Original Lecture #8: 27 October 1992
Topics: Joints and Splines via Polar Forms
Scribe: from your lecturers

Handout 19 discussed the polar form of a single polynomial curve, considered in isolation. This handout goes on to consider joints and spline curves.

1 Differentiation and polar forms

Let F be a parametric, polynomial curve of degree at most n and let f be the polar form of F. As we saw in Handout 19, converting from F to f gives us a way to compute additional values: the *polar values* $f(T_1,\ldots,T_n)$ of F, instead of just the *diagonal values* $F(T) = f(T,\ldots,T)$. But another kind of value associated with F is a value of the derivative $F'(T)$. Our first goal is to relate differentiation of a polynomial to the polar form of that polynomial. That is, we want to find some way of viewing a value $F'(T)$ of the derivative as some generalized kind of polar value of F.

By the way, in Handout 19, times were denoted by lower-case variables, such as t. That actually violated our CS348a conventions. We are modeling a curve parametrically, so the parameter space is an affine line. A *time* is a point on the parameter line. Hence, we should use an upper-case variable, such as T, to denote the single Cartesian coordinate of a time—and, in this handout, that is what we shall do.

1.1 Homogenizing polar forms

Note that a value $F(T)$ of the curve F itself is a *point*, while a value $F'(T)$ of its derivative F' is a *vector*—to be specific, a *velocity vector*. Thus, in order to handle differentiation cleanly, we must be able to distinguish between points in the object space and vectors on the object space. We therefore extend the coordinate system on our object space by adding a weight coordinate at the beginning, with the convention that points have weight 1 and vectors have weight 0. For example, let $G(T) := (T, T^2)$ be the parameterization of the parabola $Y = X^2$ that appeared as an example in Handout 19. Without a weight coordinate, we have

$$G(T) = (T, T^2) \qquad \text{and} \qquad g(T_1, T_2) = \left(\frac{T_1 + T_2}{2}, T_1 T_2\right).$$

Putting in the weight coordinate on the object space, we get instead

$$G(T) = (1; T, T^2) \qquad \text{and} \qquad g(T_1, T_2) = \left(1; \frac{T_1 + T_2}{2}, T_1 T_2\right).$$

Adding a weight coordinate to the object space is only half the job. Recall that the parameter space that we used to model our curve is an affine line and that a time T is really a point on that parameter line. In order to compute derivatives, we also have to be able to subtract one time from another, that is, to find the difference between two points on the parameter line. Such a difference is a vector on the parameter line. To handle this type of vector properly, we have to add a weight coordinate to our times as well. We will replace the time T with the pair (s,t), referring to the result as a *parameter site*. Note that converting from (T) to (s,t) when describing the parameter line is precisely analogous to converting from (X,Y) to $(w;x,y)$ when describing the object plane. (The letter that we are using for the weight coordinate comes first alphabetically, in both cases, even though we are writing the weight coordinate last. Generous readers will view this as consistency, although of a somewhat backwards sort.)

What does the parabola example look like now? Let's think about the diagonal form first, written above as $G(T) = (1;T,T^2)$. To homogenize this formula, we replace each T by t and we add as many factors of the weight coordinate s as necessary to bring the total degree of each term on the right-hand side up to two (two, in this case, because the parabola is quadratic):

$$G(t,s) = (s^2;ts,t^2).$$

Homogenizing the polar form $g(T_1,T_2) = (1;(T_1+T_2)/2,T_1T_2)$ for the parabola is a bit trickier, since there are two different weight coordinates floating around: T_1 should be replaced by (t_1,s_1) and T_2 by (t_2,s_2). The proper goal is as follows: Every term in the polynomials that define $g(T_1,T_2)$ should include exactly one coordinate from each of the two arguments, that is, should include either t_1 or s_1 but not both, and should include either t_2 or s_2 but not both. Here is the homogenized polar form:

$$g((t_1,s_1),(t_2,s_2)) = \left(s_1s_2;\frac{t_1s_2+t_2s_1}{2},t_1t_2\right).$$

Note that, when the original term included a factor of t_1 but not a factor of t_2, we have brought the degree of that term up to two by adding in a factor of s_2. It had to be s_2, not s_1, since each term must have precisely one variable with the subscript 1 and one variable with the subscript 2.

Think about the following four formulas for the parabola example:

$$\begin{array}{cc} G(T) = (1;T,T^2) & g(T_1,T_2) = \left(1;\frac{T_1+T_2}{2},T_1T_2\right) \\ G(t,s) = (s^2;ts,t^2) & g((s_1,t_1),(s_2,t_2)) = \left(s_1s_2;\frac{t_1s_2+t_2s_1}{2},t_1t_2\right) \end{array}$$

Going from a formula on the left to a formula on the right is polarization. Going from a formula on the top to a formula on the bottom is homogenization. If you want to both polarize and homogenize, you can do the two operations in either order, and you'll get the same result. The version of the polar-form theorem in Handout 19 described going from the upper left to the upper right, from a non-homogeneous polynomial to a multiaffine polynomial. There is a quite similar polar-form theorem that describes going from the lower left to the lower right, from a homogeneous polynomial to a *multilinear* polynomial. Note that the function g in the lower-right corner is a linear function of the pair (s_1,t_1) when the pair (s_2,t_2) is held fixed and vice versa; that is, g is bilinear.

1.2 The vector δ

The advantage of homogenizing a curve F and its polar form f is that we can then evaluate them, not only at a time, like $T = (1;t)$, but also at parameter sites $(s;t)$ that have $s \neq 1$. The particular case of primary interest to us for differentiation is the site $(0;1)$, which is a vector. In fact, the site $(0;1)$ is precisely the unit vector in the direction of increasing time, and we shall use the symbol δ to denote it:

$$\delta := (0;1).$$

Note that δ is the difference between the time $T = 1$ and the time $T = 0$, since we have $(1;1) - (1;0) = (0;1) = \delta$. More generally, we have $(1;t+1) - (1;t) = (0;1) = \delta$ for any t.

For brevity, let's denote the point $(1;t)$ on the time line by $\bar{t}$. With this compact notation, we can rephrase that last observation about the vector δ as the formula

$$\overline{t+1} - \bar{t} = \delta \qquad \text{for any } t.$$

More generally, we have

$$\overline{t+h} - \bar{t} = h\delta \qquad \text{for any } t \text{ and } h,$$

and hence

$$\overline{t+h} = \bar{t} + h\delta.$$

1.3 Differentiating is evaluating at δ

Differentiating a polynomial curve corresponds—except for an annoying scale factor—to evaluating that curve with one of the polar arguments equal to the vector δ. To see how this works, let's start by thinking about the cubic case.

Let F be a cubic polynomial curve, let f be its polar form, and suppose that we have homogenized both F and f. We want to compute a value of the derivative F', say $F'(\bar{t})$. By the definition of the derivative, we have

$$\begin{aligned} F'(\bar{t}) &= \lim_{h\to 0} \frac{F(\overline{t+h}) - F(\bar{t})}{h} \\ &= \lim_{h\to 0} \frac{f(\overline{t+h}, \overline{t+h}, \overline{t+h}) - f(\bar{t},\bar{t},\bar{t})}{h} \\ &= \lim_{h\to 0} \frac{f(\bar{t}+h\delta, \bar{t}+h\delta, \bar{t}+h\delta) - f(\bar{t},\bar{t},\bar{t})}{h} \end{aligned}$$

But the homogenized polar form f is a multilinear function. Using linearity in the first argument, we have

$$f(\bar{t}+h\delta, \bar{t}+h\delta, \bar{t}+h\delta) = f(\bar{t}, \bar{t}+h\delta, \bar{t}+h\delta) + hf(\delta, \bar{t}+h\delta, \bar{t}+h\delta).$$

Using linearity in the other two arguments as well and then exploiting the symmetry to the polar form f to deduce, for example, that $f(\delta,\bar{t},\bar{t}) = f(\bar{t},\bar{t},\delta)$, we end up with a formula that is very like the one for the cube of a binomial:

$$f(\bar{t}+h\delta, \bar{t}+h\delta, \bar{t}+h\delta) = f(\bar{t},\bar{t},\bar{t}) + 3hf(\bar{t},\bar{t},\delta) + 3h^2 f(\bar{t},\delta,\delta) + h^3 f(\delta,\delta,\delta).$$

When we substitute this into our formula above for $F'(\bar{t})$, we get

$$F'(\bar{t}) = \lim_{h\to 0} \frac{3hf(\bar{t},\bar{t},\delta) + 3h^2 f(\bar{t},\delta,\delta) + h^3 f(\delta,\delta,\delta)}{h}.$$

Dividing through by h and taking the limit, we arrive at our final answer:

$$F'(\bar{t}) = 3f(\bar{t},\bar{t},\delta).$$

That is, one way to differentiate F once is to replace one of the three arguments of the polar form f by δ (and multiply by an annoying factor of 3). Note that the derivative F' is itself a quadratic function, so its polar form has only two arguments: the two arguments of f that are left after we set one argument to δ.

The numeric factor of 3 in the formula $F'(\bar{t}) = 3f(\bar{t},\bar{t},\delta)$ is a bit confusing at first. Here's my best explanation of why it's there. The homogenized polar form f is trilinear. If we fix the first two arguments at the time $\bar{t}$, the result $f(\bar{t},\bar{t},\cdot)$ is a linear function of the third argument. In particular, the value $f(\bar{t},\bar{t},\delta)$ is a vector that indicates how the point $f(\bar{t},\bar{t},\bar{u})$ moves when the third argument u increases by one unit. But when differentiating, we are varying all three of the polar arguments in parallel. Varying one of the arguments by some infinitesimal distance dh, say from $\bar{t}$ to $\overline{t+dh}$ causes the result to change by the vector

$$f(\bar{t},\bar{t},\bar{t}+\delta\,dh) - f(\bar{t},\bar{t},\bar{t}) = f(\bar{t},\bar{t},\delta)\,dh.$$

Varying all three of the arguments in parallel makes the total change three times greater.

The general formula for the first derivative of an n-ic curve F is as follows:

$$F'(\bar{t}) = nf(\underbrace{\bar{t},\ldots,\bar{t}}_{n-1},\delta).$$

For second derivatives, repeating essentially the same process leads to the formula

$$F''(\bar{t}) = n(n-1)f(\underbrace{\bar{t},\ldots,\bar{t}}_{n-2},\delta,\delta).$$

For kth derivatives, we have

$$F^{(k)}(\bar{t}) = n(n-1)\cdots(n-k+1)f(\underbrace{\bar{t},\ldots,\bar{t}}_{n-k},\underbrace{\delta,\ldots,\delta}_{k}). \tag{1.1}$$

If we let k be n, we find that the (constant) value of the nth derivative function $F^{(n)}$ is given by

$$F^{(n)}(\bar{t}) = n!\,f(\underbrace{\delta,\ldots,\delta}_{n}).$$

If we differentiate an n-ic function like F more than n times, the result should be identically zero. If we let k be larger than n in Equation 1.1, we can't find k argument slots of the polar form f to put copies of δ in; but that doesn't matter, because the numeric factor outside is zero in this case, and hence guarantees the correct final result.

1.4 Easy consequences

Before going further, let's discuss some easy consequences of Equation 1.1.

For concreteness, let F be a cubic curve, and suppose that we know the Bézier points of the segment $F([0\,..\,1])$. What is the starting velocity vector $F'(\bar{0})$ in terms of those Bézier points? From Equation 1.1, we have $F'(\bar{0}) = 3f(\bar{0},\bar{0},\delta)$. Since $\delta = \bar{1}-\bar{0}$, we have

$$F'(\bar{0}) = 3\big(f(\bar{0},\bar{0},\bar{1}) - f(\bar{0},\bar{0},\bar{0})\big).$$

Thus, the starting velocity is three times the vector from the first Bézier point $f(\bar{0},\bar{0},\bar{0})$ to the second Bézier point $f(\bar{0},\bar{0},\bar{1})$.

Warning: The factor of 3 out in front is the correct factor only when the length of the parametric interval is 1. Suppose that we were given, instead, the Bézier points of the segment $F([p\,..\,q])$. The formula for the starting velocity vector $F'(\bar{p})$ in terms of those Bézier points is

$$F'(\bar{p}) = \frac{3}{q-p}\big(f(\bar{p},\bar{p},\bar{q}) - f(\bar{p},\bar{p},\bar{p})\big).$$

For a slightly harder case, let's go back to the segment $F([0\,..\,1])$ and compute the starting acceleration vector in terms of the Bézier points. From Equation 1.1, we have $F''(\bar{0}) = 6f(\bar{0},\delta,\delta)$. Substituting $\delta = \bar{1}-\bar{0}$ in both argument positions and expanding by multilinearity, we get

$$F''(\bar{0}) = 6\big(f(\bar{0},\bar{1},\bar{1}) - 2f(\bar{0},\bar{0},\bar{1}) + f(\bar{0},\bar{0},\bar{0})\big).$$

In general, the starting value of the kth derivative vector depends only on the first $k+1$ Bézier points, while the ending value of the kth derivative vector depends only on the last $k+1$ Bézier points.

1.5 Why the overbars?

The last section states some easy results, but many of the formulas have messy overbars in them. Where did those overbars come from and why do we need them?

Most people denote times simply by real numbers, with no overbar, and we did that ourselves back in Handout 19. For example, we denoted the four Bézier points of a quadratic segment $F([0\,..\,1])$ in Handout 19 as $f(0,0,0)$, $f(0,0,1)$, $f(0,1,1)$, and $f(1,1,1)$. Back then, we were thinking of the polar form f as a triaffine function whose arguments were times, and we wrote those times simply as real numbers.

Our fall from grace happened when we homogenized the polar form f. After homogenization, each argument to f is a site $(s;t)$ in parameter space. The parameter site $(s;t)$ may be a time, that is, may have $s = 1$. But we can also have sites with $s \neq 1$; consider, for example, the site $\delta = (0;1)$. Since the arguments to the homogenized polar form f are sites, we can't denote them simply by real numbers. The overbars are a convenient shorthand for writing those parameter sites that are times: $\bar{t} = (1;t)$.

That's the official story. The unofficial story is that it is a pain to write overbars all the time, and we will sometimes be sloppy about it.

1.6 The naive versus the Bézier way of writing polynomials

[The material in this section is advanced and peripheral; skip it if you get confused.]

The relationship above between differentiation and evaluation at δ also gives us an interesting connection between the Bézier theory and the naive way to write polynomials. The naive way to write a polynomial F of degree at most n is as an explicit sum of powers:

$$F(t) = c_n t^n + c_{n-1} t^{n-1} + \cdots + c_0.$$

If we want a formula for the coefficients c_k in this expansion, we can use Taylor's Theorem around 0:

$$\begin{aligned} F(\bar{t}) &= F(\bar{0}) + \frac{F'(\bar{0})}{1} t + \frac{F''(\bar{0})}{2!} t^2 + \cdots + \frac{F^{(n)}(\bar{0})}{n!} t^n \\ &= \sum_{k=0}^{n} \frac{F^{(k)}(\bar{0})}{k!} t^k. \end{aligned}$$

If f is the polar form of F, we can use Equation 1.1 to rewrite this as follows:

$$F(\bar{t}) = \sum_{k=0}^{n} \frac{n(n-1)\cdots(n-k+1)}{k!} f(\underbrace{\bar{0},\ldots,\bar{0}}_{n-k},\underbrace{\delta,\ldots,\delta}_{k}) t^k.$$

The numeric factor in this summand is such a popular quantity that it has its own special name and notation: It is a *binomial coefficient*, written

$$\binom{n}{k} = \frac{n(n-1)\cdots(n-k+1)}{k!}.$$

So we can also write

$$F(\bar{t}) = \sum_{k=0}^{n} f(\underbrace{\bar{0},\ldots,\bar{0}}_{n-k},\underbrace{\delta,\ldots,\delta}_{k}) \binom{n}{k} t^k. \tag{1.2}$$

More generally, if we expand F in a Taylor series about some arbitrary point p, instead of around 0, we get

$$F(\bar{t}) = \sum_{k=0}^{n} f(\underbrace{\bar{p},\ldots,\bar{p}}_{n-k},\underbrace{\delta,\ldots,\delta}_{k}) \binom{n}{k} (t-p)^k. \tag{1.3}$$

Equation 1.2 is interesting because it shows that the Bézier approach to polynomials is not as different from the naive approach as one might at first suppose. Recall that the Bézier points are the coefficients that result if we expand the polynomial F as a linear combination of the Bernstein basis polynomials. Using the interval $[0\,..\,1]$ as an example, we have

$$F(\bar{t}) = \sum_{k=0}^{n} f(\underbrace{\bar{0},\ldots,\bar{0}}_{n-k},\underbrace{\bar{1},\ldots,\bar{1}}_{k}) \binom{n}{k} (1-t)^{n-k} t^k.$$

Comparing this to Equation 1.2, we see that they are very similar; going from this to Equation 1.2 involves—roughly speaking—replacing $\bar{1}$ by δ. Thus, still speaking roughly, the coefficients of the naive expansion of the polynomial F in Equation 1.2 are essentially the Bézier points (more precisely, the *Bézier sites*) of the segment $F([0\,..\,\infty])$.

Here is the story in more detail. If we expand F in terms of the Bézier points of the segment $F([p\,..\,q])$, we get

$$F(\bar{t}) = \sum_{k=0}^{n} f(\underbrace{\bar{p},\ldots,\bar{p}}_{n-k},\underbrace{\bar{q},\ldots,\bar{q}}_{k}) \binom{n}{k} \left(\frac{q-t}{q-p}\right)^{n-k} \left(\frac{t-p}{q-p}\right)^{k}. \tag{1.4}$$

Take this equation and let q tend to infinity, while t and p stay fixed. The quotient $(q-t)/(q-p)$ tends to 1, and hence drops out. The polar value

$$f(\underbrace{\bar{p},\ldots,\bar{p}}_{n-k},\underbrace{\bar{q},\ldots,\bar{q}}_{k})$$

itself tends to infinity quite rapidly, since k of its arguments are tending to infinity. But the denominator also has k remaining factors of $(q-p)$, each of which is tending to infinity as well. If we use the multilinearity of f to pull one of those denominator factors in under each $\bar{q}$, each of the last k arguments to f will become $\bar{q}/(q-p)$. Recalling that $\bar{q}$ is an abbreviation for the parameter site $\bar{q} = (q,1)$, we have

$$\frac{\bar{q}}{q-p} = \left(\frac{q}{q-p}, \frac{1}{q-p}\right).$$

Letting q go to infinity, this becomes simply $(0;1) = \delta$. Thus, substituting $q = \infty$ causes Equation 1.4 to reduce to Equation 1.3. In this sense, the coefficients of the Taylor series of F around p can be thought of as the Bézier sites of the segment $F([p\,..\,\infty])$.

2 Joints

Suppose that we are given two Bézier curves. What constraints do we have to put on their Bézier points—or their polar forms—if we want to guarantee that the two segments join smoothly, say to kth order?

Actually, there are two different notions of what it means to be smooth to kth order. The first is called *parametric* continuity, and is written C^k; the other is called either *geometric* or *visual* continuity, and is written G^k or VC^k. Parametric continuity means smoothness both of the curve and of its parameterization. That is, if a curve F is parametrically smooth, we can make a movie in which a car's position at time t is $F(t)$ and the motion of the car in that movie will look smooth. Geometric continuity means simply the smoothness of the track that the car leaves in the snow after it passes by. For example, C^1 continuity means continuity of the tangent vector, while G^1 continuity continuity of slope; C^2 continuity means continuity

of the acceleration vector, while G^2 continuity means continuity of the curvature. Parametric continuity is more expensive to arrange, in the sense that it uses up more degrees of freedom, but it is mathematically much simpler to deal with. In CS348a, we will focus on parametric continuity.

Parametric continuity and polar forms work together very neatly. Suppose $F([p\,..\,q])$ and $G([q\,..\,r])$ are two cubic segments. The level of parametric continuity at the joint at q between F and G turns out to be simply the number of polar arguments that can differ from q without destroying agreement between the resulting polar values of F and G.

Consider the case of C^0 continuity to start with. The segments $F([p\,..\,q])$ and $G([q\,..\,r])$ join with C^0 continuity at q precisely when $F(q) = G(q)$, that is, when $f(q,q,q) = g(q,q,q)$. Thus, the joint has C^0 continuity precisely when the polar forms f and g agree on the particular argument triple (q,q,q). That's the same thing as saying that f and g agree on all argument triples that have zero elements different from q.

At the other extreme, consider C^3 continuity. Since the curves F and G are cubic, the only way that they can meet with C^3 continuity at q is for them to be identical curves. In this case, we surely have $f(u,v,w) = g(u,v,w)$ for any three polar arguments u, v, and w, so f and g agree on all argument triples. Thus, the number of polar arguments that can differ from q without destroying agreement of the polar values is 3.

Here's the statement and proof of the general case.

Theorem 1. *Two n-ic curve segments $F([p\,..\,q])$ and $G([q\,..\,r])$ join with C^k continuity at the joint q—that is, they agree at q parametrically to kth order—precisely when their n-polar forms agree on all sequences of polar arguments that include at most k values different from q.*

Proof: Suppose first that the polar forms of F and G do agree on all sequences of n polar arguments that include at most k values different from $\bar{q}$. Then, in particular, f and g must agree on the sequences

$$(\underbrace{\bar{q},\ldots,\bar{q}}_{n-m},\underbrace{\delta,\ldots,\delta}_{m})$$

for all m from 0 to k. From Equation 1.1, we have

$$F^{(m)}(\bar{q}) = n(n-1)\cdots(n-q+1)f(\underbrace{\bar{q},\ldots,\bar{q}}_{n-m},\underbrace{\delta,\ldots,\delta}_{m}),$$

and similarly for $G^{(m)}(\bar{q})$. Since the factor out front is the same for F as for G, we conclude that the 0th through kth derivatives of F and G agree at q, which is what we mean when we say that F and G join at q with C^k continuity.

Conversely, suppose that the 0th through kth derivatives of F and G agree at q. Working backwards through the above, we conclude that the polar forms f and g must agree on the argument sequences

$$(\underbrace{\bar{q},\ldots,\bar{q}}_{n-m},\underbrace{\delta,\ldots,\delta}_{m})$$

for m from 0 to k. We want to show that f and g agree, in fact, on any argument sequence of the form

$$(\underbrace{\bar{q},\ldots,\bar{q}}_{n-k},\bar{u}_1,\ldots,\bar{u}_k),$$

where the times $\bar{u}_1$ through $\bar{u}_k$ are arbitrary. Note that we have

$$\bar{u}_i = \bar{q} + (u_i - q)\delta \qquad \text{for } i \text{ from 1 to } k.$$

To prove the equality

$$f(\underbrace{\bar{q},\ldots,\bar{q}}_{n-k},\bar{u}_1,\ldots,\bar{u}_k) = g(\underbrace{\bar{q},\ldots,\bar{q}}_{n-k},\bar{u}_1,\ldots,\bar{u}_k),$$

we substitute $\bar{q} + (u_i - q)\delta$ for $\bar{u}_i$ for i from 1 to k and expand by multilinearity on both sides. The result on the left side will be a linear combination of polar values of the form

$$f(\underbrace{\bar{q},\ldots,\bar{q}}_{n-m},\underbrace{\delta,\ldots,\delta}_{m}),$$

where m is at most k. The result on the right will be the same linear combination of the polar values

$$g(\underbrace{\bar{q},\ldots,\bar{q}}_{n-m},\underbrace{\delta,\ldots,\delta}_{m}).$$

But we saw above that those polar values are the same for f and g. □

3 Spline curves

Now that we can control the level of parametric continuity at a joint in terms of the polar forms of the joining curves, we are ready to consider assembling a sequence of curve segments into a spline curve.

3.1 Knots

Suppose that we choose to assemble our spline curve out of segments of degree at most n. What continuity should we ask for at the joints?

A joint with C^n continuity wouldn't be a joint at all: The entering and leaving segments would be adjacent segments of the same polynomial curve. So having joints with C^n continuity is pointless. The highest continuity that we can ask for at a joint that leaves any flexibility at all is C^{n-1} continuity. At a C^{n-1} joint, the derivative vectors from the 0th up through the $(n-1)$st are smooth, but the nth derivative jumps discontinuously from one value to another.

In addition to joints with C^{n-1} continuity, we can—if we like—allow joints with C^{n-m} continuity for various values of m greater than 1. At such a joint, the 0th through $(n-m)$th

derivatives are smooth, while the remaining m derivatives may all jump discontinuously. Letting $m = n$, for example, we can allow $C^{n-n} = C^0$ joints, where only the 0th derivative, the position of the particle, is continuous. By convention, we even allow the case $m = n+1$, which leads to joints with C^{-1} continuity. At such a joint, even the position has a jump. Thus, there is no relationship at all between the incoming and outgoing segments at a C^{-1} joint.

A parameter value that corresponds to a joint is called a *knot*. If our spline curve has C^{n-1} continuity at the corresponding joint, the knot is called *simple*. Suppose that the joint has only C^{n-2} continuity. The standard convention is to say that the corresponding knot is a *double knot*, that is, it should be thought of as two separate, simple knots that have coalesced. In general, a knot of *multiplicity m* is the parameter value corresponding to a joint where the spline curve has C^{n-m} continuity, and should be thought of as a cluster of m simple knots that have coalesced.

Associating a multiplicity n with a knot in this way works out well because a multiple knot really does behave like a limiting case of a cluster of simple knots. For example, consider assembling line segments into an affine spline, that is, a polyline. A typical joint in such a spline is a vertex, where the position of the particle is continuous but the velocity undergoes a sudden jump. Such a joint has C^0 continuity. By our convention above, we also allow an affine spline to have joints with C^{-1} continuity, where even the position of the particle undergoes a jump. Consider such a C^{-1} joint. Say that our polyline L arrives at the point P at time r and then leaves from the point Q, with $P \neq Q$; that is, we have

$$\lim_{t \uparrow r} L(t) = P, \qquad \text{but} \qquad \lim_{t \downarrow r} L(t) = Q.$$

Since this joint has only C^{-1} continuity, the knot r is a double knot. Let $r_1 < r_2$ be two distinct times that are both quite close to r. We can approximate the polyline L arbitrarily closely by a polyline M with two C^0 joints, one at time r_1 with $M(r_1) = P$ and the other at time r_2 with $M(r_2) = Q$. Over the short time interval $[r_1 \,..\, r_2]$, the spline M scoots at high speed from P to Q, thus smoothing out the jump in position that occurs in L. In a similar way, one can view a knot of any multiplicity m as a cluster of m simple knots that have coalesced.

Another way to think about this issue of knot multiplicity is to say that a knot is the time at which one has the right to break a derivative. At a simple knot, we are only allowed to break a single derivative, the nth. At a double knot, each of the two simple knots buried inside the double knot gives us the right to break one derivative, so overall we can break both the nth and $(n-1)$st derivatives. And so forth. At an $(n+1)$-fold knot, we can break all of the derivatives including the 0th derivative, the position; so the resulting joint has C^{-1} continuity. It doesn't make sense to talk about a knot whose multiplicity is higher than $n+1$.

3.2 Knot sequences

With these conventions about knot multiplicity, we can incorporate all of the smoothness information about a spline curve—the time of each joint and the level of smoothness at each joint—in a single sequence of numbers, as follows. We make a list of all of the knots in which each knot is repeated according to its multiplicity, and then we sort that list into non-decreasing order. The result is called the *knot sequence* of the spline.

For example, suppose that we have a cubic spline with the knot sequence

$$(\ldots, 1, 2, 4, 4, 5, \ldots).$$

The resulting spline curve F will have a segment $F([1\,..\,2])$, a segment $F([2\,..\,4])$, and a segment $F([4\,..\,5])$, as well as possibly further segments on both ends. Since the knot 2 has multiplicity 1, the joint at time 2 between $F([1\,..\,2])$ and $F([2\,..\,4])$ will have C^2 continuity. But the knot 4 is a double knot, so the joint between $F([2\,..\,4])$ and $F([4\,..\,5])$ will be only C^1.

So far, we have been talking as if knot sequences were bi-infinite, so that there are no ends. In practice, however, we can deal only with finite sequences of knots, so we have to worry a little bit about what happens at the ends. The cleanest convention is to demand that the end knots of a finite knot sequence be knots of multiplicity $n+1$. For example, consider a finite cubic spline F with the knot sequence

$$(0, 0, 0, 0, 1, 2, 4, 4, 5, 6, 6, 6, 6).$$

The joint at the left end between whatever comes before—which we don't know anything about—and the segment $F([0\,..\,1])$ will be a C^{-1} joint. But there is no relationship between the entering and leaving segments at a C^{-1} joint, so it won't hurt us any that we don't know what came before time 0. Similarly, the fact that 6 is a quadruple knot means that the joint between $F([5\,..\,6])$ and whatever comes after—which we also don't know anything about either—is a C^{-1} joint. In fact, we can often get by with a little less: with first and last knots that have multiplicity only n, instead of $n+1$. But going all the way to multiplicity $n+1$ is certainly safe.

One common type of knot sequence is one in which all of the knots are simple knots and they are equally spaced. That is, the knots form an arithmetic progression on the parameter line. This common situation is called the *uniform case*, and a spline whose knot sequence is uniform is called a *uniform spline*. The geometry of uniform splines is particularly regular, as we will see.

3.3 The de Boor points

Suppose that we have chosen a degree n and a knot sequence (in which no knot has multiplicity greater than $n+1$). We complete our design of the spline curve F by choosing a sequence of control points, which are called *de Boor points* or *B-spline control points*. Each de Boor point is labeled by a block of n adjacent knots from the knot sequence, and successive de Boor points are labeled by blocks that are shifted by one step with respect to each other, that is, that overlap in all but one knot. That—and the de Boor generalization of the de Casteljau Algorithm—turns out to be all that you need to know to draw spline curves.

To see how this works, let's consider the affine, quadratic, and cubic cases in turn.

3.3.1 The affine case

Suppose that $n = 1$, so that we are designing an affine spline—that is, a polyline. And suppose that we have chosen the knot sequence

$$(\ldots, 1, 2, 4, 5, 5, 6, 8, \ldots).$$

Since $n = 1$, each de Boor point will be labeled by single knot, with adjacent de Boor points labeled by adjacent knots. So our spline F has, among its control points, the points P_1, P_2, P_4, P_5, P_5, P_6, and P_8. Note that there are two de Boor points labeled P_5, since 5 is a double knot.

The resulting affine spline F is quite straightforward. Over the time interval $[1\,..\,2]$, the spline F moves from P_1 to P_2 at a constant rate of speed. Over the time interval $[2\,..\,4]$, it moves from P_2 to P_4. Over $[4\,..\,5]$, it moves from P_4 to the first of the two points labeled P_5. Over $[5\,..\,6]$, it moves from the second point labeled P_5 to the point labeled P_6. And so on. Note that the joints corresponding to simple knots have C^0 continuity, as they should, while the joint corresponding to the double knot at time 5 has only C^{-1} continuity.

3.3.2 The quadratic case

To take a less trivial case, suppose that $n = 2$, so that we are assembling parabolic segments into a quadratic spline, and suppose that we have chosen the knot sequence

$$(\ldots, 0, 1, 2, 3, 5, 6, 7, 7, 8, 9, 9, 9, \ldots).$$

Since $n = 2$, each de Boor point is now labeled by a pair of adjacent knots. So reasonable names for the de Boor points are

$$P_{01}, P_{12}, P_{23}, P_{35}, P_{56}, P_{67}, P_{77}, P_{78}, P_{89}, P_{99}, \text{ and } P_{99}.$$

In order to figure out what the segments of the spline curve F are, it is helpful first to compute some auxiliary points, based on the de Boor points. For example, P_{01} and P_{12} share a common subscript. Thinking of P_{uv} rather like a 2-polar value $f(u,v)$, we will define the point P_{1t} for all t in $[0\,..\,2]$ by interpolating between P_{01} and P_{12}. In particular, we define

$$P_{11} := \frac{P_{01} + P_{12}}{2}.$$

In a similar way, we define

$$P_{22} := \frac{P_{12} + P_{23}}{2}.$$

But P_{33} is slightly different, since the knots in this part of the knot sequence are not uniformly spaced—in particular, 3 is only one-third of the way from 2 to 5:

$$P_{33} := \frac{2P_{23} + P_{35}}{3}.$$

And we define P_{55}, P_{66}, and P_{88} analogously. [Please draw a picture of all this for yourself! Start by putting the de Boor points in arbitrary locations. Sad to say, I'm afraid that I don't have time to include an example figure in this handout.]

Over the time interval $[1\,..\,2]$, our spline F will follow a parabolic segment $F([1\,..\,2])$, which is part of some overall parabola—call it G. Of course, the parabola G has a polar form g. Note that F itself is a spline curve, so it doesn't have a polar form, strictly speaking—even though the de Boor points P_{uv} almost behave like 'polar values' $P_{uv} = f(u,v)$ of the spline F. We choose to determine the parabola G by giving the Bézier points of the segment $F([1\,..\,2]) = G([1\,..\,2])$, as follows: $g(1,1) := P_{11}$, $g(1,2) := P_{12}$, and $g(2,2) := P_{22}$.

In a similar way, over each interval $[s\,..\,t]$ between two adjacent, distinct knots, the spline F will follow a parabolic segment $F([s\,..\,t])$ that is part of some parabola G, and that parabola G is determined by the conditions $g(s,s) := P_{ss}$, $g(s,t) := P_{st}$, and $g(t,t) := P_{tt}$. The way in which we computed the auxiliary points, such as P_{11} and P_{22}, from the original de Boor points guarantees that the resulting spline curve F will have the required level of continuity at each joint. For example, the condition that $P_{33} = (2P_{23} + P_{35})/3$ guarantees that the ending tangent vector of the segment $F([2\,..\,3])$ is the same as the starting tangent vector of the segment $F([3\,..\,5])$; so the joint at time 3 is a C^1, as is appropriate, since 3 is a simple knot.

In fact, we don't have to bother computing the auxiliary points P_{11}, P_{22}, and the like; we can work directly from the de Boor points. Consider the time interval $[2\,..\,3]$, for example. Above, we determined the parabola G that our spline F follows over $[2\,..\,3]$ by giving the Bézier points of the segment $G([2\,..\,3])$, which are P_{22}, P_{23}, and P_{33}. The first and last of these are auxiliary points, while the middle one is a de Boor point. Instead, we can determine the parabola G directly from three adjacent de Boor points by imposing the following three polar interpolation constraints on G: $g(1,2) = P_{12}$, $g(2,3) = P_{23}$, and $g(3,5) = P_{35}$. To prove that these three constraints are a legal way to specify a parabola G, we appeal to Theorem 4 from Handout 19, setting $r_1 := 2$, $r_2 := 1$, $s_1 := 3$, and $s_2 := 5$. From that theorem, we deduce that there exists a unique parabola G that satisfies $g(r_1,r_2) = g(1,2) = P_{12}$, $g(r_1,s_1) = g(2,3) = P_{23}$, and $g(s_1,s_2) = g(3,5) = P_{35}$.

3.3.3 The cubic case

Suppose now that $n = 3$. It's worth going through several examples of knot sequences, in the cubic case. The first knot sequence has uniformly spaced simple knots for a while, then has a single longer parameter interval at $[5\,..\,7]$, and ends with a triple knot at 9:

$$(\ldots, 0, 1, 2, 3, 4, 5, 7, 8, 9, 9, 9, \ldots).$$

The second starts with two triple knots in a row, later including a double knot in the middle of a string of simple knots:

$$(\ldots, 0, 0, 0, 1, 1, 1, 2, 3, 4, 4, 5, 6, 7, 8, \ldots).$$

[Choose nine arbitrary de Boor points and draw a spline with the first knot sequence. Choose twelve arbitrary de Boor points and draw a spline with the second knot sequence.]

Since $n = 3$, each de Boor point is labeled by a triple of adjacent knots from the knot sequence. So, in the first example, the nine de Boor points are P_{012}, P_{123}, P_{234}, P_{345}, P_{457}, P_{578}, P_{789}, P_{899}, and P_{999}.

To determine the Bézier points of the various cubic segments in the first example, we treat each de Boor point P_{uvw} like a polar value $f(u,v,w)$ of the spline F. For example, consider the cubic segment $F([2\,..\,3])$. If G is the overall cubic curve that the spline F follows over $[2\,..\,3]$, then the segment $F([2\,..\,3]) = G([2\,..\,3])$ is determined by the positions of the four Bézier points $g(2,2,2)$, $g(2,2,3)$, $g(2,3,3)$, and $g(3,3,3)$. None of the points P_{222}, P_{223}, P_{233}, or P_{333} is a de Boor point. But P_{123} and P_{234} are de Boor points, and we can find both of the points P_{223} and P_{233} by interpolating between P_{123} and P_{234}:

$$P_{223} = \tfrac{2}{3}P_{123} + \tfrac{1}{3}P_{234}$$
$$P_{233} = \tfrac{1}{3}P_{123} + \tfrac{2}{3}P_{234}$$

In a similar way, we can find P_{112} and P_{122} by interpolating between P_{012} and P_{123}, and we can find P_{334} and P_{344} by interpolating between P_{234} and P_{345}. Finally, the point P_{222} must be halfway between P_{122} and P_{223}, while the point P_{333} must be halfway between P_{233} and P_{334}. [Draw all of these points on your picture.]

The other segments of the first example and all of the segments in the second example can be found, in terms of their Bézier points, in a similar way.

Finding the Bézier points of the various segments helps to reveal the geometric structure of our spline curves. If we prefer, however, we can also dispense with the Bézier points and work straight from the de Boor points. Figure 13 in Handout 19 shows an example of determining the segment $F([4\,..\,7]) = G([4\,..\,7])$ of a cubic spline curve F on the knot sequence

$$(\ldots, 2, 3, 4, 7, 8, 9, \ldots)$$

from the four adjacent de Boor points $P_{234} = g(2,3,4)$, $P_{347} = g(3,4,7)$, $P_{478} = g(4,7,8)$, and $P_{789} = g(7,8,9)$. The instance of the de Boor Algorithm illustrated in Figure 13 is computing the point $F(5) = G(5) = g(5,5,5)$ on the resulting segment $F([4\,..\,7])$.

3.4 The basic theorem

Readers who just want to know the truth about splines should get busy and work out lots of examples, as in the previous section. But some readers may be interested in the proofs that these techniques all work as advertised: that, no matter where we put the de Boor points, we get a spline curve with the proper continuity; and that every spline curve with the proper continuity has an associated sequence of de Boor points. This section talks a bit about the theorems and proofs. For the details, see [2].

We will write a general knot sequence as an indexed collection of times, say

$$(t_i) = (\ldots, t_1, t_2, t_3, \ldots).$$

We assume that the knot sequence is sorted, so $t_i \le t_{i+1}$ for all i. If $t_i < t_{i+1}$ is a pair of adjacent knots that are not equal, then our spline curve F will follow a polynomial segment $F([t_i \,..\, t_{i+1}])$ for times in the interval $[t_i \,..\, t_{i+1}]$. Let us refer to the polynomial curve that the segment $F([t_i \,..\, t_{i+1}])$ is a part of as F_i. Thus, F_i is a polynomial curve, while F is a spline curve. Note that, for some indices i, we may have $t_i = t_{i+1}$, in which case the interval $[t_i \,..\, t_{i+1}]$ collapses to a single point and there is no corresponding polynomial curve F_i.

The segment $F_i([t_i \,..\, t_{i+1}])$ joins on, at its right end, to the left end of the segment $F_{i+m}([t_{i+m} \,..\, t_{i+m+1}])$, where m is determined by $t_{i+1} = \cdots = t_{i+m} < t_{i+m+1}$. That is, the knot $t_{i+1} = \cdots = t_{i+m}$ that separates the segments F_i and F_{i+m} has multiplicity m. In order for F to be a spline curve with the knot sequence (t_k), the joint at the knot t_{i+1} between F_i and F_{i+m} must have at least C^{n-m} continuity. By Theorem 1, this means that the polar forms f_i and f_{i+m} must agree on all argument sequences that include $(t_{i+1}, \ldots, t_{i+m})$ as a subsequence, that is, on all sequences of polar arguments that include at least m copies of t_{i+1}.

Two non-adjacent segments F_i and F_j of a spline F are still related to each other, as long as the total number $j-i$ of intervening knots is at most n. One advantage of the polar approach to splines is that it gives us a simple way to describe the relationship between F_i and F_j in general, independent of the multiplicities of the $j-i$ intervening knots: The polar forms f_i and f_j must agree on all argument sequences that include $(t_{i+1}, \ldots, t_j)$ as a subsequence. For example, let t_3 and t_4 be simple knots on a cubic spline F, so that $t_2 < t_3 < t_4 < t_5$. The segments $F_2([t_2 \,..\, t_3])$ and $F_4([t_4 \,..\, t_5])$ are then related by the fact that $f_2(u, t_3, t_4) = f_4(u, t_3, t_4)$ for all u. Here is why: f_2 and f_3 must agree on all argument sequences that include a copy of t_3, and f_3 and f_4 must agree on all argument sequences that include a copy of t_4. So f_2 and f_4 must agree on all argument sequences that include both a t_3 and a t_4. The same argument works in general.

From this, we can see that the sequences $(t_{k+1}, \ldots, t_{k+n})$ of polar arguments are very special. If i and j are any two indices in the interval $[k \,..\, k+n]$ with $t_i < t_{i+1}$ and $t_j < t_{j+1}$, the corresponding polar forms f_i and f_j will agree on the argument sequence $(t_{k+1}, \ldots, t_{k+n})$, since the sequence $(t_{k+1}, \ldots, t_{k+n})$ is bound to include as a subsequence the knots $(t_{i+1}, \ldots, t_j)$ that separate F_i and F_j. The common value that all of these polar forms, such as f_i and f_j, assign to the argument sequence $(t_{k+1}, \ldots, t_{k+n})$ is precisely a *de Boor point* of the spline F. In the last section, we would have called it

$$P_{t_{k+1} \ldots t_{k+n}}.$$

In the following theorem—the basic theorem about spline curves—we call that de Boor point simply P_k.

Theorem 2. *Let n be nonnegative, let (t_k) be a bi-infinite knot sequence in a parameter line L whose knots have multiplicity at most $n+1$, and let Q be an affine object space. There exists a one-to-one correspondence between n-ic spline curves $F : L \to Q$ with the knot sequence (t_k) and bi-infinite sequences of de Boor points (P_k) in Q, where the correspondence between a spline F and its de Boor points (P_k) is given by*

$$P_k = f_i(t_{k+1}, \ldots, t_{k+n})$$

whenever $k \le i \le k+n$ and $t_i < t_{i+1}$. In this formula, the function f_i is the polar form of the n-ic polynomial curve F_i that the spline F follows over the nontrivial parameter interval $[t_i \,..\, t_{i+1}]$.

The inequalities $k \le i \le k+n$ can be read two ways. If we think of k as fixed and i as varying, they tell us which segments F_i of the spline F are influenced by the particular de Boor point P_k. Note that, if all of the knots are simple, then each de Boor point influences $n+1$ spline segments. For example, in an affine spline ($n=1$) with C^0 joints, each vertex influences two segments: the entering one and the leaving one.

We can also turn things around and write the inequalities in the form $i-n \le k \le i$, thinking of i as fixed and k as varying. In this form, they tell us which de Boor points P_k influence the particular spline segment F_i. Since each spline segment is an n-ic segment, the number of de Boor points that influence any particular segment is $n+1$, independent of the knot multiplicities.

The proof of this basic theorem isn't too hard, but we won't go through the details here. Starting with a spline curve F, the challenge is to show that the equations

$$P_k = f_i(t_{k+1}, \ldots, t_{k+n})$$

for i in $[k\,..\,k+n]$ all agree on a common value for the de Boor point P_k. That follows from our comments above about the extent to which the polar forms f_i and f_j of a spline's segments agree. On the other hand, suppose that we start with a sequence of de Boor points (P_k), and consider some i with $t_i < t_{i+1}$. The equations

$$P_k = f_i(t_{k+1}, \ldots, t_{k+n})$$

for k in $[i-n\,..\,i]$ all constrain the segment F_i. We then appeal to Theorem 4 of Handout 19. If we let s_1 through s_n in that theorem be t_{i+1} through t_{i+n} in forwards order and we let r_1 through r_n in that theorem be t_i through t_{i-n+1} in backwards order, that theorem tells us that we have precisely enough constraints to uniquely define F_i. It then remains to check that the segments $F_i([t_i, t_{i+1}])$ thus defined do fit together properly to form a spline on the knot sequence (t_k).

4 B-splines

Every n-ic polynomial curve can be expressed as a linear combination of the polynomials in the Bernstein basis. Indeed, one way to think of Bézier points is as the coefficients in this expansion.

In a similar way, for any knot sequence (t_k), we can define a basis for the set of all spline curves with that knot sequence. The real-valued spline functions (B_k) in this basis are called *B-splines*. Another way to think about the de Boor points of a spline curve is as the coefficients that appear when we expand that spline curve as a linear combination of the B-splines:

$$F(t) = \sum_k P_k B_k(t).$$

We aren't going to say much about B-splines in this handout. For a little about them, see Farin; for a lot more, see [1]. For our purposes, it is enough to note the following: The B-spline $B_k(t)$ is the real-valued spline function that results if we set the single de Boor point

P_k to the real value 1 and all the rest of the de Boor points P_l for $l \neq k$ to 0. For a simple example, consider the affine case $n = 1$ with all simple knots. The kth B-spline B_k on such a knot sequence rises linearly from 0 to 1 over the interval $[t_k \,..\, t_{k+1}]$ and then falls linearly from 1 back to 0 over the interval $[t_{k+1} \,..\, t_{k+2}]$.

For any degree n and knot sequence (t_k), the corresponding B-splines form a partition of unity. That is, they are nonnegative functions, $B_k(t) \geq 0$ for all k and t, that sum to one: $\sum_k B_k(t) = 1$.

References

[1] Carl de Boor (1978), *A Practical Guide to Splines*, Springer-Verlag.

[2] Lyle Ramshaw (1989), Blossoms are polar forms, Computer Aided Geometric Design **6**, 323–358.