the underlying B-splines have discontinuities in the first derivative is apparent in Figures 97 and 98.

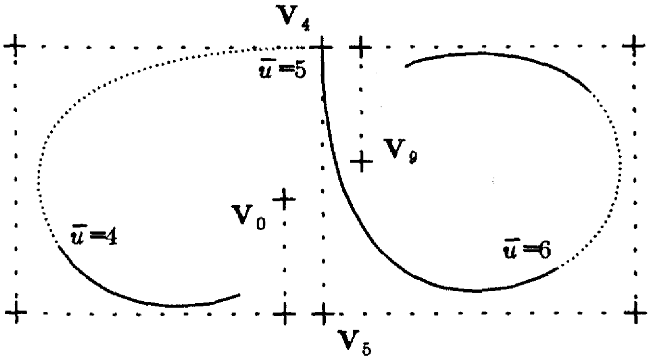


Figure 97. The breakpoints defining this curve are equally spaced, but there is a triple knot at $\bar{u}$=5. Since this is a cubic B-spline, we are left with positional continuity only at the triple knot. (See Figure 98.)

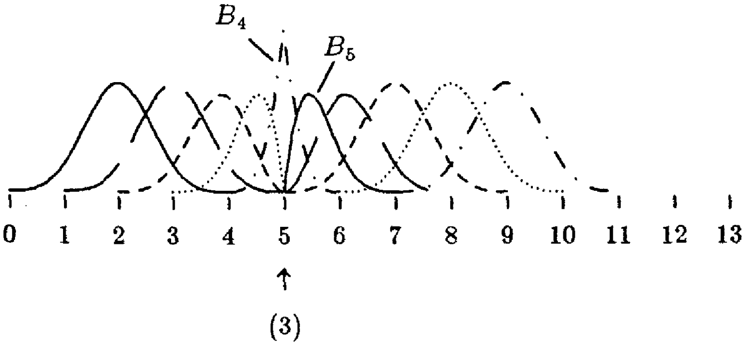


Figure 98. The B-splines used to construct the curve of Figure 97. There is a triple knot at $\bar{u}$=5.

Finally, quadrupling the knot at $\bar{u}$=5 yields the knot sequence

$$\{ 0, 1, 2, 3, 4, 5, 5, 5, 5, 6, 7, 8, 9, 10 \} ,$$

the parameter range [3,7), and four segments. Some of the underlying B-splines in Figure 100 are now discontinuous, as is the curve in Figure 99 which results.

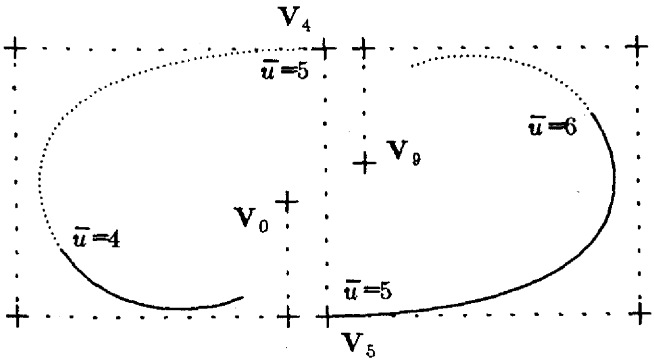


Figure 99. The breakpoints defining this curve are equally spaced, but there is a quadruple knot at $\bar{u}$=5. Since this is a cubic B-spline, we are left with no continuity whatsoever at the multiple knot. (See Figure 100.)

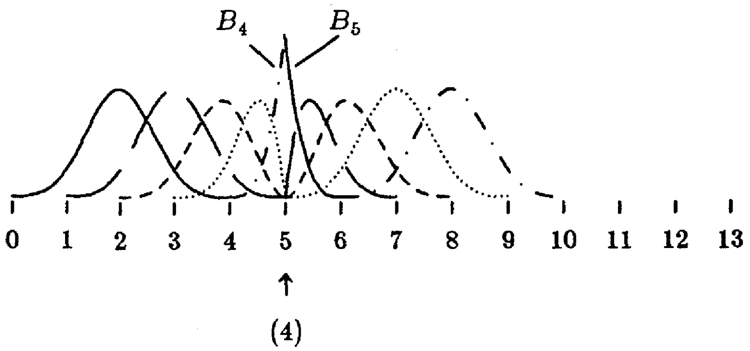


Figure 100. The B-splines used to construct the curve of Figure 99. There is a quadruple knot at $\bar{u}=5$. Although $B_{4,4}$ and $B_{5,4}$ both have the value 1 at $\bar{u}=5$, they are scaled by distinct control vertices and so a positional discontinuity will result in the curve so long as the control vertices scaling them are not identical.

## 9.6. Altering Knot Spacing — More Cubic B-splines

It is also interesting to see what effect results from changing the knot spacing rather than knot multiplicity. Figure 101 shows the original uniform cubic B-spline curve of Figure 93, generated by the B-splines of Figure 94, superimposed on the curve obtained when the knot interval $[\bar{u}_6,\bar{u}_7) = [6,7)$ defining the middle curve segment in Figure 93 is shrunk to 0.2 units in length.

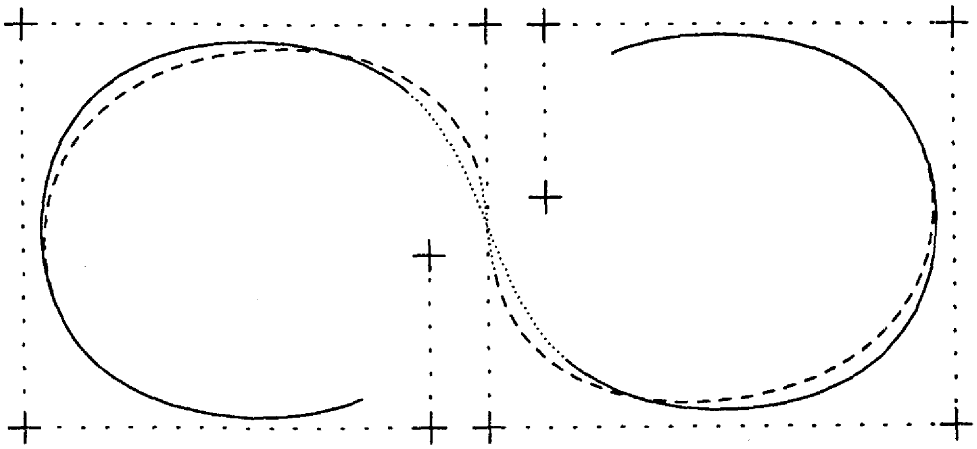


Figure 101. The middle segment in both curves is dotted. The remainder of the uniform cubic B-spline is drawn with a solid line while the remainder of the non-uniformly spaced curve is drawn dashed. The parametric length of the dotted segment is here being changed from 1.0 to 0.2.

Figure 102 illustrates what happens when the same interval is instead expanded to 5 units in length.

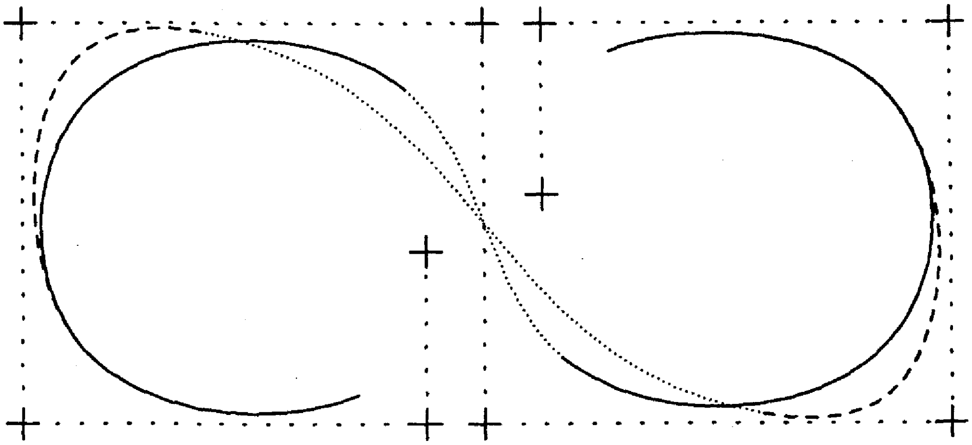


Figure 102. The middle segment in both curves is dotted. The remainder of the uniform cubic B-spline is drawn with a solid line while the remainder of the non-uniformly spaced curve is drawn dashed. The parametric length of the dotted segment is here being changed from 1.0 to 5.0.

# 10. B-Spline Properties

The previous section has given the basic definition of B-splines in general terms and showed the manner in which they can be used to construct parametric spline curves and surfaces. To proceed further, it will be necessary to discuss a few more of the theoretical properties that B-splines possess. We want to know, for example,

- whether every B-spline constructed according to the general definition is *positive* on the interval of its associated knots (something we surely want it to be if it is to be used as a "weight function" for constructing curves from control vertices);
- whether the general B-splines have *compact support* ("local control");
- whether the general B-splines sum to one (which will mean, together with the property of positivity, that general B-splines have the *convex hull* property);
- what constitutes a good way of *evaluating* the B-splines.

This section will establish some results in this regard.

### 10.1. Differencing Products — The Leibnitz Rule

We will begin by establishing that the B-splines can be evaluated by means of a recurrence.

Let us try the following exercise in creative algebra, and see where it leads. First, recall that

$$B_{i,k}(\bar{u}) = (-1)^k(\bar{u}_{i+k}-\bar{u}_i)[\bar{u}_i(k):t](\bar{u}-t)_+^{k-1} .$$

But note the obvious fact that

$$(\bar{u}-t)_+^{k-1} = (\bar{u}-t)\cdot(\bar{u}-t)_+^{k-2} ,$$

at least for $k \geq 2$. This means that $B_{i,k}(\bar{u})$ is constructed by differencing a product, for $k>1$. If the difference operator were a differentiation operator instead, it would be natural for us to follow this observation by an application of the product rule; for example,

$$\begin{aligned} D_t(\bar{u}-t)\cdot(\bar{u}-t)_+^{k-2} &= \Big[(-1)(\bar{u}-t)_+^{k-2}\Big] + \Big[(\bar{u}-t)\cdot(-1)(k-2)(\bar{u}-t)_+^{k-3}\Big] \\ &= \Big[-(\bar{u}-t)_+^{k-2}\Big] + \Big[-(k-2)(\bar{u}-t)_+^{k-2}\Big] . \end{aligned}$$

While this might not seem very productive, it would, in effect, have split a $k^{\text{th}}$-order B-spline, constructed from $(\bar{u}-t)_+^{k-1}$, into a combination of two terms involving $(\bar{u}-t)_+^{k-2}$, from which we might attempt to draw a connection with B-splines of order $k-1$. By this route, perhaps a recurrence could be established.

That is, if we can relate the $k^{\text{th}}$-order B-splines to those of order $k-1$, then we can relate those of order $k-1$ to those of order $k-2$, and so on down to $1^{\text{st}}$-order B-splines (the simple step functions).

This is precisely the approach we will take. In order to pursue this, we will have to establish a product rule for the divided difference operator.

We begin by recalling the *product rule for differentiation*:

$$D_x\{f_1(x)\cdot f_2(x)\} = \{D_x f_1(x)\}\cdot f_2(x) + f_1(x)\cdot\{D_x f_2(x)\} \ .$$

A more general version of the product rule is the *Leibniz rule*:

$$D_x^{(l)}\{f_1(x)\cdot f_2(x)\} = \sum_{r=0}^{l}\binom{l}{r}\{D_x^{(r)} f_1(x)\}\cdot\{D_x^{(l-r)} f_2(x)\} \ ,$$

where

$$\binom{l}{r} = \frac{l!}{r!(l-r)!}$$

is the binomial coefficient. That is, for example (using superscript notation for derivatives in the interests of succinctness),

$$\{f_1(x)f_2(x)\}^{(2)} = f_1^{(2)}(x)f_2^{(0)}(x) + 2f_1^{(1)}(x)f_2^{(1)}(x) + f_1^{(0)}(x)f_2^{(2)}(x) \ ,$$

and

$$\{f_1(x)f_2(x)\}^{(3)} = f_1^{(3)}(x)f_2^{(0)}(x) + 3f_1^{(2)}(x)f_2^{(1)}(x) + 3f_1^{(1)}(x)f_2^{(2)}(x) + f_1^{(0)}(x)f_2^{(3)}(x) \ ,$$

and so on through the Pascal triangle.

The corresponding Leibniz rule for divided differences is as follows:

**Theorem:** For any $z_i, \ldots, z_{i+l}$ and any appropriately differentiable functions $f_1(x)$ and $f_2(x)$:

$$[z_i(l):x]\{f_1(x)f_2(x)\}$$
$$= \sum_{r=0}^{l}\{[z_i(r):x]f_1(x)\}\{[z_{i+r}(l-r):x]f_2(x)\} \ .$$

Note that this is virtually the same as the rule for differentiating a product, save that the binomial coefficients do not appear. For example if $l=2$,

$$\begin{aligned}[z_i(2):x]\{f_1(x)\cdot f_2(x)\} &= [z_i(0):x]f_1(x)\cdot[z_i(2):x]f_2(x)\\ &+ [z_i(1):x]f_1(x)\cdot[z_{i+1}(1):x]f_2(x)\\ &+ [z_i(2):x]f_1(x)\cdot[z_{i+2}(0):x]f_2(x) \ .\end{aligned}$$

This result is important enough that we owe ourselves the struggle of seeing how it can be established.

**Argument:** Assume, for convenience, that $z_i \leq \cdots \leq z_{i+l}$.
We proceed by induction on $l$.

$l = 0$:

$$\begin{aligned}[z_i(0):x]\{f_1(x)f_2(x)\} &= f_1(z_i)f_2(z_i)\\ &= \sum_{r=0}^{0}\{[z_i(0):x]f_1(x)\}\{[z_i(0):x]f_2(x)\}\end{aligned}$$

by definition (trivially).

$l > 0$ and $z_i = z_{i+l}$:

In this case the Leibniz rule for derivatives applies:

$$[z_i(l):x]\{f_1(x)f_2(x)\} = \frac{1}{l!}D_x^{(l)}\{f_1(x)f_2(x)\}|_{x=z_i}$$

$$= \frac{1}{l!}\sum_{r=0}^{l}\frac{l!}{r!(l-r)!}\{D_x^{(r)}f_1(x)\}\{D_x^{(l-r)}f_2(x)\}|_{x=z_i}$$

$$= \sum_{r=0}^{l}\{\frac{1}{r!}D_x^{(r)}f_1(x)\}\{\frac{1}{(l-r)!}D_x^{(l-r)}f_2(x)\}|_{x=z_i}$$

$$= \sum_{r=0}^{l}\{[z_i(r):x]f_1(x)\}\{[z_{i+r}(l-r):x]f_2(x)\} \ .$$

$l > 0$ and $z_i < z_{i+l}$:

Now we make use of the inductive assumption that the theorem holds for $l-1$.
Then

$$[z_i(l):x]\{f_1(x)f_2(x)\}$$

$$= \frac{[z_{i+1}(l-1):x]\{f_1(x)f_2(x)\} \ - \ [z_i(l-1):x]\{f_1(x)f_2(x)\}}{z_{i+l} \ - \ z_i}$$

$$= \left(\sum_{r=0}^{l-1}\{[z_{i+1}(r):x]f_1(x)\}\{[z_{i+1+r}(l-1-r):x]f_2(x)\}\right.$$

$$\left. - \sum_{r=0}^{l-1}\{[z_i(r):x]f_1(x)\}\{[z_{i+r}(l-1-r):x]f_2(x)\}\right) \Big/ \ (z_{i+l} - z_i) \ .$$

We can add and subtract

(more...)

$$\sum_{r=0}^{l-1}\{[z_i(r):x]f_1(x)\}\{[z_{i+r+1}(l-1-r):x]f_2(x)\}$$

in the numerator to obtain

$$\begin{aligned}\Bigg\{&\sum_{r=0}^{l-1}\{[z_{i+1}(r):x]f_1(x)\}\{[z_{i+r+1}(l-1-r):x]f_2(x)\}\\ &\quad-\sum_{r=0}^{l-1}\{[z_i(r):x]f_1(x)\}\{[z_{i+r+1}(l-1-r):x]f_2(x)\}\\ &\quad+\sum_{r=0}^{l-1}\{[z_i(r):x]f_1(x)\}\{[z_{i+r+1}(l-1-r):x]f_2(x)\}\\ &\quad-\sum_{r=0}^{l-1}\{[z_i(r):x]f_1(x)\}\{[z_{i+r}(l-1-r):x]f_2(x)\}\Bigg\}\Big/(z_{i+l}-z_i)\ .\end{aligned}$$

The first and the second terms in the numerator can be combined, as can the third and fourth, to change the numerator into the following:

$$\begin{aligned}&\sum_{r=0}^{l-1}\Big([z_{i+r+1}(l-1-r):x]f_2(x)\Big)\Big([z_{i+1}(r):x]f_1(x)-[z_i(r):x]f_1(x)\Big)\\ &+\\ &\sum_{r=0}^{l-1}\Big([z_i(r):x]f_1(x)\Big)\Big([z_{i+r+1}(l-1-r):x]f_2(x)-[z_{i+r}(l-1-r):x]f_2(x)\Big)\ .\end{aligned}\tag{57}$$

But in the first term the recursive definition of the divided difference provides the substitution

$$\begin{aligned}&[z_{i+1}(r):x]f_1(x)-[z_i(r):x]f_1(x)\\ &\qquad=[z_i(r+1):x]f_1(x)\cdot(z_{i+r+1}-z_i)\ .\end{aligned}$$

(This, in itself, deserves a short argument. Two cases present themselves: in the first, $z_i<z_{i+r+1}$, and in the second $z_i=z_{i+r+1}$. In the former case, the substitution is obvious. In the latter, because of the assumed ordering of the $z$ values, $z_i=\cdots=z_{i+r+1}$, and consequently both

$$[z_{i+1}(r):x]f_1(x)$$

and

$$[z_i(r):x]f_1(x)$$

are equal to

$$D_x^{(r+1)}f_1(x)|_{x=z_i}\ .$$

Hence their difference is zero. On the other hand

(more...)

$$[z_i(r+1):x]\, f_1(x)\cdot(z_{i+r+1}-z_i)$$
$$= \{D_x^{(r+2)} f_1(x)|_{x=z_i}\}\cdot(z_{i+r+1}-z_i) \ ,$$

which is also zero because of the second factor.)

In the second term of (57) we similarly have the substitution

$$[z_{i+r+1}(l-1-r):x]\, f_2(x) - [z_{i+r}(l-1-r):x]\, f_2(x)$$
$$= [z_{i+r}(l-r):x]\, f_2(x)\cdot(z_{i+l}-z_{i+r}) \ .$$

After making these substitutions, we can multiply out the factors $(z_{i+r+1}-z_i)$ and $(z_{i+l}-z_{i+r})$ to obtain

$$\Bigg\{\sum_{r=0}^{l-1} z_{i+r+1}[z_i(r+1):x]\, f_1(x)[z_{i+r+1}(l-1-r):x]\, f_2(x)$$
$$- \sum_{r=0}^{l-1} z_i[z_i(r+1):x]\, f_1(x)[z_{i+r+1}(l-1-r):x]\, f_2(x)$$
$$+ \sum_{r=0}^{l-1} z_{i+l}[z_i(r):x]\, f_1(x)[z_{i+r}(l-r):x]\, f_2(x)$$
$$- \sum_{r=0}^{l-1} z_{i+r}[z_i(r):x]\, f_1(x)[z_{i+r}(l-r):x]\, f_2(x)\Bigg\} \Big/ (z_{i+l}-z_i) \ .$$

The first and the last sum collapse, leaving only the two terms

$$+z_{i+l}[z_i(l):x]\, f_1(x)[z_{i+l}(0):x]\, f_2(x)$$

and

$$-z_i[z_i(0):x]\, f_1(x)[z_i(l):x]\, f_2(x)$$

respectively. The second of these terms can be put into the second sum, and the first of these terms can be put into the third sum, to yield

$$\Bigg\{z_{i+l}\sum_{r=0}^{l}[z_i(r):x]\, f_1(x)[z_{i+r}(l-r):x]\, f_2(x)$$
$$- z_i\sum_{r=0}^{l}[z_i(r):x]\, f_1(x)[z_{i+r}(l-r):x]\, f_2(x)\Bigg\} \Big/ (z_{i+l}-z_i) \ .$$

And, finally, the term $(z_{i+l}-z_i)$ can be divided out to give

(more...)

$$\sum_{r=0}^{l}[z_i(r):x]\,f_1(x)[z_{i+r}(l-r):x]\,f_2(x) \ .$$

This completes our discussion the Leibnitz rule for divided differences.

**10.2. Establishing a Recurrence** To repeat the observations made at the beginning of this chapter, the definition of a general B-spline is:

$$B_{i,k}(\bar{u}) = (-1)^k(\bar{u}_{i+k}-\bar{u}_i)[\bar{u}_i(k):t](\bar{u}-t)_+^{k-1} \ .$$

We observe that

$$(\bar{u}-t)_+^{k-1} = (\bar{u}-t)\cdot(\bar{u}-t)_+^{k-2} \ ,$$

at least for $k>1$. So, for $k>1$, $B_{i,k}(\bar{u})$ is constructed by differencing a product. We apply the Leibniz rule:

$$\begin{aligned}
&[\bar{u}_i(k):t](\bar{u}-t)_+^{k-1} \\
&\quad = [\bar{u}_i(k):t](\bar{u}-t)(\bar{u}-t)_+^{k-2} \\
&\quad = \sum_{j=0}^{k}\{[\bar{u}_i(j):t](\bar{u}-t)\}\cdot\{[\bar{u}_{i+j}(k-j):t](\bar{u}-t)_+^{k-2}\} \\
&\quad = [\bar{u}_i(0):t](\bar{u}-t)[\bar{u}_i(k):t](\bar{u}-t)_+^{k-2} \\
&\qquad\qquad + [\bar{u}_i(1):t](\bar{u}-t)[\bar{u}_{i+1}(k-1):t](\bar{u}-t)_+^{k-2} \\
&\qquad\qquad + [\bar{u}_i(2):t](\bar{u}-t)[\bar{u}_{i+2}(k-2):t](\bar{u}-t)_+^{k-2} \\
&\qquad\qquad + \cdots + [\bar{u}_i(k):t](\bar{u}-t)[\bar{u}_{i+k}(0):t](\bar{u}-t)_+^{k-2} \ .
\end{aligned}$$

But note that

$$[\bar{u}_i(0):t](\bar{u}-t) = (\bar{u}-\bar{u}_i) \ ,$$

that

$$\begin{aligned}
[\bar{u}_i(1):t](\bar{u}-t) &= \frac{(\bar{u}-\bar{u}_{i+1})-(\bar{u}-\bar{u}_i)}{\bar{u}_{i+1}-\bar{u}_i} \ . \\
&= \frac{\bar{u}_i-\bar{u}_{i+1}}{\bar{u}_{i+1}-\bar{u}_i} = -1 \ ,
\end{aligned}$$

that

$$[\bar{u}_i(2):t](\bar{u}-t) = \frac{[\bar{u}_{i+1}(1):t](\bar{u}-t)-[\bar{u}_i(1):t](\bar{u}-t)}{\bar{u}_{i+2}-\bar{u}_i}$$

$$= \frac{(-1)-(-1)}{\bar{u}_{i+2}-\bar{u}_i} = 0 \ ,$$

and that all further differences of $(\bar{u}-t)$ are also zero. (A useful observation to make is that the $k^{\text{th}}$ divided-difference operator has the same property as does the $k^{\text{th}}$ differentiation operator in that it will cancel to zero any $k^{\text{th}}$-order polynomial.) So

$$\begin{aligned}&[\bar{u}_i(k):t](\bar{u}-t)_+^{k-1}\\ &\quad = (\bar{u}-\bar{u}_i)[\bar{u}_i(k):t](\bar{u}-t)_+^{k-2} - [\bar{u}_{i+1}(k-1):t](\bar{u}-t)_+^{k-2} \ .\end{aligned}$$

From the recursive definition of divided differences, the first term can be written as

$$\begin{aligned}&(\bar{u}-\bar{u}_i)[\bar{u}_i(k):t](\bar{u}-t)_+^{k-2}\\ &\quad = (\bar{u}-\bar{u}_i)\frac{[\bar{u}_{i+1}(k-1):t](\bar{u}-t)_+^{k-2} - [\bar{u}_i(k-1):t](\bar{u}-t)_+^{k-2}}{\bar{u}_{i+k}-\bar{u}_i} \ .\end{aligned}$$

(We assume that this expression is legal, i.e. the denominator is not zero, since if $\bar{u}_i = \bar{u}_{i+k}$, this means that $\bar{u}_i$ would have multiplicity greater than $k$, and such cases are not interesting.) Consequently,

$$\begin{aligned}B_{i,k}(\bar{u}) &= (-1)^k(\bar{u}_{i+k}-\bar{u}_i)[\bar{u}_i(k):t](\bar{u}-t)_+^{k-1}\\ &= (-1)^k(\bar{u}_{i+k}-\bar{u}_i)\frac{\bar{u}-\bar{u}_i}{\bar{u}_{i+k}-\bar{u}_i}[\bar{u}_{i+1}(k-1):t](\bar{u}-t)_+^{k-2}\\ &\quad - (-1)^k(\bar{u}_{i+k}-\bar{u}_i)\frac{\bar{u}-\bar{u}_i}{\bar{u}_{i+k}-\bar{u}_i}[\bar{u}_i(k-1):t](\bar{u}-t)_+^{k-2}\\ &\quad - (-1)^k(\bar{u}_{i+k}-\bar{u}_i)[\bar{u}_{i+1}(k-1):t](\bar{u}-t)_+^{k-2} \ .\end{aligned}$$

The first and last terms can be combined to give

$$\begin{aligned}B_{i,k}(\bar{u}) &= (-1)^k(\bar{u}-\bar{u}_{i+k})[\bar{u}_{i+1}(k-1):t](\bar{u}-t)_+^{k-2}\\ &\quad - (-1)^k(\bar{u}-\bar{u}_i)[\bar{u}_i(k-1):t](\bar{u}-t)_+^{k-2} \ .\end{aligned}$$

Note, now, that

$$(-1)^k(\bar{u}-\bar{u}_{i+k}) = (-1)^{k-1}(\bar{u}_{i+k}-\bar{u})$$

and that

$$-(-1)^k(\bar{u}-\bar{u}_i) = +(-1)^{k-1}(\bar{u}-\bar{u}_i) \ .$$

This means that

$$\begin{aligned}B_{i,k}(\bar{u}) &= (-1)^{k-1}(\bar{u}_{i+k}-\bar{u})[\bar{u}_{i+1}(k-1):t](\bar{u}-t)_+^{k-2}\\ &\quad + (-1)^{k-1}(\bar{u}-\bar{u}_i)[\bar{u}_i(k-1):t](\bar{u}-t)_+^{k-2} \ .\end{aligned} \tag{58}$$

But, in this expression, it is easy to recognize two lower-order B-splines.

Let us look carefully at the first term. If

$$\bar{u}_{i+1} = \bar{u}_{i+k}$$

then, by the definition of the divided difference,

$$[\bar{u}_{i+1}(k-1){:}t](\bar{u}-t)_+^{k-2} = \frac{1}{(k-1)!} D_t^{k-1}(\bar{u}-t)_+^{k-2} = 0 \ .$$

On the other hand, if

$$\bar{u}_{i+1} < \bar{u}_{i+k} \ ,$$

then, since by the definition of the B-splines we would have

$$B_{i+1,k-1}(\bar{u}) = (-1)^{k-1}(\bar{u}_{i+k}-\bar{u}_{i+1})[\bar{u}_{i+1}(k-1){:}t](\bar{u}-t)_+^{k-2} \ ,$$

it follows that the first term equals

$$\frac{\bar{u}_{i+k}-\bar{u}}{\bar{u}_{i+k}-\bar{u}_{i+1}} B_{i+1,k-1}(\bar{u}) \ . \tag{59}$$

That is, the first term is zero if $\bar{u}_{i+1} = \cdots = \bar{u}_{i+k}$ and it equals (59) if $\bar{u}_{i+1} < \bar{u}_{i+k}$. A similar discussion applies to the second term. This means that we can reasonably write the equation

$$B_{i,k}(\bar{u}) = \frac{\bar{u}-\bar{u}_i}{\bar{u}_{i+k-1}-\bar{u}_i} B_{i,k-1}(\bar{u}) + \frac{\bar{u}_{i+k}-\bar{u}}{\bar{u}_{i+k}-\bar{u}_{i+1}} B_{i+1,k-1}(\bar{u}) \tag{60}$$

provided that we interpret the terms

$$\frac{\bar{u}-\bar{u}_i}{\bar{u}_{i+k-1}-\bar{u}_i} B_{i,k-1}(\bar{u}) \quad \text{and} \quad \frac{\bar{u}_{i+k}-\bar{u}}{\bar{u}_{i+k}-\bar{u}_{i+1}} B_{i+1,k-1}(\bar{u})$$

as zero, respectively, whenever

$$\bar{u}_{i+k-1}-\bar{u}_i = 0 \quad \text{and} \quad \bar{u}_{i+k}-\bar{u}_{i+1} = 0 \ .$$

That is, we have discovered that the B-splines satisfy a *recurrence relation.*

### 10.3. The Recurrence and Examples

**Theorem:** For any $i \in \{0,1, \ldots, m+k\}$

$$B_{i,1}(\bar{u}) = \begin{cases} 1 & \bar{u}_i \le \bar{u} < \bar{u}_{i+1} \\ 0 & \text{otherwise} \end{cases}$$

and

$$B_{i,r}(\bar{u}) = \frac{\bar{u}-\bar{u}_i}{\bar{u}_{i+r-1}-\bar{u}_i} B_{i,r-1}(\bar{u}) + \frac{\bar{u}_{i+r}-\bar{u}}{\bar{u}_{i+r}-\bar{u}_{i+1}} B_{i+1,r-1}(\bar{u})$$

for $r = 2, 3, \ldots, k$,

(more...)

where we interpret the terms

$$\frac{\bar{u}-\bar{u}_i}{\bar{u}_{i+r-1}-\bar{u}_i}\, B_{i,r-1}(\bar{u}) \quad \text{and} \quad \frac{\bar{u}_{i+r}-\bar{u}}{\bar{u}_{i+r}-\bar{u}_{i+1}}\, B_{i+1,r-1}(\bar{u})$$

as zero respectively whenever

$$\bar{u}_{i+r-1}-\bar{u}_i \;=\; 0 \;\text{ and }\; \bar{u}_{i+r}-\bar{u}_{i+1} \;=\; 0 \;.$$

To get a feeling for this recurrence, we will use it to construct all the possible linear B-splines. Consider the general case of three knots

$$\bar{u}_i \;,\; \bar{u}_{i+1} \;,\; \text{and } \bar{u}_{i+2}$$

If these three knots are distinct, then according to the recurrence

$$\begin{aligned} B_{i,2}(\bar{u}) &= \frac{\bar{u}-\bar{u}_i}{\bar{u}_{i+1}-\bar{u}_i}\, B_{i,1}(\bar{u}) + \frac{\bar{u}_{i+2}-\bar{u}}{\bar{u}_{i+2}-\bar{u}_{i+1}}\, B_{i+1,1}(\bar{u}) \\ &= \frac{\bar{u}-\bar{u}_i}{\bar{u}_{i+1}-\bar{u}_i} \left\{ \begin{matrix} 1 & \bar{u}_i \le \bar{u} < \bar{u}_{i+1} \\ \\ 0 & \text{otherwise} \end{matrix} \right\} + \frac{\bar{u}_{i+2}-\bar{u}}{\bar{u}_{i+2}-\bar{u}_{i+1}} \left\{ \begin{matrix} 1 & \bar{u}_{i+1} \le \bar{u} < \bar{u}_{i+2} \\ \\ 0 & \text{otherwise} \end{matrix} \right\} \end{aligned}$$

That is, according to the recursive formula,

$$B_{i,2}(\bar{u}) = \begin{cases} \dfrac{\bar{u}-\bar{u}_i}{\bar{u}_{i+1}-\bar{u}_i} & \bar{u}_i \le \bar{u} < \bar{u}_{i+1} \\[2ex] \dfrac{\bar{u}_{i+2}-\bar{u}}{\bar{u}_{i+2}-\bar{u}_{i+1}} & \bar{u}_{i+1} \le \bar{u} < \bar{u}_{i+2} \\[2ex] 0 & \text{otherwise} \;. \end{cases}$$

This is just the familiar "hat function" pictured below.

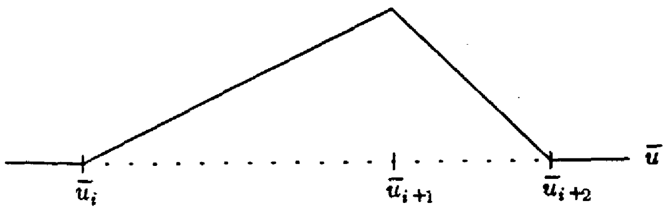


Figure 103. $B_{i,2}(\bar{u})$ for knots of multiplicity 1.

Suppose that the right-hand two knots are pushed together; i.e.

$$\bar{u}_i < \bar{u}_{i+1} = \bar{u}_{i+2} \quad .$$

Then the recursive formula becomes

$$B_{i,2}(\bar{u}) = \frac{\bar{u}-\bar{u}_i}{\bar{u}_{i+1}-\bar{u}_i} B_{i,1}(\bar{u}) + 0$$

$$= \frac{\bar{u}-\bar{u}_i}{\bar{u}_{i+1}-\bar{u}_i} \left\{ \begin{matrix} 1 & \bar{u}_i \le \bar{u} < \bar{u}_{i+1} \\ \\ 0 & \text{otherwise} \end{matrix} \right\} ,$$

which leaves us with

$$B_{i,2}(\bar{u}) = \begin{cases} \dfrac{\bar{u}-\bar{u}_i}{\bar{u}_{i+1}-\bar{u}_i} & \bar{u}_i \le \bar{u} < \bar{u}_{i+1} \\ \\ 0 & \text{otherwise} \quad . \end{cases}$$

This function has the form given in Figure 104.

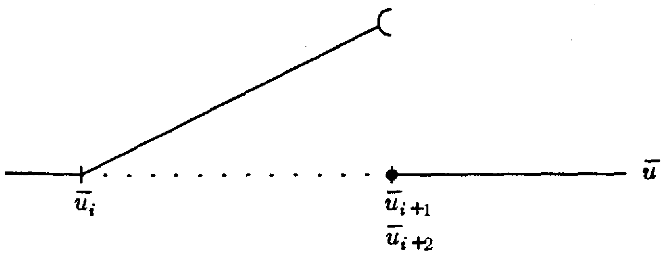


Figure 104. $B_{i,2}(\bar{u})$ with a double knot at the right.

Finally, if

$$\bar{u}_i = \bar{u}_{i+1} < \bar{u}_{i+2} \quad ,$$

then the recursive formula becomes

$$B_{i,2}(\bar{u}) = 0 + \frac{\bar{u}_{i+2}-\bar{u}}{\bar{u}_{i+2}-\bar{u}_{i+1}} B_{i+1,1}(\bar{u}) \quad .$$

$$= \frac{\bar{u}_{i+2}-\bar{u}}{\bar{u}_{i+2}-\bar{u}_{i+1}} \left\{ \begin{matrix} 1 & \bar{u}_{i+1} \le \bar{u} < \bar{u}_{i+2} \\ \\ 0 & \text{otherwise} \end{matrix} \right\} ,$$

which leaves us with

$$B_{i,2}(\bar{u}) = \begin{cases} \dfrac{\bar{u}_{i+2}-\bar{u}}{\bar{u}_{i+2}-\bar{u}_{i+1}} & \bar{u}_{i+1} \le \bar{u} < \bar{u}_{i+2} \\ & \\ 0 & \text{otherwise} \ , \end{cases}$$

which is shown in Figure 105.

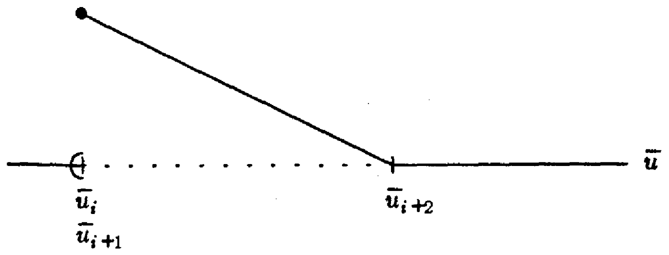


Figure 105. $B_{i,2}(\bar{u})$ with a double knot at the left.

### 10.4. Evaluating B-splines Through Recurrence

The evaluation of general B-splines may be carried out directly by this recurrence. To illustrate, we will find the value of $B_{0,3}(2)$ on the knot sequence 0,1,3,4 and on the knot sequence 0,1,1,3. These are the two examples that we used to illustrate the divided-difference evaluation of a B-spline in Chapter 9, and the pictures were

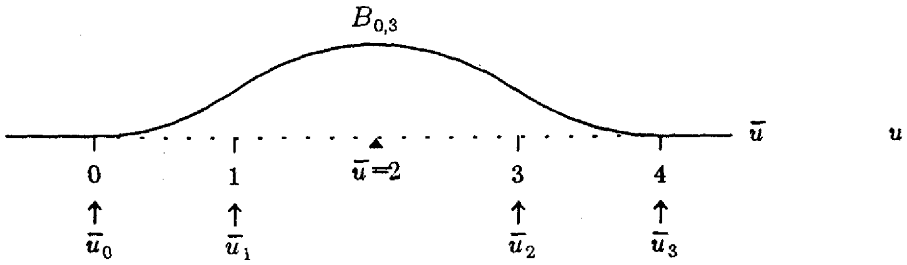


Figure 106. A quadratic B-spline with single knots.

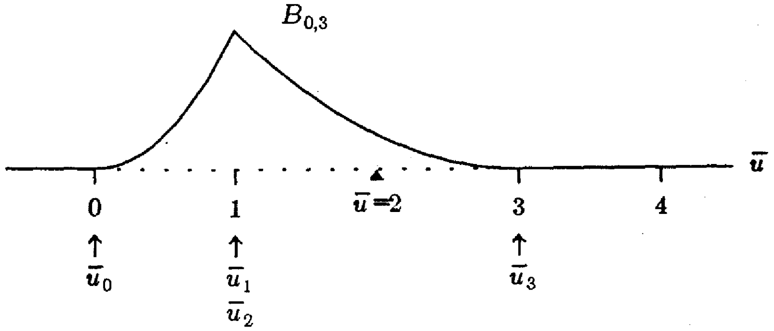


Figure 107. A quadratic B-spline with a double knot.

The tables have the same flavour as those used in the divided-difference illustration, namely that an entry in one column is obtained by performing a computation on the adjacent entries to its left. Note that $\bar{u}_i$ is needed in the computation for $i=0,\ldots,3$, but we only need the values of $B_{i,1}(\bar{u})$ for $i=0,\ldots,2$. For the distinct-knot example we have

| $\bar{u}_i$ | $B_{i,1}(\bar{u})$ | $B_{i,2}(\bar{u})$ | $B_{i,3}(\bar{u})$ |
|---|---|---|---|
| 0 | 0 | | |
| | | $\frac{(3-2)}{(3-1)}1+\frac{(2-0)}{(1-0)}0=\frac{1}{2}$ | |
| 1 | 1 | | $\frac{(4-2)}{(4-1)}\frac{1}{2}+\frac{(2-0)}{(3-0)}\frac{1}{2}=\frac{2}{3}$ |
| | | $\frac{(4-2)}{(4-3)}0+\frac{(2-1)}{(3-1)}1=\frac{1}{2}$ | |
| 3 | 0 | | |
| 4 | | | |

and for the double-knot example we have

| $\bar{u}_i$ | $B_{i,1}(\bar{u})$ | $B_{i,2}(\bar{u})$ | $B_{i,3}(\bar{u})$ |
|---|---|---|---|
| 0 | 0 | | |
| | | $0+\frac{(2-0)}{(1-0)}0=0$ | |
| 1 | 0 | | $\frac{(3-2)}{(3-1)}\frac{1}{2}+\frac{(2-0)}{(1-0)}0=\frac{1}{4}$ |
| | | $\frac{(3-2)}{(3-1)}1+0=\frac{1}{2}$ | |
| 1 | 1 | | |
| 3 | | | |

Observe that in all instances in the above two computational tables arithmetic was carried out entirely with nonnegative numbers. The implication of this is that the recurrence computation of the values of a B-spline will be numerically safe and accurate.

## 10.5. Compact Support, Positivity, and the Convex Hull Property

We will end by disposing of the remaining items that were listed at the beginning of this section.

**Theorem:**

$$B_{i,k}(\bar{u}) > 0 \quad \text{for } \bar{u}_i < \bar{u} < \bar{u}_{i+k}$$

and

$$B_{i,k}(\bar{u}) = 0 \quad \text{for } \bar{u} < \bar{u}_i \text{ and } \bar{u}_{i+k} \leq \bar{u}$$

(The value of $B_{i,k}(\bar{u}_i)$ will depend upon the multiplicity of $\bar{u}_i$ and upon the value of $k$. In particular, $B_{i,k}(\bar{u}_i)=1$ when the multiplicity of $\bar{u}_i$ is $k$, and is otherwise zero.)

**Argument:** We have deliberated so long over the divided-difference construction of the B-splines, that the $B_{i,k}(\bar{u})=0$ part of the statement should be clear. Briefly, $B_{i,k}(\bar{u})$ is constructed as the difference (i.e. a linear combination) of functions that are all zero for $\bar{u}<\bar{u}_i$. Consequently, $B_{i,k}(\bar{u})$ itself has to be zero for $\bar{u}<\bar{u}_i$. On the other hand, the difference is chosen so that $B_{i,k}(\bar{u})$ dies to zero for $\bar{u}\geq\bar{u}_{i+k}$.

To establish the positivity for $\bar{u}_i<\bar{u}<\bar{u}_{i+k}$, we work inductively from the recurrence.

$k=1$:

In this case the B-spline is just the step function whose value is 1 on the interval $[\bar{u}_i, \bar{u}_{i+1})$ and zero everywhere outside that interval.

Assumed true for $k-1$:

We have

$$B_{i,k}(\bar{u}) = \frac{\bar{u}-\bar{u}_i}{\bar{u}_{i+k-1}-\bar{u}_i} B_{i,k-1}(\bar{u}) + \frac{\bar{u}_{i+k}-\bar{u}}{\bar{u}_{i+k}-\bar{u}_{i+1}} B_{i+1,k-1}(\bar{u})$$

Notice that both of the factors

$$\bar{u}_{i+k}-\bar{u} \quad \text{and} \quad \bar{u}-\bar{u}_i$$

are positive for $\bar{u}_i<\bar{u}<\bar{u}_{i+k}$. Furthermore, each of the ratios

$$\frac{B_{i+1,k-1}(\bar{u})}{\bar{u}_{i+k}-\bar{u}_{i+1}}$$

and

$$\frac{B_{i,k-1}(\bar{u})}{\bar{u}_{i+k-1}-\bar{u}_i}$$

is, by the induction hypothesis, either positive or zero. They could not both be zero, since this would have to imply

(more...)

$$\bar{u}_i = \bar{u}_{i+k-1} \text{ and } \bar{u}_{i+1} = \bar{u}_{i+k}$$

which would force

$$\bar{u}_i = \bar{u}_{i+k}$$

and we have been working under the assumption that $\bar{u}_i < \bar{u} < \bar{u}_{i+k}$, which disallows this. Hence, $B_{i,k}(\bar{u})$ is the sum of two positive quantities (or one positive and one zero quantity) on the interval in question.

The above result establishes at least the plausibility of using the general B-splines as *weight functions* to construct curves and surfaces. A more important result is that the general B-splines *sum to one.*

The two properties of positivity and summation to one, together, are referred to as the *convex hull property*. This means that, if $\mathbf{Q}$ is any curve constructed from control vertices $\mathbf{V}_i$,

$$\mathbf{Q}(\bar{u}) = \sum_i \mathbf{V}_i B_{i,k}(\bar{u}) \quad ,$$

then each point on the curve, that is

$$\mathbf{Q}(\bar{u}) = \left(\mathbf{X}(\bar{u}), \mathbf{Y}(\bar{u})\right)$$

for each value of the curve parameter $\bar{u}$, is a weighted average (lies in the convex hull) of the vertices

$$\mathbf{V}_0, \ldots, \mathbf{V}_m \quad .$$

More specifically, since the B-splines are nonzero on *at most* $k$ successive breakpoint intervals, the curve point always lies in the convex hull of at most $k$ successive control vertices (and, in the case of multiple knots, even fewer than $k$).

**Theorem:** For any fixed value of $\bar{u} \in [\bar{u}_{k-1}, \bar{u}_{m+1})$.

$$\sum_{i=0}^{m} B_{i,k}(\bar{u}) = 1 \ .$$

**Argument:**

$k = 1$:

In this case the result is trivial. On any breakpoint interval for which $\bar{u}_i$ is the last knot in a multiplicity cluster $B_{i,1}(\bar{u})$ appears as is given below.

(more...)

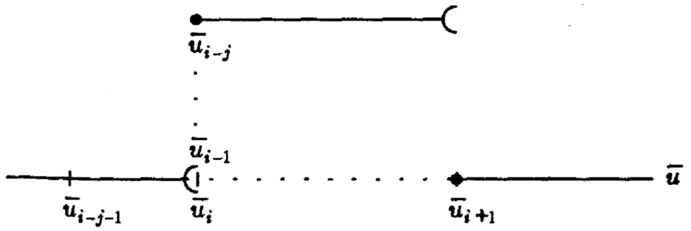


Figure 108.

$B_{i-r,1}(\bar{u})$ is vacuous for $r=1,\ldots,j$.

Assume true for $k-1$:

Then

$$\sum_{i=0}^{m} B_{i,k}(\bar{u}) = \sum_{i=0}^{m} \left\{ \frac{\bar{u}-\bar{u}_i}{\bar{u}_{i+k-1}-\bar{u}_i} B_{i,k-1}(\bar{u}) + \frac{\bar{u}_{i+k}-\bar{u}}{\bar{u}_{i+k}-\bar{u}_{i+1}} B_{i+1,k-1}(\bar{u}) \right\}$$

$$= \sum_{i=0}^{m} \frac{\bar{u}-\bar{u}_i}{\bar{u}_{i+k-1}-\bar{u}_i} B_{i,k-1}(\bar{u}) + \sum_{i=1}^{m+1} \frac{\bar{u}_{i+k-1}-\bar{u}}{\bar{u}_{i+k-1}-\bar{u}_i} B_{i,k-1}(\bar{u}) ,$$

where the sum has been broken into two parts, and then the index has been shifted by one in the second summation.

Consider the quantity

$$\frac{\bar{u}_{i+k-1}-\bar{u}}{\bar{u}_{i+k-1}-\bar{u}_i} B_{i,k-1}(\bar{u})$$

which appears in the second sum. Suppose we set the index value to $i=0$. Then this quantity would become

$$\frac{\bar{u}_{k-1}-\bar{u}}{\bar{u}_{k-1}-\bar{u}_0} B_{0,k-1}(\bar{u})$$

But $\bar{u} \in [\bar{u}_{k-1}, \bar{u}_{m+1})$, which means that $\bar{u} \geq \bar{u}_{k-1}$. On the other hand, $B_{0,k-1}(\bar{u})$ has to be zero for $\bar{u} \geq \bar{u}_{k-1}$. This means that we can add this term to the first sum without changing the value of that sum. A similar argument shows that we can add the term

(more...)

$$\frac{\bar{u} - \bar{u}_{m+1}}{\bar{u}_{m+k} - \bar{u}_{m+1}} B_{m+1,k-1}(\bar{u})$$

to the second sum, since $B_{m+1,k-1}(\bar{u})$ will be zero for all values of $\bar{u} < \bar{u}_{m+1}$. This gives

$$\sum_{i=0}^{m} B_{i,k}(\bar{u}) = \sum_{i=0}^{m+1} \frac{\bar{u}_{i+k-1} - \bar{u}_i}{\bar{u}_{i+k-1} - \bar{u}_i} B_{i,k-1}(\bar{u}) = \sum_{i=0}^{m+1} B_{i,k-1}(\bar{u}) = 1$$

by the induction assumption.

## 10.6. Practical Implications

Now that we have generalized the uniform cubic B-splines to arbitrary order, let us see how the properties of corresponding curves generalize. (The material in the remainder of this section is taken from [Barsky83].)

### 10.6.1. B-splines of Different Order

We begin with a uniform cubic B-spline curve like the one shown in Figure 109. Because it requires four vertices — and basis functions — to define a segment, there are three fewer segments than there are control vertices in a cubic B-spline curve; for a curve of order $k$ there are $k-1$ fewer segments than control vertices. If we increase the order of the B-splines for a fixed set of control vertices we therefore reduce the number of segments, and we have consequently placed a large number of initial and final vertices in the control polygon for Figure 109.

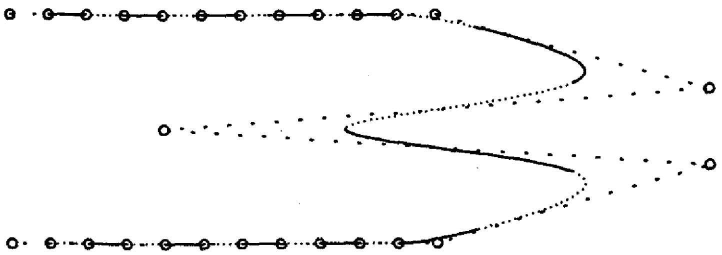


Figure 109. A uniform cubic (order 4) B-spline curve. The control vertices are circled.

Figure 110 illustrates what happens when we use tenth order B-splines. There are six fewer segments than in the fourth order curve of Figure 109: three fewer at the beginning and three fewer at the end. Also, the curve "oscillates less"; the influence of a given control vertex on any particular segment has been reduced. This follows from the fact that there are more vertices influencing the segment, since each B-spline has larger support. Each segment also lies within the convex hull of ten control vertices now instead of four.

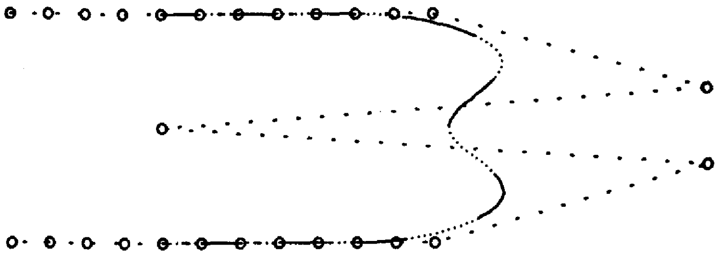


Figure 110. A uniform B-spline curve of order 10.

Conversely, if we reduce the order of the B-splines then each vertex influences fewer segments; however, its influence on these segments is stronger. For second order B-splines each segment is determined by two control vertices. Since it must lie within the convex hull of the two vertices, and the basis functions go to zero at either end, the segment is simply a straight line from the first vertex to the second. Figure 111 illustrates this.

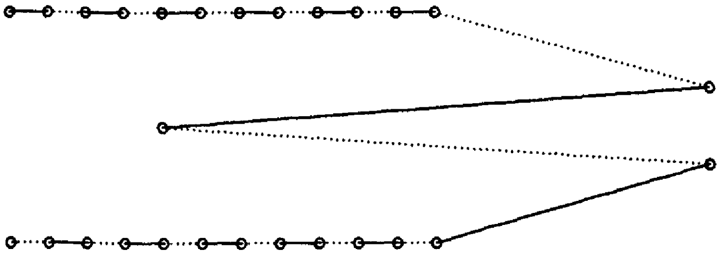


Figure 111. A uniform B-spline curve of order 2 (the control polygon is not shown). A B-spline curve of order 1 would consist simply of the control vertices.

To facilitate comparison, we show several curves of differing order in the following figure.

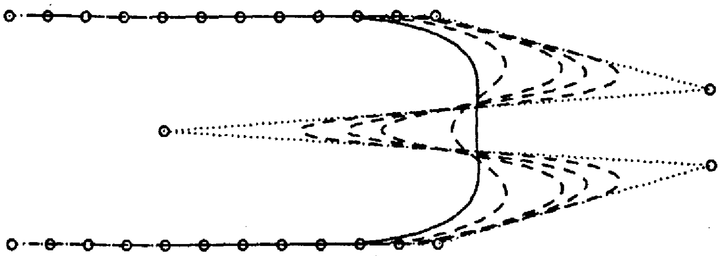


Figure 112. Uniform B-spline curves of orders 2, 3, 4, 5, 10 and 20 for the same control vertices.

The second order curve is shown with a dotted line, and connects the control vertices with straight line segments. The curve of order 20 is shown with a solid line. Intermediate curves are drawn dashed.

Higher order curves are, of course, more expensive to compute.

### 10.6.2. Multiple Knots

In a B-spline of order $k$ we may usefully associate a breakpoint with at most $k$ knots. Figure 113 illustrates why, as do Figures 99 and 100.

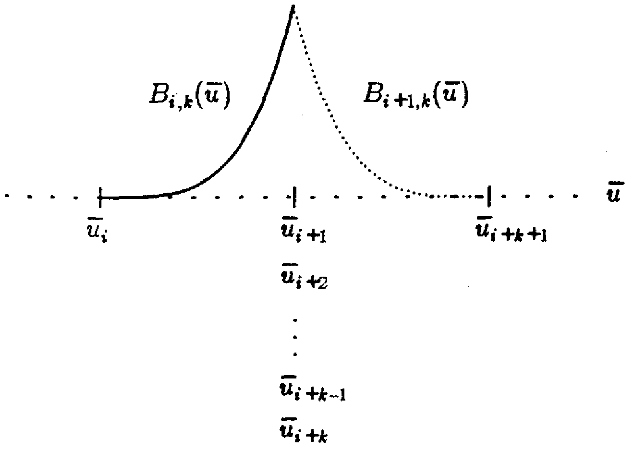


Figure 113. Since $\bar{u}_{i+1}$ is a knot of multiplicity $k$, $B_{i,k}(\bar{u})$ and $B_{i+1,k}(\bar{u})$ span only one non-vacuous interval each, as shown. All other basis functions must be zero at $\bar{u}_{i+1}$. $B_{i,k}(\bar{u})$ approaches the values 1 from the left as $\bar{u}$ approaches $\bar{u}_{i+1}$, and $B_{i+1,k}(\bar{u})$ must attain the value one at $\bar{u}_{i+1}$. Consequently $\mathbf{V}_{i+1}$ will be, and $\mathbf{V}_i$ will appear to be, interpolated at $\bar{u}_{i+1}$, and the curve $\mathbf{Q}(\bar{u})$ constructed from the control vertices $\{\ldots,\mathbf{V}_i,\mathbf{V}_{i+1},\ldots\}$ will, in general, be discontinuous.

Since the $j^{\text{th}}$ B-spline $B_{j,k}(\bar{u})$ spans the $k$ intervals from $\bar{u}_j$ to $\bar{u}_{j+k}$, there are only two B-splines whose support is associated with the breakpoint at $\bar{u}_{i+1}$. If $\bar{u}_{i+1}$ were of multiplicity $k+1$, then $B_{i+1,k}(\bar{u})$ would span the zero length interval from $\bar{u}_{i+1}$ to $\bar{u}_{i+k+1}$. (The right hand or dotted B-spline in Figure 113 would now be called $B_{i+2,k}(\bar{u})$.) Greater multiplicity would simply introduce additional vacuous B-splines.

For the breakpoint $\bar{u}=\bar{u}_{i+1}$ we have $C^{k-1-k}=C^{-1}$ continuity, i.e. no continuity at all. Moreover,

$$\lim_{\bar{u}\to\bar{u}_{i+1}} B_{i,k}(\bar{u}_{i+1}) = 1 \ .$$

Consider any infinitesimal step to the left of $\bar{u}_{i+1}$:

$$\bar{u}_{i+1} - \epsilon \quad \text{for } \epsilon>0 \ ,$$

and consider the B-splines,

$$B_{i-1,k}(\bar{u}), \ldots, B_{i-k+1,k}(\bar{u}) \ ,$$

that have support on $[\bar{u}_i,\bar{u}_{i+1})$ and that, together with $B_{i,k}(\bar{u})$, contribute to a curve segment $\mathbf{Q}(\bar{u})$ for $\bar{u}$ in this interval. Firstly, we will have

$$B_{i,k}(\bar{u}_{i+1}-\epsilon) \approx 1 \ .$$

The next B-spline to the left, $B_{i-1,k}(\bar{u})$, has a knot of multiplicity $k-1$ at $\bar{u}_{i+1}$, and is consequently $C^0$ there. Since

$$B_{i-1,k}(\bar{u}_{i+1}) = B_{i-1,k}(\bar{u}_{i+k}) = 0 \ ,$$

we will have

$$B_{i-1,k}(\bar{u}_{i+1}-\epsilon) \approx 0 \ .$$

Similarly

$$B_{i-k+1,k}(\bar{u}_{i+1}-\epsilon) \approx \cdots \approx B_{i-2,k}(\bar{u}_{i+1}) \approx 0 \ .$$

Since

$$B_{i-k+1,k}(\bar{u}_{i+1}-\epsilon) + \cdots + B_{i,k}(\bar{u}_{i+1}-\epsilon) = 1$$

for all small, positive values of $\epsilon$, we must be able to make $\mathbf{Q}(\bar{u}_{i+1}-\epsilon)$ arbitrarily close to

$$\mathbf{V}_{i-k+1}B_{i-k+1,k}(\bar{u}_{i+1}-\epsilon) + \cdots + \mathbf{V}_i B_{i,k}(\bar{u}_{i+k}-\epsilon) \approx \mathbf{V}_i$$

as $\epsilon > 0$ is made arbitrarily close to zero. As $\bar{u}$ becomes equal to $\bar{u}_{i+1}$, $B_{i,k}(\bar{u})$ drops abruptly to zero and $B_{i+1,k}(\bar{u})$ jumps abruptly to one. Since $\mathbf{Q}(\bar{u}_{i+1})$ is given by

$$\mathbf{V}_{i-k+2}B_{i-k+2,k}(\bar{u}_{i+1}) + \cdots + \mathbf{V}_i B_{i,k}(\bar{u}_{i+1}) + \mathbf{V}_{i+1}B_{i+1,k}(\bar{u}_{i+1}) + \quad ,$$

and since all of these B-spline values are zero save $B_{i+1,k}(\bar{u}_{i+1})$, which is 1,

$$\mathbf{Q}(\bar{u}_{i+1}) = \mathbf{V}_{i+1} \ ,$$

and there will be a jump in the curve from $\mathbf{V}_i$ to $\mathbf{V}_{i+1}$ (assuming $\mathbf{V}_i \neq \mathbf{V}_{i+1}$).

A slight adaptation of this argument can also be used to show that a knot of multiplicity $k-1$ will result in a positionally continuous B-spline curve $\mathbf{Q}(\bar{u})$ which interpolates $\mathbf{V}_i$. This is because $B_{i,k}(\bar{u})$ will be a function like that shown in Figure 114 which rises to one and then falls back to zero continuously, while all "surrounding" B-splines whose intervals of support are associated with the knot $\bar{u}_{i+1}$ will have the value zero at $\bar{u}_{i+1}$.

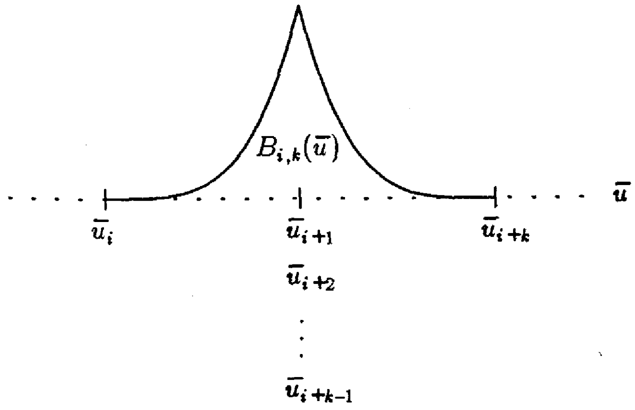


Figure 114. Since $\bar{u}_{i+1}$ is a knot of multiplicity $k-1$, $B_{i,k}(\bar{u})$ is the only B-spline among $B_{i-k+1}(\bar{u})$, $B_{i-k+2}(\bar{u})$, lated in this situation.

### 10.6.3. Collinear Control Vertices

If $k$ successive control vertices are collinear then they define a straight line segment. This follows easily from the convex hull property. If the knot sequence is such as to require the adjoining segments to meet this straight line with $C^2$ continuity (i.e. the breakpoint corresponding to that joint has multiplicity at most $k-3$) then the ends of those segments must have zero curvature there. Hence if a control polygon ends in $k-1$ collinear control vertices and the terminating breakpoint has multiplicity at most

$k-3$ (one for a cubic) then the curve will end with zero curvature.

### 10.6.4. Multiple Vertices

Just as we can repeat values in the knot sequence underlying a B-spline curve, so we can repeat vertices in the control polygon. First let us recall that the $i^{th}$ segment in a B-spline curve of order $k$ is defined by control vertices $\mathbf{V}_{i-k+1}, \cdots, \mathbf{V}_i$. Furthermore, since $B_{i-k+1}(\bar{u})$ is zero at $\bar{u}_{i+1}$, $\mathbf{V}_{i-k+1}$ does not affect the last point $\mathbf{Q}_i(\bar{u}_{i+1})$ of the $i^{th}$ segment. $\mathbf{Q}_i(\bar{u}_{i+1})$ is therefore entirely determined by $\mathbf{V}_{i-k+2}$, $\cdots, \mathbf{V}_i$. If these $k-1$ vertices are identical we have

$$\begin{aligned}\mathbf{Q}_i(\bar{u}_{i+1}) &= \sum_{r=-k+1}^{0} \mathbf{V}_{i+r} B_{i+r,k}(\bar{u}_{i+1}) = \sum_{r=-k+2}^{0} \mathbf{V}_{i+r} B_{i+r,k}(\bar{u}_{i+1}) \\ &= \mathbf{V}_i \sum_{r=-k+1}^{0} B_{i+r,k}(\bar{u}_{i+1}) = \mathbf{V}_i\end{aligned}$$

Thus a control vertex of multiplicity $k-1$ is interpolated, regardless of the knot sequence at hand. Moreover, $\mathbf{Q}_i(\bar{u})$ is guaranteed to be a straight line segment, since in this case we can factor the equation for $\mathbf{Q}_i(\bar{u}_i)$ as

$$\begin{aligned}&\mathbf{V}_{i-k+1} B_{i-k+1,k}(\bar{u}) + \mathbf{V}_i \sum_{r=-k+2}^{0} B_{i+r,k}(\bar{u}) \\ &= \mathbf{V}_{i-k+1} B_{i-k+1,k}(\bar{u}) + \mathbf{V}_i \Big(1 - B_{i-k+1,k}(\bar{u})\Big)\end{aligned}$$

which is a convex combination of $\mathbf{V}_{i-k+1}$ and $\mathbf{V}_i$ and so defines a straight line segment. $\mathbf{V}_{i-k+1}$ and $\mathbf{V}_i$ are not, in general, interpolated.

Unless the knot $\bar{u}_{i-1}$ has multiplicity $k-2$ or greater, the previous curve segment, namely $\mathbf{Q}_{i-1}(\bar{u})$, will be at least $C^2$ continuous with $\mathbf{Q}_i(\bar{u})$ at $\bar{u}_{i-1}$, and must therefore have a zero second derivative and zero curvature there since $\mathbf{Q}_i(\bar{u})$ is a straight line segment. Moreover, $\mathbf{Q}_{i-1}(\bar{u})$ must terminate somewhere on the line segment connecting $\mathbf{V}_{i-k+1}$ and $\mathbf{V}_{i-k+2} = \cdots = \mathbf{V}_i$. A similar argument establishes that $\mathbf{Q}_{i+1}(\bar{u})$ begins with zero curvature at a point lying between $\mathbf{V}_{i-k+2} = \cdots = \mathbf{V}_i$ and $\mathbf{V}_{i+1}$.

### 10.6.5. End Conditions

Our description of multiple knots and multiple vertices is applicable to any part of the curve, but is particularly useful in controlling the behaviour at the beginning and end of a curve. In general the most we can say about these endpoints is that they lie within the convex hull of the first and last $k-1$ control vertices, respectively. From earlier remarks it follows that either an initial knot of multiplicity $k-1$, or an initial control vertex of multiplicity $k-1$, will cause the curve to interpolate the first control vertex. In the latter case the first curve segment is a short straight line. Moreover, an initial vertex of multiplicity $k-2$ will cause the curve to begin somewhere on the line segment joining the first and second control vertices with zero curvature. The end of a curve may be similarly controlled. A more detailed discussion may be found in [Barsky82].

It is actually quite common to ensure interpolation of the first and last control vertices by giving them multiplicity $k-1$ and giving the first and last breakpoints multiplicity $k-1$ as well, although this is unnecessarily restrictive.

# 11. Be'zier Curves

Be'zier curves and surfaces [Bézier70, Bézier77, Forrest72] were one of the earliest attempts to develop a flexible and intuitive interface for computer aided design, and have been used for some years by Renault-Peugeot to design the "skin" or outer panels of automobiles [Be'zier74]. They are interesting in their own right, relate naturally to B-spline curves and provide a convenient context in which to introduce the idea of "subdivision."

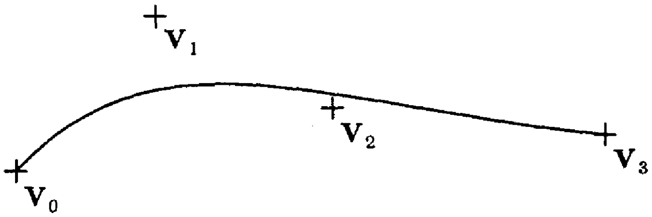


Figure 115. A cubic Be'zier curve. The endpoints are interpolated.

A degree $d$ Be'zier curve is defined, much like a B-spline curve, as

$$\mathbf{Q}(u) = \sum_{i=0}^{d} \mathbf{V}_i P_{i,d}(u) \tag{61}$$

for $0 \le u \le 1$, where

$$P_{i,d}(u) = \begin{pmatrix} d \\ i \end{pmatrix} u^i (1-u)^{d-i} \tag{62}$$

are the *Bernstein polynomials*. Using the Binomial Theorem it is is easy to show that a Be'zier curve lies within the convex hull of its defining control vertices. First we write

$$\begin{aligned} 1 &= \Big[ (1-u) + u \Big]^d = \sum_{i=0}^{d} \begin{pmatrix} d \\ i \end{pmatrix} u^i (1-u)^{d-i} \\ &= (1-u)^d + du(1-u)^{d-1} + \cdots + du^{d-1}(1-u) + u^d \\ &= P_{0,d}(u) + P_{1,d}(u) + \cdots + P_{d-1,d}(u) + P_{d,d}(u) \ . \end{aligned} \tag{63}$$

Thus the $P_{i,d}(u)$ sum to one. Because $0 \le u \le 1$ the quantities $u$ and $(1-u)$ are both nonnegative; it follows that the $P_{i,d}(u)$ are also nonnegative. Therefore a Be'zier curve must lie within the convex hull of its control vertices.

It is a fact, although we shall not bother to prove it, that the Bernstein polynomials of degree $d$ are a basis for the polynomials of degree $d$. **[ref]**

By way of example, for cubic Be'zier curves we have

$$\begin{aligned}\mathbf{Q}(u) &= \mathbf{V}_0 P_{0,3} + \mathbf{V}_1 P_{1,3} + \mathbf{V}_2 P_{2,3} + \mathbf{V}_3 P_{3,3} \\ &= \mathbf{V}_0(1-u)^3 + \mathbf{V}_1 3u(1-u)^2 + \mathbf{V}_2 3u^2(1-u) + \mathbf{V}_3 u^3 \ .\end{aligned}$$

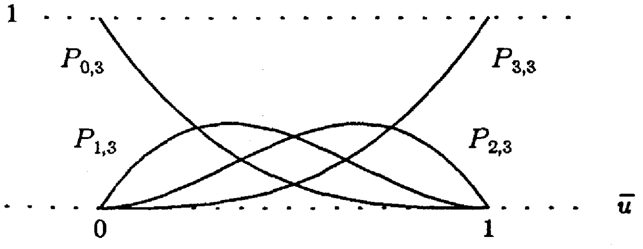


Figure 116. The cubic Be'zier basis functions.

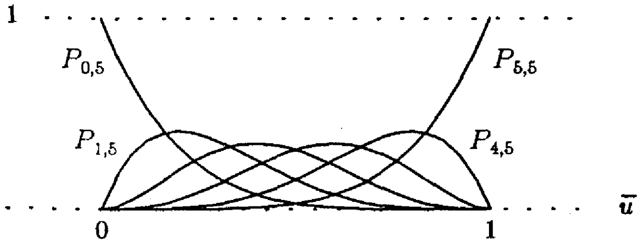


Figure 117. The quintic (degree 5) Be'zier basis functions.

By inspection it is easy to see from (61) and (62) that $\mathbf{Q}(0)=\mathbf{V}_0$ and $\mathbf{Q}(1)=\mathbf{V}_d$ — the first and last control vertices are interpolated.

### 11.1. Increasing the Degree of a Be'zier Curve

Suppose that we are unable to produce a curve of the desired shape with a degree $d$ Be'zier curve. One option is simply to use a Be'zier curve of higher degree. Having chosen to work at a higher degree, we may simply define a new curve. On the other hand, a polynomial of degree $d$ is also a polynomial of degree $d+1$; hence there exists a set of $d+2$ control vertices $\mathbf{W}_i$ which define a degree $d$ Be'zier curve originally defined by $d+1$ control vertices $\mathbf{V}_i$. The relationship between the $\mathbf{V}_i$ and the $\mathbf{W}_i$, which appears in [Forrest72, Ramshaw85], is given by the following formulae.

$$\begin{aligned}\mathbf{W}_0 &= \mathbf{V}_0 \\ \mathbf{W}_i &= \left(\frac{i}{d+1}\right)\mathbf{V}_{i-1} + \left(1-\frac{i}{d+1}\right)\mathbf{V}_i \qquad \text{for } 0 < i < d+1 \\ \mathbf{W}_{d+1} &= \mathbf{V}_d \ .\end{aligned}$$

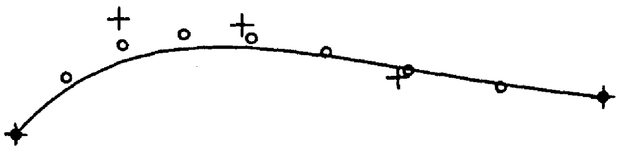


Figure 118. This is the curve of Figure 115, defined as a degree 4 Be'zier curve ("+"'s) and as a degree 8 Be'zier curve ("o"'s).

### 11.2. Composite Be'zier Curves

Using a higher degree Be'zier curve gives us more flexibility, but it also increases the cost of evaluation. Then too, the movement of any one control vertex still alters the entirety of a simple Be'zier curve. An alternative is to construct a composite curve from several simple Be'zier curves by causing the last vertex of the $i^{th}$ segment to coincide with the first vertex of the $(i+1)^{st}$ segment. Since the first and last vertices of a Be'zier curve are interpolated, this results in $C^0$ continuity. Differentiating (61) we see that

$$\mathbf{Q}^{(1)}(0) = d(\mathbf{V}_1-\mathbf{V}_0) \tag{64}$$

and

$$\mathbf{Q}^{(1)}(1) = d(\mathbf{V}_d-\mathbf{V}_{d-1}) \ , \tag{65}$$

so that the tangents at either end are collinear with the line segment between the first two and last two control vertices, respectively. Consecutive segments in a composite Be'zier curve can therefore be made $C^1$ continuous simply by arranging that the penultimate control vertex of the first curve, the shared endpoint, and the second vertex of the next curve be collinear and equally spaced. (see Figure 119).

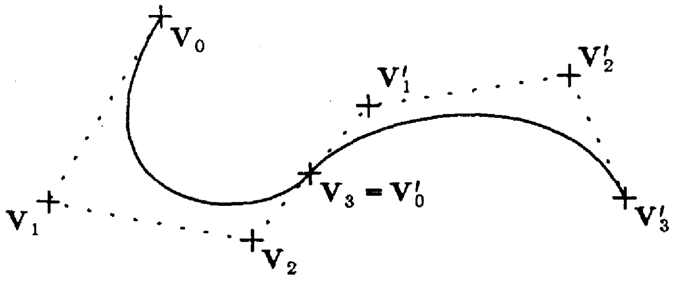


Figure 119. A composite cubic Be'zier curve. The unprimed vertices define one curve segment and the primed vertices define another. Because $\mathbf{V}_2$, $\mathbf{V}_3$=$\mathbf{V}'_0$ and $\mathbf{V}'_1$ are collinear and $|\mathbf{V}_3-\mathbf{V}_2|=|\mathbf{V}'_1-\mathbf{V}'_0|$ the composite curve will be $C^1$ continuous.

The second derivatives at the beginning and end of a simple Be'zier curve are given by

$$\mathbf{Q}^{(2)}(0) = d(d-1)\Big[(\mathbf{V}_2-\mathbf{V}_1) - (\mathbf{V}_1-\mathbf{V}_0)\Big]$$

and

$$\mathbf{Q}^{(2)}(1) = d(d-1)\Big[(\mathbf{V}_{d-2}-\mathbf{V}_{d-1}) - (\mathbf{V}_{d-1}-\mathbf{V}_d)\Big] \ .$$

Suppose that we want to link two Be'zier curves together with $C^2$ continuity. If the first segment is defined by the control vertices $\mathbf{V}_0$, $\mathbf{V}_1$, ... $\mathbf{V}_d$, then: the position of the first control vertex in the second segment is fixed at $\mathbf{V}_d$ by the requirement that the segments join; the requirement for first derivative continuity fixes the position of the second vertex at $\mathbf{V}_d+(\mathbf{V}_d-\mathbf{V}_{d-1})$; and the requirement for second derivative continuity fixes the position of the third vertex at $\mathbf{V}_{d-2}+4(\mathbf{V}_{d-1}-\mathbf{V}_d)$. Thus if we are dealing with cubic Bézier curves, when we add another curve segment we are free to position only the last of the control vertices for the new segment. The positions of the first three control vertices are fixed by the requirement of $C^2$ continuity. The use of higher degree curves leaves more of the internal control vertices for this segment free of constraints. Requiring higher degree continuity imposes constraints on additional control vertices neighboring each joint.

If we move the joint between two segments in a composite cubic Bézier curve, the above considerations tell us that if we demand only $C^1$ continuity then only the segments which meet at the joint in questions are altered; only the control vertices on either side of the joint must be moved (so as to remain collinear with the joint), and neither affects $C^1$ continuity at neighbouring joints.

If we insist on maintaining $C^2$ continuity of a cubic Bézier curve when moving a joint then the two neighboring control vertices on at least one side will need to be moved. Since these in turn determine the second derivative at the next joint, the change may ripple the entire length of the curve. We must work with fifth degree Bézier curves to localize (to two segments) the changes needed to maintain $C^2$ continuity.

Similar observations obtain from considering the movement of other control vertices in a composite cubic Bézier curve.

### 11.3. Local versus Global Control

When comparing Bézier and B-spline curves it is sometimes said that the former exhibit "global control" while the latter exhibit "local control." This is true only in the following sense. When drawing Bézier curves people very often make use of a single segment, adding control vertices (and consequently raising the degree of the curve) when they need more control or when a lower degree is unable to represent the shape they desire. Each of the control vertices then affects the entire curve — its effect is "global."

When drawing B-splines, one usually makes use of composite curves because there is no need to worry about satisfying constraints among the control vertices in order to maintain continuity. In such a curve moving a given control vertex alters only a part of the curve — the effect is "local."

Technically speaking, however, this is misleading. If we restrict ourselves to a curve consisting of a single segment then moving any control vertex alters the entire segment for both Bézier and B-spline curves. If we look at composite curves, then in either case moving a single control vertex will affect only part of the curve (if Bézier curves of sufficiently high degree are used, as discussed above).

### 11.4. Subdivision of Bézier Curves

Now let us ask ourselves the following question: can we find an easy way to break a cubic Bézier curve in half? That is, suppose we have the Bézier curve

$$\mathbf{Q}(u) = \mathbf{V}_0(1-u)^3 + \mathbf{V}_1 3u(1-u)^2 + \mathbf{V}_2 3u^2(1-u) + \mathbf{V}_3 u^3 \qquad \text{for } 0 \leq u \leq 1 \ .$$

Can we find control vertices

$$\mathbf{S}_0\,,\ \mathbf{S}_1\,,\ \mathbf{S}_2\,,\ \mathbf{S}_3\,,\ \mathbf{T}_0\,,\ \mathbf{T}_1\,,\ \mathbf{T}_2\,,\ \mathbf{T}_3$$

such that the Bézier curve

$$\mathbf{L}(s) = \mathbf{S}_0(1-s)^3 + \mathbf{S}_1 3s(1-s)^2 + \mathbf{S}_2 3s^2(1-s) + \mathbf{S}_3 s^3 \qquad \text{for } 0 \leq u \leq 1$$

is the first half of the curve defined by $\mathbf{V}_0$, $\mathbf{V}_1$, $\mathbf{V}_2$ and $\mathbf{V}_3$ (i.e. $\mathbf{Q}(u)$ for $0 \le u \le 0.5$), and

$$\mathbf{R}(t) = \mathbf{T}_0(1-t)^3 + \mathbf{T}_1 3t(1-t)^2 + \mathbf{T}_2 3t^2(1-t) + \mathbf{T}_3 t^3 \qquad \text{for } 0 \le u \le 1$$

is the second half of the curve defined by $\mathbf{V}_0$, $\mathbf{V}_1$, $\mathbf{V}_2$ and $\mathbf{V}_3$ (i.e. $\mathbf{Q}(u)$ for $0.5 \le u \le 1$)?

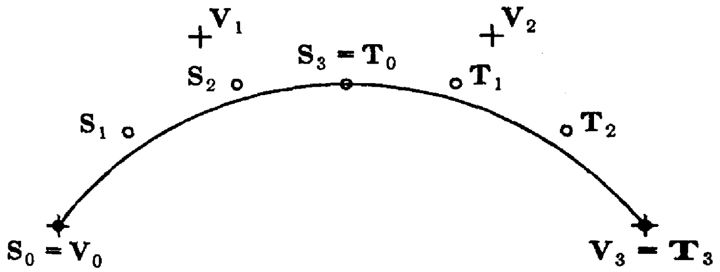


Figure 120. Subdivision of a cubic Bézier curve. The original control vertices $\mathbf{V}_0$, $\mathbf{V}_1$, $\mathbf{V}_2$ and $\mathbf{V}_3$ are represented with a "+". The new control vertices are represented with an "o".

Not surprisingly (since we bothered to ask the question), the answer is yes. There are two reasons why the question is interesting:

- doing so doubles the number of control vertices, "halving" the size of each segment and decreasing the extent of the curve which is affected by the movement of a single control vertex;
- it turns out that the resulting control vertices lie closer to the curve than the original control vertices, and one method of rendering the curve is to repeatedly subdivide until the control vertices are so close to the curve that the control polygon is an adequate approximation of the curve – and moreover the subdivision can be applied adaptively a greater number of times in regions of relatively high curvature.

We illustrate these points by considering the simple case of midpoint subdivision for cubic Bézier curves. We know that

$$\mathbf{S}_0 = \mathbf{V}_0$$

$$\mathbf{S}_3 = \mathbf{Q}(\tfrac{1}{2}) = \tfrac{1}{8}\Big(\mathbf{V}_0 + 3\mathbf{V}_1 + 3\mathbf{V}_2 + \mathbf{V}_3\Big) \ .$$

From (64) and (65) we know that

$$\mathbf{L}^{(1)}(0) = 3(\mathbf{S}_1 - \mathbf{S}_0)$$

and

$$\mathbf{L}^{(1)}(1) = 3(\mathbf{S}_3 - \mathbf{S}_2) \ .$$

Since we have $s = 2u$, by the chain rule we have

$$\frac{d}{du}\mathbf{L}(s(u)) = \frac{d}{ds}\mathbf{L}(s) \cdot \frac{d}{du}s(u) = 2\mathbf{L}^{(1)}(u)$$

whence

$$\mathbf{L}^{(1)}(0) = \frac{1}{2}\,\mathbf{Q}^{(1)}(0) = \frac{3}{2}(\mathbf{V}_1 - \mathbf{V}_0)$$

and

$$L^{(1)}(1) = \frac{1}{2}\,Q^{(1)}(\tfrac{1}{2}) = \frac{3}{8}\left(V_3+V_2-V_1-V_0\right) .$$

We now have four equations, namely

$$\begin{aligned} S_0 &= V_0 \\ 3(S_1-S_0) &= \tfrac{3}{2}(V_1-V_0) \\ 3(S_3-S_2) &= \tfrac{3}{8}\left(V_3+V_2-V_1-V_0\right) \\ S_3 &= \tfrac{1}{8}\left(V_0+3V_1+3V_2+V_3\right) . \end{aligned}$$

Solving them yields

$$\begin{aligned} S_0 &= V_0 \\ S_1 &= \tfrac{1}{2}\,(V_0+V_1) \\ S_2 &= \tfrac{1}{4}\,(V_0+2V_1+V_2) \\ S_3 &= \tfrac{1}{8}\,(V_0+3V_1+3V_2+V_3) . \end{aligned}$$

In a completely analogous way we can show that

$$\begin{aligned} T_0 &= \tfrac{1}{8}\,(V_0+3V_1+3V_2+V_3) \\ T_1 &= \tfrac{1}{4}\,(V_1+2V_2+V_3) \\ T_2 &= \tfrac{1}{2}\,(V_2+V_3) \\ T_3 &= V_3 . \end{aligned}$$

These vertices are more efficiently computed in the following order:

$$\begin{aligned} S_0 &= V_0 \\ S_1 &= \tfrac{1}{2}(V_0+V_1) \\ t &= \tfrac{1}{2}(V_1+V_2) \\ S_2 &= \tfrac{1}{2}(S_1+t) \\ T_3 &= V_3 \\ T_2 &= \tfrac{1}{2}(V_2+V_3) \\ T_1 &= \tfrac{1}{2}(t+T_2) \\ S_3 &= T_0 = \tfrac{1}{2}(S_2+T_1) . \end{aligned}$$

Again, there are two reasons why this process is interesting. Firstly, it can be shown that the new control polygons lie closer to the curve than the original control polygon [Lane80]. Thus one technique for rendering a curve is to continue subdividing until the control polygons are a sufficiently good approximation to the curve and then simply draw the control polygon.

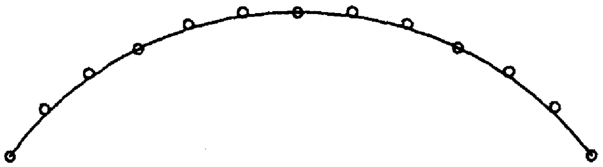


Figure 121. A second level of subdivision applied to Figure 120. There are four cubic Bézier curves here.

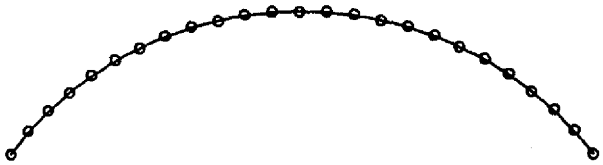


Figure 122. A third level of subdivision. There are eight cubic Bézier curves here.

Because a Bézier curve lies within the convex hull of its defining control vertices, one may test to see if the length of the control polygon is within some tolerance of the distance between the first and last control vertices [Blinn80], or whether the distance between each pair of control vertices is less than some tolerance, or whether the deviation of internal control vertices from a line segment joining the end vertices is sufficiently small [Lane80], etc. The convergence test can be applied to each subdivided curve individually, so that the subdivision process ceases adaptively when the curve has become "locally flat."

Secondly, subdivision can aid in the design of a curve since it provides more control vertices, whose movement affects the shape of a smaller portion of the curve. Of course one must be careful not to destroy the desired continuity at joints when moving control vertices; as we have seen, this can be a fairly severe restriction unless sufficiently high degree Bézier curves are used.

Midpoint subdivision of higher order Bézier curves, having degree $d$, can be accomplished using the formula

$$\mathbf{S}_i = \sum_{r=0}^{i} \binom{i}{r} \frac{\mathbf{V}_r}{2^i} \qquad \text{for } i = 0, 1, \ldots, d$$

derived in [Clark79]. From symmetry we have

$$\mathbf{T}_i = \sum_{r=i}^{d} \binom{d-i}{d-r} \frac{\mathbf{V}_r}{2^{d-i}} \qquad \text{for } i = 0, 1, \ldots, d \ .$$

Clark's proof makes use of the Binomial Theorem and various binomial identities. An induction proof of essentially the same result is given in [Lane80], where the following algorithm for efficiently computing the $\mathbf{S}_i$ and $\mathbf{T}_i$ is presented as well.

```
for $i \leftarrow 0$ step 1 until $d$ do
    $\mathbf{S}_i \leftarrow \mathbf{V}_i$
endfor
$\mathbf{R}_d \leftarrow \mathbf{V}_d$
for $j \leftarrow 1$ step 1 until $d$ do
    tmp2 $\leftarrow \mathbf{S}_{j-1}$
    for $k \leftarrow j$ step 1 until $d$ do
        tmp1 $\leftarrow$ tmp2
        tmp2 $\leftarrow \frac{1}{2}(\mathbf{S}_{k-1}+\mathbf{S}_k)$
        $\mathbf{S}_{k-1} \leftarrow$ tmp1
    endfor
    $\mathbf{S}_n \leftarrow \mathbf{T}_{n-j} \leftarrow$ tmp2
endfor
```

A general technique for directly subdividing elsewhere than at the parametric midpoint of a Bézier curve is given in [Barsky85].

The midpoint subdivision of uniform B-spline curves is also discussed in [Lane80]. This is a special case of the "Oslo Algorithm" developed subsequently by Cohen, Lyche and Riesenfeld, to which we will turn in the next chapter.

### 11.5. Bézier Curves From B-splines

There is an interesting connection between the Bernstein polynomials and the B-splines: the B-splines of order $d+1$ over a knot sequence in which each breakpoint has multiplicity $d+1$ are exactly the Bernstein polynomials of degree $d$. It is easiest to convey the idea of this connection by considering the simple knot sequence

| $\bar{u}_0$ | $\bar{u}_1$ | $\bar{u}_2$ | $\bar{u}_3$ | $\bar{u}_4$ | $\bar{u}_5$ | $\bar{u}_6$ | $\bar{u}_7$ |
|---|---|---|---|---|---|---|---|
| 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 . |

Our claim is that the four B-splines $B_{0,4}(\bar{u})$, $B_{1,4}(\bar{u})$, $B_{2,4}(\bar{u})$ and $B_{3,4}(\bar{u})$ over this knot sequence are exactly the Bernstein polynomials $P_{0,3}(\bar{u})$, $P_{1,3}(\bar{u})$, $P_{2,3}(\bar{u})$ and $P_{3,3}(\bar{u})$ given by (62).

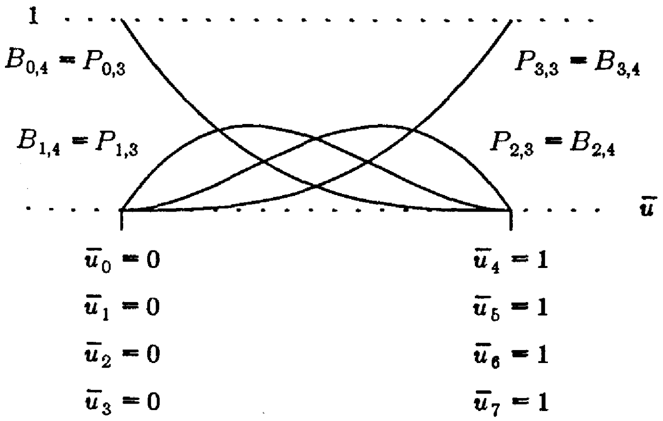
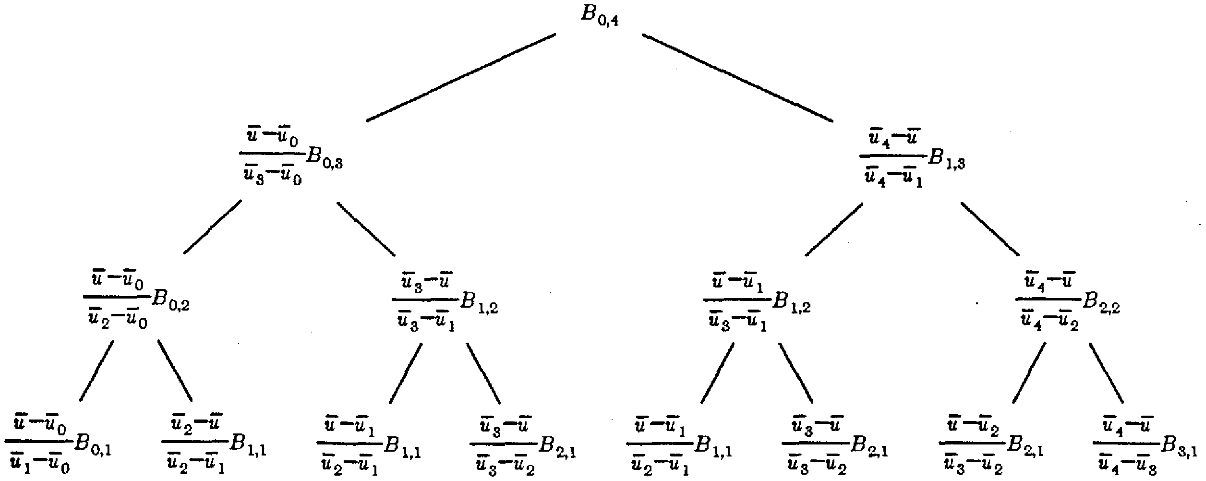


Figure 123.

Recall that

$$B_{i,k}(\bar{u}) = \frac{\bar{u} - \bar{u}_i}{\bar{u}_{i+k-1} - \bar{u}_i} B_{i,k-1}(\bar{u}) + \frac{\bar{u}_{i+k} - \bar{u}}{\bar{u}_{i+k} - \bar{u}_{i+1}} B_{i+1,k-1}(\bar{u}) \ . \tag{66}$$

Let us expand the B-spline $B_{0,4}(\bar{u})$, which has support $(\bar{u}_0,\bar{u}_4)$, all the way down to the $B_{i,1}(\bar{u})$.

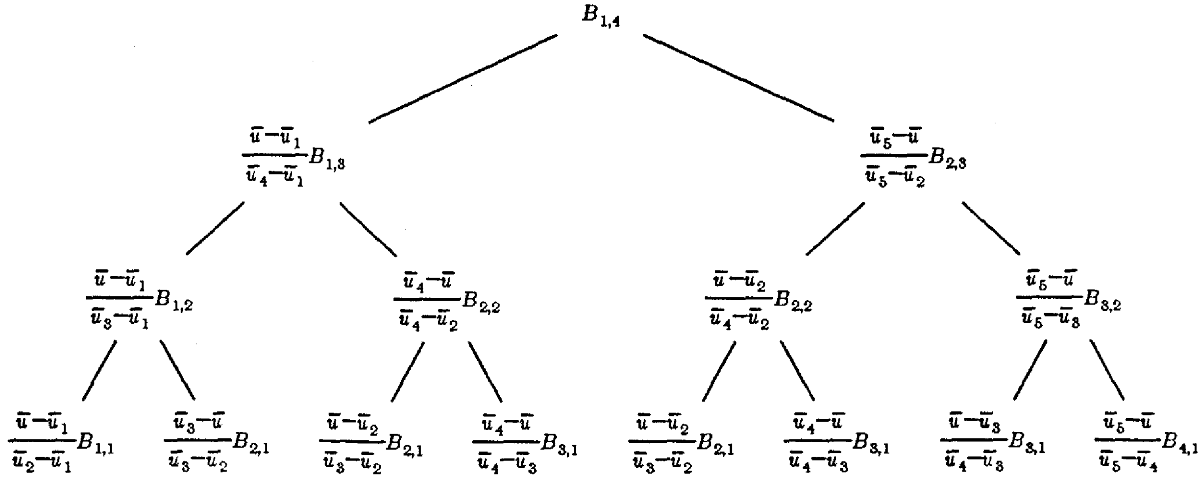

$B_{0,4}$
$\frac{\bar{u}-\bar{u}_0}{\bar{u}_3-\bar{u}_0} B_{0,3}$
$\frac{\bar{u}_4-\bar{u}}{\bar{u}_4-\bar{u}_1} B_{1,3}$
$\frac{\bar{u}-\bar{u}_0}{\bar{u}_2-\bar{u}_0} B_{0,2}$
$\frac{\bar{u}_3-\bar{u}}{\bar{u}_3-\bar{u}_1} B_{1,2}$
$\frac{\bar{u}-\bar{u}_1}{\bar{u}_3-\bar{u}_1} B_{1,2}$
$\frac{\bar{u}_4-\bar{u}}{\bar{u}_4-\bar{u}_2} B_{2,2}$
$\frac{\bar{u}-\bar{u}_0}{\bar{u}_1-\bar{u}_0} B_{0,1}$
$\frac{\bar{u}_2-\bar{u}}{\bar{u}_2-\bar{u}_1} B_{1,1}$
$\frac{\bar{u}-\bar{u}_1}{\bar{u}_2-\bar{u}_1} B_{1,1}$
$\frac{\bar{u}_3-\bar{u}}{\bar{u}_3-\bar{u}_2} B_{2,1}$
$\frac{\bar{u}-\bar{u}_1}{\bar{u}_2-\bar{u}_1} B_{1,1}$
$\frac{\bar{u}_3-\bar{u}}{\bar{u}_3-\bar{u}_2} B_{2,1}$
$\frac{\bar{u}-\bar{u}_2}{\bar{u}_3-\bar{u}_2} B_{2,1}$
$\frac{\bar{u}_4-\bar{u}}{\bar{u}_4-\bar{u}_3} B_{3,1}$

The following points are of interest.

- Since we are evaluating a B-spline of order 4, the tree must have depth 3.
- The value of $B_{0,4}$ is a sum of eight terms. Each term is the product of the coefficients in one of the eight root-leaf paths of this tree.
- A leaf of this tree will be nonzero only if its denominator is $\bar{u}_4 - \bar{u}_3$. In this case only the rightmost leaf is nonzero.
- To arrive at a leaf we begin with the root $B_{0,4}$, which has support $(\bar{u}_0,\bar{u}_4)$ and proceed down the tree. Going left as we leave a node corresponds to following the left term of the recurrence (66), and removes one knot interval from the end of the support; going right as we leave a node corresponds to following the right term of the recurrence (66), and removes one knot interval from the beginning of the support. The goal is to prune the support down to the two knots $\bar{u}_3$ and $\bar{u}_4$ which immediately surround the only non-vacuous interval.

- The denominators along the path to a nonzero leaf are one since they must include $[\bar{u}_3,\bar{u}_4)$ and $\bar{u}_4-\bar{u}_3=1$. The numerator at a node is $\bar{u}$ if the node is entered by a left branch, and $(1-\bar{u})$ if the node is entered by a right branch.
- Since in this case we begin with $(\bar{u}_0,\bar{u}_4)$ and must end with $(\bar{u}_3,\bar{u}_4)$, it is clear that we must always take a rightmost branch, pruning $(\bar{u}_0,\bar{u}_1)$, $(\bar{u}_1,\bar{u}_2)$ and $(\bar{u}_2,\bar{u}_3)$ in succession. For future reference we record this tree as being one in which the nonzero leaf is reached by taking zero left branches, and that there are exactly $\binom{3}{0}=1$ such paths.
- Hence $B_{0,4}(\bar{u}) = \begin{pmatrix}3\\0\end{pmatrix}(1-\bar{u})^3 = (1-\bar{u})^3$ .

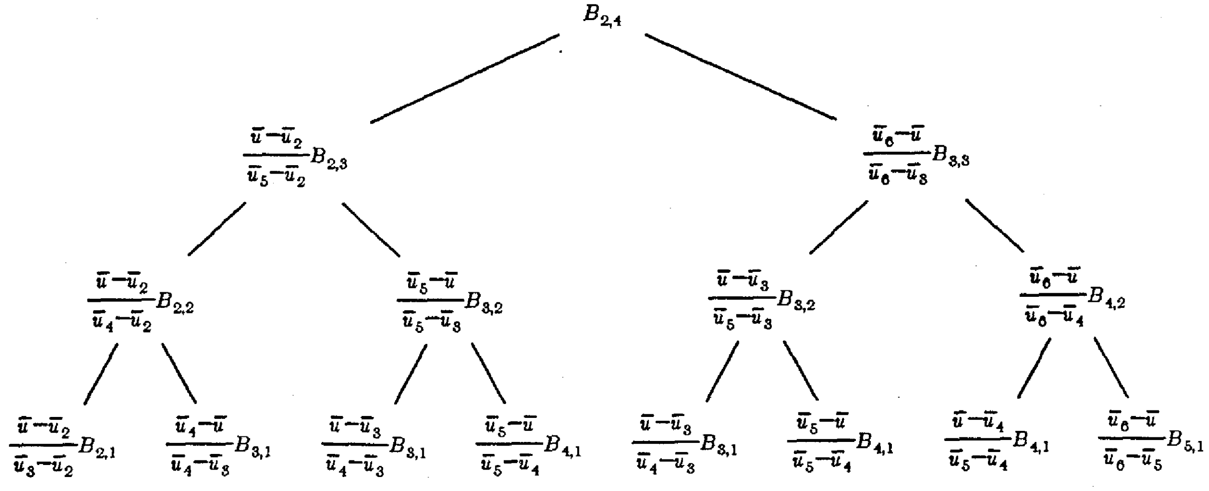
Now consider $B_{1,4}(\bar{u})$, which has support $(\bar{u}_1,\bar{u}_5)$.


$B_{1,4}$
$\frac{\bar{u}-\bar{u}_1}{\bar{u}_4-\bar{u}_1}B_{1,3}$
$\frac{\bar{u}_5-\bar{u}}{\bar{u}_5-\bar{u}_2}B_{2,3}$
$\frac{\bar{u}-\bar{u}_1}{\bar{u}_3-\bar{u}_1}B_{1,2}$
$\frac{\bar{u}_4-\bar{u}}{\bar{u}_4-\bar{u}_2}B_{2,2}$
$\frac{\bar{u}-\bar{u}_2}{\bar{u}_4-\bar{u}_2}B_{2,2}$
$\frac{\bar{u}_5-\bar{u}}{\bar{u}_5-\bar{u}_3}B_{3,2}$
$\frac{\bar{u}-\bar{u}_1}{\bar{u}_2-\bar{u}_1}B_{1,1}$
$\frac{\bar{u}_3-\bar{u}}{\bar{u}_3-\bar{u}_2}B_{2,1}$
$\frac{\bar{u}-\bar{u}_2}{\bar{u}_3-\bar{u}_2}B_{2,1}$
$\frac{\bar{u}_4-\bar{u}}{\bar{u}_4-\bar{u}_3}B_{3,1}$
$\frac{\bar{u}-\bar{u}_2}{\bar{u}_3-\bar{u}_2}B_{2,1}$
$\frac{\bar{u}_4-\bar{u}}{\bar{u}_4-\bar{u}_3}B_{3,1}$
$\frac{\bar{u}-\bar{u}_3}{\bar{u}_4-\bar{u}_3}B_{3,1}$
$\frac{\bar{u}_5-\bar{u}}{\bar{u}_5-\bar{u}_4}B_{4,1}$

In this case we must prune $(\bar{u}_1,\bar{u}_5)$ down to $(\bar{u}_3,\bar{u}_4)$ to obtain a nonzero leaf. The root-leaf paths that accomplish this are those that involve exactly one left branch, to reduce $\bar{u}_5$ to $\bar{u}_4$, and there are exactly $\binom{3}{1}=3$ such paths. Hence

$$B_{1,4}(\bar{u}) = \begin{pmatrix}3\\1\end{pmatrix}\bar{u}(1-\bar{u})^2 = 3\bar{u}(1-\bar{u})^2 \ .$$

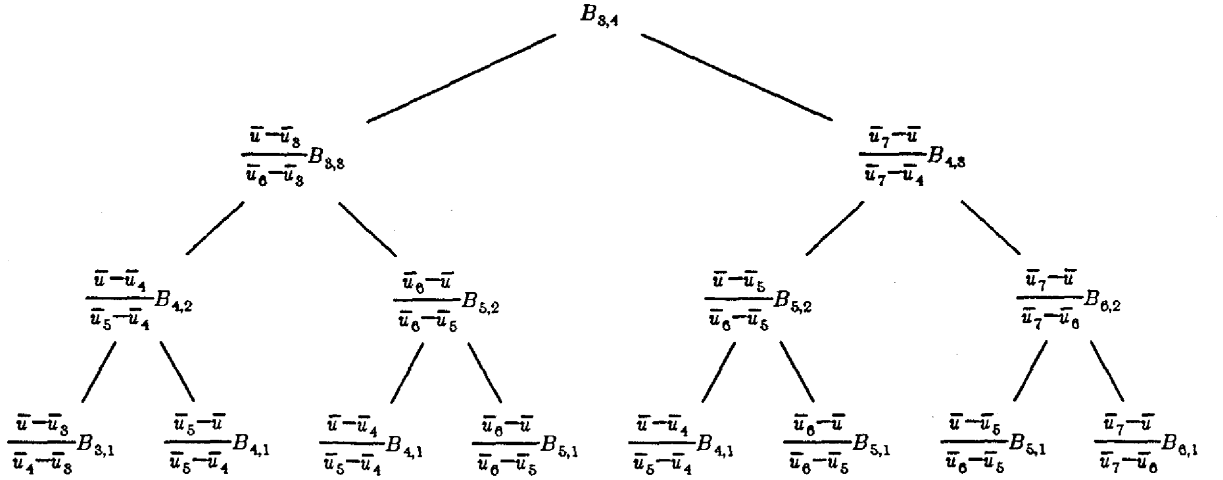
Next consider $B_{2,4}(\bar{u})$, which has support $(\bar{u}_2,\bar{u}_6)$.


$B_{2,4}$
$\frac{\bar{u}-\bar{u}_2}{\bar{u}_5-\bar{u}_2}B_{2,3}$
$\frac{\bar{u}_6-\bar{u}}{\bar{u}_6-\bar{u}_3}B_{3,3}$
$\frac{\bar{u}-\bar{u}_2}{\bar{u}_4-\bar{u}_2}B_{2,2}$
$\frac{\bar{u}_5-\bar{u}}{\bar{u}_5-\bar{u}_3}B_{3,2}$
$\frac{\bar{u}-\bar{u}_3}{\bar{u}_5-\bar{u}_3}B_{3,2}$
$\frac{\bar{u}_6-\bar{u}}{\bar{u}_6-\bar{u}_4}B_{4,2}$
$\frac{\bar{u}-\bar{u}_2}{\bar{u}_3-\bar{u}_2}B_{2,1}$
$\frac{\bar{u}_4-\bar{u}}{\bar{u}_4-\bar{u}_3}B_{3,1}$
$\frac{\bar{u}-\bar{u}_3}{\bar{u}_4-\bar{u}_3}B_{3,1}$
$\frac{\bar{u}_5-\bar{u}}{\bar{u}_5-\bar{u}_4}B_{4,1}$
$\frac{\bar{u}-\bar{u}_3}{\bar{u}_4-\bar{u}_3}B_{3,1}$
$\frac{\bar{u}_5-\bar{u}}{\bar{u}_5-\bar{u}_4}B_{4,1}$
$\frac{\bar{u}-\bar{u}_4}{\bar{u}_5-\bar{u}_4}B_{4,1}$
$\frac{\bar{u}_6-\bar{u}}{\bar{u}_6-\bar{u}_5}B_{5,1}$

Now we must prune $(\bar{u}_2,\bar{u}_6)$ down to $(\bar{u}_3,\bar{u}_4)$ to obtain a nonzero leaf. The root-leaf paths that accomplish this are those that involve exactly two left branches, thus reducing $\bar{u}_6$ to $\bar{u}_5$ and $\bar{u}_5$ to $\bar{u}_4$, and there are exactly $\binom{3}{2}=3$ such paths. Hence

$$B_{2,4}(\bar{u}) = \begin{pmatrix}3\\2\end{pmatrix}\bar{u}^2(1-\bar{u}) = 3\bar{u}^2(1-\bar{u}) \ .$$

Finally, consider $B_{3,4}(\bar{u})$, which has support $(\bar{u}_3,\bar{u}_7)$.


$B_{3,4}$
$\frac{\bar{u}-\bar{u}_3}{\bar{u}_6-\bar{u}_3}B_{3,3}$
$\frac{\bar{u}_7-\bar{u}}{\bar{u}_7-\bar{u}_4}B_{4,3}$
$\frac{\bar{u}-\bar{u}_4}{\bar{u}_5-\bar{u}_4}B_{4,2}$
$\frac{\bar{u}_6-\bar{u}}{\bar{u}_6-\bar{u}_5}B_{5,2}$
$\frac{\bar{u}-\bar{u}_5}{\bar{u}_6-\bar{u}_5}B_{5,2}$
$\frac{\bar{u}_7-\bar{u}}{\bar{u}_7-\bar{u}_6}B_{6,2}$
$\frac{\bar{u}-\bar{u}_3}{\bar{u}_4-\bar{u}_3}B_{3,1}$
$\frac{\bar{u}_5-\bar{u}}{\bar{u}_5-\bar{u}_4}B_{4,1}$
$\frac{\bar{u}-\bar{u}_4}{\bar{u}_5-\bar{u}_4}B_{4,1}$
$\frac{\bar{u}_6-\bar{u}}{\bar{u}_6-\bar{u}_5}B_{5,1}$
$\frac{\bar{u}-\bar{u}_4}{\bar{u}_5-\bar{u}_4}B_{4,1}$
$\frac{\bar{u}_6-\bar{u}}{\bar{u}_6-\bar{u}_5}B_{5,1}$
$\frac{\bar{u}-\bar{u}_5}{\bar{u}_6-\bar{u}_5}B_{5,1}$
$\frac{\bar{u}_7-\bar{u}}{\bar{u}_7-\bar{u}_6}B_{6,1}$

Now we must prune $(\bar{u}_3,\bar{u}_7)$ down to $(\bar{u}_3,\bar{u}_4)$ to obtain a nonzero leaf. The root-leaf paths which accomplish this are exactly those that involve three left branches, to reduce $\bar{u}_7$ to $\bar{u}_6$, $\bar{u}_6$ to $\bar{u}_5$ and $\bar{u}_5$ to $\bar{u}_4$, and there are exactly $\binom{3}{3}=1$ such paths. Hence

$$B_{3,4}(\bar{u}) = \begin{pmatrix}3\\3\end{pmatrix}\bar{u}^3 = \bar{u}^3 \ .$$

Summarizing, we have

$$B_{i,4}(\bar{u}) = \begin{pmatrix}3\\i\end{pmatrix}\bar{u}^i(1-\bar{u})^{3-i}$$

for $i=0,1,2,3$. By a slight generalization of this argument we have

$$B_{i,d+1}(\bar{u}) = \begin{pmatrix}d\\i\end{pmatrix}\bar{u}^i(1-\bar{u})^{d-i}$$

for $i=0,1,\cdots,d$ if the $B_{i,d+1}$ are defined over a knot sequence with uniformly spaced breakpoints of multiplicity $d+1$. We compare this with (62) and conclude that

$$B_{i,d+1}(\bar{u}) = P_{i,d}(\bar{u})$$

in these circumstances.

The Bernstein polynomials take the form

$$P_{i,d}(\bar{u}) = \begin{pmatrix}d\\i\end{pmatrix}\frac{(\bar{u}-a)^i(b-\bar{u})^{d-i}}{(b-a)^d}$$

if we are interested in the interval $[a,b)$ rather than $[0,1)$. We leave for the reader the exercise of verifying that for the knot sequence consisting of $(d+1)$ $a$'s followed by $(d+1)$ $b$'s we have

$$B_{i,d+1}(\bar{u}) = \begin{pmatrix} d \\ i \end{pmatrix} \frac{(\bar{u}-a)^i(b-\bar{u})^{d-i}}{(b-a)^d} = P_{i,d}(\bar{u}) \ .$$

## 11.6. A Matrix Formulation

It is common to use matrix notation in representing parametric curves. For example, the Bézier curve

$$\mathbf{Q}(\bar{u}) = \mathbf{U}_0(1-\bar{u})^3 + \mathbf{U}_1 3u(1-\bar{u})^2 + \mathbf{U}_2 3u^2(1-\bar{u}) + \mathbf{U}_3 u^3$$

can be written as

$$\mathbf{Q}(\bar{u}) = \begin{bmatrix} 1 & \bar{u} & \bar{u}^2 & \bar{u}^3 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ -3 & 3 & 0 & 0 \\ 3 & -6 & 3 & 0 \\ -1 & 3 & -3 & 1 \end{bmatrix} \begin{bmatrix} \mathbf{U}_0 \\ \mathbf{U}_1 \\ \mathbf{U}_2 \\ \mathbf{U}_3 \end{bmatrix} \qquad (67)$$

$$= \begin{bmatrix} 1 & \bar{u} & \bar{u}^2 & \bar{u}^3 \end{bmatrix} \cdot Bez \cdot \begin{bmatrix} \mathbf{U}_0 & \mathbf{U}_1 & \mathbf{U}_2 & \mathbf{U}_3 \end{bmatrix}^T ,$$

where $T$ is the *transpose* operator, which converts a row vector into a column vector and vice versa. The $i^{\text{th}}$ segment of a uniform cubic B-spline can be written as

$$\mathbf{Q}_i(\bar{u}) = \begin{bmatrix} 1 & \bar{u} & \bar{u}^2 & \bar{u}^3 \end{bmatrix} \frac{1}{6} \begin{bmatrix} 1 & 4 & 1 & 0 \\ -3 & 0 & 3 & 0 \\ 3 & -6 & 3 & 0 \\ -1 & 3 & -3 & 1 \end{bmatrix} \begin{bmatrix} \mathbf{V}_{i-3} \\ \mathbf{V}_{i-2} \\ \mathbf{V}_{i-1} \\ \mathbf{V}_i \end{bmatrix} \qquad (68)$$

$$= \begin{bmatrix} 1 & \bar{u} & \bar{u}^2 & \bar{u}^3 \end{bmatrix} \cdot Bspl \cdot \begin{bmatrix} \mathbf{V}_{i-3} & \mathbf{V}_{i-2} & \mathbf{V}_{i-1} & \mathbf{V}_i \end{bmatrix}^T ,$$

the Hermite interpolation formula can be represented by

$$\mathbf{Q}_i(\bar{u}) = \begin{bmatrix} 1 & \bar{u} & \bar{u}^2 & \bar{u}^3 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ -3 & -2 & 3 & -1 \\ 2 & 1 & -2 & 1 \end{bmatrix} \begin{bmatrix} \mathbf{U}_i \\ \mathbf{D}_i \\ \mathbf{U}_{i+1} \\ \mathbf{D}_{i+1} \end{bmatrix}$$

$$= \left[1 \; \bar{u} \; \bar{u}^2 \; \bar{u}^3\right] \cdot Herm \cdot \left[\mathbf{U}_i \; \mathbf{D}_i \; \mathbf{U}_{i+1} \; \mathbf{D}_{i+1}\right]^T ,$$

and, of course, the power series representation

$$Q(\bar{u}) = \mathbf{a} + \mathbf{b}\bar{u} + \mathbf{c}\bar{u}^2 + \mathbf{d}\bar{u}^3$$

is represented trivially by

$$\mathbf{Q}(\bar{u}) = \left[1 \; \bar{u} \; \bar{u}^2 \; \bar{u}^3\right] \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mathbf{a} \\ \mathbf{b} \\ \mathbf{c} \\ \mathbf{d} \end{bmatrix} .$$

As [Smith83] points out, the use of matrices emphasizes the ease with which one can render curves represented in a variety of ways. One need write only a single procedure, whose parameters are simply a coefficient matrix and a data vector. We have avoided matrix representations because they are less intuitive to the newcomer. Matrices do, however, provide a concise and powerful notation; the survey of surface modelling techniques given in [Barsky84] illustrates this nicely.

### 11.7. Converting Between Representations

Another point which it is convenient to recall here is that each of the curve representations mentioned above relies on some particular basis for the cubic polynomials, and there is consequently a transformation from each to any of the others that can conveniently be expressed in terms of a matrix. To convert the control vertices for a uniform cubic B-spline curve segment into a Bézier representation we need only equate the coefficients of the $u^i$ (which must be unique, since the $u^i$ are a basis) in (67) and (68) and solve for

$$\left[\mathbf{U}_0 \; \mathbf{U}_1 \; \mathbf{U}_2 \; \mathbf{U}_3\right] = Bez^{-1} \cdot Bspl \cdot \left[\mathbf{V}_0 \; \mathbf{V}_1 \; \mathbf{V}_2 \; \mathbf{V}_3\right]^T$$

$$= \frac{1}{6} \begin{bmatrix} 1 & 4 & 1 & 0 \\ 0 & 4 & 2 & 0 \\ 0 & 2 & 4 & 0 \\ 0 & 1 & 4 & 1 \end{bmatrix} \begin{bmatrix} \mathbf{V}_0 \\ \mathbf{V}_1 \\ \mathbf{V}_2 \\ \mathbf{V}_3 \end{bmatrix} .$$

Conversely, to convert the control vertices for a Bézier representation into a B-spline representation we compute

$$\left[\mathbf{V}_0 \; \mathbf{V}_1 \; \mathbf{V}_2 \; \mathbf{V}_3\right] = Bspl^{-1} \cdot Bez \cdot \left[\mathbf{U}_0 \; \mathbf{U}_1 \; \mathbf{U}_2 \; \mathbf{U}_3\right]^T$$

$$= \begin{bmatrix} 6 & -7 & 2 & 0 \\ 0 & 2 & -1 & 0 \\ 0 & -1 & 2 & 0 \\ 0 & 2 & -7 & 6 \end{bmatrix} \begin{bmatrix} \mathbf{U}_0 \\ \mathbf{U}_1 \\ \mathbf{U}_2 \\ \mathbf{U}_3 \end{bmatrix} .$$

To go from the power series representation to the Be'zier control vertices we can compute

$$\begin{bmatrix} \mathbf{U}_0 \, \mathbf{U}_1 \, \mathbf{U}_2 \, \mathbf{U}_3 \end{bmatrix} = Bez^{-1} \cdot \begin{bmatrix} \mathbf{a} \; \mathbf{b} \; \mathbf{c} \; \mathbf{d} \end{bmatrix}^T$$

$$= \frac{1}{3} \begin{bmatrix} 3 & 0 & 0 & 0 \\ 3 & 1 & 0 & 0 \\ 3 & 2 & 1 & 0 \\ 3 & 3 & 3 & 3 \end{bmatrix} \begin{bmatrix} \mathbf{a} \\ \mathbf{b} \\ \mathbf{c} \\ \mathbf{d} \end{bmatrix}$$

and to go from the power series representation to the B-spline control vertices involves computing

$$\begin{bmatrix} \mathbf{V}_0 \, \mathbf{V}_1 \, \mathbf{V}_2 \, \mathbf{V}_3 \end{bmatrix} = Bspl^{-1} \cdot \begin{bmatrix} \mathbf{a} \; \mathbf{b} \; \mathbf{c} \; \mathbf{d} \end{bmatrix}^T$$

$$= \frac{1}{3} \begin{bmatrix} 3 & -3 & 2 & 0 \\ 3 & 0 & -1 & 0 \\ 3 & 3 & 2 & 0 \\ 3 & 6 & 11 & 18 \end{bmatrix} \begin{bmatrix} \mathbf{a} \\ \mathbf{b} \\ \mathbf{c} \\ \mathbf{d} \end{bmatrix} . \qquad (69)$$

The above discussion tells us only how to convert a single curve segment from one representation to another. This is sufficient if we are translating into the power basis. But what happens if we convert from the power basis to Bernstein polynomials or B-splines? Do the control vertices match up so as to form a single composite curve?

Suppose that we have two consecutive cubic segments $\mathbf{Q}_3(\overline{u})$ and $\mathbf{Q}_4(\overline{u})$ that meet with $C^2$ continuity. Let $\mathbf{V}_0$, $\mathbf{V}_1$, $\mathbf{V}_2$ and $\mathbf{V}_3$ be the B-spline control vertices that define $\mathbf{Q}_3$ and let $\mathbf{W}_1$, $\mathbf{W}_2$, $\mathbf{W}_3$ and $\mathbf{W}_4$ be the B-spline control vertices that define $\mathbf{Q}_4$. Now consider the five B-spline control vertices $\mathbf{V}_0$, $\mathbf{V}_1$, $\mathbf{V}_2$, $\mathbf{V}_3$ and $\mathbf{V}_4$ that define the composite curve $\mathbf{Q}(\overline{u})$ consisting of $\mathbf{Q}_3(\overline{u})$ and $\mathbf{Q}_4(\overline{u})$. Because the B-splines are a basis, we must have $\mathbf{V}_1 = \mathbf{W}_1$, $\mathbf{V}_2 = \mathbf{W}_2$ and $\mathbf{V}_3 = \mathbf{W}_3$. Thus the B-spline control vertices must match up.

It is inefficient to compute all four control vertices for $\mathbf{Q}_4(\overline{u})$ — three of them have already been generated for $\mathbf{Q}_3(\overline{u})$. We have only to compute $\mathbf{W}_4$, which from (69) we see is exactly

$$\mathbf{W}_4 = \frac{1}{3} ( 3a + 6b + 11c + 18d ) .$$

This computation is then repeated for each segment to yield all the control vertices.

By the same sort of argument it follows that we can go uniquely from a $C^2$ Be'zier curve to the (still $C^2$) unique power representation, and thence to a B-spline representation in which the control vertices must match up. The four control vertices defining the first segment are computed as above. The additional vertex for the second and succeeding segments are given by

$$\mathbf{W}_4 = \frac{1}{3} ( 2\mathbf{U}_1 - 7\mathbf{U}_2 + 6\mathbf{U}_3 ) .$$

If the Be'zier curve is not $C^2$ then the control vertices will not match up since the curve is not, in fact, a uniform cubic B-spline. It is, of course, possible to represent it as a B-spline by using a knot vector containing multiple knots.

Conversion from a B-spline to a Be´zier curve proceeds simply by repeated application of the appropriate matrix equation given above. The Be´zier control vertices computed will necessarily satisfy the $C^2$ continuity constraints developed previously.

## 11.8. Be´zier Surfaces

Be´zier surfaces are defined from Be´zier curves in exactly the same way that B-spline surfaces are built from B-spline curves. We take the *tensor product* of two Be´zier curves:

$$\mathbf{Q}(\bar{u},\bar{v}) = \sum_{i=0}^{d}\sum_{j=0}^{l} \mathbf{V}_{i,j} P_{i,d}(\bar{u}) P_{j,l}(\bar{v}) \ .$$

The Bernstein polynomials $P_{i,d}(\bar{u})$ and $P_{j,l}(\bar{v})$ need not be of the same degree. Indeed, the same is true of the B-splines from which we constructed B-spline surfaces.

Techniques for building multipatch Be´zier surfaces with $C^1$ or $C^2$ continuity at patch boundaries are discussed in [Faux79].

# 12. Preliminaries Regarding Knot Refinement

In this section and the next we will consider the general subdivision problem for B-splines: suppose we have constructed a curve

$$\mathbf{Q}(\bar{u}) = \sum_i \mathbf{V}_i B_{i,k}(\bar{u})$$

or a surface

$$\mathbf{Q}(\bar{u},\bar{v}) = \sum_i \sum_j \mathbf{V}_{i,j}\, B_{i,k}(\bar{u})\, B_{j,l}(\bar{v})$$

using some set of control vertices, and we now wish to express the same curve or surface in terms of a larger number of control vertices. There are two different reasons for wanting to do this.

The first reason is that we may wish to "fine tune" the curve or surface by increasing the number of control vertices near a section that requires adjustment.

The second reason is that we may wish to increase the number of control vertices as an intermediate step in displaying the curve or surface. It is the case that, if control vertices are added in a "reasonable" way in the manner to be established in this chapter and the next (where "reasonable" means spread uniformly about and not allowed to bunch up only in certain regions) then the control polygon will converge to the curve. Similar results apply to surfaces constructed from tensor products of B-splines and to the control graphs that define the surfaces. This suggests that vertices can be added to such an extent that the control polygon "converges visually" to the curve or surface; i.e. the polygon "clamps down on" the curve or surface as each facet of the control polygon is replaced by more and smaller facets. This exactly follows the behaviour displayed in figures 120 through 122 of chapter 11. When the facets are "small enough," the polygon, rather than the curve or surface, can then be subjected to all the display transformations and shading needed to produce a presentable image on a graphics display.

The ideas of subdivision have been introduced in the context of Bézier curves and surfaces. In this chapter and the next we shall establish subdivision results for B-splines.

## 12.1. Knots and Vertices

Using curves as the subject of our discussion, we begin by noting the connection between control vertices and knots. Each new control vertex that we might like to add needs to be weighted by some new B-spline; each new B-spline that we might like to construct needs some knot at which to become nonzero. Thus, we can approach the problem in two ways. We want

$$\mathbf{Q}(\bar{u}) = \sum_{i=0}^{m} \mathbf{V}_i \, B_{i,k}(\bar{u}) \equiv \sum_{j=0}^{m+n} \mathbf{W}_j \, N_{j,k}(\bar{u}) \ , \tag{70}$$

where

$$n > 0$$

and

$$\{\mathbf{V}_0, \ldots, \mathbf{V}_m\} \text{ becomes } \{\mathbf{W}_0, \ldots, \mathbf{W}_{m+n}\} \ .$$

We can attempt to find the vertices **W** explicitly, or we can attempt to find the new B-splines, $N_{j,k}(\bar{u})$, and use them to determine the **W** implicitly. The latter is the route that we take. We will find that we can obtain the new B-splines merely by adding new knots to the existing knot sequence, and this addition process will implicitly define the new control vertices. This is the approach developed in [Cohen80].

Let us consider adding $n$ new knots to the existing sequence $\{\bar{u}_0, \ldots, \bar{u}_{m+k}\} = \{\bar{u}_i\}_0^{m+k}$ to obtain a new sequence $\{\bar{w}_0, \ldots, \bar{w}_{m+n+k}\} = \{\bar{w}_j\}_0^{m+n+k}$ , where $\{\bar{u}_i\}_0^{m+k} \subset \{\bar{w}_j\}_0^{m+n+k}$ , as suggested by the following picture.

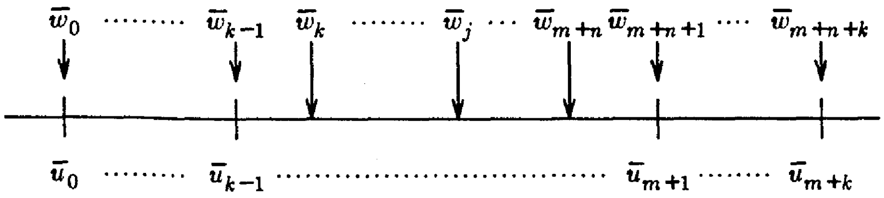

$\bar{w}_0$ ········ $\bar{w}_{k-1}$ $\bar{w}_k$ ········ $\bar{w}_j$ ··· $\bar{w}_{m+n}$ $\bar{w}_{m+n+1}$ ···· $\bar{w}_{m+n+k}$
$\bar{u}_0$ ········ $\bar{u}_{k-1}$ ································· $\bar{u}_{m+1}$ ········ $\bar{u}_{m+k}$

Figure 124. Changing the existing knot sequence by adding additional knots.

It is to be understood from this picture that each $\bar{u}_i$ is identical to one of the $\bar{w}_j$, and that some further $\bar{w}$ knots have been scattered along the parameter interval. As the picture suggests, additional knots are to be added strictly within the parameter range $[\bar{u}_{k-1}, \bar{u}_{m+1})$ so that

$$\bar{w}_0 \equiv \bar{u}_0 \ , \ \ldots \ , \ \bar{w}_{k-1} \equiv \bar{u}_{k-1}$$

$$\bar{u}_{k-1} \leq \bar{w}_k \leq \cdots \leq \bar{w}_{m+n} \leq \bar{u}_{m+1}$$

$$\bar{w}_{m+n+1} \equiv \bar{u}_{m+1} \ , \ \ldots \ , \ \bar{w}_{n+k} \equiv \bar{u}_{m+k} \ .$$

The reason for this is as follows. If we were to add a knot $\bar{w}_j$ in, for example, the interval

$$(\bar{u}_0, \bar{u}_{k-1})$$

then the resulting knot sequence would apparently allow the parameter $\bar{u}$ to vary to the left of $\bar{u}_{k-1}$, since we would now have all the conditions for constructing, by divided differences, $k$ linearly independent B-splines on an interval extending from $\bar{u}_{k-1}$ left to $\bar{w}_j$. As an example, for $k=4$ (cubics), if we add the knot $\bar{w}_3$ in the interval $[\bar{u}_2, \bar{u}_3)$, we would be permitting four linearly independent B-splines to exist on the interval

$$\bar{w}_3 \leq \bar{u} < \bar{u}_3 \ .$$

Since the given curve $\mathbf{Q}(\bar{u})$ does not exist for these values of $\bar{u}$, expanding the legal parameter range in this way would have the effect of "growing" the curve. We could ignore this effect, of course, by continuing to restrict our attention (and the parameter $\bar{u}$) to the range $[\bar{u}_3, \bar{u}_{m+1})$, but then the B-spline

$$N_{0,4}(\bar{u}) \ ,$$

constructable from the knots

$$\bar{w}_0 = \bar{u}_0 \ , \bar{w}_1 = \bar{u}_1 \ , \bar{w}_2 = \bar{u}_2 \ , \bar{w}_3 \ , \text{and } \bar{w}_4 = \bar{u}_3 \ ,$$

would be zero on $[\bar{u}_3,\bar{u}_{m+1})$. Hence, any control vertex $\mathbf{W}_0$ created implicitly by the addition of $\bar{w}_3$, which $N_{0,4}(\bar{u})$ would weight, would have absolutely no influence on $\mathbf{Q}(\bar{u})$.

While either of these artifacts — growing curves or creating useless control vertices — might elsewhere be interesting, we will not explore them here.

Restricting the introduction of new knots to the range $[\bar{u}_{k-1},\bar{u}_{m+1})$ is consistent in spirit to the discussion in [Cohen80] and to the related material on adding knots in [Schumaker81], for example. A convention followed in these references, as well, is that whenever new knots are added on top of existing knots $\bar{u}_i$, or are added multiply by themselves, it will be required that

$$\bar{w}_j < \bar{w}_{j+k}$$

for all $j$. That is, we will prohibit ourselves from adding new knots to any location on the $\bar{u}$ axis where the result of the addition would be to create a cluster (multiple knot) of multiplicity higher than $k$. If this were not observed, then we would be creating $k$-segment "knot intervals" of zero length for which the corresponding B-splines would be vacuous, clearly a futile exercise.

**Notation:** We will denote the multiplicity of each $\bar{w}_j$ by $\nu_j$.

The sense of the refinement process is that knots are "interspersed" among the knots of the $\{\bar{u}_i\}_0^{m+k}$ sequence, and then the resulting sequence is renamed using $\bar{w}$ "labels", as shown in Figure 125.

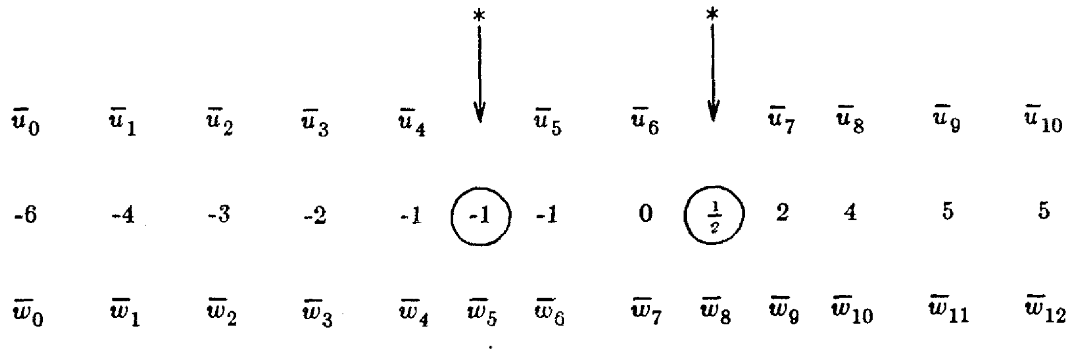


Figure 125. A specific example of the refinement and relabelling process.

Observe that a knot (which happens to have the value $-1$) is inserted between $\bar{u}_4$ and $\bar{u}_5$ (which also have the value $-1$). That new knot could have been inserted before $\bar{u}_4$ (in which case it would have received the label $\bar{w}_4$ and $\bar{u}_4$ would have been relabeled $\bar{w}_5$), or it could have been inserted after $\bar{u}_5$ (in which case it would have received the label $\bar{w}_6$, so that $\bar{u}_4$ and $\bar{u}_5$ would have become $\bar{w}_4$ and $\bar{w}_5$, respectively).

If we are firm about our policy of not adding new knots outside the legal parameter range — that is if we do not allow ourselves to add knots "to the left of $\bar{u}_{k-1}$" and "to the right of $\bar{u}_{m+1}$" — then all the knots added to $\{\bar{u}_i\}_0^{m+k}$ that are numerically equal to $\bar{u}_{k-1}$ are to be regarded as being inserted in the sequence after $\bar{u}_{k-1}$. Similarly, all the knots added to $\{\bar{u}_i\}_0^{m+k}$ that are numerically equal to $\bar{u}_{m+1}$ are to be regarded as being inserted in the sequence before $\bar{u}_{m+1}$.

**Definition:** The knot sequence

$$\{\bar{w}_j\}_0^{m+n+k} = \{ \bar{w}_0 , \ldots , \bar{w}_{m+n+k} \} \quad \text{where } n > 0$$

formed in accord with the above discussion will be called a *refinement* of the knot sequence

$$\{\bar{u}_i\}_0^{m+k} \subset \{\bar{w}_j\}_0^{m+n+k} \ .$$

We have been using the convention that, for each given index $i \in \{0, \ldots , m+k-1\}$, $\gamma_+(i)$ is defined to be the smallest index such that $\bar{u}_{\gamma_+(i)} > \bar{u}_i$. This convention will now be extended to the knot sequence $\{\bar{w}_j\}_0^{m+n+k}$ in the same way. We will use $\gamma_+(j)$ as the index of the leftmost knot in $\{\bar{w}_j\}_0^{m+n+k}$ whose value is strictly greater than $\bar{w}_j$. For example, in Figure 125 above, if $j=4$, then $\gamma_+(j)=7$.

**Notation:** For any index $j \in \{0, \ldots , m+n+k\}$

$$\gamma_+(j)$$

is defined to be the smallest index such that

$$\bar{w}_{\gamma_+(j)} > \bar{w}_j \ .$$

By convention, the index $m+n+k+1$ will be associated with $+\infty$, so

$$\gamma_+(m+n+k) = m+n+k+1$$

and

$$\bar{w}_{m+n+k+1} = +\infty \ .$$

The only caveat to be observed is that each time we use $\gamma_+$ we will now have to be clear whether we intend $\gamma_+$ to be understood as applying to the index set for the $\bar{u}$'s or as applying to the index set for the $\bar{w}$'s. For instance with respect to Figure 125 above,

$$\gamma_+(5) = 6 \ ,$$

if $\gamma_+$ is used for the $\bar{u}$ knots, and

$$\gamma_+(5) = 7 \ ,$$

if $\gamma_+$ is used for the $\bar{w}$ knots. Usually the meaning will be clear from context, and when it isn't clear we will state explicitly which meaning is intended.

A similar definition applies to $\gamma_-(j)$:

**Notation:** For any index $j \in \{0, \ldots, m+n+k-1\}$

$$\gamma_{-}(j)$$

is defined to be the largest index such that

$$\bar{w}_{\gamma_{-}(j)} < \bar{w}_j \ .$$

By convention, the index $-1$ will be associated with $-\infty$, so

$$\gamma_{-}(0) = -1$$

and

$$\bar{w}_{-1} = -\infty \ .$$

We will find it handy in this section, and in those following, to establish two further indexing notations. The first of these serves to locate the breakpoint interval in the sequence $\{\bar{u}_i\}_0^{m+k}$ into which each $\bar{w}_j$ falls. We will use $\delta(j)$ as the index of the rightmost knot in $\{\bar{u}_i\}_0^{m+k}$ whose value does not exceed $\bar{w}_j$. This will mean that, strictly according to value and not according to position in sequence, we will have

$$\bar{u}_{\delta(j)} \le \bar{w}_j < \bar{u}_{\delta(j)+1} \ ,$$

and this will serve to place $\bar{w}_j$ in its proper breakpoint interval. For example, in Figure 125 above, if $j=4$, then $\delta(j)=5$, which locates $\bar{w}_4=-1$ in the breakpoint interval

$$[\bar{u}_5,\bar{u}_6) = [-1,0) \ .$$

**Notation:** For any index $j \in \{0, \ldots, m+n+k\}$ chosen to select a knot $\bar{w}_j$,

$$\delta(j)$$

is defined to be the unique index

$$\delta(j) \in \{0, \ldots, m+k\}$$

satisfying

$$\bar{u}_{\delta(j)} \le \bar{w}_j < \bar{u}_{\delta(j)+1} \quad \text{for } j = 0, \ldots, m+n+k \ .$$

The convention holds that $\bar{u}_{m+k+1} = +\infty$, which will correctly place $\bar{w}_{m+n+k}$ in the interval $[\bar{u}_{m+k}, +\infty)$.

Finally, we introduce an indexing convention that provides a convenient way of relating the knots of $\{\bar{u}_i\}_0^{m+k}$ to the knots of $\{\bar{w}_j\}_0^{m+n+k}$ that represent them in the refinement. We will denote by $\eta(i)$ the index of the knot in $\{\bar{w}_j\}_0^{m+n+k}$ corresponding to $\bar{u}_i$. For example, in Figure 125 above, if $i=5$, then $\eta(i)=6$, since $\bar{u}_5$ was, indeed, relabled as $\bar{w}_6$.

**Notation:** For any index $i \in \{0, \ldots, m+k\}$ chosen to select a knot $\bar{u}_i$,

$$\eta(i)$$

is defined to be the unique index

$$\eta(i) \in \{0, \ldots, m+n+k\}$$

for which $\bar{w}_{\eta(i)}$ is the member of $\{\bar{w}_j\}_0^{m+n+k}$ identified with $\bar{u}_i$.

### 12.2. A Representation Result

Recall that the reason for considering knot refinement is to represent any curve defined by the B-splines $B_{0,k}(\bar{u}), \ldots, B_{m,k}(\bar{u})$ in terms of the B-splines $N_{0,k}(\bar{u}), \ldots, N_{m+n,k}(\bar{u})$. The following picture gives an overview of what happens when $k=2$ and a single knot is added.

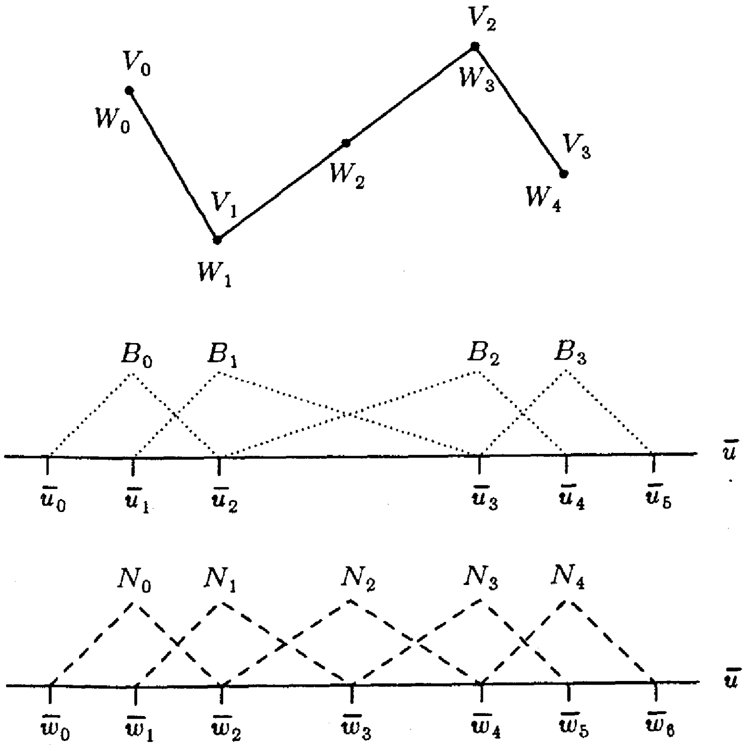


Figure 126. Adding a knot when $k=2$. Note that we show the normalized basis functions — they are not scaled by the corresponding control vertex.

The refinement process replaces the space $\mathbf{S}(\mathbf{P}^k, \{\bar{u}_i\}_0^{m+k})$ with $\mathbf{S}(\mathbf{P}^k, \{\bar{w}_j\}_0^{m+n+k})$, which contains the space $\mathbf{S}(\mathbf{P}^k, \{\bar{u}_i\}_0^{m+k})$ as a subspace. The B-splines $N_{j,k}(\bar{u})$ that can be constructed on the new knots are given by

$$N_{j,k}(\bar{u}) = (-1)^k(\bar{w}_{j+k} - \bar{w}_j)[\bar{w}_j(k):t](\bar{u}-t)_+^{k-1} \tag{71}$$

for $j = 0, \ldots, m+n$ .

Our task is to see whether we can express a curve that has been constructed in terms of the $B_{i,k}(\bar{u})$, in terms of the $N_{j,k}(\bar{u})$ instead, as is shown in equation (70) above. The next theorem assures us that we can, and it indicates that the new control vertices $\mathbf{W}_j$ will be simple linear combinations of the old control vertices $\mathbf{V}_i$.

The theorem follows easily from the fact that each $B_{i,k}(\bar{u})$ is itself a spline in $\mathbf{S}(\mathbf{P}^k, \{\bar{w}_j\}_0^{m+n+k})$, so we can "substitute" an expression in terms of $N_{j,k}(\bar{u})$ for each $B_{i,k}(\bar{u})$ and from the $\mathbf{V}_i$ deduce the $\mathbf{W}_j$. For example, it is easy to see that in Figure 126

$$\begin{aligned}
B_{0,2}(\bar{u}) &= 1.0 \cdot N_{0,2}(\bar{u}) \\
B_{1,2}(\bar{u}) &= 1.0 \cdot N_{1,2}(\bar{u}) + 0.5 \cdot N_{2,2}(\bar{u}) \\
B_{2,2}(\bar{u}) &= 0.5 \cdot N_{2,2}(\bar{u}) + 1.0 \cdot N_{3,2}(\bar{u}) \\
B_{3,2}(\bar{u}) &= 1.0 \cdot N_{4,2}(\bar{u}) \ .
\end{aligned}$$

The quadratic B-spline curve shown in Figure 127 provides a less trivial example.

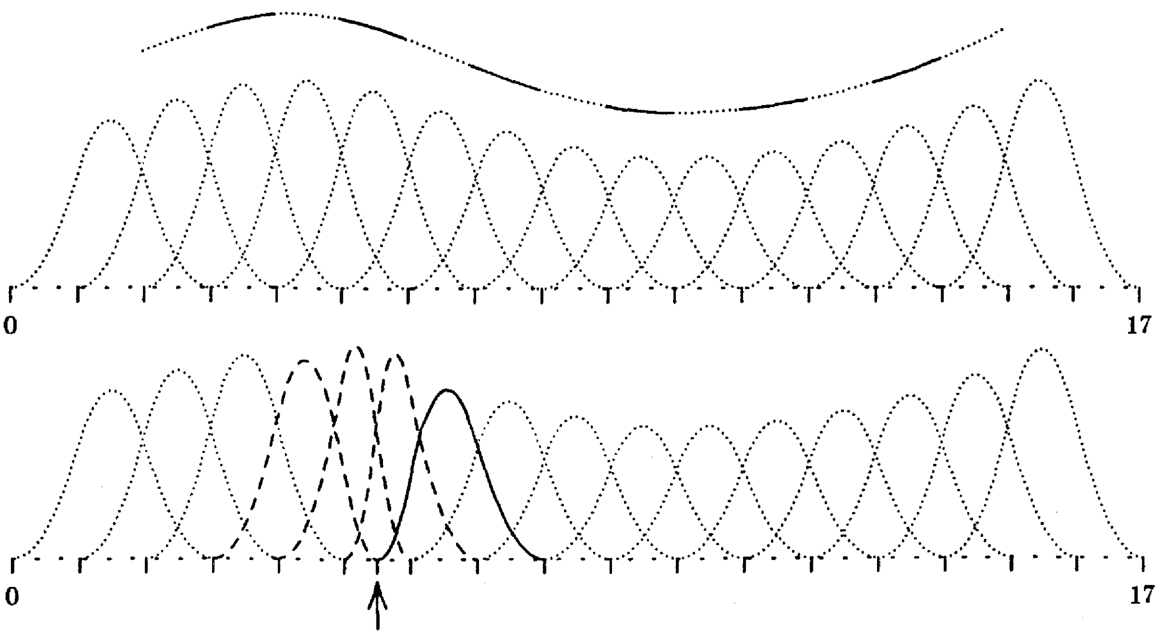


Figure 127. A thirteen segment quadratic B-spline curve. The arrow points to $\bar{u} = 5.5$, where a knot has been added, splitting the fourth segment from the left in half. That is, the fourth segment is regarded as two distinct quadratic polynomials that meet with first derivative continuity at $u = 5.5$. It follows that this curve can be represented using the lower set of B-splines, as shown.

Both sets of basis functions are shown scaled so that their sum, respectively, equals the curve shown. One new basis function is added (drawn as a solid curve), and the shape of three B-splines is changed (drawn as dashed curves). Notice also that since these are quadratic B-splines, each spans three intervals.

The curve in Figure 127 is a sum of the B-splines defined on the upper (uniform) knot sequence. The knot added at $\bar{u} = 5.5$ causes a new basis function to be added below in Figure 127, and causes three basis functions to change shape: namely the basis functions that go positive at 3, 4 and 5, these being the only old basis functions defined by divided differences that now include the new knot at 5.5.

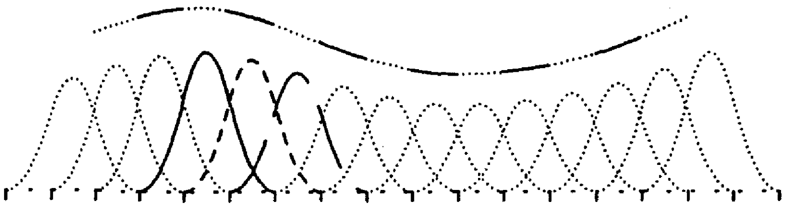


Figure 128. Figure 130 shows the representation of the solidly drawn basis function in terms of the refined knot sequence of Figure 127; Figure 131 similarly treats the basis function drawn dashed here.

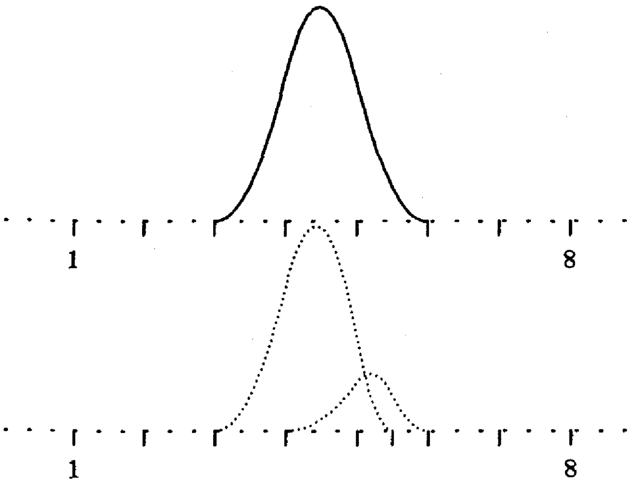


Figure 129. A detail from Figure 127. The sum of the lower two basis functions is exactly the upper.

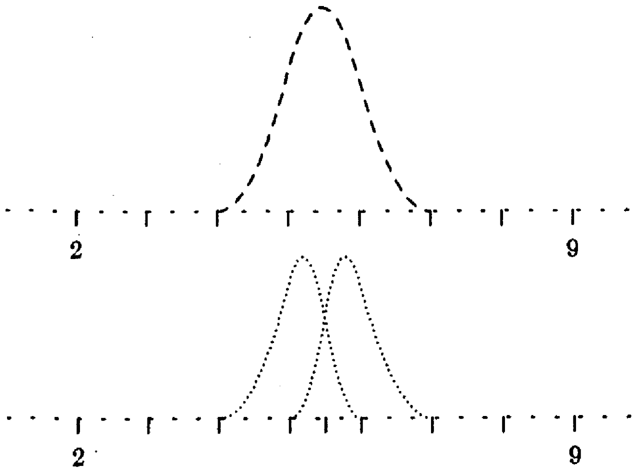


Figure 130. A detail from Figure 127. The sum of the lower two basis functions is exactly the upper.

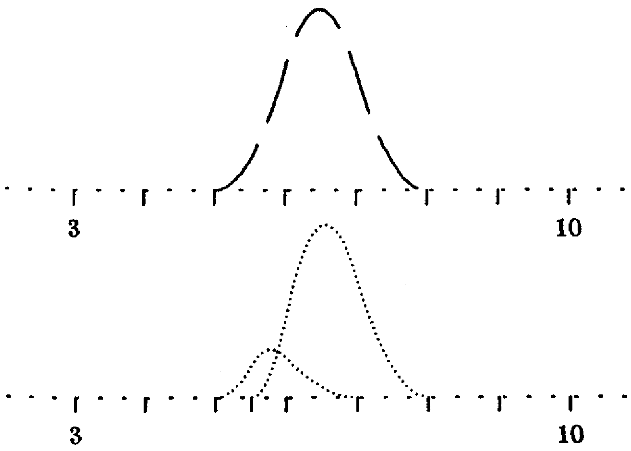


Figure 131. Another detail from Figure 127. Again, the sum of the lower two basis functions is exactly the upper.

Each of the uniform (upper) B-splines in Figure 127 is, of course, itself a piecewise quadratic curve and can therefore be represented as a scaled sum of the B-splines defined on the lower (refined) knot sequence, as shown in Figures 129, 130 and 131. (Figure 128 locates these three basis functions among those defined on the refined knot sequence in Figure 127.)

To be more general, suppose that we do this for each of the upper basis functions in Figure 127. Each lower basis function is needed some (small) number of times; add up all its contributions, and the result is the scale factor by which it is weighted in representing the curve of Figure 127. The following theorem makes this process precise.

**Theorem:** For each $j=0,\ldots,m+n$ and $i=0,\ldots,m$

$$\mathbf{W}_j = \sum_{i=0}^{m} \mathbf{V}_i \; \alpha_{i,k}(j)$$

for some collection of numbers $\alpha_{i,k}(j)$.

**Argument:** First observe that each of the "old" B-splines $B_{i,k}(\overline{u})$ can be expressed in terms of the "new" ones $N_{j,k}(\overline{u})$. This follows because the sequences $\{\overline{u}_i\}_0^{m+k}$ and $\{\overline{w}_j\}_0^{m+n+k}$ are compatible. That is, the breakpoint intervals of $\mathbf{S}(\mathbf{P}^k,\{\overline{w}_j\}_0^{m+n+k})$ must all be subintervals of the breakpoint intervals of $\mathbf{S}(\mathbf{P}^k,\{\overline{u}_i\}_0^{m+k})$. There is no crossing of boundaries or mismatch — each breakpoint of $\mathbf{S}(\mathbf{P}^k,\{\overline{u}_i\}_0^{m+k})$ must be a breakpoint of $\mathbf{S}(\mathbf{P}^k,\{\overline{w}_j\}_0^{m+n+k})$ (and, of course, $\mathbf{S}(\mathbf{P}^k,\{\overline{w}_j\}_0^{m+n+k})$ may have some additional breakpoints falling strictly within the breakpoint intervals for $\mathbf{S}(\mathbf{P}^k,\{\overline{u}_i\}_0^{m+k})$).

Now, consider $B_{i,k}(\overline{u})$ for any fixed $i$. This function is a member of $\mathbf{S}(\mathbf{P}^k,\{\overline{w}_j\}_0^{m+n+k})$: firstly, it is a polynomial throughout each breakpoint interval of $\mathbf{S}(\mathbf{P}^k,\{\overline{w}_j\}_0^{m+n+k})$; secondly, if $\overline{w}_j$ is a breakpoint of $\mathbf{S}(\mathbf{P}^k,\{\overline{w}_j\}_0^{m+n+k})$, then $B_{i,k}(\overline{u})$ will be at least $C^{k-1-\nu_j}$ at that point. This second condition is satisfied because either $\overline{w}_j$ falls between breakpoints of $\mathbf{S}(\mathbf{P}^k,\{\overline{u}_i\}_0^{m+k})$, and consequently $B_{i,k}(\overline{u})$ is

(more...)

an ordinary $k$-order polynomial at $\bar{w}_j$, or because $\bar{w}_j$ was introduced upon a breakpoint, $\bar{u}_i$ of $\mathbf{S}(\mathbf{P}^k,\{\bar{u}_i\}_0^{m+k})$, and consequently $\nu_j \geq \mu_i + 1$. In other words, the definition of $\mathbf{S}(\mathbf{P}^k,\{\bar{w}_j\}_0^{m+n+k})$ would only demand $C^{k-1-\nu_j}$ continuity at $\bar{w}_j$, and $B_{i,k}(\bar{u})$ has continuity at least one order higher than that.

Since the functions $N_{j,k}(\bar{u})$ are a basis for $\mathbf{S}(\mathbf{P}^k,\{\bar{w}_j\}_0^{m+n+k})$, and each $B_{i,k}(\bar{u}) \in \mathbf{S}(\mathbf{P}^k,\{\bar{w}_j\}_0^{m+n+k})$, we can represent $B_{i,k}(\bar{u})$ as

$$B_{i,k}(\bar{u}) = \sum_{j=0}^{m+n} \alpha_{i,k}(j)\, N_{j,k}(\bar{u}) \ . \tag{72}$$

This defines what the quantities $\alpha_{i,k}(j)$ must be.

Now consider the original problem of writing the same curve $\mathbf{Q}(\bar{u})$ in terms of the B-splines of $\mathbf{S}(\mathbf{P}^k,\{\bar{u}_i\}_0^{m+k})$ and in terms of the B-splines of $\mathbf{S}(\mathbf{P}^k,\{\bar{w}_j\}_0^{m+n+k})$ as given in (70) above. We have

$$\mathbf{Q}(\bar{u}) = \sum_{i=0}^{m} \mathbf{V}_i\, B_{i,k}(\bar{u}) = \sum_{j=0}^{m+n} \mathbf{W}_j\, N_{j,k}(\bar{u}) \ ,$$

which can be rewritten as

$$\sum_{j=0}^{m+n} \mathbf{W}_j\, N_{j,k}(\bar{u}) = \sum_{i=0}^{m} \mathbf{V}_i \left\{ \sum_{j=0}^{m+n} \alpha_{i,k}(j)\, N_{j,k}(\bar{u}) \right\}$$

or, interchanging the order of summation,

$$\sum_{j=0}^{m+n} \left\{ \sum_{i=0}^{m} \mathbf{V}_i\, \alpha_{i,k}(j) \right\} N_{j,k}(\bar{u}) = \sum_{j=0}^{m+n} \mathbf{W}_j N_{j,k}(\bar{u}) \ .$$

Since the functions $N_{j,k}(\bar{u})$ are linearly independent, the above implies that the coefficients on each side of the equality are identical. That is,

$$\mathbf{W}_j = \sum_{i=0}^{m} \mathbf{V}_i\, \alpha_{i,k}(j) \ .$$

In the following chapter we will establish precise formulas, derived from the recurrence properties of the B-splines $B_{i,k}(\bar{u})$ and $N_{j,k}(\bar{u})$, for computing the coefficients $\alpha_{i,k}(j)$. We will find that the $\alpha$'s satisfy a simple recurrence of their own.

# 13. The Oslo Algorithm

This chapter will cover the technical details of the most general refinement algorithm known for general B-splines. In the previous chapter we observed that a curve (or surface) $\mathbf{Q}(\overline{u})$ constructed

- from one set of control vertices, $\mathbf{V}_0, \ldots, \mathbf{V}_m$,
- weighted by one set of B-splines, $B_i$,
- and defined on one set of knots, $\{\overline{u}_i\}_0^{n+k}$

can be represented in terms of

- a larger set of control vertices, $\mathbf{W}_0, \ldots, \mathbf{W}_{m+n}$,
- weighted by a refined set of B-splines, $N_j$,
- and defined on a finer mesh of knots, $\{\overline{w}_j\}_0^{m+n+k}$.

The key idea is that the process of knot refinement produces a spline space $\mathbf{S}(\mathbf{P}^k, \{\overline{w}_j\}_0^{m+n+k})$ which contains the original space $\mathbf{S}(\mathbf{P}^k, \{\overline{u}_i\}_0^{m+k})$. It is directly from these observations, and from B-spline recurrence, that the detailed behaviour of the $\alpha_{i,k}(j)$ can be determined, and it is the behaviour of the $\alpha$'s upon which the B-spline and control-vertex refinements are based.

### 13.1. Discrete B-spline Recurrence

We introduced the quantities $\alpha_{i,k}(j)$ that provided a translation from the $\mathbf{V}_i$ to the $\mathbf{W}_j$ via the equation

$$\mathbf{W}_j = \sum_{i=0}^{m} \alpha_{i,k}(j)\mathbf{V}_i \quad \text{for } j = 0, \ldots, m+n \quad ,$$

and from the $B_i$ to the $N_j$ via the equation

$$B_{i,k}(\overline{u}) = \sum_{j=0}^{m+n} \alpha_{i,k}(j) N_{j,k}(\overline{u}) \quad . \tag{73}$$

Our first task in this chapter will be to establish that these coefficients $\alpha_{i,k}(j)$ satisfy a recurrence very much like the one satisfied by the B-splines:

**Theorem:**

$$\alpha_{i,1}(j) = \begin{cases} 1 & \bar{u}_i \le \bar{w}_j < \bar{u}_{i+1} \\ & \\ 0 & \text{otherwise} \end{cases}$$

and

$$\alpha_{i,r}(j) = \frac{\bar{w}_{j+r-1} - \bar{u}_i}{\bar{u}_{i+r-1} - \bar{u}_i}\,\alpha_{i,r-1}(j) + \frac{\bar{u}_{i+r} - \bar{w}_{j+r-1}}{\bar{u}_{i+r} - \bar{u}_{i+1}}\,\alpha_{i+1,r-1}(j) \qquad (74)$$

for $r = 2, 3, \ldots, k$, where $k$ is the order of the spline in question.

As usual, we interpret each ratio

$$\frac{\bar{w}_{j+r-1} - \bar{u}_i}{\bar{u}_{i+r-1} - \bar{u}_i} \quad \text{and} \quad \frac{\bar{u}_{i+r} - \bar{w}_{j+r-1}}{\bar{u}_{i+r} - \bar{u}_{i+1}}$$

to be zero if its denominator is zero.

This recurrence, as well as a related recurrence for obtaining the control vertices **W** from the control vertices **V**, were first established in [Cohen80]; they have recently been established by a much simpler method in [Prautzsch84], which is the source of the argument given here.

**Argument:**

Consider the identity (73). This is a representation formula for $B_{i,k}(\bar{u})$, regarded as an element of the space $\mathbf{S}(\mathbf{P}^k, \{\bar{w}_j\}_0^{m+n+k})$, and as such it is valid for all $\bar{u}$ in the legal parameter range:

$$\bar{u}_{k-1} = \bar{w}_{k-1} \le \bar{u} < \bar{w}_{m+n+1} = \bar{u}_{n+1} \ . \qquad (75)$$

We may apply the B-spline recurrence to $B_{i,k}(\bar{u})$ to obtain

$$B_{i,k}(\bar{u}) = \frac{\bar{u} - \bar{u}_i}{\bar{u}_{i+k-1} - \bar{u}_i}\,B_{i,k-1}(\bar{u}) + \frac{\bar{u}_{i+k} - \bar{u}}{\bar{u}_{i+k} - \bar{u}_{i+1}}\,B_{i+1,k-1}(\bar{u}) \ . \qquad (76)$$

On the other hand, we may apply the B-spline recurrence to $N_{j,k}(\bar{u})$ to obtain

$$N_{j,k}(\bar{u}) = \frac{\bar{u} - \bar{w}_j}{\bar{w}_{j+k-1} - \bar{w}_j}\,N_{j,k-1}(\bar{u}) + \frac{\bar{w}_{j+k} - \bar{u}}{\bar{w}_{j+k} - \bar{w}_{j+1}}\,N_{j+1,k-1}(\bar{u}) \ . \qquad (77)$$

Both (76) and (77) are valid for all values of $\bar{u}$.

Combining (76) and (77) with (73) yields the following identity, which is valid for all $\bar{u}$ in (75):

(more...)

$$\frac{\bar{u} - \bar{u}_i}{\bar{u}_{i+k-1} - \bar{u}_i} B_{i,k-1}(\bar{u}) + \frac{\bar{u}_{i+k} - \bar{u}}{\bar{u}_{i+k} - \bar{u}_{i+1}} B_{i+1,k-1}(\bar{u}) \tag{78}$$

$$= \sum_{j=0}^{m+n} \alpha_{i,k}(j) \left[ \frac{\bar{u} - \bar{w}_j}{\bar{w}_{j+k-1} - \bar{w}_j} N_{j,k-1}(\bar{u}) + \frac{\bar{w}_{j+k} - \bar{u}}{\bar{w}_{j+k} - \bar{w}_{j+1}} N_{j+1,k-1}(\bar{u}) \right] .$$

Let us begin with the right-hand side of (78). The summation can be regrouped in terms of $N_{j,k}(\bar{u})$; that is:

$$\sum_{j=0}^{m+n} \alpha_{i,k}(j) \left[ \frac{\bar{u} - \bar{w}_j}{\bar{w}_{j+k-1}\bar{w} - {}_j} N_{j,k-1}(\bar{u}) + \frac{\bar{w}_{j+k} - \bar{u}}{\bar{w}_{j+k} - \bar{w}_{j+1}} N_{j+1,k-1}(\bar{u}) \right] \tag{79}$$

$$= \sum_{j=0}^{m+n+1} \left[ \frac{\bar{w}_{j+k-1} - \bar{u}}{\bar{w}_{j+k-1} - \bar{w}_j} \alpha_{i,k}(j-1) + \frac{\bar{u} - \bar{w}_j}{\bar{w}_{j+k-1} - \bar{w}_j} \alpha_{i,k}(j) \right] N_{j,k-1}(\bar{u}) .$$

The second summation of (79) contains two spurious $\alpha$'s: $\alpha_{i,k}(-1)$ and $\alpha_{i,k}(m+n+1)$. However, the terms $N_{0,k-1}(\bar{u})$ and $N_{m+n+1,k-1}(\bar{u})$, respectively, that multiply these $\alpha$'s are zero on the parameter range (75); consequently, we may regard $\alpha_{i,k}(-1)$ and $\alpha_{i,k}(m+n+1)$ to be zero.

Returning to (78), consider the left-hand side of the equation. The representation of $B$'s in terms of $\alpha$'s and $N$'s given by (73) for $k^{\text{th}}$-order splines can be equally well used for $k-1^{\text{st}}$-order splines. This can be seen from the fact that $\{\bar{w}_j\}_0^{m+n+k}$ is still a refinement of $\{\bar{u}_i\}_0^{n+k}$, so the spline space for which the functions $N_{j,k-1}(\bar{u})$ form a basis still contains the space for which the functions $B_{i,k-1}(\bar{u})$ form a basis. The only peculiarity worth mention is the fact that the multiplicity some of the knots in $\{\bar{u}_i\}_0^{m+k}$ and/or $\{\bar{w}_j\}_0^{m+n+k}$ may be higher than we have usually admitted for $k-1^{\text{st}}$-order spline spaces. This is consistent with the nature of the B-spline recurrence, however; we merely accept ratios with zero denominators to be zero, and we must accept the fact that some of the B-splines that formally appear in our summations will be vacuous.

Applying the representation (73) to $B_{i,k-1}(\bar{u})$ and $B_{i+1,k-1}(\bar{u})$ separately transforms the left-hand side of (78) into:

$$\frac{\bar{u} - \bar{u}_i}{\bar{u}_{i+k-1} - \bar{u}_i} \sum_{j=0}^{m+n+1} \alpha_{i,k-1}(j)\, N_{j,k-1}(\bar{u}) + \frac{\bar{u}_{i+k} - \bar{u}}{\bar{u}_{i+k} - \bar{u}_{i+1}} \sum_{j=0}^{m+n+1} \alpha_{i+1,k-1}(j)\, N_{j,k-1}(\bar{u}) .$$

Since we are limited to the parameter range (75), we may consider the spurious quantities $\alpha_{i,k}(m+n+1)$ and $\alpha_{i+1,k}(m+n+1)$ to be zero, since they are multiplied by $N_{m+n+1,k-1}(\bar{u})$, which is zero on the parameter range.

The two summations above may be combined to obtain

(more...)

$$\sum_{j=0}^{m+n+1} \left[ \frac{\bar{u} - \bar{u}_i}{\bar{u}_{i+k-1} - \bar{u}_i} \alpha_{i,k-1}(j) + \frac{\bar{u}_{i+k} - \bar{u}}{\bar{u}_{i+k} - \bar{u}_{i+1}} \alpha_{i+1,k-1}(j) \right] N_{j,k-1}(\bar{u}) \ . \tag{80}$$

Combining (80) with (79) yields

$$\begin{aligned} &\sum_{j=0}^{m+n+1} \left[ \frac{\bar{u} - \bar{u}_i}{\bar{u}_{i+k-1} - \bar{u}_i} \alpha_{i,k-1}(j) + \frac{\bar{u}_{i+k} - \bar{u}}{\bar{u}_{i+k} - \bar{u}_{i+1}} \alpha_{i+1,k-1}(j) \right] N_{j,k-1}(\bar{u}) \\ &= \sum_{j=0}^{m+n+1} \left[ \frac{\bar{w}_{j+k-1} - \bar{u}}{\bar{w}_{j+k-1} - \bar{w}_j} \alpha_{i,k}(j-1) + \frac{\bar{u} - \bar{w}_j}{\bar{w}_{j+k-1} - \bar{w}_j} \alpha_{i,k}(j) \right] N_{j,k-1}(\bar{u}) \end{aligned} \tag{81}$$

The identity (81) is valid for all values of $\bar{u}$ in the parameter range (75).

It is tempting to conclude that the corresponding expressions in brackets in the two summations of (81) are equal for all $\bar{u}$, that is:

$$\begin{aligned} &\frac{\bar{u} - \bar{u}_i}{\bar{u}_{i+k-1} - \bar{u}_i} \alpha_{i,k-1}(j) + \frac{\bar{u}_{i+k} - \bar{u}}{\bar{u}_{i+k} - \bar{u}_{i+1}} \alpha_{i+1,k-1}(j) \\ &= \frac{\bar{w}_{j+k-1} - \bar{u}}{\bar{w}_{j+k-1} - \bar{w}_j} \alpha_{i,k}(j-1) + \frac{\bar{u} - \bar{w}_j}{\bar{w}_{j+k-1} - \bar{w}_j} \alpha_{i,k}(j) \end{aligned} \tag{82}$$

In fact this is true. We will sketch the reasoning for this in the simplest case, that in which there are no multiple knots. A complete justification would require a limiting process to produce multiple knots by the confluence of distinct knots, and that is outside the mathematical scope of this work. An alternative argument for the recurrence, without an omission of this nature, will be given in 15.5.3.

Let us, therefore, assume that all of the refined knots, $\bar{w}$, are distinct. If we pick any segment interval in the legal parameter range,

$$\bar{w}_j \leq \bar{u} < \bar{w}_{j+1} \ ,$$

then only $k-1$ successive terms of the summations in (81) could be nonzero owing to the locality of the support of the $N$'s. This means that equation (81) could be written in the form

$$\sum_{r=j-k+2}^{j} a_r(\bar{u}-b_r) N_{r,k-1}(\bar{u}) = 0 \quad \text{for all } \bar{u} \in [\bar{w}_j, \bar{w}_{j+1}) \ , \tag{83}$$

where we have put both summations on one side of the equal sign and abbreviated each of the resulting linear expressions in brackets by writing it as:

(more...)

$$a_r(\bar{u}+b_r) = \frac{\bar{u}-\bar{u}_i}{\bar{u}_{i+k-1}-\bar{u}_i}\,\alpha_{i,k-1}(r) + \frac{\bar{u}_{i+k}-\bar{u}}{\bar{u}_{i+k}-\bar{u}_{i+1}}\,\alpha_{i+1,k-1}(r)$$
$$- \frac{\bar{w}_{r+k-1}-\bar{u}}{\bar{w}_{r+k-1}-\bar{w}_r}\,\alpha_{i,k}(r-1) + \frac{\bar{u}-\bar{w}_r}{\bar{w}_{r+k-1}-\bar{w}_r}\,\alpha_{i,k}(r) \ . \tag{84}$$

It is worth noting that every $\alpha$ is represented in at least one of the segment intervals

$$[\bar{w}_{k-1},\bar{w}_k) \quad [\bar{w}_k,\bar{w}_{k+1}) \quad ,\ldots, \quad [\bar{w}_{m+n},\bar{w}_{m+n+1}) \ .$$

Hence, all of the terms appearing in brackets in (81) will be covered in (84).

Since we are restricting our attention to a single segment interval in the legal parameter range, the functions $N_{r,k-1}(\bar{u})$ for $r = j-k+2,\ldots,j$ are simply polynomials. Moreover, they are linearly independent and positive throughout the interval. As a result of this, the coefficients, $a_r$, can all be shown to be zero.

Suppose that some coefficient is nonzero; for example, $a_\lambda \neq 0$. Then, we can rewrite (83) as

$$a_\lambda(\bar{u}-b_\lambda)N_{\lambda,k-1}(\bar{u}) = -\sum_{\substack{r=j-k+2\\ r\neq\lambda}}^{j-1} a_r(\bar{u}-b_r)N_{r,k-1}(\bar{u}) \ .$$

But $N_{\lambda,k-1}(\bar{u}) > 0$, so we may divide:

$$a_\lambda(\bar{u}-b_\lambda) = -\sum_{\substack{r=j-k+2\\ r\neq\lambda}}^{j-1} a_r(\bar{u}-b_r)\left[\frac{N_{r,k-1}(\bar{u})}{N_{\lambda,k-1}(\bar{u})}\right] \ .$$

But the expression on the left of this equality is linear, which demands that the ratios

$$\frac{N_{r,k-1}(\bar{u})}{N_{\lambda,k-1}(\bar{u})}$$

be constant, and this contradicts the linear independence of the $N$'s.

The consequence of this is that the expression in (84) is zero for all values of $\bar{u}$, which means that (82) is true for all values of $\bar{u}$. If we merely substitute the particular value

$$\bar{u} = \bar{w}_{j+k-1}$$

into (82), the result given by (74) follows immediately.

To see that the starting values for the recurrence are correct; that is, the definitions for the quantities $\alpha_{i,1}(j)$, it is merely necessary to consider a picture. Since

$$B_{i,k}(\bar{u}) = \sum_{j=0}^{m+n} \alpha_{i,k}(j)N_{j,k}(\bar{u}) \quad \text{for } j=0,\ldots,m+n \ ,$$

this tells us what the contribution of a particular B-spline, $N_{j,k}(\bar{u})$ is to the B-spline $B_{i,k}(\bar{u})$. For the 1st-order B-splines we see that $B_{i,k}(\bar{u})$ and $N_{j,k}(\bar{u})$ must, in fact, be exactly equal if $\bar{u}_i \leq \bar{w}_j < \bar{u}_{i+1}$

(more...)

for values of $\bar{u}$ within the interval $[\bar{w}_j,\bar{w}_{j+1})$. This makes sense, as an example will illustrate: the single 1st-order B-spline shown in Figure 132 is replaced by the four 1st-order B-splines shown in Figure 133 when we insert the knots at 0.20, 0.50 and 0.75.

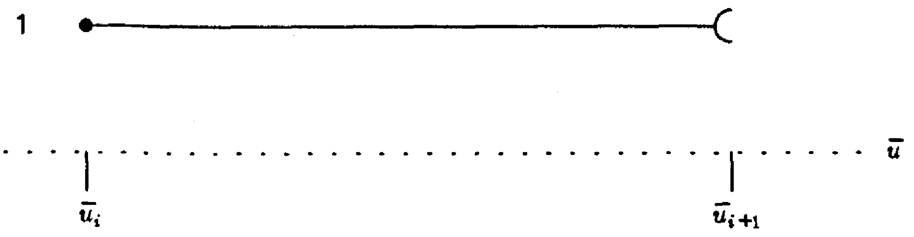


Figure 132. The single 1st-order B-spline $B_{i,1}(\bar{u})$ which has the value 1.0 on the interval $[\bar{u}_i,\bar{u}_{i+1})$.

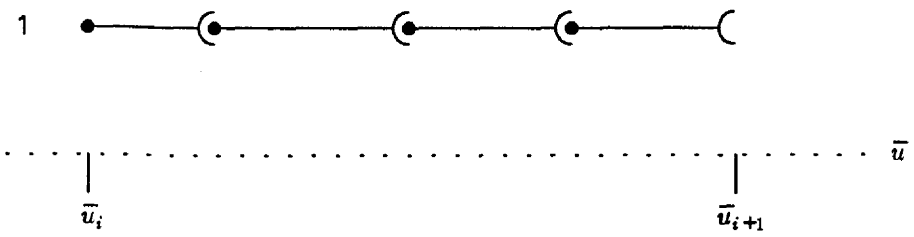


Figure 133. The four constant B-splines that replace $B_{i,1}(\bar{u})$ of Figure 132 when we insert knots at the values 0.20, 0.50 and 0.75 in the interval $[\bar{u}_i,\bar{u}_{i+1})$.

An inspection of (74) suggests that $\alpha_{i,k}(j)$ must have the character that the B-spline $B_{i,k}(\bar{u})$ would have if it were defined, not on the continuum of $\bar{u}$ values, but on the discrete collection $\{\bar{w}_j\}_0^{m+n+k}$ instead. This observation is the justification for the name given to the $\alpha$'s; they are known as the *discrete B-splines**.

## 13.2. Discrete B-spline properties

We see by the above that the behaviour of the $\alpha$'s parallels very closely the behaviour of the B-splines. What do the $\alpha$'s look like? Are they "hump-shaped" like the continuous B-splines? Are they local? Are they nonnegative? Do they sum to one? On the following pages we will plot a few low-order discrete B-splines to gain some feeling for them.

*The reader should be cautioned that the $\alpha_{i,k}(j)$ are not precisely the discrete B-splines defined in [Schumaker81].

For $k=1$ we are at the bottom of the alpha recurrence. $\alpha_{i,1}(j)$ is a function defined over the indices of the $\bar{w}$ knots, and it is clear from the recurrence that the interval of indices on which this function is nonzero is that corresponding to the $\bar{w}$ values falling into the interval $[\bar{u}_i,\bar{u}_{i+1})$. A typical table of $\bar{u}$, $\bar{w}$, and $\alpha_{i,1}$ values would be

| $\bar{u}$ | $\bar{w}$ | $\alpha_{i,1}$ |
|---|---|---|
| -3.0000 ($\bar{u}_{i-3}$) | -3.0000 | 0 |
| -2.0000 ($\bar{u}_{i-2}$) | -2.0000 | 0 |
| | -1.5000 | 0 |
| -1.0000 ($\bar{u}_{i-1}$) | -1.0000 | 0 |
| | -0.5000 | 0 |
| | -0.2500 | 0 |
| 0.0000 ($\bar{u}_i$) | 0.0000 | 1 |
| | 0.2000 | 1 |
| | 0.5000 | 1 |
| | 0.7500 | 1 |
| 1.0000 ($\bar{u}_{i+1}$) | 1.0000 | 0 |
| | 1.3333 | 0 |
| | 1.6250 | 0 |
| | 1.7500 | 0 |
| 2.0000 ($\bar{u}_{i+2}$) | 2.0000 | 0 |
| | 2.5000 | 0 |
| 3.0000 ($\bar{u}_{i+3}$) | 3.0000 | 0 |
| 4.0000 ($\bar{u}_{i+4}$) | 4.0000 | 0 |

which produces the following graph.

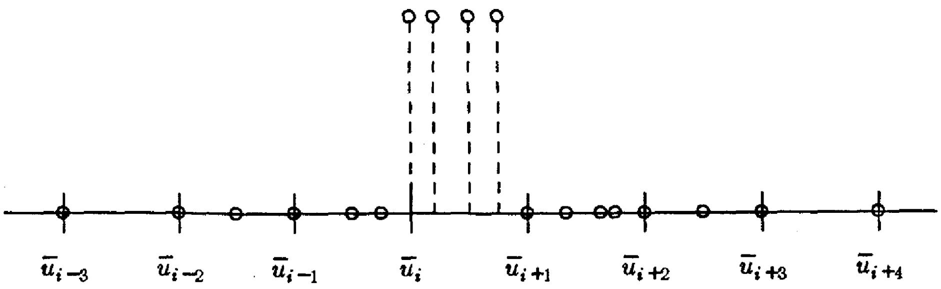


Figure 134. The first-order discrete B-spline. We have not explicitly labeled the $\bar{w}$ values; they are visible only as the locations of the values plotted for $\alpha_{i,1}$. Notice that the nonzero alphas all fall within one half-open interval.

Since

$$\mathbf{W}_j = \sum_{i=0}^{m} \alpha_{i,k}(j)\mathbf{V}_i$$

such graphs and tables indicate precisely how many **W**'s depend in what way upon which **V**'s. In particular, the table above and Figure 134 show that $\mathbf{W}_j$, $\mathbf{W}_{j+1}$, $\mathbf{W}_{j+2}$, and $\mathbf{W}_{j+3}$ are each $1\times\mathbf{V}_i$, where the index $j$ corresponds to $\bar{w}_j = 0.0000$ and the index $i$ corresponds to $\bar{u}_i = 0.0000$.

For $k=2$ there are several configurations, depending on multiplicities. We show three examples. A table of $\bar{u}$, $\bar{w}$, and $\alpha_{i,1}$ values for simple $\bar{u}$ knots would be

| $\bar{u}$ | $\bar{w}$ | $\alpha_{i,2}$ |
|---|---|---|
| -3.0000 ($\bar{u}_{i-3}$) | -3.0000 | 0 |
| -2.0000 ($\bar{u}_{i-2}$) | -2.0000 | 0 |
| | -1.5000 | 0 |
| -1.0000 ($\bar{u}_{i-1}$) | -1.0000 | 0 |
| | -0.5000 | 0 |
| | -0.2500 | 0 |
| 0.0000 ($\bar{u}_i$) | 0.0000 | 1/5 |
| | 0.2000 | 1/2 |
| | 0.5000 | 3/4 |
| | 0.7500 | 1 |
| 1.0000 ($\bar{u}_{i+1}$) | 1.0000 | 2/3 |
| | 1.3333 | 3/8 |
| | 1.6250 | 1/4 |
| | 1.7500 | 0 |
| 2.0000 ($\bar{u}_{i+2}$) | 2.0000 | 0 |
| | 2.5000 | 0 |
| 3.0000 ($\bar{u}_{i+3}$) | 3.0000 | 0 |
| 4.0000 ($\bar{u}_{i+4}$) | 4.0000 | 0 |

which produces the graph

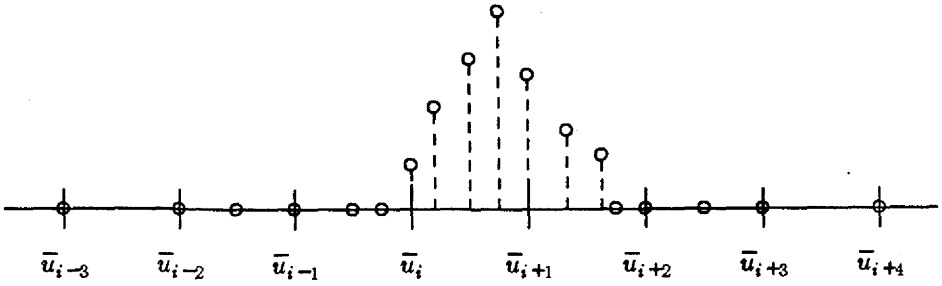


Figure 135. Values of the second-order discrete B-spline with no multiplicities. Recall that these values graph the contribution of $\mathbf{V}_i$, which is weighted by the B-spline $B_{i,2}(\bar{u})$, to various $\mathbf{W}_j$'s. This time the nonzero alphas all fall within two successive half-open intervals.

Observe that this graph and table specify that

$$
\begin{aligned}
\mathbf{W}_j &= \tfrac{1}{5}\mathbf{V}_i \ + \text{other } \mathbf{V}\text{'s} \\
\mathbf{W}_{j+1} &= \tfrac{1}{2}\mathbf{V}_i \ + \text{other } \mathbf{V}\text{'s} \\
\cdot \quad & \cdot \quad \cdot \quad \cdot \\
\cdot \quad & \cdot \quad \cdot \quad \cdot \\
\cdot \quad & \cdot \quad \cdot \quad \cdot \\
\mathbf{W}_{j+6} &= \tfrac{1}{4}\mathbf{V}_i \ + \text{other } \mathbf{V}\text{'s} \ ,
\end{aligned}
$$

where $j$ is the index for which $\bar{w}_j = 0.0000$ and $i$ is the index for which $\bar{u}_i = 0.0000$.

Placing a double knot at $\bar{u}_i$ produces the table

| $\bar{u}$ | $\bar{w}$ | $\alpha_{i,2}$ |
|---|---|---|
| -3.0000 ($\bar{u}_{i-3}$) | -3.0000 | 0 |
| -2.0000 ($\bar{u}_{i-2}$) | -2.0000 | 0 |
| | -1.5000 | 0 |
| -1.0000 ($\bar{u}_{i-1}$) | -1.0000 | 0 |
| | -0.5000 | 0 |
| | -0.2500 | 0 |
| 0.0000 ($\bar{u}_i$) | 0.0000 | 1 |
| 0.0000 ($\bar{u}_{i+1}$) | 0.0000 | 4/5 |
| | 0.2000 | 1/2 |
| | 0.5000 | 1/4 |
| | 0.7500 | 0 |
| 1.0000 ($\bar{u}_{i+2}$) | 1.0000 | 0 |
| | 1.3333 | 0 |
| | 1.6250 | 0 |
| | 1.7500 | 0 |
| 2.0000 ($\bar{u}_{i+3}$) | 2.0000 | 0 |
| | 2.5000 | 0 |
| 3.0000 ($\bar{u}_{i+4}$) | 3.0000 | 0 |
| 4.0000 ($\bar{u}_{i+5}$) | 4.0000 | 0 |

and the corresponding graph is

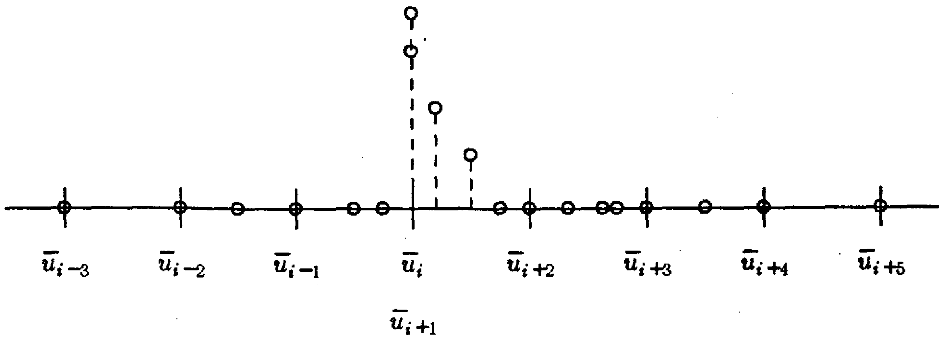


Figure 136. The second-order discrete B-spline with $\bar{u}_i = \bar{u}_{i+1}$. The introduction of a double knot means that $\alpha_{i,2}$ will contribute to $\mathbf{W}_i$ and $\mathbf{W}_{i+1}$. Since both of these weight refined B-splines that become positive at the

same breakpoint, our graph shows two alpha values aligned over that breakpoint.

The message contained in Figure 136 and its accompanying table is that

$$\begin{aligned} \mathbf{W}_j &= 1\mathbf{V}_i \\ \mathbf{W}_{j+1} &= \frac{4}{5}\mathbf{V}_i \ + \text{other } \mathbf{V}\text{'s} \\ \mathbf{W}_{j+2} &= \frac{1}{2}\mathbf{V}_i \ + \text{other } \mathbf{V}\text{'s} \\ \mathbf{W}_{j+3} &= \frac{1}{4}\mathbf{V}_i \ + \text{other } \mathbf{V}\text{'s} \ , \end{aligned}$$

where $j$ is the smallest index for which $\overline{w}_j$=0.0000, and $i$ is likewise the smallest index for which $\overline{u}_i$=0.0000. A double knot at the next $\overline{u}$ position to the right, on the other hand, gives the table

| $\overline{u}$ | $\overline{w}$ | $\alpha_{i,2}$ |
|---|---|---|
| -3.0000 ($\overline{u}_{i-3}$) | -3.0000 | 0 |
| -2.0000 ($\overline{u}_{i-2}$) | -2.0000 | 0 |
| | -1.5000 | 0 |
| -1.0000 ($\overline{u}_{i-1}$) | -1.0000 | 0 |
| | -0.5000 | 0 |
| | -0.2500 | 0 |
| 0.0000 ($\overline{u}_i$) | 0.0000 | 1/5 |
| | 0.2000 | 1/2 |
| | 0.5000 | 3/4 |
| | 0.7500 | 1 |
| 1.0000 ($\overline{u}_{i+1}$) | 1.0000 | 0 |
| 1.0000 ($\overline{u}_{i+2}$) | 1.0000 | 0 |
| | 1.3333 | 0 |
| | 1.6250 | 0 |
| | 1.7500 | 0 |
| 2.0000 ($\overline{u}_{i+3}$) | 2.0000 | 0 |
| | 2.5000 | 0 |
| 3.0000 ($\overline{u}_{i+4}$) | 3.0000 | 0 |
| 4.0000 ($\overline{u}_{i+5}$) | 4.0000 | 0 |

for which the graph is

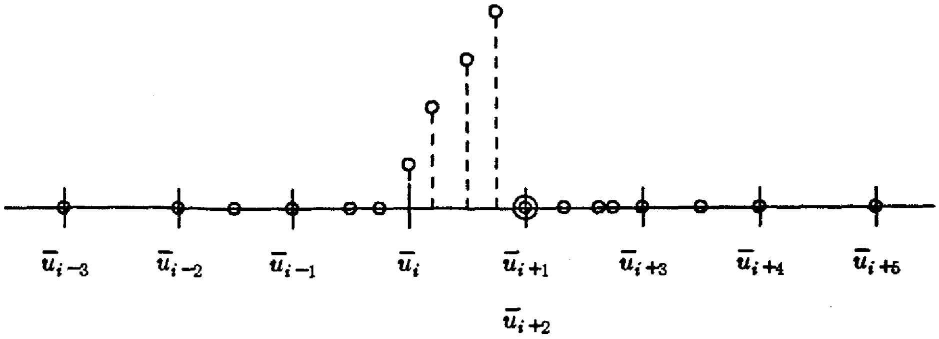


Figure 137. The second-order discrete B-spline with $\overline{u}_{i+1} = \overline{u}_{i+2}$. This time the two aligned alpha values happen to both have the value zero.

Finally, we give a few representative configuations for cubic ($k=4$) discrete B-splines. The simple set of $\overline{u}$ knots which we have been using in these examples yields the table

| $\overline{u}$ | $\overline{w}$ | $\alpha_{i,4}$ |
|---|---|---|
| -3.0000 ($\overline{u}_{i-1}$) | -3.0000 | 0 |
| -2.0000 ($\overline{u}_i$) | -2.0000 | 1/8 |
| | -1.5000 | 7/16 |
| -1.0000 ($\overline{u}_{i+1}$) | -1.0000 | 5/8 |
| | -0.5000 | 41/60 |
| | -0.2500 | 19/30 |
| 0.0000 ($\overline{u}_{i+2}$) | 0.0000 | 119/240 |
| | 0.2000 | 5/16 |
| | 0.5000 | 5/32 |
| | 0.7500 | 1/24 |
| 1.0000 ($\overline{u}_{i+3}$) | 1.0000 | 1/96 |
| | 1.3333 | 0 |
| | 1.6250 | 0 |
| | 1.7500 | 0 |
| 2.0000 ($\overline{u}_{i+4}$) | 2.0000 | 0 |
| | 2.5000 | 0 |
| 3.0000 ($\overline{u}_{i+5}$) | 3.0000 | 0 |
| 4.0000 ($\overline{u}_{i+6}$) | 4.0000 | 0 |

and the corresponding graph is

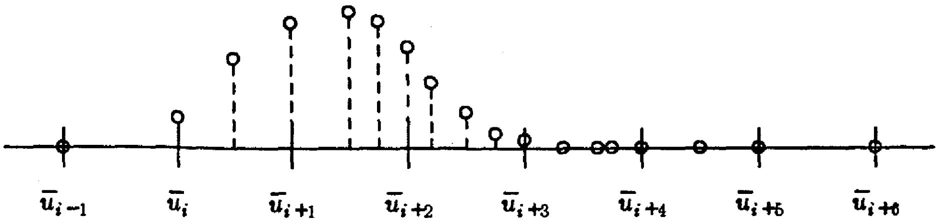


Figure 138. The discrete cubic (order 4) B-spline with no multiplicity. This time the nonzero alpha's span four successive intervals, namely $[\bar{u}_i,\bar{u}_{i+4})$.

Doubling the knot at $\bar{u}_{i+2}$ yields the table

| $\bar{u}$ | $\bar{w}$ | $\alpha_{i,4}$ |
|---|---|---|
| -3.0000 ($\bar{u}_{i-1}$) | -3.0000 | 0 |
| -2.0000 ($\bar{u}_i$) | -2.0000 | 3/16 |
| | -1.5000 | 21/32 |
| -1.0000 ($\bar{u}_{i+1}$) | -1.0000 | 3/4 |
| | -0.5000 | 5/8 |
| | -0.2500 | 2/5 |
| 0.0000 ($\bar{u}_{i+2}$) | 0.0000 | 1/5 |
| 0.0000 ($\bar{u}_{i+3}$) | 0.0000 | 1/20 |
| | 0.2000 | 0 |
| | 0.5000 | 0 |
| | 0.7500 | 0 |
| 1.0000 ($\bar{u}_{i+4}$) | 1.0000 | 0 |
| | 1.3333 | 0 |
| | 1.6250 | 0 |
| | 1.7500 | 0 |
| 2.0000 ($\bar{u}_{i+5}$) | 2.0000 | 0 |
| | 2.5000 | 0 |
| 3.0000 ($\bar{u}_{i+6}$) | 3.0000 | 0 |
| 4.0000 ($\bar{u}_{i+7}$) | 4.0000 | 0 |

and the graph

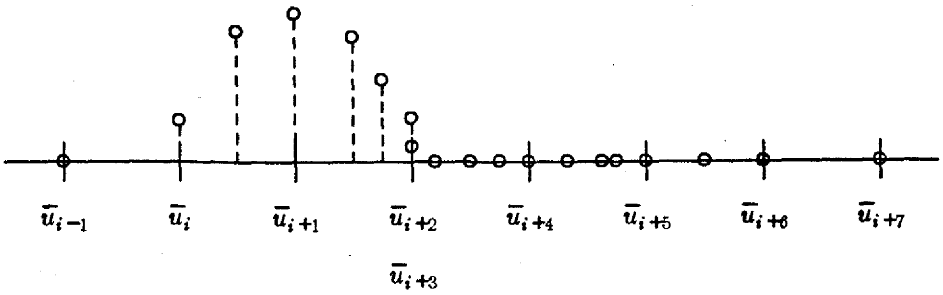


Figure 139. The cubic discrete B-spline with $\bar{u}_{i+2} = \bar{u}_{i+3}$.

Tripling the knot at the same position results in the table

| $\bar{u}$ | $\bar{w}$ | $\alpha_{i,4}$ |
|---|---|---|
| -3.0000 ($\bar{u}_{i-1}$) | -3.0000 | 0 |
| -2.0000 ($\bar{u}_i$) | -2.0000 | 3/16 |
| | -1.5000 | 21/32 |
| -1.0000 ($\bar{u}_{i+1}$) | -1.0000 | 9/16 |
| | -0.5000 | 1/4 |
| | -0.2500 | 0 |
| 0.0000 ($\bar{u}_{i+2}$) | 0.0000 | 0 |
| 0.0000 ($\bar{u}_{i+3}$) | 0.0000 | 0 |
| 0.0000 ($\bar{u}_{i+4}$) | 0.0000 | 0 |
| | 0.2000 | 0 |
| | 0.5000 | 0 |
| | 0.7500 | 0 |
| 1.0000 ($\bar{u}_{i+5}$) | 1.0000 | 0 |
| | 1.3333 | 0 |
| | 1.6250 | 0 |
| | 1.7500 | 0 |
| 2.0000 ($\bar{u}_{i+6}$) | 2.0000 | 0 |
| | 2.5000 | 0 |
| 3.0000 ($\bar{u}_{i+7}$) | 3.0000 | 0 |
| 4.0000 ($\bar{u}_{i+8}$) | 4.0000 | 0 |

for which the graph is

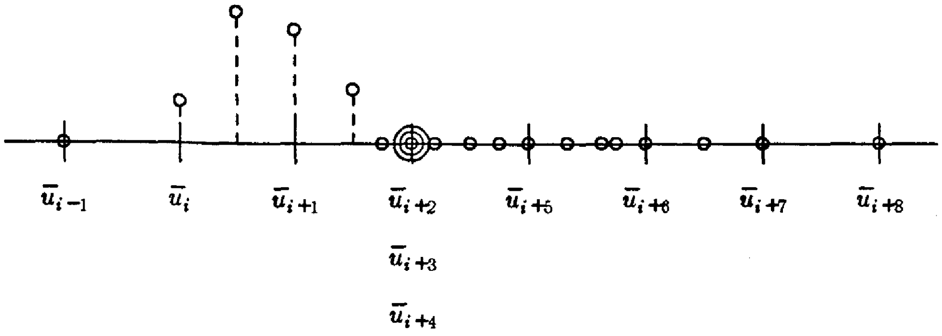


Figure 140. The cubic discrete B-spline with $\bar{u}_{i+2} = \bar{u}_{i+3} = \bar{u}_{i+4}$.

In none of the above examples did we explore the effect of increasing the multiplicity of one or more of the $\bar{w}$ knots. The effect of doing this, like the effect of increasing the multiplicity of the $\bar{u}$ knots, is to "shorten" the interval on which one or more of the $\alpha$'s is nonzero.

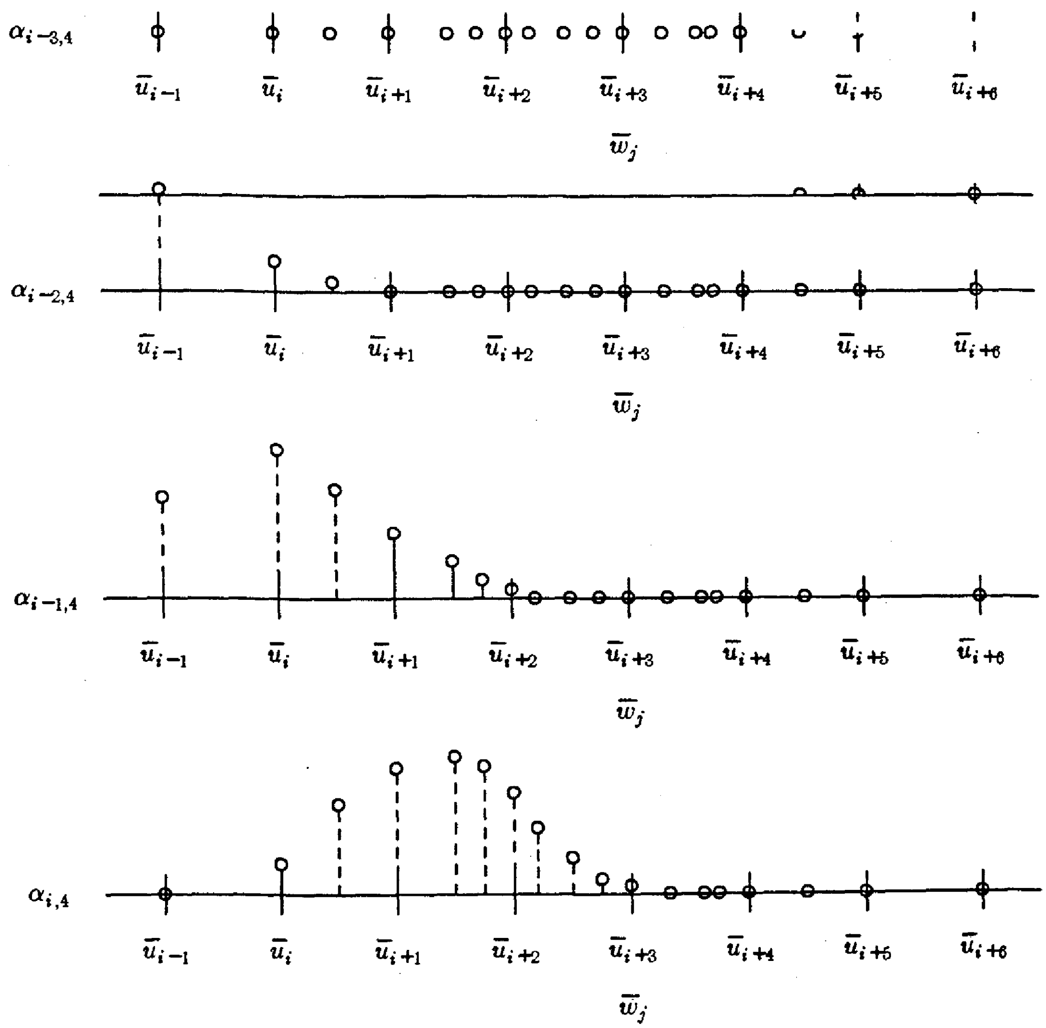
The above diagrams and tables have served, we hope, to convey a feeling for the behaviour of $\alpha_{i,k}(j)$ for fixed $i$ as a function of $j$. It is equally useful to observe how the $\alpha$'s behave for fixed $j$ and a sequence of successive $i$'s. Since

$$\mathbf{W}_j = \sum_{i=0}^{m} \mathbf{V}_i \alpha_{i,k}(j) \quad \text{for} \quad j = 0, \ldots, m+n \ ,$$

this tells us how various $\mathbf{V}_i$'s are weighted in computing a particular $\mathbf{W}_j$. The following is an example for $k=4$ using the $\bar{u}$ and $\bar{w}$ knots of Figure 138. Fixing our attention on $\bar{w}_j = 1.0000$ we compute succesive $\alpha$'s from $\alpha_{i-3,4}(\bar{w}_j)$ to $\alpha_{i+4,4}(\bar{w}_j)$ to be

| $\bar{w}_j = 1.0000$ | |
|---|---|
| $\alpha_{i-3,4}(j)$ | 0 |
| $\alpha_{i-2,4}(j)$ | 0 |
| $\alpha_{i-1,4}(j)$ | 0 |
| $\alpha_{i,4}(j)$ | 1/96 |
| $\alpha_{i+1,4}(j)$ | 251/576 |
| $\alpha_{i+2,4}(j)$ | 19/36 |
| $\alpha_{i+3,4}(j)$ | 5/192 |
| $\alpha_{i+4,4}(j)$ | 0 |

The graphs of these $\alpha$'s in a neighbourhood of $\bar{w}_j$ are shown below. The above table should be read as a "vertical slice" through those plots at the position indicated by the label $\bar{w}_j$.



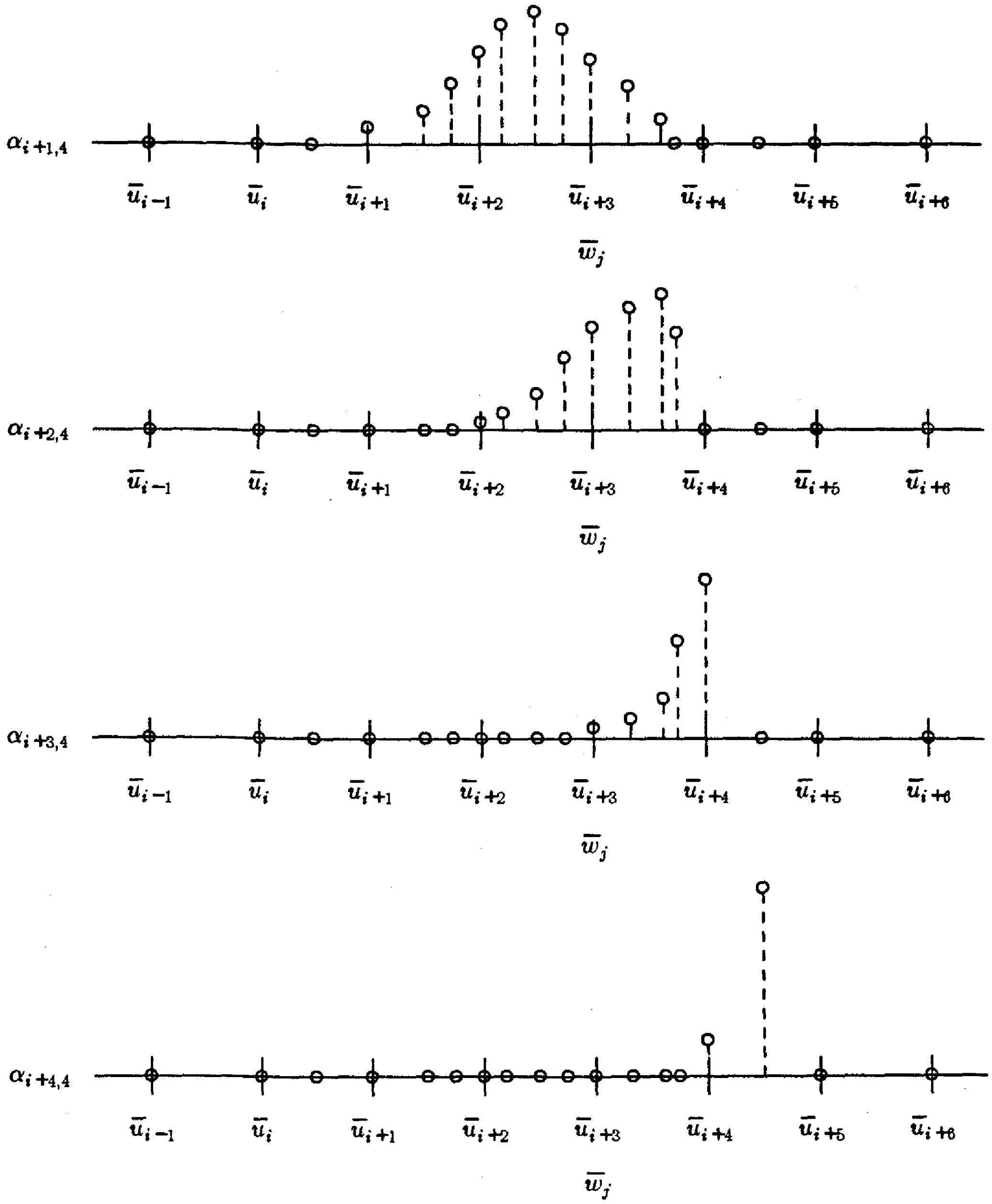


Figure 141. A sequence of successive $\alpha$'s.

Figure 141 and its accompanying table, indicate that

$$\mathbf{W}_j = \frac{1}{96}\mathbf{V}_i + \frac{251}{576}\mathbf{V}_{i+1} + \frac{19}{36}\mathbf{V}_{i+2} + \frac{5}{192}\mathbf{V}_{i+3} .$$

The value of $\bar{w}_j$ being considered is $\bar{w}_j = \bar{u}_{i+3}$. Consequently, it falls in the interval

$$\bar{u}_{i+3} \le \bar{w}_j < \bar{u}_{i+4}$$

Letting $\delta = i+3$ be the index such that $\bar{u}_\delta \le \bar{w}_j < \bar{u}_{\delta+1}$, we observe that

$$\alpha_{0,4}(\overline{w}_j) = \cdots = \alpha_{\delta-4,4}(\overline{w}_j) = 0$$
$$\alpha_{\delta+1,4}(\overline{w}_j) = \cdots = \alpha_{m,4}(\overline{w}_j) = 0$$

and

$$\alpha_{\delta-3,4}(\overline{w}_j)\ ,\ \alpha_{\delta-2,4}(\overline{w}_j)\ ,\ \alpha_{\delta-1,4}(\overline{w}_j)\ ,\ \alpha_{\delta,4}(\overline{w}_j) \neq 0$$

Remarkably, these nonzero values sum to one.

In the light of these preliminary remarks and demonstrations we state the following

**Properties:**

1. For any given $j$ let $\delta(j) = \delta$ be such that $\overline{u}_\delta \le \overline{w}_j < \overline{u}_{\delta+1}$. Then $\alpha_{i,k}(j) = 0$ for $i \notin \{\delta-k+1, \ldots, \delta\}$, for $0 \le i \le m$.
2. $\alpha_{i,k}(j) \ge 0$ for all $i,j,k$.
3. $\sum_{i=0}^{m} \alpha_{i,k}(j) = 1$.

Property 1 establishes the locality of the refinement process by saying that at most $k$ discrete B-splines can be nonzero for any fixed value of $j$, meaning that the refinement process will produce new control vertices $\mathbf{W}_j$, each of which depends on no more than $k$ of the original control vertices $\mathbf{V}_i$. More specifically, $\mathbf{W}_j$ will depend upon some subset of $\mathbf{V}_{\delta-k+1}, \ldots, \mathbf{V}_\delta$, these being the control vertices that are weighted by the B-splines whose support includes the new knot. In particular, introducing a new knot (and a therefore a new control vertex) changes at most $k$ of the old vertices.

Properties 2 and 3 together establish a geometric containment property, namely that $\mathbf{W}_j$ will be a weighted average of the members of $\mathbf{V}_{\delta-k+1}, \ldots, \mathbf{V}_\delta$. That is, $\mathbf{W}_j$ will lie in their convex hull. We will establish these properties formally using the argument to be found in [Cohen80].

**Argument:**

We will establish properties 1 and 2 first, using an induction argument. Property 3 will follow from the fact that the functions $N_{j,k}(\overline{u})$ constitute a basis for $\mathrm{S}(\mathbf{P}^k, \{\overline{w}_j\}_0^{m+n+k})$, that they sum to one, and that the functions $B_{i,k}(\overline{u})$ are representable in terms of the $N$'s. We begin with property 1.

For first-order $\alpha$'s (that is, for $k = 1$), properties 1 and 2 may be taken as evident, by inspection, from the consideration of pictures such as Figure 134. Alternatively, note from the recurrence that

$$\alpha_{i,1}(j) = 0 \quad \text{for} \quad \overline{w}_j < \overline{u}_i \quad \text{and} \quad \overline{w}_j \ge \overline{u}_{i+1}$$

and that

$$\alpha_{i,1}(j) = 1 \quad \text{for} \quad \overline{w}_j \ge \overline{u}_i \quad \text{and} \quad \overline{w}_j < \overline{u}_{i+1} \ .$$

In this context it follows directly from the recurrence that

- for fixed $j$, $\alpha_{i,1}(j) = 0$ when we do not have $\overline{u}_i \le \overline{w}_j < \overline{u}_{i+1}$ (that is, property 1 holds);

(more...)

- $\alpha_{i,1}(j) \geq 0$; that is, property 2 holds.

We next establish inductively that properties 1 and 2 hold for higher-order $\alpha$'s. Assume that they hold for $\alpha_{i,k-1}(j)$ for all $i$ and $j$ and for some $k > 1$.

For property 1 we wish to show that, for fixed $j$ and for $\delta = \delta(j)$ defined by $\bar{u}_\delta \leq \bar{w}_j < \bar{u}_{\delta+1}$, it is true that $\alpha_{i,k}(j) = 0$ whenever $i \notin \{\delta-k+1, \ldots, \delta\}$.

Recall the recurrence

$$\alpha_{i,k}(j) = \frac{\bar{w}_{j+k-1}-\bar{u}_i}{\bar{u}_{i+k-1}-\bar{u}_i}\,\alpha_{i,k-1}(j) + \frac{\bar{u}_{i+k}-\bar{w}_{j+k-1}}{\bar{u}_{i+k}-\bar{u}_{i+1}}\,\alpha_{i+1,k-1}(j) \ .$$

By the induction hypothesis the factor $\alpha_{i,k-1}(j)$ is zero for all $i \notin \{\delta-(k-1)+1, \ldots, \delta\}$; that is, for all $i \notin \{\delta-k+2, \ldots, \delta\}$. Similarly, the factor $\alpha_{i+1,k-1}(j)$ is zero for all $i+1 \notin \{\delta-k+2, \ldots, \delta\}$; that is, for all $i \notin \{\delta-k+1, \ldots, \delta-1\}$. Taking the union of these two index sets, we see that both terms in the recurrence are zero when

$$i \notin \{\delta-k+1, \ldots, \delta\} \ ,$$

which establishes property 1.

To establish property 2, it is instrumental to establish a stronger version of property 1; namely, we show that

$$\alpha_{i,k}(j) = 0 \quad \text{when} \quad \bar{w}_j < \bar{u}_i \quad \text{or} \quad \bar{w}_{j+k-1} \geq \bar{u}_{i+k} \ .$$

The fact that $\alpha_{i,k}(j) = 0$ when $\bar{w}_j < \bar{u}_i$ is immediate from property 1, since this implies that $i \geq \delta+1$. For the other part of the desired result, we work by induction.

Note that by definition

$$\alpha_{i,1}(j) = 0 \quad \text{when} \quad \bar{w}_j \geq \bar{u}_{i+1} \ .$$

But, when $k = 1$,

$$\bar{w}_j = \bar{w}_{j+k-1} \quad \text{and} \quad \bar{u}_{i+1} = \bar{u}_{i+k} \ ,$$

and again the desired result follows directly from the definition.

Suppose that $\alpha_{i,k-1}(j) = 0$ when $\bar{w}_{j+(k-1)-1} \geq \bar{u}_{i+(k-1)}$ for some $k > 1$. Assume, now, that $\bar{w}_{j+k-1} \geq \bar{u}_{i+k}$. Certain conclusions can be drawn from this assumption by the way in which the knot refinement proceeds. Firstly,

$$\bar{w}_{j+k-1} \geq \bar{w}_{j+k-2} = \bar{w}_{j+(k-1)-1} \ ,$$

since the $\bar{w}$ knots are indexed in monotone order. Similarly,

$$\bar{u}_{i+k} \geq \bar{u}_{i+k-1} = \bar{u}_{i+(k-1)} \ .$$

Secondly,

(more...)

$$\bar{w}_{j+k-2} \geq \bar{u}_{i+k-1} \ ,$$

since the $\bar{w}$ knots contain all of the $\bar{u}$ knots.

Now consider the expression for $\alpha_{i,k}(j)$ from the recurrence:

$$\alpha_{i,k}(j) = \frac{\bar{w}_{j+k-1} - \bar{u}_i}{\bar{u}_{i+k-1} - \bar{u}_i} \alpha_{i,k-1}(j) + \frac{\bar{u}_{i+k} - \bar{w}_{j+k-1}}{\bar{u}_{i+k} - \bar{u}_{i+1}} \alpha_{i+1,k-1}(j) \ .$$

Since $\bar{w}_{j+k-1} \geq \bar{u}_{i+k}$ implies that $\bar{w}_{j+(k-1)-1} \geq \bar{u}_{i+(k-1)}$, this causes $\alpha_{i,k-1}(j)$ to be zero, by assumption. Hence, the first term of the recurrence is zero. For the second term, notice that, if

$$\bar{w}_{j+k-1} = \bar{u}_{i+k} \ ,$$

then the numerator of the fraction in the second term is zero. On the other hand, if

$$\bar{w}_{j+k-1} > \bar{u}_{i+k} \ ,$$

then

$$\bar{w}_{j+k-2} \geq \bar{u}_{i+k} \ ,$$

because the $\bar{w}$ and $\bar{u}$ knots are monotonically indexed, and the $\bar{w}$'s contain the $\bar{u}$'s. But since $\bar{w}_{j+k-2} = \bar{w}_{j+(k-1)-1}$ and $\bar{u}_{i+k} = \bar{u}_{(i+1)+(k-1)}$, this implies that $\alpha_{i+1,k-1}(j) = 0$.

This establishes the stronger version of property 1.

Property 2 is now easily established. We assume that the $\alpha$'s of order $k-1$ are nonnegative. Again, we consider the recurrence. For the first term, if it is possible that $\alpha_{i,k-1}(j)$ is not zero, then we must have $i \in \{\delta-k+2, \ldots, \delta\}$, which means that

$$\bar{u}_i \leq \bar{w}_j$$

by the definition of $\delta$; consequently

$$\bar{u}_i < \bar{w}_{j+k-1} \ .$$

This means that the numerator of the fraction in the first term is positive.

For the second term, if $\bar{w}_{j+k-1} \geq \bar{u}_{i+k}$, then we have shown that the value of $\alpha_{i+1,k-1}(j)$ is zero. In all other cases, the numerator of the fraction in the second term is positive. Further, by assumption, the values of $\alpha_{i,k-1}(j)$ and $\alpha_{i+1,k-1}(j)$ are nonnegative. Consequently, the recurrence shows that $\alpha_{i,k}(j)$ can be written as the sum of two nonnegative terms.

This establishes property 2.

Property 3 is immediately established for all $k$ by an observation about linear independence. We know that

$$\sum_{j=0}^{m+n} N_{j,k}(\bar{u}) = 1 \ .$$

Furthermore, by the linear independence of the $N$'s the only linear combination of $N_{j,k}(\bar{u})$ which can sum to 1 is the combination shown here, i.e. the one having all coefficients equal to 1. But recall that

(more...)

$$B_{i,k}(\bar{u}) = \sum_{j=0}^{m+n} \alpha_{i,k}(j) N_{j,k}(\bar{u})$$

for each i. Summing both sides of this equation on $i$ yields

$$\begin{aligned} 1 &= \sum_{i=0}^{m} B_{i,k}(\bar{u}) = \sum_{i=0}^{m} \sum_{j=0}^{m+n} \alpha_{i,k}(j) N_{j,k}(\bar{u}) \\ &= \sum_{j=0}^{m+n} \left[ \sum_{i=0}^{m} \alpha_{i,k}(j) \right] N_{j,k}(\bar{u}) \ . \end{aligned}$$

By the uniqueness of the coefficients which will yield a linear combination of 1 with the $N_{j,k}(\bar{u})$, it follows that

$$\sum_{i=0}^{m} \alpha_{i,k}(j) = 1 \ ,$$

which establishes property 3.

### 13.3. Control-vertex Recurrence

The final theoretical remark which we have to make concerns the control vertices themselves. Recall that

$$\mathbf{W}_j = \sum_{i=0}^{m} \alpha_{i,k}(j) \mathbf{V}_i \ .$$

We now see that

$$\mathbf{W}_j = \sum_{i=\delta-k+1}^{\delta} \alpha_{i,k}(j) \mathbf{V}_i \ ,$$

which means that the $\mathbf{W}_j$ "depend locally" on the $\mathbf{V}_i$ in the sense that adding knots in a certain region of $\bar{u}$ will only change the control vertices being weighted by the B-splines whose nonzero intervals are touched by these new knots. Moreover, since

$$\sum_{i=\delta-k+1}^{\delta} \alpha_{i,k}(j) = 1$$

and the alpha values are nonnegative, $\mathbf{W}_j$ must be a weighted average of the vertices $\mathbf{V}_i$. Like the spline curve which both the **V**'s and the **W**'s define, each $\mathbf{W}_j$ lies in the convex hull of $k$ succesive vertices $\mathbf{V}_i$.

Finally, note that the recurrence for the $\alpha$'s can be applied to produce

$$\begin{aligned} \mathbf{W}_j &= \sum_{i=\delta-k+1}^{\delta} \alpha_{i,k}(j) \mathbf{V}_i \qquad (85) \\ &= \sum_{i=\delta-k+1}^{\delta} \left[ \frac{\bar{w}_{j+k-1} - \bar{u}_i}{\bar{u}_{i+k-1} - \bar{u}_i} \alpha_{i,k-1}(j) + \frac{\bar{u}_{i+k} - \bar{w}_{j+k-1}}{\bar{u}_{i+k} - \bar{u}_{i+1}} \alpha_{i+1,k-1}(j) \right] \mathbf{V}_i \ . \end{aligned}$$

This develops into a recurrence for the control vertices themselves. Property 3 of the previous section guarantees that

$$\alpha_{\delta-k+1,k-1}(j) = \alpha_{\delta+1,k-1}(j) = 0 \ ,$$

which permits us to rearrange and collect terms in (85) to obtain

$$\mathbf{W}_j = \sum_{i=\delta-k+2}^{\delta} \alpha_{i,k-1}(j)\mathbf{V}_{i,2}$$

where

$$\mathbf{V}_{i,2} = \Big[ (\overline{w}_{j+k-1}-\overline{u}_i)\mathbf{V}_i + (\overline{u}_{i+k-1}-\overline{w}_{j+k-1})\mathbf{V}_{i-1} \Big] / (\overline{u}_{i+k-1}-\overline{u}_i)$$

This may be repeated to yield

**Control-Vertex Recurrence:**

Let

$$\mathbf{V}_{i,1} = \mathbf{V}_i$$

and

$$\mathbf{V}_{i,r} = \Big[ (\overline{w}_{j+k-r+1}-\overline{u}_i)\mathbf{V}_{i,r-1} + (\overline{u}_{i+k-r+1}-\overline{w}_{j+k-r+1})\mathbf{V}_{i-1,r-1} \Big] / (\overline{u}_{i+k-r+1}-\overline{u}_i)$$

for $r=2, \ldots, k$ (interpreted as zero when the denominator is zero).

Then

$$\mathbf{W}_j = \mathbf{V}_{\delta,k} \ ,$$

where $\delta = \delta(j)$ is the unique index for which $\overline{u}_\delta \le \overline{w}_j < \overline{u}_{\delta+1}$.

This permits the direct computation of the $\mathbf{W}$ vertices from the $\mathbf{V}$ vertices using only the knots $\{\overline{u}_i\}_0^{m+k}$ and $\{\overline{w}_j\}_0^{m+n+k}$.

## 13.4. Illustrations

We close with some examples of this process. For the first example consider the curve of Figure 142.

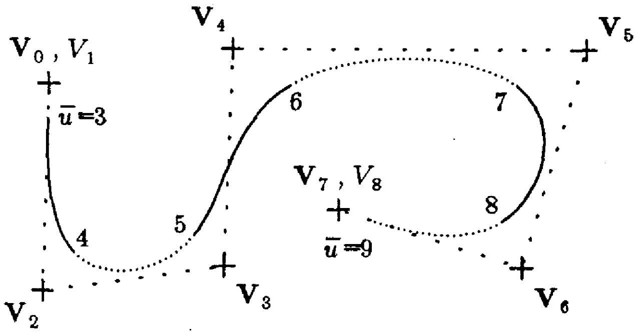


Figure 142. A uniform B-spline curve with doubled initial and final control vertices.

The B-splines which weight the nine control vertices are defined on the uniform knots $\overline{u}_i = i$, and the beginning and ending control vertices are repeated once. The coordinates $x_i$, $y_i$ of the $\mathbf{V}_i$ and the knots $\bar{u}_i$ are given below:

| $x$ | $y$ | |
|---|---|---|
| 0.4568 | 1.3369 | |
| 0.4568 | 1.3369 | |
| 0.4122 | 0.2562 | $\mathbf{V}_2$ |
| 1.3482 | 0.3788 | $\mathbf{V}_3$ |
| 1.4100 | 1.5153 | |
| 3.2199 | 1.4930 | |
| 2.8746 | 0.3565 | |
| 1.9387 | 0.6685 | |
| 1.9387 | 0.6685 | |

We have flagged control vertices $\mathbf{V}_2$ and $\mathbf{V}_3$ because, if we introduce a new knot at $\bar{u}=4.5$, precisely these vertices change. The new control vertices $\mathbf{W}_j$ prove to be

| $x$ | $y$ | |
|---|---|---|
| 0.4568 | 1.3369 | |
| 0.4568 | 1.3369 | |
| 0.4196 | 0.4363 | $\mathbf{W}_2$ |
| 0.8802 | 0.3175 | $\mathbf{W}_3$ |
| 1.3585 | 0.5682 | $\mathbf{W}_4$ |
| 1.4100 | 1.5153 | |
| 3.2199 | 1.4930 | |
| 2.8746 | 0.3565 | |
| 1.9387 | 0.6685 | |
| 1.9387 | 0.6685 | |

and the corresponding control polygon is shown in Figure 143.

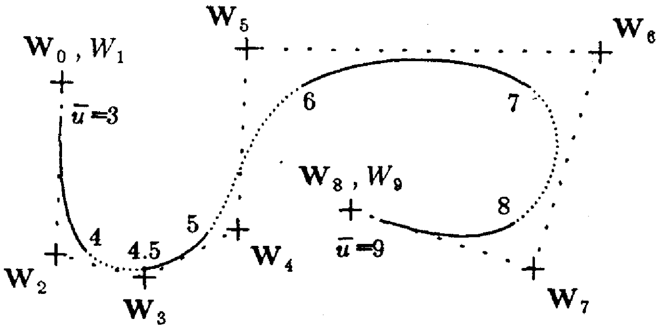


Figure 143. The curve of Figure 142 after the addition of a knot at $\bar{u}=4.5$.

In particular, notice that the three new control vertices $\mathbf{W}_2$, $\mathbf{W}_3$, and $\mathbf{W}_4$ lie closer to the curve than did the two vertices $\mathbf{V}_2$ and $\mathbf{V}_3$ which they replace.

For a second illustration of subdivision, suppose we consider uniform knots $\bar{u}_i$ and add a new knot at the midpoint of each knot interval of the parameter range: $[\bar{u}_i, \bar{u}_{i+1})$, $i = 3, \ldots, m$. For example, when $m = 8$

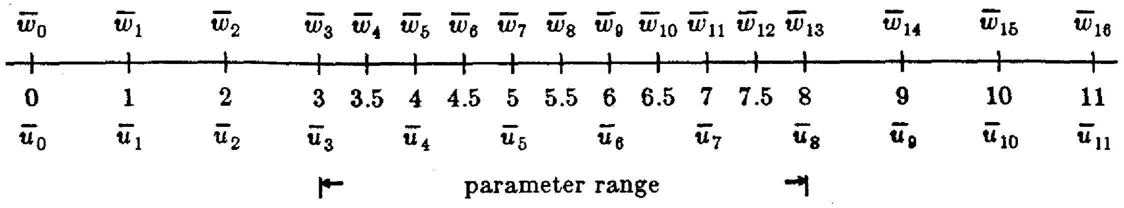

$\bar{w}_0$ $\bar{w}_1$ $\bar{w}_2$ $\bar{w}_3$ $\bar{w}_4$ $\bar{w}_5$ $\bar{w}_6$ $\bar{w}_7$ $\bar{w}_8$ $\bar{w}_9$ $\bar{w}_{10}$ $\bar{w}_{11}$ $\bar{w}_{12}$ $\bar{w}_{13}$ $\bar{w}_{14}$ $\bar{w}_{15}$ $\bar{w}_{16}$
0 1 2 3 3.5 4 4.5 5 5.5 6 6.5 7 7.5 8 9 10 11
$\bar{u}_0$ $\bar{u}_1$ $\bar{u}_2$ $\bar{u}_3$ $\bar{u}_4$ $\bar{u}_5$ $\bar{u}_6$ $\bar{u}_7$ $\bar{u}_8$ $\bar{u}_9$ $\bar{u}_{10}$ $\bar{u}_{11}$
parameter range

Figure 144. The special case of uniform knot spacing and refinement by midpoints.

Then the V's and W's will be related as follows

| | $\mathbf{W}_0$ | $\mathbf{W}_1$ | $\mathbf{W}_2$ | $\mathbf{W}_3$ | $\mathbf{W}_4$ | $\mathbf{W}_5$ | $\mathbf{W}_6$ | $\mathbf{W}_7$ | $\mathbf{W}_8$ | $\mathbf{W}_9$ | $\mathbf{W}_{10}$ | $\mathbf{W}_{11}$ | $\mathbf{W}_{12}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $\mathbf{V}_0$ | 1 | $\frac{1}{6}$ | | | | | | | | | | | |
| $\mathbf{V}_1$ | | $\frac{5}{6}$ | $\frac{1}{2}$ | $\frac{1}{8}$ | | | | | | | | | |
| $\mathbf{V}_2$ | | | $\frac{1}{2}$ | $\frac{3}{4}$ | $\frac{1}{2}$ | $\frac{1}{8}$ | | | | | | | |
| $\mathbf{V}_3$ | | | | $\frac{1}{8}$ | $\frac{1}{2}$ | $\frac{3}{4}$ | $\frac{1}{2}$ | $\frac{1}{8}$ | | | | | |
| $\mathbf{V}_4$ | | | | | | $\frac{1}{8}$ | $\frac{1}{2}$ | $\frac{3}{4}$ | $\frac{1}{2}$ | $\frac{1}{8}$ | | | |
| $\mathbf{V}_5$ | | | | | | | | $\frac{1}{8}$ | $\frac{1}{2}$ | $\frac{3}{4}$ | $\frac{1}{2}$ | | |
| $\mathbf{V}_6$ | | | | | | | | | | $\frac{1}{8}$ | $\frac{1}{2}$ | $\frac{5}{6}$ | |
| $\mathbf{V}_7$ | | | | | | | | | | | | $\frac{1}{6}$ | 1 |

Figure 145. The control-vertex chart for the refinement shown in Figure 144.

This chart indicates, for example, that $\mathbf{W}_3$ is a weighted average of the V's given by

$$\begin{aligned}\mathbf{W}_3 &= \alpha_{1,4}(3)\mathbf{V}_1 + \alpha_{2,4}(3)\mathbf{V}_2 + \alpha_{3,4}(3)\mathbf{V}_3 \\ &= \frac{1}{8}\mathbf{V}_1 + \frac{3}{4}\mathbf{V}_2 + \frac{1}{8}\mathbf{V}_3 \ .\end{aligned}$$

In Figure 146 we see the control graph introduced in Figure 36 of section 4.6, which has tripled vertices its perimeter.

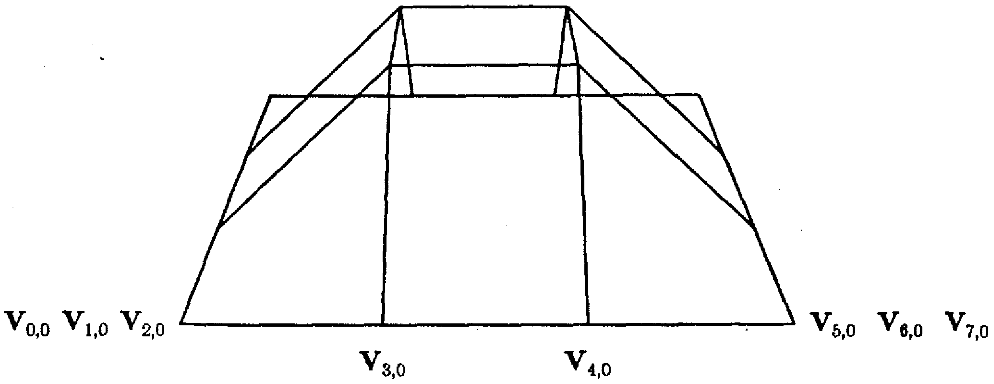


Figure 146. The given, unrefined control graph. Only a sampling of the tripled boundary vertices, varying in the first subscript alone, are labeled to keep the picture uncluttered.

If we halve the knot intervals in both parametric directions as illustrated in Figures 144 and 145, we obtain the control graph of Figure 147.

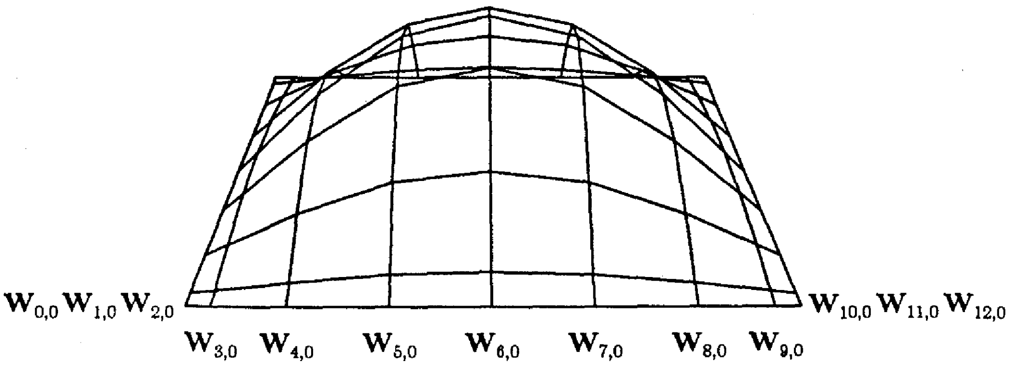

$W_{0,0}$ $W_{1,0}$ $W_{2,0}$
$W_{3,0}$ $W_{4,0}$ $W_{5,0}$ $W_{6,0}$ $W_{7,0}$ $W_{8,0}$ $W_{9,0}$
$W_{10,0}$ $W_{11,0}$ $W_{12,0}$

Figure 147. Refinement achieved by halving the knot intervals.

The surface defined by this control graph, which is swept out by values of $3 \le \bar{u} < 8$ and $3 \le \bar{v} < 8$, can easily be partitioned into four subsurfaces whose parameter ranges are

Surface 1: $3 \le \bar{u} < 5.5$ , $3 \le \bar{v} < 5.5$

Surface 2: $3 \le \bar{u} < 5.5$ , $5.5 \le \bar{v} < 8$

Surface 3: $5.5 \le \bar{u} < 8$ , $3 \le \bar{v} < 5.5$

Surface 4: $5.5 \le \bar{u} < 8$ , $5.5 \le \bar{v} < 8$

whose control vertices are, respectively,

$$\text{Surface 1:}\quad \begin{array}{ccc} \mathbf{W}_{0,0} & \cdots & \mathbf{W}_{7,0} \\ \vdots & \vdots & \vdots \\ \mathbf{W}_{0,7} & \cdots & \mathbf{W}_{7,7} \end{array}$$

$$\text{Surface 2:}\quad \begin{array}{ccc} \mathbf{W}_{0,5} & \cdots & \mathbf{W}_{7,5} \\ \vdots & \vdots & \vdots \\ \mathbf{W}_{0,12} & \cdots & \mathbf{W}_{7,12} \end{array}$$

$$\text{Surface 3:}\quad \begin{array}{ccc} \mathbf{W}_{5,0} & \cdots & \mathbf{W}_{12,0} \\ \vdots & \vdots & \vdots \\ \mathbf{W}_{5,7} & \cdots & \mathbf{W}_{12,7} \end{array}$$

$$\text{Surface 4:}\quad \begin{array}{ccc} \mathbf{W}_{5,5} & \cdots & \mathbf{W}_{12,5} \\ \vdots & \vdots & \vdots \\ \mathbf{W}_{5,12} & \cdots & \mathbf{W}_{12,12} \end{array} .$$

This observation lays the foundation for a process of "subdivided refinement." Specifically, each of surfaces 1 through 4 can be regarded as totally separate from the other three and can be subjected independently to further applications of the Oslo algorithm. It is in this manner that the Oslo refinement can be brought into cooperation with the subdivision schemes introduced by Catmull [Catmull75, Catmull74].

Considering once more the complete surface, if we again halve the knot intervals then the control graph of Figure 148 results.

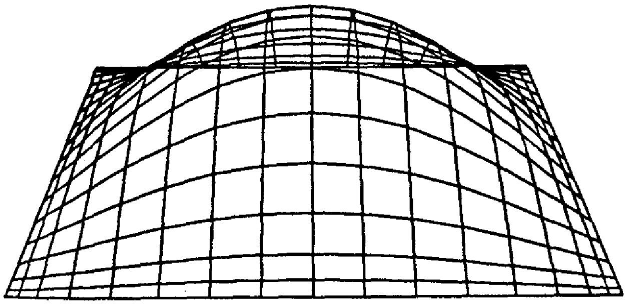


Figure 148. A further halving of the knot intervals.

Note that this particular sequence of refinements is creating control graphs which are quite clearly converging to the spline surface which the graphs define (refer to Figure 38 in Section 4.7.1). In fact, any sequence of refinements in which the spacing between the knots tends to zero throughout the parameter range will show this pattern of convergence.

# 14. Rendering and Evaluation

We will cover ways to obtain values of B-splines and their derivatives. This will provide the background with which to cover some of the methods that have been proposed in the literature to render B-spline curves and surfaces graphically. Because we are concerned with interactive applications, we will lay the main emphasis on evaluation techniques that are as rapid as possible. But one rendering technique, ray-tracing, which is by no stretch of the imagination suitable for interactive systems, has been included because it offers an interesting use of the Oslo algorithm.

## 14.1. Derivatives of B-splines

Many of the evaluation and rendering processes used upon parametric curves and surfaces in computer graphics require the knowledge of parametric derivatives, i.e.

$$\frac{d^r}{d\bar{u}^r}\,\mathbf{Q}(\bar{u})\ ,\quad \frac{\partial^r}{\partial\bar{u}^r}\,\mathbf{Q}(\bar{u},\bar{v})\ ,\quad \frac{\partial^r}{\partial\bar{v}^r}\,\mathbf{Q}(\bar{u},\bar{v})\ \ .$$

For parametric curves and surfaces generated by B-splines, these derivatives require the evaluation of the derivatives of individual B-splines:

$$D_{\bar{u}}^{(r)}B_{i,k}(\bar{u}) \;=\; B_{i,k}^{(r)}(\bar{u})\ \ .$$

When the segment polynomials of $B_{i,k}(\bar{u})$ are known explicitly, finding derivatives is no more than a simple exercise in calculus. For example, with uniform knot spacing, $\bar{u}_{i+1} = \bar{u}_i + 1$, the cubic B-spline has the first derivative

$$B_{i,4}^{(1)}(\bar{u}) \;=\; \begin{cases} b_{-0}^{(1)}(u) \;=\; \dfrac{1}{2}\,u^2 & \text{for } \bar{u}_i \le \bar{u} < \bar{u}_{i+1} \text{ and } u = \bar{u}-\bar{u}_i \\ b_{-1}^{(1)}(u) \;=\; \dfrac{1}{6}\,(3+6u-9u^2) & \text{for } \bar{u}_{i+1} \le \bar{u} < \bar{u}_{i+2} \text{ and } u = \bar{u}-\bar{u}_{i+1} \\ b_{-2}^{(1)}(u) \;=\; \dfrac{1}{6}\,(-12u+9u^2) & \text{for } \bar{u}_{i+2} \le \bar{u} < \bar{u}_{i+3} \text{ and } u = \bar{u}-\bar{u}_{i+2} \\ b_{-3}^{(1)}(u) \;=\; \dfrac{1}{6}\,(-3+6u-3u^2) & \text{for } \bar{u}_{i+3} \le \bar{u} < \bar{u}_{i+4} \text{ and } u = \bar{u}-\bar{u}_{i+3}\ . \end{cases}$$

In general situations, however, the individual B-splines are available only through their definitions as divided differences or as the results of a recurrence process. The segment polynomials are not tabulated; hence they can't be called upon to provide derivatives.

We shall show how the derivatives of a B-spline can be obtained from the B-spline recurrence.

Indeed, it is through means of the derivatives, obtained from the recurrence, that the coefficients of the segment polynomials can be tabulated. Since some rendering and evaluation schemes work more efficiently in terms of segment polynomials, this will be an important result.

We remarked in the development of chapter 8 that differentiation with respect to $\bar{u}$ can pass through differencing with respect to $t$. Consequently,

$$\begin{aligned} D^{(1)}_{\bar{u}}B_{i,k}(\bar{u}) &= D^{(1)}_{\bar{u}}(-1)^k(\bar{u}_{i+k}-\bar{u}_i)[\bar{u}_i(k){:}t](\bar{u}-t)_+^{k-1} \\ &= (-1)^k(\bar{u}_{i+k}-\bar{u}_i)[\bar{u}_i(k){:}t]D^{(1)}_{\bar{u}}(\bar{u}-t)_+^{k-1} \\ &= (-1)^k(\bar{u}_{i+k}-\bar{u}_i)[\bar{u}_i(k){:}t](k-1)(\bar{u}-t)_+^{k-2} \ . \end{aligned}$$

If we ignore the terms $(-1)^k$ and $(k-1)$ for a moment, what remains can be written as follows:

$$\begin{aligned} &(\bar{u}_{i+k}-\bar{u}_i)[\bar{u}_i(k){:}t](\bar{u}-t)_+^{k-2} \\ &\quad= (\bar{u}_{i+k}-\bar{u}_i)\frac{[\bar{u}_{i+1}(k-1){:}t](\bar{u}-t)_+^{k-2}-[\bar{u}_i(k-1){:}t](\bar{u}-t)_+^{k-2}}{\bar{u}_{i+k}-\bar{u}_i} \ . \end{aligned} \tag{86}$$

This is true directly from the definition of the divided difference $[\bar{u}_i(k){:}t]$, if $\bar{u}_{i+k} > \bar{u}_i$ (as we generally assume to be the case), but it is also true if $\bar{u}_{i+k}$ and $\bar{u}_i$ have the same value. To see this, note that $\bar{u}_{i+k} = \bar{u}_i$ will make the left side of (86) be equal to

$$(\bar{u}_{i+k}-\bar{u}_i)[\bar{u}_i(k){:}t](\bar{u}-t)_+^{k-2} = 0\cdot\frac{1}{k!}D_t^{(k)}(\bar{u}-t)_+^{k-2}\big|_{t=\bar{u}_i} = 0 \ .$$

For the right side of (86) we may consider $\bar{u}_{i+k} = \bar{u}_i$ to be the limiting case of $\bar{u}_{i+k} \to \bar{u}_i$. Since we should have cancelled out the term $(\bar{u}_{i+k}-\bar{u}_i)$, this will make the right side of (86) be equal to

$$\begin{aligned} &[\bar{u}_{i+1}(k-1){:}t](\bar{u}-t)_+^{k-2} - [\bar{u}_i(k-1){:}t](\bar{u}-t)_+^{k-2} \\ &\quad= \frac{1}{(k-1)!}D_t^{(k-1)}(\bar{u}-t)_+^{k-2}\big|_{t=\bar{u}_{i+1}} - \frac{1}{(k-1)!}D_t^{(k-1)}(\bar{u}-t)_+^{k-2}\big|_{t=\bar{u}_i} \ , \end{aligned}$$

which is zero, since both terms in the difference will be equal in the limit. The result is that both the right and the left side of (86) are zero, hence the equality expressed by (86) is valid. (The case $\bar{u}_{i+k} = \bar{u}_i$ would be pertinent in fully understanding the validity of a recurrence for B-spline derivatives that will appear below. However, we will not be setting up a formal argument for the derivative recurrence, so the case in which $\bar{u}_{i+k} = \bar{u}_i$ will rest here merely as an observation.)

Applying (86) to the differentiation result and then applying the divided difference definition of $k-1^{st}$- order B-splines to the result gives the following:

$$\begin{aligned} D^{(1)}_{\bar{u}}B_{i,k}(\bar{u}) &= (-1)(k-1)(-1)^{k-1}[\bar{u}_{i+1}(k-1){:}t](\bar{u}-t)_+^{k-2} \\ &\quad - (-1)(k-1)(-1)^{k-1}[\bar{u}_i(k-1){:}t](\bar{u}-t)_+^{k-2} \\ &= (k-1)\left[\frac{B_{i,k-1}(\bar{u})}{\bar{u}_{i+k-1}-\bar{u}_i} - \frac{B_{i+1,k-1}(\bar{u})}{\bar{u}_{i+k}-\bar{u}_{i+1}}\right] \ . \end{aligned} \tag{87}$$

For higher derivatives this result could be repeated recursively, but the expressions which result rapidly become quite complicated. A more productive approach, to be found in [deBoor78] and [Schumaker81], involves looking at the derivatives of linear combinations of B-splines:

$$\sum_{i=0}^{m} c_i B_{i,k}(\bar{u}) \ .$$

Since the $x$, $y$, and $z$ components of any B-spline curve or surface will be functions of this form, linear combinations are frequently what we want to work with, rather than with the individual B-splines themselves. If the occasion should arise in which the derivatives of only a single B-spline $B_{s,k}(\bar{u})$ are wanted, this can be handled by noting that, for any specific index, $s$,

$$B_{s,k}(\bar{u}) = \sum_{i=0}^{m} c_i B_{i,k}(\bar{u}) \text{ for } c_0 = \cdots = c_{s-1} = c_{s+1} = \cdots = c_m = 0 \text{ and } c_s = 1 \ .$$

Applying the derivative result, (87), to a linear combination gives:

$$D_{\bar{u}} \sum_{i=0}^{m} c_i B_{i,k}(\bar{u}) = \sum_{i=0}^{m} (k-1) c_i \left[ \frac{B_{i,k-1}(\bar{u})}{\bar{u}_{i+k-1} - \bar{u}_i} - \frac{B_{i+1,k-1}(\bar{u})}{\bar{u}_{i+k} - \bar{u}_{i+1}} \right] .$$

By rearranging the summation to collect together terms which are common to each B-spline, we obtain

$$D_{\bar{u}} \sum_{i=0}^{m} c_i B_{i,k}(\bar{u}) = \sum_{i=0}^{m+1} (k-1) \frac{c_i - c_{i-1}}{\bar{u}_{i+k-1} - \bar{u}_i} B_{i,k-1}(\bar{u}) \ .$$

Notice that this sum calls for values of $c_{-1}$ (for $i=0$) and $c_{m+1}$ (for $i=m+1$). These fictitious coefficients were introduced to unify the summation; we define them to have the value zero.

The gain in taking this approach is the following: we can define

$$c_{i,1} = c_i \quad \text{for } i = 0, \ldots, m$$

and

$$c_{i,2} = (k-1) \frac{c_{i,1} - c_{i-1,1}}{\bar{u}_{i+k-1} - \bar{u}_i} \quad \text{for } i = 0, \ldots, m+1 \ ,$$

and we have the start of a recurrence which can be carried on to higher derivatives.

**Theorem:**

$$D_{\bar{u}}^{(r)} \sum_{i=0}^{m} c_i B_{i,k}(\bar{u}) = \sum_{i=0}^{m+r} c_{i,r+1} B_{i,k-r}(\bar{u})$$

where

$$c_{i,1} = c_i \quad \text{for } i = 0, \ldots, m$$

and for each $j = 2, \ldots, r+1$,

$$c_{i,j} = (k-j+1) \frac{c_{i,j-1} - c_{i-1,j-1}}{\bar{u}_{i+k-j+1} - \bar{u}_i} \tag{88}$$

(more...)

$$\text{for } i = 0, \ldots, m+j-1 \ .$$

The convention is adopted that values of zero are adjoined as necessary at both ends of the index range $i$ to handle values of $c$ which might not otherwise be defined in (88); that is,

$$c_{-1,1} = c_{-1,2} = \cdots = c_{-1,r} = 0$$

$$c_{m+1,1} = c_{m+2,2} = \cdots = c_{m+r,r} = 0 \ .$$

The convention that ratios with zero denominators are taken equal to zero is, of course, also in force; that is,

$$c_{i,j} = 0 \text{ if } \bar{u}_{i+k-j+1} = \bar{u}_i \ .$$

The above theorem is stated for general $\bar{u}$. In a computational setting we will want the value of a derivative at some specific $\bar{u}$ in the legal parameter range. At any specific value of $\bar{u}$, however, many of the terms in the linear combination will disappear, because the corresponding values of the B-splines will be zero. Let us take this into account.

Suppose $\delta$ is the unique index satisfying

$$\bar{u}_\delta \le \bar{u} < \bar{u}_{\delta+1} \ .$$

Then

$$B_{i,k}(\bar{u}) = 0 \text{ for } i \notin \{\delta-k+1, \ldots, \delta\} \ .$$

Consequently,

$$s(\bar{u}) = \sum_{i=0}^{m} c_{i,1} B_{i,k}(\bar{u}) = \sum_{i=\delta-k+1}^{\delta} c_{i,1} B_{i,k}(\bar{u}) \ ,$$

and

$$D_{\bar{u}}^{(1)} s(\bar{u}) = \sum_{i=\delta-k+1}^{\delta+1} c_{i,2} B_{i,k-1}(\bar{u}) \ .$$

But

$$B_{i,k-1}(\bar{u}) = 0 \text{ for } i \notin \{\delta-k+2, \ldots, \delta\} \ ,$$

so the summation for $s^{(1)}(\bar{u})$ reduces to

$$\sum_{i=\delta-k+2}^{\delta} c_{i,2} B_{i,k-1}(\bar{u}) \ .$$

This means that the linear combination which defines $s^{(1)}(\bar{u})$ has one fewer term than that which defines $s(\bar{u})$. Furthermore, the coefficients

$$c_{\delta-k+2,2} \ , \ \ldots \ , \ c_{\delta,2}$$

defining $s^{(1)}(\bar{u})$ will only depend upon

$$c_{\delta-k+1,1}\ , \dots ,\ c_{\delta,1}\ ,$$

all of which are defined. That means that there is no longer any reason to deal with fictitious coefficients.

As a final observation, it should be noted that $D_{\bar{u}}^{(r)} s(\bar{u}) = 0$ if $r \geq k$ for any $\bar{u}$.

The result of these considerations is that the theorem can be abbreviated to serve as a computational process.

**Derivative-Value Recurrence:**

Begin with the coefficients $c_i$ of the linear combination

$$s(\bar{u}) = \sum_{i=0}^{m} c_i B_{i,k}(\bar{u}) \ .$$

Let $\bar{u}$ and $\delta$ be given, with

$$\bar{u}_\delta \leq \bar{u} < \bar{u}_{\delta+1} \ ,$$

and let $r$ be given in the range $1 \leq r < k$.

Set

$$c_{i,1} = c_i \ \text{ for } \ i = \delta-k+1, \dots, \delta \ .$$

For each $j = 2, \dots, r+1$, let

$$c_{i,j} = (k+j-1) \frac{c_{i,j-1} - c_{i-1,j-1}}{\bar{u}_{i+k-j+1} - \bar{u}_i} \tag{89}$$

$$\text{for } \ i = \delta-k+j, \dots, \delta \ .$$

Then the $r^{th}$ derivative of $s(\bar{u})$ at the specific value of $\bar{u}$ in question is given by the linear combination

$$s^{(r)}(\bar{u}) = \sum_{i=\delta-k+r+1}^{\delta} c_{i,r+1} B_{i,k-r}(\bar{u}) \ .$$

The complaint might be raised, now, that this recurrence pushes one computational problem onto another, since a summation must be evaluated in order to define the value of $s^{(r)}(\bar{u})$. We will consider the evaluation of such summations in the next section.

## 14.2. B-Spline Evaluation

We have remarked in section 14.1 that linear combinations of B-splines are the objects of primary concern in rendering and evaluation. This is because they constitute the $x$, $y$, and $z$ coordinates of curves and surfaces, for example,

$$X(\bar{u}) = \sum_i x_i B_{i,k}(\bar{u}) \ ,$$

or

$$X(\bar u,\bar v) = \sum_j d_j(\bar u) B_{j,k}(\bar v) \ ,$$

where for each $j$ and each fixed value of $\bar u$

$$d_j(\bar u) = \sum_i x_{i,j} B_{i,k}(\bar u) \ ,$$

and similarly for $Y$ and $Z$. In this section we will use the following general notation for any such linear combination:

$$s(\bar U) = \sum_{i=0}^{m} C_i B_{i,K}(\bar U) \ . \tag{90}$$

We have used capitol $C$'s for the coefficients, we have replaced $\bar u$ by $\bar U$, and we are using a capitol $K$ for the order of the B-splines, all for reasons that will become clear in the next section.

For any given value of $\bar U$, the corresponding value of (90) can be obtained from the B-spline recurrence without dealing with the values of the individual B-splines. Consider the following:

$$\sum_{i=1}^{m} C_i B_{i,K}(\bar U) = \sum_{i=1}^{m} C_i \left[ \frac{\bar U - \bar u_i}{\bar u_{i+K-1} - \bar u_i} B_{i,K-1}(\bar U) + \frac{\bar u_{i+K} - \bar U}{\bar u_{i+K} - \bar u_{i+1}} B_{i+1,K-1}(\bar U) \right]$$

$$= \sum_{i=0}^{m} C_i \frac{\bar U - \bar u_i}{\bar u_{i+K-1} - \bar u_i} B_{i,K-1}(\bar U) + \sum_{i=0}^{m} C_i \frac{\bar u_{i+K} - \bar U}{\bar u_{i+K} - \bar u_{i+1}} B_{i+1,K-1}(\bar U) \ .$$

The index in the second sum can be shifted, and the two sums can be recombined, to give

$$\sum_{i=0}^{m+1} \left[ \frac{\bar U - \bar u_i}{\bar u_{i+K-1} - \bar u_i} C_i + \frac{\bar u_{i+K-1} - \bar U}{\bar u_{i+K-1} - \bar u_i} C_{i-1} \right] B_{i,K-1}(\bar U) \ ,$$

where the values of $C_{-1}$ and $C_{m+1}$ are taken to be zero.

If we set

$$C_{i,1} = C_i \quad \text{for } i = 0, \ldots, m$$

and

$$C_{i,2} = \frac{\bar U - \bar u_i}{\bar u_{i+K-1} - \bar u_i} C_{i,1} + \frac{\bar u_{i+K-1} - \bar U}{\bar u_{i+K-1} - \bar u_i} C_{i-1,1} \quad \text{for } i = 0, \ldots, m+1 \ , \tag{91}$$

then we have

$$s(\bar U) = \sum_{i=0}^{m+1} C_{i,2} B_{i,K-1}(\bar U)$$

on which we can repeat the above development. This clearly produces yet another recurrence.

**Theorem:**

Let $\bar{U}$ be a given fixed value in the legal parameter range. Let $\delta$ be the unique index such that

$$\bar{u}_\delta \leq \bar{U} < \bar{u}_{\delta+1} \ .$$

Then the value

$$s(\bar{U}) = \sum_{i=0}^{m} C_i B_{i,K}(\bar{U})$$

is given by

$$s(\bar{U}) = C_{\delta,K} \ ,$$

where

$$C_{i,1} = C_i \quad \text{for } i = 0, \ldots, m$$

and

$$C_{i,j} = \frac{\bar{U}-\bar{u}_i}{\bar{u}_{i+K-j+1}-\bar{u}_i} C_{i,j-1} + \frac{\bar{u}_{i+K-j+1}-\bar{U}}{\bar{u}_{i+K-j+1}-\bar{u}_i} C_{i-1,j-1} \quad \text{for } i = 0, \ldots, m+j-1$$

for $j = 2, \ldots, K$, As in the theorem (88) for derivatives, the convention is adopted that values of zero are adjoined as necessary at both ends of the index range $i$ to handle values of $C_{i,j-1}$ and $C_{i-1,j-1}$ which might not otherwise be defined; that is,

$$C_{-1,1} = C_{-1,2} = \cdots = C_{-1,K-1} = 0$$

$$C_{m+1,1} = C_{m+2,2} = \cdots = C_{m+K-1,K-1} = 0 \ .$$

The convention that ratios with zero denominators are taken equal to zero is, of course, also in force; that is,

$$C_{i,j} = 0 \text{ if } \bar{u}_{i+K-j+1} = \bar{u}_i \ .$$

The formal result of the recurrence is actually

$$s(\bar{U}) = \sum_{i=0}^{m+K-1} C_{i,K} B_{i,1}(\bar{U}) \ ,$$

but all of the values $B_{i,1}(\bar{U})$ are zero on the interval $\bar{u}_\delta \leq \bar{U} < \bar{u}_{\delta+1}$ except for $B_{\delta,1}(\bar{U})$, which has the value 1.

The observations which were made in the previous section about computational economies which arise when the specific value of $\bar{U}$ is taken into account may be echoed here to yield another computational recurrence.

**Spline-Value Recurrence:**

Let $\overline{U}$ be given, and take $\delta$ to be the index satisfying

$$\overline{u}_\delta \leq \overline{U} < \overline{u}_{\delta+1} .$$

Let

$$C_{i,1} = C_i \text{ for } i = \delta - K+1, \ldots, \delta .$$

For each for $j = 2, \ldots, K$, let

$$C_{i,j} = \frac{\overline{U}-\overline{u}_i}{\overline{u}_{i+K-j+1}-\overline{u}_i} C_{i,j-1} + \frac{\overline{u}_{i+K-j+1}-\overline{U}}{\overline{u}_{i+K-j+1}-\overline{u}_i} C_{i-1,j-1} \qquad (92)$$

$$\text{for } i = \delta-K+j, \ldots, \delta .$$

Then

$$s(\overline{U}) = C_{\delta,K} .$$

### 14.3. Conversion to Segment Polynomials

Now we will revert to our previous notation:

$$s(\overline{u}) = \sum_{i=0}^{m} c_i B_{i,k}(\overline{u}) . \qquad (93)$$

Given a fixed value of $\overline{u}$, consider the breakpoint interval

$$\overline{u}_\delta \leq \overline{u} < \overline{u}_{\delta+1} .$$

Within such an interval, any linear combination (93) becomes simply a polynomial; hence, it could be expressed as

$$\begin{aligned} s(\overline{u}) &= \sum_{r=0}^{k-1} a_r(\delta)(\overline{u}-\overline{u}_\delta)^r \qquad (94) \\ &= \sum_{r=0}^{k-1} a_r(\delta) u^r \end{aligned}$$

for some collection of coefficients $a_r(\delta)$, where $u = \overline{u} - \overline{u}_\delta$. We will present a computational scheme for making the conversion from the B-spline representation of $s(\overline{u})$, (93), which is valid for all $\overline{u} \in [\overline{u}_{k-1}, \overline{u}_{m+1})$, to the power representation, (94), whose coefficients $a_r(\delta)$ are only valid on the specific breakpoint interval $\overline{u}_\delta \leq \overline{u} < \overline{u}_{\delta+1}$. The reason for wishing to make this conversion will become more evident in the next section, but it derives from the observation that representation (94) can be evaluated more efficiently than (93). The conversion is costly, however, so it will only be interesting when we are faced with the task of evaluating $s(\overline{u})$ several times in succession on a breakpoint interval. In [Schumaker81] it is observed that two or more values of $s(\overline{u})$ on a breakpoint interval would already make the conversion worthwhile for cubic splines.

Note that

$$
\begin{aligned}
D_{\bar{u}}^{(r)} s(\bar{u}) &= D_{\bar{u}}^{(r)} \sum_{j=0}^{k-1} a_j(\delta)(\bar{u}-\bar{u}_\delta)^j \\
&= \sum_{j=r}^{k-1} a_j(\delta)(j)\cdots(j-r+1)(\bar{u}-\bar{u}_\delta)^{j-r} \quad ,
\end{aligned}
$$

and

$$
D_{\bar{u}}^{(r)} s(\bar{u}) \Big|_{\bar{u}=\bar{u}_\delta} = a_r(\delta) r! \quad .
$$

The consequence of this is that it is merely necessary to "dovetail" recurrences (89) and (91) to produce

$$
a_0(\delta) \, , \ldots , \, a_{k-1}(\delta) \quad .
$$

**Segment-Polynomial Conversion Recurrence:**

Let a breakpoint interval $[\bar{u}_\delta, \bar{u}_{\delta+1})$ be given. Let

$$
c_{i,1} = c_i \quad \text{for } i = \delta+k-1, \ldots, \delta \quad .
$$

For each $r = 1, \ldots, k-1$ in turn:

(1) Let $C_i = c_{i,r}$ for $i = \delta+k-r, \ldots, \delta$.

Use recurrence (92) with $\bar{U} = \bar{u}_\delta$, $K = k-r+1$, and coefficients $C_i$ to obtain a value $C_{\delta,K}$.

(2) Set $a_{r-1}(\delta) = \dfrac{1}{(r-1)!} C_{\delta,K}$.

(3) Use one step of recurrence (89) to produce $c_{i,r+1}$ for $i = \delta+k-r+1, \ldots, \delta$.

### 14.4. Rendering Curves: Horner's Rule & Forward Differencing

Our problem is to evaluate piecewise cubic polynomials so that we can display the curve they define. Let us consider the polynomial

$$
p(u) = a + bu + cu^2 + du^3 \quad . \tag{95}
$$

If $p(u)$ is to be approximated by $N$ line segments then we will need to evaluate $p(u)$ at $N+1$ values of $u$. Direct application of equation (95) requires $6(N+1)$ multiplications and $3(N+1)$ additions. However, we can rearrange equation (95) to obtain

$$
p(u) = a + u(b + u(c + du)) \quad . \tag{96}
$$

Evaluating $p(u)$ at $N+1$ values of $u$ using (96), which is called *Horner's rule* or *nested multiplication*, requires $3(N+1)$ multiplications and $3(N+1)$ additions — an improvement. Indeed, Horner's rule is optimal with respect to the number of arithmetic operations if we are evaluating $p(u)$ at a single $u$ value [Aho74].

We can do even better if we are evaluating $p(u)$ at a sequence of equally spaced $u$'s. Suppose that we wish to evaluate $p(u)$ at the $N+1$ positions

$$
u = i \cdot h \quad \text{for } i = 0,1,\ldots,N \quad ,
$$

where $h$ is the *step size*. If we were dealing with a linear polynomial, say

$$q(u) = a + bu$$

we would simply observe that

$$q(u+h) - q(u) = a + b(u+h) - a - bu = bh$$

so that we could efficiently generate the desired $N+1$ points $q_i = (u_i, y_i)$ by computing

```
$u_0 = 0$
$y_0 = a$
for $i \leftarrow 1$ step 1 , . . . , $N$ do
    $u_i = u_{i-1} + h$
    $y_i = y_{i-1} + bh$
endfor
```

Of course, to avoid redundant computation of $bh$ we would precompute it outside the loop:

```
$u_0 = 0$
$y_0 = a$
$\Delta_1 = bh$
for $i \leftarrow 1$ step 1 , . . . , $N$ do
    $u_i = u_{i-1} + h$
    $y_i = y_{i-1} + \Delta_1$
endfor
```

This technique can be generalized to polynomials of higher order. Suppose that $r(u)$ is a quadratic polynomial, say

$$r(u) = a + bu + cu^2 \ .$$

Then

$$\begin{aligned} \Delta_1(u) &= r(u+h) - r(u) \\ &= a + b(u+h) + c(u^2+2hu+h^2) - a - bu - cu^2 \\ &= (bh+ch^2) + (2ch)u \end{aligned}$$

so that

$$r(u+h) = r(u) + \Delta_1(u)$$

where

$$\Delta_1(u) = (bh + ch^2) + (2ch)u \; .$$

However, the technique used to evaluate $q(u+h)$ can't be applied immediately because $\Delta_1(u)$ is a function of $u$, and therefore changes value at each iteration of the loop. We could, of course, simply evaluate $\Delta_1(u)$ at $u=0,h,2h,3h,...Nh$ and use the results to compute $r(u)$ at each of these points. Notice, however, that $\Delta_1(u)$ is a linear polynomial. We already know how to evaluate a linear polynomial efficiently for such a sequence — we simply compute $\Delta_1(0) = bh + ch^2$ and then add

$$\Delta_2(u) = \Delta_1(u+h) - \Delta_1(u) = 2ch^2$$

to $\Delta_1(ih)$ to obtain $\Delta_1(ih+h)$. Altogether, then, our computation now looks like this:

```
$u_0 = 0$
$y_0 = a$
$\Delta_1 = bh + ch^2$
$\Delta_2 = 2ch^2$
for $i \leftarrow 1$ step 1 , . . . , $N$ do
    $u_i = u_{i-1} + h$
    $y_i = y_{i-1} + \Delta_1$
    $\Delta_1 = \Delta_1 + \Delta_2$
endfor
```

Let us recapitulate. When $r(u)$ is a quadratic, the value $\Delta_1(u) = r(u+h) - r(u)$ that must be added to $r(u)$ to obtain $r(u+h)$ is not a constant – it changes value as we move from $u$ to $u+h$. Fortuitously, though, $\Delta_1(u)$ itself is easy to update after we have reached $u+h$ so as to obtain the increment that will be needed to compute $u+2h$ from $u+h$; we simply need to increment $\Delta_1$ by $\Delta_2$, which is the constant $2ch^2$.

We may extend this approach to our cubic polynomial $p(u)$ in much the same way. In this case

$$\begin{aligned}
\Delta_1(u) &= p(u+h) - p(u) = (bh + ch^2 + dh^3) + (2ch + 3dh^2)u + (3dh)u^2 \\
\Delta_2(u) &= \Delta_1(u+h) - \Delta_1(u) = (2ch^2 + 6dh^3) + (6dh^2)u \\
\Delta_3(u) &= \Delta_2(u+h) - \Delta_2(u) = 6dh^3 \; .
\end{aligned}$$

Suppose that we know $\Delta_1(u)$, $\Delta_2(u)$ and $\Delta_3(u)$. Then these equations tell us that

$$\begin{aligned}
p(u+h) &= p(u) + \Delta_1(u) \\
\Delta_1(u+h) &= \Delta_1(u) + \Delta_2(u) \\
\Delta_2(u+h) &= \Delta_2(u) + \Delta_3(u) \\
\Delta_3(u+h) &= 6dh^3 \; .
\end{aligned}$$

Having obtained $p(u+h)$, $\Delta_1(u+h)$, $\Delta_2(u+h)$ and $\Delta_3(u+h)$, the same equations tell us how to compute

$p(u+2h)$, $\Delta_1(u+2h)$, $\Delta_2(u+2h)$ and $\Delta_3(u+2h)$, and so on since they are valid for any $u$, and in particular are valid for $u'=u+h$:

$$\begin{aligned} p(u'+h) &= p(u') + \Delta_1(u') \\ \Delta_1(u'+h) &= \Delta_1(u') + \Delta_2(u') \\ \Delta_2(u'+h) &= \Delta_2(u') + \Delta_3(u') \\ \Delta_3(u'+h) &= 6dh^3 \ . \end{aligned}$$

Thus we may use the following algorithm to compute the desired $N+1$ points on $p(u)$, beginning from $p(0)$, $\Delta_1(0)$, $\Delta_2(0)$ and $\Delta_3(0)$:

```
$u_0 = 0$
$y_0 = a$
$\Delta_1 = bh + ch^2 + dh^3$
$\Delta_2 = 2ch^2 + 6dh^3$
$\Delta_3 = 6dh^3$
for $i \leftarrow 1$ step 1 , . . . , $N$ do
    $u_i = u_{i-1} + h$
    $y_i = y_{i-1} + \Delta_1$
    $\Delta_1 = \Delta_1 + \Delta_2$
    $\Delta_2 = \Delta_2 + \Delta_3$
endfor
```

Aside from initialization, this method of computing the $y_i$ requires no multiplications and only $3N$ additions (plus $N$ additions to compute the $u_i$). This is a substantial improvement, especially when multiplications are expensive.

Of course, we are actually interested in parametric polynomials. Thus a 2D curve is represented by

$$\mathbf{Q}(u) = (\, X(u), Y(u) \,)$$

where $X(u)$ and $Y(u)$ are each cubic polynomials of the parameter $u$. Typically $u$ varies between 0 and some maximum value $u_{\max}$; for the time being we shall assume that $u_{\max} = 1$, as is the case for each segment of a uniform cubic B-spline curve. Evaluation of the curve is then performed by forward differencing the equations

$$\begin{aligned} X(u) &= a_x + b_x u + c_x u^2 + d_x u^3 \\ Y(u) &= a_y + b_y u + c_y u^2 + d_y u^3 \end{aligned}$$

simultaneously, using the same step size $h$:

```
$x_0 = a_x$
$\Delta x_1 = b_x h + c_x h^2 + d_x h^3$
$\Delta x_2 = 2c_x h^2 + 6d_x h^3$
$\Delta x_3 = 6d_x h^3$
$y_0 = a_y$
$\Delta y_1 = b_y h + c_y h^2 + d_y h^3$
$\Delta y_2 = 2c_y h^2 + 6d_y h^3$
$\Delta y_3 = 6d_y h^3$
for $i \leftarrow 1$ step $1 , \ldots , N$ do
    $x_i = x_{i-1} + \Delta x_1$
    $\Delta x_1 = \Delta x_1 + \Delta x_2$
    $\Delta x_2 = \Delta x_2 + \Delta x_3$
    $y_i = y_{i-1} + \Delta y_1$
    $\Delta y_1 = \Delta y_1 + \Delta y_2$
    $\Delta y_2 = \Delta y_2 + \Delta y_3$
endfor
```

A 3D cubic curve

$$\mathbf{Q}(u) = ( X(u), Y(u), Z(u) )$$

would be computed analogously.

We should note that forward differencing is not a universal panacea, owing to the cumulative error that arises from the finiteness of our arithmetic. This is particularly a problem on machines that lack floating point hardware, as is usually true of the microprocessors one finds integrated with displays. To see why this is so, suppose that each of $x_0$, $\Delta_1$, $\Delta_2$ and $\Delta_3$ is at most one unit in error. After $i$ iterations, $\Delta_3$ is still correct within one unit because it is a constant.

The possible error in $\Delta_2$, however, is larger. At each iteration of the loop its error may increase by one unit because we are adding $\Delta_2$ to it, so that at the end of $j$ iterations it may be in error by as much as $1+j$.

The maximum error in $\Delta_1$ after $k$ steps is its initial error, plus the sum of the amounts by which $\Delta_2$ may be in error at each step of the iteration. Hence the maximum possible error in $\Delta_1$ is given by

$$1 + \sum_{j=1}^{k}(1+j) = 1 + k + \frac{k(k+1)}{2} .$$

Finally, the maximum error in $x$ or $y$ after $N$ steps is its initial error plus the error that may have been contributed at each iteration by $\Delta_1$. This amounts to

$$1 + \sum_{k=1}^{N}\left(1 + k + \frac{k(k+1)}{2}\right) = \frac{11N + 6N^2 + N^3}{6} .$$

If we are working on a 512×512 raster display, the coefficients $a$, $b$, $c$ and $d$ require at least 9 bits of accuracy, and since they can differ in sign, will require more. Two to three additional bits of subpixel accuracy are desirable for antialiasing. If $N = 2^6$ then the total error might be as much as $2^{16}$, requiring that we maintain at least 16 bits of extra fractional precision to avoid an error or more than one pixel, for a total of about 28 bits. If $N = 2^8$ then the total error might be as much as $2^{22}$ and we cannot be sure of preserving the 12 or more bits of accuracy we desire, even on a 32 bit machine.

There are two obvious ways of dealing with this problem: one may use multiple precision arithmetic, or one may scale $\Delta_1$ and $\Delta_2$ so as to provide "guard" bits, and shift them away before adding them to $y_i$ and $\Delta_1$, respectively. In this way the error that is added to $y_i$ does not, practically speaking, grow as the computation proceeds.

Suppose that we maintain $\Delta_1$ and $\Delta_2$ scaled up by $N = 2^n = 1/h$ and shift them right by $n$ (written $\gg n$) before adding them to $y_i$ and $\Delta_1$, respectively, so that (roughly speaking) errors will be restricted to the bits that are discarded. Notice that there is no reason to scale up $\Delta_3$ since it is a constant.

```
$y_0 = a$
$\Delta_1 = b + c \gg n + d \gg 2n$
$\Delta_2 = 2c \gg n + 6d \gg 2n$
$\Delta_3 = 6d \gg 3n$
for $i \leftarrow 1$ step $1, \ldots, N$ do
    $y_i = y_{i-1} + \Delta_1 \gg n$
    $\Delta_1 = \Delta_1 + \Delta_2 \gg n$
    $\Delta_2 = \Delta_2 + \Delta_3$
endfor
```

Now the maximum possible error in $y_i$ is approximately $2^n$. If we assume that we will never want $n > 8$, and maintain $y_i$ with 10-12 fractional bits of precision our computation will be satisfactorily accurate using 32 bit integer arithmetic. If $n \leq 6$ then we can even squeeze the computation into 16 bits.

Of course, a machine with floating point hardware performs this scaling for us automatically, and if $N \leq 2^8$ we are unlikely to have problems with cubic polynomials. Nevertheless, it is apparent that cumulative error could become a problem, even on machines with floating point hardware, if we were to try differencing significantly higher degree polynomials.

### 14.5. Partial Derivatives and Normals

To perform solid area shading and hidden surface processing on raster devices, the simplest approach to take is to obtain a "wire frame" approximation to a spline surface. One way this may be done is by using the points $\mathbf{Q}(\bar{u}_i,\bar{v}_j)$ generated from the grid of values

$$\bar{u}_i = \bar{u}_0 + i\,\Delta\bar{u} \quad \text{and} \quad \bar{v}_i = \bar{v}_0 + i\,\Delta\bar{v} \ .$$

These are the positions in space at which the lines of constant parameter on the surface intersect, and these positions may be taken as the vertices of polygons used to approximate the surface. We have rendered most of the spline surfaces in our figures by this method.

Another way of obtaining a wire frame approximation is to use the vertices $\mathbf{W}_{i,j}$ of a suitably refined control graph as the polygonal mesh.

After a wire frame approximation has been obtained, by whatever means, standard polygonal techniques can be used to determine visibility, compute shading (if desired), and render the polygons. With respect to shading, however, a word of caution is in order. Since the surface $\mathbf{Q}(\bar{u})$ is not planar, the rectangles formed in the obvious way from the points $\mathbf{Q}(\bar{u}_i,\bar{v}_j)$ or $\mathbf{W}_{i,j}$ are not necessarily planar. It is sometimes advisable to render shaded, spline-generated polyhedral surfaces by dividing each rectangle into two triangles along one of the diagonals.

For smooth shading computations one may, of course, simply average the polygon normals for each of the polygons sharing a vertex. However, it is straightforward to compute the cross-product of the partial derivatives with respect to $u$ and $v$, so as to obtain an accurate normal vector at each polygon. The example of uniform cubic B-splines is instructive to bring out some of the computational issues in doing this.

From equation (31) we can see that to compute

$$\frac{\partial}{\partial v}\,\mathbf{Q}_{i,j}(u,v)$$

we simply evaluate a point on each of four uniform cubic B-spline curves to obtain what we called $\mathbf{W}_0$, $\mathbf{W}_1$, $\mathbf{W}_2$ and $\mathbf{W}_3$ in equation (31), and then we use these to scale $b_{-3}^{(1)}(v)$, $b_{-2}^{(1)}(v)$, $b_{-1}^{(1)}(v)$, and $b_{-0}^{(1)}(v)$. To compute

$$\frac{\partial}{\partial u}\,\mathbf{Q}_{i,j}(u,v)$$

we factor out $b_{-0}^{(1)}(u)$, $b_{-1}^{(1)}(u)$, $b_{-2}^{(1)}(u)$, and $b_{-3}^{(1)}(u)$ instead and proceed analogously.

It is worth pointing out that there are a variety of ways to give the user effective cues about the shape of a surface, and "realistic" shading is only one of them. Robin Forrest compares a number of others in [Forrest79].

### 14.6. Locality

The locality evidenced by uniform cubic B-splines has two advantages.

- It allows the designer of a complex curve to alter its shape in one region without affecting the shape of remote portions of the curve that have already been satisfactorily defined; the same is, of course, true of surfaces.
- Because only a part of the curve changes when a control vertex is moved, only a part of the curve must be recomputed. This facilitates real-time interaction.

Recall that the uniform cubic B-splines are translates of one another; that is, they are identically shaped.

If we choose to approximate each curve segment by $s$ consecutive chords whose end points are equally spaced $h = 1/s$ apart in $u$, then it is sufficient to compute values of the four basis segments at

$$u = 0, h, 2h \cdots, (s-1)h, sh = 1$$

and store them in arrays $b_{-3}[k]$, $b_{-2}[k]$, $b_{-1}[k]$ and $b_{-0}[k]$ before beginning to draw the curve and simply look up these values as we compute and draw each segment using

$$\mathbf{Q}_i(kh) = \sum_{r=-3}^{r=0} \mathbf{V}_{i+r} b_r[k] = \mathbf{V}_{i-3} b_{-3}[k] + \mathbf{V}_{i-2} b_{-2}[k] + \mathbf{V}_{i-1} b_{-1}[k] + \mathbf{V}_i b_{-0}[k] \ .$$

These precomputed values can also be used when we alter the position of some control vertex $\mathbf{V}_i$ and need to recompute the four segments $\mathbf{Q}_i(u)$, $\mathbf{Q}_{i+1}(u)$, $\mathbf{Q}_{i+2}(u)$ and $\mathbf{Q}_{i+3}(u)$. Since we usually need to recompute these four segments several times as we move $\mathbf{V}_i$, it is advantageous to add together the terms not involving $\mathbf{V}_i$ so that we need only perform a single multiplication and addition in order to obtain each new coordinate. For example, on the $i^{th}$ segment we would precompute

$$C[k] = \mathbf{V}_{i-3} b_{-3}[k] + \mathbf{V}_{i-2} b_{-2}[k] + \mathbf{V}_{i-1} b_{-1}[k]$$

as soon as the user had selected $\mathbf{V}_i$ for alteration, and then compute

$$\mathbf{Q}_i(kh) = C[k] + \mathbf{V}_i b_{-0}[k]$$

each time we wanted to redraw this segment.

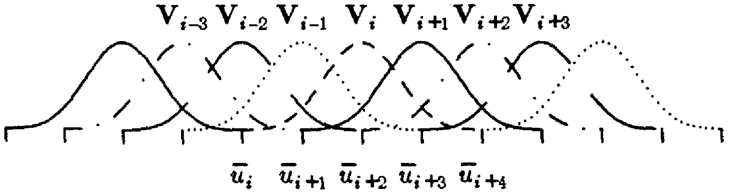

$\mathbf{V}_{i-3}$ $\mathbf{V}_{i-2}$ $\mathbf{V}_{i-1}$ $\mathbf{V}_i$ $\mathbf{V}_{i+1}$ $\mathbf{V}_{i+2}$ $\mathbf{V}_{i+3}$
$\bar{u}_i$ $\bar{u}_{i+1}$ $\bar{u}_{i+2}$ $\bar{u}_{i+3}$ $\bar{u}_{i+4}$

Figure 149. It is occasionally helpful to recall our indexing conventions. The $i^{th}$ interval is $[\bar{u}_i,\bar{u}_{i+1})$. It is determined by $\mathbf{V}_{i-3}$, $\mathbf{V}_{i-2}$, $\mathbf{V}_{i-1}$ and $\mathbf{V}_i$. On the other hand, $\mathbf{V}_i$ contributes to $\mathbf{Q}_i$, $\mathbf{Q}_{i+1}$, $\mathbf{Q}_{i+2}$ and $\mathbf{Q}_{i+3}$.

These observations extend to surfaces in the obvious way.

### 14.7. Scan-line Methods

One would prefer to work directly with the curved boundaries of these spline surfaces, rather than approximating them with straight line segments as one does by reducing them to polygons. There are two sources of complication in this approach, both resulting from the way in which scan-line algorithms are organized.

Firstly, because patch boundaries are parametric curves, altering $\bar{u}$ or $\bar{v}$ by a fixed amount does not everywhere result in movement of the same distance along the surface, or even in the same direction. Hence there is no simple incremental way to compute the intersection of an edge with scan-line $n$ from its intersection with scan-line $n-1$. Instead one has to use iterative numerical techniques. A typical approach is to solve for a zero of $Y(0,\bar{v})-n=0$, $Y(1,\bar{v})-n=0$, $Y(\bar{u},0)-n=0$, or $Y(\bar{u},1)-n=0$ (depending on which boundary is involved) by performing Newton iteration, using the known intersection of $Y(\bar{u})$ with scan-line $n-1$ as the initial guess. Given that we know the values $(\bar{u}_l,\bar{v}_l)$ and $(\bar{u}_r,\bar{v}_r)$ at which the left and right edges of a patch intersect the current scan-line, an analogous technique must be used to compute the values of $(\bar{u},\bar{v})$ corresponding to each interior pixel on that scan-line.

A more serious problem, arising because the patches are themselves curved, is that the apparent visible boundary of a patch on the display screen need not be an actual boundary of the patch. These are called ***silhouette edges***, and to perform the sort of scan conversion described above one must first identify the silhouette edges. A given boundary also need not be monotonic; it may intersect a scan-line more than once. Moreover, if we are generating a picture in top to bottom order, the highest point on a curved patch need not occur at a vertex, as it does for a polygon. Completely acceptable methods for dealing with these problems have yet to be developed. The state of the art is ably discussed in [Blinn80, Schweitzer82].

## 14.8. Ray-Tracing Cubic B-spline Surfaces

In this section we outline a method for intersecting rays with B-spline surfaces, based on the recurrence properties of B-splines [Cohen80] and on the fractal intersection algorithm of Kajiya [Kajiya83].

The material presented here is a brief account of that in [Sweeney84]. The method given there requires the aid of two preprocessing steps. The first step employs control-vertex refinement to produce local information about the surface suitable for use in starting Newton's iteration. The second step builds a tree of nested bounding boxes to be used for a version of hierarchical intersection testing that derives from Kajiya's work on ray-tracing fractals

### 14.8.1. Refinement Preprocessing

The first step in preprocessing a spline surface involves using the Oslo algorithm to replace the representation of the surface in terms of given control vertices by a refined representation in terms of more control vertices. The easiest version of this refinement process, which is the one presented in [Sweeney84], restricts the surfaces to those generated by uniform cubic B-splines and carries out Oslo refinement only by the introduction of $d$ equally-spaced knots within each interval of the legal parameter range. For example, Figures 144, 145, 146, and 147 give an instance of $d = 2$, and Figures 146 and 148 give an instance of $d = 4$. General B-splines could be used, of course, and more general refinements are possible. Moreover, the "subdivided refinements" mentioned in connection with figures 146, 147, and 148 appearing in section 13.4 could be employed. That is, an entire surface can be regarded as the union of two or more smaller surfaces, with separate control graphs derived from the single graph of the entire surface, and each of the smaller surfaces can be independently refined. Subdivided refinements provide the flexibility of dynamically adjusting a control graph to account for surface areas of local high or low variation.

The control vertices must be refined in advance of the ray tracing process sufficiently so that:

- the projection of each refined facet

$$\begin{matrix} \mathbf{W}_{r,s+1} & \mathbf{W}_{r+1,s+1} \\ \mathbf{W}_{r,s} & \mathbf{W}_{r+1,s} \end{matrix} \tag{97}$$

  covers no more than a few hundred pixels on the screen;

- the refined knots $\bar{w}_{r+2}$ (for parameter $\bar{u}$) and $\bar{t}_{s+2}$ (for parameter $\bar{v}$) associated with a control vertex $\mathbf{W}_{r,s}$ resulting from the refinement constitute acceptably good starting guesses for the Newton iteration, which is used to locate a ray's intersection with the spline surface.

### 14.8.2. Tree Construction

The refinement process described above constitutes a first step in the preprocessing of each spline surface. The second step in preprocessing involves building a tree of nested rectilinear bounding boxes containing the refined vertices. (Rectilinear bounding boxes; that is, boxes whose sides are aligned with the coordinate planes, are advocated in [Sweeney84] because intersections of rays with such boxes are easy to compute [Weghorst84]. contains further discussions on building nested structures of bounding volumes for the purposes of ray tracing.) The smallest bounding boxes in the tree, the leaf bounding boxes, must satisfy two containment requirements:

- each leaf of the tree should represent a small bounding box that is cented on one particular refined vertex or facet and is large enough to guarantee the inclusion of a piece of the underlying surface.
- the union of the leaf boxes should include the entire surface.

Each other bounding box in the tree; that is, any one associated with a non-leaf node, must provide nested containment of all boxes associated with its children nodes.

A rectilinear bounding box is defined by two points $(xmin, ymin, zmin)$ and $(xmax, ymax, zmax)$. An secure way of meeting the requirements just stated is to build the bounding box at each leaf around one of the facets (97) with $xmax$, $xmin$, $ymax$, $ymin$, $zmax$, and $zmin$ set just large enough to contain the convex hull of the 16 vertices associated with this facet:

$$\begin{array}{llll}
\mathbf{W}_{r-1,s+2} & \mathbf{W}_{r,s+2} & \mathbf{W}_{r+1,s+2} & \mathbf{W}_{r+2,s+2} \\[1em]
\mathbf{W}_{r-1,s+1} & \mathbf{W}_{r,s+1} & \mathbf{W}_{r+1,s+1} & \mathbf{W}_{r+2,s+1} \\[1em]
\mathbf{W}_{r-1,s} & \mathbf{W}_{r,s} & \mathbf{W}_{r+1,s} & \mathbf{W}_{r+2,s} \\[1em]
\mathbf{W}_{r-1,s-1} & \mathbf{W}_{r,s-1} & \mathbf{W}_{r+1,s-1} & \mathbf{W}_{r+2,s-1} \ .
\end{array}$$

An alternative to this is the construction advocated in [Sweeney84], where a box is built around a single vertex $\mathbf{W}_{r,s}$, and $xmax, \ldots, zmax$ are set large enough to include the 4 surrounding vertices

$$\begin{array}{lll}
 & \mathbf{W}_{r,s+1} & \\
\mathbf{W}_{r-1,s} & \mathbf{W}_{r,s} & \mathbf{W}_{r+1,s} \ . \\
 & \mathbf{W}_{r,s-1} &
\end{array} \tag{98}$$

A predefined overlap is included in the setting of $xmin, \ldots, zmax$ to include volume beyond that containing the vertices (98). For a sufficiently large overlap, the requirements concerning containment set out above will hold well enough to serve the purposes of ray tracing.

Finally, a pair of values for $\bar{u}, \bar{v}$ is associated with each leaf box. This pair of parameter values should give a point on the surface cented within the box. For example (recalling that the knots $\bar{w}_r$ are those which result in refining along the $\bar{u}$ parameter axis and the knots $\bar{t}_s$ are those which result in refining along the $\bar{v}$ axis), the pair of values

$$\begin{aligned}
\bar{u} &= \frac{1}{2}(\bar{w}_{r+2} + \bar{w}_{r+3}) \\
\bar{v} &= \frac{1}{2}(\bar{t}_{s+2} + \bar{t}_{s+3})
\end{aligned} \tag{99}$$

are reasonable ones to store for a bounding box defined on the 16 vertices around the facet (97), while the pair of values

$$\begin{aligned} \bar{u} &= \bar{w}_{r+2} \\ \bar{v} &= \bar{t}_{s+2} \end{aligned} \tag{100}$$

is reasonable for the scheme which uses 4 vertices about (98) plus overlap. Such a $(\bar{u},\bar{v})$ pair will serve as the starting values of a Newton process to be described below.

Finally, each internal node of the tree should represent a bounding box that is just large enough to contain the bounding boxes of its four children.

The leaves can be organized into a tree by a procedure that recursively subdivides the $\bar{u},\bar{v}$ parameter rectangle. At each level of recursion, the procedure allocates a node of the tree (the *current node*) and connects to it the four nodes to be allocated at the next level. The current node is associated with a rectangular section of the $\bar{u},\bar{v}$ range (in particular the root node is associated with the entire $\bar{u},\bar{v}$ rectangle), and the current node's rectangular section is quartered by halving its sides to produce the subrectangles given to the current node's children.

The recursion terminates when the current node's rectangular section of the $\bar{u},\bar{v}$ plane contains only the pair of values (99), in case the 16-vertex leaf box is used, or the pair of values (100) in the other case. The current node is tagged as a leaf node, and a leaf bounding box is calculated as described above.

As the procedure returns through the recursion, the parent nodes are tagged as internal nodes, and ever larger bounding boxes are calculated to contain the bounding boxes of the children.

The memory requirements for a spline surface are determined largely by the size of the tree of bounding boxes, and this is dictated, in turn, by the number of given control vertices and the level of refinement.

### 14.8.3. Intersection Processing

Kajiya [Kajiya83] has reported on a method for finding the intersection of rays with fractals and other surfaces he calls *height fields*. His algorithm has the property that it correctly handles surfaces that intersect rays at more than one spot. The algorithm is not limited to height fields but can be applied to any three dimensional surface.

Recall that the leaf nodes of the tree of bounding boxes contain starting values for a Newton iteration. Kajiya's algorithm, as applied to the subdivision trees described above, selects candidate leaf nodes for further processing by Newton's iteration, or else it rejects the ray as having no intersection with the surface if the ray fails to intersect any bounding box at some level of depth in the tree.

Briefly, for each ray one may maintain a linked list of active nodes. Attached to those nodes are various subtrees of the tree of bounding boxes described above. With each node is associated a distance from the ray origin to the closest intersection with the bounding box of the root of the attached subtree. One may maintain the list of active nodes sorted by increasing distance. The algorithm would proceed as follows:

- Choose the first (closest) node on the active node list, and remove it.
- If the root of the attached subtree is interior to the tree consider in turn each of its four children.
- If the ray hits the bounding box of a child, then attach the child to an active node, and sort the node into the the active node list.
- If the root of the attached subtree is a leaf, use the contained $(\bar{u},\bar{v})$ parameter values to initiate a Newton process.

This algorithm will terminate when the active node list is empty (failure), or the distance to the surface, as returned by the Newton iteration routine, is less than the distance to the first (closest) node on the active node list (success).

### 14.8.4. The Newton Iteration

The goal of an intersection computation is that of finding a pair of parameter values $\bar{u},\bar{v}$ such that a point $\mathbf{Q}(\bar{u},\bar{v})$ on the surface is also a point contained in a given ray. The two unknowns, $\bar{u}$ and $\bar{v}$, can be expressed as the roots of a pair of polynomial equations by the trick of formulating the desired intersections as the locus of all points on the surface that lie simultaneously in two planes containing the ray. This formulation was borrowed from [Kajiya82] although the rest of intersection process to be described is entirely different from the one he presented. We have

$$\text{Plane 1: } (A_1,B_1,C_1)\cdot(x,y,z) = D_1$$

$$\text{Plane 2: } (A_2,B_2,C_2)\cdot(x,y,z) = D_2 \ ,$$

where

$$(x,y,z) = (X(\bar{u},\bar{v}),Y(\bar{u},\bar{v}),Z(\bar{u},\bar{v})) = \mathbf{Q}(\bar{u},\bar{v}) \ .$$

In particular, for a ray given parametrically as

$$(x_a,y_a,z_a) + t\ (x_b,y_b,z_b) \ ,$$

we have

$$\begin{aligned}(A_1,B_1,C_1) &= (x_a,y_a,z_a)\times(x_b,y_b,z_b)\\ (A_2,B_2,C_2) &= (A_1,B_1,C_1)\times(x_b,y_b,z_b)\\ D_1 &= (A_1,B_1,C_1)\cdot(x_a,y_a,z_a)\\ D_2 &= (A_2,B_2,C_2)\cdot(x_a,y_a,z_a) \ .\end{aligned}$$

Using

$$\mathbf{Q}(\bar{u},\bar{v}) = \sum_{i=0}^{m}\sum_{j=0}^{n}\mathbf{V}_{i,j}B_{i,4}(\bar{u})B_{j,4}(\bar{v}) \ ,$$

this gives two equations in two unknowns to be solved:

$$E_k(\bar{u},\bar{v}) \equiv \sum_{i=0}^{m}\sum_{j=0}^{n}\Big[(A_k,B_k,C_k)\cdot\mathbf{V}_{i,j}\Big]B_{i,4}(\bar{u})B_{j,4}(\bar{v}) - D_k = 0 \qquad (101)$$

for $k = 1,2$.

Let $\bar{u}^{[0]},\bar{v}^{[0]}$ stand for the values stored in a leaf node. Newton's method starts with these values as an approximation to the solution of (101) and refines them

$$\bar{u}^{[0]} \rightarrow \cdots \rightarrow \bar{u}^{[l]} \rightarrow \bar{u}^{[l+1]} \rightarrow \cdots$$

$$\bar{v}^{[0]} \rightarrow \cdots \rightarrow \bar{v}^{[l]} \rightarrow \bar{v}^{[l+1]} \rightarrow \cdots$$

by taking each $\bar{u}^{[l]},\bar{v}^{[l]}$ and solving the $2\times2$ system

$$\begin{bmatrix} \dfrac{\partial E_1}{\partial \bar{u}} & \dfrac{\partial E_1}{\partial \bar{v}} \\ \dfrac{\partial E_2}{\partial \bar{u}} & \dfrac{\partial E_2}{\partial \bar{v}} \end{bmatrix} \begin{bmatrix} \bar{u}^{[l+1]} \\ \bar{v}^{[l+1]} \end{bmatrix} = \begin{bmatrix} E_1(\bar{u}^{[l]},\bar{v}^{[l]}) \\ E_2(\bar{u}^{[l]},\bar{v}^{[l]}) \end{bmatrix}$$

to produce the (usually) more accurate solution of (101) given by $\bar{u}^{[l+1]},\bar{v}^{[l+1]}$. The partial derivatives $\dfrac{\partial E_k}{\partial \bar{u}}$ for $k = 1,2$ are given by

$$\frac{\partial E_k}{\partial \bar{u}} = \sum_{i=0}^{m}\sum_{j=0}^{n}\Big[(A_k,B_k,C_k)\cdot\mathbf{V}_{i,j}\Big]B^{(1)_{i,4}}(\bar{u})B_{j,4}(\bar{v})$$

and similarly for $\dfrac{\partial E_k}{\partial \bar{v}}$. Note that control vertices used in the iteration should be the original, unrefined set $\mathbf{V}_{i,j}$ in order to reduce computation.

The Newton iteration can be terminated, and $\bar{u}^{[l+1]},\bar{v}^{[l+1]}$ can be taken as defining an intersection, if

$$|E_1(\bar{u}^{[l+1]},\bar{v}^{[l+1]})| + |E_2(\bar{u}^{[l+1]},\bar{v}^{[l+1]})| < \textit{tolerance} \ .$$

Failures should be registered (that is, a ray strike should be regarded as not occurring) if the Newton iterates $\bar{u}^{[l+1]},\bar{v}^{[l+1]}$ wander outside the bounds of the parametric intervals; *i.e.*

$$\bar{u}^{[l+1]} < \bar{u}_{k-1} \quad \text{or} \quad \bar{u}^{[l+1]} > \bar{u}_{m+1} \quad \text{or} \quad \bar{v}^{[l+1]} < \bar{v}_{k-1} \quad \text{or} \quad \bar{v}^{[l+1]} > \bar{v}_{p+1}$$

or if

$$l > \textit{allowance}$$

and the value of

$$|E_1(\bar{u}^{[l+1]},\bar{v}^{[l+1]})| + |E_2(\bar{u}^{(l+1)},\bar{v}^{(l+1)})|$$

has increased over that of the preceding iteration step.

# 15. A Retrospective and Selected Applications

Before proceeding to the discussion of Beta-splines, we will pause to recall and emphasize a few results concerning the representation of splines. Our theme will be "alternative representations;" we will present several somewhat peripheral topics, some with practical application and some of purely intellectual interest.

Splines are assembled from polynomial pieces, all of a common order and joined with a degree of continuity specified at each joint individually. That continuity has been specified implicitly, by the knot multiplicity underlying each joint.

One obvious way of specifying a spline, in fact the first method we discussed, is by listing the individual coefficients of individual segment polynomials on the individual breakpoint intervals. The price to be paid for this method of representation is vigilance. Only some lists of coefficients are acceptable, for not all lists define segment polynomials that meet with the requisite continuity. Indeed, continuity is enforced by the linear equations that we have often imposed upon the coefficients of adjacent segment polynomials.

We have given some evidence that segment-polynomial representation may be useful, despite its awkwardness. We have used it to construct splines that interpolate control vertices, for example, and we have suggested that splines can be evaluated efficiently using segment-polynomial representation.

We have not made much comment about the use of different basis polynomials for representing the segment polynomials, but it is worth mentioning here. We have generally represented segment polynomials as a sum of coefficients times powers of $u$, which means that we have implicitly used a different powers basis

$$u^j \equiv (\overline{u}-\overline{u}_i)^j$$

in each breakpoint interval $[\overline{u}_i,\overline{u}_{i+1})$. Other choices exist. In section 15.1, for example, we will introduce a basis for describing cubic segment polynomials that automatically enforces $C^1$ continuity at the joints and facilitates the interpolation of control vertices. This basis also makes it trivial to impose any desired tangent vector upon a parametric spline curve at each joint, and the application chosen to illustrate this will be the use of cubic $C^1$ splines as a method of key-frame inbetweening for computer animation.

Another method of representing splines was given by the one-sided power functions. Since these functions were themselves splines, and since they formed a basis for the spline space we were considering, it sufficed to simply supply a list of coefficients for the one-sided power functions. The continuity at joints was automatically enforced by the continuity inherent in the one-sided power functions themselves. However, this does not produce something for nothing. The one-sided power functions were, in effect, a mathematical mechanism for "hiding" the segments from view, but the inherent piecewise nature of a spline appears when values of a spline are needed. The evaluation of a spline at any fixed $\bar{u}$ requires an "if-then-else" test on each power to determine whether it is or is not zero at that $\bar{u}$-value, and the spline's value is given by a summation of the remaining powers times their respective coefficients.

The most computationally useful collection of splines that we presented, which like the one-sided power functions constitutes a basis for the splines of a given space, was the B-splines. In terms of these special splines, any spline had a representation involving a single list of coefficients, and the satisfaction of the continuity conditions on general splines in the space was a consequence of the continuity properties of the individual B-splines. The segmentwise nature of a spline, hidden by the notation of B-spline representation, arises here as well when one attempts the evaluation of a given spline, for one must ultimately decide (e.g. at the bottom level of a recurrence) in which interval the point of evaluation lies.

This chapter's guiding theme — alternative representations — prompts us to recall that there are many different ways of representing any particular spline. A spline may be given, for example, by its segment-polynomial representation, or in terms of one-sided power functions, or in terms of the B-splines. There are algorithms to transform a spline from one representation to another. We have not discussed transformations between segment polynomials and one-sided power functions; they have little interest for us either practically or pedagogically. But the divided-difference material covered in chapters 8 and 9 provides the description of a basis-change process that transforms a one-sided representation into a B-spline representation. Material in chapter 14 went in another direction, establishing an algorithm to transform the B-spline representation of a spline into its segment-polynomial representation. Because it is an important topic that appears in most mathematical books on splines, we develop "Marsden's Lemma" in section 15.5. This result gives, in somewhat disguised form, a transformation that reverses the divided-difference process and transforms B-spline representations of splines into one-sided power representations.

The bases given above are not the only possible ones to use in representing splines. A special class of splines that make the solution of interpolating problems trivial to solve, borrowing the tricks embodied in the basis to be discussed in section 15.1, is the class of the "cardinal-spline" bases. An example of these will be presented briefly in section 15.2.

Extensions of the control-vertex construction process that we have been using will also be explored: if the control vertices are replaced by certain vector functions, the result is a constructive process whose products are called "Catmull-Rom" splines. These will be presented in section 15.4.

Section 15.2 will have the secondary purpose of illustrating how problems solved using one representation of a spline can be solved in a different way using another representation. The cardinal-spline representation provides a solution to the interpolation problem, which was solved in the initial chapters using a segment-polynomial representation. Section 15.3 will solve the interpolation problem again, this time using B-splines. A sample application in picture processing will be given.

Finally, section 15.6 will illustrate the use of B-splines in another form of control-vertex approximation: least-squares fitting. We will present an application of this form of approximation to the job of capturing gestures made with a tablet or mouse, for use in graphical free-hand-drawing software.

We close this introduction with a reminder that the Oslo algorithm presents yet another example of alternative representations. If $\mathbf{S}(\mathbf{P}^k, \{\bar{u}_i\}_0^{m+k})$ is one spline space and $\mathbf{S}(\mathbf{P}^k, \{\bar{w}_j\}_0^{m+n+k})$ is another, having polynomial segments of the same order and for the same legal parameter range, and if the knot sequence $\{\bar{w}_j\}_0^{m+n+k}$ contains the knot sequence $\{\bar{u}_i\}_0^{m+k}$, then $\mathbf{S}(\mathbf{P}^k, \{\bar{u}_i\}_0^{m+k})$ is a subspace of $\mathbf{S}(\mathbf{P}^k, \{\bar{w}_j\}_0^{m+n+k})$. This means that any spline in $\mathbf{S}(\mathbf{P}^k, \{\bar{u}_i\}_0^{m+k})$ can be represented by any of the descriptive mechanisms for $\mathbf{S}(\mathbf{P}^k, \{\bar{w}_j\}_0^{m+n+k})$. The Oslo algorithm rested, of course, on the idea of representing the splines of $\mathbf{S}(\mathbf{P}^k,\{\bar{u}_i\}_0^{m+k})$ (more specifically, of representing the B-splines of $\mathbf{S}(\mathbf{P}^k,\{\bar{u}_i\}_0^{m+k})$) in terms of the B-splines of $\mathbf{S}(\mathbf{P}^k,\{\bar{w}_j\}_0^{m+n+k})$.

Another observation of this sort was made in chapter 11 (on Be'zier techniques). For the uniform-knot case we showed that any B-spline curve could be represented as a Be'zier curve, and alluded to the fact that this was true in general. The argument is something like the following: Suppose that $\mathbf{S}(\mathbf{P}^k,\{\bar{u}_i\}_0^{m+k})$ is an arbitrary spline space. Then each knot in $\{\bar{u}_i\}_0^{m+k}$ may have multiplicity 1 up to $k-1$. Suppose that $\mathbf{S}(\mathbf{P}^k,\{\bar{w}_j\}_0^{m+n+k})$ is formed by adding knots to $\{\bar{u}_i\}_0^{m+k}$ to bring each of the knots $\bar{u}_{k-1}, \ldots, \bar{u}_{m+1}$ up to multiplicity $k-1$. Then the B-spline basis for $\mathbf{S}(\mathbf{P}^k,\{\bar{w}_j\}_0^{m+n+k})$ will constitute the $k^{\text{th}}$-order Bernstein polynomials; i.e. the basis polynomials for $k^{\text{th}}$-order Be'zier curves. From this we may conclude that any spline curve has a Be'zier representation. Because Be'zier subdivision is efficient and simple to implement, Barsky has used this observation (jointly with DeRose and Dippe') to construct a subdivision-based rendering algorithm [Barsky85], and (jointly with Thomas) in designing an interactive design system [Barsky81].

Since there are a number of options for the representation of splines, questions of choice have to be resolved by convenience, efficiency, numerical accuracy, or compactness of storage. Sometimes there is no single solution. In large computer-aided design systems, for example, rendering might best be done using segment polynomials or using Be'zier-based subdivision. Shaping interactions might best be done using the Beta-spline techniques, to be discussed in the final chapters, backed up by Oslo-based refinement to make adjustments to the locality over which the shaping takes effect. And the description of a resulting surface might best be recorded — for the subsequent process of engineering analysis or fabrication — using B-splines. Only the specifics of an application, and a software and hardware environment, will tell.

For a graphical illustration of the points we have been making above, the reader is referred to the cover of the January 1985 issue of *Computer Graphics*, where the outline of a flamingo is defined in three equivalent ways [Stone85]: one (designated "interpolating") is essentially an illustration of the segment-polynomial representation; one (labeled "B-spline") shows the B-spline control polygon defining the same piecewise curve; and the third (labeled "Be'zier") shows the equivalent Be'zier polygon for the curve.

## 15.1. The Hermite Basis and $C^1$ Key-Frame Inbetweening

Material in this section derives from [Kochanek84].

One of the oldest techniques used in computer animation is the automatic generation of *inbetweens* (intermediate frames) based on a set of *key frames* supplied by the animator. This same method is frequently used in computer-assisted special effects where camera and positions of objects are defined only at key points in the action, leaving the calculation of intermediate positions to the computer. Linear interpolation has been used in many such systems, but it produces undesirable side effects that give the animation a mechanical look, often referred to as the "computer signature." The most objectionable characteristic of this type of animation is a lack of smoothness in the motion. The key frames may be clearly visible in the animation because of sudden changes in the direction of motion (Figure 150).

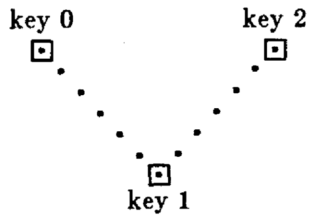


Figure 150. Discontinuity in direction with linear interpolation.

Discontinuities in the speed of motion may also be visible with linear interpolation, for example when the animator requests a different number of frames between successive keys (Figure 151).

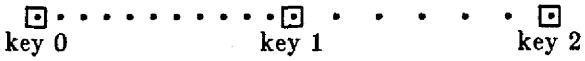


Figure 151. Discontinuity in speed with linear interpolation.

A third common problem is distortion, which may occur whenever the movement has a rotational component (Figure 152).

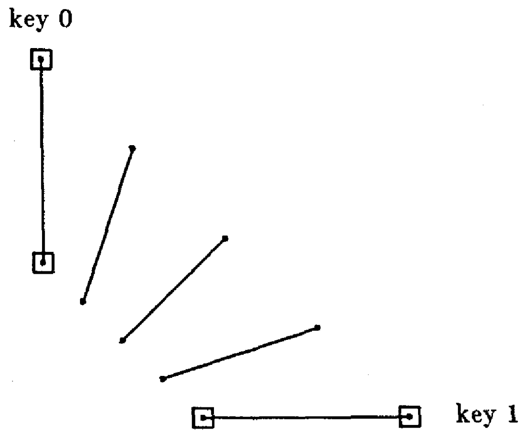


Figure 152. Distortion in length when rotation is simulated linearly.

Inbetweening systems usually begin with the assumption that each of the objects in the $i^{\text{th}}$ key frame in a sequence can be described by a collection of *key points* (e.g., the two designated endpoints of the line segment in key 0 shown in Figure 152 completely define the segment) and that to each point in one key frame there will be a corresponding point in all the other key frames of a motion sequence (e.g., the same two endpoints reappear in key 1 of Figure 152 to specify a later position of the line segment.) If we choose one such point at the $i^{\text{th}}$ key frame,

$$\mathbf{P}_i = (x_i, y_i, z_i) \; ,$$

called the $i^{\text{th}}$ *key position*, then the corresponding points in all of the key frames constitute a sequence of key positions

$$\cdots, \mathbf{P}_{i-1}, \mathbf{P}_i, \mathbf{P}_{i+1}, \cdots,$$

that we want to interpolate using a simple smooth curve to ameliorate the above problems. In Figure 152, for example, the top endpoint of the line segment in key 0 could be the chosen key position, $\mathbf{P}_0$; hence, the key position, $\mathbf{P}_1$, would be the right-hand endpoint of the line segment in key 1, and the sequence we must interpolate is merely

$$\mathbf{P}_0, \mathbf{P}_1 .$$

The result of the interpolation should be a parametric curve $\mathbf{Q}(\bar{u})$ with segments

$$\mathbf{Q}_i(u) = (X_i(u), Y_i(u), Z_i(u)) \text{ for } 0 \le u \le 1 ,$$

where

$$\mathbf{Q}_i(0) = (X_i(0), Y_i(0), Z_i(0)) = (x_i, y_i, z_i) = \mathbf{P}_i$$

and

$$\mathbf{Q}_i(1) = (X_i(1), Y_i(1), Z_i(1)) = (x_{i+1}, y_{i+1}, z_{i+1}) = \mathbf{P}_{i+1} .$$

The positions in the inbetween frames on each such segment will be given by $\mathbf{Q}_i(u)$ for some sequence of $u$-values between 0 and 1. The entire curve $\mathbf{Q}(\bar{u})$ defines the ***trajectory*** followed by the point whose key positions we have interpolated. The *motion dynamics* of the the point's transit over this trajectory will be determined by the sequence of $\bar{u}$-values chosen. The discussion to be given here covers only the trajectory aspects of inbetweening; there is much work left to be done on the the aspects of motion dynamics.

If we decide to use cubic splines with $C^1$ continuity at each joint, and we consider each of the segments in turn, we recall from chapter 3 that each of the component polynomials $X_i(u)$, $Y_i(u)$, and $Z_i(u)$ can be defined uniquely by Hermite interpolation. Two constraints are given directly by the interpolation conditions and the other two constraints are given by specifying derivatives at $u = 0$ and $u = 1$. Thus, $\mathbf{Q}_i(u)$ is completely determined by

$$\mathbf{P}_i \quad \text{and} \quad \mathbf{D}_i = \left( \frac{dX_i(0)}{du}, \frac{dY_i(0)}{du}, \frac{dZ_i(0)}{du} \right)$$

and

$$\mathbf{P}_{i+1} \quad \text{and} \quad \mathbf{D}_{i+1} = \left( \frac{dX_{i+1}(1)}{du}, \frac{dY_{i+1}(1)}{du}, \frac{dZ_{i+1}(1)}{du} \right) .$$

This is shown in Figure 153.

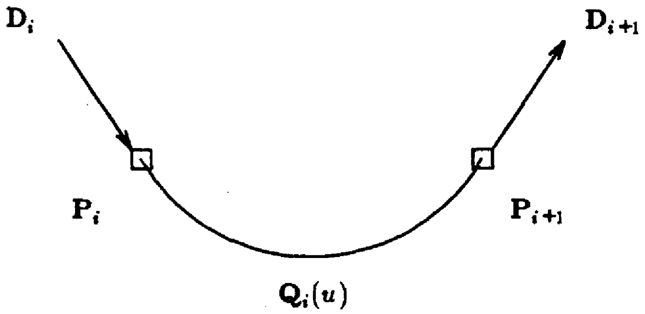
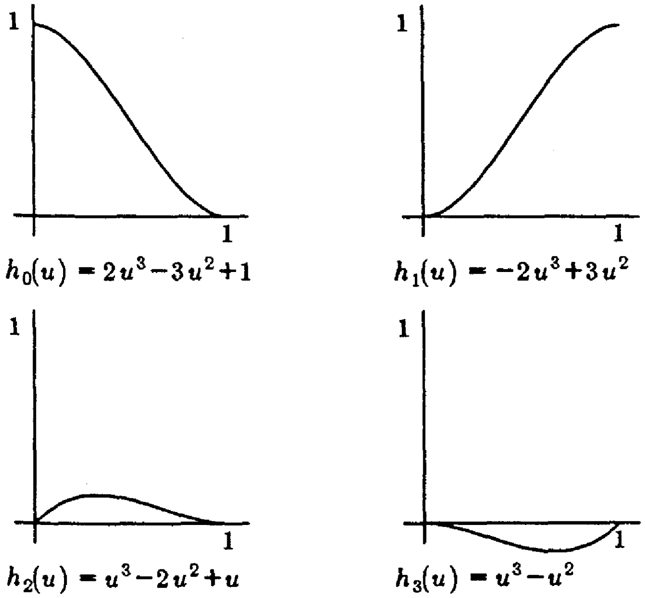


Figure 153. The trajectory between two key positions.

An automatic inbetweening system may choose $\mathbf{D}_i$ and $\mathbf{D}_{i+1}$ by some geometric information derived from the surrounding keys, or by human input, or by some combination of both.

The specification of $X_i(u)$, $Y_i(u)$, and $Z_i(u)$ is most conveniently made in terms of the *Hermite interpolation basis functions*, which are shown in Figure 154:



$h_0(u) = 2u^3-3u^2+1$

$h_1(u) = -2u^3+3u^2$

$h_2(u) = u^3-2u^2+u$

$h_3(u) = u^3-u^2$

Figure 154. Basis polynomials for Hermite interpolation.

These functions have the following properties.

| | $h_0$ | $h_1$ | $h_2$ | $h_3$ |
|---|---|---|---|---|
| function value at $u = 0$ | 1 | 0 | 0 | 0 |
| function value at $u = 1$ | 0 | 1 | 0 | 0 |
| derivative at $u = 0$ | 0 | 0 | 1 | 0 |
| derivative at $u = 1$ | 0 | 0 | 0 | 1 |

Consider any expression of the form

$$p(u) = a\,h_0(u) + b\,h_1(u) + c\,h_2(u) + d\,h_3(u) \;,$$

where $a$, $b$, $c$, and $d$ are arbitrary coefficients. Note that $h_0(u)$ alone determines the function value of $p(u)$ at the start of the interval; that is, $p(0) = a$. Similarly, $h_1(u)$ determines $p(u)$ at the end of the interval; that is, $p(1) = b$. The derivatives of $p(u)$ at the beginning and end of the interval are determined by $h_2(u)$ and $h_3(u)$, respectively; that is, $p'(0) = c$ and $p'(1) = d$. These observations lead to the representation

$$\begin{aligned}\mathbf{Q}_i(u) &= (X_i(u), Y_i(u), Z_i(u)) \\ &= h_0(u)\mathbf{P}_i + h_1(u)\mathbf{P}_{i+1} + h_2(u)\mathbf{D}_i + h_3(u)\mathbf{D}_{i+1} \;.\end{aligned}$$

In matrix form this expression reduces to

$$\begin{aligned}\mathbf{Q}_i(u) &= \mathbf{u}\cdot\mathbf{h}\cdot\mathbf{C} \qquad (102)\\ &= \begin{bmatrix} u^3 & u^2 & u & 1\end{bmatrix} \cdot \begin{bmatrix} 2 & -2 & 1 & 1 \\ -3 & 3 & -2 & -1 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix} \cdot \begin{bmatrix} \mathbf{P}_i \\ \mathbf{P}_{i+1} \\ \mathbf{D}_i \\ \mathbf{D}_{i+1} \end{bmatrix} \;.\end{aligned}$$

Note that the vector $\mathbf{u}$ changes only from one frame in the animation to the next. Within a given frame it applies to the $x$, $y$, and $z$ components of all key positions which are being interpolated. The matrix $\mathbf{h}$ contains the coefficients of the Hermite interpolation basis functions and is therefore constant for all frames and all key positions. In practice, $\mathbf{u}\cdot\mathbf{h}$ need only be calculated once per frame for each collection of key positions that are moving as a unit. By contrast, each $\mathbf{C}$, which is a 4×3 matrix, corresponds to a single key position and is independent of the $\mathbf{C}$ associated with any of the other key positions being interpolated. It does not change from one frame to another (except at a key frame), and the independence implies that all key positions can be interpolated "in parallel."

To give an example, the product $\mathbf{u}\cdot\mathbf{h}$ would be the same for both the top (key 0), respectively right (key 1), and the bottom (key 0), respectively left (key 1), endpoint of the line segment in Figure 152. However, there would be one version of $\mathbf{C}$ for the top/right endpoint and another for the bottom/left endpoint. If the animation sequence included a circle whose motion differed from that of the line segment, $\mathbf{h}$ would remain the same, but a different version of $\mathbf{u}$ and $\mathbf{C}$ would be needed.

Using this formulation as a framework, we offer some suggestions about finding values for the components of $\mathbf{D}_i$ and $\mathbf{D}_{i+1}$ (the tangent vectors at the key positions) purely from local geometric information. The tangent vector at $\mathbf{P}_i$ may be calculated as

$$\mathbf{D}_i = \frac{1}{2}(\mathbf{P}_{i+1} - \mathbf{P}_{i-1}) = \frac{1}{2}\Big((\mathbf{P}_{i+1} - \mathbf{P}_i) + (\mathbf{P}_i - \mathbf{P}_{i-1})\Big) \;, \qquad (103)$$

which is simply the average of the *source chord* $\mathbf{P}_i - \mathbf{P}_{i-1}$ and the destination chord $\mathbf{P}_{i+1} - \mathbf{P}_i$. We will refer to this average as the *default* (Figure 155). (This method of derivative generates what are commonly called the "Catmull-Rom splines," although we shall see in section 15.4 that this method is actually just a particular instance of the family of splines defined by Catmull and Rom.)

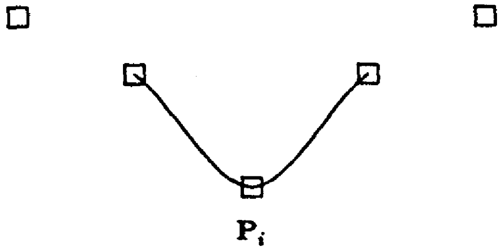


Figure 155. An example of the default interpolation.

At the beginning of a motion sequence, i.e. at $\mathbf{P}_0$, some arbitrary choice for the source chord must be made. Similarly, the destination chord must be specified arbitrarily at the end of the sequence. Alternatively a specification of the beginning and ending tangent vectors can be made without regard to any chords.

A standard smooth motion through a given set of keys does not always produce the effect desired by the animator. In certain cases a wider, more exaggerated curve may be desired, while in other cases the desired path may be much tighter. This suggests that some sort of "tension" in the trajectory as it passes through a key position, such as that shown in Figure 156, would be desirable.

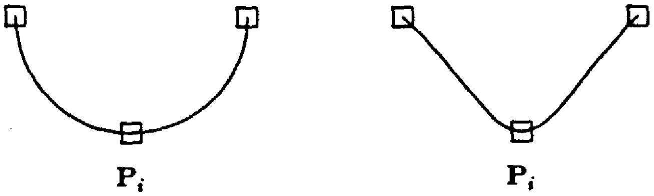


Figure 156. Two interpolations; the one on the right being more tense at $\mathbf{P}_i$ than the one at the left.

The animator may wish to have a trajectory anticipate or overshoot a key position by a certain amount. This suggests that the sort of "bias" illustrated in Figure 157 would be useful.

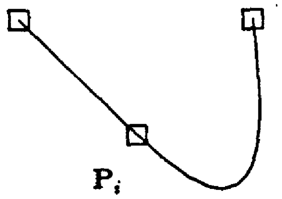


Figure 157. A biased interpolation at $\mathbf{P}_i$.

Even continuity in the direction and speed of motion is not always desirable. Animating a bouncing ball, for example, actually requires the introduction of a discontinuity in the motion at the point of impact. Variation of "continuity" is illustrated in Figure 158.

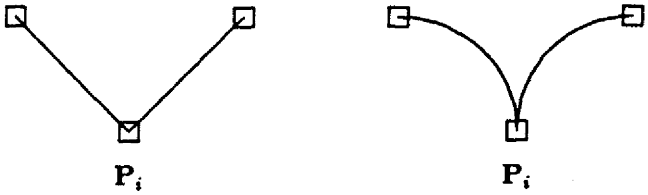


Figure 158. Two interpolations; the one on the right being more discontinuous than the one on the left at $\mathbf{P}_i$.

We can introduce *tension*, *continuity*, and *bias* parameters by separating each tangent at the $i^{\text{th}}$ key position into an *incoming* and an *outgoing* part, respectively the *source derivative* $\mathbf{DS}_i$ and the *destination derivative* $\mathbf{DD}_i$ as indicated in Figure 159.

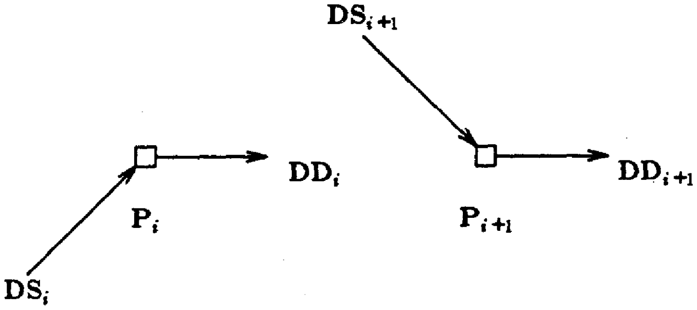


Figure 159. Incoming and outgoing tangents of two key positions.

These replace the single tangent vector in the default spline at each $\mathbf{P}$. Furthermore, the default average (103) is relaxed in favour of a more selective average of the source and destination chord.

A tension parameter $t_i$ to control how sharply the curve bends may be implemented as a scale factor which changes the length of both the incoming and outgoing parts of the tangent vector equally at $\mathbf{P}_i$:

$$\mathbf{DS}_i = \mathbf{DD}_i = (1-t_i)\frac{1}{2}\Big((\mathbf{P}_{i+1}-\mathbf{P}_i)+(\mathbf{P}_i-\mathbf{P}_{i-1})\Big) \ . \tag{104}$$

Setting $t_i = 0$ produces the default; the tangent vector is the average of the two adjacent chords. Increasing the tension to $t_i = 1$ (Figure 160) reduces the length of the tangent vector to zero and tightens the curve to a corner.

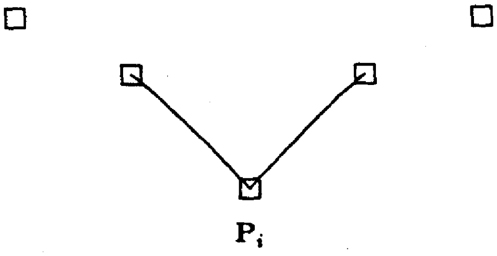


Figure 160. An interpolation at $\mathbf{P}_i$ with $t_i = 1$. The value of $t$ is zero at all other points $\mathbf{P}$.

Reducing the tension to $t_i = -1$ increases the tangent vector to twice its default length and produces more slack in the curve (Figure 161).

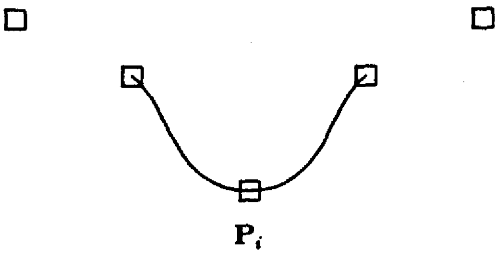


Figure 161. An interpolation at $\mathbf{P}_i$ with $t_i = -1$. The value of $t$ is zero at all other points $\mathbf{P}$.

The value of $t_i$ can be set at other values for more pronounced effects. For example, $t_i > 1$ will produce loops (Figure 162):

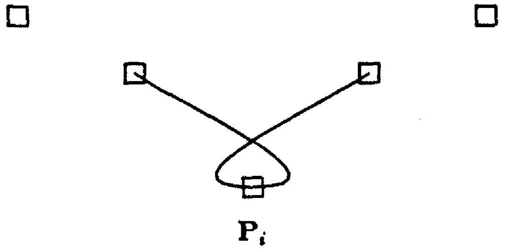


Figure 162. An interpolation at $\mathbf{P}_i$ with $t_i = 4$, which results in a loop. The value of $t$ is zero at all other points $\mathbf{P}$.

The principal reason for using splines in key frame animation is to avoid discontinuities in the direction and speed of motion which are produced by linear interpolation. However, in animation discontinuities are sometimes necessary to create realistic effects such as punching, bouncing, etc. A common technique to introduce such a discontinuity into an otherwise continuous spline is to repeat a key position or to simply terminate the spline at a key and start an entirely independent spline to interpolate the next sequence of key frames.

Neither of these approaches is very satisfactory because the discontinuity cannot be controlled. While it is true that, mathematically speaking, a spline's derivative is either continuous or discontinuous, the artist's view is quite different. He or she would like to have more control over continuity than a simple on/off switch. In fact, from the animator's point of view two curve segments which have very different tangent vectors at their joint appear "more discontinuous" than two curve segments which have fairly similar tangent vectors.

Using $c_i$ to denote a continuity parameter, we may allow the source and destination components of the tangent vector to differ from each other according to:

$$\mathbf{DS}_i = \left(\frac{1-c_i}{2}(\mathbf{P}_i-\mathbf{P}_{i-1}) + \frac{1+c_i}{2}(\mathbf{P}_{i+1}-\mathbf{P}_i)\right) \tag{105}$$

$$\mathbf{DD}_i = \left(\frac{1+c_i}{2}(\mathbf{P}_i-\mathbf{P}_{i-1}) + \frac{1-c_i}{2}(\mathbf{P}_{i+1}-\mathbf{P}_i)\right) . \tag{106}$$

Note that with $c_i = 0$ we obtain $\mathbf{DS}_i = \mathbf{DD}_i$, which produces a spline with tangent vector continuity at the keys. (In fact, this choice reproduces the default interpolation.) As the magnitude $|c_i|$ of $c_i$

increases, the two tangent vectors become increasingly distinct. When $c_i = -1$ (Figure 163), the source tangent vector $\mathbf{DS}_i$ reduces to the source chord, and the destination tangent vector $\mathbf{DD}_i$ reduces to the destination chord, producing a pronounced corner in the curve, if the two chords are not colinear and of equal length. In fact, Figure 160 is exactly reproduced with this setting. As $c_i$ is made more negative, the corner becomes more acute, and the curve buckles inward (Figure 163).

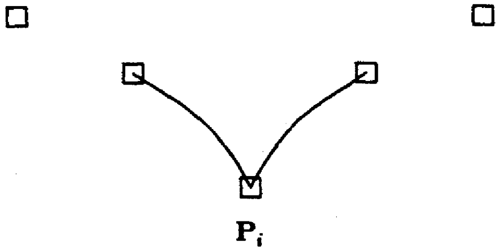


Figure 163. An interpolation at $\mathbf{P}_i$ with $c_i = -2$. The value of $c$ is zero at all other points $\mathbf{P}$.

For positive values of $c_i$ corners pointing in the opposite direction are produced (Figure 164).

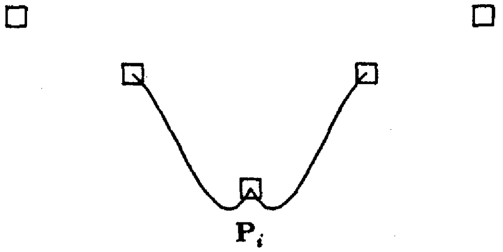


Figure 164. An interpolation at $\mathbf{P}_i$ with $c_i = 2$. The value of $c$ is zero at all other points $\mathbf{P}$.

Finally, we can introduce a bias parameter $b_i$ to control the direction of the path as it passes through $\mathbf{P}_i$. Both incoming and outgoing parts of the tangent are formed as an average of the incoming and outgoing chords, but the bias assigns different weights to the two chords when forming the average.

$$\mathbf{DS}_i = \mathbf{DD}_i = \frac{1+b_i}{2}(\mathbf{P}_i - \mathbf{P}_{i-1}) + \frac{1-b_i}{2}(\mathbf{P}_{i+1} - \mathbf{P}_i) \ . \tag{107}$$

Note that with $b_i = 0$ the two chords are weighted equally, and the default interpolation is produced. When $b_i = -1$, the tangent vector is completely determined by the destination chord, and when $b_i = 1$, the tangent vector is completely determined by the source chord. The more negative $b_i$ is made, the more the trajectory "bends" to one side of $\mathbf{P}_i$ (Figure 165).

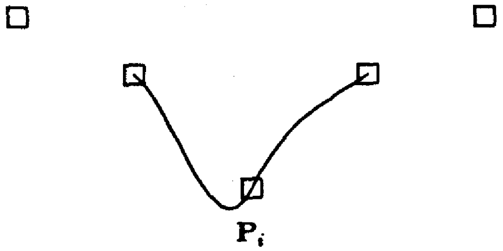


Figure 165. An interpolation at $\mathbf{P}_i$ with $b_i = -2$. The value of $b$ is zero at all other points $\mathbf{P}$.

The more positive $b_i$ is made, the more the trajectory bends to the other side of $\mathbf{P}_i$ (Figure 166):

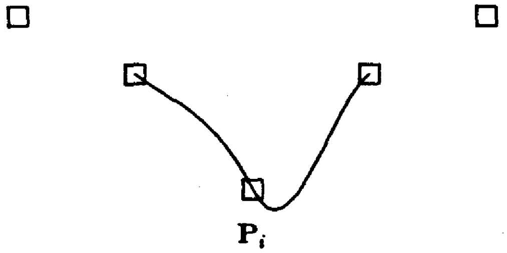


Figure 166. An interpolation at $\mathbf{P}_i$ with $b_i = 2$. The value of $b$ is zero at all other points $\mathbf{P}$.

The bias parameter easily simulates the traditional animation effect of following through after an action by "overshooting" the key position or exaggerating a movement by "undershooting" a key position.

Combining the tension, continuity, and bias control parameters we obtain the following general equations for the source and destination tangent vectors at the key position $\mathbf{P}_i$.

$$\mathbf{DS}_i = \frac{(1-t_i)(1-c_i)(1+b_i)}{2}(\mathbf{P}_i-\mathbf{P}_{i-1}) + \frac{(1-t_i)(1+c_i)(1-b_i)}{2}(\mathbf{P}_{i+1}-\mathbf{P}_i) \qquad (108)$$

$$\mathbf{DD}_i = \frac{(1-t_i)(1+c_i)(1+b_i)}{2}(\mathbf{P}_i-\mathbf{P}_{i-1}) + \frac{(1-t_i)\cdot(1-c_i)\cdot(1-b_i)}{2}\cdot(\mathbf{P}_{i+1}-\mathbf{P}_i) \qquad (109)$$

This composite formula provides considerable flexibility in the construction of trajectories (Figure 167).

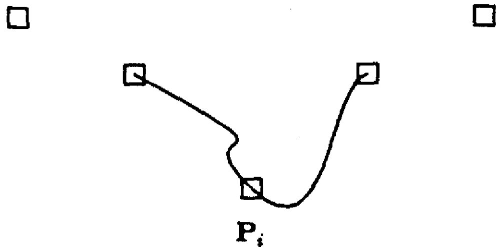


Figure 167. An interpolation at $\mathbf{P}_i$ with $b_{i-1} = 2$, $b_i = 1$, $b_{i+1} = 0$, $c_{i-1} = 2$, $c_i = 0$, $c_{i+1} = 2$, $t_{i-1} = 0$, $t_i = -1$, and $t_{i+1} = 0$.

## 15.2. A Cardinal-Spline Basis for Interpolation

The functions $h_0(u)$, $h_1(u)$, $h_2(u)$, and $h_3(u)$ which we used in section 15.1 to achieve cubic $C^1$ spline interpolation are not, themselves, cubic $C^1$ splines. That is, while they may constitute basis functions for the segment polynomials of the spline curves we constructed, they are not basis splines. The most profound name that we could give to them, since they "blend" together any collection of given data values and derivative values into a cubic $C^1$ spline, is *blending functions*.

The term "blending functions" is generally used to denote a linearly independent collection of functions which, like the Hermite basis polynomials of 15.1, serve to create piecewise functions that have at their joints given values and a number of given consecutive derivatives (first, second, third, etc.). Such functions are used frequently in graphics. The simplest are the *linear blending functions*

$$L_0(u) = 1-u \quad \text{and} \quad L_1(u) = u \quad \text{for } 0 \leq u \leq 1 \ ,$$

which constituted the segment polynomials for the uniform linear B-splines (the hat functions of chapter 4).

The linear B-splines are an example of the fact that blending functions can also be basis splines. The hat functions clearly "blend" given data values together into an interpolating linear $C^0$ spline. Such splines (those which act simultaneously as blending functions and basis functions) are individually called *cardinal splines*, and are collectively called a *cardinal-spline basis*. Once the idea is presented, the construction of cardinal bases is a straightforward exercise. For example, taking the hint from the fact that the hat functions are merely the linear blending functions "placed back-to-back," it is easy to see that translates of the two cubic $C^1$ functions shown in Figure 168 could be used to form a cardinal basis for cubic $C^1$ Hermite interpolation.

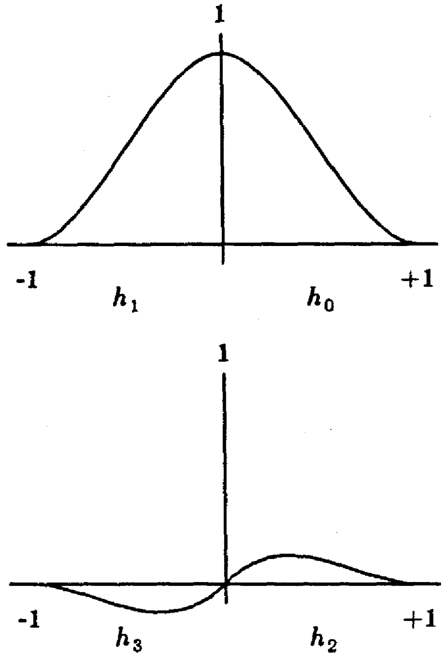


Figure 168. The two distinct members of the cubic, $C^1$, cardinal spline basis, suitable for use in piecewise cubic Hermite interpolation.

In closing we present two members of a cardinal cubic $C^2$ basis that are suited for the interpolation of data values. These basis elements produce interpolating splines satisfying the end conditions used by [Forsythe77] and discussed in chapter 3 (see Figure 8). They are generated by interpolating the data values

$$1 , 0 , 0 , \ldots , , 0 , 0 , 0$$

$$0 , 1 , 0 , \ldots , , 0 , 0 , 0$$

$$0 , 0 , 1 , \ldots , , 0 , 0 , 0$$

and so on, finishing with

$$0 , 0 , 0 , \ldots , , 0 , 1 , 0$$

$$0 , 0 , 0 , \ldots , , 0 , 0 , 1 \ .$$

That is, the first basis function is the spline satisfying the Forsythe, Malcolm and Moler end conditions that interpolates

$$1 , 0 , 0 , \ldots , , 0 , 0 , 0 \ ,$$

the second basis function is the spline satisfying the Forsythe, Malcolm and Moler end conditions that interpolates

$$0 , 1 , 0 , \ldots , , 0 , 0 , 0 \ ,$$

and so on. It is easy to see that if each is multiplied by a control vertex then their sum interpolates all the control vertices.

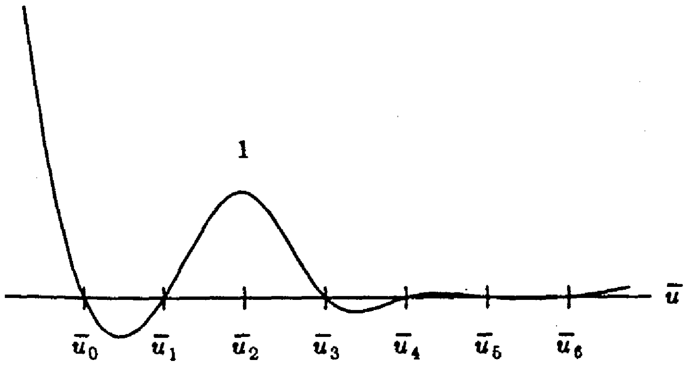


Figure 169. The cardinal, cubic, $C^2$ basis spline, with Forsythe, Malcolm and Moler end conditions, suitable for interpolating a data value at $\bar{u}_2$.

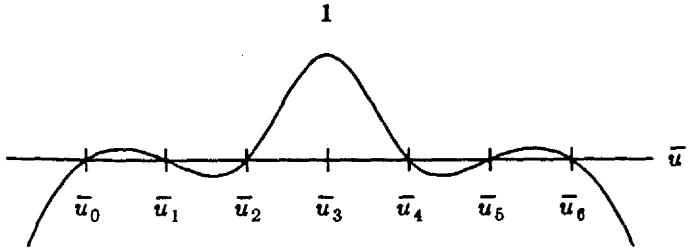


Figure 170. The cardinal, cubic, $C^2$ basis spline, with Forsythe, Malcolm and Moler end conditions, suitable for interpolating a data value at $\bar{u}_3$.

### 15.3. Interpolation Using B-Splines

We will present the material of this section primarily in terms of curves. Some of the material in [Wu77] amplifies what we will present, and a more complete treatment of surface interpolation may be found in [Barsky80].

The problem to be solved is the following: given some points $\mathbf{P}_j$, find control vertices $\mathbf{V}_i$ such that at each knot $\bar{u}_j$ in the range $[\bar{u}_{k-1},\bar{u}_{m+1}]$ the curve attains a specified point. That is, we want to compute $\mathbf{V}_i$ such that

$$\mathbf{Q}(\bar{u}_j) = \sum_{i=0}^{m} \mathbf{V}_i B_{i,k}(\bar{u}_j) = \mathbf{P}_j \tag{110}$$

for all $j = k-1, \ldots, m+1$.

If $\mathbf{V}_i = (x_i,y_i,z_i)$ and $\mathbf{P}_j = (r_j,s_j,t_j)$, then (110) can be written in terms of individual components as follows (using the $y$ components for illustration):

$$Y(\bar{u}_j) = \sum_{i=0}^{m} y_i B_{i,k}(\bar{u}_j) = s_j \quad \text{for} \quad j = k-1, \ldots, m+1 \ .$$

This constitutes a system of $m-k+3$ equations in $m+1$ unknowns:

$$\begin{bmatrix} B_{0,k}(\bar{u}_{k-1}) & \cdots & B_{m,k}(\bar{u}_{k-1}) \\ \cdot & & \cdot \\ \cdot & & \cdot \\ \cdot & & \cdot \\ B_{0,k}(\bar{u}_{m+1}) & \cdots & B_{m,k}(\bar{u}_{m+1}) \end{bmatrix} \begin{bmatrix} y_0 \\ \cdot \\ \cdot \\ \cdot \\ y_m \end{bmatrix} = \begin{bmatrix} s_{k-1} \\ \cdot \\ \cdot \\ \cdot \\ s_{m+1} \end{bmatrix} . \qquad (111)$$

We are short $k-2$ equations. For cubic splines $k$ is 4, and we are short 2 equations, just as we were in sections 3.1 and 3.2; we might use any of the end conditions of sections 3.1 and 3.2 may be considered here. For example, the two extra equations

$$\sum_{i=0}^{m} y_i B_{i,4}^{(2)}(\bar{u}_{k-1}) = 0$$

and

$$\sum_{i=0}^{m} y_i B_{i,4}^{(2)}(\bar{u}_{m+1}) = 0$$

may be added to provide a system of equations defining the B-spline representation of the natural cubic interpolating spline.

In the B-spline formulation of the interpolation problem, however, an alternative selection of extra conditions becomes evident. We may choose auxiliary data values associated with knots in the range $[\bar{u}_0,\bar{u}_{k-2}]$ and in the range $[\bar{u}_{m+2},\bar{u}_{m+k}]$ to form the extra equations needed to fill out system (111). For example, in the cubic case, we could select values $s_{k-2}$ and $s_{m+2}$ as "boundary values" to produce the full, nonsingular system of equations. (This amounts to selecting two additional points $\mathbf{P}_{k-2}$ and $\mathbf{P}_{m+2}$ to be interpolated on extensions of the curve at $\mathbf{Q}(\bar{u}_{k-2})$ and at $\mathbf{Q}(\bar{u}_{m+2})$.)

$$\begin{bmatrix} B_{0,k}(\bar{u}_{k-2}) & \cdots & B_{m,k}(\bar{u}_{k-2}) \\ B_{0,k}(\bar{u}_{k-1}) & & B_{m,k}(\bar{u}_{k-1}) \\ \cdot & & \cdot \\ \cdot & & \cdot \\ \cdot & & \cdot \\ B_{0,k}(\bar{u}_{m+1}) & & B_{m,k}(\bar{u}_{m+1}) \\ B_{0,k}(\bar{u}_{m+2}) & \cdots & B_{m,k}(\bar{u}_{m+2}) \end{bmatrix} \begin{bmatrix} y_0 \\ \\ \cdot \\ \cdot \\ \cdot \\ \\ y_m \end{bmatrix} = \begin{bmatrix} s_{k-2} \\ s_{k-1} \\ \cdot \\ \cdot \\ \cdot \\ s_{m+1} \\ s_{m+2} \end{bmatrix} . \qquad (112)$$

These auxiliary data values can be chosen at will, though the shape of the resulting curve will depend, nonlocally, upon the specific values chosen.

An interesting candidate for auxiliary values are those which approximate the standard first-derivative and second-derivative end conditions. Considering only the cubic case for illustration, suppose we wished to approximate the natural end condition:

$$\sum_{i=0}^{m} y_i B_{i,4}^{(2)}(\bar{u}_{k-1}) = 0 .$$

We can use the fact that, by Taylor expansion,

$$B_{i,k}(\bar{u}_{k-2}) \approx B_{i,k}(\bar{u}_{k-1}) + (\bar{u}_{k-2}-\bar{u}_{k-1})B_{i,k}^{(1)}(\bar{u}_{k-1}) \\ + \frac{1}{2}(\bar{u}_{k-2}-\bar{u}_{k-1})^2 B_{i,k}^{(2)}(\bar{u}_{k-1}) \ .$$

Setting $B_{i,k}^{(2)}(\bar{u}_{k-1})$ to zero yields

$$B_{i,k}(\bar{u}_{k-2}) \approx B_{i,k}(\bar{u}_{k-1}) + (\bar{u}_{k-2}-\bar{u}_{k-1})B_{i,k}^{(1)}(\bar{u}_{k-1}) \ .$$

If this approximate equality is multiplied by $y_i$ and the result is summed up for $i = 0, \ldots, m$, we obtain the approximate equality:

$$\sum_{i=0}^{m} y_i B_{i,k}(\bar{u}_{k-2}) \approx \sum_{i=0}^{m} y_i B_{i,k}(\bar{u}_{k-1}) + \sum_{i=0}^{m} y_i (\bar{u}_{k-2}-\bar{u}_{k-1})B_{i,k}^{(1)}(\bar{u}_{k-1}) \ .$$

Observe that the first summation on the right is equal to $s_{k-1}$, from (111), so we may take

$$s_{k-2} = s_{k-1} + \sum_{i=0}^{m} y_i (\bar{u}_{k-2}-\bar{u}_{k-1})B_{i,k}^{(1)}(\bar{u}_{k-1})$$

as an auxiliary value that defines an approximately natural cubic interpolating spline — for surfaces, this observation is essentially what underlies the material in [Barsky80].

Note that system (112) will have zero entries in the matrix except in a band $k-1$ entries wide along the main diagonal. The follows from the locality of the B-splines. For the uniform cubic B-splines, the matrix of (112) reduces to

$$\frac{1}{6}\begin{bmatrix} 4 & 1 & & & \\ 1 & 4 & 1 & & \\ \cdot & \cdot & \cdot & \cdot & \\ & \cdot & \cdot & \cdot & \\ & & \cdot & \cdot & \cdot \\ & & 1 & 4 & 1 \\ & & & 1 & 4 \end{bmatrix} ,$$

which is familiar from 3.1 and 3.2.

The same discussion may be carried out for surfaces. In this case the problem to be solved is that of finding control vertices $\mathbf{V}_{i,j}$ such that

$$\mathbf{Q}(\bar{u}_f,\bar{v}_g) = \sum_{i=0}^{m}\sum_{j=0}^{n} \mathbf{V}_{i,j} B_{i,k}(\bar{u}_f) B_{j,l}(\bar{v}_g) = \mathbf{P}_{f,g}$$

for given points $\mathbf{P}_{f,g}$. (Note that the orders of the B-splines could be different in each parametric direction.) There are many fewer equations than unknowns, and extra equations can be selected to specify derivatives (tangents, curvature, etc.) around the periphery of the surface. As in the case of curves we may also fill in extra equations by specifying auxiliary data corresponding to knot-value pairs with $\bar{u}$ components in the range $\bar{u}_0, \ldots, \bar{u}_{k-2}$ and $\bar{u}_{m+2}, \ldots, \bar{u}_{m+k}$, and with $\bar{v}$ components in the range $\bar{v}_0, \ldots, \bar{v}_{l-2}$ and $\bar{v}_{n+2}, \ldots, \bar{v}_{n+l}$. For the bicubic case, the number of points $\mathbf{P}_{f,g}$ that must be given is $(m-1)\times(n-1)$, and the number of equations needed to make a complete system is $2n+2m$, which just happens to be the number of knots bordering the lattice of knot pairs, $(\bar{u}_f,\bar{v}_g)$, in the legal parameter range.

To illustrate the use of the material outlined in this section we display the following three pictures. The first, Figure 171, shows a milk drop striking a surface. The data for this picture was obtained as a 256×256 raster of 256 grey levels. The grey levels were then imagined to be one-dimensional points $\mathbf{P}_{f,g}$, for $f = 0, \ldots, 255$ $g = 0, \ldots, 255$ floating in the range

$$0.0 \leq \mathbf{P}_{f,g} \leq 255.0$$

and associated with the knots

$$\bar{u}_f = f+2 \quad \text{and} \quad \bar{v}_g = g+2 \ .$$

The totality of knots in $\bar{u}$ and $\bar{v}$ was taken to be

$$\{0,1,2,3, \ldots ,257,258,259,260\} \ ,$$

the grey levels corresponding to the bordering raster points; i.e. those along the horizontal and vertical pixel lines corresponding to $\bar{u} = 2$, $\bar{u} = 258$, $\bar{v} = 2$, and $\bar{v} = 258$. were taken to be the auxiliary data values, and the points $\mathbf{P}_{f,g}$ were interpolated. In effect, a one-pixel "frame" around the raster image was taken for use a interpolational "boundary" values, and the remaining 254×254 pixels were taken as interpolational "central-data" values.

Figure 171.

Figure 172 shows a selected portion of the interpolating "surface" that results. This represents a quarter-section of the original raster image, and it has been evaluated over 256×256 equally-spaced points in $\bar{u}$ and $\bar{v}$ to achieve the effect of magnifying the original photograph. (The effect of compression or of distortions can be obtained, likewise, by the density with which the surface is evaluated.)

Figure 172.

To provide a comparison between interpolation and control-graph approximation, Figure 173 displays the same portion of the bicubic spline produced by taking the grey levels as control vertices and merely constructing the "surface"

$$\sum_i \sum_j \mathbf{P}_{i,j} B_{i,4}(\bar{u}_i) B_{i,4}(\bar{v}_j) \ .$$

As in the interpolating case uniform cubic B-splines were used. Furthermore, in evaluating the both surfaces, any values that fell outside of the interval [0,255] were cut down (or up) to size.

Figure 173.

### 15.4. Catmull-Rom Splines

The material in this section derives from [DeRose84].

It has usually been our practice to define a curve in the form

$$\mathbf{Q}(\bar{u}) = \sum_{i=0}^{m} \mathbf{V}_i B_{i,k}(\bar{u}) \ , \tag{113}$$

but Catmull and Rom have noted in [Catmull74] that a more general formulation would be

$$\mathbf{Q}(\bar{u}) = \frac{\sum_{i=0}^{m} \mathbf{P}_i(\bar{u}) W_i(\bar{u})}{\sum_{i=0}^{m} W_i(\bar{u})} \ , \tag{114}$$

where the $W_i$ are a set of basis splines and the $\mathbf{P}_i$ are vector-valued functions. The summation in the denominator was included in the formulation to provide a normalization in case the $W_i$ did not sum to one, which is necessary for translation invariance. While

$$W_0(\bar{u}) \ , \ \ldots \ , \ W_m(\bar{u})$$

can be chosen as any basis splines, it is most reasonable to select them to have local support. The reason for this, which will become clearer as we proceed, is that the flexibility of the class of Catmull-Rom splines is kept under control by specifying the functions $\mathbf{P}_i$ to have "useful" values on the support $W_i$. We may ignore the values of $\mathbf{P}_i$ outside of this support, which makes the task of specifying them to achieve some useful effect manageable. These remarks imply that the B-splines and the Hermite cardinal-spline basis are reasonable choices for the basis splines, since they have local support, while the interplating cardinal-splines are a less reasonable choice. The following assumptions will be made:

- The functions $W_i(\bar{u})$ are nonzero over the parametric interval from $\bar{u}_i$ to $\bar{u}_{i+d}$ (excluding $\bar{u}_{i+d}$ but possibly including $\bar{u}_i$).
- The functions $\mathbf{P}_i$ satisfy $\mathbf{P}_i(\bar{u}_q) = \mathbf{V}_q$ for $q = i, i+1, \ldots, i+r$.

Our usual construction of curves is of the form (114), if we choose

$$W_i(\bar{u}) = B_{i,k}(\bar{u})$$

(for which $d = k$), and if we choose

$$\mathbf{P}_i(\bar{u}) = \mathbf{V}_i \quad \text{for all } \bar{u}$$

(for which $r = 0$).

In general it is the interaction of the width of the support, $d$, of the basis functions, $W_i$, and the - number of control vertices, $r$, interpolated by the vector-valued functions, $\mathbf{P}_i$, which dictates the character of the resulting curve. The example $W = B$ and $\mathbf{P} = \mathbf{V}$ that was given above demonstrates that the curves we have usually been studying, which have the character of approximating the control vertices $\mathbf{V}_i$, are a special case of the Catmull-Rom splines. We now show that, if $r$ is increased, any curve of the form (114) can be made to interpolate the control vertices. Observe that the only functions $W$ that are nonzero over the interval $[\bar{u}_q, \bar{u}_{q+1})$ are

$$W_{q-d+1}(\bar{u}), \ldots, W_q(\bar{u}) .$$

Hence, for

$$\bar{u}_q \le \bar{u} < \bar{u}_{q+1} ,$$

(114) reduces to

$$\mathbf{Q}(\bar{u}) = \frac{\sum_{i=q-d+1}^{q} \mathbf{P}_i(\bar{u}) W_i(\bar{u})}{\sum_{i=q-d+1}^{q} W_i(\bar{u})} ,$$

and at $\bar{u} = \bar{u}_q$

$$\mathbf{Q}(\bar{u}_q) = \frac{\sum_{i=q-d+1}^{q} \mathbf{P}_i(\bar{u}_q) W_i(\bar{u}_q)}{\sum_{i=q-d+1}^{q} W_i(\bar{u}_q)} . \tag{115}$$

Now, if $r \ge d-1$, then

$$\mathbf{P}_i(\bar{u}_q) = \mathbf{V}_q ,$$

and (115) reduces to

$$\mathbf{Q}(\bar{u}_q) = \frac{\mathbf{V}_q \sum_{i=q-d+1}^{q} W_i(\bar{u}_q)}{\sum_{i=q-d+1}^{q} W_i(\bar{u}_q)} = \mathbf{V}_q .$$

Any vector-valued functions $\mathbf{P}_i$ satisfying the condition that

$$\mathbf{P}_i(\bar{u}_q) = \mathbf{V}_q$$

for $q = i, i+1, \ldots, i+r$ will serve to define a Catmull-Rom spline. Catmull and Rom themselves chose to use

$$\mathbf{P}_i(\bar{u}) = \sum_{j=0}^{r} \mathbf{V}_{i+j} L_j(\bar{u}) \ ,$$

where $L_j$ is the classical *Lagrange polynomial*:

$$L_j(\bar{u}) = \prod_{\substack{p=0 \\ p \neq j}}^{r} \left( \frac{\bar{u}-\bar{u}_p}{\bar{u}_j-\bar{u}_p} \right) \ ,$$

but one might choose to replace the Lagrange polynomials with the interpolating cardinal-spline basis, with step functions, or with other convenient functions. In [DeRose84] the Lagrange polynomials are replaced with functions specially chosen to introduce ***shape parameters*** of the sort already mentioned in section 15.1 and to be discussed more fully in the chapters on ***geometric continuity***.

### 15.5. Representing Powers and One-Sided Powers by B-Splines

We will be concerned here with two important items in the material of the foregoing chapters:

- We have created a basis (the B-splines) from the one-sided power functions, which themselves formed a basis for $S(\mathbf{P}^k, \{\bar{u}_i\}_0^{m+k})$. The "basis-change transformation" consisted of the divided-difference operation, as we used it to achieve cancellation.
- The $k^{\text{th}}$-order polynomials form a subspace of $S(\mathbf{P}^k, \{\bar{u}_i\}_0^{m+k})$; consequently, any $k^{\text{th}}$-order polynomial must have a representation in terms of B-splines.

This chapter will be used to close up a circle of representational identities by establishing the following:

- The one-sided power functions have a representation in terms of the B-splines (vector-space theory assures us must be the case), and this representation will be explicitly given.
- The members of the power basis $(\bar{u}-\bar{u}_i)^j$ for $0 \leq j \leq k-1$ for $k^{\text{th}}$-order polynomials have a representation in terms of B-splines, and this representation will be explicitly given.

This material is included for a couple of reasons: it provides an overview for a number of important topics usually found in the mathematical treatment of splines, and it provided the background from which the authors of [Cohen80] first established Oslo subdivision. In connection with this latter reason, we will end the section by establishing an alternative, but equivalent, formula for the discrete B-splines, one which uses the divided-difference operator. Briefly, the material in this section will be presented in three stages:

1. We begin by determining that $(\bar{u}-t)^{k-1}$ can be represented in terms of the B-splines as

$$(\bar{u}-t)^{k-1} = \sum_{j=0}^{m+n} \psi_{j,k}(t) N_{j,k}(\bar{u})$$

for the coefficient functions

$$\psi_{j,k}(t) = \prod_{r=1}^{k-1} (\bar{w}_{j+r} - t) \ .$$

(This result is called Marsden's Lemma. See Figure 174.)

2. We next show that

$$(\bar{u}-t)_+^{k-1} = \sum_{j=0}^{m+n} \phi_{j,k}(t) N_{j,k}(\bar{u}) \qquad (116)$$

for

$$t \in \{\bar{u}_0, \ldots, \bar{u}_{m+k}\} = \{\bar{u}_i\}_0^{m+k}$$

and

$$\bar{u}_{k-1} = \bar{w}_{k-1} \leq \bar{u} < \bar{w}_{m+n+1} = \bar{u}_{m+1}$$

where

$$\phi_{j,k}(t) = (\bar{w}_j + \epsilon_j - t)_+^0 \psi_{j,k}(t)$$

for (small) values of $\epsilon_j$ to be specified. (See Figure 175.) The term $(\bar{w}_j + \epsilon_j - t)_+^0$ is introduced simply to force $\psi_{j,k}(t)$ to zero left of $t$.

3. Since

$$B_{i,k}(\bar{u}) = (-1)^k (\bar{u}_{i+k} - \bar{u}_i)[\bar{u}_i(k):t](\bar{u} - t)_+^{k-1}$$

by substituting (116) we have

$$B_{i,k}(\bar{u}) = (-1)^k (\bar{u}_{i+k} - \bar{u}_i)[\bar{u}_i(k):t] \sum_{j=0}^{m+n} \phi_{j,k}(t) N_{j,k}(\bar{u})$$

and then rearranging yields

$$B_{i,k}(\bar{u}) = \sum_{j=0}^{m+n} \left\{ (-1)^k (\bar{u}_{i+k} - \bar{u}_i)[\bar{u}_i(k):t]\ \phi_{j,k}(t) \right\} N_{j,k}(\bar{u}) \ . \tag{117}$$

Comparing equation (117) with equation (72), we see that the $\alpha_{i,k}(j)$ are given by

$$\alpha_{i,k}(j) = (-1)^k (\bar{u}_{i+k} - \bar{u}_i)[\bar{u}_i(k):t]\ \phi_{j,k}(t) \tag{118}$$

and this result will end the section.

Now for the details.

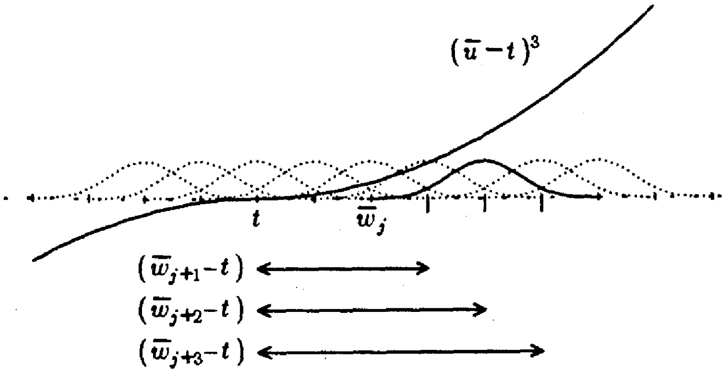


Figure 174. Marsden's Lemma for the cubic case. The coefficient function $\psi_{j,4}(t)$ is a product of the three terms $(\bar{w}_{j+1} - t)$, $(\bar{w}_{j+2} - t)$, and $(\bar{w}_{j+3} - t)$ shown here.

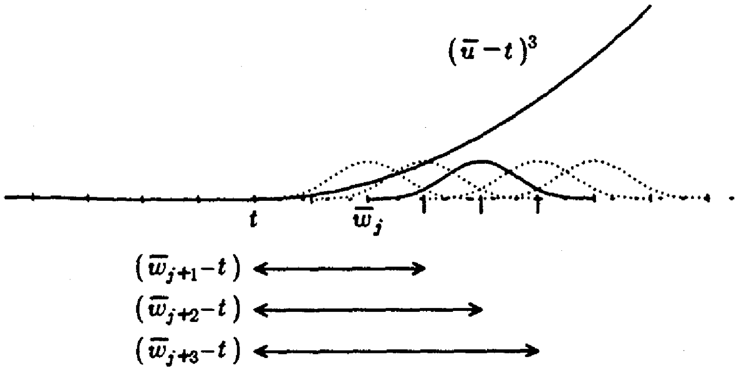


Figure 175. The extension of Marsden's Lemma to the onesided cubic. Notice that the B-splines which go positive to the left of $\bar{w}_j$ in Figure 174 have disappeared — the term $(\bar{w}_j+\epsilon_j-t)_+^0$ (think of $\epsilon_j$ as being zero for the time being) is responsible for this since it is zero to the left of $t$.

### 15.5.1. Marsden's Lemma

We will try to establish some insight into the way in which $(\bar{u}-t)_+^{k-1}$ can be represented by the "refined" B-splines, $N_{i,k}(\bar{u})$. Since multiple knots add complexity to the issues we wish to motivate, we will ignore them for the time being.

For the case $k=1$, $S(P^k,\{\bar{w}_j\}_0^{m+n+k})$ can be represented by the basis shown in Figure 176.

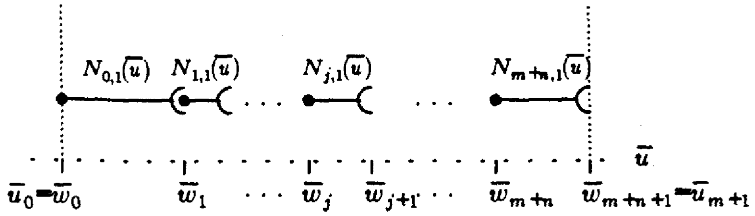


Figure 176. The B-spline basis of $S(P^1,\{\bar{w}_j\}_0^{m+n+k})$.

Consider $(\bar{u}-t)^0$, which is shown in Figure 177.

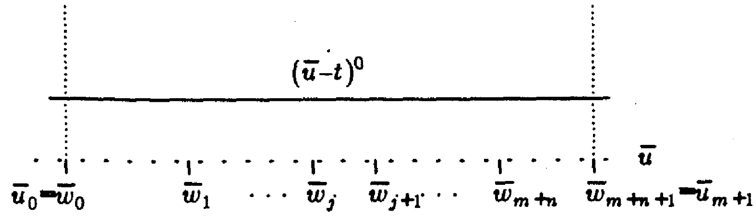


Figure 177. The power function $(\bar{u}-t)^0$.

The representation of $(\bar{u}-t)^0$ using the basis of Figure 176 is represented in Figure 178.

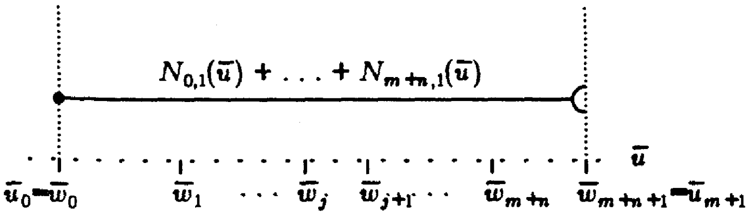


Figure 178. The representation of $(\bar{u}-t)^0$ using the first-order B-splines $N_{j,1}(\bar{u})$.

We can write the representation in Figure 178 in the form

$$(\bar{u}-t)^0 = \sum_{j=0}^{m+n} \psi_{j,1}(t) N_{j,1}(\bar{u})$$

where (rather trivially)

$$\psi_{j,1}(t) = 1 \quad \text{for all } t,\ j=0,\ldots,m+n \ .$$

Now, consider the one-sided power function

$$(\bar{u}-t)_+^0$$

where $t=\bar{u}_i$.

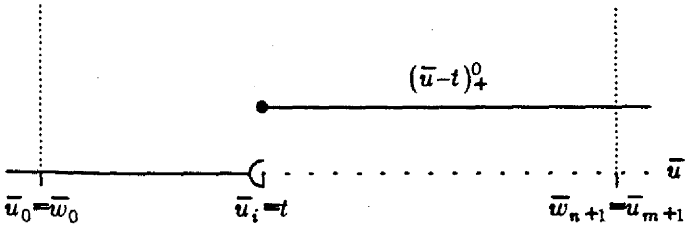


Figure 179. The one-sided power function $(\bar{u}-t)_+^0$

By inspection, for any $\bar{u}$ in the parameter range, i.e. $\bar{u}_0=\bar{w}_0 \le \bar{u} < \bar{w}_{m+n+1}=\bar{u}_{m+1}$,

$$(\bar{u}-t)_+^0 = (\bar{u}-\bar{u}_i)_+^0 = N_{\eta,1}(\bar{u}) + \cdots + N_{m+n,1}(\bar{u})$$

where $\eta=\eta(i)$ is the unique index such that $\bar{u}_i=\bar{w}_{\eta(j)}$. (Also note the $\bar{w}_j<\bar{u}_{i+1}$.) This is shown in Figure 180.

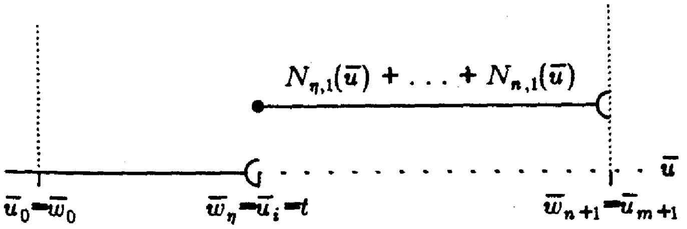

$N_{\eta,1}(\bar{u}) + \ldots + N_{n,1}(\bar{u})$
$\bar{u}$
$\bar{u}_0 = \bar{w}_0$
$\bar{w}_\eta = \bar{u}_i = t$
$\bar{w}_{n+1} = \bar{u}_{m+1}$

Figure 180. The representation of $(\bar{u}-t)_+^0$ using the first-order B-splines $N_{j,1}(\bar{u})$.

Recall that $t = \bar{u}_i = \bar{w}_\eta$. Hence the representation for $(\bar{u}-t)_+^0$ is merely a "truncated version" of the representation for $(\bar{u}-t)^0$ in which the first $\eta$ terms in the $N_{j,1}(\bar{u})$-representation of $(\bar{u}-t)^0$ have been suppressed with a zero coefficient. This means that we could consider the representation of $(\bar{u}-t)_+^0$ to be

$$(\bar{u}-t)_+^0 = (\bar{u}-\bar{u}_i)_+^0 = \sum_{j=0}^{m+n} \phi_{j,1}(t)\, N_{j,1}(\bar{u})$$

where the coefficient functions $\phi_{j,1}(t)$ satisfy

$$\phi_{j,1}(t) = \begin{cases} 0 & \text{for } j < \eta \\ 1 & \text{for } j \geq \eta \end{cases} .$$

An easy way to generate these coefficients $\phi$ is to notice that "to the left" of $t = \bar{u}_i = \bar{w}_\eta$ (that is, for $j < \eta$),

$$(\bar{w}_j - t)_+^0 = 0 = 0 \cdot \psi_{j,1}(t) .$$

Similarly, "directly at" $\bar{u}_i$ (that is, for $j = \eta$)

$$(\bar{w}_j - t)_+^0 = (\bar{w}_\eta - \bar{u}_i)_+^0 = (\bar{u}_i - \bar{u}_i)_+^0 = 0_+^0 = 1 = 1 \cdot \psi_{j,1}(t) .$$

Finally, "to the right" of $\bar{u}_i$ (that is, for $j > \eta$),

$$(\bar{w}_j - t)_+^0 = 1 = 1 \cdot \psi_{j,1}(t) .$$

So we may consider the following way of arranging to suppress terms in the representation of $(\bar{u}-t)^0$ to get the representation of $(\bar{u}-t)_+^0$ for $t = \bar{u}_i = \bar{w}_\eta$:

$$(\bar{u}-t)_+^0 = \sum_{j=0}^{m+n} (\bar{w}_j - t)_+^0\, \psi_{j,1}(t)\, N_{j,1}(\bar{u}) .$$

In fact, it should be pointed out that $(\bar{w}_j - t)_+^0$ has the same "0-1 properties" as

$$(\bar{w}_j + \epsilon_j - t)_+^0$$

has for any set of values $\epsilon_j \geq 0$, provided only that

$$\bar{w}_j + \epsilon_j < \bar{w}_{j+1} \quad \text{for all} \quad j < m+n+1 .$$

Figure 181 gives an example of each of the relevant cases.

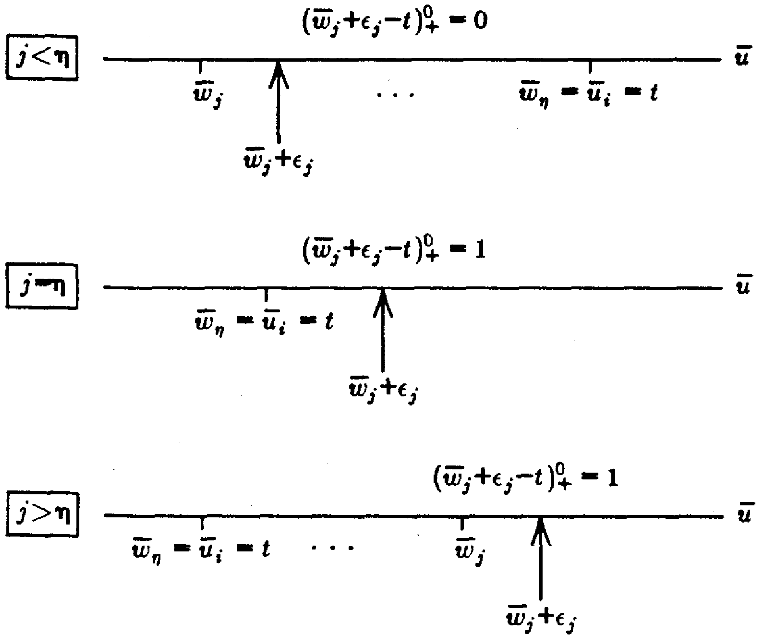

$j<\eta$
$(\bar{w}_j+\epsilon_j-t)^0_+ = 0$
$\bar{w}_j$
$\bar{w}_\eta = \bar{u}_i = t$
$\bar{w}_j+\epsilon_j$
$\bar{u}$
$j=\eta$
$(\bar{w}_j+\epsilon_j-t)^0_+ = 1$
$\bar{w}_\eta = \bar{u}_i = t$
$\bar{w}_j+\epsilon_j$
$\bar{u}$
$j>\eta$
$(\bar{w}_j+\epsilon_j-t)^0_+ = 1$
$\bar{w}_\eta = \bar{u}_i = t$
$\bar{w}_j$
$\bar{w}_j+\epsilon_j$
$\bar{u}$

Figure 181. The 0-1 properties of $(\bar{w}_j+\epsilon_j-t)^0_+$

Consequently,

$$(\bar{u}-t)^0_+ \equiv \sum_{j=0}^{m+n}(\bar{w}_j+\epsilon_j-t)^0_+ \,\psi_{j,1}(t)\, N_{j,1}(\bar{u})$$

for each $t \in \{\bar{u}_0, \ldots, \bar{u}_{m+k}\}$, and we let

$$\phi_{j,1}(t) = (\bar{w}_j+\epsilon_j-t)^0_+ \,\psi_{j,1}(t) \ .$$

The reason for adding the complication of the $\epsilon$'s is that it permits us to differentiate this equation with respect to $t$ in the appropriate (left-handed) fashion; i.e.

$$(\bar{w}_j+\epsilon_j-t)^0_+ = (\bar{w}_j-[t-\epsilon])^0_+$$

and therefore

$$\lim_{\substack{\epsilon \to 0 \\ \epsilon > 0}} \frac{(\bar{w}_j-[t-\epsilon])^0_+-(\bar{w}_j-t)^0_+}{-\epsilon}$$

will exist. Since we plan to take divided differences of our representation for $(\bar{u}-t)^{k-1}_+$ to obtain the quantities $\alpha$ that were mentioned at the close of the last subsection, and since divided differences become derivatives at multiple knots, this is necessary.

We will give a further illustration. Consider the case $k=2$. Suppose that we have found a representation for $(\bar{u}-t)^1$ when $t=\bar{u}_i\equiv\bar{w}_\eta$ as

$$(\bar{u}-t)^1 = \sum_{j=0}^{m+n} \psi_{j,2}(t)\, N_{j,2}(\bar{u})$$

and we wish to convert this into a representation for $(\bar{u}-t)^1_+$ by "truncating on the left." If this is feasible, then it is obvious from inspection how the chopping would have to proceed. The following picture indicates that the representation for $(\bar{u}-t)^1_+$ will not contain a contribution from $N_{j,2}(\bar{u})$ for any $j<\eta$, and it will include a contribution from $N_{j,2}(\bar{u})$ for all $j\geq\eta$.

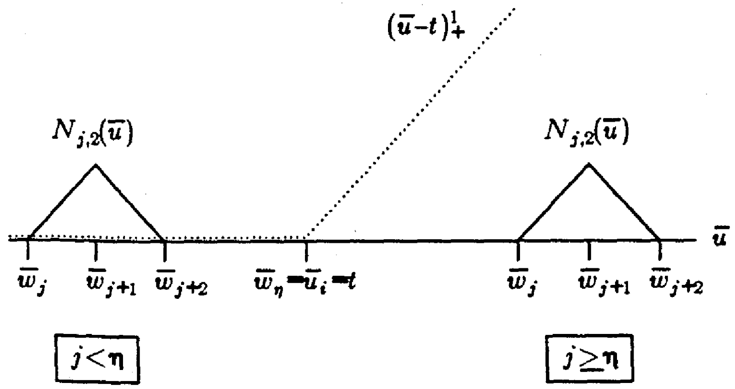


Figure 182. The one-sided power function $(\bar{u}-t)^1_+$ and two representative second-order B-splines on the knot sequence $\{\bar{w}_j\}_0^{m+n+k}$.

The method of arranging 0 and 1 coefficients used for the $k=1$ can surely be used here:

$$
\begin{aligned}
(\bar{u}-t)^1_+ &= \sum_{j=0}^{m+n} (\bar{w}_j+\epsilon_j-t)^0_+ \, \psi_{j,2}(t) \, N_{j,2}(\bar{u}) \\
&= \sum_{j=0}^{m+n} \phi_{j,2}(t) \, N_{j,2}(\bar{u}) \quad .
\end{aligned}
$$

Whether this will work, of course, is a question to be settled formally later. However, this hints at the approach we would like to take. We propose to express $(\bar{u}-t)^{k-1}$ as a linear combination of the $N_{i,k}(\bar{u})$. This is always possible since $(\bar{u}-t)^{k-1}$ is simply a polynomial of order $k$ in the variable $\bar{u}$ (a member of $\mathbf{P}^k$) and therefore must be a member of $\mathbf{S}(\mathbf{P}^k,\{\bar{w}_j\}_0^{m+n+k})$. Consequently,

$$
(\bar{u}-t)^{k-1} = \sum_{j=0}^{m+n} \psi_{j,k}(t) \, N_{j,k}(\bar{u})
$$

for some coefficients $\psi_{j,k}(t)$ depending upon $t$. Then we expect to find a representation for $(\bar{u}-t)^{k-1}_+$ by truncating the representation for $(\bar{u}-t)^{k-1}$, as was indicated above for the cases $k=1$ and $k=2$. Only one further insight is required. We need to discover a formula for these coefficients $\psi_{j,k}(t)$. The case $k=1$ is too simple to give us any clues, so we will explore the cases $k=2$ and $k=3$.

Consider Figure 183.

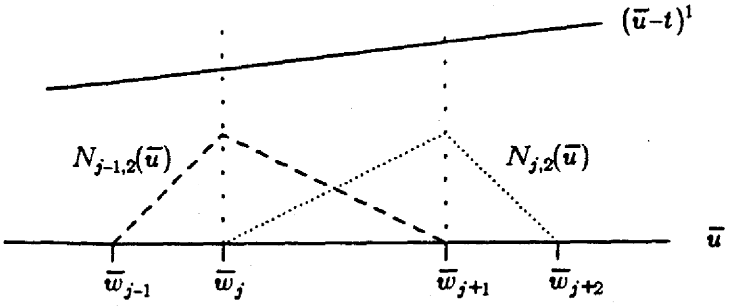


Figure 183. The power function $(\bar{u}-t)^1$ viewed over the interval $[\bar{w}_j,\bar{w}_{j+1})$ in $S(P^2,\{\bar{w}_j\}_0^{m+n+k})$.

By inspection, if $\bar{w}_j \leq \bar{u} < \bar{w}_{j+1}$, then we must have

$$(\bar{u}-t)^1 = A N_{j-1,2}(\bar{u}) + B N_{j,2}(\bar{u})$$

for some coefficients $A$and$B$ since $N_{j-1,2}(\bar{u})$ and $B N_{j,2}(\bar{u})$ are the only nonzero B-splines on this interval. But $N_{j-1,2}$ and $N_{j,2}$ can be produced from $N_{j-1,1}$, $N_{j,1}$, and $N_{j+1,1}$ by recurrence. In particular,

$$N_{j-1,2}(\bar{u}) = \frac{(\bar{w}_{j+1}-\bar{u})}{(\bar{w}_{j+1}-\bar{w}_j)} N_{j,1}(\bar{u}) + \frac{(\bar{u}-\bar{w}_{j-1})}{(\bar{w}_j-\bar{w}_{j-1})} N_{j-1,1}(\bar{u}) \ .$$

But for $\bar{w}_j \leq \bar{u} < \bar{w}_{j+1}$,

$$N_{j-1,1}(\bar{u}) = 0 \ \text{ and } \ N_{j,1}(\bar{u}) = 1 \ .$$

Hence

$$N_{j-1,2}(\bar{u}) = \frac{\bar{w}_{j+1}-\bar{u}}{\bar{w}_{j+1}-\bar{w}_j} \ .$$

Similarly,

$$N_{j,2}(\bar{u}) = \frac{\bar{u}-\bar{w}_j}{\bar{w}_{j+1}-\bar{w}_j}$$

for $\bar{w}_j \leq \bar{u} < \bar{w}_{j+1}$. This means that

$$\begin{aligned} \bar{u}-t &= A N_{j-1,2}(\bar{u}) + B N_{j,2}(\bar{u}) \\ &= A \frac{\bar{w}_{j+1}-\bar{u}}{\bar{w}_{j+1}-\bar{w}_j} + B \frac{\bar{u}-\bar{w}_j}{\bar{w}_{j+1}-\bar{w}_j} \ . \end{aligned} \qquad (119)$$

If we expand the right-hand side of (119) in powers of $\bar{u}$ and equate coefficients on the left and right, we get the equations

$$1 = \frac{B-A}{\bar{w}_{j+1}-\bar{w}_j}$$

and

$$-t = \frac{\bar{w}_{j+1}A - \bar{w}_j B}{\bar{w}_{j+1} - \bar{w}_j} \ .$$

This means that

$$\begin{aligned} \bar{w}_{j+1}A - \bar{w}_j B &= -(\bar{w}_{j+1} - \bar{w}_j)t \\ -A + B &= \bar{w}_{j+1} - \bar{w}_j \ . \end{aligned}$$

Eliminating $A$ gives

$$B = \bar{w}_{j+1} - t$$

and eliminating $B$ gives

$$A = \bar{w}_j - t \ .$$

For $\bar{u} \in [\bar{u}_{k-1}, \bar{u}_{m+1}) = [\bar{w}_{k-1}, \bar{w}_{m+n+1})$ in general, then, we would expect to have

$$(\bar{u} - t)^1 = \sum_{j=0}^{m+n} (\bar{w}_{j+1} - t)\, N_{j,2}(\bar{u}) \ .$$

Figure 184 shows a typical configuration when $k = 3$.

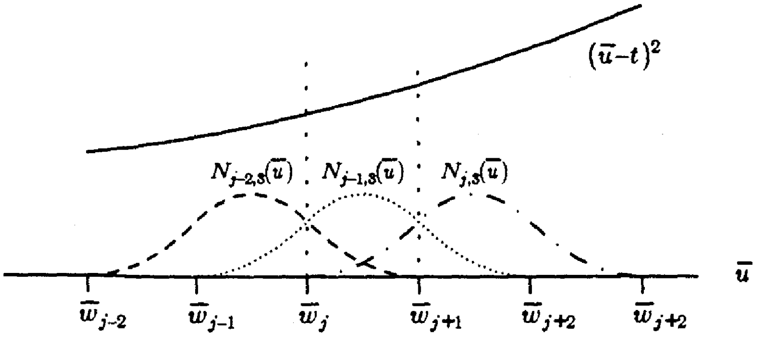


Figure 184. The power function $(\bar{u} - t)^2$ viewed over the interval $[\bar{w}_j, \bar{w}_{j+1})$ in $S(P^3, \{\bar{w}_j\}_0^{m+n+k})$.

We expect that, if $\bar{w}_j \le \bar{u} < \bar{w}_{j+1}$, then

$$(\bar{u} - t)^2 = A N_{j-2,3}(\bar{u}) + B N_{j-1,3}(\bar{u}) + C N_{j,3}(\bar{u}) \ . \qquad (120)$$

But each of the functions $N$ in this equation is derivable from the recurrence; i.e.

$$N_{j,3}(\bar{u}) = \frac{\bar{w}_{j+3} - \bar{u}}{\bar{w}_{j+3} - \bar{w}_{j+1}} N_{j+1,2}(\bar{u}) + \frac{\bar{u} - \bar{w}_j}{\bar{w}_{j+2} - \bar{w}_j} N_{j,2}(\bar{u})$$

and similarly for $N_{j-1,3}$ and $N_{j-2,3}$. But, if $\bar{u} \in [\bar{w}_j, \bar{w}_{j+1})$, then

$$N_{j+1,2}(\bar{u}) = 0$$

and

$$N_{j,2}(\bar{u}) = \frac{\bar{u}-\bar{w}_j}{\bar{w}_{j+1}-\bar{w}_j} \; .$$

Similarly,

$$N_{j-1,2}(\bar{u}) = \frac{\bar{w}_{j+1}-\bar{u}}{\bar{w}_{j+1}-\bar{w}_j}$$

and

$$N_{j-2,2}(\bar{u}) = 0 \; ,$$

and these terms appear in the recurrences for $N_{j-1,3}$ and $N_{j-2,3}$. We find that, for $\bar{w}_j \leq \bar{u} < \bar{w}_{j+1}$,

$$N_{j-2,3}(\bar{u}) = \frac{(\bar{w}_{j+1}-\bar{u})^2}{(\bar{w}_{j+1}-\bar{w}_{j-1})(\bar{w}_{j+1}-\bar{w}_j)}$$

$$N_{j-1,3}(\bar{u}) = \frac{(\bar{u}-\bar{w}_j)(\bar{w}_{j+2}-\bar{u})}{(\bar{w}_{j+1}-\bar{w}_j)(\bar{w}_{j+2}-\bar{w}_j)} + \frac{(\bar{u}-\bar{w}_{j-1})(\bar{w}_{j+1}-\bar{u})}{(\bar{w}_{j+1}-\bar{w}_{j-1})(\bar{w}_{j+1}-\bar{w}_j)}$$

and

$$N_{j,3}(\bar{u}) = \frac{(\bar{u}-\bar{w}_j)^2}{(\bar{w}_{j+1}-\bar{w}_j)(\bar{w}_{j+2}-\bar{w}_j)} \; .$$

If (120) is now written out, we obtain

$$\begin{aligned}
(\bar{u}-t)^2 &= \bar{u}^2 - 2\bar{u}t + t^2 \\
&= A\frac{(\bar{w}_{j+1}-\bar{u})^2}{(\bar{w}_{j+1}-\bar{w}_{j-1})(\bar{w}_{j+1}-\bar{w}_j)} \\
&\quad + B\left[\frac{(\bar{u}-\bar{w}_j)(\bar{w}_{j+2}-\bar{u})}{(\bar{w}_{j+1}-\bar{w}_j)(\bar{w}_{j+2}-\bar{w}_j)} + \frac{(\bar{u}-\bar{w}_{j-1})(\bar{w}_{j+1}-\bar{u})}{(\bar{w}_{j+1}-\bar{w}_{j-1})(\bar{w}_{j+1}-\bar{w}_j)}\right] . \\
&\quad + C\frac{(\bar{u}-\bar{w}_j)^2}{(\bar{w}_{j+1}-\bar{w}_j)(\bar{w}_{j+2}-\bar{w}_j)} \; .
\end{aligned}$$

If we expand in powers of $\bar{u}$ and equate coefficients, we obtain the following for the $t^2$ term:

$$\begin{aligned}
t^2 &= B[-\frac{\bar{w}_j\bar{w}_{j+2}}{(\bar{w}_{j+1}-\bar{w}_j)(\bar{w}_{j+2}-\bar{w}_j)} - \frac{\bar{w}_{j-1}\bar{w}_{j+1}}{(\bar{w}_{j+1}-\bar{w}_{j-1})(\bar{w}_{j+1}-\bar{w}_j)}] \\
&\quad + C\frac{\bar{w}_j{}^2}{(\bar{w}_{j+1}-\bar{w}_j)(\bar{w}_{j+2}-\bar{w}_j)} + A\frac{\bar{w}_{j+1}{}^2}{(\bar{w}_{j+1}-\bar{w}_{j-1})(\bar{w}_{j+1}-\bar{w}_j)}
\end{aligned}$$

the following for the $t^1$ term:

$$-2t = B[\frac{\bar{w}_{j+2}}{(\bar{w}_{j+1}-\bar{w}_j)(\bar{w}_{j+2}-\bar{w}_j)} + \frac{\bar{w}_j}{(\bar{w}_{j+1}-\bar{w}_j)(\bar{w}_{j+2}-\bar{w}_j)}$$

$$+ \frac{\bar{w}_{j+1}}{(\bar{w}_{j+1}-\bar{w}_{j-1})(\bar{w}_{j+1}-\bar{w}_j)} + \frac{\bar{w}_{j-1}}{(\bar{w}_{j+1}-\bar{w}_{j-1})(\bar{w}_{j+1}-\bar{w}_j)}]$$

$$-C\frac{2\bar{w}_j}{(\bar{w}_{j+1}-\bar{w}_j)(\bar{w}_{j+2}-\bar{w}_j)} - A\frac{2\bar{w}_{j+1}}{(\bar{w}_{j+1}-\bar{w}_{j-1})(\bar{w}_{j+1}-\bar{w}_j)}$$

and the following for the $t^0$ term:

$$1 = B[-\frac{1}{(\bar{w}_{j+1}-\bar{w}_j)(\bar{w}_{j+2}-\bar{w}_j)} - \frac{1}{(\bar{w}_{j+1}-\bar{w}_{j-1})(\bar{w}_{j+1}-\bar{w}_j)}]$$

$$+ C\frac{1}{(\bar{w}_{j+1}-\bar{w}_j)(\bar{w}_{j+2}-\bar{w}_j)}$$

$$+ A\frac{1}{(\bar{w}_{j+1}-\bar{w}_{j-1})(\bar{w}_{j+1}-\bar{w}_j)} \quad .$$

Solving these equations for $A$, $B$, and $C$ yields the simple expressions

$$\begin{aligned} A &= (\bar{w}_{j-1}-t)(\bar{w}_j-t) \\ B &= (\bar{w}_j-t)(\bar{w}_{j+1}-t) \\ C &= (\bar{w}_{j+1}-t)(\bar{w}_{j+2}-t) \ , \end{aligned}$$

which leads us to conclude that

$$(\bar{u}-t)^2 = \sum_{j=0}^{m+n}(\bar{w}_{j+1}-t)\,(\bar{w}_{j+2}-t)\,N_{j,3}(\bar{u}) \quad .$$

In general we now have cause to believe that

$$(\bar{u}-t)^{k-1} = \sum_{j=0}^{m+n}\psi_{j,k}(t)\,N_{j,k}(\bar{u})$$

where

$$\psi_{j,k}(t) = (\bar{w}_{j+1}-t)(\bar{w}_{j+2}-t)\cdots(\bar{w}_{j+k-1}-t) = \prod_{r=1}^{k-1}(\bar{w}_{j+r}-t) \quad .$$

This is true, and our method of arriving at this insight suggests the way in which it is most easily established; that is, by induction using the recurrence. We will try the first couple of steps to set the stage.

First consider the case $k=1$. On the one hand

$$(\bar{u}-t)^{k-1} = (\bar{u}-t)^0 = 1 \ ,$$

and on the other hand

$$\sum_{j=0}^{m+n} N_{j,1}(\bar{u}) = 1$$

for each value of $\bar{u}\in[\bar{u}_0,\bar{u}_{m+1}) = [\bar{w}_0,\bar{w}_{m+n+1})$. So need simply set

$$\psi_{j,1}(t) = 1$$

for all $j$ and $t$ in order that

$$(\bar{u}-t)^0 = \sum_{j=0}^{m+n} \psi_{j,1}(t) N_{j,1}(\bar{u}) \ .$$

Next consider the case $k=2$. We must study the summation

$$\sum_{j=0}^{m+n} \psi_{j,2}(t) N_{j,2}(\bar{u})$$

and we must find some expression for $\psi_{j,2}(t)$ which makes this sum equal to $(\bar{u}-t)^1$. The recurrence relation for the B-splines gives the clue. Briefly, we can rewrite the above sum as

$$\sum_{j=0}^{m+n} \psi_{j,2}(t) \left\{ (\bar{u}-\bar{w}_j)\frac{N_{j,1}(\bar{u})}{(\bar{w}_{j+1}-\bar{w}_j)} + (\bar{w}_{j+2}-\bar{u})\frac{N_{j+1,1}(\bar{u})}{(\bar{w}_{j+2}-\bar{w}_{j+1})} \right\} \tag{121}$$

where the convention applies that each of the ratios is to be interpreted as zero if its denominator is zero.

Observe that $N_{0,1}(\bar{u})$ is nonzero only when $\bar{w}_0 \le \bar{u} < \bar{w}_1$. Hence,

$$N_{0,1}(\bar{u}) = 0 \quad \text{for } \bar{u} \ge \bar{w}_1 \ .$$

But the parameter range for $k=2$ is $[\bar{w}_1, \bar{w}_{n+1})$, and so this restriction is in force. This means that we can ignore the left-hand ratio in (121) when $j=0$. For $j=m+n$, it should be noted that a term containing $N_{m+n+1,1}(\bar{u})$ is nominally present. But $N_{m+n+1,1}(\bar{u})$ is nonzero only when $\bar{w}_{m+n+1} \le \bar{u} < \bar{w}_{m+n+2}$, which is also outside of the parameter range. Hence

$$N_{m+n+1,1}(\bar{u}) = 0 \quad \text{for } \bar{u} < \bar{w}_{m+n+1} \ .$$

This means that we can ignore the term in the summation in which this function appears. So the summation expands as follows for any $\bar{u}$ in the parameter range:

$$\begin{aligned}
&0 + \psi_{0,2}(t)(\bar{w}_2-\bar{u})\frac{N_{1,1}(\bar{u})}{(\bar{w}_2-\bar{w}_1)} \\
&\quad + \psi_{1,2}(t)(\bar{u}-\bar{w}_1)\frac{N_{1,1}(\bar{u})}{(\bar{w}_2-\bar{w}_1)} + \psi_{1,2}(t)(\bar{w}_3-\bar{u})\frac{N_{2,1}(\bar{u})}{(\bar{w}_3-\bar{w}_2)} \\
&\quad + \psi_{2,2}(t)(\bar{u}-\bar{w}_2)\frac{N_{2,1}(\bar{u})}{(\bar{w}_3-\bar{w}_2)} + \psi_{2,2}(t)(\bar{w}_4-\bar{u})\frac{N_{3,1}(\bar{u})}{(\bar{w}_4-\bar{w}_3)} \\
&\quad + \cdots + \psi_{m+n,2}(t)(\bar{u}-\bar{w}_{m+n})\frac{N_{m+n,1}(\bar{u})}{(\bar{w}_{m+n+1}-\bar{w}_{m+n})} \\
&\quad + \psi_{m+n,2}(t)(\bar{w}_{m+n+2}-\bar{u})\frac{N_{m+n+1,1}(\bar{u})}{(\bar{w}_{m+n+2}-\bar{w}_{m+n+1})} + 0 \ .
\end{aligned}$$

Grouping terms in common indices for $N$ together, our summation becomes

$$\sum_{j=1}^{m+n} \left\{ \psi_{j-1,2}(t)\,(\bar{w}_{j+1}-\bar{u}) + \psi_{j,2}(t)\,(\bar{u}-\bar{w}_j) \right\} \frac{N_{j,1}(\bar{u})}{\bar{w}_{j+1}-\bar{w}_j} \ .$$

If only we can find an expression for the quantities $\psi$ which makes

$$\psi_{j-1,2}(t)\,(\overline{w}_{j+1}-\overline{u}) + \psi_{j,2}(t)\,(\overline{u}-\overline{w}_j) = (\overline{u}-t)\cdot(\overline{w}_{j+1}-\overline{w}_j)\ , \tag{122}$$

then our summation would become

$$\sum_{j=1}^{m+n} (\overline{u}-t)(\overline{w}_{j+1}-\overline{w}_j)\frac{N_{j,1}(\overline{u})}{\overline{w}_{j+1}-\overline{w}_j}$$

$$= (\overline{u}-t)\sum_{j=1}^{m+n} N_{j,1}(\overline{u}) = (\overline{u}-t)\cdot 1$$

for any $\overline{u}_{k-1}=\overline{w}_{k-1}\leq\overline{u}<\overline{w}_{m+n+1}=\overline{u}_{m+1}$, which is exactly what we want. But if we let

$$\psi_{j-1,2}(t) = (\overline{w}_j-t) \text{ and } \psi_{j,2}(t) = (\overline{w}_{j+1}-t)$$

it is easily verified that (122) holds.

**Theorem:** For any $t$ and any $\overline{u}\in[\overline{w}_{k-1},\overline{w}_{m+n+1}) = [\overline{u}_{k-1},\overline{u}_{m+1})$

$$(\overline{u}-t)^{k-1} = \sum_{j=0}^{m+n}\psi_{j,k}(t)\,N_{j,k}(\overline{u})$$

where

$$\psi_{j,k}(t) = 1 \quad \text{if } k=1$$

and

$$\psi_{j,k}(t) = (\overline{w}_{j+1}-t)(\overline{w}_{j+2}-t)\cdots(\overline{w}_{j+k-1}-t)$$

$$= \prod_{r=1}^{k-1}(\overline{w}_{j+r}-t) \quad \text{if } k>1\ .$$

This theorem is known as Marsden's Lemma. It establishes precisely the formula by which a power function can be represented in terms of B-splines.

**Argument:** The preliminary discussion gives the outline of an inductive proof. We have already seen that the result is true when $k=1$ and $k=2$.

For general $k>1$ the induction assumption is made that

$$\sum_{j=0}^{m+n}\psi_{j,k-1}(t)N_{j,k-1} = (\overline{u}-t)^{(k-1)-1} = (\overline{u}-t)^{(k-2)}$$

for all $t\in\{\overline{u}_i\}_0^{m+k}$ and all $\overline{u}\in[\overline{w}_{k-1},\overline{w}_{m+n+1}) = [\overline{u}_{k-1},\overline{u}_{m+1})$.

We take

$$\sum_{j=0}^{m+n}\psi_{j,k}(t)N_{j,k}(\overline{u})$$

and use the recurrence result to obtain

(more...)

$$\sum_{j=0}^{m+n} \psi_{j,k}(t) \left\{ (\bar{u}-\bar{w}_j)\frac{N_{j,k-1}(\bar{u})}{\bar{w}_{j+k-1}-\bar{w}_j} + (\bar{w}_{j+k}-\bar{u})\frac{N_{j+1,k-1}(\bar{u})}{\bar{w}_{j+k}-\bar{w}_{j+1}} \right\} .$$

With $\bar{u}$ restricted to $\bar{u} \in [\bar{w}_{k-1},\bar{w}_{m+n+1}) = [\bar{u}_{k-1},\bar{u}_{m+1})$, the terms with $N_{0,k-1}(\bar{u})$ and $N_{m+n+1,k-1}(\bar{u})$ can be ignored, and the summation can be regrouped to yield

$$\sum_{j=1}^{m+n} \left\{ \psi_{j-1,k}(t)\,(\bar{w}_{j+k-1}-\bar{u}) + \psi_{j,k}(t)\,(\bar{u}-\bar{w}_j) \right\} \frac{N_{j,k-1}(\bar{u})}{\bar{w}_{j+k-1}-\bar{w}_j} .$$

But

$$\psi_{j-1,k}(t) = (\bar{w}_j-t)(\bar{w}_{j+1}-t)\cdots(\bar{w}_{j+k-2}-t)$$

and

$$\psi_{j,k}(t) = (\bar{w}_{j+1}-t)\cdots(\bar{w}_{j+k-2}-t)(\bar{w}_{j+k-1}-t) .$$

Both of these have the common factor

$$(\bar{w}_{j+1}-t)\cdots(\bar{w}_{j+k-2}-t) = \psi_{j,k-1}(t) .$$

Consequently

$$\begin{aligned} &\psi_{j-1,k}(t)(\bar{w}_{j+k-1}-\bar{u})+\psi_{j,k}(t)(\bar{u}-\bar{w}_j) \\ &\quad = \Big[(\bar{w}_{j+k-1}-\bar{u})(\bar{w}_j-t) + (\bar{u}-\bar{w}_j)(\bar{w}_{j+k-1}-t)\Big]\,\psi_{j,k-1}(t) . \end{aligned}$$

And if the expression in brackets is multiplied out, terms cancel to give

$$\begin{aligned} &\psi_{j-1,k}(t)(\bar{w}_{j+k-1}-\bar{u})+\psi_{j,k}(t)(\bar{u}-\bar{w}_j) \\ &\quad = (\bar{w}_{j+k-1}-\bar{w}_j)(\bar{u}-t)\,\psi_{j,k-1}(t) . \end{aligned}$$

This converts the summation

$$\sum_{j=0}^{m+n} \psi_{j,k}(t)N_{j,k}(\bar{u})$$

into the summation

$$(\bar{u}-t)\sum_{j=1}^{m+n} \psi_{j,k-1}(t)N_{j,k-1}(\bar{u}) .$$

We may add the $j=0$ term to obtain

$$(\bar{u}-t)\sum_{j=0}^{m+n} \psi_{j,k-1}(t)N_{j,k-1}(\bar{u})$$

on the parameter range, since $N_{0,k-1}(\bar{u})$ is zero on that range.

Finally, by the induction hypothesis this equals

(more...)

$$(\bar{u}-t)\cdot(\bar{u}-t)^{k-2}$$

for any $\bar{u}\in[\bar{w}_{k-1},\bar{w}_{m+n+1})$.

### 15.5.2. Discrete B-splines

We now return to the process of truncating the coefficients $\psi$.

**Definition:** Let

$$\phi_{j,k}(t) = (\bar{w}_j+\epsilon_j-t)^0_+ \quad \text{if} \quad k=1$$

and

$$\begin{aligned}\phi_{j,k}(t) &= (\bar{w}_j+\epsilon_j-t)^0_+\psi_{j,k}(t)\\ &= (\bar{w}_j+\epsilon_j-t)^0_+(\bar{w}_{j+1}-t)(\bar{w}_{j+2}-t)\cdots(\bar{w}_{j+k-1}-t) \quad \text{if} \quad k>1 \ ,\end{aligned}$$

where the numbers $\epsilon_j$ satisfy

$$\bar{w}_j \le \bar{w}_j+\epsilon_j < \bar{w}_{\gamma_+(j)} \quad \text{for} \quad j<n+k$$

and

$$\epsilon_{n+k} \ge 0 \ .$$

(Recall that $\gamma_+(j)$ is the smallest index such that $\bar{w}_j < \bar{w}_{\gamma_+(j)}$.)

With the aid of the $\phi_{j,k}(t)$, the one-sided functions can be represented in terms of the $N_{j,k}(\bar{u})$:

**Theorem:** For any $t\in\{\bar{u}_0,\ldots,\bar{u}_{m+k}\} = \{\bar{u}_i\}_0^{m+k}$ and $\bar{u}\in[\bar{w}_{k-1},\bar{w}_{m+n+1}) = [\bar{u}_{k-1},\bar{u}_{m+1})$,

$$(\bar{u}-t)^{k-1}_+ = \sum_{j=0}^{m+n}\phi_{j,k}(t)N_{j,k}(\bar{u}) \ .$$

**Argument:**

($k=1$)

We have already established that

$$(\bar{u}-t)^0_+ = \sum_{j=0}^{m+n}(\bar{w}+\epsilon_j-t)^0_+N_{j,1}(\bar{u}) \ .$$

($k>1$)

(more...)

We divide this case into two parts.

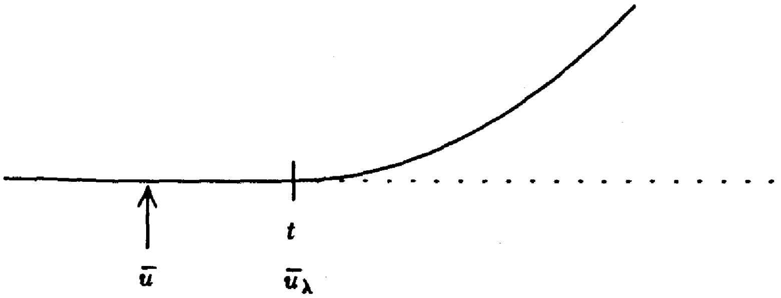


Figure 185. The first subcase of the induction step, in which case $\bar{u} < t$.

$(\bar{u} < t \in \{\bar{u}_i\}_0^{m+k})$

Notice that

$$(\bar{w}_j + \epsilon_j - t)_+^0 = 0 \quad \text{for all } j \text{ such that } \bar{w}_j < t$$

and

$$(\bar{w}_j + \epsilon_j - t)_+^0 = 1 \quad \text{for all } j \text{ such that } \bar{w}_j \geq t \ .$$

Consequently

$$\sum_{j=0}^{m+n} \phi_{j,k}(t) N_{j,k}(\bar{u}) = \sum_{j \text{ s.t. } \bar{w}_j \geq t}^{m+n} \phi_{j,k}(t) N_{j,k}(\bar{u}) \ .$$

But

$$N_{j,k}(\bar{u}) = 0 \text{ for all } \bar{u} < \bar{w}_j \ .$$

And, since

$$\bar{u} < t \text{ and } \bar{w}_j \geq t \ ,$$

the sum is zero.

On the other hand

$$(\bar{u} - t)_+^{k-1} = 0$$

for the values of $\bar{u}$ in question.

Since $(\bar{u} - t)_+^{k-1}$ and the summation expression are consistent, this part of the $k>1$ case is complete.

(more...)

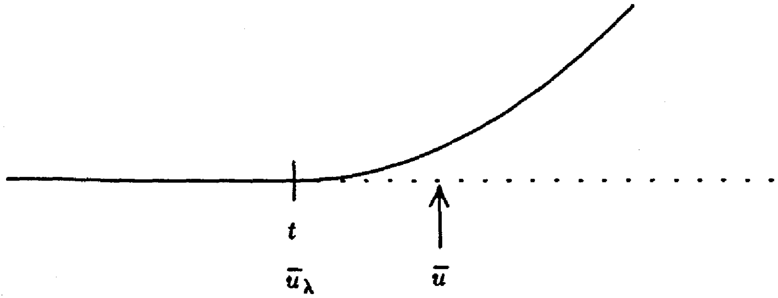


Figure 186. The second subcase of the induction step, in which case $\bar{u} \geq t$.

$(\bar{u} \geq t \in \{\bar{u}_i\}_0^{m+k})$

As in the argument for Marsden's Lemma just before,

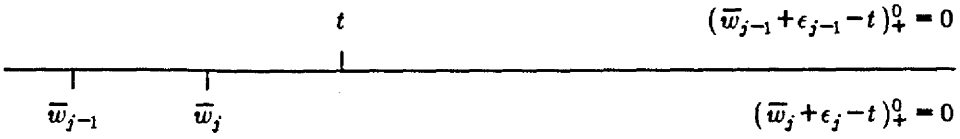
$$
\begin{aligned}
&\sum_{j=0}^{m+n} \phi_{j,k}(t) N_{j,k}(\bar{u}) \\
&\quad = \sum_{j=0}^{m+n} \phi_{j,k}(t) \left\{ \frac{\bar{w}_{j+k}-\bar{u}}{\bar{w}_{j+k}-\bar{w}_{j+1}} N_{j+1,k-1}(\bar{u}) + \frac{\bar{u}-\bar{w}_j}{\bar{w}_{j+k-1}-\bar{w}_j} N_{j,k-1}(\bar{u}) \right\} \\
&\quad = \sum_{j=1}^{m+n} \left\{ \phi_{j-1,k}(t)(\bar{w}_{j+k-1}-\bar{u}) + \phi_{j,k}(t)(\bar{u}-\bar{w}_j) \right\} \frac{N_{j,k-1}(\bar{u})}{\bar{w}_{j+k-1}-\bar{w}_j} \quad .
\end{aligned}
$$

Further, just as in the argument for Marsden's Lemma,

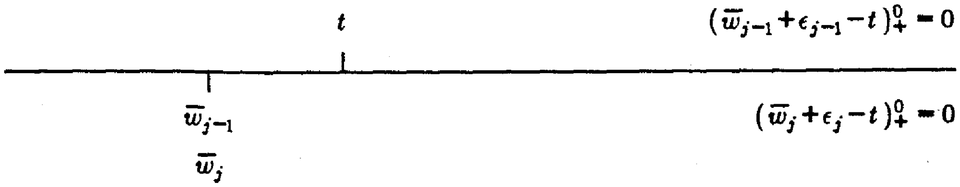
$$
\begin{aligned}
&\phi_{j-1,k}(t)(\bar{w}_{j+k-1}-\bar{u}) + \phi_{j,k}(t)(\bar{u}-\bar{w}_j) \\
&\quad = \left[ (\bar{w}_{j-1}+\epsilon_{j-1}-t)_+^0 (\bar{w}_{j+k-1}-\bar{u})(\bar{w}_j-t) + (\bar{w}_j+\epsilon_j-t)_+^0 (\bar{u}-\bar{w}_j)(\bar{w}_{j+k-1}-t) \right] \psi_{j,k-1}(t) \quad .
\end{aligned} \tag{123}
$$

The following are the only configurations possible.

[1]

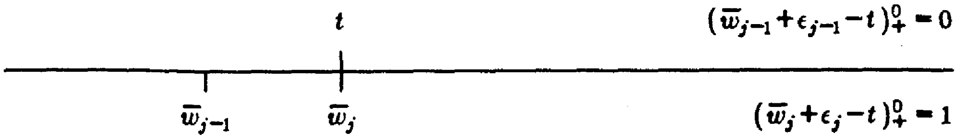

$t$
$(\bar{w}_{j-1}+\epsilon_{j-1}-t)_+^0 = 0$
$\bar{w}_{j-1}$
$\bar{w}_j$
$(\bar{w}_j+\epsilon_j-t)_+^0 = 0$

[2]


$t$
$(\bar{w}_{j-1}+\epsilon_{j-1}-t)_+^0 = 0$
$\bar{w}_{j-1}$
$(\bar{w}_j+\epsilon_j-t)_+^0 = 0$
$\bar{w}_j$

(more...)

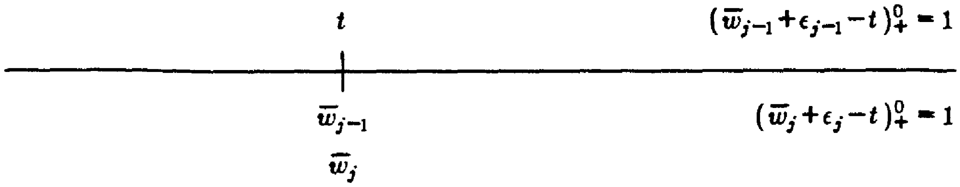
[3]



$$(\bar{w}_{j-1}+\epsilon_{j-1}-t)_+^0=0$$

$$(\bar{w}_j+\epsilon_j-t)_+^0=1$$

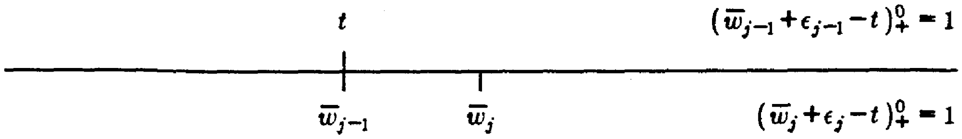
[4]



$$(\bar{w}_{j-1}+\epsilon_{j-1}-t)_+^0=1$$

$$(\bar{w}_j+\epsilon_j-t)_+^0=1$$

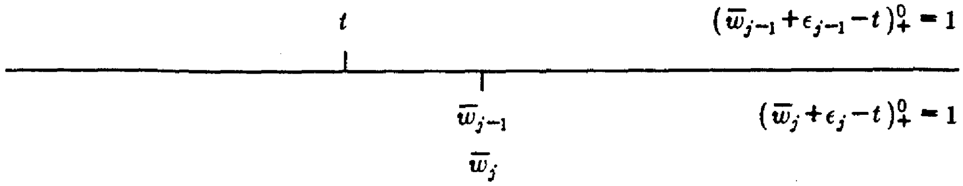
[5]



$$(\bar{w}_{j-1}+\epsilon_{j-1}-t)_+^0=1$$

$$(\bar{w}_j+\epsilon_j-t)_+^0=1$$

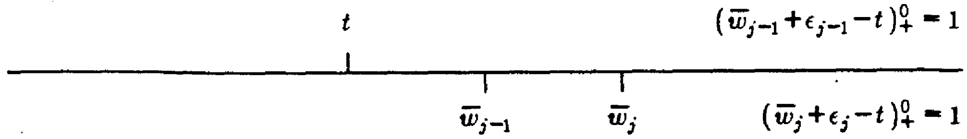
[6]



$$(\bar{w}_{j-1}+\epsilon_{j-1}-t)_+^0=1$$

$$(\bar{w}_j+\epsilon_j-t)_+^0=1$$

[7]



$$(\bar{w}_{j-1}+\epsilon_{j-1}-t)_+^0=1$$

$$(\bar{w}_j+\epsilon_j-t)_+^0=1$$

A complete case analysis is tedious and repetitive. By way of example, however, for configuration [3] it is easily checked that (123) becomes

$$(\bar{u}-t)(\bar{w}_{j+k-1}-\bar{w}_j)(\bar{w}_j+\epsilon_j-t)_+^0+\psi_{j,k-1}(t) \;,$$

where we have used the identities

$$(\bar{w}_j-t)=0 \;,\; (\bar{u}-\bar{w}_j)=(\bar{u}-t) \;,\; \text{and} \; (\bar{w}_{j+k-1}-t)=(\bar{w}_{j+k-1}-\bar{w}_j) \;.$$

For the other configurations, using the same algebra as was used in the argument for Marsden's Lemma, (123) becomes

$$(\bar{u}-t)(\bar{w}_{j+k-1}-\bar{w}_j)(\bar{w}_j+\epsilon_j-t)_+^0+\psi_{j,k-1}(t) \;.$$

Hence by the same inductive reasoning as was used in Marsden's Lemma, the summation becomes

(more...)

$$(\bar{u}-t)\sum_{j=1}^{m+n}\phi_{j,k-1}N_{j,k-1} = (\bar{u}-t)(\bar{u}-t)_+^{k-2} \ .$$

This completes the argument.

We have almost reached our goal. We know that

$$B_{i,k}(\bar{u}) = \sum_{j=0}^{m+n}\alpha_{i,k}(j)N_{j,k}(\bar{u})$$

$$B_{i,k}(\bar{u}) = (-1)^k(\bar{u}_{i+k}-\bar{u}_i)[\bar{u}_i(k):t](\bar{u}-t)_+^{k-1}$$

and

$$(\bar{u}-t)_+^{k-1} = \sum_{j=0}^{m+n}\phi_{j,k}(t)N_{j,k}(\bar{u})$$

for any $t \in \{\bar{u}_i\}_0^{m+k}$ and $\bar{u} \in [\bar{w}_{k-1},\bar{w}_{m+n+1}) = [\bar{u}_{k-1},\bar{u}_{m+1})$. We put these together in the obvious way to obtain equation (118), an expression for the $\alpha$'s.

**Definition:**

$$\alpha_{i,k}(j) = (-1)^k(\bar{u}_{i+k}-\bar{u}_i)[\bar{u}_i(k):t]\phi_{j,k}(t) \ .$$

**Theorem:**

$$B_{i,k}(\bar{u}) = \sum_{j=0}^{m+n}\alpha_{i,k}(j)N_{j,k}(\bar{u}) \ .$$

Notice that the functions

$$\phi_{j,k}(t) = (\bar{w}_j+\epsilon_j-t)_+^0\prod_{r=1}^{k-1}(\bar{w}_{j+r}-t) \qquad (\epsilon_j \approx 0)$$

look very much like the functions

$$(\bar{u}-t)_+^0\prod_{r=1}^{k-1}(\bar{u}-t) = (\bar{u}-t)_+^{k-1}$$

would look if the variable $\bar{u}$ were restricted to the discrete values in the sequence of knots $\{\bar{w}_j\}$. This presents another justification for regarding, $\alpha_{i,k}(j)$ to be a "discretized" version of $B_{i,k}(\bar{u})$; that is, a "revision" of the *continuous B-splines* for $\bar{u}$ restricted to a sequence of values.

### 15.5.3. The Discrete B-Spline Recurrence Again

Having established the connection between the discrete B-splines and divided differences, we close the material on discrete B-splines by establishing their recurrence properties strictly from the properties of the divided difference.

**Theorem:**

$$\alpha_{i,1}(j) = \begin{cases} 1 & \bar{u}_i \leq \bar{w}_j < \bar{u}_{i+1} \\ 0 & \text{otherwise} \end{cases}$$

and

$$\alpha_{i,r}(j) = \frac{\bar{w}_{j+r-1} - \bar{u}_i}{\bar{u}_{i+r-1} - \bar{u}_i}\,\alpha_{i,r-1}(j) + \frac{\bar{u}_{i+r} - \bar{w}_{j+r-1}}{\bar{u}_{i+r} - \bar{u}_{i+1}}\,\alpha_{i+1,r-1}(j)$$

for $r = 2, 3, \ldots, k$, where $k$ is the order of the spline in question.

**Argument:**

To establish the result for $k = 1$, it helps to keep a picture of $\phi_{j,1}(t) = (\bar{w}_j + \epsilon_j - t)_+^0$ in mind (where $\epsilon_j \geq 0$, and is to be regarded as "vanishingly small" if positive). This function is of interest only for $t$ values in the set of knots $\{\bar{u}_i\}_0^{m+k}$. Its values in a neighbourhood of $\bar{w}_j$ would typically look as follows:

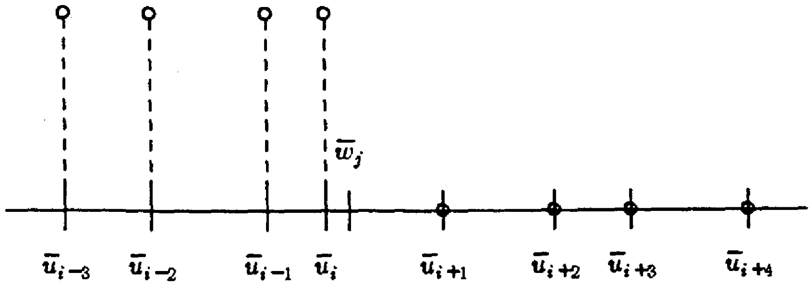


Figure 187. A graph of $\phi_{j,1}(t)$ for values of $t \in \{\bar{u}_i\}_0^{m+k}$ near $\bar{w}_j$.

This picture is not changed materially if $\bar{w}_j$ falls upon $\bar{u}_i$ or if $\bar{u}_i = \bar{u}_{i+1}$. The essential feature of $\phi_{j,1}(t)$ is that

$$\phi_{j,1}(t) = 1 \;\; \mathit{for} \;\; t = \bar{u}_i \leq \bar{w}_j$$

and

(more...)

$$\phi_{j,1}(t) \;=\; 0 \quad \mathit{for} \quad t = \bar{u}_i > \bar{w}_j$$

We split the case $k=1$ into three mutually exclusive possibilities.

$k=1$, Case a:

$$(\bar{u}_i < \bar{u}_{i+1} \;\; \mathit{and} \;\; \bar{u}_i \leq \bar{w}_j < \bar{u}_{i+1})$$

Then

$$\begin{aligned}
\alpha_{i,1}(j) \;&=\; (-1)(\bar{u}_{i+1}-\bar{u}_i)[\bar{u}_i(1):t]\phi_{j,1}(t) \\
&=\; (-1)(\bar{u}_{i+1}-\bar{u}_i)\frac{\phi_{j,1}(\bar{u}_{i+1})-\phi_{j,1}(\bar{u}_i)}{\bar{u}_{i+1}-\bar{u}_i} \\
&=\; \phi_{j,1}(\bar{u}_i)-\phi_{j,1}(\bar{u}_{i+1}) \\
&=\; (\bar{w}_j-\bar{u}_i)_+^0 - (\bar{w}_j-\bar{u}_{i+1})_+^0 \\
&=\; 1 - 0 \;=\; 1
\end{aligned}$$

$k=1$, Case b:

$$(\bar{w}_j < \bar{u}_i \;\; \text{or} \;\; \bar{w}_j \geq \bar{u}_{i+1}, \;\; \text{and} \;\; \bar{u}_i < \bar{u}_{i+1})$$

In the manner of case a,

$$\alpha_{i,1}(j) \;=\; (\bar{w}_j-\bar{u}_i)_+^0 - (\bar{w}_j-\bar{u}_{i+1})_+^0 \;=\; 0$$

because both of the terms of the difference have the same value. (Both are equal to 1 or else both are equal to 0, depending on whether or not $\bar{w}_j$ lies to the left of $\bar{u}_j$.)

$k=1$, Case c:

$$(\bar{u}_i = \bar{u}_{i+1}) \;\; \text{for arbitrary} \;\; \bar{w}_j$$

Strictly from the definition we have

$$\begin{aligned}
\alpha_{i,1}(j) \;&=\; (-1)(\bar{u}_{i+1}-\bar{u}_i)[\bar{u}_i(1):t]\phi_{j,1}(t) \\
&=\; (-1)\cdot 0\cdot D_t\phi_{j,1}(t)|_{t=\bar{w}_j} \\
&=\; (-1)\cdot 0\cdot \lim_{\substack{\epsilon_j \to 0 \\ \epsilon_j > 0}} \frac{(\bar{w}_j+\epsilon_j-\bar{w}_j)_+^0-(\bar{w}_j-\bar{w}_j)_+^0}{-\epsilon_j}
\end{aligned}$$

(more...)

$$= (-1)\cdot 0\cdot \lim_{\substack{\epsilon_j \to 0 \\ \epsilon_j > 0}} \frac{(\epsilon_j)_+^0 - (0)_+^0}{-\epsilon_j}$$

$$= (-1)\cdot 0\cdot \lim_{\substack{\epsilon_j \to 0 \\ \epsilon_j > 0}} \frac{1-1}{-\epsilon_j} = 0$$

(It is for precisely this case of multiple $\overline{u}$ knots that the definition of $\phi_{j,k}(t)$ includes $\epsilon_j$ — it makes the differentiation with respect to $t$ legal.)

Since the "otherwise" part of the recurrence also specifies $\alpha_{i,1}(j)$ to be zero, and since we have covered all cases, we see that the formula given by the definition

$$\alpha_{i,1}(j) = (-1)(\overline{u}_{i+1}-\overline{u}_i)[\overline{u}_i(1):t]\phi_{j,1}(t)$$

and the alternative formula proposed for the recurrence

$$\alpha_{i,1}(j) = \begin{cases} 1 & \overline{u}_i \le \overline{w}_j < \overline{u}_{i+j} \text{ and } \overline{u}_i < \overline{u}_{i+1} \\ 0 & \text{otherwise} \end{cases} .$$

are consistent with each other. This establishes the validity of the recurrence for $k=1$.

$k>1$:
Consider first the case in which $\overline{u}_i < \overline{u}_{i+r}$.

Notice that

$$\alpha_{i,r}(j) = (-1)^r(\overline{u}_{i+r}-\overline{u}_i)[\overline{u}_i(r):t]\phi_{j,r}(t)$$

But

$$\begin{aligned}\phi_{j,r}(t) &= (\overline{w}_j-t)_+^0(\overline{w}_{j+1}-t)\cdots(\overline{w}_{j+r-1}-t) \\ &= (\overline{w}_{j+r-1}-t)\phi_{j,r-1}(t)\end{aligned}$$

that is, $\phi_{j,r}(t)$ can be viewed as a product. Consequently,

$$\begin{aligned}&[\overline{u}_i(r):t]\phi_{j,r}(t) \\ &\quad = \sum_{s=0}^{r}[\overline{u}_i(s):t](\overline{w}_{j+r-1}-t)\cdot[\overline{u}_{i+s}(r-s):t]\phi_{j,r-1}(t) \quad ,\end{aligned}$$

using the Leibniz rule.

But we discover that

(more...)

$$[\bar{u}_i(0):t](\bar{w}_{j+r-1}-t) = (\bar{w}_{j+r-1}-\bar{u}_i)$$
$$[\bar{u}_i(1):t](\bar{w}_{j+r-1}-t) = -1$$

and

$$[\bar{u}_i(s):t](\bar{w}_{j+r-1}-t) = 0$$

for $s > 1$.

This produces

$$[\bar{u}_i(r):t]\phi_{j,r}(t)$$
$$= (\bar{w}_{j+r-1}-\bar{u}_i)[\bar{u}_i(r):t]\phi_{j,r-1}(t) - [\bar{u}_{i+1}(r-1):t]\phi_{j,r-1}(t)$$

Now we invoke the recursive definition of the divided difference on the term

$$[\bar{u}_i(r):t]\phi_{j,r-1}(t)$$

and this will allow us to write

$$\alpha_{i,r}(j) = (-1)^r(\bar{u}_{i+r}-\bar{u}_i)[\bar{u}_i(r):t]\phi_{j,r}(t)$$
$$= (-1)^r\{(\bar{w}_{j+r-1}-\bar{u}_i)[A-B]-(\bar{u}_{i+r}-\bar{u}_i)C\}$$

where

$$A = [\bar{u}_{i+1}(r-1):t]\phi_{j,r-1}(t)$$
$$B = [\bar{u}_i(r-1):t]\phi_{j,r-1}(t)$$

and

$$C = [\bar{u}_{i+1}(r-1):t]\phi_{j,r-1}(t) \ .$$

These terms may be multiplied out and collected to yield

$$(\bar{w}_{j+r-1}-\bar{u}_i)(-1)^{r-1}[\bar{u}_i(r-1):t]\phi_{j,r-1}(t)$$
$$+ (\bar{u}_{i+r}-\bar{w}_{j+r-1})(-1)^{r-1}[\bar{u}_{i+1}(r-1):t]\phi_{j,r-1}(t) \ ,$$

which yields the relationship to be shown.

In the case where $\bar{u}_i = \bar{u}_{i+r}$, notice that

$$\alpha_{i,r}(j) = (-1)^r(\bar{u}_{i+r}-\bar{u}_i)[\bar{u}_i(r):t]\phi_{j,r}(t) = 0$$

according to the definition and the properties of the divided difference, whereas the recurrence formula becomes

(more...)

$$\alpha_{i,r}(j) = \frac{\overline{w}_{j+r-1}-\overline{u}_i}{\overline{u}_{i+r-1}-\overline{u}_i}\,\alpha_{i,r-1}(j) + \frac{\overline{u}_{i+r}-\overline{w}_{j+r-1}}{\overline{u}_{i+r}-\overline{u}_{i+1}}\,\alpha_{i+1,r-1}(j) \ ,$$

and both terms on the right are zero by convention, since $\overline{u}_i = \overline{u}_{i+1} = \cdots = \overline{u}_{i+r-1} = \overline{u}_{i+r}$.

(We have presented this argument somewhat tersely, since it is a very close parallel to the one which established the recurrence for the $B_{i,k}(\overline{u})$.)

## 15.6. B-Splines and Least-Squares Fitting

Often it is most natural to begin constructing a curve by simply sketching a rough approximation of it, perhaps with tablet and stylus or puck. The tablet is periodically sampled to obtain tens or hundreds of data points representing the curve. (See Figure 188 for an example.) This data is generally a bit noisy, both because of electronic glitches in the puck and because the user's hand motions are jittery. Hence we would like to approximate the data by a piecewise polynomial curve having a relatively small number of segments, the exact number depending on the complexity of the curve.

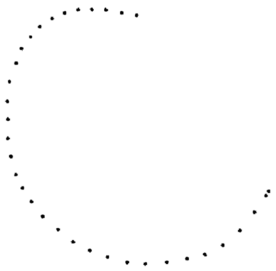


Figure 188. The data for a simple Benesh movement line, generated by a tablet and puck. This is actually filtered data; sample points were thrown away if they were less than three pixels in $x$ or $y$ from the previously accepted data point.

What we shall do is perform a *least squares fit* of the data by a B-spline curve, which can then be manipulated in the usual way to fine-tune its shape. Our treatment follows that of [Forsythe77]; our examples are movement lines generated by an interactive editor for Benesh Dance Notation [Dransch85]. A more general treatment of least squares approximations by parametric cubic splines is given by [Plass83].

Suppose that we are given $p+1$ points $\mathbf{P}_i = (x_i, y_i)$. We want to find a set of $n+1$ control vertices $\mathbf{V}_j$ that minimize the distance between the cubic B-spline curve they define and the data points. If we use enough control vertices (namely $p+1$) we can arrange to actually interpolate the data points; instead we select a smaller value of $n$, yielding $n-2$ segments which "adequately" represent the curve. We assume that $n$ is given.

For the sake of efficiency we will make some simplifying assumptions whose legitimacy is discussed later. Recall that

$$\mathbf{Q}(\overline{u}) = (\, X(\overline{u}), Y(\overline{u}) \,)$$

$$= \sum_{j=0}^{n} \mathbf{V}_j B_{j,4}(\bar{u})$$

$$= \sum_{j=0}^{n} \left( X_j B_{j,4}(\bar{u}), Y_j B_{j,4}(\bar{u}) \right) ,$$

where the position of the $j^{\text{th}}$ control vertex is represented by $(X_j,Y_j)$ so as to distinguish them from the data points $(x_i,y_i)$ we are fitting. What we shall actually minimize is the expression

$$\sum_{i=0}^{p} |\mathbf{Q}(\bar{u}_i) - \mathbf{P}_i|^2 = \sum_{i=0}^{n} \Big( (X(\bar{u}_i)-x_i)^2 + (Y(\bar{u}_i)-y_i)^2 \Big) = R , \tag{124}$$

where $u_i$ is a parameter value associated with the $i^{\text{th}}$ data point. Since equation (124) is quadratic its minimum occurs for those values of $X_j$ and $Y_j$ such that

$$\frac{\partial}{\partial X_k} R = 0$$

$$\frac{\partial}{\partial Y_k} R = 0 ,$$

where $k$ ranges between 0 and $n$. As usual we will consider just the $Y_j$, the $X_j$ being treated analogously.

If we compute a typical such partial derivative we obtain

$$\frac{\partial}{\partial Y_k} R = \sum_{j=0}^{n} \left( \sum_{i=0}^{p} B_{j,4}(\bar{u}_i)\, B_{k,4}(\bar{u}_i) \right) Y_j = \sum_{i=0}^{p} y_i B_{k,4}(\bar{u}_i) .$$

(Note that terms involving $X_j$ and $x_i$ disappear.) If we do this for each $Y_k$, we have a set of $n+1$ simultaneous linear equations in $n+1$ unknowns, which can be solved by the usual techniques. (Although [Forsythe77] warn that in general this system of equations is prone to numerical error, in fact our particular formulation is safe because we are using B-splines rather than the power functions $\bar{u}^j$.)

We have still to indicate how the $u_i$ are associated with the data points. We let

$$S = \sum_{i=1}^{p} |\mathbf{P}_i - \mathbf{P}_{i-1}|$$

($S$ is thus the total length of the line segments connecting the data points), and then set

$$u_0 = 3$$

$$\bar{u}_{i+1} = \bar{u}_i + (n-2)\frac{|\mathbf{P}_{i+1} - \mathbf{P}_i|}{S} .$$

As a result the spacing between the $\bar{u}_i$ is proportional to the Euclidean distance between their associated data points. This does not, of course, ensure that $\mathbf{Q}(\bar{u}_i)$ is the point at which the curve is closest to the $i^{\text{th}}$ data point, but in practise it produces better results than uniform spacing.

For the application discussed in [Dransch85] it was important that the first and last data points be interpolated, and that the user have explicit control over the endpoints of the curve when manipulating control vertices. Hence the initial and final knots (3 and $n+1$) were given multiplicity 4, $Y_0$ was given the value $y_0$, and $Y_n$ was given the value $y_p$. This leaves $n-1$ equations

$$\sum_{j=1}^{n-1}\left(\sum_{i=0}^{p}B_{j,4}(\bar{u}_i)\,B_{k,4}(\bar{u}_i)\right)Y_j$$

$$= \sum_{i=0}^{p}y_iB_{k,4}(\bar{u}_i) - \sum_{i=0}^{p}B_{0,4}(\bar{u}_i)\,B_{k,4}(\bar{u}_i)\,Y_0 - \sum_{i=0}^{p}B_{n,4}(\bar{u}_i)\,B_{k,4}(\bar{u}_i)\,Y_n \ .$$

$$= \sum_{i=0}^{p}y_iB_{k,4}(\bar{u}_i) - \sum_{i=0}^{p}B_{0,4}(\bar{u}_i)\,B_{k,4}(\bar{u}_i)\,y_0 - \sum_{i=0}^{p}B_{n,4}(\bar{u}_i)\,B_{k,4}(\bar{u}_i)\,y_n \ .$$

in the $n-1$ unknowns $Y_1$ through $Y_{n-1}$.

Figures 189, 190 and 191 show one, three and five segment cubic B-spline curves fit to the data of Figure 188 using this technique.

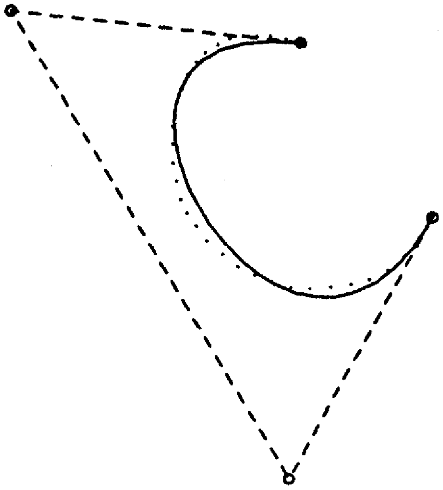


Figure 189. A one segment cubic B-spline curve, fit to the 34 data points shown with dots. Multiplicity 4 knots are used at either end to force interpolation of the ending data points; otherwise the parametric spacing is proportional to the Euclidean spacing between data points. The control polygon is shown with a dashed line to avoid confusion with the data points.

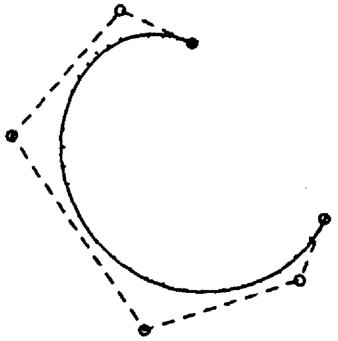


Figure 190. A three segment cubic B-spline curve, fit to the data of Figure 188. Compare the resulting curve to that of Figure 189.

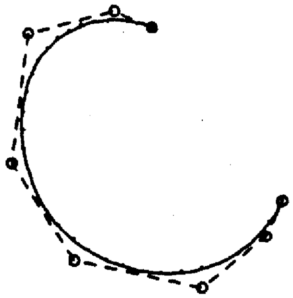


Figure 191. A five segment cubic B-spline curve, fit to the data of Figure 188. Compare the resulting curve to those of Figures 189 and 190.

The curves of Figures 193-195 show the results of fitting the more complicated data shown in Figure 192.

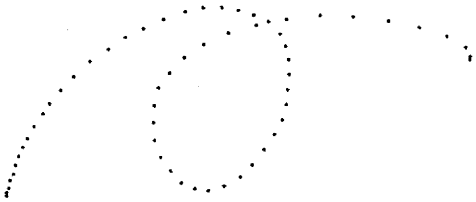


Figure 192. The data for a more complex Benesh movement line. This is also filtered data. Even so there is a marked change in the spacing between data points as the user's hand changes speed at the ends of the curve. There are 60 data points.

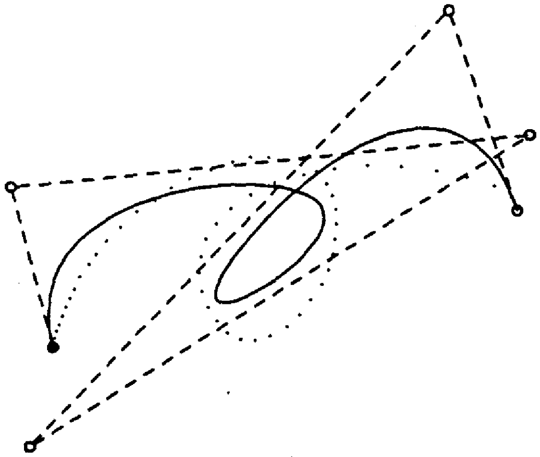


Figure 193. A three segment cubic B-spline curve, fit to the data of Figure 192.

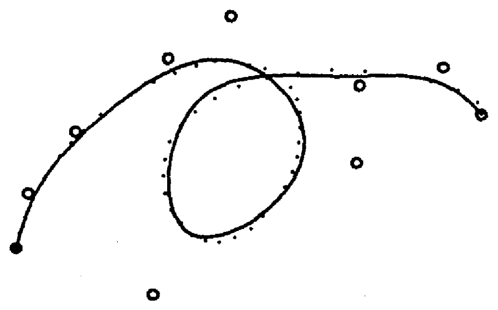


Figure 194. A seven segment cubic B-spline curve, fit to the data of Figure 192. The control polygon has been omitted to avoid clutter.

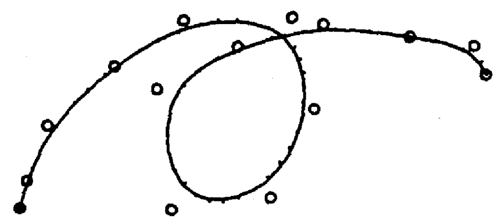


Figure 195. A 12 segment cubic B-spline curve, fit to the data of Figure 192. We are well past the point of needing more segments, since we have begun to mimic jitter in the user's hand. A curve of fewer segments should be used and reshaped slightly by moving a few of the control vertices.

The quality of the fit, as measured by the sum-square of the residuals $R$, can generally be improved by intelligently selecting where the joints between successive segments occur, and by adjusting the knot values $\overline{u}_i$ associated with each data point $\mathbf{P}_i$. Techniques for doing so are discussed in [Plass83]. Doing so is, of course, more expensive. In an interactive environment it may well be preferable to rapidly compute a somewhat inferior fit and adjust it by manipulating control vertices.

There are a variety of means other than the use of multiple knots for obtaining interpolation of the endpoints, as discussed in 4.5. Also, the initial fit by an open curve will generally be quite close to the data points at either end, and for some applications that may be sufficient.

# 16. Interlude

Our next objective is to explore an alternative generalization of the uniform cubic B-splines called the *Beta-splines*. They are motivated by the fact that parametrizing curves can result in various non-intuitive effects; we give some examples in the next section. If we focus on the physical notions of derivative continuity in which we are really interested, namely continuity of direction and curvature, instead of continuity of first and second parametric derivatives, then a class of curves results which is expressed in terms of two *shape parameters* $\beta_1$ and $\beta_2$.

A particularly novel and useful aspect of these shape parameters is that they can be used to locally control *tension* in a piecewise polynomial curve. That is, they can be manipulated so as to pull a curve flat against the control polygon without altering distant portions of the curve in any way.

Our initial development of the Beta-splines will be based on the presentation of uniform cubic B-splines given in Section . In particular, they will be expressed in terms of a uniform knot sequence. Following this we will see that there is a simple way of associating distinct values of the shape parameters with each knot, and interpolating between them as we move along the curve in such a way as to achieve local control of the shape parameters. These are the *continuously-shaped Beta-splines*. Finally, we will present a different generalization of the Beta-splines analogous to the divided difference definition of the general B-splines. For these *discretely-shaped Beta-splines* the shape parameters can be thought of as an intuitive and useful way of controlling the parametric discontinuities which result from multiple knots in cubic B-spline curves.

# 17. Parametric vs. Geometric Continuity

Our objective in this section is to show that the relationship between the continuity of parametric derivatives and the physically meaningful notions of "unit tangent vector continuity" and "curvature vector continuity" is subtle. This will motivate a generalization of the uniform cubic B-splines, called the uniformly-shaped cubic Beta-splines, which we shall introduce in the next section. Most of what follows may be found in [Barsky81, Barsky85].

## 17.1. Geometric Continuity

Intuitively a first derivative vector tells us the direction in which a curve is headed and, by virtue of its length, the speed with which we are moving in that direction. Geometrically, two curves are joined smoothly with respect to their first derivatives if their respective derivative vectors at the joint point in the same direction. Hence we extract the direction from a parametric first derivative by normalizing its length: the *unit tangent vector* of a curve $\mathbf{Q}(\bar{u})$ is

$$\hat{\mathbf{T}}(\bar{u}) = \frac{\mathbf{Q}^{(1)}(\bar{u})}{|\mathbf{Q}^{(1)}(\bar{u})|} \quad . \tag{125}$$

As we have discussed in Section 4.5.1, the *curvature vector* is

$$\mathbf{K}(\bar{u}) = \kappa(\bar{u})\hat{\mathbf{N}}(\bar{u}) = \kappa(\bar{u})\frac{\hat{\mathbf{T}}^{(1)}(\bar{u})}{|\hat{\mathbf{T}}^{(1)}(\bar{u})|}$$

where $\kappa(\bar{u})$ is the *curvature* of $\mathbf{Q}(\bar{u})$ at $u$ and $\hat{\mathbf{N}}(\bar{u})$ is the unit vector pointing from $\mathbf{Q}(\bar{u})$ towards the centre of the osculating circle at $\mathbf{Q}(\bar{u})$. $\mathbf{K}(\bar{u})$ records the extent to which the curve is "bent" away from a straight line; its direction tells us how that bending is oriented with respect to the direction in which the curve is headed.

It is shown in [Barsky81, Barsky85, Barsky85a] that

$$\mathbf{K}(\bar{u}) = \frac{\mathbf{Q}^{(1)}(\bar{u})\times\mathbf{Q}^{(2)}(\bar{u})\times\mathbf{Q}^{(1)}(\bar{u})}{|\mathbf{Q}^{(1)}(\bar{u})|^4} \quad . \tag{126}$$

We shall say that a curve whose unit tangent vector and curvature vector are everywhere continuous has $G^2$ or *second degree geometric continuity.*

Next we shall see that we can rig the way in which curve segments are parametrized so as to cause the parametric first and second derivative vectors to incorrectly represent our intuition as to the physical continuity of a curve. Many curves are $G^2$ continuous but not $C^2$ continuous, and $C^2$ continuous curves can fail to appear geometrically continuous. A more rigorous and general discussion of this material can be found in [Barsky84].

## 17.2. First Derivative Continuity

First let us see that a first derivative may be continuous even though the curve itself has a discontinuous tangent. We have already seen an illustration of this in Figure 32, but we can give an even simpler example. The idea is to arrange for the first derivative vector to be (0,0) at the point in question, so that the unit tangent vector is discontinuous even though the first derivative is continuous. In such a case we may easily arrange that for $\epsilon > 0$, the limit from the left

$$\lim_{\epsilon \to 0^+} \frac{\mathbf{Q}^{(1)}(\bar{u}-\epsilon)}{|\mathbf{Q}^{(1)}(\bar{u}-\epsilon)|}$$

and the limit from the right

$$\lim_{\epsilon \to 0^+} \frac{\mathbf{Q}^{(1)}(\bar{u}+\epsilon)}{|\mathbf{Q}^{(1)}(\bar{u}+\epsilon)|}$$

be distinct.

Consider the two line segments $\mathbf{Q}_1(u)$ and $\mathbf{Q}_2(u)$ defined by

$$\mathbf{Q}_1(u) = (\, 2u - u^2, 2u - u^2 \,) \qquad 0 \le u \le 1$$
$$\mathbf{Q}_2(u) = (\, 1 + u^2, 1 - u^2 \,) \qquad 0 \le u \le 1 \;.$$

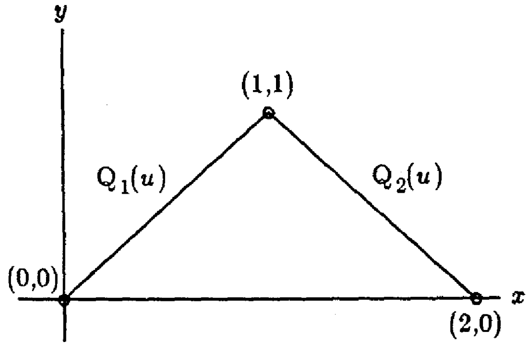


Figure 196. A continuous first derivative with a discontinuous unit tangent vector.

These line segments are positionally continuous since $\mathbf{Q}_1(1) = \mathbf{Q}_2(0) = (1,1)$. Their first derivative vectors are

$$\mathbf{Q}_1^{(1)}(u) = (\, 2-2u, 2-2u \,) = (2-2u)(1,1)$$
$$\mathbf{Q}_2^{(1)}(u) = (\, 2u, -2u \,) \qquad = 2u\,(1,-1) \;.$$

Since $\mathbf{Q}_1^{(1)}(1) = \mathbf{Q}_2^{(1)}(0) = (0,0)$, the first derivative vectors are continuous at the joint (as well as being continuous elsewhere), even though the unit tangent vectors for $\mathbf{Q}_1(u)$ and $\mathbf{Q}_2(u)$, given by

$$\hat{\mathbf{T}}_1(1) = \frac{1}{\sqrt{2}}(1,1)$$

$$\hat{\mathbf{T}}_2(0) = \frac{1}{\sqrt{2}}(1,-1) \ ,$$

clearly point in different directions are therefore not continuous.

It is also quite possible for the first derivative vector to be discontinuous even though the curve possesses a physically continuous unit tangent vector throughout its length. Consider

$$\mathbf{Q}_1(u) = (\ 12u, 9u\ ) \qquad 0 \le u \le 1$$

$$\mathbf{Q}_2(u) = (\ 4(u+3), 3(u+3)\ ) \qquad 0 \le u \le 1$$

whose first derivative vectors are

$$\mathbf{Q}_1^{(1)}(u) = (\ 12, 9\ )$$

$$\mathbf{Q}_2^{(1)}(u) = (\ 4, 3\ ) \ .$$

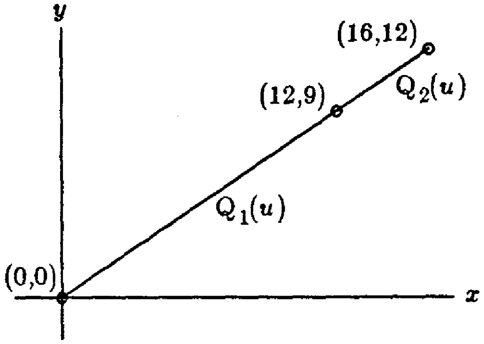


Figure 197. A discontinuous first derivative with a continuous unit tangent vector.

These line segments are collinear, and have a continuous unit tangent vector (namely $(\frac{4}{5},\frac{3}{5})$) even though there is a jump in the first derivative vector at the joint.

## 17.3. Second Derivative Continuity

We can find instances of the same sort of phenomena for the second parametric derivative as well. First we show that the existence of a continuous second derivative vector need not ensure that the curvature vector is continuous. Consider

$$\mathbf{Q}_1(u) = \left( cos(\frac{\pi}{2}(1-u)^3)\ ,\ sin(\frac{\pi}{2}(1-u)^3) \right) \qquad 0 \le u \le 1$$

$$\mathbf{Q}_2(u) = \left( 3-2cos(\frac{\pi}{2}u^3)\ ,\ -2sin(\frac{\pi}{2}u^3) \right) \qquad 0 \le u \le 1$$

which define two circles of radius one and two centred at (0,0) and at (3,0), respectively, which meet at (1,0). Because they have different radii, there is a change in the curvature where they meet, and consequently a jump in the curvature vector.

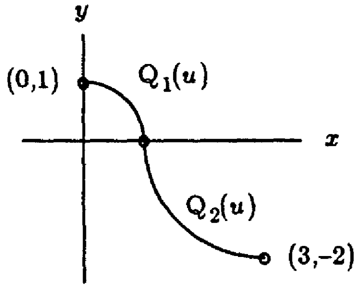


Figure 198. A continuous second derivative with a discontinuous curvature vector.

On the other hand, their first derivative vectors are

$$Q_1^{(1)}(u) = \frac{1}{2}\left( 3\pi sin(\frac{\pi(1-u)^3}{2})(1-u)^2 \, , \; -3\pi cos(\frac{\pi(1-u)^3}{2})(1-u)^2 \right)$$

$$Q_2^{(1)}(u) = \left( 3\pi u^2 sin(\frac{\pi u^3}{2}) \, , \; -3\pi u^2 cos(\frac{\pi u^3}{2}) \right)$$

and their second derivative vectors are

$$Q_1^{(2)}(u) = \frac{1}{4}\left( -9\pi^2 cos(\frac{\pi(1-u)^3}{2})(1-u)^4 - 12\pi sin(\frac{\pi(1-u)^3}{2})(1-u) \, , \right.$$
$$\left. 12\pi cos(\frac{\pi(1-u)^3}{2})(1-u) - 9\pi^2 \sin(\frac{\pi(1-u)^3}{2})(1-u)^4 \right)$$

$$Q_2^{(2)}(u) = \frac{1}{2}\left( 12\pi u \, sin(\frac{\pi u^3}{2}) + 9\pi^2 u^4 cos(\frac{\pi u^3}{2}) \, , \right.$$
$$\left. 9\pi^2 u^4 sin(\frac{\pi u^3}{2}) - 12\pi u \, cos(\frac{\pi u^3}{2}) \right) .$$

In particular,

$$Q_1^{(2)}(1) = (\, 0 \, , 0 \,)$$
$$Q_2^{(2)}(0) = (\, 0 \, , 0 \,)$$

so that the second derivative vectors for the two curve segments are continuous at their common joint, even though the curvature vector has a jump there both in direction and in magnitude since

$$K_1(1) = (\, -1.0 \, , 0 \,)$$
$$K_2(0) = (\, +0.5 \, , 0 \,) \; .$$

It is also possible for the curvature vector to be continuous even if the second derivative vector is not. Consider the following two curve segments, which define successive portions of a circle of radius 1 centered at the origin (so that the curvature vector must be continuous).

$$Q_1(u) = \left( sin(\frac{\pi}{2}u^2) \, , \; cos(\frac{\pi}{2}u^2) \right) \qquad 0 \leq u \leq 1$$

$$Q_2(u) = \left[ \cos(\frac{\pi}{2}u^2) \,,\; -\sin(\frac{\pi}{2}u^2) \right] \qquad 0 \le u \le 1 \;.$$

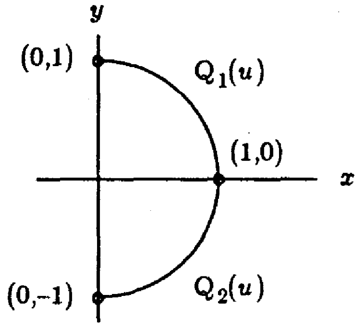


Figure 199. Discontinuous second derivatives with a continuous curvature vector.

The first derivative vectors are

$$Q_1^{(1)}(u) = \left[ \pi u \cos(\frac{\pi u^2}{2}) \,,\; -\pi u \sin(\frac{\pi u^2}{2}) \right]$$

$$Q_2^{(1)}(u) = \left[ -\pi u \sin(\frac{\pi u^2}{2}) \,,\; -\pi u \cos(\frac{\pi u^2}{2}) \right]$$

and the second derivative vectors are

$$Q_1^{(2)}(u) = \left[ \pi \cos(\frac{\pi u^2}{2}) - \pi^2 u^2 \sin(\frac{\pi u^2}{2}) \,, \right.$$
$$\left. -\pi \sin(\frac{\pi u^2}{2}) - \pi^2 u^2 \cos(\frac{\pi u^2}{2}) \right]$$

$$Q_2^{(2)}(u) = \left[ -\pi \sin(\frac{\pi u^2}{2}) - \pi^2 u^2 \cos(\frac{\pi u^2}{2}) \,, \right.$$
$$\left. \pi^2 u^2 \sin(\frac{\pi u^2}{2}) - \pi \cos(\frac{\pi u^2}{2}) \right] .$$

In particular, the second derivative vectors at the joint between the two segments are

$$Q_1^{(2)}(1) = ( -\pi^2 \,, -\pi )$$

$$Q_2^{(2)}(0) = ( 0 \,, -\pi )$$

although the curvature vector is clearly continuous since the two curves together are simply a half circle of radius one centred at the origin. Indeed, the reader may care to verify that the less exotic parametric representation of such a semicircle as

$$\mathbf{Q}_1(u) = \left[ \, sin(\frac{\pi}{2}u) \, , \; cos(\frac{\pi}{2}u) \, \right]$$

$$\mathbf{Q}_2(u) = \left[ \, cos(\frac{\pi}{2}u) \, , \; - sin(\frac{\pi}{2}u) \, \right]$$

is $C^2$ continuous since

$$\mathbf{K}_1(1) = \mathbf{K}_2(0) = ( -1 \, , 0 ) \; .$$

# 18. Uniformly-Shaped Beta-splines

We have seen that $\hat{\mathbf{T}}(\bar{u})$ and $\mathbf{K}(\bar{u})$ capture the physically meaningful notions of direction and curvature. The parametric first and second derivative vectors, on the other hand, may be changed by reparametrization without altering the curve, and moreover their continuity may not reflect the actual "physical continuity" of the curve. Hence it is reasonable to ask if one can define curves in which geometric rather than parametric continuity is required. We will do so in this chapter, drawing upon material which appears in [Barsky81, Barsky85, Barsky85a]. We shall see that in doing so we can gain additional control over the shape of the cubic piecewise polynomial curves in which we are interested.

Of course, $\mathbf{Q}(u)$, $\hat{\mathbf{T}}(u)$ and $\mathbf{K}(u)$ are easily seen to be continuous away from the joints of a piecewise polynomial; what we want is to develop a means of enforcing positional, unit tangent and curvature vector continuity *at* the joint between two successive curve segments as well. Our approach is very much analogous to the way in which we previously derived the uniform cubic B-splines.

Obtaining positional continuity is easy. We have simply to require that

$$\mathbf{Q}_{i-1}(1) = \mathbf{Q}_i(0) \quad . \tag{127}$$

Next we observe that two curves will have the same unit tangent vector at their common joint if their first derivative vectors are collinear and have the same sense, which is to say if one is a positive multiple of the other. The following equation captures this notion easily.

$$\beta_1 \mathbf{Q}_{i-1}^{(1)}(1) = \mathbf{Q}_i^{(1)}(0) \qquad \beta_1 > 0 \tag{128}$$

There is an instantaneous change in velocity at the joint, but not a change in direction.

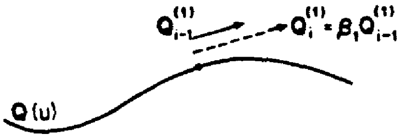


Figure 200. The basic idea behind the notion of $G^1$ continuity is that the tangent vectors at the joint between two successive segments need only be collinear; their magnitudes may differ by a positive factor which we are calling $\beta_1$.

Obtaining continuity of the curvature vector is somewhat more involved. Equation (126) gives us a way of computing the curvature vector at an arbitrary point. Let us use it to obtain the curvature of two consecutive segments $\mathbf{Q}_{i-1}(u)$ and $\mathbf{Q}_i(u)$ at their common joint and equate the two expressions.

$$\frac{\mathbf{Q}_{i-1}^{(1)}(1)\times\mathbf{Q}_{i-1}^{(2)}(1)\times\mathbf{Q}_{i-1}^{(1)}(1)}{|\mathbf{Q}_{i-1}^{(1)}(1)|^4} = \frac{\mathbf{Q}_i^{(1)}(0)\times\mathbf{Q}_i^{(2)}(0)\times\mathbf{Q}_i^{(1)}(0)}{|\mathbf{Q}_i^{(1)}(0)|^4}$$

If we substitute for $\mathbf{Q}_i^{(1)}(0)$ using equation (128) this becomes

$$\frac{\mathbf{Q}_{i-1}^{(1)}(1)\times\mathbf{Q}_{i-1}^{(2)}(1)\times\mathbf{Q}_{i-1}^{(1)}(1)}{|\mathbf{Q}_{i-1}^{(1)}(1)|^4} = \frac{\beta_1\mathbf{Q}_{i-1}^{(1)}(1)\times\mathbf{Q}_i^{(2)}(0)\times\beta_1\mathbf{Q}_{i-1}^{(1)}(1)}{|\beta_1\mathbf{Q}_{i-1}^{(1)}(1)|^4}$$

$$= \frac{\mathbf{Q}_{i-1}^{(1)}(1)\times\dfrac{\mathbf{Q}_i^{(2)}(0)}{\beta_1^2}\times\mathbf{Q}_{i-1}^{(1)}(1)}{|\mathbf{Q}_{i-1}^{(1)}(1)|^4} .$$

Clearly equality is ensured if $\beta_1^2\mathbf{Q}_{i-1}^{(2)}(1) = \mathbf{Q}_i^{(2)}(0)$. However, since the cross product of a vector with itself is zero, $\mathbf{Q}_i^{(2)}(0)$ may have an additional component along $\mathbf{Q}_{i-1}^{(1)}(1)$. Hence equality still results if, for any real numbers $\beta_1$ and $\beta_2$,

$$\beta_1^2\mathbf{Q}_{i-1}^{(2)}(1) + \beta_2\mathbf{Q}_{i-1}^{(1)}(1) = \mathbf{Q}_i^{(2)}(0) \qquad \beta_1 > 0 . \tag{129}$$

Equation (129) has a natural physical interpretation: $\mathbf{Q}_i^{(2)}(0)$ may have an additional component directed along the tangent since acceleration along the tangent does not "deflect" a point traveling along the curve, and so does not affect the curvature there.

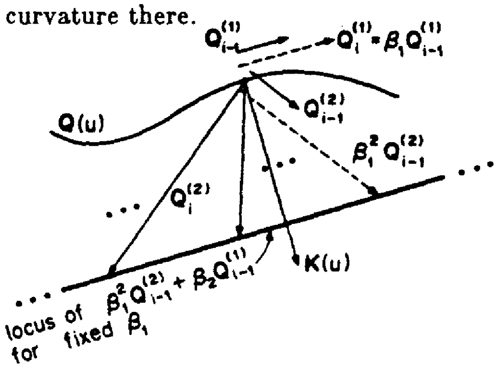

$Q_{i-1}^{(1)}$
$Q_i^{(1)} = \beta_1 Q_{i-1}^{(1)}$
$Q(u)$
$Q_{i-1}^{(2)}$
$\beta_1^2 Q_{i-1}^{(2)}$
$Q_i^{(2)}$
$K(u)$
$\beta_1^2 Q_{i-1}^{(2)} + \beta_2 Q_{i-1}^{(1)}$
locus of for fixed $\beta_1$
...

Figure 201. The left and right second parametric derivatives at the joint between two segments may differ by an arbitrary component along their common unit tangent since applying force along the direction of motion does not deflect a moving point.

Geometric continuity results, then, if equations (127), (128) and (129) hold at every knot $u_i$, for any positive $\beta_1$ and for any $\beta_2$. These equations are, by definition, less restrictive than simple continuity of position and parametric derivatives, which is the special case in which $\beta_1 = 1$ and $\beta_2 = 0$, yielding the uniform cubic B-spline curves.

If $\mathbf{Q}(\bar{u})$ is defined using cubic splines then these equations must hold for the basis functions as well since

$$\mathbf{Q}(\bar{u}) = \sum_i \mathbf{V}_i B_i(\bar{u})$$

$$\mathbf{Q}^{(1)}(\bar{u}) = \sum_i \mathbf{V}_i B_i^{(1)}(\bar{u})$$

and any of the control vertices may be $\mathbf{0}$. Conversely, if the basis functions satisfy these equations, as for example

$$\beta_1\, b_{-0}^{(1)}(\bar{u}_i) = b_{-1}^{(1)}(\bar{u}_i)$$

then so will the curves they define. We shall now arrange that this be so.

Let us again consider a basis function composed of four cubic polynomial basis segments, so that

$$\mathbf{Q}_i(\bar{u}) = \mathbf{V}_{i-3}B_{i-3}(\bar{u}) + \mathbf{V}_{i-2}B_{i-2}(\bar{u}) + \mathbf{V}_{i-1}B_{i-1}(\bar{u}) + \mathbf{V}_{i-0}B_{i-0}(\bar{u})$$

(just as for the uniform cubic B-splines in Chapter 4), but this time ask that they satisfy the geometric constraints (127), (128) and (129) instead of the parametric constraints (6), (7) and (8). The equations which result are

$$\begin{array}{rclrcl}
0 &=& b_{-0}(0) & 0 &=& b^{(1)}_{-0}(0) \\
b_{-0}(1) &=& b_{-1}(0) & \beta_1\, b^{(1)}_{-0}(1) &=& b^{(1)}_{-1}(0) \\
b_{-1}(1) &=& b_{-2}(0) & \beta_1\, b^{(1)}_{-1}(1) &=& b^{(1)}_{-2}(0) \\
b_{-2}(1) &=& b_{-3}(0) & \beta_1\, b^{(1)}_{-2}(1) &=& b^{(1)}_{-3}(0) \\
b_{-3}(1) &=& 0 & \beta_1\, b^{(1)}_{-3}(1) &=& 0 \\
 & & & 0 &=& b^{(2)}_{-0}(0) \\
 & & & \beta_1^2\, b^{(2)}_{-0}(1) + \beta_2\, b^{(1)}_{-0}(1) &=& b^{(2)}_{-1}(0) \\
 & & & \beta_1^2\, b^{(2)}_{-1}(1) + \beta_2\, b^{(1)}_{-1}(1) &=& b^{(2)}_{-2}(0) \\
 & & & \beta_1^2\, b^{(2)}_{-2}(1) + \beta_2\, b^{(1)}_{-2}(1) &=& b^{(2)}_{-3}(0) \\
 & & & \beta_1^2\, b^{(2)}_{-3}(1) + \beta_2\, b^{(1)}_{-3}(1) &=& 0
\end{array} \qquad (130)$$

To obtain sixteen equations we again require, in order (we hope) to obtain the convex hull property, that

$$\begin{aligned}
\sum_{r=-3}^{0} B_{i+r}(\bar{u}) &= b_{-0}(0) + b_{-1}(0) + b_{-2}(0) + b_{-3}(0) \\
&= b_{-1}(0) + b_{-2}(0) + b_{-3}(0) = 1 \ ,
\end{aligned}$$

yielding a total of sixteen equations in sixteen unknowns. For any particular values of $\beta_1$ and $\beta_2$ these equations can be solved numerically (as in the B-spline case) to obtain explicit formulae for the polynomials comprising the basis segments. This is not very practical, however, since we do not want to solve a new system every time we wish to alter one of the $\beta$ parameters. Instead we can solve this system symbolically, using a symbolic manipulation system such as Vaxima [Bogen77, Fateman82], to obtain the following symbolic representation of the basis segments for all values of $\beta_1$ and $\beta_2$.

$$\begin{aligned}
b_{-0}(u) &= \frac{1}{\delta}\Big(\, 2u^3 \,\Big) \qquad (131) \\
b_{-1}(u) &= \frac{1}{\delta}\Big(\, 2 + (6\beta_1)u + (3\beta_2+6\beta_1^2)u^2 - (2\beta_2+2\beta_1^2+2\beta_1+2)u^3 \,\Big) \\
b_{-2}(u) &= \frac{1}{\delta}\Big(\, (\beta_2+4\beta_1^2+4\beta_1) + (6\beta_1^3-6\beta_1)u \\
&\qquad - (3\beta_2+6\beta_1^3+6\beta_1^2)u^2 + (2\beta_2+2\beta_1^3+2\beta_1^2+2\beta_1)u^3 \,\Big) \\
b_{-3}(u) &= \frac{1}{\delta}\Big(\, (2\beta_1^3) - (6\beta_1^3)u + (6\beta_1^3)u^2 - (2\beta_1^3)u^3 \,\Big)
\end{aligned}$$

where

$$\delta = \beta_2 + 2\beta_1^3 + 4\beta_1^2 + 4\beta_1 + 2 \neq 0 \ .$$

Notice that if we substitute $\beta_1 = 1$ and $\beta_2 = 0$ into the Beta-spline constraint equations (130) we obtain the B-spline constraint equations (9), and that substituting these values into the Beta-spline basis segments (131) we obtain the B-spline basis segments (11). For other values of $\beta 1$ and $\beta 2$ the Beta-spline basis segments fail to be $C^2$ continuous at knots, although they do satisfy equations (130) and are therefore $G^2$ continuous.

Equations (131) can, of course, be evaluated more rapidly if they are factored. For any particular values of $\beta 1$ and $\beta 2$ they are cubic polynomials in $u$, so forward differencing can also be used where appropriate. The efficient evaluation of these equations is discussed in [Barsky81, Barsky85, Barsky85b].

We will refer to the basis functions whose segments are defined by equations (131) as *uniformly-shaped Beta-splines* in order to distinguish them from the more general Beta-splines which will be defined subsequently.

Increasing $\beta 1$ increases the "velocity" with which we traverse a curve immediately after a joint, with respect to the "velocity" just previous to the joint, thus serving to *bias* the curve; values in excess of one cause the unit tangent vector at the joint (which is, of course, continuous) to have greater influence to the right than to the left, in that the curve will "continue in the direction of the tangent" longer in the rightmost segment. Values of $\beta 1$ ranging from one down to zero have the reciprocal effect, causing the curve to lie close to the tangent longer to the left of a joint than to the right.

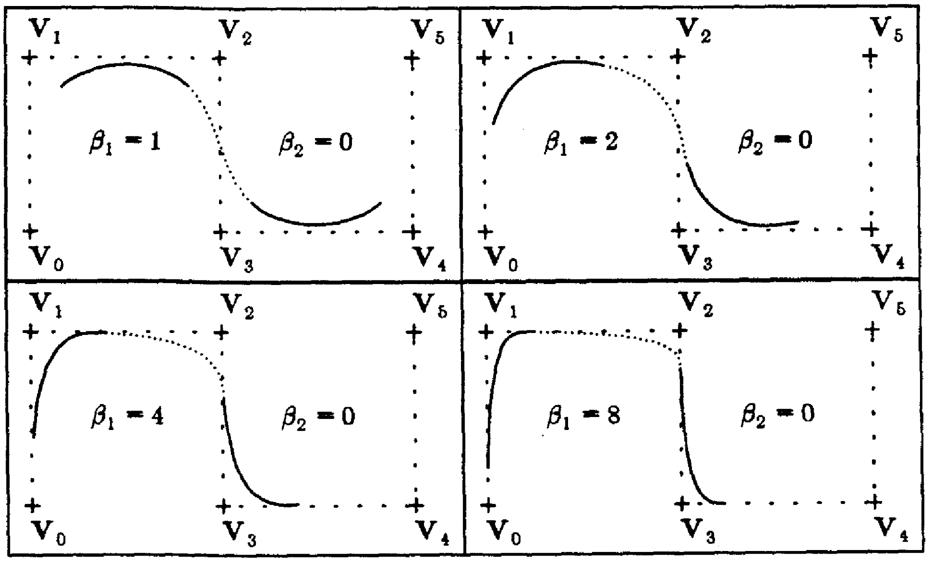


Figure 202. This sequence of curves illustrates the effect of increasing $\beta 1$ on a uniformly-shaped Beta-spline curve.

It is instructive to examine the basis functions used to define these curves:

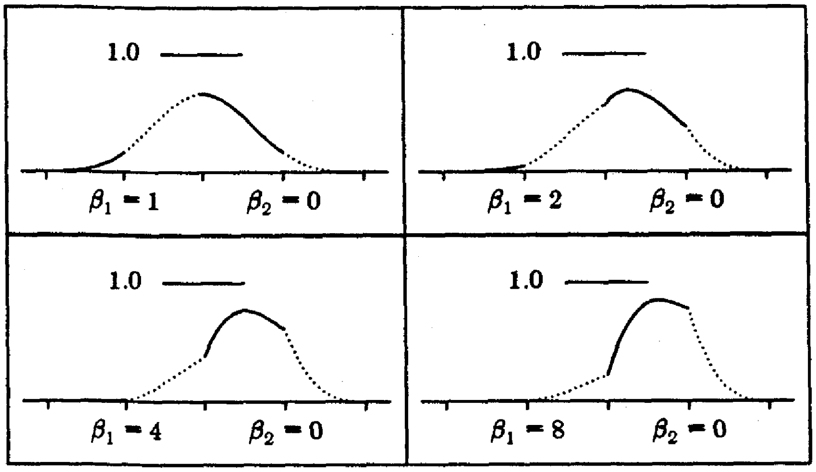


Figure 203. These are the basis functions corresponding to the curves of Figure 202.

Each is computed for a distinct value of $\beta 1$, which determines the ratio of the slopes to the left and right of each joint. Notice that since the same basis function is used for each of $X(\bar{u})$ and $Y(\bar{u})$, any continuous basis function whose first derivative is continuous except for a positive jump of some arbitrary value ($\beta 1$) at the knots suffices to define a curve with unit tangent continuity.

The $\beta 2$ parameter serves to control *tension* in the curve: altering the value of $\beta 2$ moves the joint between $Q_{i-1}(u)$ and $Q_i(u)$ along a vector that passes through the $(i-2)^{nd}$ control vertex, and this happens simultaneously for all the joints in a uniformly-shaped curve. For example, increasingly positive values move each joint towards its corresponding control vertex and flatten the curve against the control polygon.

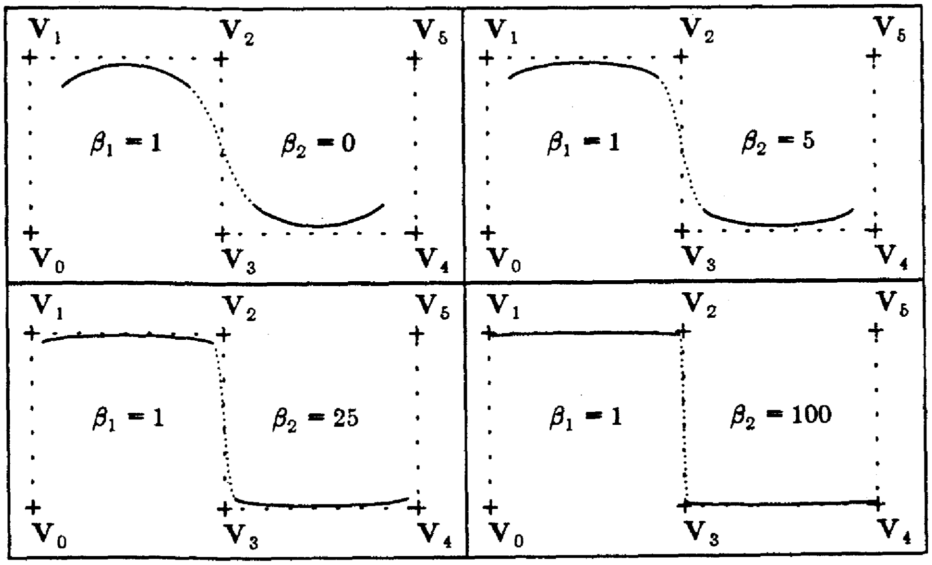


Figure 204. This sequence of curves illustrates the effect of increasing $\beta 2$ on a uniformly-shaped Beta-spline curve.

The corresponding basis functions are:

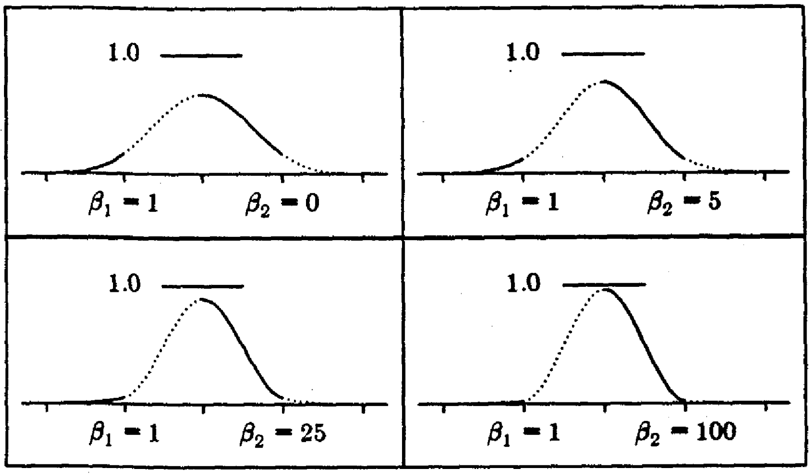


**Figure 205. These are the basis functions corresponding to the curves of Figure 204.**

Notice that as $\beta2$ increases the peak of the basis function approaches one and the "tails" of the basis function, lying in the leftmost and rightmost intervals of its support, approach zero. Since our indexing convention is that the $i^{\text{th}}$ basis function is scaled by $\mathbf{V}_i$ and has support $(\bar{u}_i, \bar{u}_{i+4})$, this peak is at $\bar{u}_{i+2}$. Again by convention this is the joint between $\mathbf{Q}_{i+1}(u)$ and $\mathbf{Q}_{i+2}(u)$.

More generally, the curve itself converges to the control polygon as $\beta2$ goes to infinity, the joints between segments converging to the control vertices. This behaviour is predictable from equations (131). As $\beta2$ is increased, the basis segments converge to

$$
\begin{aligned}
b_{-0}(u) &= 0 \\
b_{-1}(u) &= (3u^2 - 2u^3) \\
b_{-2}(u) &= 1 - (3u^2 - 2u^3) \\
b_{-3}(u) &= 0
\end{aligned}
$$

for any value of $\beta1$. If we let $t = (3u^2 - 2u^3)$, it is easy to see that in the limit we obtain a curve that varies linearly between each successive pair of control vertices.

$\beta1$ also serves, to some extent, as an "asymmetric tension parameter." If for any value of $\beta2$ we allow $\beta1$ to become arbitrarily large then the basis segments converge to

$$
\begin{aligned}
b_{-0}(u) &= 0 \\
b_{-1}(u) &= 0 \\
b_{-2}(u) &= (3u - 3u^2 + u^3) \\
b_{-3}(u) &= 1 - (3u - 3u^2 + u^3) \ .
\end{aligned}
$$

If these are scaled by $\mathbf{V}_i$, $\mathbf{V}_{i-1}$, $\mathbf{V}_{i-2}$ and $\mathbf{V}_{i-3}$, respectively, to define the $i^{\text{th}}$ segment $\mathbf{Q}_i(u)$ then this segment of the curve converges to a straight line between $\mathbf{V}_{i-3}$ and $\mathbf{V}_{i-2}$.

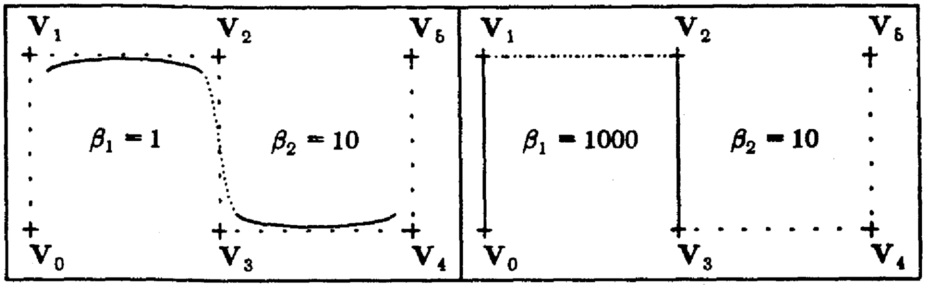


Figure 206. The effect of making $\beta 1$ very large for any value of $\beta 2$.

If $\beta 2$ has the value zero and we allow $\beta 1$ to approach zero then we obtain symmetrical behaviour:

$$
\begin{aligned}
b_{-0}(u) &= u^3 \\
b_{-1}(u) &= 1 - u^3 \\
b_{-2}(u) &= 0 \\
b_{-3}(u) &= 0 \ .
\end{aligned}
$$

In this case $\mathbf{Q}_i(u)$ is, in the limit, a straight line running from $\mathbf{V}_{i-1}$ to $\mathbf{V}_i$.

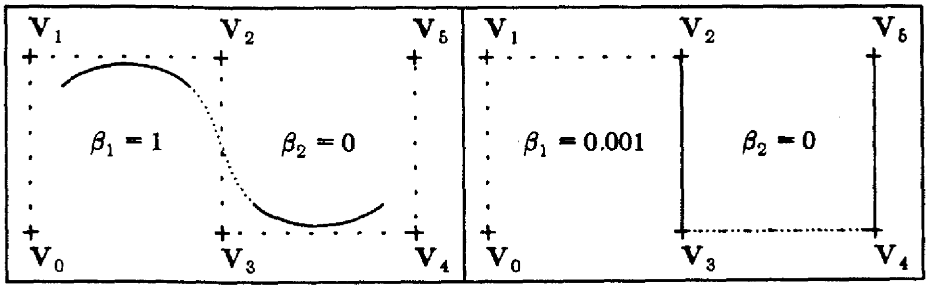


Figure 207. Decreasing $\beta 1$ to zero does draw the curve flat against the control polygon when $\beta 2$ is zero.

Curiously enough, however, if $\beta 2$ is nonzero then as $\beta 1$ approaches zero the basis segments converge to

$$
\begin{aligned}
b_{-0}(u) &= \frac{1}{\beta 2+2} \ 2u^3 \\
b_{-1}(u) &= \frac{1}{\beta 2+2} \left( 2 + 3\beta 2 u^2 - (2\beta 2+2)u^3 \right) \\
b_{-2}(u) &= \frac{1}{\beta 2+2} \left( \beta 2 - 3\beta 2 u^2 + 2\beta 2 u^3 \right) \\
b_{-3}(u) &= 0 \ .
\end{aligned}
$$

Thus as $\beta 1$ approaches zero $\mathbf{Q}_i(u)$ does not, in general, approach a straight line unless $\beta 2$ is zero.

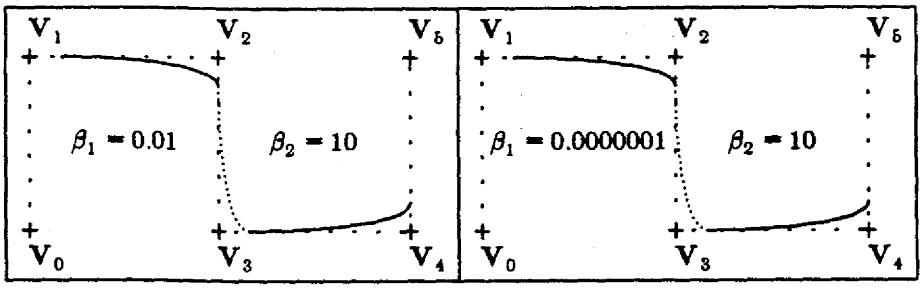


Figure 208. If $\beta 2$ is not zero the curve does not converge to the control polygon as $\beta 1$ approaches zero.

$\beta 1$ and $\beta 2$ may be altered, independent of the control vertices, to change the shape of the curve. In the curves we have been discussing a single value of $\beta 1$ is used for the entire curve, and similarly for $\beta 2$. We would prefer, if possible, to specify distinct values of $\beta 1$ and $\beta 2$ at each joint. Before discussing how this can be done, we indicate briefly how uniformly-shaped Beta-spline surfaces can be constructed from uniformly-shaped Beta-spline curves.

## 18.1. Uniformly-shaped Beta-spline Surfaces

The formation of uniformly-shaped Beta-spline surfaces is completely analogous to our earlier construction of uniform cubic B-spline surfaces. Once again our surface is a scaled sum of basis functions in which $X$, $Y$ and $Z$ are functions of two independent variables:

$$\begin{aligned} \mathbf{Q}(\bar{u},\bar{v}) &= \sum_{i,j} \mathbf{V}_{i,j} B_{i,j}(\bar{u},\bar{v}) \\ &= \sum_{i,j} \left( \ x_{i,j}B_{i,j}(\bar{u},\bar{v}),\ y_{i,j}B_{i,j}(\bar{u},\bar{v}),\ z_{i,j}B_{i,j}(\bar{u},\bar{v}) \ \right) . \end{aligned} \tag{132}$$

For coefficients we again use the $x$-, $y$- and $z$-coordinates of a two-dimensional array of control vertices that we have called the control mesh or control graph. To obtain locality we want the new basis functions $B_{i,j}(\bar{u},\bar{v})$ to be nonzero only for a small range of $\bar{u}$ and $\bar{v}$. One way of arranging this is to let $B_{i,j}(\bar{u},\bar{v}) = B_i(\bar{u})B_j(\bar{v})$, where $B_i(\bar{u})$ and $B_j(\bar{v})$ are simply the univariate basis functions (131) that we developed for the Beta-spline curves. Since each is nonzero only over four successive intervals, if $\bar{u}_i \le \bar{u} \le \bar{u}_{i+1}$ and $\bar{v}_j \le \bar{v} \le \bar{v}_{j+1}$ we can rewrite (132) as

$$\mathbf{Q}_{i,j}(\bar{u},\bar{v}) = \sum_{r=-3}^{0} \sum_{s=-3}^{0} \mathbf{V}_{i+r,j+s} B_{i+r}(\bar{u}) B_{j+s}(\bar{v}) \ . \tag{133}$$

If we rewrite this in terms of basis segments instead of basis functions and recall our convention that the portion of $\mathbf{Q}(u,v)$ defined by this set of values for $u$ and $v$ is denoted by $\mathbf{Q}_{i,j}(u,v)$, then we can write

$$\mathbf{Q}_{i,j}(u,v) = \sum_{r=-3}^{0} \sum_{s=-3}^{0} \mathbf{V}_{i+r,j+s} b_r(u) b_s(v) \tag{134}$$

so that $\mathbf{Q}_{i,j}(u,v)$, the $i,j^{th}$ *patch*, is completely determined by sixteen control vertices. As before, the separability of $B_{i,j}(\bar{u},\bar{v})$ into $B_i(\bar{u})$ and $B_j(\bar{v})$ can be used to establish that the resulting surfaces are $G^2$ continuous across patch boundaries. For example, we can expand (134) as

$$
\begin{aligned}
&\mathbf{Q}_{i,j}(u,v) = \qquad\qquad (135)\\
&[\mathbf{V}_{i-3,j}\ b_{-3}(u) + \mathbf{V}_{i-2,j}\ b_{-2}(u) + \mathbf{V}_{i-1,j}\ b_{-1}(u) + \mathbf{V}_{i,j}\ b_{-0}(u)]\, b_{-0}(v) +\\
&[\mathbf{V}_{i-3,j-1} b_{-3}(u) + \mathbf{V}_{i-2,j-1} b_{-2}(u) + \mathbf{V}_{i-1,j-1} b_{-1}(u) + \mathbf{V}_{i,j-1} b_{-0}(u)]\, b_{-1}(v) +\\
&[\mathbf{V}_{i-3,j-2} b_{-3}(u) + \mathbf{V}_{i-2,j-2} b_{-2}(u) + \mathbf{V}_{i-1,j-2} b_{-1}(u) + \mathbf{V}_{i,j-2} b_{-0}(u)]\, b_{-2}(v) +\\
&[\mathbf{V}_{i-3,j-3} b_{-3}(u) + \mathbf{V}_{i-2,j-3} b_{-2}(u) + \mathbf{V}_{i-1,j-3} b_{-1}(u) + \mathbf{V}_{i,j-3} b_{-0}(u)]\, b_{-3}(v)\ .
\end{aligned}
$$

From this it is clear that if we fix $u$ at some arbitrary value between 0 and 1 then we can write (135) as

$$\mathbf{Q}_{i,j,u}(v) = \mathbf{W}_0 b_{-3}(v) + \mathbf{W}_1 b_{-2}(v) + \mathbf{W}_2 b_{-1}(v) + \mathbf{W}_3 b_{-0}(v)$$

where

$$
\begin{aligned}
\mathbf{W}_3 &= \mathbf{V}_{i-3,j}\ b_{-3}(u) + \mathbf{V}_{i-2,j}\ b_{-2}(u) + \mathbf{V}_{i-1,j}\ b_{-1}(u) + \mathbf{V}_{i,j}\ b_{-0}(u)\\
\mathbf{W}_2 &= \mathbf{V}_{i-3,j-1} b_{-3}(u) + \mathbf{V}_{i-2,j-1} b_{-2}(u) + \mathbf{V}_{i-1,j-1} b_{-1}(u) + \mathbf{V}_{i,j-1} b_{-0}(u)\\
\mathbf{W}_1 &= \mathbf{V}_{i-3,j-2} b_{-3}(u) + \mathbf{V}_{i-2,j-2} b_{-2}(u) + \mathbf{V}_{i-1,j-2} b_{-1}(u) + \mathbf{V}_{i,j-2} b_{-0}(u)\\
\mathbf{W}_0 &= \mathbf{V}_{i-3,j-3} b_{-3}(u) + \mathbf{V}_{i-2,j-3} b_{-2}(u) + \mathbf{V}_{i-1,j-3} b_{-1}(u) + \mathbf{V}_{i,j-3} b_{-0}(u)\ .
\end{aligned}
$$

Thus $\mathbf{Q}_{i,j,u}(v)$ is simply the uniformly-shaped Beta-spline curve segment defined by the "control vertices" $\mathbf{W}_0$, $\mathbf{W}_1$, $\mathbf{W}_2$ and $\mathbf{W}_3$. It is not hard to see that $\mathbf{Q}_{i,j+1,u}(v)$, in the next patch "up", is given by

$$\mathbf{Q}_{i,j+1,u}(v) = \mathbf{W}_1 b_{-3}(v) + \mathbf{W}_2 b_{-2}(v) + \mathbf{W}_3 b_{-1}(v) + \mathbf{W}_4 b_{-0}(v)$$

where

$$\mathbf{W}_4 = \mathbf{V}_{i-3,j+1} b_{-3}(u) + \mathbf{V}_{i-2,j+1} b_{-2}(u) + \mathbf{V}_{i-1,j+1} b_{-1}(u) + \mathbf{V}_{i,j+1} b_{-0}(u)\ .$$

This is simply the second segment in a uniformly-shaped Beta-spline curve defined by the "control vertices" $\mathbf{W}_0$, $\mathbf{W}_1$, $\mathbf{W}_2$, $\mathbf{W}_3$ and $\mathbf{W}_4$. It follows immediately that this curve is $G^2$ continuous. Since a completely analogous argument can be made with respect to $u$ by factoring the $b_r(u)$ out of (134) instead of the $b_s(v)$, the uniformly-shaped Beta-spline surface we have defined is $G^2$ continuous along lines of constant $u$ and $v$.

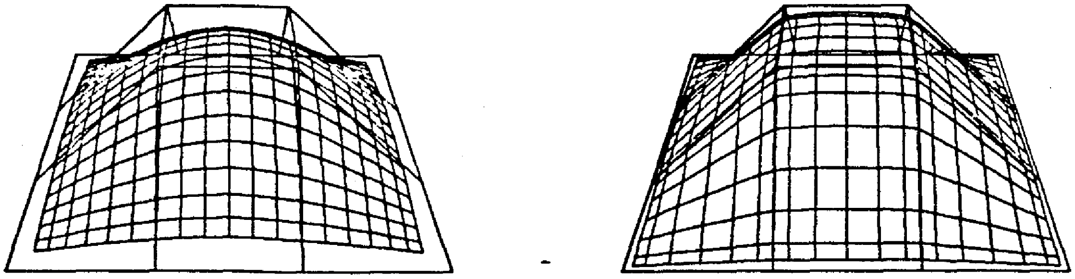


Figure 209. Uniformly shaped Beta-spline surfaces: $\beta 2$ is 0 on the left and 25 on the right.

# 19. Continuously-Shaped Beta-splines

Now we want to see how to generalize the uniformly-shaped Beta-splines so as to obtain local control of the shape parameters $\beta 1$ and $\beta 2$. The material in this chapter is taken from [Barsky82, Barsky83].

Let $\beta 1_i$ and $\beta 2_i$ be the values of $\beta 1$ and $\beta 2$, respectively, to be associated with the joint between $\mathbf{Q}_{i-1}(u)$ and $\mathbf{Q}_i(u)$. We would like to use the basis segments given by equations (131), making $\beta 1$ and $\beta 2$ functions of $u$ in such a way as to interpolate between the $\beta 1_i$'s and $\beta 2_i$'s at each end of a segment while preserving $G^2$ continuity of the curve.

Let us consider the following derivative with respect to $u$ of a representative term of (131),[3]

$$\frac{c\,[\beta(u)]^p\,u^q}{\delta(u)}\ , \tag{136}$$

where $c$ is a constant. Its first parametric derivative with respect to $u$ is

$$\frac{c\,q\,[\beta(u)]^p\,u^{q-1}}{\delta(u)} + \frac{c\,p\,[\beta(u)]^{p-1}\,\beta^{(1)}(u)\,u^q}{\delta(u)} - \frac{c\,[\beta(u)]^p\,\delta^{(1)}(u)\,u^q}{\delta(u)^2}\ , \tag{137}$$

where

$$\begin{aligned} \delta(u) &= \beta 2(u) + 2[\beta 1(u)]^3 + 4[\beta 1(u)]^2 + 4\beta 1(u) + 2\ . \\ \delta^{(1)}(u) &= \beta 2^{(1)}(u) + 6[\beta 1(u)]^2\beta 1^{(1)}(u) + 8\beta 1(u)\beta 1^{(1)}(u) + 4\beta 1^{(1)}(u) \end{aligned} \tag{138}$$

Examination of (137) and (138) reveals that the second and third terms of (137) involve products with $\beta 1^{(1)}(u)$ or $\beta 2^{(1)}(u)$, while the first term of (137) would constitute the complete parametric derivative if $\beta 1$ and $\beta 2$ were not functions of $u$. If we were to compute $\beta 1(u)$ and $\beta 2(u)$ by interpolating between the $\beta 1_i$'s and $\beta 2_i$'s in such a way as to cause $\beta 1^{(1)}(u)$ and $\beta 2^{(1)}(u)$ to be zero at each joint then equations (128) would hold and $G^1$ continuity would be preserved.

Similarly, the second parametric derivative of (136) is

$$\frac{c\,(q-1)q\,[\beta(u)]^p\,u^{q-2}}{\delta(u)} \tag{139}$$

[3] We will use $\beta(u)$ rather than $\beta 1(u)$ or $\beta 2(u)$ when the argument applies to both. No confusion can occur because products of $\beta 1$ and $\beta 2$ do not arise. Similarly, $\beta_i$ will be used to represent both $\beta 1_i$ and $\beta 2_i$

$$- \frac{c\,[\beta(u)]^p\,\delta^{(2)}(u)\,u^q}{\delta(u)^2} + \frac{2c\,[\beta(u)]^p\,\delta^{(1)}(u)^2\,u^q}{\delta(u)^3}$$
$$- \frac{2cp\,[\beta(u)]^{p-1}\,\beta^{(1)}(u)\,\delta^{(1)}(u)\,u^q}{\delta(u)^2} - \frac{2cq\,[\beta(u)]^p\,\delta^{(1)}(u)\,u^{q-1}}{\delta(u)^2}$$
$$+ \frac{cp\,[\beta(u)]^{p-1}\,\beta^{(2)}(u)\,u^q}{\delta(u)} + \frac{c(p-1)p\,[\beta(u)]^{p-2}\,\beta^{(1)}(u)^2\,u^q}{\delta(u)}$$
$$+ \frac{2cpq\,[\beta(u)]^{p-1}\,\beta^{(1)}(u)\,u^{q-1}}{\delta(u)} ,$$

where

$$\delta^{(2)}(u) = \beta 2^{(2)}(u) + 6\,[\beta 1(u)]^2\,\beta 1^{(2)}(u) + 8\,\beta 1(u)\,\beta 1^{(2)}(u)$$
$$+ 4\,\beta 1^{(2)}(u) + 12\,\beta 1(u)\,\beta 1^{(1)}(u)^2 + 8\,\beta 1^{(1)}(u)^2 .$$

Again, only the first term of (139) lacks a product with at least one of $\beta 1^{(1)}(u)$, $\beta 2^{(1)}(u)$, $\beta 1^{(2)}(u)$ or $\beta 2^{(2)}(u)$, and the first term would constitute the complete second parametric derivative if $\beta 1$ and $\beta 2$ were not functions of $u$. Thus arranging that all four derivatives have the value zero at joints should be sufficient to preserve $G^2$ continuity of the curve. This is easily accomplished in the following manner.

Suppose that we use a polynomial $H(\beta_{i-1},\beta_i;u)$ to interpolate between $\beta_{i-1}$ and $\beta_i$.

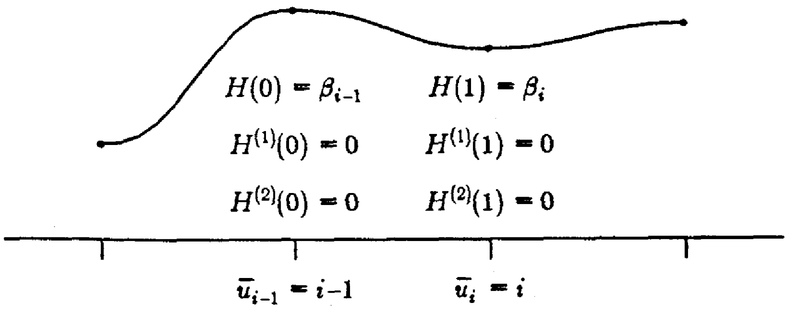


Figure 210. The idea is to interpolate the $\beta_i$ in such a way as to cause the first and second derivatives at the knots to be zero.

We have six constraints, since we would like

$$H(\beta_{i-1},\beta_i;0) = \beta_{i-1}$$
$$H(\beta_{i-1},\beta_i;1) = \beta_i$$
$$H^{(1)}(\beta_{i-1},\beta_i;0) = 0$$
$$H^{(1)}(\beta_{i-1},\beta_i;1) = 0$$
$$H^{(2)}(\beta_{i-1},\beta_i;0) = 0$$
$$H^{(2)}(\beta_{i-1},\beta_i;1) = 0 .$$

This suggests the use of a fifth degree polynomial (which has, of course, six coefficients). If

$$H(\beta_{i-1},\beta_i;u) = a + bu + cu^2 + du^3 + eu^4 + fu^5$$

then the above equations take the form

$$
\begin{aligned}
H(\beta_{i-1},\beta_i;0) &= \beta_{i-1} = a \\
H(\beta_{i-1},\beta_i;1) &= \beta_i = a + b + c + d + e + f \\
H^{(1)}(\beta_{i-1},\beta_i;0) &= 0 = b \\
H^{(1)}(\beta_{i-1},\beta_i;1) &= 0 = b + 2c + 3d + 4e + 5f \\
H^{(2)}(\beta_{i-1},\beta_i;0) &= 0 = 2c \\
H^{(2)}(\beta_{i-1},\beta_i;1) &= 0 = 2c + 6d + 12e + 20f \ .
\end{aligned}
$$

It is straightforward to obtain the polynomial

$$
\begin{aligned}
\beta_i(u) = H(\beta_{i-1},\beta_i;u) &= \beta_{i-1} + 10(\beta_i-\beta_{i-1})u^3 - 15(\beta_i-\beta_{i-1})u^4 + 6(\beta_i-\beta_{i-1})u^5 \\
&= \beta_{i-1} + (\beta_i-\beta_{i-1})[10u^3 - 15u^4 + 6u^5] \qquad (140)
\end{aligned}
$$

which satisfies these equations; this is, in fact, a special case of quintic Hermite interpolation. By the argument given above, the use of (140) to interpolate $\beta 1$ and $\beta 2$ in (131) preserves $G^2$ continuity of the curve.

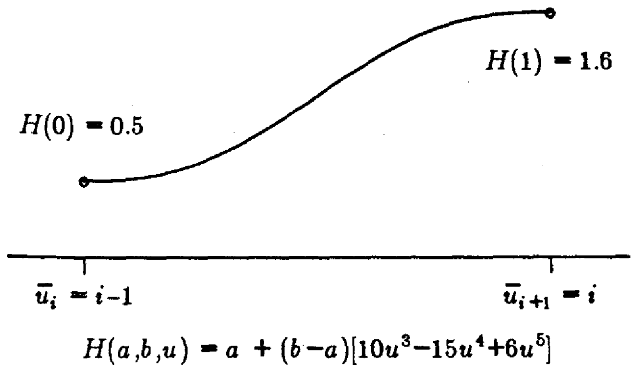


Figure 211. An application of the formula for interpolating $\beta$ values between joints.

It is, of course, possible that the derivative terms appearing in (136) and (139) might sum in such a way as to yield $G^2$ continuous curves even though the derivatives were nonzero; we have not ruled this out for all other interpolation schemes. However, using Vaxima [Bogen77, Fateman82] it is not hard to produce examples that demonstrate that neither linear interpolation nor cubic Hermite interpolation work. Moreover, geometric continuity is not necessarily preserved if we use general quintic Hermite interpolation, even if the same two nonzero values are used for the first and second derivatives of $\beta 1(u)$ at the joints (and similarly for $\beta 2(u)$). Thus $C^2$ continuity of $\beta 1(u)$ and $\beta 2(u)$ is not sufficient to ensure $G^2$ continuity. (See [Barsky82] for an example.)

We shall refer to the curves whose segments are defined by equations (13) and (131), where $\beta 1$(u) and $\beta 2$(u) are interpolated by equation (140), as *continuously-shaped Beta-spline curves.*

## 19.1. Locality

Just as for the uniformly-shaped Beta-splines, each basis function is nonzero only over four successive intervals. Since each basis function is used to weight a particular control vertex, moving a control vertex will alter only the four corresponding curve segments. These are, of course, consecutive.

The effect of altering a $\beta_i$ is more localized still. The $\beta_i$ at a particular joint determines how $\beta$ is interpolated over the segments that meet at that joint, so that only two curve segments are changed.

### 19.2. Bias

The following figure illustrates a few of the effects that can be obtained by altering $\beta 1$'s.

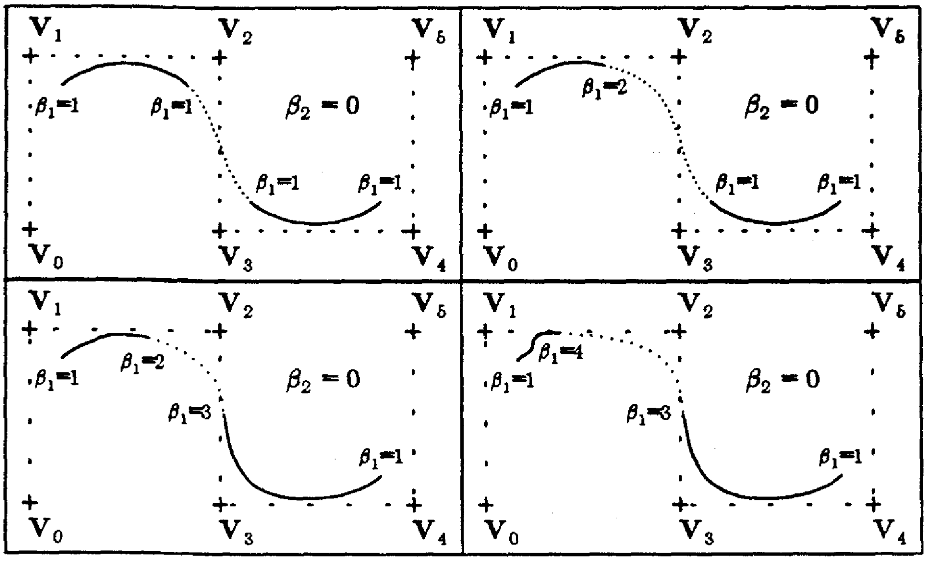


Figure 212. For these curves $\beta 2$ is held constant at zero while $\beta 1$ interpolates the $\beta 1$ values shown.

Although the resulting curves are often visually satisfying, their extreme locality with respect to changes in the shape parameters can result in "kinks" if there are large differences in the $\beta$ values for consecutive control vertices. A modest reduction in the size of the jumps ameliorates the effect.

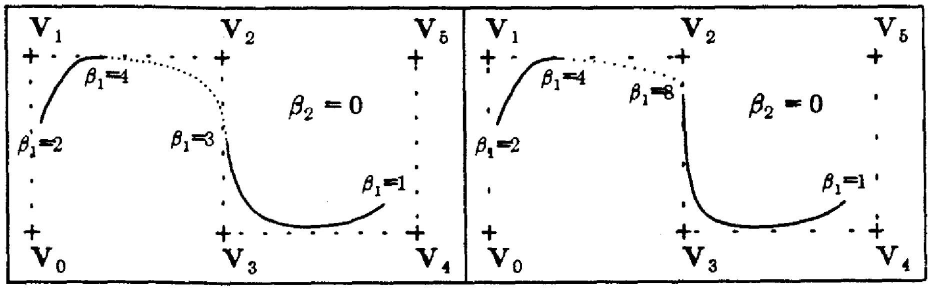


Figure 213. Less abrupt variation of $\beta 1$ can be used to smooth out the kink visible in the lower right frame of Figure 212, if that is desirable.

### 19.3. Tension

Since this scheme interpolates the $\beta_i$, the discussion of tension in [Barsky81, Barsky85, Barsky85a] is equally applicable here. We already know that the effect of increasing $\beta 2$ is to draw the curve towards the control polygon. Let us examine the path followed by a particular joint, say the joint between the $(i-1)^{st}$ and $i^{th}$ segments.

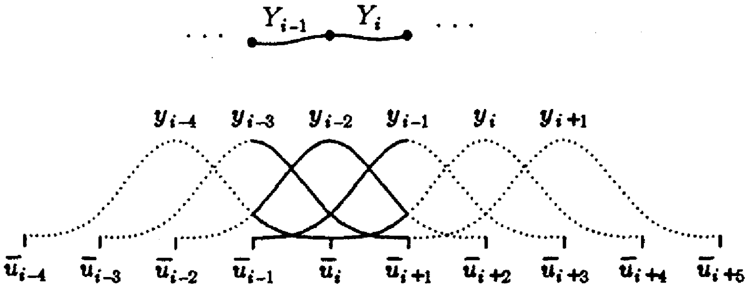


Figure 214. This illustration may help in keeping track of the indices. The $i^{th}$ interval runs from $\bar{u}_i$ to $\bar{u}_{i+1}$. The $i^{th}$ control vertex (actually the $i^{th}$ $y$ coordinate here) scales the basis function whose support begins at $\bar{u}_i$. The $i^{th}$ uniform cubic B-spline peaks at $\bar{u}_{i+2}$; the peak of $i^{th}$ Beta-spline $B_i(\bar{u})$ may be displaced left or right of $\bar{u}_{i+2}$ by decreasing or increasing $\beta 1$.

The difference

$$\mathbf{Q}_i(0) - \mathbf{V}_{i-2} = \mathbf{Q}_{i-1}(1) - \mathbf{V}_{i-2} = \frac{(\mathbf{C} - c\,\mathbf{V}_{i-2})}{(c + \beta 2_i)}$$

is the vector from the $(i-2)^{nd}$ control vertex to this joint, where

$$\mathbf{C} = 2\beta 1_i^3 \mathbf{V}_{i-2} + 4\beta 1_i(\beta 1_i + 1)\mathbf{V}_{i-1} + 2\mathbf{V}_i$$

$$c = 2\beta 1_i^3 + 4\beta 1_i^2 + 4\beta 1_i + 2 \ .$$

Altering $\beta 2_i$ merely changes the length of this vector: values approaching $-c$ "push" the joint arbitrarily far away from $\mathbf{V}_{i-2}$; large positive or negative values draw the joint arbitrarily close to $\mathbf{V}_{i-2}$, pulling the two segments meeting at that joint flat against the control polygon. Hence $\beta 2_i$ serves as a tension parameter, just as for uniformly-shaped Beta-spline curves.

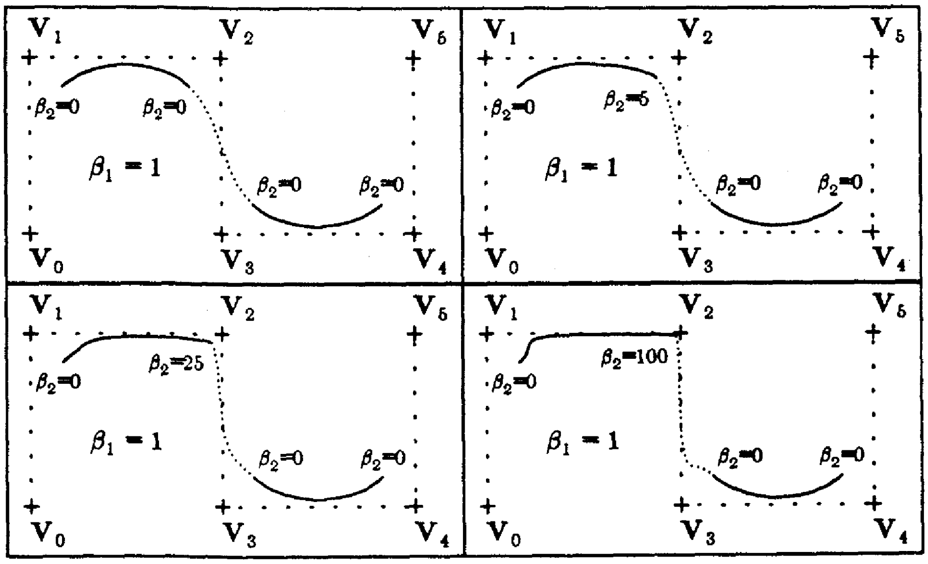


Figure 215. The value of $\beta 2$ at the joint nearest to $\mathbf{V}_2$ is increased from 0 to 100 in three steps, pulling the joint towards $\mathbf{V}_2$. In the limit this joint converges to $\mathbf{V}_2$.

Again, wildly disparate values of $\beta 2$ for adjacent control vertices can produce kinks. These can be removed, if that is desirable, by smaller adjustments in neighbouring $\beta$ values, as shown below.

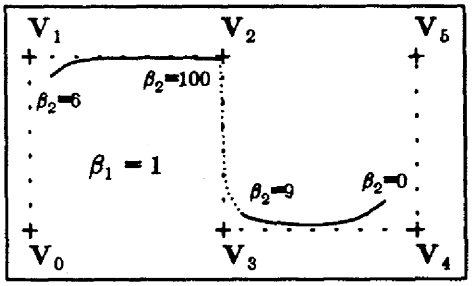


Figure 216. Altering the value of $\beta 2$ at a joint affects only the two curve segments that meet there. Making one such $\beta 2$ very large in comparison with its neighbors, as in Figure 215, causes these two segments to be abruptly pulled close to the control polygon. The value of $\beta 2$ at adjacent joints can be adjusted to smooth out the curve.

For comparison with Figures 203 and 205 we give some examples of continuously-shaped basis functions.

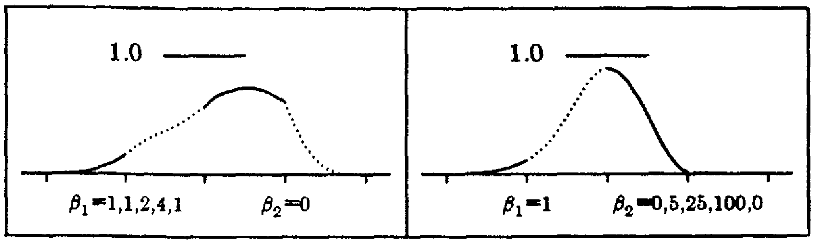


Figure 217. Here we illustrate the effect of interpolating $\beta$ values on the basis functions. On the left $\beta 1$ changes value from joint to joint, while on the right $\beta 2$ changes value.

Notice that each point on a continuously-shaped Beta-spline curve $\mathbf{Q}(u)$ also lies on the uniformly-shaped curve $\mathbf{R}(u)$ defined by the same control vertices and the values of $\beta 1$ and $\beta 2$ at that point on $\mathbf{Q}(u)$. The behaviour of $\mathbf{Q}(u)$ as $\beta 2$ is varied can therefore be inferred from the behaviour of the corresponding uniformly-shaped curves. Thus $\beta 2$ values can be used to locally force a curve to converge to the control polygon if they are increased arbitrarily.

## 19.4. Convex Hull

Like the uniformly-shaped Beta-spline curves, continuously-shaped Beta-spline curves possess a convex hull property in that the $i^{\text{th}}$ segment lies within the convex hull of control vertices $\mathbf{V}_{i-3}$, $\mathbf{V}_{i-2}$, $\mathbf{V}_{i-1}$ and $\mathbf{V}_i$, so long as both $\beta 1$ and $\beta 2$ are nonnegative. The argument, as we shall see, is straightforward. Recall that because each basis function is nonzero over four intervals, we have

$$\mathbf{Q}_i(u) = \mathbf{V}_{i-3}b_{-3}(u) + \mathbf{V}_{i-2}b_{-2}(u) + \mathbf{V}_{i-1}b_{-1}(u) + \mathbf{V}_i b_{-0}(u) \ . \qquad (141)$$

Now for any given value of $u$, $\beta 1(u)$ and $\beta 2(u)$ yield some particular value of $\beta 1$ and $\beta 2$. By simply summing equations (131) we see that for every such $\beta 1$, $\beta 2$ and $u$

$$b_{-0}(u) + b_{-1}(u) + b_{-2}(u) + b_{-3}(u) = 1 \ .$$

Next we must verify that these basis segments are nonnegative for all $u$ in the interval [0,1]. If we rewrite equations (131) in the form

$$b_{-0}(u) = \frac{1}{\delta}\left(\,2u^3\,\right)$$

$$b_{-1}(u) = \frac{1}{\delta}\left(\,2\beta1^2u^2(3-u) + 2\beta1u(3-u^2) + \beta2u^2(3-2u) + 2(1-u^3)\,\right)$$

$$b_{-2}(u) = \frac{1}{\delta}\left(\,2\beta1^3u((1-u)(2-u)+1) + 2\beta1^2(u^3-3u^2+2)\right.$$
$$\left. + 2\beta1(\,u^3-3u+2) + \beta2(\,2u^3-3u^2+1)\,\right)$$

$$b_{-3}(u) = \frac{1}{\delta}\left(\,2\beta1^3(1-u)^3\,\right)$$

where

$$\delta = \beta2 + 2\beta1^3 + 4\beta1^2 + 4\beta1 + 2 \neq 0$$

for $\beta1 \geq 0$, $\beta2 \geq 0$, and $u \in [0,1]$, it is easy to see by inspection that $b_{-0}(u)$, $b_{-1}(u)$, and $b_{-3}(u)$ are nonnegative. For $b_{-2}(u)$, elementary consideration of the zeros of the derivatives $3u(u-2)$, $3(u-1)(u-1)$ and $6u(u-1)$ of $u^3-3u^2+2$, $u^3-3u+2$, and $2u^3-3u^2+1$ yields the same conclusion. Since $\beta1$ and $\beta2$ are actually interpolated by (140), it is necessary to show that

$$\beta_i(u) = \beta_{i-1} + (\beta_i - \beta_{i-1})[10u^3 - 15u^4 + 6u^5] \geq 0$$

if $\beta_{i-1} \geq 0$, $\beta_i \geq 0$, and $u \in [0,1]$. Consider

$$\beta_i^{(1)}(u) = 30(\beta_i - \beta_{i-1})u^2(1-u)^2 \; .$$

Clearly the slope changes sign only at $u=0$ and $u=1$. Since

$$\beta_i(0.5) = \frac{\beta_i + \beta_{i+1}}{2} \geq 0 \quad \text{if } \beta_{i-1}, \beta_i \geq 0,$$

$\beta_i(u)$ must be nonnegative on [0,1] so long as the $\beta_i$ are nonnegative.

Hence so long as $\beta1_i \geq 0$ and $\beta2_i \geq 0$, $\mathbf{Q}_i(u)$ lies within the convex hull of $\mathbf{V}_{i-3}$, $\mathbf{V}_{i-2}$, $\mathbf{V}_{i-1}$ and $\mathbf{V}_i$.

### 19.5. End Conditions

Just as for the uniform cubic B-splines, a properly defined continuously-shaped Beta-spline curve segment is the linear combination of four basis functions, as in equation (141). Thus $m+1$ control vertices $\mathbf{V}_0, \cdots, \mathbf{V}_m$ can be used to define $m-2$ segments, which we index as $\mathbf{Q}_3(u), \cdots, \mathbf{Q}_m(u)$. The Beta-spline curve then begins[4] (see [Barsky81] and [Barsky85]) at

$$\mathbf{Q}_3(0) = \frac{1}{\delta(0)}\left(\,2\beta1_i^3\mathbf{V}_0 + (\delta(0) - 2\beta1_i^3 - 2)\mathbf{V}_1 + 2\mathbf{V}_2\,\right) .$$

[4] The terminal point of the curve is analyzed in an exactly analogous manner, and we therefore omit explicit treatment of it.

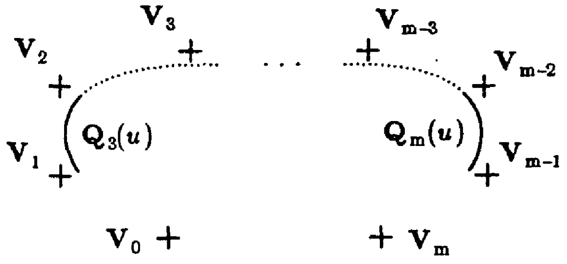


Figure 218. It is hard to say more than that a typical Beta-spline curve begins and ends "in the vicinity" of the first and last control vertices. For example it does not, in general, interpolate any of them.

Thus the curve does not, in general, begin at a control vertex, or even at a point along the line segment from $\mathbf{V}_0$ to $\mathbf{V}_1$. In order to obtain better control of the beginning of the curve, one therefore often treats the ends of the curve specially.

Let $\mathbf{Q}(\bar{u})$ be a continuously-shaped Beta-spline curve with $\beta 1 = \beta 1_i$ and $\beta 2 = \beta 2_i$ at the joint between the $i^{th}$ and $i+1^{st}$ segments. Let $\mathbf{R}(\bar{u})$ be a uniformly-shaped Beta-spline curve defined by the same control vertices, but with $\beta 1 = \beta 1_i$ and $\beta 2 = \beta 2_i$ throughout. By the definition of $\mathbf{Q}(\bar{u})$ we must have $\mathbf{Q}(\bar{u}) = \mathbf{R}(\bar{u})$, $\mathbf{Q}^{(1)}(\bar{u}) = \mathbf{R}^{(1)}(\bar{u})$ and $\mathbf{Q}^{(2)}(\bar{u}) = \mathbf{R}^{(2)}(\bar{u})$ at the joint in question. Hence the analysis of end conditions in [Barsky81] applies immediately to continuously-shaped Beta-spline curves. In fact these techniques for controlling Beta-spline end conditions turn out to be identical to the techniques discussed earlier for uniform cubic B-splines (which are a special case of the Beta-splines), although the analysis is more complicated. For convenience we summarize them here, but do not give a detailed development.

- *A Double First Vertex*. We define an additional segment at the beginning of the curve by

  $$\mathbf{Q}_2(u) = \mathbf{V}_0[b_{-3}(u)+b_{-2}(u)] + \mathbf{V}_1 b_{-1}(u) + \mathbf{V}_2 b_{-0}(u) \ .$$

  $\mathbf{Q}_2(u)$ begins at a point lying along the line segment from $\mathbf{V}_0$ to $\mathbf{V}_1$, at which point it is tangent to that line and has zero curvature.

- *A Triple First Vertex*. We define two additional segments at the beginning of the curve by

  $$\mathbf{Q}_1(u) = \mathbf{V}_0[b_{-3}(u)+b_{-2}(u)+b_{-1}(u)] + \mathbf{V}_2 b_{-0}(u)$$
  $$\mathbf{Q}_2(u) = \mathbf{V}_0[b_{-3}(u)+b_{-2}(u)] + \mathbf{V}_1 b_{-1}(u) + \mathbf{V}_2 b_{-0}(u) \ .$$

  The curve then begins at $\mathbf{Q}_1(0) = \mathbf{V}_0$ and the first segment of the curve is a short straight line. The behaviour of the second segment $\mathbf{Q}_2(u)$, which has a double first vertex, is described above.

The analysis of double and triple vertices is equally applicable on the interior of a curve. Triple interior vertices are particularly interesting since they can result in a cusp:

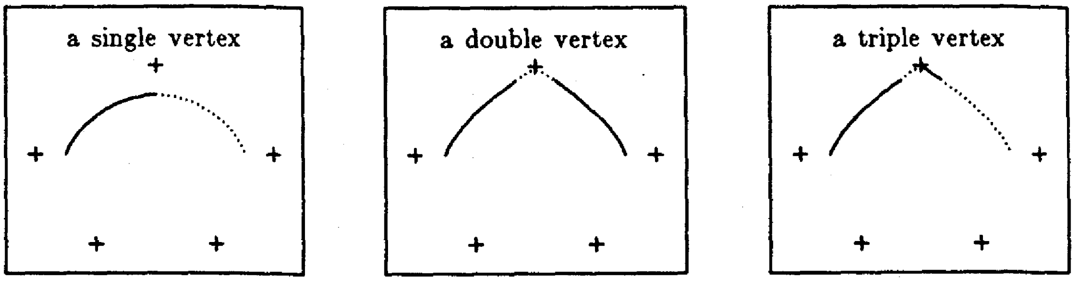


Figure 219. $\beta 1$ is one and $\beta 2$ is zero at all joints; these are in fact simply uniform cubic B-spline curves, although a cusp results at a triple vertex for any values of $\beta 1$ and $\beta 2$ unless the control vertices immediately preceding and following the vertex are both collinear with it. The double control vertex is not interpolated, while the triple vertex is.

This cusp is not a violation of $G^2$ continuity because, or at least in the sense that, the first parametric derivative vector has the value (0,0) at the joint that coincides with the interpolated control vertex where the cusp occurs, so that the unit tangent vector is not defined. Multiple vertices give a tension-like effect, and it is instructive to compare the effect of repeating a vertex with the effect of altering $\beta 2$ there:

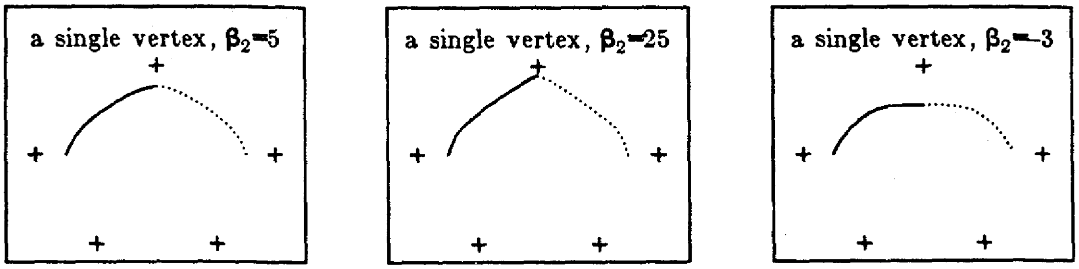


Figure 220. Here $\beta 1$ is one at all joints and $\beta 2$ is zero except as indicated.

An alternative way of controlling the beginning of a curve is to automatically define a *phantom vertex* $\mathbf{V}_{-1}$ and a corresponding initial segment

$$\mathbf{Q}_2(u) = \mathbf{V}_{-1} b_{-3}(u) + \mathbf{V}_0 b_{-2}(u) + \mathbf{V}_1 b_{-1}(u) + \mathbf{V}_2 b_{-0}(u)$$

in such a way as to satisfy some requirement. We may ask that:

- $\mathbf{Q}_2(0)$ interpolate some furnished point (generally resulting in nonzero curvature);
- $\mathbf{Q}_2(0)$ interpolate $\mathbf{V}_0$ (at which point the curvature is then zero);
- $\mathbf{Q}_2^{(1)}(0)$ have some specified value (generally resulting in nonzero curvature);
- $\mathbf{Q}_2^{(2)}(0)$ have some specified value (generally resulting in nonzero curvature);
- $\mathbf{Q}_2^{(2)}(0)$ be zero, resulting in zero curvature at $\mathbf{Q}_2(0)$.

All these techniques involve extending the curve by one or two segments at either end. This implies the existence of additional joints and associated $\beta$ values. Hence the sequence of control vertices is extended in order to specify behaviour at the ends of the curve, and additional $\beta 1$ and $\beta 2$ values must be specified as well. These may take any value without affecting the behaviour described above. In practice it is probably easiest simply to replicate $\beta$ values as well as vertices.

The curves we have discussed so far are *open* curves, which is to say that the two endpoints do not, in general, coincide. A $G^2$-continuous *closed* curve whose endpoints do meet and which is $G^2$-continuous is obtained if the first three control vertices are identical to the last three and the same values of $\beta 1$ and $\beta 2$ are used at the joint between the beginning and the ending of the curve.

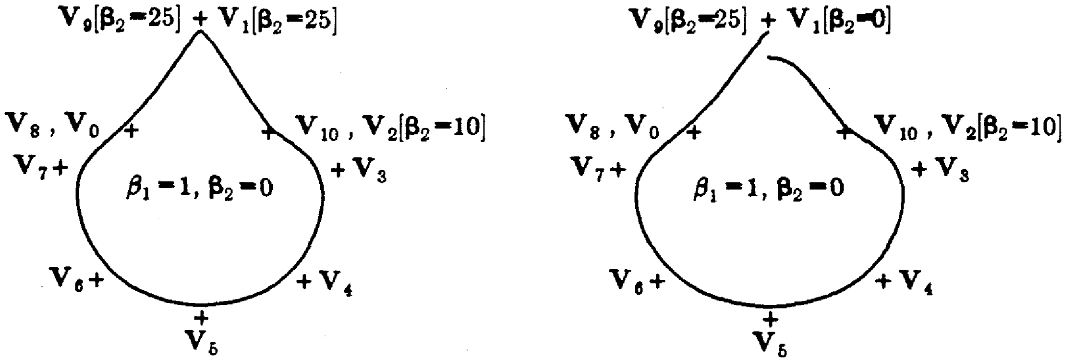


Figure 221. On the left is a continuously-shaped Beta-spline curve in which the first three and the last three control vertices are identical and the values of $\beta 1$ and $\beta 2$ at the second and penultimate control vertices are also identical; a closed $G^2$ continuous curve results. The right hand curve is defined identically except that the second and penultimate control vertices, whose positions coincide, have distinct values of $\beta 2$; a discontinuity results.

Although it may appear in this figure that the joint near $\mathbf{V}_1$ in the left-hand curve is a cusp, by zooming in on the joint we can see that in fact curvature continuity is maintained.

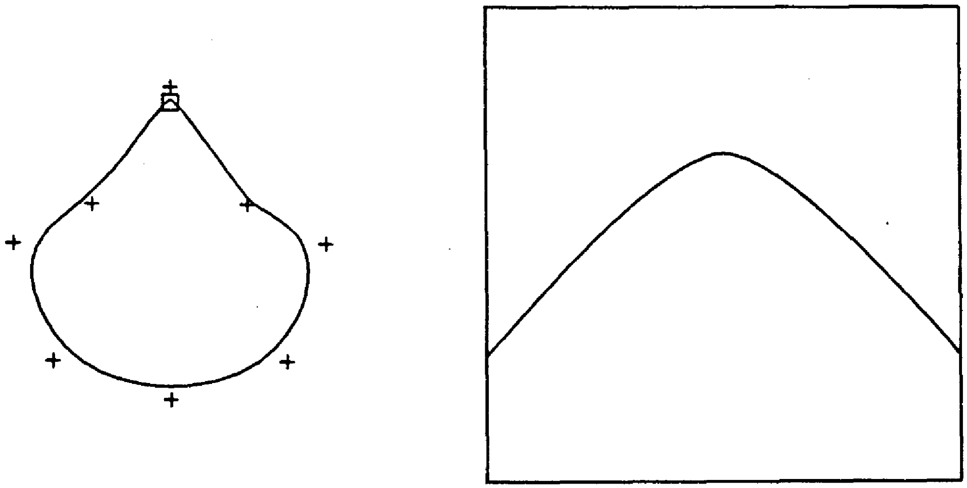


Figure 222. On the right is a magnified image of the indicated portion of the curve shown on the left.

Again, the arguments establishing these results appear in [Barsky81] and the details have therefore been omitted.

## 19.6. Evaluation

Using factorizations given in [Barsky81] and [Barsky85b], the Beta-spline basis segments (131) can be evaluated in 28 multiplication/divisions and 21 addition/subtractions. If a *single point* on $Q(\bar{u})$ is to be determined, the evaluation of the right hand side of (13) in $d$ dimensions then requires $4d$ multiplications and $3d$ additions. The total cost for evaluating a point on a uniformly-shaped 2D Beta-spline curve is therefore 36 multiplication/divisions and 27 addition/subtractions; a 3D uniformly-shaped Beta-spline curve requires 40 multiplication/divisions and 30 addition/subtractions.

For a continuously-shaped Beta-spline curve, equation (140) can be evaluated in 6 multiplications and 4 addition/subtractions if it is factored into the form

$$H(\beta_{i-1},\beta_i,u) = \beta_{i-1} + (\beta_i - \beta_{i-1})[10+(6u-15)u]u^3 \ .$$

Since both $\beta_1(\bar{u})$ and $\beta_2(\bar{u})$ must be computed, both $H(\beta 1_{i-1},\beta 1_i,u)$ and $H(\beta 2_{i-1},\beta 2_i,u)$ must be evaluated. However, since $[10+(6u-15)u]u^3$ need only be evaluated once, the total cost of interpolation is 7 multiplications and 6 addition/subtractions. The additional cost for a single evaluation by this technique of a continuously-shaped Beta-spline curve, beyond that required to evaluate a uniformly-shaped curve, is therefore about 20%.

More often we wish to evaluate a sequence of points along each segment in order to render a curve. If we compute these points by repeatedly evaluating the basis functions as described above, then a uniformly-shaped 2D Beta-spline segment can be evaluated at $r$ values of $u$ in $16+20r$ multiplication/divisions and $14+13r$ addition/subtractions while its 3D counterpart requires $16+24r$ multiplication/divisions and $14+16r$ addition/subtractions. The corresponding cost to evaluate a 2D continuously-shaped Beta-spline curve is $36r$ multiplication/divisions and $2+31r$ addition/subtractions, while in 3D the cost is $41r$ multiplication/divisions and $2+34r$ addition/subtractions. The difference between the evaluation of uniformly- and continuously-shaped Beta-spline curves results from the need to re-evaluate the coefficients of the polynomials forming the basis segments, owing to the fact that $\beta_1$ and $\beta_2$ are no longer constant, as well as from the cost of actually performing the interpolation [Barsky81, Barsky85b].

If instead we first sum the terms in equations (14) so as to compute the coefficients of $X(\bar{u})$ and $Y(\bar{u})$, and then use Horner's rule (nested multiplication), then the evaluation of a 2D uniformly-shaped Beta-spline segment at $r$ points requires $49+6r$ multiplication/divisions and $38+6r$ addition/subtractions while the 3D curve requires $65+9r$ multiplication/divisions and $50+9r$ addition/subtractions. A modified version of this algorithm which computes continuously-shaped Beta-spline curves requires $55r$ multiplication/divisions and $2+48r$ addition/subtractions in 2D and $75r$ multiplication/divisions and $2+63r$ addition/subtractions in 3D.

A third alternative is to use forward differencing techniques. For large values of $r$ the evaluation of a 2D uniformly-shaped curve in this way is almost a factor of 17 faster than the evaluation of a continuously-shaped curve using Horner's rule, although it is subject to cumulative roundoff error. While in principle forward differencing is applicable to the continuously-shaped Beta-splines as well, in fact it is impractical since each coordinate is the quotient of an 18th and a 15th degree polynomial. Where cost is a crucial factor it may be desirable to fix $\beta_1$ at one and manipulate $\beta_2$ alone. Doing so significantly reduces the expense of evaluating equations (131) after interpolating $\beta_2$; each coordinate is then the quotient of an 8th and a 5th degree polynomial.

There are other possibilities. Uniformly-shaped Beta-splines are translates of one another, and need only be evaluated for the first segment drawn if they are saved and reused. In the case of continuously-shaped Beta-splines, each joint is associated with distinct values of $\beta_1$ and $\beta_2$, so that in general each basis function has a different shape and must be individually evaluated. (The rendering of curves by "subdivision" will be discussed in Chapter 11.)

An existing curve can be altered much more efficiently than a new curve can be drawn. If a control vertex is moved then only four segments of the curve must be recomputed, since the basis function that the vertex weights is nonzero on only four successive intervals. Because the vertex is usually moved several times in succession, it is advantageous to save the basis segments as they are first evaluated to avoid recomputing them. Moreover, the portions of the computation for each segment that are actually dependent on the vertex being moved may be segregated from those portions of the computation that are not, and which therefore need not be recomputed.

Altering a $\beta$ parameter necessitates recomputing only two intervals, although all the basis segments in each must be re-evaluated.

### 19.7. Continuously-Shaped Beta-spline Surfaces

Continuously-shaped Beta-spline curves can be elegantly generalized to define surfaces that preserve $G^2$ continuity at the boundaries between adjacent patches. The generalization we shall present allows the user to specify a bias and tension parameter at each corner of a patch; of course, patches that share a corner make use of the same $\beta$ values at that corner. The technique is to generalize the univariate interpolation formula (140) to a bivariate formula in such a way that:

- the $\beta$ values at the four corners of a patch are interpolated;
- two patches which share an edge will have the same $\beta$ values along that edge;
- the first and second partial derivatives of $\beta 1(u,v)$ and $\beta 2(u,v)$ across a patch boundary will be zero.

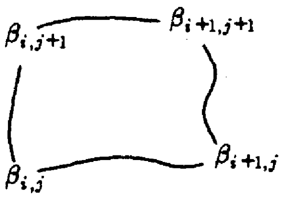
This last property will allow us to ignore (at boundaries) all but one of the terms which arise in computing the partial derivatives of a Beta-spline surface in which $\beta 1(u,v)$ and $\beta 2(u,v)$ are allowed to vary, so that the properties of a uniformly-shaped Beta-spline surface will be inherited by our continuously-shaped surface.

Thus our first consideration is to develop a bivariate interpolation formula. It is at least plausible that we would like lines of constant $u$ or of constant $v$ on a continuously-shaped surface to be continuously-shaped curves. Along such curves we would then expect $\beta 1$ and $\beta 2$ to vary as they do along continuously-shaped Beta-spline curves. For convenience let us write equation (140) in two pieces as

$$s = 10u^3 - 15u^4 + 6u^5$$

$$H(\beta_{i-1},\beta_i;u) = (1-s)\beta_{i-1} + s\beta_i$$

and along the top and bottom boundaries of the patch interpolate the $\alpha$ values



with our customary formula to obtain

$$\beta_{top} = H(\beta_{i-1,j},\beta_{i,j};u) = (1-s)\beta_{i-1,j} + s\beta_{i,j}$$

$$\beta_{bot} = H(\beta_{i-1,j-1},\beta_{i,j-1};u) = (1-s)\beta_{i-1,j-1} + s\beta_{i,j-1} \ .$$

This yields values of $\beta$ at parametric distance $u$ from the left edge along the top and bottom of the patch. To interpolate in the $v$ direction across the interior of the patch it is natural to again use the formula

$$H(\beta_{bot},\beta_{top};v) = (1-t)\beta_{bot} + t\beta_{top}$$

with

$$t = 10v^3 - 15v^4 + 6v^5$$

Substituting, we obtain the desired bivariate interpolation formula

$$\beta_{i,j}(u,v) = (1-s)(1-t)\beta_{i-1,j-1} + s(1-t)\beta_{i,j-1} + (1-s)t\beta_{i-1,j} + st\beta_{i,j} \qquad (142)$$

with

$$s = 10u^3 - 15u^4 + 6u^5$$

$$t = 10v^3 - 15v^4 + 6v^5 \ .$$

(We emphasize that $s$ and $t$ are used here for notational convenience.) $\beta_{i,j}(u,v)$ has some rather attractive properties:

- it interpolates $\beta_{i-1,j-1}$, $\beta_{i,j-1}$, $\beta_{i-1,j}$ and $\beta_{i,j}$;
- along any of the four borders of a patch it reduces to the univariate interpolating formula (140);
- the first and second partial derivatives of $\beta_{i,j}(u,v)$ with respect to $v$ for $v=0$ and $v=1$ (*i.e.* across a vertical patch boundary) are zero, as are the first and second partial derivatives with respect to $u$ for $u=0$ and $u=1$.

Now let us define a continuously-shaped Beta-spline surface patch $\mathbf{Q}_{i,j}$ by equation (134) except that we let $\beta 1$ and $\beta 2$ be functions of $u$ and $v$, using equation (142) to interpolate between $\beta$ values associated with the corners of each patch. To simplify the notation we shall actually discuss $\mathbf{Q}_{3,3}$ and $\mathbf{Q}_{3,4}$, which are defined by the control vertex mesh

$$\begin{array}{cccc}
\mathbf{V}_{0,4} & \mathbf{V}_{1,4} & \mathbf{V}_{2,4} & \mathbf{V}_{3,4} \\
\mathbf{V}_{0,3} & \mathbf{V}_{1,3} & \mathbf{V}_{2,3} & \mathbf{V}_{3,3} \\
\mathbf{V}_{0,2} & \mathbf{V}_{1,2} & \mathbf{V}_{2,2} & \mathbf{V}_{3,2} \\
\mathbf{V}_{0,1} & \mathbf{V}_{1,1} & \mathbf{V}_{2,1} & \mathbf{V}_{3,1} \\
\mathbf{V}_{0,0} & \mathbf{V}_{1,0} & \mathbf{V}_{2,0} & \mathbf{V}_{3,0}
\end{array} \ .$$

(The generalization for an arbitrary patch is straightforward.) Since the $b_r(u)$ and $b_s(v)$ are now functions of $\beta 1(u,v)$ and $\beta 2(u,v)$, we write equation (135) for $\mathbf{Q}_{3,4}$ as

$$\mathbf{Q}_{3,4}(u,v) = \qquad (143)$$

$$[\,\mathbf{V}_{0,4}b_{-3}(\beta_1,\beta_2;u) + \mathbf{V}_{1,4}b_{-2}(\beta_1,\beta_2;u) + \mathbf{V}_{2,4}b_{-1}(\beta_1,\beta_2;u) + \mathbf{V}_{3,4}b_{-0}(\beta_1,\beta_2;u)\,]\ b_{-0}(\beta_1,\beta_2;v) \quad +$$

$$[\,\mathbf{V}_{0,3}b_{-3}(\beta_1,\beta_2;u) + \mathbf{V}_{1,3}b_{-2}(\beta_1,\beta_2;u) + \mathbf{V}_{2,3}b_{-1}(\beta_1,\beta_2;u) + \mathbf{V}_{3,3}b_{-0}(\beta_1,\beta_2;u)\,]\ b_{-1}(\beta_1,\beta_2;v) \quad +$$

$$[\,\mathbf{V}_{0,2}b_{-3}(\beta_1,\beta_2;u) + \mathbf{V}_{1,2}b_{-2}(\beta_1,\beta_2;u) + \mathbf{V}_{2,2}b_{-1}(\beta_1,\beta_2;u) + \mathbf{V}_{3,2}b_{-0}(\beta_1,\beta_2;u)\,]\ b_{-2}(\beta_1,\beta_2;v) \quad +$$

$$[\,\mathbf{V}_{0,1}b_{-3}(\beta_1,\beta_2;u) + \mathbf{V}_{1,1}b_{-2}(\beta_1,\beta_2;u) + \mathbf{V}_{2,1}b_{-1}(\beta_1,\beta_2;u) + \mathbf{V}_{3,1}b_{-0}(\beta_1,\beta_2;u)\,]\ b_{-3}(\beta_1,\beta_2;v) \ .$$

$\mathbf{Q}_{3,3}$ is similarly defined by

$$\mathbf{Q}_{3,3}(u,v) = \tag{144}$$

$$[\mathbf{V}_{0,3} b_{-3}(\beta_1,\beta_2;u) + \mathbf{V}_{1,3} b_{-2}(\beta_1,\beta_2;u) + \mathbf{V}_{2,3} b_{-1}(\beta_1,\beta_2;u) + \mathbf{V}_{3,3} b_{-0}(\beta_1,\beta_2;u)]\, b_{-0}(\beta_1,\beta_2;v) \quad +$$

$$[\mathbf{V}_{0,2} b_{-3}(\beta_1,\beta_2;u) + \mathbf{V}_{1,2} b_{-2}(\beta_1,\beta_2;u) + \mathbf{V}_{2,2} b_{-1}(\beta_1,\beta_2;u) + \mathbf{V}_{3,2} b_{-0}(\beta_1,\beta_2;u)]\, b_{-1}(\beta_1,\beta_2;v) \quad +$$

$$[\mathbf{V}_{0,1} b_{-3}(\beta_1,\beta_2;u) + \mathbf{V}_{1,1} b_{-2}(\beta_1,\beta_2;u) + \mathbf{V}_{2,1} b_{-1}(\beta_1,\beta_2;u) + \mathbf{V}_{3,1} b_{-0}(\beta_1,\beta_2;u)]\, b_{-2}(\beta_1,\beta_2;v) \quad +$$

$$[\mathbf{V}_{0,0} b_{-3}(\beta_1,\beta_2;u) + \mathbf{V}_{1,0} b_{-2}(\beta_1,\beta_2;u) + \mathbf{V}_{2,0} b_{-1}(\beta_1,\beta_2;u) + \mathbf{V}_{3,0} b_{-0}(\beta_1,\beta_2;u)]\, b_{-3}(\beta_1,\beta_2;v) \quad .$$

We shall discuss the behaviour of these patches at their common ("horizontal") boundary, which is $\mathbf{Q}_{3,4}(u,0)$ and $\mathbf{Q}_{3,3}(u,1)$. (The argument for common "vertical" boundaries is analogous, and is therefore omitted.)

First, of course, we must verify that the curves $\mathbf{Q}_{3,4}(u,0)$ and $\mathbf{Q}_{3,3}(u,1)$ are actually identical. For any fixed $u$ we may rewrite (143) and (144) as

$$\mathbf{Q}_{bot}(v) = \mathbf{W}_0 b_{-3}(v) + \mathbf{W}_1 b_{-2}(v) + \mathbf{W}_2 b_{-1}(v) + \mathbf{W}_3 b_{-0}(v) \tag{145}$$

and

$$\mathbf{Q}_{top}(v) = \mathbf{W}_1 b_{-3}(v) + \mathbf{W}_2 b_{-2}(v) + \mathbf{W}_3 b_{-1}(v) + \mathbf{W}_4 b_{-0}(v) \tag{146}$$

where

$$\mathbf{W}_4 = \mathbf{V}_{0,4} b_{-3}(u) + \mathbf{V}_{1,4} b_{-2}(u) + \mathbf{V}_{2,4} b_{-1}(u) + \mathbf{V}_{3,4} b_{-0}(u) \tag{147}$$

$$\mathbf{W}_3 = \mathbf{V}_{0,3} b_{-3}(u) + \mathbf{V}_{1,3} b_{-2}(u) + \mathbf{V}_{2,3} b_{-1}(u) + \mathbf{V}_{3,3} b_{-0}(u)$$

$$\mathbf{W}_2 = \mathbf{V}_{0,2} b_{-3}(u) + \mathbf{V}_{1,2} b_{-2}(u) + \mathbf{V}_{2,2} b_{-1}(u) + \mathbf{V}_{3,2} b_{-0}(u)$$

$$\mathbf{W}_1 = \mathbf{V}_{0,1} b_{-3}(u) + \mathbf{V}_{1,1} b_{-2}(u) + \mathbf{V}_{2,1} b_{-1}(u) + \mathbf{V}_{3,1} b_{-0}(u)$$

$$\mathbf{W}_0 = \mathbf{V}_{0,0} b_{-3}(u) + \mathbf{V}_{1,0} b_{-2}(u) + \mathbf{V}_{2,0} b_{-1}(u) + \mathbf{V}_{3,0} b_{-0}(u) \quad .$$

As we have seen, along the common border $\beta_{2,3}(u,0)$ and $\beta_{2,2}(u,1)$ both reduce to $H(\beta_{2,1},\beta_{2,2};u)$. Hence the $\beta 1$ and $\beta 2$ which appear in (143) and (144) are identical, so that (145) and (146) are simply two successive segments on a uniformly-shaped Beta-spline curve. Hence $\mathbf{Q}_{bot}(1) = \mathbf{Q}_{top}(0)$, which is to say that $\mathbf{Q}_{3,3}(u,1) = \mathbf{Q}_{3,4}(u,0)$, as desired.

Tangent and curvature continuity between patches follow similarly if we apply the argument used earlier. Recall that the partial derivatives of $\beta 1(u,v)$ and $\beta 2(u,v)$ with respect to $v$ for $v=0$ and $v=1$ are zero. If we fully expand equations (143) or (144), a typical term has the form

$$\frac{c\,[\beta 1(u,v)]^m\,[\beta 2(u,v)]^n\,u^p\,v^q}{[\beta 2(u,v)] + 2[\beta 1(u,v)]^3 + 4[\beta 1(u,v)]^2 + 4[\beta 1(u,v)] + 2} \quad .$$

If we then compute the first partial derivative of this term with respect to $v$ we find, after repeated application of the product, quotient and chain rules, that the only resulting term that does not contain a product with at least one of

$$\frac{d}{dv}\,\beta 1(u,v) \quad \text{and} \quad \frac{d}{dv}\,\beta 2(u,v) \ ,$$

both of which are known to be zero by construction, is

Hence our surface has tangent continuity along its "horizontal" boundaries. The same argument works, *mutatis mutandis,* for the "vertical" boundaries as well, and generalizes to arbitrary patch boundaries, so that our surface is everywhere $G^1$ continuous.

An analogous argument suffices to establish curvature vector continuity.

Alternatively, $G^2$ continuity can be directly verified using Vaxima by evaluating the Beta-spline constraint equations if (142) is used to compute the values of $\beta_1$ and $\beta_2$. The algebra involved is, however, rather extensive...

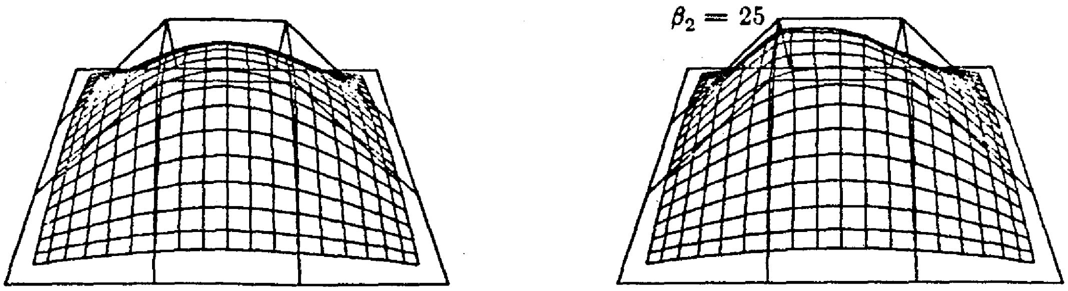


Figure 223. On the left is a Beta-spline surface in which $\beta_1 = 1$ and $\beta_2 = 0$ — a uniform bicubic B-spline surface. On the right the $\beta_2$ value at the joint corresponding to the indicated control vertex has been increased to 25. The twelve boundary vertices in the control graph have been "doubled" so as to define a total of 9 patches; otherwise the sixteen control vertices shown would define only a single patch lying close to the four central control vertices.

# 20. Discretely-Shaped Beta-splines

The continuously-shaped Beta-splines provide very local control of shape, but at significant cost. An obvious question is whether we can obtain local control of the shape parameters in strictly cubic splines. In this section we will see that this can be accomplished by suitably generalizing the one-sided functions from which we obtained the B-splines. In the "discretely-shaped Beta-splines" that result, alteration of $\beta 1$ or $\beta 2$ at a joint causes a change in the shape of four segments, as compared to the two segments of a continuously-shaped Beta-spline that are affected. In most other respects the uniformly-shaped, continuously-shaped and discretely-shaped Beta-splines display the same general behaviour.

We seek a simple and computationally efficient means: (a) to attach distinct values of $\beta 1$ and $\beta 2$ to each joint in a piecewise cubic polynomial curve, in such a way that changing a single $\beta$ parameter will alter only a local portion of the curve being defined; and (b) to generalize the uniformly-shaped Beta-splines to non-uniform knot sequences. Our approach is analogous to the development of cubic B-splines sketched earlier.

## 20.1. A Truncated Power Basis for the Beta-splines

Our first task is to define an analog of the one-sided function. The function $(\bar{u}-t)_+^3$ itself will not do, because its first and second derivatives are continuous across all knots. What we want is a function that undergoes a jump in its first and second derivatives as it crosses each knot, sufficient to satisfy the geometric continuity constraints (127), (128) and (129). Consider a function of the form

$$p(\bar{u}) \ + \ a_{i,i+1}(\bar{u}-\bar{u}_{i+1})_+^1 \ + \ b_{i,i+1}(\bar{u}-\bar{u}_{i+1})_+^2 \ .$$

(See Figure 224.) Its first and second derivatives from the left at $\bar{u}_{i+1}$ are simply $p^{(1)}(\bar{u}_i)$ and $p^{(2)}(\bar{u}_i)$. (We assume that these exist.) Its first and second derivatives from the right at $\bar{u}_{i+1}$ are

$$p^{(1)}(\bar{u}_{i+1}) + a_{i,i+1}$$

$$p^{(2)}(\bar{u}_{i+1}) + 2b_{i,i+1} \ .$$

Thus there is a jump of $a_{i,i+1}$ in the first derivative and of $2b_{i,i+1}$ in the second derivative. If we want to satisfy (128) then we must have

$$\beta_{1,i+1}p^{(1)}(\bar{u}_{i+1}) \ = \ p^{(1)}(\bar{u}_{i+1}) + a_{i,i+1}$$

or

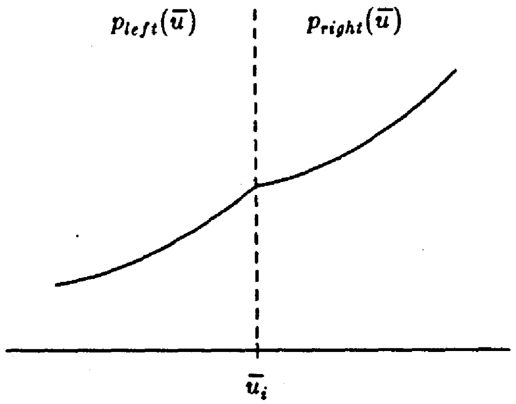


Figure 224. Adding $a(\bar{u}-\bar{u}_i)^1$ and $b(\bar{u}-\bar{u}_i)^2$ to *pleft* create a new function with discontinuities in the first and second derivative.

$$a_{i,i+1} = (\beta_{1,i+1}-1)\, p^{(1)}(\bar{u}_{i+1}) \ . \tag{148}$$

To satisfy (129) we must have

$$\beta_{1,i+1}^2 p^{(2)}(\bar{u}_{i+1}) + \beta_{2,i+1} p^{(1)}(\bar{u}_{i+1}) = p^{(2)}(\bar{u}_{i+1}) + 2b_{i,i+1}$$

or

$$b_{i,i+1} = \frac{1}{2}\Big[ (\beta_{1,i+1}^2-1) p^{(2)}(\bar{u}_{i+1}) + \beta_{2,i+1} p^{(1)}(\bar{u}_{i+1}) \Big] \ . \tag{149}$$

These equations tell us how to modify an arbitrary function so that it will satisfy our $G^2$ continuity conditions as it crosses a knot. To construct a one-sided basis for the Beta-splines, we begin with the one-sided function $(\bar{u}-\bar{u}_i)_+^3$, since it introduces the necessary third derivative discontinuity at $\bar{u}_i$, and modify it as above each time we cross a knot. Consider the function

$$\begin{aligned} g_i(\bar{u}) = (\bar{u}-\bar{u}_i)_+^3 &+ a_{i,i+1}(\bar{u}-\bar{u}_{i+1})_+^1 + \cdots + a_{i,m+3}(\bar{u}-\bar{u}_{m+3})_+^1 \\ &+ b_{i,i+1}(\bar{u}-\bar{u}_{i+1})_+^2 + \cdots + b_{i,m+3}(\bar{u}-\bar{u}_{m+3})_+^2 \ . \end{aligned}$$

Since (127), (128) and (129) will necessarily be satisfied by any linear combination of functions individually satisfying (127), (128) and (129), it is sufficient to ensure that the functions $g_i(\bar{u})$ each do so.

The function $(\bar{u}-\bar{u}_i)_+^3$ itself has zero value, as well as zero first and second derivatives, at $\bar{u}_i$ and at all knots left of $\bar{u}_i$, and so trivially satisfies our $G^2$ constraints for all $\bar{u} \le \bar{u}_i$. It is therefore sufficient to define the $a_{i,j}$ and $b_{i,j}$ from left to right, for $i < j \le m+3$, using equations (148) and (149). Thus when computing $a_{i,i+1}$ and $b_{i,i+1}$, $p(\bar{u})$ is simply $(\bar{u}-\bar{u}_i)_+^3$. The values $a_{i,i+1}$ and $b_{i,i+1}$ are given by (148) and (149). More generally, when computing $a_{i,j}$ and $b_{i,j}$, $\bar{u}$ is at the knot $\bar{u}_j$ and $p(\bar{u})$ has the value

$$(\bar{u}-\bar{u}_i)^3 + \sum_{k=i+1}^{j-1} a_{i,k}(\bar{u}-\bar{u}_k)^1 + \sum_{k=i+1}^{j-1} b_{i,k}(\bar{u}-\bar{u}_k)^2 \ ,$$

the preceding $a_i$'s and $b_i$'s having already been computed. Consequently the first derivative $p_{left}^{(1)}$ of $p(\bar{u})$ at $\bar{u}_j$ is

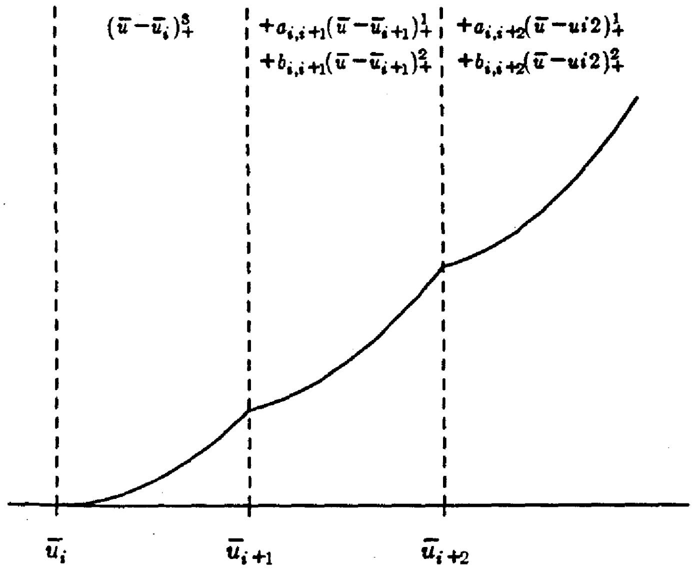


Figure 225. The function $g_i(\bar{u})$.

$$3(\bar{u}_j-\bar{u}_i)^2 + \sum_{k=i+1}^{j-1} a_{i,k} + 2\sum_{k=i+1}^{j-1} b_{i,k}(\bar{u}_j-\bar{u}_k)^1$$

and the second derivative $p^{(2)}_{\{e\}t}$ is

$$6(\bar{u}_j-\bar{u}_i)^1 + 2\sum_{k=i+1}^{j-1} b_{i,k} \quad .$$

Equations (148) and (149), with a suitable change of indices, then yield $a_{i,j}$ and $b_{i,j}$. The following algorithm computes the $a_{i,j}$ and $b_{i,j}$.

**Algorithm I**

```
for $i \leftarrow 0$ step 1 until $m+2$ do
    $Sa \leftarrow 0$
    $Sb \leftarrow 0$
    for $j \leftarrow i+1$ step 1 until $m+3$ do
        $p^{(1)}_{left} \leftarrow 3(\bar{u}_j - \bar{u}_i)^2 + Sa + 2\sum_{k=i+1}^{j-1} b_{i,k}(\bar{u}_j - \bar{u}_k)$
        $p^{(2)}_{left} \leftarrow 6(\bar{u}_j - \bar{u}_i) + Sb$
        $a_{i,j} \leftarrow (\beta_{1,j} - 1)p^{(1)}_{left}$
        $b_{i,j} \leftarrow \frac{1}{2}\Big[(\beta_{1,j}^2 - 1)p^{(2)}_{left} + \beta_{2,j}p^{(1)}_{left}\Big]$
        $Sa \leftarrow Sa + a_{i,j}$
        $Sb \leftarrow Sb + 2b_{i,j}$
    endfor
endfor
```

The outer loop steps through the $g_i(\bar{u})$ in turn. For each $g_i(\bar{u})$ the inner loop computes the $a_{i,j}$'s and $b_{i,j}$'s; *Sa* and *Sb* keep a running total of the $a_{i,j}$'s and $b_{i,j}$'s that have been computed thus far.

It is not hard to see that the functions $g_i(\bar{u})$ form a basis for the $G^2$ splines over some particular knot sequence and associated shape parameters $\beta_{1,i}$ and $\beta_{2,i}$ — the argument is very much analogous to that given in the case of $C^2$ splines for the one-sided cubics, and is therefore omitted.

### 20.2. A Local Basis for the Beta-splines

The $g_i(\bar{u})$ have the same deficiencies — namely rapid growth and non-locality — that the one-sided basis for the $C^2$ splines suffers from. The obvious next step, then, is to see whether some form of differencing can be applied to the $g_i(\bar{u})$ so as to obtain a local basis.

Just as when constructing the B-splines, the cubic term in each $g_i(\bar{u})$ is easily cancelled for $\bar{u}$ sufficiently far to the right. We need only compute

$$g_{i+1}(\bar{u}) - g_i(\bar{u}) \ , \tag{150}$$

$$g_{i+2}(\bar{u}) - g_{i+1}(\bar{u}) \ , \tag{151}$$

and so on. In order to cancel the quadratic terms in (150) and (151) by computing a further difference we need to arrange for the coefficient of $\bar{u}^2$ in (150) and (151) to have the same value. Unfortunately, these coefficients depend not only on the knot spacing (as was true for the B-splines), but also on the particular knot interval containing $\bar{u}$ since we pick up an additional $a_{i,j}$ and $b_{i,j}$ each time we move rightwards across a knot. In particular, if $\bar{u}_j \le \bar{u} < \bar{u}_{j+1}$ and $i < j$ then

$$g_i(\bar{u}) = \bar{u}^3 \quad + \bar{u}^2\Big[(b_{i,i+1} + \cdots + b_{i,j}) - 3\bar{u}_i\Big]$$

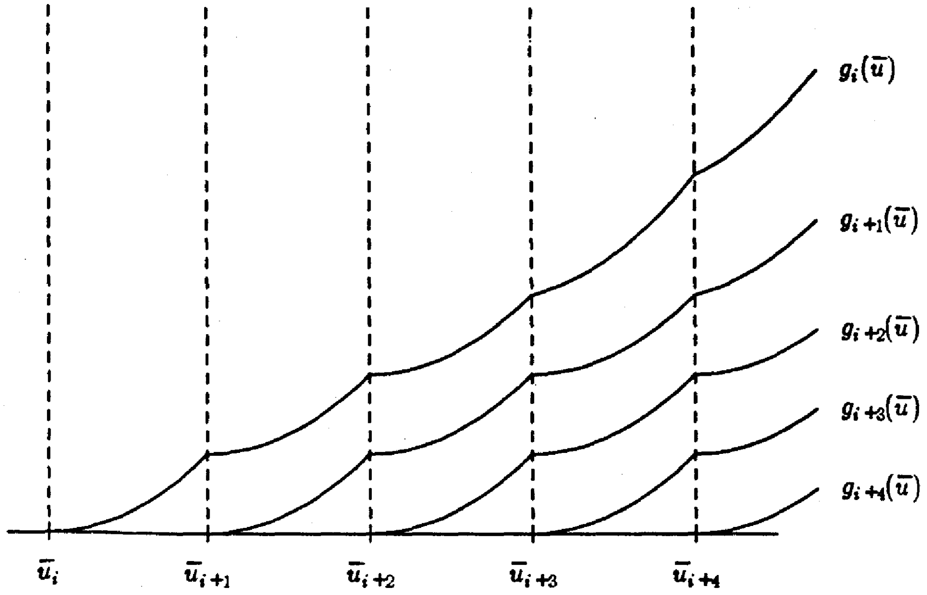


Figure 226. The basis functions shown can be combined to form a $G^2$ function which replaces $g_i(\bar{u})$.

$$\begin{aligned} + \bar{u} \Big[ &(a_{i,i+1} + \cdots + a_{i,j}) \\ &\qquad - 2(b_{i,i+1}\bar{u}_{i+1} + \cdots + b_{i,j}\bar{u}_j) + 3\bar{u}_i^2 \Big] \\ + \Big[ &(b_{i,i+1}\bar{u}_{i+1}^2 + \cdots + b_{i,j}\bar{u}_j^2) \\ &\qquad - (a_{i,i+1}\bar{u}_{i+1} + \cdots + a_{i,j}\bar{u}_j) - \bar{u}_i^3 \Big] , \end{aligned}$$

while the one-sided basis used for the B-splines is simply

$$\begin{aligned} (\bar{u} - \bar{u}_i)^3 = \bar{u}^3 \quad &+ \bar{u}^2 \Big[ -3\bar{u}_i \Big] \\ &+ \bar{u} \Big[ + 3\bar{u}_i^2 \Big] \\ &+ \Big[ - \bar{u}_i^3 \Big] \end{aligned}$$

for all $\bar{u} > \bar{u}_i$. Thus for $(\bar{u} - \bar{u}_i)_+^3$ the coefficient of $\bar{u}^2$ is a constant: if we divide $(\bar{u} - \bar{u}_{i+1})_+^3 - (\bar{u} - \bar{u}_i)_+^3$ by $-3(\bar{u}_{i+1} - \bar{u}_i)$ then the coefficient of the quadratic term is simply 1, no matter what the value of $i$. For $g_i(\bar{u})$ the coefficient of $\bar{u}^2$ alters each time we move rightwards across a knot. Hence we cannot divide $g_{i+1}(\bar{u}) - g_i(\bar{u})$ by any single constant and expect that the coefficient will be constant — in general it will change each time we cross a knot.

This difficulty can be overcome, however. For the B-splines we needed to take a fourth difference in order to obtain a local function, and the $B_i(\bar{u})$ became zero for $\bar{u} \geq \bar{u}_{i+4}$. At each step we arranged for the leading coefficients to be identical for $\bar{u} \geq \bar{u}_{i+4}$ so that they would cancel when performing the next difference.

For the Beta-splines, then, we will normalize the leading coefficients after each difference so that for $\bar{u}_{i+4} \leq \bar{u} < \bar{u}_{i+5}$ these coefficients will be identical. In particular, for the fourth difference they will be identically zero on this interval. From the equation for $g_i(\bar{u})$ above it is apparent that we will need the constants

$$\begin{aligned}
A_{i,i} &= \quad ( b_{i,i+1} + b_{i,i+2} + b_{i,i+3} + b_{i,i+4} ) \\
&\quad - 3\bar{u}_i \\
B_{i,i} &= \quad ( a_{i,i+1} + a_{i,i+2} + a_{i,i+3} + a_{i,i+4} ) \\
&\quad - 2( b_{i,i+1}\bar{u}_{i+1} + b_{i,i+2}\bar{u}_{i+2} + b_{i,i+3}\bar{u}_{i+3} + b_{i,i+4}\bar{u}_{i+4} ) \\
&\quad + 3\bar{u}_i^2 \\
C_{i,i} &= \quad ( b_{i,i+1}\bar{u}_{i+1}^2 + b_{i,i+2}\bar{u}_{i+2}^2 + b_{i,i+3}\bar{u}_{i+3}^2 + b_{i,i+4}\bar{u}_{i+4}^2 ) \\
&\quad - ( a_{i,i+1}\bar{u}_{i+1} + a_{i,i+2}\bar{u}_{i+2} + a_{i,i+3}\bar{u}_{i+3} + a_{i,i+4}\bar{u}_{i+4} ) \\
&\quad - \bar{u}_i^3
\end{aligned}$$

$$\begin{aligned}
A_{i,i+1} &= \quad ( b_{i+1,i+2} + b_{i+1,i+3} + b_{i+1,i+4} ) \\
&\quad - 3\bar{u}_{i+1} \\
B_{i,i+1} &= \quad ( a_{i+1,i+2} + a_{i+1,i+3} + a_{i+1,i+4} ) \\
&\quad - 2( b_{i+1,i+2}\bar{u}_{i+2} + b_{i+1,i+3}\bar{u}_{i+3} + b_{i+1,i+4}\bar{u}_{i+4} ) \\
&\quad + 3\bar{u}_{i+1}^2 \\
C_{i,i+1} &= \quad ( b_{i+1,i+2}\bar{u}_{i+2}^2 + b_{i+1,i+3}\bar{u}_{i+3}^2 + b_{i+1,i+4}\bar{u}_{i+4}^2 ) \\
&\quad - ( a_{i+1,i+2}\bar{u}_{i+2} + a_{i+1,i+3}\bar{u}_{i+3} + a_{i+1,i+4}\bar{u}_{i+4} ) \\
&\quad - \bar{u}_{i+1}^3
\end{aligned}$$

$\dots$

$$\begin{aligned}
A_{i,i+4} &= -3\bar{u}_{i+4} \\
B_{i,i+4} &= +3\bar{u}_{i+4}^2 \\
C_{i,i+4} &= -\bar{u}_{i+4}^3 \quad .
\end{aligned}$$

Then we may write

$$g_j(\bar{u}) = \bar{u}^3 + A_{i,i}\bar{u}^2 + B_{i,i}\bar{u} + C_{i,i} \qquad \text{for } \bar{u}_{i+4} \leq \bar{u} < \bar{u}_{i+5} \quad .$$

Similarly we have

$$\begin{aligned}
g_{i+1}(\bar{u}) &= \bar{u}^3 + A_{i,i+1}\bar{u}^2 + B_{i,i+1}\bar{u} + C_{i,i+1} &&\text{for } \bar{u}_{i+4} \leq \bar{u} < \bar{u}_{i+6} \\
g_{i+2}(\bar{u}) &= \bar{u}^3 + A_{i,i+2}\bar{u}^2 + B_{i,i+2}\bar{u} + C_{i,i+2} &&\text{for } \bar{u}_{i+4} \leq \bar{u} < \bar{u}_{i+6} \\
g_{i+3}(\bar{u}) &= \bar{u}^3 + A_{i,i+3}\bar{u}^2 + B_{i,i+3}\bar{u} + C_{i,i+3} &&\text{for } \bar{u}_{i+4} \leq \bar{u} < \bar{u}_{i+5} \\
g_{i+4}(\bar{u}) &= \bar{u}^3 + A_{i,i+4}\bar{u}^2 + B_{i,i+4}\bar{u} + C_{i,i+4} &&\text{for } \bar{u}_{i+4} \leq \bar{u} < \bar{u}_{i+6} \quad .
\end{aligned}$$

From these we form the four functions $\Delta_i^1 g_i(\bar{u})$, $\Delta_i^1 g_{i+1}(\bar{u})$, $\Delta_i^1 g_{i+2}(\bar{u})$ and $\Delta_i^1 g_{i+3}(\bar{u})$ defined by

$$\begin{aligned}
\Delta_i^1 g_j(\bar u) &= \frac{g_{j+1}(\bar u) - g_j(\bar u)}{A_{i,j+1} - A_{i,j}} && \text{for all } \bar u \text{ and } j = i,\, i+1,\, i+2,\, i+3 \\
&= \bar u^2 + \frac{B_{i,j+1} - B_{i,j}}{A_{i,j+1} - A_{i,j}}\bar u + \frac{C_{i,j+1} - C_{i,j}}{A_{i,j+1} - A_{i,j}} && \text{for } \bar u_{i+4} \le \bar u < \bar u_{i+6} \\
&= \bar u^2 + D_{i,j}\bar u + E_{i,j} \ ,
\end{aligned}$$

thus implicitly defining the $D_{i,j}$ and $E_{i,j}$. The index $i$ with which we subscript $\Delta$ reminds us that we are eventually going to replace $g_i(\bar u)$ with an appropriate linear combination $G_i(\bar u)$ of $g_i(\bar u)$, $g_{i+1}(\bar u)$, $g_{i+2}(\bar u)$, $g_{i+3}(\bar u)$ and $g_{i+4}(\bar u)$, computed in such a way as to ensure that $G_i(\bar u)$ will be zero on $\bar u_{i+4} \le \bar u < \bar u_{i+5}$.

We can now cancel the quadratic term by forming the three functions $\Delta_i^2 g_i(\bar u)$, $\Delta_i^2 g_{i+1}(\bar u)$ and $\Delta_i^2 g_{i+2}(\bar u)$ as

$$\begin{aligned}
\Delta_i^2 g_j(\bar u) &= \frac{\Delta_i^1 g_{j+1}(\bar u) - \Delta_i^1 g_j(\bar u)}{D_{i,j+1} - D_{i,j}} && \text{for all } \bar u \text{ and } j = i,\, i+1,\, i+2 \\
&= \bar u + \frac{E_{i,j+1} - E_{i,j}}{D_{i,j+1} - D_{i,j}} && \text{for } \bar u_{i+4} \le \bar u < \bar u_{i+5} \\
&= \bar u + F_{i,j} && \text{for } \bar u_{i+4} \le \bar u < \bar u_{i+5}
\end{aligned}$$

and then cancel the linear term by forming the two functions $\Delta_i^3 g_i(\bar u)$ and $\Delta_i^3 g_{i+1}(\bar u)$ as

$$\begin{aligned}
\Delta_i^3 g_j(\bar u) &= \frac{\Delta_i^2 g_{j+1}(\bar u) - \Delta_i^2 g_i(\bar u)}{F_{i,j+1} - F_{i,j}} && \text{for all } \bar u \text{ and } j = i,\, i+1 \\
&= 1 && \text{for } \bar u_{i+4} \le \bar u < \bar u_{i+5} \ .
\end{aligned}$$

Finally we compute the function

$$\Delta_i^4 g_i(\bar u) = -\left[\Delta_i^3 g_{i+1}(\bar u) - \Delta_i^3 g_i(\bar u)\right] \qquad \text{for } \bar u_{i+4} \le \bar u < \bar u_{i+5} \ ,$$

with which we replace $g_i(\bar u)$. The pattern of this computation is shown in the following diagram.

$$\begin{array}{ccccccccc}
g_i(\bar u) & & g_{i+1}(\bar u) & & g_{i+2}(\bar u) & & g_{i+3}(\bar u) & & g_{i+4}(\bar u) \\
& \Delta_i^1 g_i(\bar u) & & \Delta_i^1 g_{i+1}(\bar u) & & \Delta_i^1 g_{i+2}(\bar u) & & \Delta_i^1 g_{i+3}(\bar u) & \\
& & \Delta_i^2 g_i(\bar u) & & \Delta_i^2 g_{i+1}(\bar u) & & \Delta_i^2 g_{i+2}(\bar u) & & \\
& & & \Delta_i^3 g_i(\bar u) & & \Delta_i^3 g_{i+1}(\bar u) & & & \\
& & & & \Delta_i^4 g_i(\bar u) & & & &
\end{array}$$

Now $\Delta_i^4 g_i(\bar u)$ is defined for any value of $\bar u$, but we have only ensured that it is zero when $\bar u$ lies between $\bar u_{i+4}$ and $\bar u_{i+5}$, or is less than $\bar u_i$. To arrange for locality we simply define our ***discretely-shaped Beta-splines*** $G_i(\bar u)$ to be

$$G_i(\bar{u}) = \begin{cases} 0 & \bar{u} < \bar{u}_i \ \ or \ \ \bar{u} \geq \bar{u}_{i+4} \\ \\ \Delta_i^4 g_i(\bar{u}) & \bar{u}_i \leq \bar{u} < \bar{u}_{i+4} \ . \end{cases}$$

Since by construction $\Delta_i^4 g_i(\bar{u})$ is zero on $[\bar{u}_4,\bar{u}_5)$, the rightwards extension by zero leaves us with a function satisfying the $G^2$ continuity constraints.

We have still to argue that the $G_i(\bar{u})$ can be used as a basis for $G^2$ polynomial splines. Since each of the $G_i(\bar{u})$ is a linear combination of $G^2$ functions, each of the $G_i(\bar{u})$ is a $G^2$ function, and there are as many $G_i(\bar{u})$ in $[\bar{u}_0,\bar{u}_m)$ as there are $g_i(\bar{u})$. If we consider the $G_i(\bar{u})$ in turn from left to right, each is nonzero on an interval for which all the $G_i(\bar{u})$ to the left are zero, and it is therefore plausible that no single $G_i(\bar{u})$ can be written as a linear combination of basis functions lying to the left, so that the $G_i(\bar{u})$ are linearly independent. This argument breaks down only as we approach the end of the curve; a rigourous proof of linear independence may be found in [Bartels84].

*Discretely-shaped Beta-spline curves* are now defined by

$$\mathbf{Q}(\bar{u}) = \sum_i \mathbf{V}_i G_i(\bar{u}) = \sum_i (\ x_i G_i(\bar{u}),\ y_i G_i(\bar{u})\ )\ . \tag{152}$$

The $i^{\text{th}}$ curve segment is

$$\mathbf{Q}_i = \sum_{r=-3}^{0} \mathbf{V}_{i+r} G_{i+r}(\bar{u})$$

$$= \mathbf{V}_{i-3}G_{i-3}(\bar{u}) + \mathbf{V}_{i-2}G_{i-2}(\bar{u}) + \mathbf{V}_{i-1}G_{i-1}(\bar{u}) + \mathbf{V}_i G_i(\bar{u})\ .$$

## 20.3. Evaluation

For the $C^2$ splines we defined basis functions $B_{i,k}(\bar{u})$ of arbitrary order $k$, and developed a recursive definition of $B_{i,k}(\bar{u})$ in terms of $B_{i,k-1}(\bar{u})$ and $B_{i+1,k-1}(\bar{u})$. This provided both an efficient means of computing the value of a basis function, and of computing its derivatives. Given the latter one can then develop an efficient algorithm for converting from a "control vertex" representation such as (152) to a power representation

$$c_0 + c_1(\bar{u}-\bar{u}_i)^1 + c_2(\bar{u}-\bar{u}_i)^2 + c_3(\bar{u}-\bar{u}_i)^3\ ,$$

from which one can efficiently compute points along a curve segment by using forward differences. Unfortunately we have not been able to develop such a recursive definition of the Beta-splines, and indeed we rather doubt that a natural such definition exists. This is not, however, a fatal obstacle. One can simply pre-compute the coefficients $a_{i,j}$, $b_{i,j}$, $A_{i,j}$, $B_{i,j}$, $C_{i,j}$, $D_{i,j}$, $E_{i,j}$, $F_{i,j}$ and then compute the difference $\Delta_i^4 g_i(\bar{u})$ directly whenever a point on the curve is required. Doing so does not require an $a_{i,j}$ or $b_{i,j}$ for any value of $j$ other than $i+1$, $i+2$, $i+3$ or $i+4$. Hence Algorithm I can be made somewhat more efficient by replacing the expression $m+3$ in line 4 by $\min(i+4,m+3)$.

Moreover, since differencing and differentiation commute, we may compute derivatives of the $G_i(\bar{u})$ by differencing derivatives of the $g_i(\bar{u})$, and so obtain a power representation of the basis segments that can be evaluated by using Horner's rule or forward differences.

### 20.4. Properties

Practically speaking the most important properties of spline basis functions are summation to one (because this ensures translation invariance) and positivity (because together with summation to one this provides a convex hull property). From a theoretical point of view it is necessary to show that the basis functions we have constructed are, in fact, linearly independent. Positivity and linear independence are established in [Bartels84]. The arguments are somewhat involved as yet, and we shall not repeat them here.

Summation to one is both easier and harder. Because the constant polynomial 1 is trivially a $G^2$ spline for every knot sequence, no matter what values of $\beta 1$ and $\beta 2$ are required at the joints, it is clear that scale factors $c_i$ exist such that $\sum c_i G_i(\bar{u}) = 1$.[6] We are certain from computational experience, although we do not yet have a proof, that in fact $c_i = 1$ for all $i$.

### 20.5. Locality

Consider a $\beta$ value at the joint corresponding to the knot $\bar{u}_i$ (See Figure 227). By construction it is clear that no basis function prior to $G_{i-4}(\bar{u})$ or subsequent to $G_i(\bar{u})$ could possibly be affected by a change in $\beta_{1,i}$ or $\beta_{2,i}$ because no use is made of them in the one-sided functions from which the $G_j(\bar{u})$ are formed. Hence we know immediately that the effect of changing $\beta$ values at $\bar{u}_i$ must be restricted at least to the eight intervals comprising $[\bar{u}_{i-4}, \bar{u}_{i+4})$.

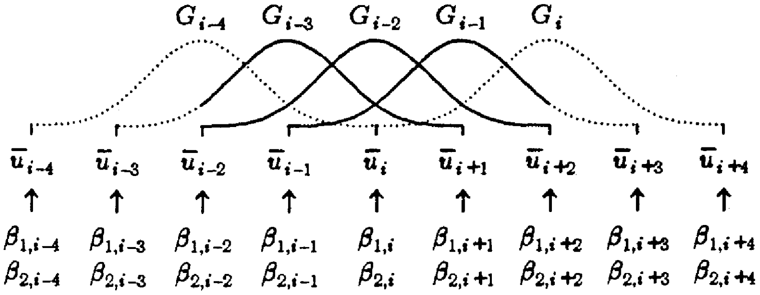

$G_{i-4}$
$G_{i-3}$
$G_{i-2}$
$G_{i-1}$
$G_i$
$\bar{u}_{i-4}$
$\bar{u}_{i-3}$
$\bar{u}_{i-2}$
$\bar{u}_{i-1}$
$\bar{u}_i$
$\bar{u}_{i+1}$
$\bar{u}_{i+2}$
$\bar{u}_{i+3}$
$\bar{u}_{i+4}$
$\beta_{1,i-4}$ $\beta_{2,i-4}$
$\beta_{1,i-3}$ $\beta_{2,i-3}$
$\beta_{1,i-2}$ $\beta_{2,i-2}$
$\beta_{1,i-1}$ $\beta_{2,i-1}$
$\beta_{1,i}$ $\beta_{2,i}$
$\beta_{1,i+1}$ $\beta_{2,i+1}$
$\beta_{1,i+2}$ $\beta_{2,i+2}$
$\beta_{1,i+3}$ $\beta_{2,i+3}$
$\beta_{1,i+4}$ $\beta_{2,i+4}$

Figure 227. The basis segments that are affected by a change in $\beta 1$ or $\beta 2$ at $\bar{u}_i$.

We can show substantially greater locality if we assume that

$$G_{i-3}(\bar{u}) + G_{i-2}(\bar{u}) + G_{i-1}(\bar{u}) + G_i(\bar{u}) = 1$$

on $[\bar{u}_i, \bar{u}_{i+1})$ without the need of further scale factors, as is almost certainly the case. We make this assumption throughout the remainder of this subsection.

It is clear from our construction that no use is made of $\beta_{1,i}$ or $\beta_{2,i}$ in constructing $G_i(\bar{u})$ and $G_i(\bar{u})$ must therefore be independent of $\beta_{1,i}$ and $\beta_{2,i}$. With somewhat more effort we can also show that $G_{i-4}(\bar{u})$ is independent of $\beta_{1,i}$ and $\beta_{2,i}$.

- We are assuming that

$$G_{i-7}(\bar{u}_{i-3}) + G_{i-6}(\bar{u}_{i-3}) + G_{i-5}(\bar{u}_{i-3}) + G_{i-4}(\bar{u}_{i-3}) = 1 \ .$$

- We already know that $G_{i-7}(\bar{u})$, $G_{i-6}(\bar{u})$ and $G_{i-5}(\bar{u})$ are independent of $\beta_{1,i}$ and $\beta_{2,i}$.

[6] We are indebted to Tony DeRose for pointing this out.

- Hence $G_{i-4}(\overline{u}_{i-3})$ has some fixed value $K$, regardless of the value assigned to $\beta_{1,i}$ or $\beta_{2,i}$.
- But $G_{i-4}(\overline{u})$ is composed of the four segment polynomials $s_{i-4,-0}(\overline{u})$, $s_{i-4,-1}(\overline{u})$, $s_{i-4,-2}(\overline{u})$ and $s_{i-4,-3}(\overline{u})$ (having sixteen coefficients) which are the necessarily unique solution to the fifteen equations obtained by applying the constraints (127), (128) and (129) at $\overline{u}_{i-4}$, $\overline{u}_{i-3}$, $\overline{u}_{i-2}$, $\overline{u}_{i-1}$ and $\overline{u}_i$ together with the (scaling) constraint that $s_{i-4,-1}(\overline{u}_{i-3}) = K$.
- Hence $G_{i-4}(\overline{u})$ cannot change for any value of $\overline{u}$ if $\beta_{1,i}$ or $\beta_{2,i}$ is altered.

Since neither $G_{i-4}(\overline{u})$ not $G_i(\overline{u})$ is dependent on the shape parameters at $\overline{u}_i$, we may conclude that the effect of changing these values at $\overline{u}_i$ must be restricted at least to the six intervals comprising $[\overline{u}_{i-3},\overline{u}_{i+3})$.

By a similar argument we can easily show that $G_{i-3}(\overline{u})$ is independent of $\beta_{1,i}$ and $\beta_{2,i}$ on $[\overline{u}_{i-3},\overline{u}_{i-2})$, and that $G_{i-1}(\overline{u})$ is independent of $\beta_{1,i}$ and $\beta_{2,i}$ on $[\overline{u}_{i+2},\overline{u}_{i+3})$. Finally, then, we conclude that the affect of changing $\beta$ values at $\overline{u}_i$ is restricted to the four intervals comprising $[\overline{u}_{i-2},\overline{u}_{i+2})$, under the assumption that the $G_i(\overline{u})$ sum to one without further scaling.

Thus the amount of re-computation required by the change of a shape parameter is independent of the number of control vertices defining the curve, and the change in shape is local.

## 20.6. Uniform Cubic Discretely-Shaped Beta-splines

Analyzing the properties of the discretely-shaped Beta-splines, and the rendering of discretely-shaped Beta-spline curves, would be facilitated if we could obtain a compact symbolic representation of their segments. Unfortunately we have not, as yet, been able to do so for non-uniform knot sequences, and it seems likely that any such representation will prove to be quite complicated. However, we have been able to directly analyze the discretely-shaped Beta-splines over a uniform knot sequence.

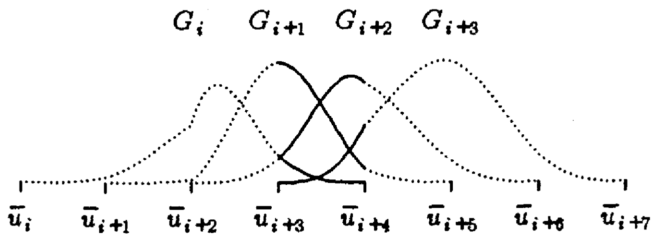


Figure 228. The four basis functions $G_i(\overline{u})$, $G_{i+1}(\overline{u})$, $G_{i+2}(\overline{u})$ and $G_{i+3}(\overline{u})$ that are nonzero on the particular interval $[\overline{u}_{i+3},\overline{u}_{i+4})$. The knots are equally spaced, but a variety of different $\beta 1$ and $\beta 2$ values have been used.

Assume that the underlying knot sequence is uniform, so that $\overline{u}_i - \overline{u}_{i-1} = 1$ for all $i$. If, for an arbitrary set of positive $\beta_{1,i}$'s and $\beta_{2,i}$'s, we compute the four segments that are non-zero on the interval $[\overline{u}_{i+3},\overline{u}_{i+4})$ (see Figure 228) and sum them using Vaxima, we find that they sum to one. We can show directly that they are non-negative. Two of the basis segments are trivial. We find that

$$s_{i+3,-0}(\overline{u}) = \frac{1}{\delta 1}\left[2(\beta_{2,i+5}+2\beta_{1,i+5}^2+2\beta_{1,i+5})u^3\right]$$

where

$$\delta 1 = \beta_{2,i+4}\beta_{2,i+5} + 2\beta_{1,i+4}^2\beta_{2,i+5} + 4\beta_{1,i+4}\beta_{2,i+5} + 2\beta_{2,i+5}$$

$$+ 2\beta^3_{1,i+5}\beta_{2,i+4} + 4\beta^2_{1,i+5}\beta_{2,i+4} + 2\beta_{1,i+5}\beta_{2,i+4} + 4\beta^2_{1,i+4}\beta^3_{1,i+5}$$
$$+ 4\beta_{1,i+4}\beta^3_{1,i+5} + 8\beta^2_{1,i+4}\beta^2_{1,i+5} + 12\beta_{1,i+4}\beta^2_{1,i+5} + 4\beta^2_{1,i+5}$$
$$+ 4\beta^2_{1,i+4}\beta_{1,i+5} + 8\beta_{1,i+4}\beta_{1,i+5} + 4\beta_{1,i+5}$$

and

$$s_{i+3,-3}(\overline{u}) = \frac{1}{\delta 4}\left[2\beta^3_{1,i+3}(\beta_{2,i+2} + 2\beta^2_{1,i+2} + 2\beta_{1,i+2})(1-u)^3\right]$$

where

$$\delta 4 = \beta_{2,i+2}\beta_{2,i+3} + 2\beta^2_{1,i+2}\beta_{2,i+3} + 4\beta_{1,i+2}\beta_{2,i+3} + 2\beta_{2,i+3}$$
$$+ 2\beta^3_{1,i+3}\beta_{2,i+2} + 4\beta^2_{1,i+3}\beta_{2,i+2} + 2\beta_{1,i+3}\beta_{2,i+2} + 4\beta^2_{1,i+2}\beta^3_{1,i+3}$$
$$+ 4\beta_{1,i+2}\beta^3_{1,i+3} + 8\beta^2_{1,i+2}\beta^2_{1,i+3} + 12\beta_{1,i+2}\beta^2_{1,i+3} + 4\beta^2_{1,i+3}$$
$$+ 4\beta^2_{1,i+2}\beta_{1,i+3} + 8\beta_{1,i+2}\beta_{1,i+3} + 4\beta_{1,i+3} \ .$$

It is easy to see that these two basis segments will be positive since all the $\beta$'s are positive and $0 \le u < 1$. The denominators $\delta 1$ and $\delta 4$ can be factored further, but we have left them in this form for simplicity.

The remaining two segments require more effort. The segment $s_{i+3,-1}(\overline{u})$ may be written as

$$s_{i+3,-1}(u) = \left(c_0 + c_1 u + c_2 u^2\right) - \left(c_3 u^3\right) \ . \tag{153}$$

The following argument establishes that $s_{i+3,-1}(u)$ is positive on (0,1):

- $c_1$, $c_2$, $c_3$, $s_{i+3,-1}(0) = c_0$ and $s_{i+3,-2}(1) = c_0 + c_1 + c_2 - c_3 \equiv d_3$ are all sums of products of positive values, like $\delta 1$ and $\delta 4$, and are therefore themselves positive;
- hence we may represent $c_3$ as $c_0 + c_1 + c_2 - d_3$;
- since $0 < u < 1$, we have $1 > u > u^2 > u^3$;
- therefore
$$\begin{array}{lcl} c_0 & > & c_0 u^3 \\ c_1 u & > & c_1 u^3 \\ c_2 u^2 & > & c_2 u^3 \\ 0 & > & -d_3 u^3 \ ; \end{array}$$
- therefore
$$c_0 + c_1 u + c_2 u^2 + 0 > \left(c_0 + c_1 + c_2 - d_3\right) u^3 = c_3 u^3 \ ;$$
- therefore $s_{i+3,-1}(u)$ is positive on (0,1), as desired.

An exactly analogous argument suffices for the right middle segment if it is written in the form

$$s_{i+3,-2}(u) = \left(c_0 + c_1(1-u) + c_2(1-u)^2\right) - \left(c_3(1-u)^3\right) \ .$$

A variety of important properties follow from the fact that the discretely-shaped cubic Beta-splines are non-negative and sum to one:

- the $i^{\text{th}}$ segment $\mathbf{Q}_i$ lies within the convex hull of $\mathbf{V}_{i-3}$, $\mathbf{V}_{i-2}$, $\mathbf{V}_{i-1}$ and $\mathbf{V}_i$;
- if $\mathbf{V}_{i-3} = \mathbf{V}_{i-2} = \mathbf{V}_{i-1}$ then this point will be interpolated, and the curve segment defined by these

three points and $\mathbf{V}_i$ will be a straight line;

- if $\mathbf{V}_{i-3}=\mathbf{V}_{i-2}$ then the first point on the curve segment defined by these two points, together with $\mathbf{V}_{i-1}$ and $\mathbf{V}_i$ must lie on the line segment joining $\mathbf{V}_{i-2}$ and $\mathbf{V}_{i-1}$ and the curvature there will be zero.

(If one assumes that the $G_i(\bar{u})$ do not need further scaling in order to sum to one, or failing that if one computes the scale factors that produce a partition of unity, then these results apply also to discretely-shaped Beta-splines over a non-uniform knot sequence.)

It is possible to verify, with the aid of Vaxima, that as $\beta_{2,i+2}$ is made arbitrarily large $\mathbf{Q}(\bar{u}_{i+2})$ converges to $\mathbf{V}_i$. This behaviour, which the uniformly-shaped and continuously-shaped Beta-splines display as well, naturally associates the joint at $\bar{u}_{i+2}$ with the control vertex $\mathbf{V}_i$, and so we sometimes speak loosely of the "$\beta 2$ value associated with $\mathbf{V}_i$" when referring to $\beta_{2,i+2}$ (and similarly for $\beta 1$).

If $\beta_{1,i}=\beta 1$ and $\beta_{2,i}=\beta 2$ for all $i$, we then obtain the uniformly-shaped Beta-spline for $\beta 1$ and $\beta 2$.

In many applications the ability to manipulate $\beta 2$ may be sufficient [Barsky83], and we therefore list the basis segments on the interval $[\bar{u}_i,\bar{u}_{i+1})$ for the special case in which the knots are spaced one unit apart and the $\beta 1$ values all have the value one. (These are the four segments drawn as solid curves in Figure 228.)

$$s_{i+3,-1}(u) = \frac{2(\beta_{2,i+2}+4)u^3}{\delta 1}$$

$$\begin{aligned} s_{i+2,-2}(u) = & -\frac{2(\beta_{2,i+1}+4)}{\delta 1\ \delta 2}\Big[\beta_{2,i}\beta_{2,i+1}\beta_{2,i+2}+8\beta_{2,i}\beta_{2,i+1}+3\beta_{2,i+1}\beta_{2,i+2}+8\beta_{2,i}\beta_{2,i+2} \\ & \qquad +44\beta_{2,i}+24\beta_{2,i+1}+28\beta_{2,i+2}+144\Big]u^3 \\ & + \frac{(\beta_{2,i+1}+4)}{\delta 2}\Big[3(\beta_{2,i}+2)u^2+6u+2\Big] \end{aligned}$$

$$\begin{aligned} s_{i+1,-3}(u) = & -\frac{2(\beta_{2,i}+4)}{\delta 2\ \delta 3}\Big[\beta_{2,i-1}\beta_{2,i}\beta_{2,i+1}+3\beta_{2,i-1}\beta_{2,i}+8\beta_{2,i}\beta_{2,i+1}+8\beta_{2,i-1}\beta_{2,i+1} \\ & \qquad +28\beta_{2,i-1}+24\beta_{2,i}+44\beta_{2,i+1}+144\Big](1-u)^3 \\ & + \frac{(\beta_{2,i}+4)}{\delta 3}\Big[3(\beta_{2,i+1}+2)(1-u)^2+6(1-u)+2\Big] \end{aligned}$$

$$s_{i,-4}(u) = \frac{2(\beta_{2,i-1}+4)(1-u)^3}{\delta 3}$$

where

$$\begin{aligned} \delta 1 &= (\beta_{2,i+1}\beta_{2,i+2}+8\beta_{2,i+1}+8\beta_{2,i+2}+48) \\ \delta 2 &= (\beta_{2,i}\beta_{2,i+1}+8\beta_{2,i}+8\beta_{2,i+1}+48) \\ \delta 3 &= (\beta_{2,i-1}\beta_{2,i}+8\beta_{2,i-1}+8\beta_{2,i}+48)\ . \end{aligned}$$

By inspection it is clear that so long as $\beta_{1,i}$ and $\beta_{2,i}$ are non-negative the above representation for discretely-shaped Beta-splines over a uniform knot sequence are necessarily well-defined — the denominators cannot vanish, even though the differencing representation of the discretely-shaped Beta-splines admits of this possibility.

## 20.7. Examples

Generally speaking the discretely-shaped Beta-splines behave much as the uniformly- and continuously-shaped Beta-splines do. Figures 229-238 illustrate this. It is illuminating to see how changes in the discretely-shaped basis functions shown in Figures 230, 232 and 236 produce the curves shown in Figures 229, 231, 233 and 235.

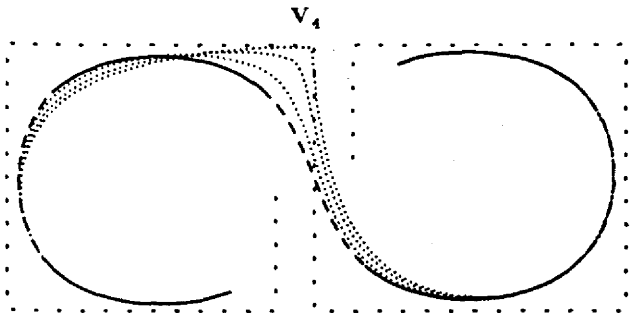


Figure 229. The solid/dashed line is a uniform cubic B-spline curve ($\beta_1$ and $\beta_2$ have the values 1 and 0 at every joint). The dotted curves result when the value of $\beta_2$ at the joint nearest $V_4$ is set to 2, 10 and 100, respectively. Increasing values of $\beta_2$ draw the joint in question towards $V_4$. For clarity the control polygon is shown, but not the control vertices.

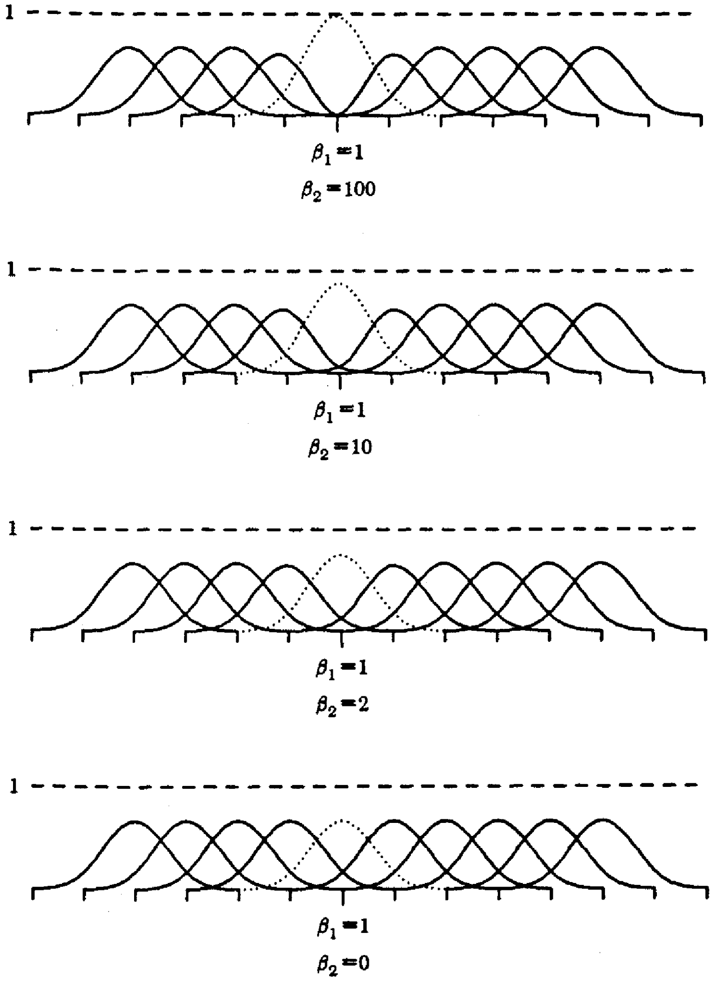


Figure 230. The discretely-shaped Beta-splines corresponding to the four curves of Figure 229. $\beta_1$ and $\beta_2$ have the values 1 and 0 at all knots except the one explicitly labeled.

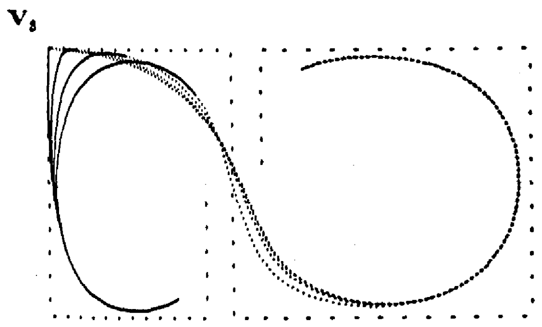


Figure 231. Starting from the same uniform cubic B-spline curve as appears in Figure 229, we successively increase $\beta_1$ at the joint between the solid and dotted portions of the curve, so that it has the values 1, 4, 16 and 256. As $\beta_1$ is increased the joint is pulled towards $V_8$.

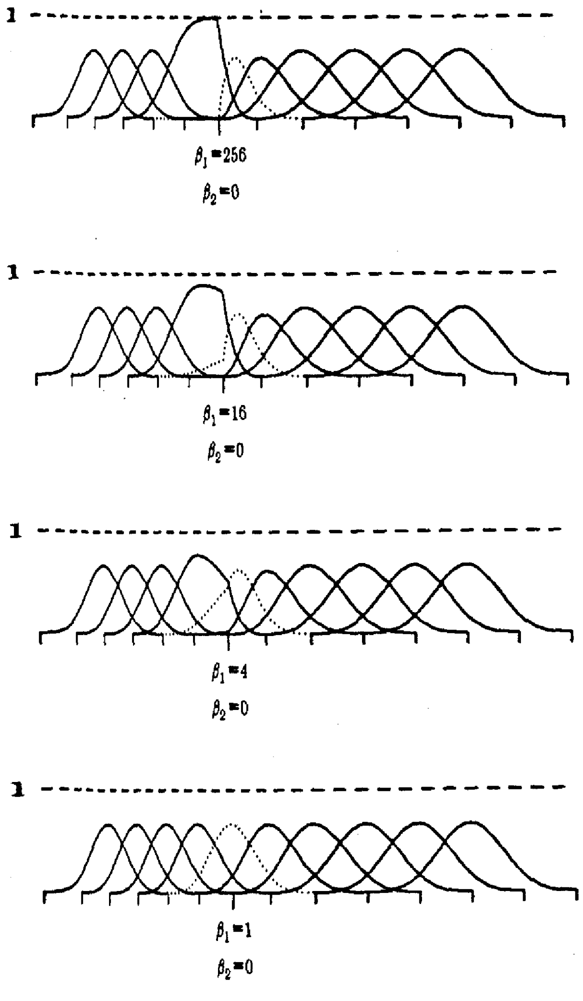


Figure 232. The discretely-shaped Beta-splines corresponding to the four curves of Figure 231. $\beta_1$ and $\beta_2$ have the values 1 and 0 at all knots except the one explicitly labeled.

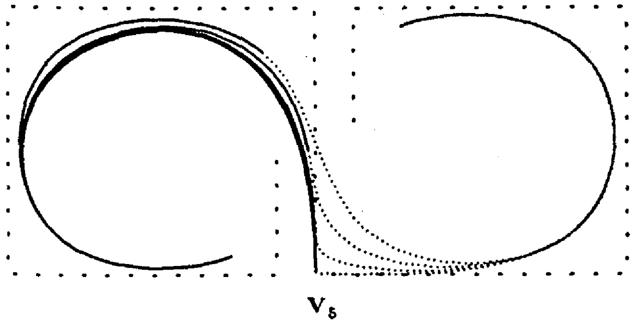


Figure 233. Symmetric behaviour occurs if we set $\beta_1$ to the values 1, 1/4, 1/16 and 1/256, respectively, with $\beta_2$=0. This time the joint is pulled towards $V_5$.

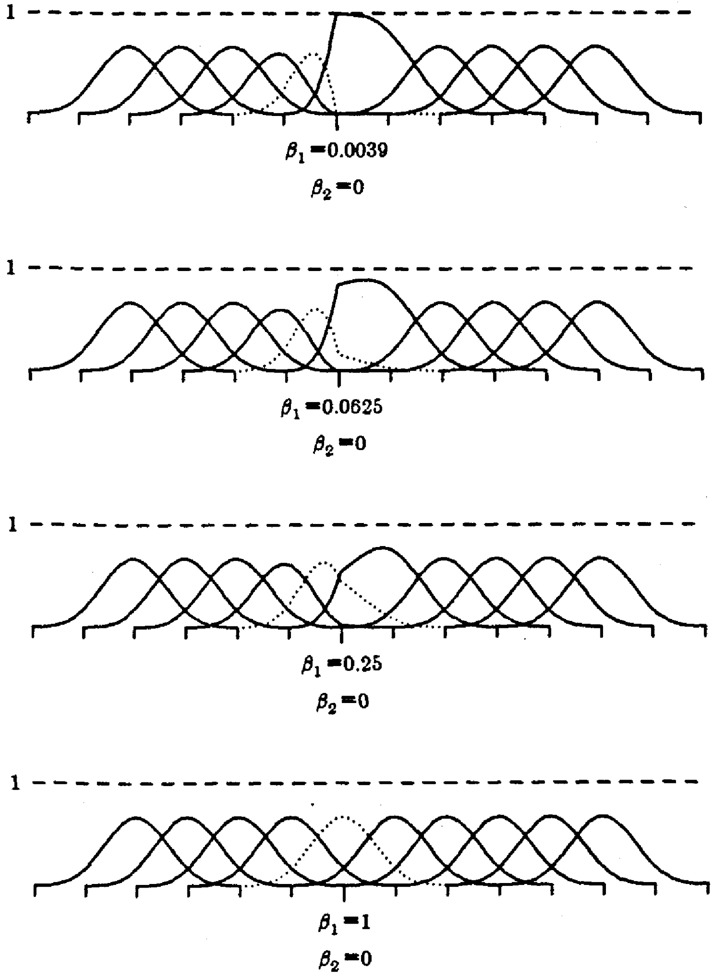


Figure 234. The discretely-shaped Beta-splines corresponding to the four curves of Figure 233. $\beta_1$ and $\beta_2$ have the values 1 and 0 at all knots except the one explicitly labeled.

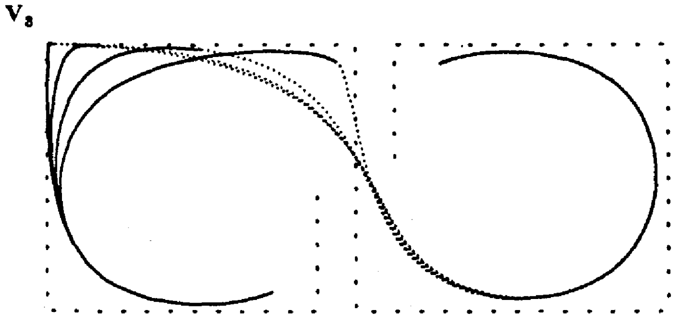


Figure 235. The $\beta_1$ values here are the same as in figure 231 except that the value of $\beta_2$ at the joint in question is 10 in each case instead of 0. Again the joint is pulled towards $\mathbf{V}_8$. Recall that increasing $\beta_2$ at that joint has the effect of pulling the curve towards $\mathbf{V}_4$.

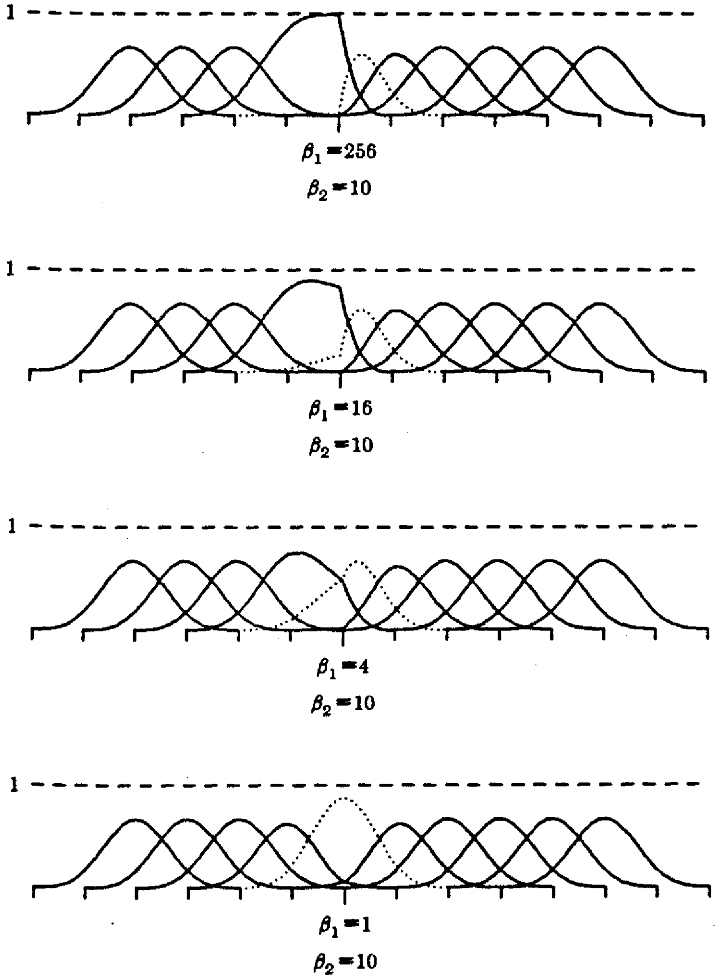


Figure 236. The discretely-shaped Beta-splines corresponding to the four curves of Figure 235. $\beta_1$ and $\beta_2$ have the values 1 and 0 at all knots except the one explicitly labeled.

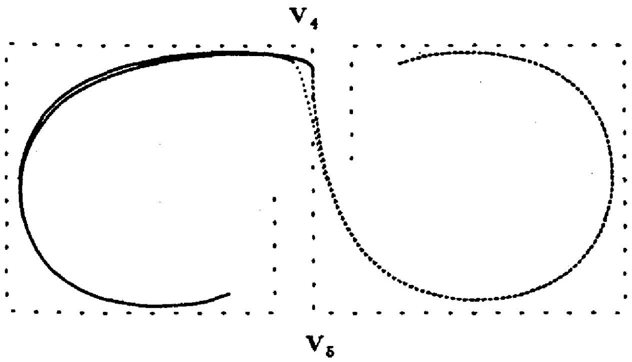


Figure 237. The $\beta_1$ values here are the same as in figure 233 except that the value of $\beta_2$ at the joint in question is 10 in each case instead of 0. Note that in this case the joint does not converge to $V_5$. Tensing the curve toward $V_4$ by setting a high value on $\beta_2$ at the joint has inhibited the convergence.

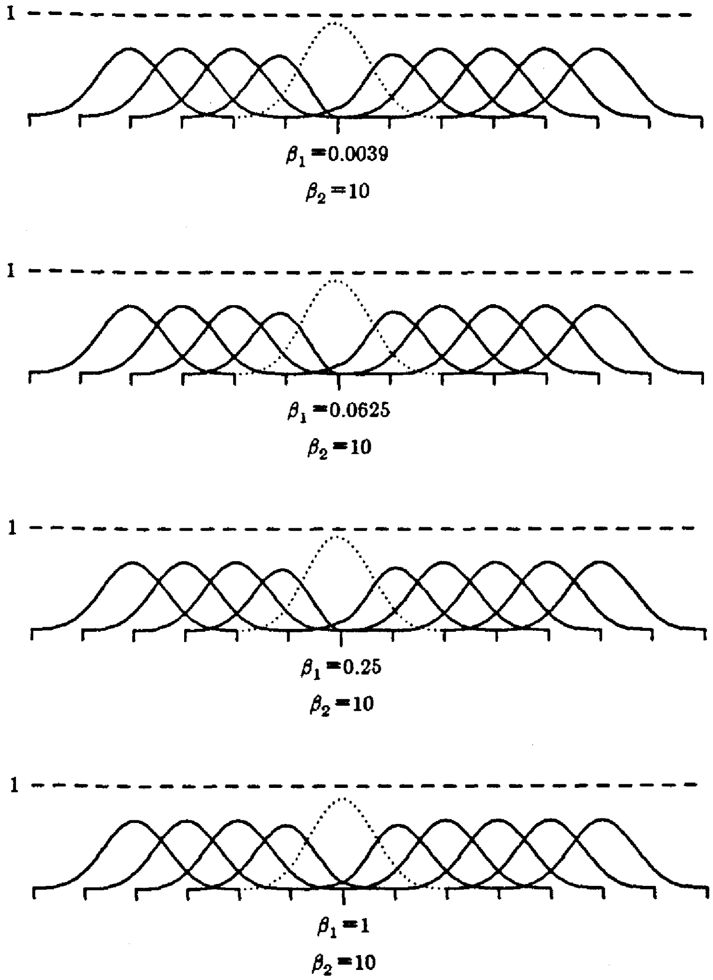


Figure 238. The discretely-shaped Beta-splines corresponding to the four curves of Figure 237. $\beta_1$ and $\beta_2$ have the values 1 and 0 at all knots except the one explicitly labeled.

One sees also how the curves lie within the convex hull of their corresponding control vertices; Figures

239 and 240 illustrate the failure of a curve to lie within the convex hull of its control points when a $\beta_2$ value is negative.

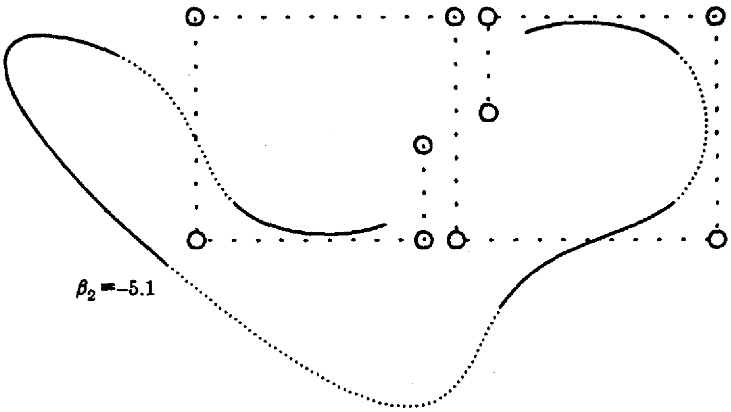


Figure 239. For negative values of $\beta_2$ the curve may pass outside the convex hull. $\beta_1$ has the value 1 and $\beta_2$ the value 0 at every joint except the one explicitly indicated.

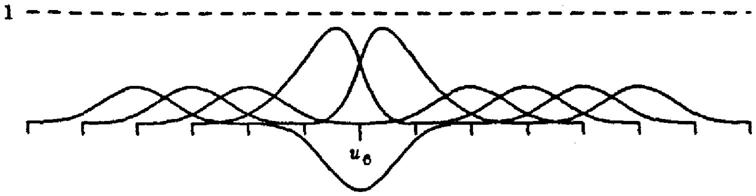


Figure 240. These are the (unscaled) Beta-splines with which the curve of Figure 239 is defined. Notice the negative basis function centered over the knot at which $\beta_2$=−5.1. This is not a violation of the convex hull property established in the text, which holds only for positive values of $\beta_1$ and $\beta_2$.

Figure 241 demonstrates the similar, though not identical, tension-like effects produced by manipulating $\beta_1$ and $\beta_2$. Figure 242 is produced by varying several shape parameters simultaneously.

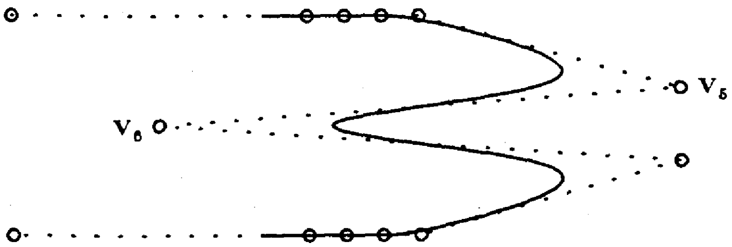


Figure 241. A uniform discretely-shaped Beta-spline curve. Actually this is a $C^2$ spline curve since $\beta_1$ and $\beta_2$ have the values 1 and 0 throughout the curve, which should be compared with the curves in Figure 242.

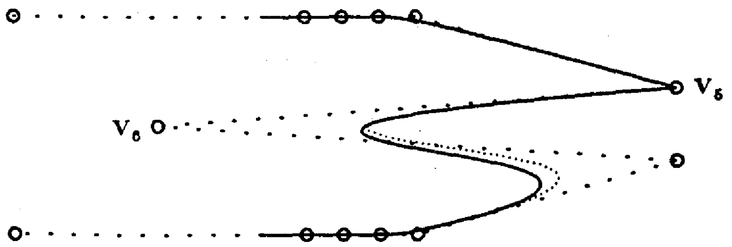


Figure 242. The solid curve here is obtained from the curve of Figure 241 by increasing $\beta_1$ at $V_5$ from 1 to 10,000. The dotted curve is obtained from Figure 241 by instead increasing $\beta_2$ at $V_5$ from 0 to 10,000. In both cases a further increase in the shape parameter produces no observable change in the figure.

Figures 243 and 244 illustrate the locality provided by the discretely-shaped Beta-splines.

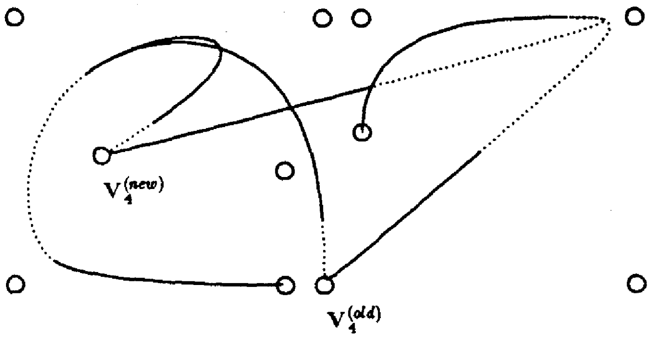


Figure 243. Here we see the effect produced by moving one of the control vertices defining a curve. Notice that only four curve segments are altered. (The control polygon has been omitted here to enhance visibility of the curves.)

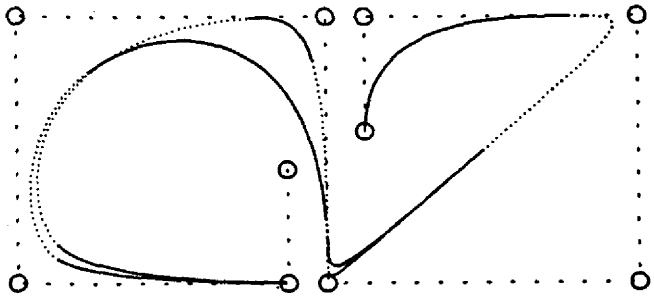


Figure 244. In this case we have changed the knot spacing for the third segment.

## 21. Acknowledgments

The Vaxima symbolic manipulation system [Bogen77, Fateman82] has been extremely useful in this work, and the authors are grateful to Bruce Char for assisting us in using it effectively.

We would like to thank the students in CS284 at the University of California, Berkeley and CS779 at the University of Waterloo for meticulous proofreading of preliminary manuscripts.

# 22. References

[Aho74] Alfred V. Aho, John E. Hopcroft, and Jeffrey D. Ullman (1974), ***The Design and Analysis of Computer Algorithms***, Addison-Wesley.

[Barsky80] Brian A. Barsky and Donald P. Greenberg (1980), Determining a Set of Control Vertices to Generate an Interpolating Surface, *Computer Graphics and Image Processing* 14(3), November, 203-226.

[Barsky81a] Brian A. Barsky and Spencer W. Thomas (1981), TRANSPLINE — A System for Representing Curves Using Transformations Among Four Spline Formulations, *The Computer Journal* 24(3), 271-277.

[Barsky81] Brian A. Barsky (1981), The Beta-spline: A Local Representation Based On Shape Parameters and Fundamental Geometric Measures, PhD dissertation, Department of Computer Science, University of Utah, December.

[Barsky82] Brian A. Barsky (1982), End Conditions and Boundary Conditions for Uniform B-Spline Curve and Surface Representations, *Computers in Industry* 3(1/2), March/June, 17-29 [the Steven A. Coons memorial issue].

[Barsky82a] Brian A. Barsky and John C. Beatty (1982), Varying the Betas in Beta-splines, TR-82-49, Department of Computer Science, University of Waterloo, Waterloo, Ontario, December, 42 pages [also available as TR UCB/CSD 82/112 from the Computer Science Division of the University of California, Berkeley].

[Barsky83] Brian A. Barsky and John C. Beatty (1983), Controlling the Shape of Parametric B-spline and Beta-spline Curves, *Graphics Interface '83*, May, 223-232.

[Barsky83a] Brian A. Barsky and John C. Beatty (1983), Local Control of Bias and Tension in Beta-splines, *ACM Transactions on Graphics* 2(2), April, 27-52 [also available in the proceedings of the 1983 Siggraph Conference, which are the July 1983 issue of the Computer Graphics Quarterly, 17(3) pp 193-218].

[Barsky83b] Brian A. Barsky and Tony D. DeRose (1983), The Beta2-spline: A Special Case of the Beta-spline Curve and Surface Representation, UCB/CSD 83/152, Computer Science

Division, University of California, Berkeley, California 94720, November.

[Barsky84] Brian A. Barsky (1984), A Description and Evaluation of Various 3-D Models, *IEEE Computer Graphics and Applications* 4(1), January, 38-52 [an earlier version of this paper appeared in the *Proceedings of InterGraphics '83*, Japan Management Association, Tokyo, 11-14 April 1983, pp (B2-5) 1-21 and was reprinted in *Computer Graphics — Theory and Applications*, edited by Tosiyasu L. Kunii, Springer-Verlag, Tokyo, 1984].

[Barsky84a] Brian A. Barsky and Tony D. DeRose (1984), Geometric Continuity of Parametric Curves, UCB/CSD 84/205, Computer Science Division, University of California, Berkeley, California 94720, October.

[Barsky85] Brian A. Barsky (1985), *Computer Aided Geometric Design and Computer Graphics Using Beta-splines*, Springer-Verlag, Tokyo.

[Barsky85a] Brian A. Barsky (1985), Arbitrary Subdivision of Bézier Curves, Computer Science Division, University of California, Berkeley, California 94720, 17 pages [submitted for publication].

[Barsky85b] Brian A. Barsky, Tony D. DeRose, and Mark D. Dippé (1985), An Adaptive Subdivision Method With Crack Prevention for Rendering Beta-spline Objects, Computer Science Division, University of California, Berkeley, California 94720 [submitted for publication].

[Barsky85c] Brian A. Barsky (1985), The Beta-spline: A Curve and Surface Representation for Computer Graphics and Computer Aided Geometric Design, Computer Science Division, University of California, Berkeley, California 94720 [submitted for publication].

[Barsky85d] Brian A. Barsky (1985), Algorithms for the Evaluation and Perturbation of Beta-splines, Computer Science Division, University of California, Berkeley, California 94720 [submitted for publication].

[Bartels84] Richard H. Bartels and John C. Beatty (1984), Beta-splines With A Difference, CS-83-40, Department of Computer Science, University of Waterloo, Waterloo, Ontario, Canada N2L 3G1, May, 61 pages.

[Bézier70] Pierre E. Bézier (1970), *Emploi des machines à commande numérique*, Masson et Cie., Paris [translated by A. Robin Forrest and Anne F. Pankhurst as *Numerical Control — Mathematics and Applications*, John Wiley and Sons, Ltd., London, 1972, and available from University Microfilms, Ann Arbor, Michigan].

[Bézier74] Pierre E. Bézier (1974), Mathematical and Practical Possibilities of UNISURF, *Computer Aided Geometric Design*, Robert E. Barnhill & Richard F. Riesenfeld (eds.), Academic Press, New York, 127-152.

[Bézier77] Pierre E. Bézier (1977), Essai de définition numérique des courbes et des surfaces expérimentales, PhD dissertation, l'Université Pierre et Marie Curie, Paris, February.

[Bogen77] Richard Bogen, Jeffrey Golden, Michael Genesereth, and Alexander Doohovskoy (1977), *MACSYMA Reference Manual*, Massachusetts Institute of Technology [version nine].

[deBoor72] Carl de Boor (1972), On Calculating with B-splines, *Journal of Approximation Theory* 6(1), July, 50-62.

[deBoor78] Carl de Boor (1978), *A Practical Guide to Splines*, Applied Mathematical Sciences

Volume 27, Springer-Verlag.

[Catmull74] Edwin E. Catmull (1974), A Subdivision Algorithm for Computer Display of Curved Surfaces, Computer Science Department, University of Utah, Salt Lake City, Utah 84112, December.

[Catmull75] Edwin E. Catmull (1975), Computer Display of Curved Surfaces, *Proceedings of the Conference on Computer Graphics, Pattern Recognition, and Data Structures*, May, 11-17 [also available in **Tutorial and Selected Readings in Interactive Computer Graphics**, ed. Herbert Freeman, IEEE Catalog No. EHO 156-0.].

[Clark79] James H. Clark (1979), A Fast Scan-Line Algorithm for Rendering Parametric Surfaces, *Computer Graphics* 14(3), August, 7-12 [addendum to the proceedings of the 1979 Siggraph Conference — "papers to be published in the CACM"].

[Cohen80] Elaine Cohen, Tom Lyche, and Richard Riesenfeld (1980), Discrete B-splines and Subdivision Techniques in Computer-Aided Geometric Design and Computer Graphics, *Computer Graphics and Image Processing* 14(2), October, 87-111.

[Coons64] Steven A. Coons (1964), Surfaces for Computer Aided Design, Design Division, Mech. Engin. Dept., M.I.T., Cambridge, Massachusetts.

[Coons67] Steven A. Coons (1967), Surfaces for Computer-Aided Design of Space Forms, MAC-TR-41, Project MAC, M.I.T., Cambridge, Massachusetts, June [available as AD-663 504 from NTIS, Springfield, Virginia].

[Cox71] Morris G. Cox (1971), The Numerical Evaluation of B-splines, NPL-DNACS-4, Division of Numerical Analysis and Computing, National Physical Laboratory, Teddington, Middlesex, England, August [Also in *J. Inst. Maths. Applics.*, Vol. 10, 1972, pp. 134-149.].

[Dransch85] Detlef Dransch (1985), An Editor for Benesh Dance Notation, MMath thesis dissertation, Department of Computer Science, University of Waterloo [in preparation].

[Fateman82] Richard J. Fateman (1982), Addendum to the MACSYMA Reference Manual for the VAX, University of California, Berkeley.

[Faux79] Ivor D. Faux and Michael J. Pratt (1979), *Computational Geometry for Design and Manufacture*, John Wiley & Sons.

[Foley82] James D. Foley and Andries van Dam (1982), *Fundamentals of Interactive Computer Graphics*, Addison Wesley.

[Forrest72] A. Robin Forrest (1972), Interactive Interpolation and Approximation by Be´zier Polynomials, *Cmputer Journal* 15(1), 71-79.

[Forrest79] A. Robin Forrest (1979), On the Rendering of Surfaces, *Computer Graphics* 13(2), August, 253-259 [proceedings of the 1979 Siggraph Conference].

[Forsythe77] George E. Forsythe, Michael A. Malcolm, and Cleve B. Moler (1977), *Computer Methods for Mathematical Computations*, Prentice-Hall.

[Gordon74] William J. Gordon and Richard F. Riesenfeld (1974), B-spline Curves and Surfaces, *Computer Aided Geometric Design*, Robert E. Barnhill & Richard F. Riesenfeld (eds.), Academic Press, 95-126.

[Kajiya82] James T Kajiya (July, 1982), Ray Tracing Parametric Patches, *Computer Graphics*

16(3), 245-254.

[Kajiya83] James T. Kajiya (July, 1983), New Techniques for Ray Tracing Procedurally Defined Objects, *Transactions on Graphics* 2(3), 161-181.

[Kochanek82] Doris H. U. Kochanek, Richard Bartels, and Kellogg S. Booth (1982), A Computer System for Smooth Keyframe Animation, CS-82-42, Department of Computer Science, University of Waterloo, Waterloo, Ontario, Canada N2L 3G1, 84 pages.

[Kochanek84] Doris H. U. Kochanek and Richard H. Bartels (1984), Interpolating Splines with Local Tension, Continuity and Bias Control, *Computer Graphics*, vol. 18, no. 3, July, 33-41 [proceedings of the 1984 Siggraph Conference].

[Lane80] Jeffrey M. Lane and Richard F. Riesenfeld (1980), A Theoretical Development for the Computer Generation of Piecewise Polynomial Surfaces, *IEEE Transactions on Pattern Analysis and Machine Intelligence* PAMI-2(1), January, 35-46.

[Blinn80] Jeffrey M. Lane, Loren C. Carpenter, Turner Whitted, and James F. Blinn (1980), Scan Line Methods for Displaying Parametrically Defined Surfaces, *Communications of the ACM* 23(1), January, 23-34.

[Newman73] William M. Newman and Robert F. Sproull (1973), *Principles of Interactive Computer Graphics*, McGraw-Hill.

[Plass83] Michael Plass and Maureen Stone (July, 1983), Curve-Fitting with Piecewise Parametric Cubics, *Computer Graphics* 17(3), 229-239.

[Prautzsch84] Hartmut Prautzsch (1984), A Short Proof of the Oslo Algorithm, *Computer Aided Geometric Design* 1(1), July, 95-96.

[Ramshaw85] Lyle Ramshaw (1985), A Euclidean View of Joints Between Bézier Curves, 19 pages [submitted for publication].

[Riesenfeld73] Richard F. Riesenfeld (1973), Applications of B-spline Approximation to Geometric Problems of Computer-Aided Design, PhD dissertation, Department of Systems and Information Science, Syracuse University, May.

[Rogers76] David F. Rogers and J. Alan Adams (1976), *Mathematical Elements for Computer Graphics*, McGraw-Hill.

[Schumaker81] Larry L. Schumaker (1981), *Spline Functions: Basic Theory*, John Wiley & Sons.

[Schweitzer82] Dino Schweitzer and Elizabeth S. Cobb (1982), Scanline Rendering of Parametric Surfaces, *Computer Graphics* 16(3), July, 265-271 [proceedings of the 1982 Siggraph Conference].

[Smith83] Alvy Ray Smith (1983), Spline Tutorial Notes, Technical Memo No. 77, Computer Graphics Project, Lucasfilm Ltd., May, 12 [available in the notes for the Introduction to Computer Animation Tutorial at the 1983 Siggraph Conference].

[Stone85] Maureen Stone (1985), One Curve, Three Definitions, *Computer Graphics* 19(1), January, cover.

[Sweeney84] Michael A. J. Sweeney (1984), The Waterloo CGL Ray Tracing Package, M.Math dissertation, Department of Computer Science, University of Waterloo, September.

[Weghorst84] Hank Weghorst, Gary Hooper, and Donald Greenberg (1984), Improved Computational Methods for Ray Tracing, *ACM Transactions on Graphics* 3(1), January, 52-69.

[Wu77] Sheng-Chuan Wu, John F. Abel, and Donald P. Greenberg (1977), An Interactive