1 Basic Theory

1.1 Points and Vectors

Real life methods for constructing curves and surfaces often start with points and vectors, which is why we start with a short discussion of the properties of these mathematical entities. The material in this section applies to both two-dimensional and three-dimensional points and vectors, while the examples are given in two-dimensions.

Points and vectors are different mathematical entities. A point has no dimensions; it represents a location in space. A vector, on the other hand, has no well-defined location and its only attributes are direction and magnitude. People tend to confuse points and vectors because it is natural to associate a point $\mathbf{P}$ with the vector $\mathbf{v}$ that points from the origin to $\mathbf{P}$ (Figure 1.1a). This association is useful, but the reader should bear in mind that $\mathbf{P}$ and $\mathbf{v}$ are different.

Both points and vectors are represented by pairs or triplets of real numbers, but these numbers have different meanings. A point with coordinates $(3,4)$ is located 3 units to the right of the y axis and 4 units above the x axis. A vector with components $(3,4)$, however, points in direction $4/3$ (it moves 3 units in the x direction for every 4 units in the y direction, so its slope is $4/3$) and its magnitude is $\sqrt{3^2+4^2}=5$. It can be located anywhere.

In mathematics, entities are always associated with operations. An entity that cannot be operated on is generally not useful. Thus, we discuss operations on points and vectors. The first operation is to multiply a point $\mathbf{P}$ by a real number α. The product $\alpha\mathbf{P}$ is a point on the line connecting $\mathbf{P}$ to the origin (Figure 1.1b). Note that this line is infinite and $\alpha\mathbf{P}$ can be located anywhere on it, depending on the value of α.

The next operation is subtracting points. Let $\mathbf{P}_0=(x_0,y_0)$ and $\mathbf{P}_1=(x_1,y_1)$ be two points. The difference $\mathbf{P}_1-\mathbf{P}_0=(x_1-x_0,y_1-y_0)=(\Delta x,\Delta y)$ is well defined. It is the vector (the direction and distance) from $\mathbf{P}_0$ to $\mathbf{P}_1$ (Figure 1.1b).

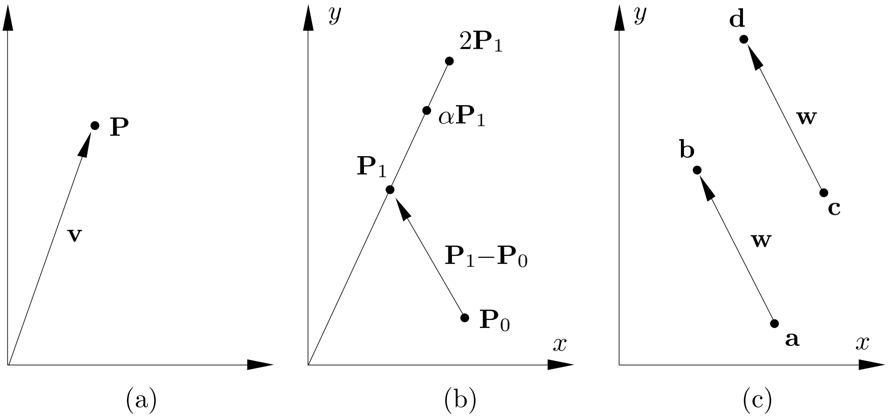


Figure 1.1: Operations on Points.

Figure 1.1c shows two pairs of points **a b** and **c d**. Points **a** and **c** are different and so are **b** and **d**. The vectors $\mathbf{b}-\mathbf{a}$ and $\mathbf{d}-\mathbf{c}$, however, are identical

Example: The two points $\mathbf{P}_0=(5,4)$ and $\mathbf{P}_1=(2,6)$ are subtracted to produce the pair $\mathbf{P}_1-\mathbf{P}_0=(-3,2)$. The new pair is a vector, because it represents a direction and a distance. To get from $\mathbf{P}_0$ to $\mathbf{P}_1$, we need to move -3 units in the x direction and 2 units in the y direction. Similarly, $\mathbf{P}_0-\mathbf{P}_1$ is the direction from $\mathbf{P}_1$ to $\mathbf{P}_0$. The distance between the points is $\sqrt{(-3)^2+2^2}$. These properties do not depend on the particular coordinate axes used. If we translate the origin—or, equivalently, translate the points—m units in the x direction and n units in the y direction, the points will have new coordinates, but the difference will not change. The same property (the difference of points being independent of the coordinate axes) holds after rotation, scaling, shearing, and reflection: the so-called *affine transformations* (or mappings). This is why the operation of subtracting two points is affinely invariant. (Note that the product $\alpha\mathbf{P}$ is also affinely invariant.)

The sum of a point and a vector is well defined and is a point. Figure 1.2a shows the two sums $\mathbf{P}_1^*=\mathbf{P}_1+\mathbf{v}$ and $\mathbf{P}_2^*=\mathbf{P}_2+\mathbf{v}$. It is easy to see that the relative positions of $\mathbf{P}_1^*$ and $\mathbf{P}_2^*$ are the same as those of $\mathbf{P}_1$ and $\mathbf{P}_2$. Another way to look at the sum $\mathbf{P}+\mathbf{v}$ is to observe that it moves us away from $\mathbf{P}$, which is a point, in a certain direction and by a certain distance, thereby bringing us to another point. Yet another way of showing the same thing is to rewrite the relation $\mathbf{a}-\mathbf{b}=\mathbf{v}$ as $\mathbf{a}=\mathbf{b}+\mathbf{v}$, which shows that the sum of point $\mathbf{b}$ and vector $\mathbf{v}$ is a point $\mathbf{a}$.

Given any two points $\mathbf{P}_0$ and $\mathbf{P}_2$, the expression $\mathbf{P}_0+\alpha(\mathbf{P}_2-\mathbf{P}_0)$ is the sum of a point and a vector, so it is a point that we can denote by $\mathbf{P}_1$. The vector $\mathbf{P}_2-\mathbf{P}_0$ points from $\mathbf{P}_0$ to $\mathbf{P}_2$, so adding it to $\mathbf{P}_0$ produces a point on the line connecting $\mathbf{P}_0$ to $\mathbf{P}_2$. Thus, we conclude that the three points $\mathbf{P}_0$, $\mathbf{P}_1$, and $\mathbf{P}_2$ are collinear. Note that the expression $\mathbf{P}_1=\mathbf{P}_0+\alpha(\mathbf{P}_2-\mathbf{P}_0)$ can be written $\mathbf{P}_1=(1-\alpha)\mathbf{P}_0+\alpha\mathbf{P}_2$, showing that $\mathbf{P}_1$ is a linear combination of $\mathbf{P}_0$ and $\mathbf{P}_2$. In general, any of three collinear points can be written as a linear combination of the other two. Such points are not independent.

◇ **Exercise 1.1:** Given the three points $\mathbf{P}_0=(1,1)$, $\mathbf{P}_1=(2,2.5)$, and $\mathbf{P}_2=(3,4)$, are they collinear?

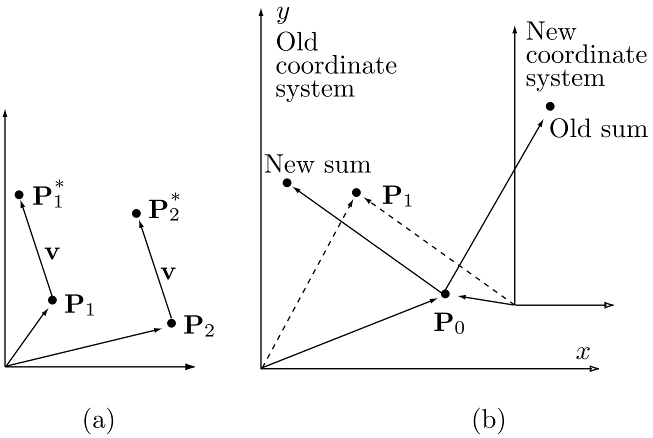


Figure 1.2: (a) Adding a Point and a Vector. (b) Adding Points.

◇ **Exercise 1.2:** What can we say about four collinear points?

The next operation to consider is the sum of points. In general this operation is not well defined. We intuitively feel that adding two points should be done like adding vectors. The lines connecting the points with the origin should be added, to produce a sum vector. In fact, as Figure 1.2b shows, this operation depends on the coordinate axes. Moving the origin (or moving the points) will move the sum of the vectors a different distance or in a different direction, thereby changing the sum of the points. This is why the sum of points is, in general, undefined.

Example: Given the two points $(5,3)$ and $(7,-2)$, we add them to produce $(12,1)$. We now move the two points one unit to the left to become $(4,3)$ and $(6,-2)$. Their new sum is $(10,1)$, a point located two units to the left of the original sum.

There is, however, one important special case where the sum of points is well defined, the so-called *barycentric sum.* If we multiply each point by a weight and if the weights add up to 1, then the sum of the weighted points is affinely invariant, i.e., it is a valid point. Here is the (simple) proof: If $\sum_{i=0}^{n} w_i = 1$, then

$$\begin{aligned}\sum_{i=0}^{n} w_i\mathbf{P}_i &= \mathbf{P}_0 + \sum_{i=1}^{n} w_i\mathbf{P}_i - (1-w_0)\mathbf{P}_0\\ &= \mathbf{P}_0 + w_1\mathbf{P}_1 + w_2\mathbf{P}_2 + \cdots + w_n\mathbf{P}_n - (w_1+\cdots+w_n)\mathbf{P}_0\\ &= \mathbf{P}_0 + w_1(\mathbf{P}_1-\mathbf{P}_0) + w_2(\mathbf{P}_2-\mathbf{P}_0) + \cdots + w_n(\mathbf{P}_n-\mathbf{P}_0)\\ &= \mathbf{P}_0 + \sum_{i=1}^{n} w_i(\mathbf{P}_i-\mathbf{P}_0). \qquad (1.1)\end{aligned}$$

This is the sum of the point $\mathbf{P}_0$ and the vector $\sum_{i=1}^{n} w_i(\mathbf{P}_i - \mathbf{P}_0)$, and we already know that the sum of a point and a vector is a point.

Notice that the proof above does not assume that the weights are nonnegative and barycentric weights can in fact be negative. A little experiment may serve to convince the sceptics. Given two points (a, b) and (c, d) we construct the barycentric sum $(x, y) = -0.5(a, b) + 1.5(c, d)$. If we now translate both points by the vector (α, β), the sum is modified to

$$-0.5(a+\alpha, b+\beta) + 1.5(c+\alpha, d+\beta) = -0.5(a, b) + 1.5(c, d) + (\alpha, \beta) = (x, y) + (\alpha, \beta).$$

The barycentric sum (x, y) is translated by the same vector.

Mathematically-savvy readers may be familiar with the concept of normalization. Given a set of weights w_i that add up to $\alpha \neq 1$, they can be normalized by dividing each weight by the sum α. Thus, if we need a barycentric sum of certain quantities P_i and we are given nonbarycentric weights w_i, we can compute

$$\sum_{i=1}^{n} \frac{w_i}{\sum_{j=1}^{n} w_j} P_i = \sum_{i=1}^{n} \left(\frac{w_i}{\alpha}\right) P_i = \sum_{i=1}^{n} r_i P_i,$$

where the new, normalized weights r_i are barycentric.

Barycentric sums are common in curve and surface design. This book has numerous examples of curves and surfaces that are constructed as weighted sums of points, and they all must be barycentric. When a curve consists of a non-barycentric weighted sum of points, its shape depends on the particular coordinate system used. The shape changes when either the curve or the coordinate axes are moved or are affinely transformed. Such a curve is ill conditioned and cannot be used in practice.

The Isotropic Principle

Given a curve that's constructed as the sum

$$\mathbf{P}(t) = \sum w_i \mathbf{P}_i + \sum u_i \mathbf{v}_i,$$

where $\mathbf{P}_i$ are points and $\mathbf{v}_i$ are vectors, the curve is independent of the particular coordinate system used if and only if the weights w_i are barycentric. There is no similar requirement for the u_i weights. Notice that the points can be data points, control points, or any other points. The vectors can be tangents, second derivatives or any other vectors, but the statement above is always true. This statement is sometimes known as the *isotropic principle.*

A special case is the barycentric sum of two points $(1-t)\mathbf{P}_0 + t\mathbf{P}_1$. This is a point on the line from $\mathbf{P}_0$ to $\mathbf{P}_1$. In fact, the entire straight segment from $\mathbf{P}_0$ to $\mathbf{P}_1$ is obtained when t is varied from 0 to 1 (Figure 1.3a). To see this, we write $\mathbf{P}(t) = (1-t)\mathbf{P}_0 + t\mathbf{P}_1$. Clearly, $\mathbf{P}(0) = \mathbf{P}_0$ and $\mathbf{P}(1) = \mathbf{P}_1$. Also, since $\mathbf{P}(t) = t(\mathbf{P}_1 - \mathbf{P}_0) + \mathbf{P}_0$, $\mathbf{P}(t)$ is a linear function of t, which implies a straight line in t. The tangent vector is the derivative $\frac{d\mathbf{P}}{dt}$ and it is the constant $\mathbf{P}_1 - \mathbf{P}_0$, the direction from $\mathbf{P}_0$ to $\mathbf{P}_1$. Notice that this derivative is a vector, not a number. Selecting $t = 1/2$ yields $\mathbf{P}(0.5) = 0.5\mathbf{P}_1 + 0.5\mathbf{P}_0$, the midpoint between $\mathbf{P}_0$ and $\mathbf{P}_1$.

The concept of barycentric weights is so useful that the two numbers $1-t$ and t are termed the *barycentric coordinates* of point $\mathbf{P}(t)$ with respect to $\mathbf{P}_0$ and $\mathbf{P}_1$.

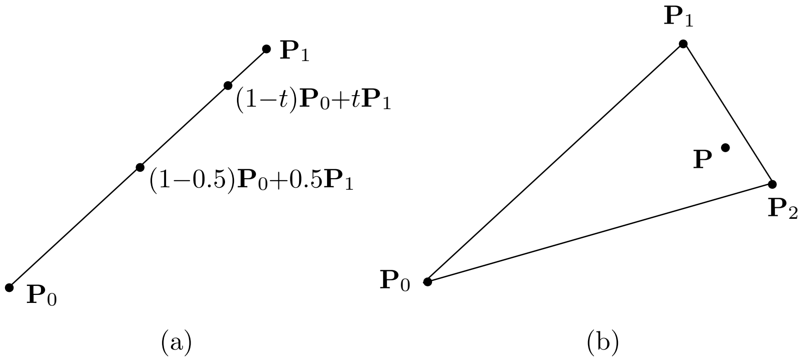


Figure 1.3: Line and Triangle.

The word *barycentric* seems to have first been used in [Dupuy 48]. It is derived from *barycenter*, meaning "center of gravity," because such weights are used to calculate the center of gravity of an object. Barycentric weights have many uses in geometry in general and in curve and surface design in particular.

Another useful example is the barycentric coordinates of a two-dimensional point with respect to the three corners of a triangle. Imagine a triangle with corners $\mathbf{P}_0$, $\mathbf{P}_1$, and $\mathbf{P}_2$ (Figure 1.3b). Any point $\mathbf{P}$ inside the triangle can be expressed as the weighted combination

$$\mathbf{P} = u\mathbf{P}_0 + v\mathbf{P}_1 + w\mathbf{P}_2, \quad \text{where} \quad u + v + w = 1. \tag{1.2}$$

The proof is that Equation (1.2) can be written explicitly as three equations in the three unknowns u, v, and w:

$$\begin{aligned} P_x &= uP_{0x} + vP_{1x} + wP_{2x}, \\ P_y &= uP_{0y} + vP_{1y} + wP_{2y}, \\ 1 &= u + v + w. \end{aligned} \tag{1.3}$$

The solutions are unique provided that the three equations are independent. ◂

◇ **Exercise 1.3:** Show that Equation (1.3) consists of three independent equations if the three points $\mathbf{P}_0$, $\mathbf{P}_1$, and $\mathbf{P}_2$ are independent.

◇ **Exercise 1.4:** Show that the barycentric coordinates of point $\mathbf{P}_0$ with respect to $\mathbf{P}_0$, $\mathbf{P}_1$, and $\mathbf{P}_2$ are $(1, 0, 0)$. Also discuss the barycentric coordinates of points outside the triangle.

Example: Let $\mathbf{P}_0 = (1, 1)$, $\mathbf{P}_1 = (2, 3)$, $\mathbf{P}_2 = (5, 1)$, and $\mathbf{P} = (2, 2)$. Equation (1.3) becomes

$$(2, 2) = u(1, 1) + v(2, 3) + w(5, 1); \quad u + v + w = 1,$$

or

$$\begin{aligned}2&=u+2v+5w,\\ 2&=u+3v+w,\\ 1&=u+v+w,\end{aligned}\qquad\text{which yield}\qquad\left\{\begin{aligned}u&=3/8,\\ v&=1/2,\\ w&=1/8.\end{aligned}\right.$$

◇ **Exercise 1.5:** For a given triangle, calculate the (x, y, z) coordinates of the point with barycentric coordinates $(1/3, 1/3, 1/3)$. This point is called the *centroid* and is one of many centers that can be defined for a triangle. (Imagine cutting the triangle out of a piece of cardboard. If you try to support it at the centroid, it will balance.)

(This material is useful for the triangular Bézier surface patches described in Section 6.23.)

The barycentric combination is the most fundamental operation on points; so much so that it is used to define affine transformations. The definition is: a transformation of points in space is affine if it leaves barycentric combinations invariant. Hence, if $\mathbf{P} = \sum w_i\mathbf{P}_i$ and $\sum w_i = 1$, and if $\mathbf{T}$ is an affine transformation, then $\mathbf{TP} = \sum w_i\mathbf{TP}_i$. All common geometric transformations—such as scaling, shearing, rotation, and reflection—are affine.

Note: The difference of two points is a vector. We can consider such a difference a weighted sum where the weights add up to zero (they are $+1$ and -1). It turns out that a weighted sum of points where the weights add up to zero is a vector. To prove this, let

$$\mathbf{Q} = \sum_{i=1}^{n} w_i\mathbf{P}_i, \qquad \text{where} \qquad \sum w_i = 0,$$

and let $\mathbf{P}$ be a point. The sum $\mathbf{R} = \mathbf{Q} + \mathbf{P}$ is barycentric (since its coefficients add up to 1) and is therefore a point. The difference $\mathbf{R} - \mathbf{P} = \mathbf{Q}$ is a difference of points and is therefore a vector.

Note: Multiplying a point by a number produces a point, so if $\mathbf{P}$ is a point, then $-\mathbf{P}$ is also a point. It is located on the line connecting $\mathbf{P}$ with the origin, on the other side of the origin from $\mathbf{P}$. Once this is understood, we notice that the sum of points $\mathbf{P} + \mathbf{Q}$ can be written as the difference of points $\mathbf{P} - (-\mathbf{Q})$. This difference is, of course, the vector from point $-\mathbf{Q}$ to point $\mathbf{P}$ (Figure 1.4), so we conclude that the sum $\mathbf{P} + \mathbf{Q}$ of two points is well defined but is not very useful, since it tells us something about the relative positions of $\mathbf{P}$ and $-\mathbf{Q}$, not of $\mathbf{P}$ and $\mathbf{Q}$. Assuming that Figure 1.4 depicts the points $\mathbf{Q} = (-5, -1)$ and $\mathbf{P} = (4, 3)$, the sum $\mathbf{P} + \mathbf{Q}$ equals $(-5, -1) + (4, 3) = (-1, 2)$. This shows that in order to get from point $-\mathbf{Q}$ to point $\mathbf{P}$, we need to move one negative step in the x direction for every two steps in the y direction.

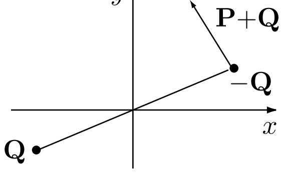


Figure 1.4: Adding Two Points.

◇ **Exercise 1.6:** Let **P** and **Q** be points and let **v** and **w** be vectors. What is the sum $\mathbf{P}-\mathbf{Q}+\mathbf{v}+\mathbf{w}$?

1.1.1 Operations on Vectors

The notation $|\mathbf{P}|$ indicates the magnitude (or absolute value) of vector **P**. Vector addition is defined by adding the individual elements of the vectors being added: $\mathbf{P}+\mathbf{Q}=(P_x,P_y,P_z)+(Q_x,Q_y,Q_z)=(P_x+Q_x,P_y+Q_y,P_z+Q_z)$. This operation is both commutative $\mathbf{P}+\mathbf{Q}=\mathbf{Q}+\mathbf{P}$ and associative $\mathbf{P}+(\mathbf{Q}+\mathbf{T})=(\mathbf{P}+\mathbf{Q})+\mathbf{T}$. Subtraction of vectors $(\mathbf{P}-\mathbf{Q})$ is done similarly and results in the vector from **Q** to **P**.

Vectors can be multiplied in three different ways as follows:

1. The product of a real number α by a vector **P** is denoted by $\alpha\mathbf{P}$ and produces the vector $(\alpha x,\alpha y,\alpha z)$. It changes the magnitude of **P** by a factor α, but does not change its direction.

2. The dot product of two vectors is denoted by $\mathbf{P}\bullet\mathbf{Q}$ and is defined as the scalar

$$(P_x,P_y,P_z)(Q_x,Q_y,Q_z)^T=\mathbf{P}\mathbf{Q}^T=P_xQ_x+P_yQ_y+P_zQ_z.$$

This also equals $|\mathbf{P}|\,|\mathbf{Q}|\cos\theta$, where θ is the angle between the vectors. The dot product of perpendicular vectors (also called *orthogonal vectors*) is therefore zero. The dot product is commutative, $\mathbf{P}\bullet\mathbf{Q}=\mathbf{Q}\bullet\mathbf{P}$.

The triple product $(\mathbf{P}\bullet\mathbf{Q})\mathbf{R}$ is sometimes useful. It can be represented as

$$\begin{aligned}(\mathbf{P}\bullet\mathbf{Q})\mathbf{R}&=(P_xQ_x+P_yQ_y+P_zQ_z)(R_x,R_y,R_z)\\&=\big((P_xQ_x+P_yQ_y+P_zQ_z)R_x,(P_xQ_x+P_yQ_y+P_zQ_z)R_y,\\&\qquad(P_xQ_x+P_yQ_y+P_zQ_z)\big)R_z\\&=(Q_x,Q_y,Q_z)\begin{pmatrix}P_xR_x&P_yR_x&P_zR_x\\P_xR_y&P_yR_y&P_zR_y\\P_xR_z&P_yR_z&P_zR_z\end{pmatrix}\\&=\mathbf{Q}(\mathbf{PR}),\end{aligned}\tag{1.4}$$

where the notation (**PR**) stands for the 3×3 matrix of Equation (1.4).

3. The cross product of two vectors (also called the *vector product*) is denoted by $\mathbf{P}\times\mathbf{Q}$ and is defined as the vector

$$(P_2Q_3-P_3Q_2,-P_1Q_3+P_3Q_1,P_1Q_2-P_2Q_1).\tag{1.5}$$

It is easy to show that $\mathbf{P}\times\mathbf{Q}$ is perpendicular to both **P** and **Q**.

◇ **Exercise 1.7:** Show it!

The following expressions show how $\mathbf{P}\times\mathbf{Q}$ can be expressed by means of a determinant:

$$\mathbf{P}\times\mathbf{Q}=\begin{vmatrix}\mathbf{i}&\mathbf{j}&\mathbf{k}\\P_1&P_2&P_3\\Q_1&Q_2&Q_3\end{vmatrix}=\mathbf{i}\begin{vmatrix}P_2&P_3\\Q_2&Q_3\end{vmatrix}-\mathbf{j}\begin{vmatrix}P_1&P_3\\Q_1&Q_3\end{vmatrix}+\mathbf{k}\begin{vmatrix}P_1&P_2\\Q_1&Q_2\end{vmatrix}$$

$$= (P_2Q_3 - P_3Q_2, -P_1Q_3 + P_3Q_1, P_1Q_2 - P_2Q_1),$$

or, alternatively, by means of a matrix

$$= (Q_1, Q_2, Q_3)\begin{pmatrix} 0 & P_3 & -P_2 \\ -P_3 & 0 & P_1 \\ P_2 & -P_1 & 0 \end{pmatrix}. \tag{1.6}$$

◇ **Exercise 1.8:** The cross-product $\mathbf{P}\times\mathbf{Q}$ is perpendicular to both $\mathbf{P}$ and $\mathbf{Q}$. In what direction does it point?

The cross-product is not commutative and is not associative. It is, however, distributive with respect to addition or subtraction of vectors. Hence, $\mathbf{P}\times(\mathbf{Q}\pm\mathbf{T}) = \mathbf{P}\times\mathbf{Q}\pm\mathbf{P}\times\mathbf{T}$.

The magnitude of $\mathbf{P}\times\mathbf{Q}$ equals $|\mathbf{P}|\,|\mathbf{Q}|\sin\theta$, where θ is the angle between the two vectors. The cross-product, therefore, has a simple geometric interpretation. Its magnitude equals the area of the parallelogram defined by the two vectors.

◇ **Exercise 1.9:** Given that $\mathbf{P}\times\mathbf{Q} = 0$, what does it tell us about the vectors involved?

◇ **Exercise 1.10:** Derive the vector line equation for the straight segment between two given points $\mathbf{P}_1$ and $\mathbf{P}_2$.

1.1.2 The Scalar Triple Product

The scalar triple product of three vectors, $\mathbf{P}$, $\mathbf{Q}$, and $\mathbf{R}$, is defined as

$$\begin{aligned} S = \mathbf{P}\bullet(\mathbf{Q}\times\mathbf{R}) &= P_1(Q_2R_3 - Q_3R_2) + P_2(Q_3R_1 - Q_1R_3) + P_3(Q_1R_2 - Q_2R_1) \\ &= \begin{vmatrix} P_1 & P_2 & P_3 \\ Q_1 & Q_2 & Q_3 \\ R_1 & R_2 & R_3 \end{vmatrix}. \end{aligned} \tag{1.7}$$

Interchanging two rows in a determinant changes its sign, so interchanging rows twice leaves the determinant unchanged. This is why the triple product is not affected by a cyclic permutation of its three components. We can therefore write

$$S = \mathbf{P}\bullet(\mathbf{Q}\times\mathbf{R}) = \mathbf{Q}\bullet(\mathbf{R}\times\mathbf{P}) = \mathbf{R}\bullet(\mathbf{P}\times\mathbf{Q}).$$

The triple product has a simple geometric interpretation. It equals the volume of the parallelepiped defined by the three vectors. An important corollary is: if the three vectors are coplanar, then the parallelepiped defined by them has zero volume, implying that their scalar triple product is zero. This property is used in Section 2.2.1 to determine whether or not a given polygon is planar.

1.1.3 Projecting a Vector

A common and useful operation on vectors is projecting a vector $\mathbf{a}$ on another vector $\mathbf{b}$. The idea is to break vector $\mathbf{a}$ up into two perpendicular components $\mathbf{c}$ and $\mathbf{d}$, such that $\mathbf{c}$ is in the direction of $\mathbf{b}$.

Figure 1.5a shows that $\mathbf{a} = \mathbf{c} + \mathbf{d}$ and $|\mathbf{c}| = |\mathbf{a}|\cos\alpha$. On the other hand, $\mathbf{a}\bullet\mathbf{b} = |\mathbf{a}|\,|\mathbf{b}|\cos\alpha$, yielding the magnitude of $\mathbf{c}$:

$$|\mathbf{c}| = |\mathbf{a}|\frac{(\mathbf{a}\bullet\mathbf{b})}{|\mathbf{a}|\,|\mathbf{b}|} = \frac{(\mathbf{a}\bullet\mathbf{b})}{|\mathbf{b}|}. \tag{1.8}$$

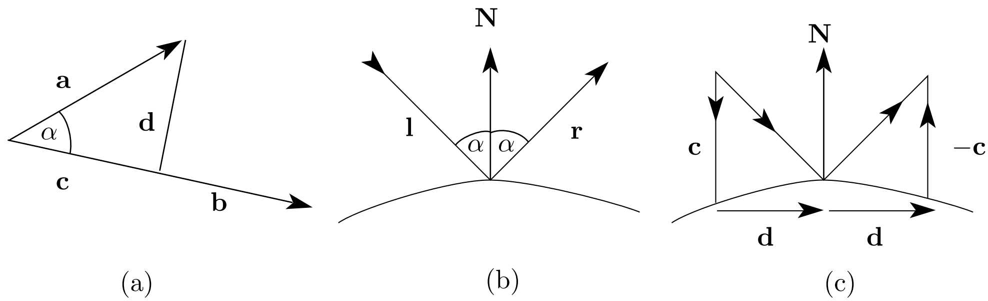


Figure 1.5: Projecting a Vector.

The direction of **c** is identical to the direction of **b**, so we can write vector **c** as

$$\mathbf{c} = |\mathbf{c}|\frac{\mathbf{b}}{|\mathbf{b}|} = \frac{(\mathbf{a}\bullet\mathbf{b})}{|\mathbf{b}|^2}\mathbf{b}. \tag{1.9}$$

Example: Given vectors $\mathbf{a} = (2, 1)$ and $\mathbf{b} = (1, 0)$, we compute the projection of a on b.

$$\mathbf{c} = \frac{(\mathbf{a}\bullet\mathbf{b})}{|\mathbf{b}|^2}\mathbf{b} = \frac{2\times 1 + 1\times 0}{1^2 + 0^2}(2, 0) = (4, 0), \qquad \mathbf{d} = \mathbf{a} - \mathbf{c} = (-2, 1).$$

◇ **Exercise 1.11:** The projection method works also for three-dimensional vectors. Given vectors $\mathbf{a} = (2, 1, 3)$ and $\mathbf{b} = (1, 0, -1)$, calculate the projection of **a** on **b**.

Summary: The following operations have been discussed in this section:

$$\begin{array}{lll} \text{point} - \text{point} = \text{vector}, & \text{scalar}\times\text{point} = \text{point}, & \text{vector} \pm \text{vector} = \text{vector}, \\ \text{scalar}\times\text{vector} = \text{vector}, & \text{point} + \text{vector} = \text{point}, & \text{vector} \bullet \text{vector} = \text{scalar}, \\ & \text{vector} \times \text{vector} = \text{vector}. & \end{array}$$

The operation point + point is left undefined (since it is not useful). A barycentric sum of points is a point, and a weighted sum of points where the weights add up to zero is a vector.

From the dictionary

Vector: (1) A variable quantity that can be resolved into components. (2) A straight line segment whose length is magnitude and whose orientation in space is direction. (3) Any agent (person or animal or microorganism) that carries and transmits a disease.

1.2 Parametric Blending

Parametric blending is a family of techniques that make it possible to vary the value of some quantity in small steps, without any discontinuities. Blending can be thought of as averaging or interpolating. The following are examples:

1. Numbers. The average of the two numbers 15 and 18 is $(15+18)/2 = 16.5$. This can also be written as $0.5\times 15 + 0.5\times 18$, which can be interpreted as the *blend*, or the weighted sum, of the two numbers, where each is assigned a weight of 0.5. When the weights are different, such as $0.9\times 15 + 0.1\times 18$, the result is a blend of 90% 15 and 10% 18.

2. Points. If $\mathbf{P}_1$ and $\mathbf{P}_2$ are points, then the expression $\alpha\mathbf{P}_1 + \beta\mathbf{P}_2$ is a blend of the two points, in which α and β are the weights (or the coefficients). If α and β are nonnegative and $\alpha+\beta = 1$, then the blend is a point on the straight segment connecting $\mathbf{P}_1$ and $\mathbf{P}_2$.

3. Rotations. A rotation in three dimensions is described by means of the rotation angle (one number) and the axis of rotation (three numbers). These four numbers can be combined into a mathematical entity called quaternion and two quaternions can also be blended, resulting in a smooth sequence of rotations that proceeds in small, equal steps from an initial rotation to a final one. This type of blending is useful in computer animation.

4. Curve construction. Given a number of points, a curve can be created as a weighted sum of the points. It has the form $\sum w_i(t)\mathbf{P}_i$, where the weights $w_i(t)$ are barycentric. Such a curve is a *blend* of the points. For each value of t, the blend is different, but we have to make sure that the sum of the weights is always 1. It is possible to blend vectors, in addition to points, as part of the curve, and the weights of the vectors don't have to satisfy any particular requirement. Most of the curve methods described in this book generate a curve as a blend of points, vectors, or both.

A special case of curve construction is the linear blending of two points, which can be expressed as $(1-t)\mathbf{P}_1 + t\,\mathbf{P}_2$ for $0 \le t \le 1$ [this is Equation (2.1)].

5. Surfaces. Using the same principle, points, vectors, and curves can be blended to form a surface patch.

6. Images. Various types of image processing, such as sharpening, blurring, and embossing, are performed by blending an image with a special mask image.

7. It is possible to blend points in nonlinear ways. An intuitive way to get, for example, quadratic blending is to square the two weights of the linear blend. However, the result, which is $\mathbf{P}(t) = (1-t)^2\mathbf{P}_1 + t^2\mathbf{P}_2$, depends on the particular coordinate axes used, since the two coefficients $(1-t)^2$ and t^2 are not barycentric. It turns out that the sum $(1-t)^2 + 2t(1-t) + t^2$ equals 1. As a result, we can use quadratic blending to blend three points, but not two.

Similarly, if we try a cubic blend by simply writing $\mathbf{P}(t) = (1-t)^3\mathbf{P}_1 + t^3\mathbf{P}_2$, we end up with the same problem. Cubic blending can be achieved by adding four terms with weights t^3, $3t^2(1-t)$, $3t(1-t)^2$, and $(1-t)^3$.

We therefore conclude that Bézier methods (Chapter 6) can be used for blending. The Bézier curve is a result of blending several points with the Bernstein polynomials, which add up to unity. Quadratic and cubic blending are special cases of the Bézier blending (or the Bézier interpolation).

1.3 Parametric Curves

As mentioned in the Preface, the main aim of computer graphics is to display an arbitrary surface so that it looks real. The first step toward this goal is an understanding of curves. Once we have an algorithm to calculate and display any curve, we may try to extend it to a surface.

In practice, curves (and surfaces) are specified by the user in terms of points and are constructed in an interactive process. The user starts by entering the coordinates of points, either by scanning a rough image of the desired shape and digitizing certain points on the image, or by drawing a rough shape on the screen and selecting certain points with a pointing device such as a mouse. After the curve has been drawn, the user may want to modify its shape by moving, adding, or deleting points. Such points can be employed in two different ways:

1. We may want the curve to pass through them. Such points are called *data points* and the curve is called an interpolating curve.

2. We may want the points to control the shape of the curve by exerting a “pull” on it. A point may pull part of the curve toward it, allowing the user to change the shape of the curve by moving the point. Generally, however, the curve does not pass through the point. Such points are called *control points* and the curve is called an approximating curve.

A mathematical function $y = f(x)$ can be plotted as a curve. Such a function is the *explicit* representation of the curve. The explicit representation is not general, since it cannot represent vertical lines and is also single-valued. For each value of x, only a single value of y is normally computed by the function.

The *implicit* representation of a curve has the form $F(x, y) = 0$. It can represent multivalued curves (more than one y value for an x value). A common example is the circle, whose implicit representation is $x^2 + y^2 - R^2 = 0$.

The explicit and implicit curve representations can be used only when the function is known. In practical applications—where complex curves such as the shape of a car or of a toaster are needed—the function is normally unknown, which is why a different approach is required.

The curve representation used in practice is called the *parametric representation.* A two-dimensional parametric curve has the form $\mathbf{P}(t) = \big(f(t), g(t)\big)$ or $\mathbf{P}(t) = \big(x(t), y(t)\big)$. The functions f and g become the (x, y) coordinates of any point on the curve, and the points are obtained when the parameter t is varied over a certain interval $[a, b]$, normally $[0, 1]$.

A simple example of a two-dimensional parametric curve is $\mathbf{P}(t) = (2t - 1, t^2)$. When t is varied from 0 to 1, the curve proceeds from the initial point $\mathbf{P}(0) = (-1, 0)$ to the final point $\mathbf{P}(1) = (1, 1)$. The x coordinate is linear in t and the y coordinate varies as t^2.

The first derivative $\frac{d\mathbf{P}(t)}{dt}$ is denoted by $\mathbf{P}^t(t)$, or by $\dot{\mathbf{P}}$, or by $(P^t_x(t), P^t_y(t))$. This derivative is the tangent vector to the curve at any point. The derivative is a vector and not a point because it is the limit of the difference $(\mathbf{P}(t + \Delta) - \mathbf{P}(t))/\Delta$, and the difference of points is a vector. As a vector, the tangent possesses a direction (the direction of the curve at the point) and a magnitude (which indicates the speed of the curve at the point). The tangent, however, is not the slope of the curve. The tangent is a pair of numbers, whereas the slope is a single number. The slope equals $\tan\theta$, where θ is the angle between the tangent vector and the x axis. The slope of a two-dimensional parametric curve is obtained by

$$\frac{dy}{dx}=\frac{\frac{dy}{dt}}{\frac{dx}{dt}}=\frac{P^t_y(t)}{P^t_x(t)}.$$

Example: The curve $\mathbf{P}(t)=(x(t),y(t))=(1+t^2/2,t^2)$. Its tangent vector is $\mathbf{P}^t(t)=(t,2t)$ and the slope is $2t/t=2$. The slope is constant, which indicates that the curve is a straight line. This is also easy to see from the tangent vector. The direction of this vector is always the same since it can be described by saying "for every t steps in the x direction, move $2t$ steps in the y direction."

Example: A circle. Because of its high symmetry, a circle can be represented in different ways. We list four different parametric representations of a circle of radius R centered on the origin.

1. $\mathbf{P}(t)=R(\cos t,\sin t)$, where $0\le t\le 2\pi$. This is identical to the polar representation.

2. Substituting $t=\tan(u/2)$ yields $\mathbf{P}(t)=R[(1-t^2)/(1+t^2),2t/(1+t^2)]$. When $0\le t\le 1$, this generates the first quadrant from $(R,0)$ to $(0,R)$ (see also Figure 1.6a).

3. $\mathbf{P}(t)=R(t,\pm\sqrt{1-t^2})$. When $0\le t\le 1$ this generates the first quadrant from $(0,R)$ to $(R,0)$ and, simultaneously, the third quadrant from $(0,-R)$ to $(-R,0)$.

4. $\mathbf{P}(t)=(0.441,-0.441)t^3+(-1.485,-0.162)t^2+(0.044,1.603)t+(1,0)$. When $0\le t\le 1$, this generates (approximately) the first quadrant from $(1,0)$ to $(0,1)$.

(See also circle example in Section 6.15, and Equation (Ans.31).)

◇ **Exercise 1.12:** Explain how representation 4 is derived.

◇ **Exercise 1.13:** Figure 1.6b shows a polygon inscribed in a circle. It is clear that adding sides to the polygon brings it closer to the circle. Calculate the difference $R-d$ as a function of n, the number of polygon sides.

The particle paradigm: Better insight into the behavior of parametric functions can be gained by thinking of the curve $\mathbf{P}(t)=(x(t),y(t))$ as a path traced out by a hypothetical particle. The parameter t can then be interpreted as time and the first two derivatives $\mathbf{P}^t(t)$ and $\mathbf{P}^{tt}(t)$ can be interpreted as the velocity and acceleration of the particle, respectively. It turns out that different parametric representations of the same curve may have different "speeds." The particle represented by $(\cos t,\sin t)$, for example, "moves" along the circle at speed $\mathbf{P}^t(t)=(-\sin t,\cos t)$, which is constant since $|\mathbf{P}^t(t)|=\sqrt{\sin^2 t+\cos^2 t}=1$. The particle of circle representation 2, on the other hand, moves at the variable velocity

$$\mathbf{P}^t(t)=\left(\frac{-4t}{(1+t^2)^2},\frac{2(1-t^2)}{(1+t^2)^2}\right).$$

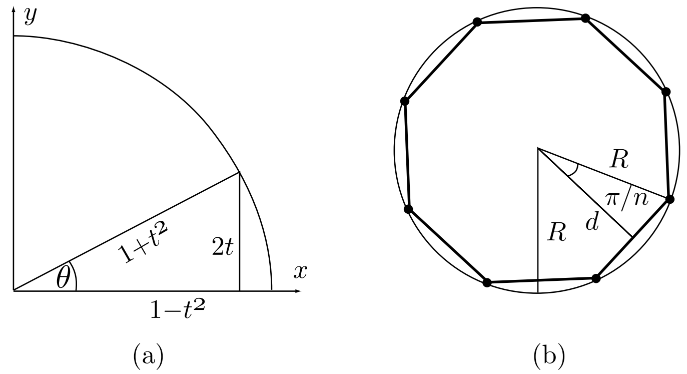


Figure 1.6: (a) A Parametric Representation.
(b) A Polygon Inscribed in a Circle.

◇ **Exercise 1.14:** Show that this velocity does vary with t.

◇ **Exercise 1.15:** What three-dimensional curve is described by the parametric function $(\cos t, \sin t, t)$? (Hint: see Section 2.4.1).

See also page 354 for the parametric representations of the sphere, the ellipsoid, and of the torus as a small circle rotating around a larger circle.

1.4 Properties of Parametric Curves

Generally, it is impossible to tell much about the behavior of a parametric curve $\mathbf{P}(t) = (x(t), y(t))$ by examining the two components $x(t)$ and $y(t)$ separately. Each of the two functions may have features that do not exist in the combination. The reverse is also true—the combined curve may have features not found in any of the two components.

Here is an example of two smooth curves whose combination is a parametric plane curve with a cusp (a sharp corner). The following two curves are polynomials in t:

$$x(t) = -18t^2 + 18t + 2, \qquad y(t) = -16t^3 + 24t^2 - 12t + 5, \quad \text{where} \quad 0 \le t \le 1.$$

They are smooth, since their derivatives $x'(t) = -36t + 18$ and $y'(t) = -48t^2 + 48t - 12$ are continuous in the range $0 \le t \le 1$. However, the combined curve

$$\mathbf{P}(t) = (0, -16)t^3 + (-18, 24)t^2 + (18, -12)t + (2, 5)$$

has a sharp corner (a cusp or a kink), because its tangent vector

$$\mathbf{P}^t(t) = 3(0, -16)t^2 + 2(-18, 24)t + (18, -12)$$

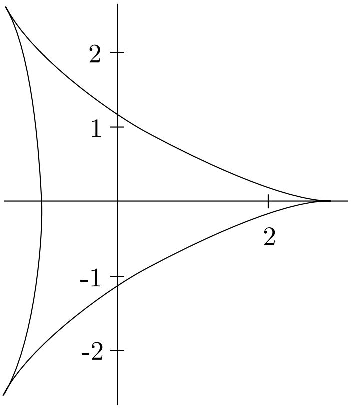
satisfies $\mathbf{P}^t(0.5) = (0, 0)$.

◇ **Exercise 1.16:** Find two curves $x(t)$ and $y(t)$, each with a cusp, such that the combined curve $\mathbf{P}(t) = (x(t), y(t))$ is smooth.

The parametric curves used in computer graphics are normally based on polynomials, since polynomials are simple functions that are easy to calculate and are flexible enough to create many different shapes. However, in principle, any functions can be used to create a parametric curve. Here is an example that uses the smooth sine and cosine curves to create the nonsmooth parametric curve shown on the right. It is defined by the simple expression

$$\mathbf{P}(t) = (2\cos(t) + \cos(2t), 2\sin(t) - \sin(2t)),$$

where $0 \le t \le 2\pi$. This curve has cusps at $t = 0$, $t = 0.261799$, and $t = 0.523599$. Another example of a parametric curve that's not a simple polynomial is the circular Bézier curve, Equation (4.141) of [Salomon 99].



◇ **Exercise 1.17:** Find three curves $x(t)$, $y(t)$, and $z(t)$, each a cubic polynomial, such that the combined curve $\mathbf{P}(t) = (x(t), y(t), z(t))$ is not a cubic polynomial.

Note. A word about the notation used here. We have used the letter $\mathbf{P}$ to denote both points and curves. The same letter is later used to denote surfaces. In spite of using the same letter, the notation is unambiguous. It is always easy to tell what a particular $\mathbf{P}$ stands for by counting the number of free parameters. Something like $\mathbf{P}(u, w)$ denotes a surface since it depends on two variable parameters, whereas $\mathbf{P}(0, w)$ is a curve and $\mathbf{P}(u_0, 1)$ (for a fixed u_0) is a point.

One important feature of curves is *independence of the coordinate axes.* We don't want the curve to change shape when the coordinate axes (or the points defining the curve) are moved rigidly or rotated. Here is an example of how such a thing can happen. Consider the parametric curve

$$\mathbf{P}(t) = (1-t)^3\mathbf{P}_0 + t^3\mathbf{P}_1 = \big((1-t)^3x_0 + t^3x_1, (1-t)^3y_0 + t^3y_1\big).$$

It is easy to see that $\mathbf{P}(0) = \mathbf{P}_0$ and $\mathbf{P}(1) = \mathbf{P}_1$ (the curve passes through the two points). What kind of a curve is $\mathbf{P}(t)$? The tangent vector of our curve is

$$\left(\frac{dx}{dt}, \frac{dy}{dt}\right) = \big(-3(1-t)^2x_0 + 3t^2x_1, -3(1-t)^2y_0 + 3t^2y_1\big).$$

To calculate the slope, we have to select actual points. We start with the two points $\mathbf{P}_0 = (0,0)$ and $\mathbf{P}_1 = (5,6)$. The slope of the curve is

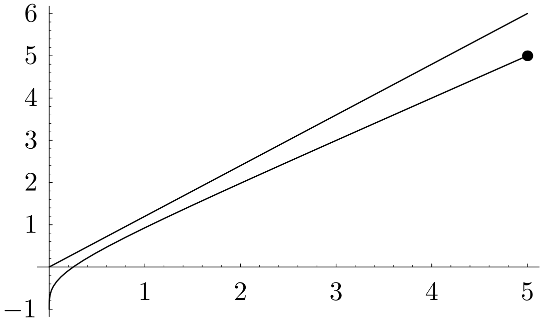
$$\frac{dy}{dx} = \frac{dy}{dt}\Big/\frac{dx}{dt} = \frac{-3(1-t)^2 0 + 3t^2 \times 6}{-3(1-t)^2 0 + 3t^2 \times 5} = \frac{6}{5} = \text{constant},$$

so the curve is a straight line.

Next, we translate both points by the same amount $(0,-1)$, so the new points are $\mathbf{P}_0=(0,-1)$ and $\mathbf{P}_1=(5,5)$. The new slope is

$$\frac{3(1-t)^2+15t^2}{15t^2}=\frac{1}{5}\left(\frac{1}{t}-1\right)+1.$$

It is no longer constant and therefore the curve is no longer a straight line (Figure 1.7). The curve has changed its shape just because its endpoints have been moved!



```
(* non-barycentric weights example *)
Clear[p0,p1,g1,g2,g3,g4];
p0={0,0}; p1={5,6};
g1=ParametricPlot[(1-t)^3 p0+t^3 p1,{t,0,1},PlotRange->All, Compiled->False,
 DisplayFunction->Identity];
g3=Graphics[{AbsolutePointSize[4], {Point[p0],Point[p1]} }];
p0={0,-1}; p1={5,5};
g2=ParametricPlot[(1-t)^3 p0+t^3 p1,{t,0,1},PlotRange->All, Compiled->False,
 PlotStyle->AbsoluteDashing[{2,2}], DisplayFunction->Identity];
g4=Graphics[{AbsolutePointSize[4], {Point[p0],Point[p1]} }];
Show[g2,g1,g3,g4, DisplayFunction->$DisplayFunction, DefaultFont->{"cmr10", 10}];
```

Figure 1.7: Effect of Nonbarycentric Weights.

It turns out that a curve of the form $\mathbf{P}(t)=\sum_{i=0}^{n} w_i(t)\mathbf{P}_i$, is independent of the particular coordinate axes used if $\sum_{i=0}^{n} w_i(t)=1$. This is arguably the most important property of barycentric weights.

It is easy to extend the concept of parametric curves to three dimensions (space curves) with two minor differences (1) $\mathbf{P}(t)$ should be of the form $\big(x(t),y(t),z(t)\big)$ and (2) the slope of a three-dimensional curve is undefined. Such a curve has a tangent vector $d\mathbf{P}/dt$, but not a slope.

◇ **Exercise 1.18:** Show that the parametric curve

$$\mathbf{P}(t)=\mathbf{P}+2\alpha(\mathbf{Q}-\mathbf{P})t+(1-2\alpha)(\mathbf{Q}-\mathbf{P})t^2,\quad 0\le t\le 1 \qquad (1.10)$$

(where α is any real number) is a straight line, even though it is a polynomial of degree 2 in t. Note that the curve goes from point $\mathbf{P}$ to point $\mathbf{Q}$.

1.4.1 Uniform and Nonuniform Parametric Curves

So far, we have assumed that the parameter t of a parametric curve $\mathbf{P}(t) = (x(t), y(t))$ varies in the interval $[0, 1]$. It is also possible to vary t in other ranges, and such curves may be useful in special applications. This idea arises naturally when we try to fit a curve to a given set of data points. One question that should be answered in such a case is what value should the parameter t have at each point. It turns out that this is both a problem and an opportunity. A practical, interactive algorithm for curve design should make it possible to treat the values of t at the data points as parameters, and therefore to produce an entire family of curves, all of whose members pass through the given data points (but behave differently between points). This gives the designer an extra tool that can be used to construct the right curve.

The two approaches to this problem are (1) increment t by one for each point and (2) increment t by different values. The former approach yields a *uniform* parametric curve, while the latter results in a *nonuniform* parametric curve. Uniform parametric curves are normally easy to calculate and they produce good results when the points are roughly equally spaced. However, when the spacing of the points is very different, a uniform curve may look strange and unnatural, even though it passes through all the data points. This is when a nonuniform parametric curve should be used.

If the spacings of the points are far from uniform, it is common to increase the value of t at point $\mathbf{P}_i$ by the distance $|\mathbf{P}_i - \mathbf{P}_{i-1}|$. Notice that this distance is the chord length from point $\mathbf{P}_{i-1}$ to point $\mathbf{P}_i$. If this convention is used, then t starts at zero and is assigned the accumulated chord length at every data point. If the curve does not oscillate much between data points, the chord length is a good approximation to the arc length of the curve, with the result that t is assigned, in such a case, values that are close to the arc length. A curve $\mathbf{P}(s)$ where the parameter is the arc length s has a tangent vector $\mathbf{P}^s(s)$ of magnitude one (it's a unit vector). If we express such a curve as $\mathbf{P}(s) = (x(s), y(s))$, then $(x^s(s), y^s(s))$ is a unit vector, which implies that $|x^s(s)| \leq 1$ and $|y^s(s)| \leq 1$. This, in turn, means that the slopes of both curves $x(s)$ and $y(s)$ are bounded between -1 and $+1$, so the two curves are never too steep and are generally well behaved.

1.4.2 Curve Continuity

In practice, a complete curve is often made up of segments, so it is important to understand how individual segments can be connected. There are two types of curve continuities: *geometric* and *parametric*. If two consecutive segments meet at a point, the total curve is said to have G^0 geometric continuity. (It may look as in Figure 1.8a.) If, in addition, the directions of the tangent vectors of the two segments are the same at the point, the curve has G^1 geometric continuity at the point. The two segments connect smoothly (Figure 1.8b). In general, a curve has geometric continuity G^n at a join point if every pair of the first n derivatives of the two segments have the same direction at the point. If the same derivatives also have identical magnitudes at the point, then the curve is said to have C^n parametric continuity at the point.

We can refer to C^0, C^1, and C^2 as point, tangent, and curvature continuities, respectively. Figure 1.9 illustrates the geometric meanings of the three types. In part C^0 of the figure, the curve is continuous at the interior point, but its tangent is not. The curve changes its direction abruptly at the point; it has a kink. In part C^1, both

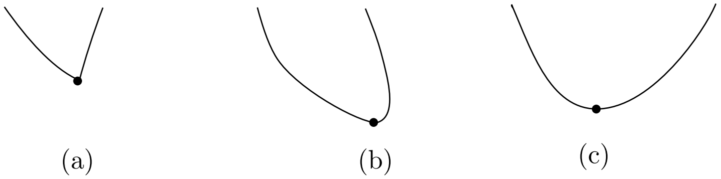


Figure 1.8: (a) G^0 Continuity (a Sharp Corner). (b) G^1 Continuity (a Smooth Connection). (c) G^2 Continuity (a Tight Curve).

the curve and its tangent are continuous at the interior point, but the curve changes its shape at the point from a straight line (zero curvature) to a curved line (nonzero curvature). Thus, the curvature is discontinuous at the point. In part C^2 the curve starts curving before it reaches the interior point, in order to preserve its curvature at the point. Generally, high continuity results in a smoother curve.

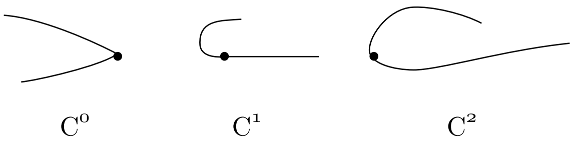


Figure 1.9: Three Curve Continuities.

A C^k continuity is more restrictive than G^k, so a curve that has C^k continuity at a join point also has G^k continuity at the point, but there is an exception. Imagine two segments connecting at a point, where both have tangent vectors of $(0,0,0)$ at the point. The vectors are identical, so the curve has C^1 continuity at the point. However, Exercise 5.3 (page 146) shows that the two segments may move in different directions at the point, in which case the curve will not have G^1 continuity.

Parameter Substitution

Instead of naming the parameter t, we can give it a different name. Moreover, we can use a function of t as the parameter. It can be shown that if $g(t)$ is a function that increases monotonically with t (i.e., if $t_2 > t_1$ implies $g(t_2) > g(t_1)$), then the curve $\mathbf{P}(g(t))$ will have the same shape as $\mathbf{P}(t)$ (although $g(t)$ will normally have to vary in a different range than t).

For two-dimensional curves, the substitution does not affect the slope of the curve since

$$\frac{\frac{dy(g)}{dg}\Big/\frac{dg(t)}{dt}}{\frac{dx(g)}{dg}\Big/\frac{dg(t)}{dt}} = \frac{\frac{dy(t)}{dt}}{\frac{dx(t)}{dt}} = \frac{dy(t)}{dx(t)}.$$

The reason for having two types of continuities has to do with parameter substitution (see box). Given a curve segment $\mathbf{P}(t)$ where $0 \le t \le 1$, we can substitute $T = t^2$.

The new segment $\mathbf{Q}(T) = \mathbf{Q}(t^2)$, where $0 \le T \le 1$, is identical in shape to $\mathbf{P}(t)$. The two identical curves must, of course, have the same tangents. However, their calculated tangent vectors have different magnitudes because

$$\frac{d\mathbf{Q}(t^2)}{dt} = 2t\frac{d\mathbf{Q}(t)}{dt} = 2t\frac{d\mathbf{P}(t)}{dt}.$$

This is why we separate the direction and the magnitude of the tangent vectors when considering curve continuities. If the directions of the tangent vectors are equal, they produce a smooth join and we call this case G^1 continuity (which is often all that is required in practice).

Example: Consider the two straight segments $\mathbf{P}(t) = (8t, 6t)$ and $\mathbf{Q}(t) = (4(t+2), 3(t+2))$. The first goes from $(0,0)$ to $(8,6)$ and the second goes from $(8,6)$ to $(12,9)$. Their tangent vectors are $\mathbf{P}^t(t) = (8,6)$ and $\mathbf{Q}^t(t) = (4,3)$. The segments connect smoothly at $(8,6)$ (in fact, they look like one straight segment), but their tangent vectors are different at that point! Thus, the total curve has G^1 continuity at point $(8,6)$, but not C^1 continuity.

It is interesting to note, however, that the unit tangent vectors **are** equal at the joint. The magnitude of $\mathbf{P}^t(t)$ is $\sqrt{8^2+6^2} = 10$ and that of $\mathbf{Q}^t(t) = \sqrt{4^2+3^2} = 5$. The two unit tangent vectors are therefore equal $(8/10, 6/10) = (4/5, 3/5)$. Thus, the unit tangent vector provides a better measure of the direction of the curve than the tangent vector itself. Another natural vector that's associated with every point of a smooth curve is the curvature, a basic concept that's discussed in Section 1.6.

A curve whose tangent vector and curvature vector (Section 1.6.6) are everywhere continuous is said to have G^2 (second-order geometric) continuity.

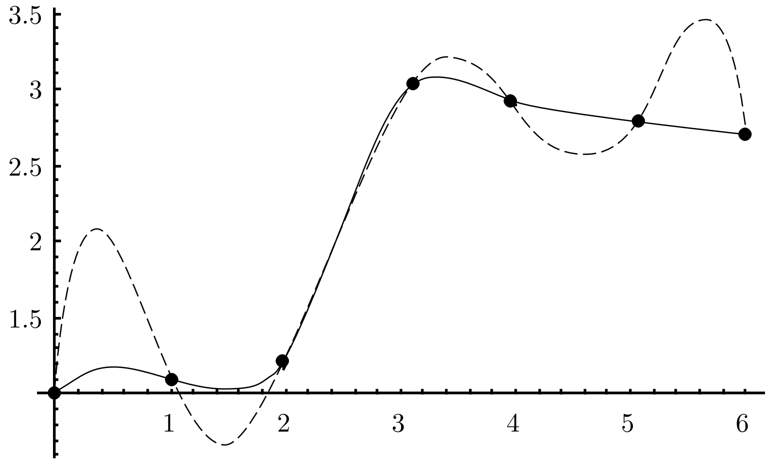
> You can do anything you like with me except paint me, Hughie dear. I have to draw the line somewhere. But that's just what you *can't* do—draw a line, I mean. I like you in every way, as you well know, except as a painter. You would have been a good painter if you had never painted—did I invent that?
>
> —L. P. Hartley, *The Hireling*

1.5 PC Curves

Parametric curves used in computer graphics are based on polynomials. A polynomial of degree one has the form $\mathbf{P}_1(t) = \mathbf{A}t + \mathbf{B}$ and is, therefore, a straight line so it can only be used in limited cases. A parametric polynomial of degree 2 (quadratic) has the form $\mathbf{P}_2(t) = \mathbf{A}t^2 + \mathbf{B}t + \mathbf{C}$ and is always a parabola (see next paragraph and Appendix A). A polynomial of degree 3 (cubic) has the form $\mathbf{P}_3(t) = \mathbf{A}t^3 + \mathbf{B}t^2 + \mathbf{C}t + \mathbf{D}$ and is the simplest curve that can have complex shapes and can also be a space curve. (The complexity of this polynomial is limited, though. It can have at most one loop, and, if it does not have a loop, it can have at most two inflection points, see Section 1.6.8). Polynomials of higher degrees are sometimes needed, but they generally wiggle too much and are difficult to control. They also have more coefficients, so they require more input data to determine all the coefficients. As a result, a complete curve is often constructed from segments, each a parametric cubic polynomial (also called a PC). The complete curve is a piecewise polynomial curve, sometimes also called a *spline* (see definition on page 141).

Plane curves described by degree-2 polynomials are conic sections, but this is true only for the implicit representation. A plane curve described parametrically by a degree-2 polynomial can only be a parabola. Given such a curve $\mathbf{P}(t) = \mathbf{a}\,t^2 + \mathbf{b}\,t + \mathbf{c}$ we observe that it has a single value for any value of t and that it grows without limit when t becomes very large (positive or negative). Thus, when t approaches $\pm\infty$, $\mathbf{P}(t)$ also approaches ∞ or $-\infty$ (depending on the sign of $\mathbf{a}$) but there is only one branch that goes toward ∞ and one branch that goes toward $-\infty$. We therefore conclude that $\mathbf{P}(t)$ cannot be an ellipse because ellipses are finite, and it cannot be a hyperbola because these curves approach $\pm\infty$ in two directions. It must therefore be a parabola. A more rigorous proof, using parameter substitution, can be found in [Gallier 00], page 66.

Figure 1.10 shows seven data points and two curves that fit them. The dashed curve is a polynomial of degree 6; the solid curve is a spline. It is easy to see that the polynomial oscillates, whereas the spline curve is tight and is therefore more pleasing to the eye.



```
Clear[points];
points={{0,1},{1,1.1},{2,1.2},{3,3},{4,2.9},{5,2.8},{6,2.7}};
InterpolatingPolynomial[points,x];
Interpolation[points,InterpolationOrder->3];
Show[ListPlot[points,Prolog->AbsolutePointSize[5]],
 Plot[%%,{x,0,6},PlotStyle->Dashing[{0.05,0.05}]],
 Plot[%[x],{x,0,6}]]
```

Figure 1.10: Polynomial and Spline Fit.

◇ **Exercise 1.19:** Show that a quadratic polynomial must be a plane curve.

◇ **Exercise 1.20:** Why does a high-degree polynomial wiggle?

Question: The word “quad” comes from Latin for “four,” so why is a degree-2 polynomial called quadratic? While we are at it, why is a degree-3 polynomial called cubic?
Answer: A square of side length n has four sides (it is quadratic), but its area is n^2 and this is associated with a degree-2 polynomial, which has terms up to x^2. Similarly, a cube of side length n has volume n^3, which is why the term “cubic” has become associated with a degree-3 polynomial.

A single PC segment is determined by means of points (data or control) or tangent vectors. Continuity considerations are also used sometimes to constrain the curve. Regardless of the input data, the segment always has the form $\mathbf{P}(t) = \mathbf{A}t^3 + \mathbf{B}t^2 + \mathbf{C}t + \mathbf{D}$. Thus, four unknown coefficients have to be calculated, which requires four equations. The equations must depend on four known quantities, points or vectors, that we denote by $\mathbf{G}_1$ through $\mathbf{G}_4$. The PC segment is expressed as the product

$$\mathbf{P}(t) = (t^3, t^2, t, 1)\begin{pmatrix} m_{11} & m_{12} & m_{13} & m_{14}\\ m_{21} & m_{22} & m_{23} & m_{24}\\ m_{31} & m_{32} & m_{33} & m_{34}\\ m_{41} & m_{42} & m_{43} & m_{44}\end{pmatrix}\begin{pmatrix}\mathbf{G}_1\\ \mathbf{G}_2\\ \mathbf{G}_3\\ \mathbf{G}_4\end{pmatrix} = \mathbf{T}(t)\cdot\mathbf{M}\cdot\mathbf{G},$$

where $\mathbf{M}$ is the basis matrix that depends on the method used and $\mathbf{G}$ is the geometry vector, consisting of the four given quantities. The segment can also be written as the weighted sum

$$\begin{aligned}\mathbf{P}(t) &= (t^3m_{11} + t^2m_{21} + tm_{31} + m_{41})\mathbf{G}_1 + (t^3m_{12} + t^2m_{22} + tm_{32} + m_{42})\mathbf{G}_2\\ &\quad + (t^3m_{13} + t^2m_{23} + tm_{33} + m_{43})\mathbf{G}_3 + (t^3m_{14} + t^2m_{24} + tm_{34} + m_{44})\mathbf{G}_4\\ &= B_1(t)\mathbf{G}_1 + B_2(t)\mathbf{G}_2 + B_3(t)\mathbf{G}_3 + B_4(t)\mathbf{G}_4 = \mathbf{B}(t)\cdot\mathbf{G} = \mathbf{T}(t)\cdot\mathbf{N}\cdot\mathbf{G},\end{aligned}$$

where $\mathbf{B}(t)$ equals the product $\mathbf{T}(t)\cdot\mathbf{M}$ and the $B_i(t)$ are the weights. They are also called the *blending functions*, since they blend the four given quantities. If any of the quantities being blended are points, their weights should be barycentric. In the case where all four quantities are points, this requirement implies that the sum of the elements of matrix $\mathbf{M}$ should equal 1 (because the 16 elements of $\mathbf{M}$ are also the elements of the $B_i(t)$’s).

A PC segment can also be written in the form

$$\mathbf{P}(t) = \mathbf{A}t^3 + \mathbf{B}t^2 + \mathbf{C}t + \mathbf{D} = (t^3, t^2, t, 1)\begin{pmatrix} A_x & A_y & A_z\\ B_x & B_y & B_z\\ C_x & C_y & C_z\\ D_x & D_y & D_z\end{pmatrix} = \mathbf{T}(t)\cdot\mathbf{C},$$

where $\mathbf{A} = (A_x, A_y, A_z)$ and similarly for $\mathbf{B}$, $\mathbf{C}$, and $\mathbf{D}$. Its first derivative is

$$\frac{d\mathbf{P}(t)}{dt} = \frac{d\mathbf{T}(t)}{dt}\cdot\mathbf{C} = (3t^2, 2t, 1, 0)\mathbf{C}$$

and this is the tangent vector of the curve. This vector points in the direction of the tangent to the curve, but its magnitude is also important. It describes the *speed* of the curve.

In physics, if the function $x(t)$ describes the position of an object at time t, then $dx(t)/dt$ describes its velocity, and $d^2x(t)/dt^2$ gives its acceleration. This is also true for curves, but the speed in question is not the speed of drawing the curve on the screen! Rather, it is the distance covered on the curve when t is incremented in equal steps (see the particle paradigm of Section 1.3).

This concept is important in computer animation. Imagine a camera moving along the curve while t is incremented in equal steps. The speed of the camera at a point is given by the magnitude of the tangent vector at that point. If we want the camera to move at a constant speed, all tangent vectors must have the same magnitude. For this to happen, the tangent vector must be independent of t, a constant. This implies that the second derivative (the acceleration) is the zero vector, and the curve itself must be a linear function of t, a straight line. Any other curve has a tangent vector that depends on t, implying that the curve itself moves at variable speed.

1.5.1 Fast Computation of a PC

This section employs the method of *forward differences*, together with the Taylor series representation, to speed up the calculation of a point on a parametric curve $\mathbf{P}(t)$. Once this method is implemented, an entire curve can be drawn in a loop where t is incremented from 0 to 1 in small, equal steps of Δ. In iteration $i+1$, a point $\mathbf{P}([i+1]\Delta)$ is computed and is connected to the previous point $\mathbf{P}(i\Delta)$ by a short, straight segment. Section 6.3 applies this method to the Bézier curve.

The principle of forward differences is to find a quantity $\mathbf{dP}$ such that $\mathbf{P}(t+\Delta) = \mathbf{P}(t) + \mathbf{dP}$ for any value of t. If such a $\mathbf{dP}$ can be found, then it is enough to calculate $\mathbf{P}(0)$, and use forward differences to compute

$$\begin{aligned}\mathbf{P}(0+\Delta) &= \mathbf{P}(0)+\mathbf{dP},\\ \mathbf{P}(2\Delta) &= \mathbf{P}(\Delta)+\mathbf{dP} = \mathbf{P}(0)+2\mathbf{dP},\\ &\ \vdots\\ \mathbf{P}([i+1]\Delta) &= \mathbf{P}\bigl(i\Delta\bigr)+\mathbf{dP} = \mathbf{P}(0)+(i+1)\,\mathbf{dP}.\end{aligned}$$

The point is that $\mathbf{dP}$ should not depend on t. If $\mathbf{dP}$ turns out to depend on t, then as we advance t from 0 to 1, we have to use different values of $\mathbf{dP}$, slowing down the calculations. The fastest way to calculate the curve is to precalculate $\mathbf{dP}$ before the loop starts and repeatedly add this precalculated value to $\mathbf{P}(0)$ inside the loop.

We calculate $\mathbf{dP}$ from the *Taylor series* representation of the curve. The Taylor series of a function $f(t)$ at a point $f(t+\Delta)$ is the infinite sum

$$f(t+\Delta) = f(t) + f'(t)\Delta + \frac{f''(t)\Delta^2}{2!} + \frac{f'''(t)\Delta^3}{3!} + \cdots.$$

In order to avoid dealing with an infinite sum, we limit our discussion to the popular PC curves. The mathematical treatment for any other type of curve (a different-degree polynomial or a nonpolynomial) is similar, although normally more complex. A general PC curve has the form $\mathbf{P}(t) = \mathbf{a}t^3 + \mathbf{b}t^2 + \mathbf{c}t + \mathbf{d}$, so only its first three derivatives are nonzero. These derivatives are

$$\mathbf{P}^t(t) = 3\mathbf{a}t^2 + 2\mathbf{b}t + \mathbf{c}, \quad \mathbf{P}^{tt}(t) = 6\mathbf{a}t + 2\mathbf{b}, \quad \mathbf{P}^{ttt}(t) = 6\mathbf{a},$$

so the Taylor series representation produces

$$\begin{aligned}\mathbf{dP} &= \mathbf{P}(t+\Delta) - \mathbf{P}(t)\\ &= \mathbf{P}^t(t)\Delta + \frac{\mathbf{P}^{tt}(t)\Delta^2}{2} + \frac{\mathbf{P}^{ttt}(t)\Delta^3}{6}\\ &= 3\mathbf{a}\,t^2\Delta + 2\mathbf{b}\,t\Delta + \mathbf{c}\Delta + 3\mathbf{a}\,t\Delta^2 + \mathbf{b}\Delta^2 + \mathbf{a}\Delta^3.\end{aligned}$$

This seems a failure since $\mathbf{dP}$ is a function of t (it should therefore be denoted by $\mathbf{dP}(t)$ instead of just $\mathbf{dP}$) and is also slow to calculate. However, the original PC curve $\mathbf{P}(t)$ is a degree-3 polynomial, whereas $\mathbf{dP}(t)$ is only a degree-2 polynomial. This suggests a way out of our difficulty. We can try to express $\mathbf{dP}(t)$ by means of the Taylor series, similar to what we did with the original curve $\mathbf{P}(t)$. This should result in a forward difference $\mathbf{ddP}(t)$ that's a polynomial of degree 1 in t. The quantity $\mathbf{ddP}(t)$ can, in turn, be represented by another Taylor series to produce a forward difference $\mathbf{dddP}$ that's a degree-0 polynomial in t, i.e., a constant. Once this is done, we hope to end up with an algorithm of the form

```
Compute $\mathbf{P}(0)$, dP, ddP, and dddP;
P = $\mathbf{P}(0)$;
for t:=0 to 1 step $\Delta t$ do
PN:=P+dP; dP:=dP+ddP; ddP:=ddP+dddP;
line(P,PN);
P:=PN;
endfor;
```

The quantity $\mathbf{ddP}(t)$ is obtained by

$$\mathbf{dP}(t+\Delta) = \mathbf{dP}(t) + \mathbf{ddP}(t) = \mathbf{dP}(t) + \mathbf{dP}^t(t)\Delta + \frac{\mathbf{dP}(t)^{tt}\Delta^2}{2},$$

yielding

$$\begin{aligned}\mathbf{ddP}(t) &= \mathbf{dP}^t(t)\Delta + \frac{\mathbf{dP}(t)^{tt}\Delta^2}{2}\\ &= (6\mathbf{a}\,t\Delta + 2\mathbf{b}\Delta + 3\mathbf{a}\Delta^2)\Delta + \frac{6\mathbf{a}\Delta\Delta^2}{2}\\ &= 6\mathbf{a}\,t\Delta^2 + 2\mathbf{b}\Delta^2 + 6\mathbf{a}\Delta^3.\end{aligned}$$

Finally, $\mathbf{dddP}$ is similarly obtained by $\mathbf{ddP}(t+\Delta) = \mathbf{ddP}(t) + \mathbf{dddP} = \mathbf{ddP}(t) + \mathbf{ddP}^t(t)\Delta$, yielding $\mathbf{dddP} = \mathbf{ddP}^t(t)\Delta = 6\mathbf{a}\Delta^3$, a constant.

The four quantities involved in the calculation of the curve are therefore

$$\begin{aligned}
\mathbf{P}(t) &= \mathbf{a}t^3 + \mathbf{b}t^2 + \mathbf{c}t + \mathbf{d},\\
\mathbf{dP}(t) &= 3\mathbf{a}\,t^2\Delta + 2\mathbf{b}\,t\Delta + \mathbf{c}\Delta + 3\mathbf{a}\,t\Delta^2 + \mathbf{b}\Delta^2 + \mathbf{a}\Delta^3,\\
\mathbf{ddP}(t) &= 6\mathbf{a}\,t\Delta^2 + 2\mathbf{b}\Delta^2 + 6\mathbf{a}\Delta^3,\\
\mathbf{dddP} &= 6\mathbf{a}\Delta^3.
\end{aligned}$$

They have to be calculated at $t = 0$ before the loop starts, then each iteration computes the first three quantities from those of the previous iteration (**dddP** doesn't depend on t). Here are the details

$$\begin{gathered}
\mathbf{P}(0) = \mathbf{d}, \quad \mathbf{dP}(0) = \mathbf{a}\Delta^3 + \mathbf{b}\Delta^2 + \mathbf{c}\Delta, \quad \mathbf{ddP}(0) = 6\mathbf{a}\Delta^3 + 2\mathbf{b}\Delta^2, \quad \mathbf{dddP} = 6\mathbf{a}\Delta^3.\\
\mathbf{P}(\Delta) = \mathbf{a}\Delta^3 + \mathbf{b}\Delta^2 + \mathbf{c}\Delta + \mathbf{d} = \mathbf{P}(0) + \mathbf{dP}(0),\\
\mathbf{dP}(\Delta) = \mathbf{a}\Delta^3 + 2\mathbf{b}\Delta^2 + \mathbf{c}\Delta + 3\mathbf{a}\Delta^3 + \mathbf{b}\Delta^2 + \mathbf{a}\Delta^3 = \mathbf{dP}(0) + \mathbf{ddP}(0),\\
\mathbf{ddP}(\Delta) = 6\mathbf{a}\Delta^3 + 2\mathbf{b}\Delta^2 + 6\mathbf{a}\Delta^3 = \mathbf{ddP}(0) + \mathbf{dddP},\\
\ldots\\
\mathbf{P}([i+1]\Delta) = \mathbf{P}(i\Delta) + \mathbf{dP}(i\Delta),\\
\mathbf{dP}([i+1]\Delta) = \mathbf{dP}(i\Delta) + \mathbf{ddP}(i\Delta),\\
\mathbf{ddP}([i+1]\Delta) = \mathbf{ddP}(i\Delta) + \mathbf{dddP}.
\end{gathered}$$

Thus, each iteration computes a point $\mathbf{P}([i+1]\Delta)$ on the curve by performing six simple operations, three additions and three assignments. No multiplications are needed.

1.5.2 Subdividing a Parametric Curve

Parametric curves are defined by means of points (data or control) and sometimes also vectors. Editing such a curve is normally done by moving points around and by adding new points. Intuitively, it is clear that adding points allows for finer control of the shape of the curve. On the other hand, adding points results in a curve that's a high-degree polynomial, and such polynomials tend to oscillate. Also, more points implies more calculations to compute and display the curve.

It therefore seems that a reasonable method to obtain the right curve is to start with a few points, and if these are not enough to obtain the desired shape of the curve, to add a point (or a few points) at a time until the desired shape is achieved.

This section discusses a different approach whereby the correct curve is achieved by subdividing a parametric curve into two segments. Together, the two segments have the same shape as the original curve, but they are defined by more entities (points or vectors), thereby making it possible to fine-tune the curve. This approach is applied in Section 6.8 to the Bézier curve. Section 1.12 extends this approach to surface patches.

> The control of large numbers is possible, and like unto that of small numbers, if we subdivide them.
>
> —Sun Tze

We limit our discussion to cubic curves, but the method illustrated here applies to polynomial curves of any degree. Let

$$\mathbf{P}(t)=(t^3,t^2,t,1)\mathbf{M}\begin{pmatrix}\mathbf{P}_0\\ \mathbf{P}_1\\ \mathbf{P}_2\\ \mathbf{P}_3\end{pmatrix} \tag{1.11}$$

be any cubic parametric curve defined by four nonscalar entities (points or vectors) where the parameter t varies from 0 to 1. We construct the two halves $\mathbf{P}_1(t)$ and $\mathbf{P}_2(t)$ of this curve by varying the parameter in the intervals $[0,0.5]$ and $[0.5,1]$ (Section 6.8 shows how the unequal ranges $[0,\alpha]$ and $[\alpha,1]$ can be used instead).

Each of the two new curves should have the same shape as half of the original curve. Each half should therefore be written as an expression similar to Equation (1.11) but based on a new set of entities $\mathbf{Q}_i$ computed from the original set $\mathbf{P}_i$. To construct the first half $\mathbf{P}_1(t)$, we define a new parameter $u=2t$. When t varies in the range $[0,0.5]$, u varies from 0 to 1. The first half of the curve is obtained from Equation (1.11) by substituting $t=u/2$

$$\begin{aligned}\mathbf{P}_1(u)&=(u^3/8,u^2/4,u/2,1)\mathbf{M}\begin{pmatrix}\mathbf{P}_0\\ \mathbf{P}_1\\ \mathbf{P}_2\\ \mathbf{P}_3\end{pmatrix}\\ &=(u^3,u^2,u,1)\begin{pmatrix}\frac{1}{8}&0&0&0\\ 0&\frac{1}{4}&0&0\\ 0&0&\frac{1}{2}&0\\ 0&0&0&1\end{pmatrix}\mathbf{M}\begin{pmatrix}\mathbf{P}_0\\ \mathbf{P}_1\\ \mathbf{P}_2\\ \mathbf{P}_3\end{pmatrix}\\ &=(u^3,u^2,u,1)\mathbf{LM}\begin{pmatrix}\mathbf{P}_0\\ \mathbf{P}_1\\ \mathbf{P}_2\\ \mathbf{P}_3\end{pmatrix}\\ &=(u^3,u^2,u,1)\mathbf{M}\begin{pmatrix}\mathbf{Q}_0\\ \mathbf{Q}_1\\ \mathbf{Q}_2\\ \mathbf{Q}_3\end{pmatrix}.\end{aligned} \tag{1.12}$$

The last line of Equation (1.12) expresses $\mathbf{P}_1(u)$ in terms of new entities $\mathbf{Q}_i$. It shows that these entities can be calculated from the equation

$$\mathbf{M}\begin{pmatrix}\mathbf{Q}_0\\ \mathbf{Q}_1\\ \mathbf{Q}_2\\ \mathbf{Q}_3\end{pmatrix}=\mathbf{LM}\begin{pmatrix}\mathbf{P}_0\\ \mathbf{P}_1\\ \mathbf{P}_2\\ \mathbf{P}_3\end{pmatrix},\ \text{whose solution is}\ \begin{pmatrix}\mathbf{Q}_0\\ \mathbf{Q}_1\\ \mathbf{Q}_2\\ \mathbf{Q}_3\end{pmatrix}=\mathbf{M}^{-1}\mathbf{LM}\begin{pmatrix}\mathbf{P}_0\\ \mathbf{P}_1\\ \mathbf{P}_2\\ \mathbf{P}_3\end{pmatrix}. \tag{1.13}$$

◇ **Exercise 1.21:** Why does $\mathbf{P}_1(t)$ have the same shape as the first half of $\mathbf{P}(t)$?

The second half, $\mathbf{P}_2(t)$ is calculated similarly. We first define a new parameter $u=2t-1$. When t varies in the range $[0.5,1]$, u varies from 0 to 1. The second half of the curve is obtained from Equation (1.11) by substituting $t = (u+1)/2$:

$$\begin{aligned}
\mathbf{P}_2(u) &= \bigl((u+1)^3/8, (u+1)^2/4, (u+1)/2, 1\bigr)\mathbf{M}\begin{pmatrix}\mathbf{P}_0\\ \mathbf{P}_1\\ \mathbf{P}_2\\ \mathbf{P}_3\end{pmatrix}\\
&= (u^3, u^2, u, 1)\begin{pmatrix}\frac{1}{8} & 0 & 0 & 0\\ \frac{3}{8} & \frac{1}{4} & 0 & 0\\ \frac{3}{8} & \frac{2}{4} & \frac{1}{2} & 0\\ \frac{1}{8} & \frac{1}{4} & \frac{1}{2} & 1\end{pmatrix}\mathbf{M}\begin{pmatrix}\mathbf{P}_0\\ \mathbf{P}_1\\ \mathbf{P}_2\\ \mathbf{P}_3\end{pmatrix}\\
&= (u^3, u^2, u, 1)\mathbf{R}\mathbf{M}\begin{pmatrix}\mathbf{P}_0\\ \mathbf{P}_1\\ \mathbf{P}_2\\ \mathbf{P}_3\end{pmatrix}\\
&= (u^3, u^2, u, 1)\mathbf{M}\begin{pmatrix}\mathbf{Q}_4\\ \mathbf{Q}_5\\ \mathbf{Q}_6\\ \mathbf{Q}_7\end{pmatrix}. \qquad (1.14)
\end{aligned}$$

The new entities $\mathbf{Q}_i$ are calculated for this second half by

$$\begin{pmatrix}\mathbf{Q}_4\\ \mathbf{Q}_5\\ \mathbf{Q}_6\\ \mathbf{Q}_7\end{pmatrix} = \mathbf{M}^{-1}\mathbf{R}\mathbf{M}\begin{pmatrix}\mathbf{P}_0\\ \mathbf{P}_1\\ \mathbf{P}_2\\ \mathbf{P}_3\end{pmatrix}. \qquad (1.15)$$

Given matrix $\mathbf{M}$ and four entities $\mathbf{P}_i$, the eight new entities $\mathbf{Q}_i$ can be calculated from Equations (1.13) and (1.15). The generalization of this method to higher-degree curves is straightforward. As an example, we apply this method to the cubic Bézier curve, Equation (6.8). Matrix $\mathbf{M}$ and its inverse are

$$\mathbf{M} = \begin{pmatrix}-1 & 3 & -3 & 1\\ 3 & -6 & 3 & 0\\ -3 & 3 & 0 & 0\\ 1 & 0 & 0 & 0\end{pmatrix}, \qquad \mathbf{M}^{-1} = \begin{pmatrix}0 & 0 & 0 & 1\\ 0 & 0 & \frac{1}{3} & 1\\ 0 & \frac{1}{3} & \frac{2}{3} & 1\\ 1 & 1 & 1 & 1\end{pmatrix}.$$

The matrix products of Equations (1.13) and (1.15) now become

$$\mathbf{M}^{-1}\mathbf{L}\mathbf{M} = \begin{pmatrix}1 & 0 & 0 & 0\\ \frac{1}{2} & \frac{1}{2} & 0 & 0\\ \frac{1}{4} & \frac{2}{4} & \frac{1}{4} & 0\\ \frac{1}{8} & \frac{3}{8} & \frac{3}{8} & \frac{1}{8}\end{pmatrix}, \qquad \mathbf{M}^{-1}\mathbf{R}\mathbf{M} = \begin{pmatrix}\frac{1}{8} & \frac{3}{8} & \frac{3}{8} & \frac{1}{8}\\ 0 & \frac{1}{4} & \frac{2}{4} & \frac{1}{4}\\ 0 & 0 & \frac{1}{2} & \frac{1}{2}\\ 0 & 0 & 0 & 1\end{pmatrix}. \qquad (1.16)$$

The eight new entities (which in this case are control points) are

$$\mathbf{Q}_0 = \mathbf{P}_0,$$

$$\begin{aligned}
\mathbf{Q}_1 &= \frac{1}{2}\mathbf{P}_0 + \frac{1}{2}\mathbf{P}_1 = \frac{1}{2}(\mathbf{P}_0 + \mathbf{P}_1),\\
\mathbf{Q}_2 &= \frac{1}{4}\mathbf{P}_0 + \frac{2}{4}\mathbf{P}_1 + \frac{1}{4}\mathbf{P}_2 = \frac{1}{2}\left(\frac{1}{2}(\mathbf{P}_0 + \mathbf{P}_1) + \frac{1}{2}(\mathbf{P}_1 + \mathbf{P}_2)\right),\\
\mathbf{Q}_3 &= \frac{1}{8}\mathbf{P}_0 + \frac{3}{8}\mathbf{P}_1 + \frac{3}{8}\mathbf{P}_2 + \frac{1}{8}\mathbf{P}_3\\
&= \frac{1}{2}\left(\frac{1}{2}\left(\frac{1}{2}(\mathbf{P}_0 + \mathbf{P}_1) + \frac{1}{2}(\mathbf{P}_1 + \mathbf{P}_2)\right) + \frac{1}{2}\left(\frac{1}{2}(\mathbf{P}_1 + \mathbf{P}_2) + \frac{1}{2}(\mathbf{P}_2 + \mathbf{P}_3)\right)\right),\\
\mathbf{Q}_4 &= \frac{1}{8}\mathbf{P}_0 + \frac{3}{8}\mathbf{P}_1 + \frac{3}{8}\mathbf{P}_2 + \frac{1}{8}\mathbf{P}_3\\
&= \frac{1}{2}\left(\frac{1}{2}\left(\frac{1}{2}(\mathbf{P}_0 + \mathbf{P}_1) + \frac{1}{2}(\mathbf{P}_1 + \mathbf{P}_2)\right) + \frac{1}{2}\left(\frac{1}{2}(\mathbf{P}_1 + \mathbf{P}_2) + \frac{1}{2}(\mathbf{P}_2 + \mathbf{P}_3)\right)\right),\\
\mathbf{Q}_5 &= \frac{1}{4}\mathbf{P}_1 + \frac{2}{4}\mathbf{P}_2 + \frac{1}{4}\mathbf{P}_3 = \frac{1}{2}\left(\frac{1}{2}(\mathbf{P}_1 + \mathbf{P}_2) + \frac{1}{2}(\mathbf{P}_2 + \mathbf{P}_3)\right),\\
\mathbf{Q}_6 &= \frac{1}{2}\mathbf{P}_1 + \frac{1}{2}\mathbf{P}_2 = \frac{1}{2}(\mathbf{P}_1 + \mathbf{P}_2),\\
\mathbf{Q}_7 &= \mathbf{P}_3.
\end{aligned}$$

Section 6.8 shows a different approach, using the mediation operator, to the problem of subdividing a curve. That approach is applied to the Bézier curve.

1.6 Curvature and Torsion

The first derivative $\mathbf{P}^t(t)$ of a parametric curve $\mathbf{P}(t)$ is the tangent vector of the curve. In this section, we denote the unit tangent vector at point $\mathbf{P}(i)$ by $\mathbf{T}(i)$. Thus,

$$\mathbf{T}(i) = \frac{\mathbf{P}^t(i)}{|\mathbf{P}^t(i)|}.$$

The tangent vector is an example of an *intrinsic property* of a curve. An intrinsic property of a geometric figure depends only on the figure and not on the particular choice of the coordinate axes. Any geometric figure may have intrinsic and extrinsic properties. A triangle has three angles and a quadrilateral has four edges, regardless of the choice of coordinates. The tangent vector of a curve, as well as its curvature, does not depend on the particular coordinate system used. In contrast, the slope of a curve depends on the particular coordinates chosen, which makes it an extrinsic property of the curve.

◇ **Exercise 1.22:** Give a few more intrinsic and extrinsic properties of geometric figures.

This section discusses the important intrinsic properties of parametric curves. They include the principal vectors (the tangent, normal, and binormal vectors), the principal planes (the osculating, rectifying, and normal planes), and the concepts of curvature and torsion. These properties are all local and they vary from point to point on the curve.

They are therefore functions of the parameter t. Notice that these properties exist for all curves, but the discussion here is limited to parametric curves.

> Newton was seeking better methods—more general—for finding the slope of a curve at any particular point, as well [as] another quantity, related but once removed, the degree of curvature, rate of bending, "the crookedness in lines." He applied himself to the tangent, the straight line that grazes the curve at any point. The straight line that the curve would become at that point, if it could be seen through an infinitely powerful microscope.
>
> —James Gleick, *Isaac Newton* (2003)

1.6.1 Normal Plane

The normal plane to a curve $\mathbf{P}(t)$ at point $\mathbf{P}(i)$ is the plane that's perpendicular to the tangent $\mathbf{P}^t(i)$ and contains point $\mathbf{P}(i)$. If $\mathbf{Q}$ is an arbitrary point on the normal plane, then Figure 1.11 shows that $(\mathbf{Q} - \mathbf{P}(i)) \bullet \mathbf{P}^t(i) = 0$. This can be written $\mathbf{Q} \bullet \mathbf{P}^t(i) - \mathbf{P}(i) \bullet \mathbf{P}^t(i) = 0$ or

$$x \cdot x_i^t + y \cdot y_i^t + z \cdot z_i^t - (x_i \cdot x_i^t + y_i \cdot y_i^t + z_i \cdot z_i^t) = 0, \tag{1.17}$$

an expression that has the familiar form $Ax + By + Cz + D = 0$ (Section 2.2.2).

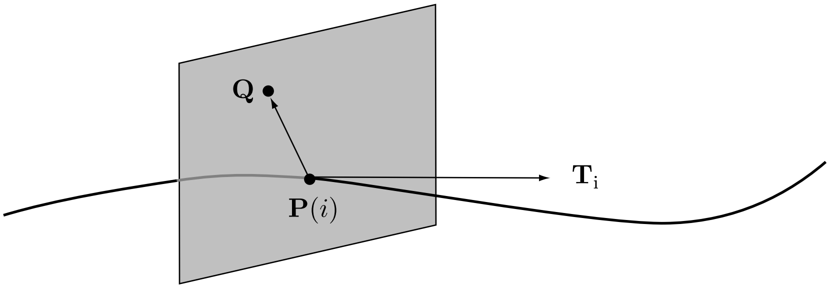


Figure 1.11: The Normal Plane.

1.6.2 Principal Normal Vector

Another important vector associated with a curve is the *principal normal vector* $\mathbf{N}(t)$. This unit vector is normal to the curve (and is therefore contained in the normal plane and is also perpendicular to the tangent vector), but it is called the principal normal since it points in a special direction, the direction in which the curve is turning. The principal normal vector points toward a point called the *center of curvature* of the curve. To express $\mathbf{N}(t)$ in terms of the curve and its derivatives, we select two nearby points, t and $t + \Delta t$, on the curve. The tangent vectors at the two points are $\mathbf{a} = \mathbf{P}^t(t)$ and $\mathbf{b} = \mathbf{P}^t(t + \Delta t)$, respectively. If we subtract them as in Figure 1.12a, we get $\mathbf{c} = \mathbf{b} - \mathbf{a}$. The difference vector $\mathbf{c}$ can be interpreted in two ways. On one hand, we can say that it is a small change in the tangent vector $\mathbf{P}^t(t)$, so we can denote it $\Delta\mathbf{P}^t(t)$. On the other hand, since the tangent vector can be interpreted as the velocity of the curve, any changes in it can be interpreted as acceleration, that is, the second derivative $\mathbf{P}^{tt}(t)$.

Thus, we can write $\mathbf{c} = \Delta\mathbf{P}^t(t) = \mathbf{P}^{tt}(t)$. The two vectors $\mathbf{a} = \mathbf{P}^t(t)$ and $\mathbf{b} = \mathbf{P}^t(t+\Delta t)$ define a plane and the principal normal vector lies at the intersection of this plane and the normal plane. Our task is therefore to compute a vector that is perpendicular to the tangent $\mathbf{a} = \mathbf{P}^t(t)$ and that is contained in the plane defined by $\mathbf{a}$ and $\mathbf{b}$.

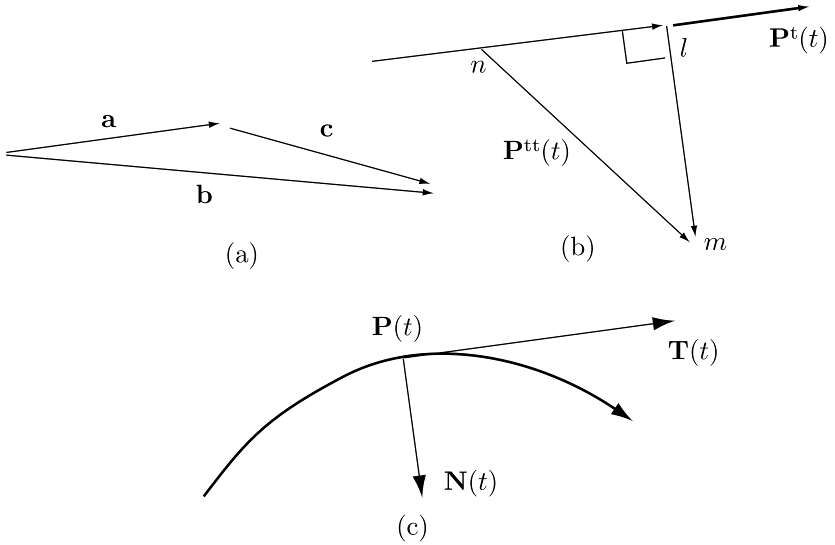


Figure 1.12: The Principal Normal Vector.

Figure 1.12b shows vector $\vec{nl}$, which is the projection of $\mathbf{P}^{tt}(t)$ (vector $\vec{nm}$) onto $\mathbf{P}^t(t)$. Equation (1.8) tells us that the length of $\vec{nl}$ is

$$\frac{\mathbf{P}^{tt}(t) \bullet \mathbf{P}^t(t)}{|\mathbf{P}^t(t)|}.$$

Since $\vec{nl}$ is in the direction of $\mathbf{P}^t(t)$, we can write the vector $\vec{nl}$ as

$$\vec{nl} = \frac{\mathbf{P}^{tt}(t) \bullet \mathbf{P}^t(t)}{|\mathbf{P}^t(t)|} \cdot \frac{\mathbf{P}^t(t)}{|\mathbf{P}^t(t)|} = \frac{\mathbf{P}^{tt}(t) \bullet \mathbf{P}^t(t)}{|\mathbf{P}^t(t)|^2}\mathbf{P}^t(t).$$

We denote the vector $\vec{lm}$ by $\mathbf{K}(t)$ and compute it from the relation $\vec{nl} + \vec{lm} = \vec{nm} = \mathbf{P}^{tt}(t)$:

$$\mathbf{K}(t) = \mathbf{P}^{tt}(t) - \vec{nl} = \mathbf{P}^{tt}(t) - \frac{\mathbf{P}^{tt}(t) \bullet \mathbf{P}^t(t)}{|\mathbf{P}^t(t)|^2}\mathbf{P}^t(t). \tag{1.18}$$

The principal normal vector $\mathbf{N}(t)$ is a unit vector in the direction of $\mathbf{K}(t)$, so it is given by

$$\mathbf{N}(t) = \frac{\mathbf{K}(t)}{|\mathbf{K}(t)|}.$$

◇ **Exercise 1.23:** What can we say about the nature of the principal normal vector of a straight line?

◇ **Exercise 1.24:** Calculate the principal normal vector of the PC curve $\mathbf{P}(t) = (-1,0)t^3 + (1,-1)t^2 + (1,1)t$. Notice that this curve is Equation (4.10), so we know that it goes from $(0,0)$ to $(1,0)$ with start and end tangents $(1,1)$, $(0,-1)$, respectively. Use this to check your results.

1.6.3 Binormal Vector

The third important vector associated with a curve is the *binormal vector* $\mathbf{B}(t)$. It is defined as the vector perpendicular to both the tangent and principal normal, so its definition is simply $\mathbf{B}(t) = \mathbf{T}(t) \times \mathbf{N}(t)$. Notice that it is a unit vector. Since the binormal is perpendicular to the tangent, it is contained in the normal plane. The three vectors $\mathbf{T}(t)$, $\mathbf{N}(t)$, and $\mathbf{B}(t)$ therefore constitute an orthogonal coordinate system that moves along the curve as t varies, except at cusps, where they are undefined.

1.6.4 The Osculating Plane

Imagine three points h, i, and j, located close to each other on a curve. If they are not collinear, they define a plane. Now, move h and j independently closer and closer to i. As these points move, the plane may change. The plane obtained at the limit is called the *osculating plane* at point i (Figure 1.13). It contains the tangent vector $\mathbf{T}(i)$ and the principal normal $\mathbf{N}(i)$. If $\mathbf{Q}$ is an arbitrary point on the osculating plane, then the plane equation is given by the determinant $|(\mathbf{Q} - \mathbf{P}(i))\ \mathbf{P}^t(i)\ \mathbf{P}^{tt}(i)| = 0$, which can be written explicitly as

$$(x - x_i)(y_i^t z_i^{tt} - y_i^{tt} z_i^t) - (y - y_i)(x_i^t z_i^{tt} - x_i^{tt} z_i^t) + (z - z_i)(x_i^t y_i^{tt} - x_i^{tt} y_i^t) = 0.$$

Another way to obtain the plane equation is to use the fact that point $\mathbf{P}(i)$ and vectors $\mathbf{T}(i)$ and $\mathbf{N}(i)$ are contained in the osculating plane. Any general point $\mathbf{Q}$ in the osculating plane can, therefore, be expressed as $\mathbf{Q} = \mathbf{P}(i) + \alpha\mathbf{T}(i) + \beta\mathbf{N}(i)$, where α and β are real parameters. The osculating plane of a plane curve is, of course, the plane of the curve. The osculating plane of a straight line is undefined.

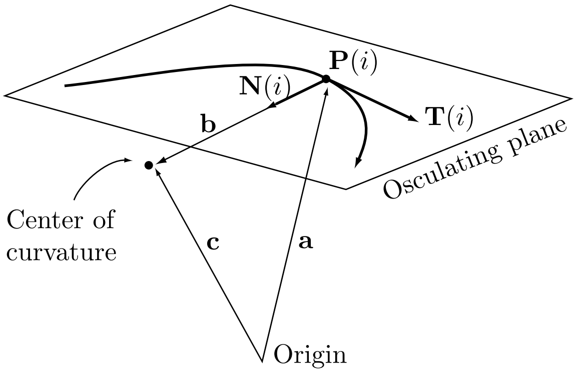


Figure 1.13: The Osculating Plane.

Incidentally, two curves joined at a point have C^2 continuity (Section 1.4.2) at the point if they have the same osculating planes and the same curvature vectors at the point.

◇ **Exercise 1.25:** (1) Calculate the Bézier curve for the four points $\mathbf{P}_0 = (0,0,0)$, $\mathbf{P}_1 = (1,0,0)$, $\mathbf{P}_2 = (2,1,0)$, and $\mathbf{P}_3 = (3,0,1)$. [Those unfamiliar with this curve should use Equation (6.8).] Notice that this is a space curve since the first three points are in the $z = 0$ plane, while the fourth one is outside that plane. (2) Calculate the (unnormalized) principal normal vector of the curve and find its values for $t = 0, 0.5$, and 1. (3) Calculate the osculating plane of the curve and find its equations for $t = 0, 0.5$, and 1 as above.

1.6.5 Rectifying Plane

The plane perpendicular to the principal normal vector of a curve is called the rectifying plane of the curve. If the curve is $\mathbf{P}(t)$, $\mathbf{N}(t)$ is its principal normal, and $\mathbf{Q}$ is an arbitrary point on the rectifying plane, then the equation of the rectifying plane at point $\mathbf{P}(i)$ is $[\mathbf{Q} - \mathbf{P}(i)] \bullet \mathbf{N}(i) = 0$. Another equation is obtained when we realize that both the tangent and binormal vectors are contained in the rectifying plane. A general point on this plane can therefore be expressed as $\mathbf{Q} = \mathbf{P}(i) + \alpha\mathbf{T}(i) + \beta\mathbf{B}(i)$.

Figure 1.14 shows the three unit vectors and three planes associated with a particular point $\mathbf{P}(i)$ on a curve. They constitute intrinsic properties of the curve and together they form the *moving trihedron* of the curve, which can be considered a local coordinate system for the curve. The three vectors constitute the local coordinate axes and the three planes divide the space around point $\mathbf{P}(i)$ into eight octants. The curve passes through the normal plane and is tangent to both the osculating and rectifying planes.

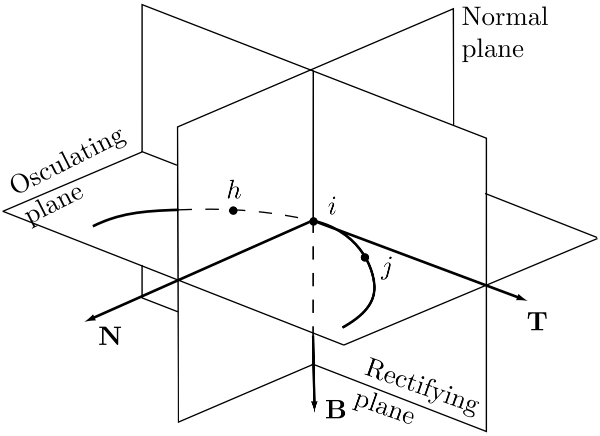


Figure 1.14: The Moving Trihedron.

1.6.6 Curvature

The curvature of a curve is a useful entity, so it deserves to be rigorously defined. Intuitively, the curvature should be a number that measures how much the curve deviates from a straight line at any point. It should be large in areas where the curve wiggles, oscillates, or makes a sudden direction change; it should be small in areas where the curve is close to a straight line. It is also useful to associate a direction with the curvature, i.e., to make it a vector.

Given a parametric curve $\mathbf{P}(t)$ and a point $\mathbf{P}(i)$ on it, we calculate the first two derivatives $\mathbf{P}^t(i)$ and $\mathbf{P}^{tt}(i)$ of the curve at the point. We then construct a circle that has these same first and second derivatives and move it so it grazes the point. This is called the *osculating circle* of the curve at the point. The curvature is now defined as the vector $\kappa(i)$ whose direction is from point $\mathbf{P}(i)$ to the center of this circle and whose magnitude is the reciprocal of the radius of the circle.

Using differential geometry, it can be shown that the vector

$$\frac{\mathbf{P}^t(t)\times\mathbf{P}^{tt}(t)}{|\mathbf{P}^t(t)|^3}$$

has the right magnitude. However, this vector is perpendicular to both $\mathbf{P}^t(t)$ and $\mathbf{P}^{tt}(t)$, so it is perpendicular to the osculating plane. To bring it into the plane, we need to cross-product it with $\mathbf{P}^t(t)/|\mathbf{P}^t(t)|$, so the result is

$$\kappa(t)=\frac{\mathbf{P}^t(t)\times\mathbf{P}^{tt}(t)\times\mathbf{P}^t(t)}{|\mathbf{P}^t(t)|^4}. \tag{1.19}$$

Figure 1.13 shows that the curvature (vector $\mathbf{b}$) is in the direction of the binormal $\mathbf{N}(t)$, so it can be expressed as $\kappa(t)=\rho(t)\mathbf{N}(t)$ where $\rho(t)$ is the *radius of curvature* at point $\mathbf{P}(t)$.

Given a curve $\mathbf{P}(t)$ with an arc length $s(t)$, we assume that $d\mathbf{P}/ds$ is a *unit* tangent vector:

$$\frac{d\mathbf{P}(t)}{ds}=\frac{d\mathbf{P}(t)}{dt}\frac{ds(t)}{dt}=\frac{\mathbf{P}^t(t)}{s^t(t)}. \tag{1.20}$$

Equation (1.20) shows the following:

1. $d\mathbf{P}(t)/ds$ and $\mathbf{P}^t(t)$ point in the same direction. Therefore, since $d\mathbf{P}(t)/ds$ is a unit vector, we get

$$\frac{d\mathbf{P}(t)}{ds}=\frac{\mathbf{P}^t(t)}{|\mathbf{P}^t(t)|}.$$

2. $s^t(t)=|\mathbf{P}^t(t)|$.

We now derive the expression for curvature from a different point of view. The curvature k is defined by $d^2\mathbf{P}(t)/ds^2=k\mathbf{N}$, where $\mathbf{N}$ is the unit principal normal vector (Section 1.6.2). The problem is to express k in terms of the curve $\mathbf{P}(t)$ and its derivatives, not involving the (normally unknown) function $s(t)$. We start with

$$\begin{aligned}\frac{d^2\mathbf{P}(t)}{ds^2}&=\frac{d}{ds}\left(\frac{\mathbf{P}^t(t)}{|\mathbf{P}^t(t)|}\right)=\frac{\frac{d}{dt}\left(\frac{\mathbf{P}^t(t)}{|\mathbf{P}^t(t)|}\right)}{s^t(t)}\\&=\frac{\frac{\mathbf{P}^{tt}(t)}{|\mathbf{P}^t(t)|}-\frac{\mathbf{P}^t(t)}{|\mathbf{P}^t(t)|^2}\cdot\frac{d|\mathbf{P}^t(t)|}{dt}}{|\mathbf{P}^t(t)|}.\end{aligned} \tag{1.21}$$

The identity $\mathbf{A} \bullet \mathbf{A} = |\mathbf{A}|^2$ is true for any vector $\mathbf{A}(t)$ and it implies

$$\mathbf{A}(t) \bullet \mathbf{A}^t(t) = |\mathbf{A}(t)| \frac{d|\mathbf{A}(t)|}{dt}.$$

When we apply this to the vector $\mathbf{P}^t(t)$, we get

$$\frac{d^2\mathbf{P}(t)}{ds^2} = \frac{\mathbf{P}^{tt}(t)}{\mathbf{P}^t(t) \bullet \mathbf{P}^t(t)} - \frac{\mathbf{P}^t(t) \bullet \mathbf{P}^{tt}(t)}{\big(\mathbf{P}^t(t) \bullet \mathbf{P}^t(t)\big)^2} \mathbf{P}^t(t), \tag{1.22}$$

which can also be written

$$k\mathbf{N} = \frac{d^2\mathbf{P}(t)}{ds^2} = \frac{\mathbf{P}^t(t) \times \big(\mathbf{P}^{tt}(t) \times \mathbf{P}^t(t)\big)}{\big(\mathbf{P}^t(t) \bullet \mathbf{P}^t(t)\big)^2}. \tag{1.23}$$

1.6.7 Torsion

Torsion is a measure of how much a given curve deviates from a plane curve. The torsion $\tau(i)$ of a curve at a point $\mathbf{P}(i)$ is defined by means of the following two quantities:

1. Imagine a point h close to i. The curve has rectifying planes at points h and i (Figure 1.15). Denote the angle between them by θ.

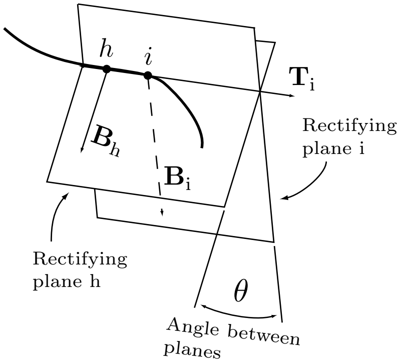


Figure 1.15: Torsion.

2. Denote by s the arc length from point h to point i.

The torsion of the curve at point i is defined as the limit of the ratio θ/s when h approaches i. Figure 1.15 shows how the rectifying plane rotates about the tangent as we move on the curve from h to i. The torsion can be expressed by means of the derivatives of the curve and by means of the curvature

$$\tau(t) = \frac{|\mathbf{P}^t(t)\,\mathbf{P}^{tt}(t)\,\mathbf{P}^{ttt}(t)|}{|\mathbf{P}^t(t) \times \mathbf{P}^t(t)|^2} = \frac{|\mathbf{P}^t(t)\,\mathbf{P}^{tt}(t)\,\mathbf{P}^{ttt}(t)|}{|\mathbf{P}^t(t)|^6} \rho(t)^2.$$

(The numerator is a determinant and the denominator is an absolute value. This expression is meaningful only when $\rho(t) < \infty$.) The torsion of a plane curve is zero.

It is interesting to note that a curve can be fully defined by specifying its curvature and torsion as functions of its arc length s. The functions $\kappa = f(s)$ and $\tau = g(s)$ uniquely define the shape of a curve (although not its location in space). An alternative is the single (implicit) function $F(\kappa, \tau, s) = 0$.

An alternative representation can be derived for a plane curve. Assume that $\mathbf{P}(t) = (x(t), y(t))$ is a curve in the xy plane. Figure 1.16 shows that its shape can be determined if its start point $\mathbf{P}(0)$ and its slope (or, equivalently, angle θ) are known as functions of the arc length s. Since θ is the angle between the tangent and the x axis, functions $x(s)$ and $y(s)$ must satisfy

$$\frac{dx}{ds} = \cos\theta, \qquad \frac{dy}{ds} = \sin\theta.$$

Differentiating produces

$$\frac{d^2x}{ds^2} = -\sin\theta\frac{d\theta}{ds} = -\frac{dy}{ds}\frac{d\theta}{ds}, \quad \frac{d^2y}{ds^2} = \cos\theta\frac{d\theta}{ds} = \frac{dx}{ds}\frac{d\theta}{ds}. \tag{1.24}$$

Figure 1.16 also shows that $d\theta/ds$ is the magnitude of the curvature κ, so the conclusion is that, given the curvature $\kappa(s)$ of a curve as a function of its arc length, the two functions $x(s)$ and $y(s)$ can be calculated, either analytically, or point by point numerically, from the differential equations (1.24).

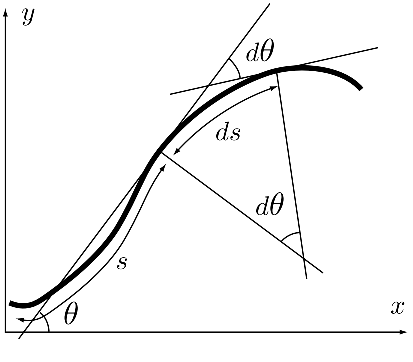


Figure 1.16: A Plane Curve.

◇ **Exercise 1.26:** Given $\kappa(s) = R$ (a constant), solve Equation (1.24) for $x(s)$ and $y(s)$. What kind of a curve is this?

1.6.8 Inflection Points

An inflection point is a point on a curve where the curvature is zero. On a straight line, every point is an inflection point. On a typical curve, an inflection point is created when the curve reverses its direction of turning (for example, from a clockwise direction to a counterclockwise direction). From the definition of curvature [Equation (1.19)] it follows that an inflection point satisfies

$$0 = |\mathbf{P}^t(t) \times \mathbf{P}^{tt}(t)| = \sqrt{(\mathbf{P}^t(t) \times \mathbf{P}^{tt}(t)) \bullet (\mathbf{P}^t(t) \times \mathbf{P}^{tt}(t))}.$$

Therefore,

$$(\mathbf{P}^t(t) \times \mathbf{P}^{tt}(t)) \bullet (\mathbf{P}^t(t) \times \mathbf{P}^{tt}(t)) = 0,$$

which is equivalent to

$$\begin{aligned} &(\mathbf{P}^t(t) \times \mathbf{P}^{tt}(t))_x^2 + (\mathbf{P}^t(t) \times \mathbf{P}^{tt}(t))_y^2 + (\mathbf{P}^t(t) \times \mathbf{P}^{tt}(t))_z^2 = 0, \\ \text{or} \qquad &(y^t z^{tt} - z^t y^{tt})^2 + (z^t x^{tt} - x^t z^{tt})^2 + (x^t y^{tt} - y^t x^{tt})^2 = 0. \qquad (1.25) \end{aligned}$$

This is the sum of three nonnegative quantities, so each must be zero. Since

$$\frac{dy}{dx} = \frac{dy}{dt} \Big/ \frac{dx}{dt} = \frac{y^t}{x^t},$$

we get

$$\frac{d^2y}{dx^2} = \frac{d}{dt}\left(\frac{y^t}{x^t}\right)\frac{dt}{dx} = \frac{x^t y^{tt} - x^{tt} y^t}{(x^t)^3}.$$

Therefore, saying that the three quantities above are zero is the same as saying that

$$\frac{d^2y}{dx^2} = \frac{d^2x}{dz^2} = \frac{d^2z}{dy^2} = 0.$$

Equation (1.25) can be used to show that a two-dimensional parametric cubic can have at most two inflection points. We denote a general PC by

$$\mathbf{P}(t) = \mathbf{a}t^3 + \mathbf{b}t^2 + \mathbf{c}t + \mathbf{d} = (a_x, a_y)t^3 + (b_x, b_y)t^2 + (c_x, c_y)t + (d_x, d_y),$$

which implies $x^t = 3a_x t^2 + 2b_x t + c_x$ and $x^{tt} = 6a_x t + b_x$, and similarly for y^t and y^{tt}. Using this notation, we write Equation (1.25) explicitly (notice that for a two-dimensional PC, only the third part is nonzero) as

$$\begin{aligned} 0 &= x^t y^{tt} - y^t x^{tt} \\ &= (3a_x t^2 + 2b_x t + c_x)(6a_y t + b_y) - (3a_y t^2 + 2b_y t + c_y)(6a_x t + b_x) \\ &= 6(a_y b_x - a_x b_y)t^2 + 6(a_y c_x - a_x c_y)t + 2(b_y c_x - b_x c_y). \end{aligned}$$

This is a quadratic equation in t, so there can be at most two solutions.

1.7 Special and Degenerate Curves

Parametric curves may exhibit unusual behavior when their derivatives satisfy certain conditions. Such curves are referred to as special or degenerate. Here are four examples:

1. If the first derivative $\mathbf{P}^t(t)$ of a curve $\mathbf{P}(t)$ is zero for all values of t, then $\mathbf{P}(t)$ degenerates to the point $\mathbf{P}(0)$.

2. If $\mathbf{P}^t(t) \neq 0$ and $\mathbf{P}^t(t) \times \mathbf{P}^{tt}(t) = 0$ (i.e., the tangent vector points in the direction of the acceleration vector), then $\mathbf{P}(t)$ is a straight line.

3. If $\mathbf{P}^t(t) \times \mathbf{P}^{tt}(t) \neq 0$ and $|\mathbf{P}^t(t)\,\mathbf{P}^{tt}(t)\,\mathbf{P}^{ttt}(t)| = 0$, then $\mathbf{P}(t)$ is a plane curve. (The notation $|\mathbf{a}\,\mathbf{b}\,\mathbf{c}|$ refers to the determinant whose three columns are $\mathbf{a}$, $\mathbf{b}$, and $\mathbf{c}$.)

4. Finally, if both $\mathbf{P}^t(t) \times \mathbf{P}^{tt}(t)$ and $|\mathbf{P}^t(t)\,\mathbf{P}^{tt}(t)\,\mathbf{P}^{ttt}(t)|$ are nonzero, the curve $\mathbf{P}(t)$ is nonplanar (i.e., it is a space curve).

1.8 Basic Concepts of Surfaces

Section 1.3 mentions the explicit, implicit, and parametric representations of curves. Surfaces can also be represented in these three ways. The explicit representation of a surface is $z = f(x, y)$ and the implicit representation is $F(x, y, z) = 0$ (Figure C.3). In practice, however, the parametric representation is used almost exclusively, for the same reasons that parametric curves are so important.

A simple, intuitive way to grasp the concept of a parametric surface is to visualize it as a set of curves. Figure 1.17a shows a single curve and Figure 1.17b shows how it is duplicated several times to create a family of identical curves. The brain finds it natural to interpret such a family as a surface. If we denote the curve by $\mathbf{P}(u)$, we can denote each of its copies in the family by $\mathbf{P}_i(u)$, where i is an integer index.

Taking this idea a step further, a solid surface is obtained by creating infinitely many copies of the curve and placing them next to each other without any gaps in between. It makes sense to replace the integer index i of each curve by a real (continuous) index w. The solid version of the surface of Figure 1.17b can therefore be denoted by $\mathbf{P}_w(u)$, where varying u moves us along a curve and varying w moves us from curve to curve in steps that can be arbitrarily small.

The next step is to obtain a general surface by varying the shape of the curves so they are not identical (Figure 1.17c). The shape of a curve should therefore depend on w, which suggests a notation such as $\mathbf{P}(u, w)$ for the surface. The shape of each curve depends on both u and w but in a special way. Each of the two parameters moves us along a different direction on the surface, so we can talk about the u direction and the w direction (Figure 1.17d).

The general form of a parametric surface is $\mathbf{P}(u, w) = (f_1(u, w), f_2(u, w), f_3(u, w))$. The surface depends on two parameters, u and w, that vary independently in some interval $[a, b]$ (normally, but not always, limited to $[0, 1]$). For each pair (u, w), the expression above produces the three coordinates of a point on the surface.

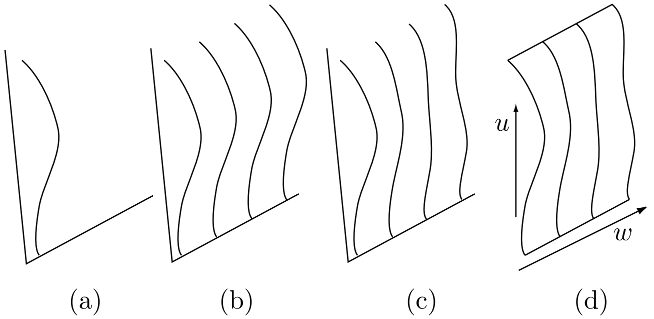


Figure 1.17: A Surface as a Family of Curves.

◇ **Exercise 1.27:** A curve can be either two-dimensional or three-dimensional. A surface, however, exists only in three dimensions, and each surface point has three coordinates. Why is it that the expression for the surface depends on two, and not on three, parameters? We would expect the surface to be of the form $\mathbf{P}(u, v, w)$, a function of three parameters. What's the explanation?

A simple example of a parametric surface is

$$\mathbf{P}(u, w) = [0.5(1-u)w + u, w, (1-u)(1-w)] \tag{1.26}$$

[this is also Equation (2.11)]. Such a surface is called *bilinear* since it is linear in both parameters. We use this example to discuss the concept of a surface patch and to show how a wire-frame surface can be displayed.

1.8.1 A Surface Patch

The expression $\mathbf{P}(u, 0.2)$ (where w is held fixed and u varies) depends on just one parameter and is therefore a curve on the surface. The four curves $\mathbf{P}(u, 0)$, $\mathbf{P}(u, 1)$, $\mathbf{P}(0, w)$, and $\mathbf{P}(1, w)$ are of special interest. They are the *boundary curves* of the surface (Figure 1.18a). Since there are four such curves, our surface is a *patch* that has a (roughly) rectangular shape. Of special interest are the four quantities $\mathbf{P}(0, 0)$, $\mathbf{P}(0, 1)$, $\mathbf{P}(1, 0)$, and $\mathbf{P}(1, 1)$. They are the corner points of the surface patch and are sometimes denoted by $\mathbf{P}_{ij}$.

We say that the curve $\mathbf{P}(u, 0.2)$ lies on the surface in the u direction. It is an *isoparametric curve*. Similarly, any curve $\mathbf{P}(u_0, w)$ where u_0 is fixed, lies in the w direction and is an isoparametric curve. These are the two main directions on a rectangular surface patch.

Two more special curves, the *surface diagonals*, are $\mathbf{P}(u, 1-u)$ and $\mathbf{P}(u, u)$. The former goes from $\mathbf{P}_{01}$ to $\mathbf{P}_{10}$ and the latter goes from $\mathbf{P}_{00}$ to $\mathbf{P}_{11}$.

A large surface is obtained by constructing a number of patches and connecting them. The method used to construct the patch should allow for smooth connection of patches.

◇ **Exercise 1.28:** Compute the corner points, boundary curves, and diagonals of the bilinear surface patch of Equation (1.26).

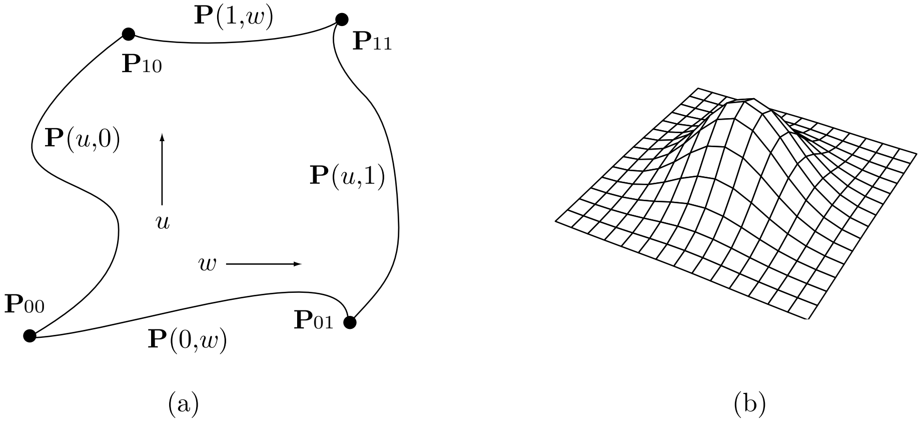


Figure 1.18: (a) A Surface Patch. (b) A Wire Frame.

◇ **Exercise 1.29:** Calculate the corner points and boundary curves of the surface patch

$$\mathbf{P}(u,w) = \big((c-a)u + a, (d-b)w + b, 0\big),$$

where a, b, c, and d are given constants and the parameters u and w vary independently in the range $[0,1]$. What kind of a surface is this?

1.8.2 Displaying a Surface Patch

A surface patch can be displayed either as a wire frame (Figure 1.18b) or as a solid surface. The pseudo-code of Figure 1.19 shows how to display a surface patch as a wire frame. The code consists of two similar loops—one drawing the curves in the w direction and the other drawing the curves in the u direction. The first loop varies u from 0 to 1 in steps of 0.2, thereby drawing six curves. Each of the six is drawn by varying w in small steps (0.01 in the example). The second loop is similar and draws six curves in the u direction.

Procedure `SurfacePoint` receives the current values of u and w, and calculates the coordinates (x,y,z) of one surface point. Procedure `PersProj` uses these coordinates to calculate the screen coordinates (xs, ys) of a pixel (it projects the three-dimensional pixel on the two-dimensional screen using perspective projection). Finally, procedure `Pixel` actually displays the pixel in the desired color. Better results are obtained by eliminating those parts of the surface that are hidden by other parts, but this topic is outside the scope of this book.

To display a solid surface, the *normal vector* of the surface (Section 1.13) has to be calculated at every point and a shading algorithm applied to compute the amount of light reflected from the point. Most texts on computer graphics discuss shading models and algorithms.

```
for u:=0 to 1 step 0.2 do                for w:=0 to 1 step 0.2 do
 begin                                     begin
 for w:=0 to 1 step 0.01 do                for u:=0 to 1 step 0.01 do
  begin                                     begin
  SurfacePoint(u,w,x,y,z);                  SurfacePoint(u,w,x,y,z);
  PersProj(x,y,z,xs,ys);                    PersProj(x,y,z,xs,ys);
  Pixel(xs,ys,color)                        Pixel(xs,ys,color)
  end;                                      end;
 end;                                      end;
```

Figure 1.19: Procedure for a Wire-Frame Surface.

1.9 The Cartesian Product

The concept of blending was introduced in Section 1.2. This is an important concept that is used in many curve and surface algorithms. This section shows how blending can be used in surface design. We start with two parametric curves $\mathbf{Q}(u) = \sum_{i=1}^{n} f_i(u)\mathbf{Q}_i$ and $\mathbf{R}(w) = \sum_{i=1}^{m} g_i(w)\mathbf{R}_i$ where $\mathbf{Q}_i$ and $\mathbf{R}_i$ can be points or vectors. Now examine the function

$$\mathbf{P}(u,w) = \sum_{i=1}^{n}\sum_{j=1}^{m} f_i(u)g_j(w)\mathbf{P}_{ij} = \sum_{i=1}^{n}\sum_{j=1}^{m} h_{ij}(u,w)\mathbf{P}_{ij}, \tag{1.27}$$

where $h_{ij}(u,w) = f_i(u)g_j(w)$. The function $\mathbf{P}(u,w)$ describes a surface, since it is a function of the two independent parameters u and w. For any value of the pair (u,w), the function computes a weighted sum of the quantities $\mathbf{P}_{ij}$. These quantities—which are normally points, but can also be vectors—are triplets, so $\mathbf{P}(u,w)$ returns a triplet (x,y,z) that are the three-dimensional coordinates of a point on the surface. When u and w vary over their ranges independently, $\mathbf{P}(u,w)$ computes all the three-dimensional points of a surface patch.

> I don't blend in at a family picnic.
> —Batman in *Batman Forever*, 1995.

The technique of blending quantities $\mathbf{P}_{ij}$ into a surface by means of weights taken from two curves is called the *Cartesian product*, although the terms *tensor product* and *cross-product* are also sometimes used. The quantities $\mathbf{P}_{ij}$ can be points, tangent vectors, or second derivatives. Equation (1.27) can also be written in the compact form

$$\mathbf{P}(u,w) = \bigl(f_1(u),\ldots,f_n(u)\bigr)\begin{pmatrix}\mathbf{P}_{11} & \mathbf{P}_{12} & \ldots & \mathbf{P}_{1m}\\ \vdots & \vdots & & \vdots\\ \mathbf{P}_{n1} & \mathbf{P}_{n2} & \ldots & \mathbf{P}_{nm}\end{pmatrix}\begin{pmatrix}g_1(w)\\ \vdots\\ g_m(w)\end{pmatrix}. \tag{1.28}$$

Notice that it uses a matrix whose elements are nonscalar quantities (triplets). Even more important, Equation (1.27), combined with the isotropic principle (Section 1.1), tells us that if all $\mathbf{P}_{ij}$ are points, then the surface $\mathbf{P}(u,w)$ is independent of the particular coordinate axes used if $\sum_{ij} h_{ij}(u,w) = 1$. If the two original curves $\mathbf{Q}(u)$ and $\mathbf{R}(w)$ are isotropic, then it's easy to see that the surface is also isotropic because

$$\sum_{ij} h_{ij}(u,w) = \sum_i \sum_j f_i g_j = \Bigl(\sum_j g_j\Bigr)\Bigl(\sum_i f_i\Bigr) = 1.$$

The following two examples illustrate the importance of the Cartesian product. The first example applies this technique to derive the equation of the bilinear surface (Section 2.3) from that of a straight segment. The parametric representation of the line segment from $\mathbf{P}_0$ to $\mathbf{P}_1$ is Equation (2.1)

$$\begin{aligned}\mathbf{P}(t) &= (1-t)\mathbf{P}_0 + t\mathbf{P}_1 = \mathbf{P}_0 + (\mathbf{P}_1 - \mathbf{P}_0)t\\ &= [1-t, t]\begin{bmatrix}\mathbf{P}_0\\ \mathbf{P}_1\end{bmatrix} = [B_{10}(t), B_{11}(t)]\begin{bmatrix}\mathbf{P}_0\\ \mathbf{P}_1\end{bmatrix},\end{aligned} \tag{1.29}$$

where $B_{1i}(t)$ are the Bernstein polynomials of degree 1 [Equation (6.5)]. The Cartesian product of Equation (1.29) with itself is

$$\begin{aligned}\mathbf{P}(u,w) &= [B_{10}(u), B_{11}(u)]\begin{bmatrix}\mathbf{P}_{00} & \mathbf{P}_{01}\\ \mathbf{P}_{10} & \mathbf{P}_{11}\end{bmatrix}\begin{bmatrix}B_{10}(w)\\ B_{11}(w)\end{bmatrix}\\ &= [1-u, u]\begin{bmatrix}\mathbf{P}_{00} & \mathbf{P}_{01}\\ \mathbf{P}_{10} & \mathbf{P}_{11}\end{bmatrix}\begin{bmatrix}1-w\\ w\end{bmatrix}\\ &= \mathbf{P}_{00}(1-u)(1-w) + \mathbf{P}_{01}(1-u)w + \mathbf{P}_{10}u(1-w) + \mathbf{P}_{11}uw,\end{aligned}$$

and this is the parametric expression of the bilinear surface patch, Equation (2.8).

The second example starts with the parametric cubic polynomial that passes through four given points. This curve is derived from first principles in Section 3.1 and is given by Equation (3.6), duplicated here

$$\begin{aligned}\mathbf{P}(t) &= (t^3, t^2, t, 1)\begin{bmatrix}-4.5 & 13.5 & -13.5 & 4.5\\ 9.0 & -22.5 & 18 & -4.5\\ -5.5 & 9.0 & -4.5 & 1.0\\ 1.0 & 0 & 0 & 0\end{bmatrix}\begin{bmatrix}\mathbf{P}_1\\ \mathbf{P}_2\\ \mathbf{P}_3\\ \mathbf{P}_4\end{bmatrix}\\ &= (t^3, t^2, t, 1)\mathbf{N}\begin{bmatrix}\mathbf{P}_1\\ \mathbf{P}_2\\ \mathbf{P}_3\\ \mathbf{P}_4\end{bmatrix}.\end{aligned} \tag{3.6}$$

The principle of Cartesian product is now applied to multiply this curve by itself in order to obtain a bicubic surface patch that passes through 16 given points. The result is obtained immediately

$$\mathbf{P}(u,w) = (u^3, u^2, u, 1)\mathbf{N}\begin{bmatrix}\mathbf{P}_{33} & \mathbf{P}_{32} & \mathbf{P}_{31} & \mathbf{P}_{30}\\ \mathbf{P}_{23} & \mathbf{P}_{22} & \mathbf{P}_{21} & \mathbf{P}_{20}\\ \mathbf{P}_{13} & \mathbf{P}_{12} & \mathbf{P}_{11} & \mathbf{P}_{10}\\ \mathbf{P}_{03} & \mathbf{P}_{02} & \mathbf{P}_{01} & \mathbf{P}_{00}\end{bmatrix}\mathbf{N}^T\begin{bmatrix}w^3\\ w^2\\ w\\ 1\end{bmatrix}. \tag{1.30}$$

Note that this result is also obtained in Section 3.6.1 [Equation (3.27)], where it is derived from first principles and requires the solution of a system of 16 equations. Cartesian product is obviously a useful, simple, and elegant method to easily derive the expressions of many types of surfaces.

1.10 Connecting Surface Patches

Often, a complex surface is constructed of individual patches that have to be connected smoothly, which is why this short section examines the conditions required for the smooth connection of two rectangular patches. Figure 1.20 illustrates two patches $\mathbf{P}(u,w)$ and $\mathbf{Q}(u,w)$ connected along the w direction such that $\mathbf{P}(1,w)=\mathbf{Q}(0,w)$ for $0\le w\le 1$. Specifically, the two corner points $\mathbf{Q}_{00}$ and $\mathbf{P}_{10}$ are identical and so are $\mathbf{Q}_{01}$ and $\mathbf{P}_{11}$. The two patches will connect smoothly if any of the following conditions are met:

1. $\mathbf{Q}^u(0,w)=\mathbf{P}^u(1,w)$ for $0\le w\le 1$.
2. $\mathbf{Q}^u(0,w)=f(w)\mathbf{P}^u(1,w)$ for $0\le w\le 1$ and a positive function $f(w)$.
3. $\mathbf{Q}^u(0,w)=f(w)\mathbf{P}^u(1,w)+g(w)\mathbf{P}^w(1,w)$ for $0\le w\le 1$ and positive functions $f(w)$ and $g(w)$.

These conditions involve the three tangent vectors:

1. $\mathbf{Q}^u(0,w)$, the tangent in the u direction of patch $\mathbf{Q}$ at $u=0$.
2. $\mathbf{P}^u(1,w)$, the tangent in the u direction of $\mathbf{P}$ at $u=1$.
3. $\mathbf{P}^w(1,w)$, the tangent in the w direction of $\mathbf{P}$ at $u=1$.

Condition 1 implies that tangents 1 and 2 are equal. Condition 2 implies that they point in the same direction but their sizes differ. Condition 3 means that tangent 1 does not point in the direction of tangent 2, but lies in the plane defined by tangents 2 and 3.

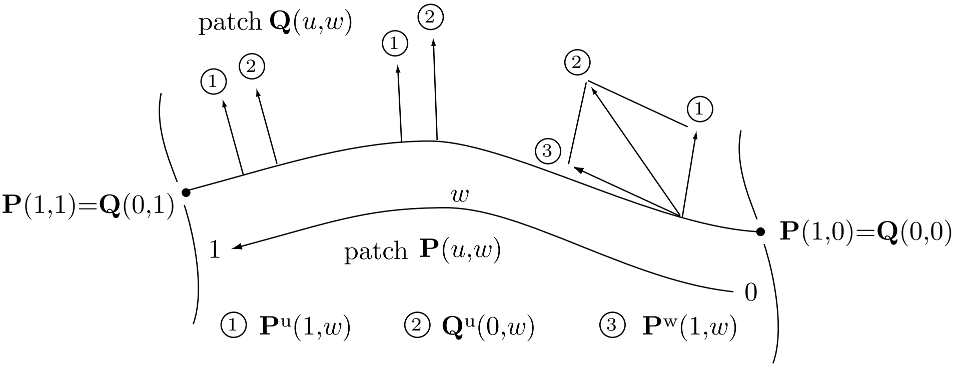

patch $\mathbf{Q}(u,w)$
①
②
①
②
②
①
③
$\mathbf{P}(1,1)=\mathbf{Q}(0,1)$
w
$\mathbf{P}(1,0)=\mathbf{Q}(0,0)$
1
patch $\mathbf{P}(u,w)$
0
① $\mathbf{P}^{\mathrm{u}}(1,w)$
② $\mathbf{Q}^{\mathrm{u}}(0,w)$
③ $\mathbf{P}^{\mathrm{w}}(1,w)$

Figure 1.20: Tangent Vectors For Smooth Connection.

Note that condition 3 includes condition 2 (in the special case $g(w)=0$) and condition 2 includes condition 1 (in the special case $f(w)=1$).

1.11 Fast Computation of a Bicubic Patch

A complete rectangular surface patch is displayed as a wireframe by drawing two families of curves, in the u and w directions, as pointed out in Section 1.8.2. This section shows how to apply the technique of forward differences to the problem of fast computation of these curves. The material presented here is an extension of the ideas and methods presented in Section 1.5.1. We limit this discussion to a general bicubic surface patch, whose expression is

$$\mathbf{P}(u,w)=(u^3,u^2,u,1)\begin{bmatrix}\mathbf{M}_{00}&\mathbf{M}_{01}&\mathbf{M}_{02}&\mathbf{M}_{03}\\ \mathbf{M}_{10}&\mathbf{M}_{11}&\mathbf{M}_{12}&\mathbf{M}_{13}\\ \mathbf{M}_{20}&\mathbf{M}_{21}&\mathbf{M}_{22}&\mathbf{M}_{23}\\ \mathbf{M}_{30}&\mathbf{M}_{31}&\mathbf{M}_{32}&\mathbf{M}_{33}\end{bmatrix}\begin{bmatrix}w^3\\ w^2\\ w\\ 1\end{bmatrix}. \tag{1.31}$$

(Where matrix elements $\mathbf{M}_{ij}$ are derived from the 16 points $\mathbf{P}_{ij}$ and from the elements of matrix $\mathbf{N}$. Compare with Equation (3.21).)

For a fixed w, the surface $\mathbf{P}(u,w)$ reduces to a PC curve in the u direction $\mathbf{P}_w(u)=\mathbf{A}u^3+\mathbf{B}u^2+\mathbf{C}u+\mathbf{D}$. Each of the four coefficients is a cubic polynomial in w as follows:

$$\begin{aligned}\mathbf{A}(w)&=\mathbf{M}_{00}w^3+\mathbf{M}_{01}w^2+\mathbf{M}_{02}w+\mathbf{M}_{03},\\ \mathbf{B}(w)&=\mathbf{M}_{10}w^3+\mathbf{M}_{11}w^2+\mathbf{M}_{12}w+\mathbf{M}_{13},\\ \mathbf{C}(w)&=\mathbf{M}_{20}w^3+\mathbf{M}_{21}w^2+\mathbf{M}_{22}w+\mathbf{M}_{23},\\ \mathbf{D}(w)&=\mathbf{M}_{30}w^3+\mathbf{M}_{31}w^2+\mathbf{M}_{32}w+\mathbf{M}_{33}.\end{aligned}$$

Applying the forward differences technique of Section 1.5.1, we can compute the n points $\mathbf{P}_w(0)$, $\mathbf{P}_w(\Delta)$, $\mathbf{P}_w(2\Delta)$,..., $\mathbf{P}_w([n-1]\Delta)$ [where $(n-1)\Delta=1$] with three additions and three assignments for each point. This, however, requires that the four quantities $\mathbf{A}(w)$, $\mathbf{B}(w)$, $\mathbf{C}(w)$, and $\mathbf{D}(w)$ be computed first, which involves multiplications and exponentiations. Moreover, to display the entire surface patch we need to compute and display U curves $\mathbf{P}_w(u)$ for U values of w in the interval $[0,1]$. The natural solution is to apply forward differences to the computations of $\mathbf{A}(w)$, $\mathbf{B}(w)$, $\mathbf{C}(w)$, and $\mathbf{D}(w)$ for each value of w.

To compute $\mathbf{A}(w)=\mathbf{M}_{00}w^3+\mathbf{M}_{01}w^2+\mathbf{M}_{02}w+\mathbf{M}_{03}$ we compute the following

$$\begin{gathered}\mathbf{A}(0)=\mathbf{M}_{03},\quad \mathbf{dA}(0)=\mathbf{M}_{00}\Delta^3+\mathbf{M}_{01}\Delta^2+\mathbf{M}_{02}\Delta,\quad \mathbf{ddA}(0)=6\mathbf{M}_{00}\Delta^3+2\mathbf{M}_{01}\Delta^2,\\ \mathbf{dddA}=6\mathbf{M}_{00}\Delta^3,\\ \mathbf{A}(\Delta)=\mathbf{A}(0)+\mathbf{dA}(0),\quad \mathbf{dA}(\Delta)=\mathbf{dA}(0)+\mathbf{ddA}(0),\quad \mathbf{ddA}(\Delta)=\mathbf{ddA}(0)+\mathbf{dddA},\\ \mathbf{A}([j+1]\Delta)=\mathbf{A}(j\Delta)+\mathbf{dA}(j\Delta),\\ \mathbf{dA}([j+1]\Delta)=\mathbf{dA}(j\Delta)+\mathbf{ddA}(j\Delta),\\ \mathbf{ddA}([j+1]\Delta)=\mathbf{ddA}(j\Delta)+\mathbf{dddA},\end{gathered}$$

and similarly for $\mathbf{B}(w)$, $\mathbf{C}(w)$, and $\mathbf{D}(w)$. Each requires three additions and three assignments, for a total of 12 additions and 12 assignments.

Thus, a complete curve $\mathbf{P}(u,j\Delta)$ is drawn in the u direction on the surface in the following two steps:

1. Compute $\mathbf{A}(j\Delta)$ from $\mathbf{A}([j-1]\Delta)$, $\mathbf{dA}([j-1]\Delta)$, and $\mathbf{ddA}([j-1]\Delta)$ and similarly for $\mathbf{B}(j\Delta)$, $\mathbf{C}(j\Delta)$, and $\mathbf{D}(j\Delta)$, in 12 additions and 12 assignments.

2. Use these four quantities to compute the n points $\mathbf{P}(0,j\Delta)$, $\mathbf{P}(\Delta,j\Delta)$, $\mathbf{P}(2\Delta,j\Delta)$, up to $\mathbf{P}(1,j\Delta)$, in three additions and three assignments for each point.

The total number of simple operations required for drawing curve $\mathbf{P}(u,j\Delta)$ is therefore $12+12+n(3+3)=6n+24$. If U such curves are drawn in the u direction, the total number of operations is $(6n+24)U$.

To complete the wireframe, another family of W curves of the form $\mathbf{P}(i\Delta,w)$ should be computed and displayed. We assume that m points are computed for each curve, which brings the total number of operations for this family of curves to $(6m+24)W$.

A PC curve $\mathbf{P}_u(w)$ in the w direction on the surface has the form $\mathbf{P}_u(w)=\mathbf{E}w^3+\mathbf{F}w^2+\mathbf{G}w+\mathbf{H}$, where each of the four coefficients is a cubic polynomial in u as follows:

$$\begin{aligned}
\mathbf{E}(u) &= \mathbf{M}_{00}u^3+\mathbf{M}_{10}u^2+\mathbf{M}_{20}u+\mathbf{M}_{30},\\
\mathbf{F}(u) &= \mathbf{M}_{01}u^3+\mathbf{M}_{11}u^2+\mathbf{M}_{21}u+\mathbf{M}_{31},\\
\mathbf{G}(u) &= \mathbf{M}_{02}u^3+\mathbf{M}_{12}u^2+\mathbf{M}_{22}u+\mathbf{M}_{32},\\
\mathbf{H}(u) &= \mathbf{M}_{03}u^3+\mathbf{M}_{13}u^2+\mathbf{M}_{23}u+\mathbf{M}_{33}.
\end{aligned}$$

Thus, $\mathbf{E}$, $\mathbf{F}$, $\mathbf{G}$, and $\mathbf{H}$ are similar to $\mathbf{A}(w)$, $\mathbf{B}(w)$, $\mathbf{C}(w)$, and $\mathbf{D}(w)$, but are computed with the transpose of matrix $\mathbf{M}$.

A complete curve $\mathbf{P}(i\Delta,w)$ is drawn in the w direction on the surface in the following two steps:

1. Compute $\mathbf{E}(i\Delta)$, $\mathbf{F}(i\Delta)$, $\mathbf{G}(i\Delta)$, and $\mathbf{H}(i\Delta)$ from the corresponding quantities for $[i-1]\Delta$ in 12 additions and 12 assignments.

2. Use these four quantities to compute the m points $\mathbf{P}(i\Delta,0)$, $\mathbf{P}(i\Delta,\Delta)$, $\mathbf{P}(i\Delta,2\Delta)$, up to $\mathbf{P}(i\Delta,1)$, in three additions and three assignments for each point.

The total number of simple operations required to compute the m points for curve $\mathbf{P}(i\Delta,w)$ is therefore $6m+24$. If W such curves are drawn in the w direction, the total number of operations is $(6m+24)W$.

Thus, it seems that the entire wireframe can be computed and drawn with $(6n+24)U+(6m+24)W$ operations. For $m=n$ and $U=W$ this becomes $2(6n+24)U$. Typical values of these parameters may be $m=n=100$ and $U=W=15$, which results in $624\times30=18{,}720$ operations.

However, as Figure 1.21 illustrates, some of the points traversed by the curves of the two families are identical, so a sophisticated algorithm may identify them and store them in memory to eliminate double computations and thereby reduce the total number of operations. The figure shows seven curves in the w direction, with 13 points each (the white circles) and five curves in the u direction, consisting of 19 points each (the black circles). Thus, $n=19$, $m=13$, $W=7$, and $U=5$. The total number of points is $19\times5+13\times7=186$, and of these, 7×5, or about 19%, are identical (the $U\times W$ squares).

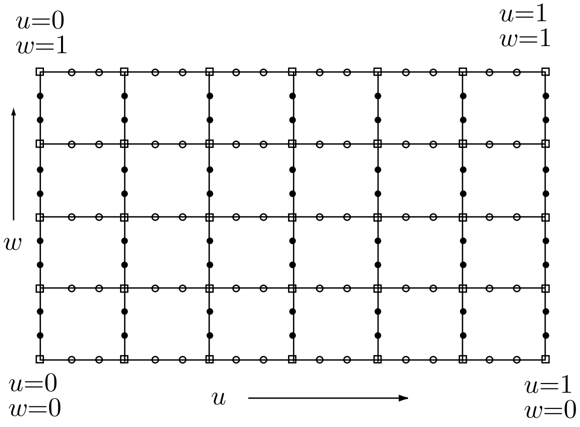


Figure 1.21: A Rectangular Wireframe With 186 Points.

1.12 Subdividing a Surface Patch

The surface subdivision method illustrated here is based on the approach employed in Section 1.5.2 to subdivide a curve. Hence, the reader is advised to read and understand Section 1.5.2 before tackling the material presented here.

Imagine a user trying to construct a surface patch with an interactive algorithm. The patch is based on quantities $\mathbf{P}_{ij}$ that are normally points (some of these quantities may be tangent vectors, but we'll refer to them as points), but the surface refuses to take the desired shape even after the points $\mathbf{P}_{ij}$ have been moved about, shuffled, and manipulated endlessly. This is a common case and it indicates that more points are needed. Just adding new points is a bad approach, because the extra points will modify the shape of the surface and will therefore require the designer to start afresh. A better solution is to add points in such a way that the new surface will have the same shape as the original one. A surface subdivision method takes a surface patch defined by n points $\mathbf{P}_{ij}$ and partitions it into several smaller patches such that together those patches have the same shape as the original surface, and each is defined by n points $\mathbf{Q}_{ij}$, each of which is computed from the original points.

We illustrate this approach to surface subdivision using the bicubic surface patch as an example. The general expression of such a patch is Equation (3.21), duplicated here

$$\mathbf{P}(u,w) = (u^3,u^2,u,1)\mathbf{N}\begin{bmatrix}\mathbf{P}_{33} & \mathbf{P}_{32} & \mathbf{P}_{31} & \mathbf{P}_{30}\\ \mathbf{P}_{23} & \mathbf{P}_{22} & \mathbf{P}_{21} & \mathbf{P}_{20}\\ \mathbf{P}_{13} & \mathbf{P}_{12} & \mathbf{P}_{11} & \mathbf{P}_{10}\\ \mathbf{P}_{03} & \mathbf{P}_{02} & \mathbf{P}_{01} & \mathbf{P}_{00}\end{bmatrix}\mathbf{N}^T\begin{bmatrix}w^3\\ w^2\\ w\\ 1\end{bmatrix} = \mathbf{UNPN}^T\mathbf{W}^T,$$

where both u and w vary independently over the interval $[0,1]$. We now select four numbers u_1, u_2, w_1, and w_2 that satisfy $0 \le u_1 < u_2 \le 1$ and $0 \le w_1 < w_2 \le 1$. The expression $\mathbf{P}(u,w)$ where u and w vary in the intervals $[u_1,u_2]$ and $[w_1,w_2]$, respectively, is a rectangle on this surface (Figure 1.22a).

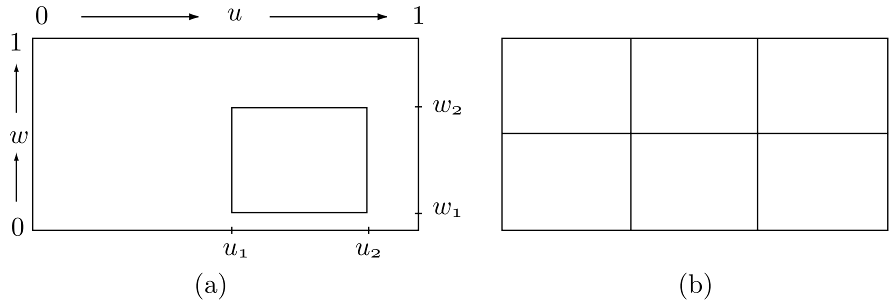


Figure 1.22: Rectangles On A Bicubic Surface Patch.

The next step is to substitute new parameters t and v for u and w, respectively, and express rectangle $\mathbf{P}(u,w)$ as $\mathbf{P}(t,v)$ where both t and v vary independently in $[0,1]$. If the original rectangle is expressed as

$$\mathbf{P}(u,w)=\mathbf{UNPN}^T\mathbf{W}^T,\quad u_1\le u\le u_2,\quad w_1\le w\le w_2,$$

then after the substitutions its shape will be the same and its form will be

$$\mathbf{P}(t,v)=\mathbf{TNQN}^T\mathbf{V}^T,\ \text{for}\ \ 0\le t\le 1,\quad 0\le v\le 1.$$

Both rectangles have the same shape, but $\mathbf{P}(t,v)$ is defined by means of new points $\mathbf{Q}_{ij}$, and the main task is to figure out how to compute the $\mathbf{Q}_{ij}$'s from the original points $\mathbf{P}_{ij}$ while preserving the shape.

Once this is clear, a surface patch can be divided into several rectangles, as in Figure 1.22b, and each expressed in terms of new points. Each new rectangle has the same shape as that part of the surface from which it came, but is defined by the same number of points as the entire original surface. Each rectangle can now be reshaped because of the extra points.

The parameter substitutions from u and w to t and v are the linear relations $t=(u-u_1)/(u_2-u_1)$ and $v=(w-w_1)/(w_2-w_1)$. These imply

$$u=(u_2-u_1)\left[t+\frac{u_1}{u_2-u_1}\right]\ \text{and}\ w=(w_2-w_1)\left[v+\frac{w_1}{w_2-w_1}\right].$$

The rectangle is expressed by means of the new parameters in the form

$$\begin{aligned}&\mathbf{P}(t,v)\\&=\left[(u_2-u_1)^3\left[t+\frac{u_1}{u_2-u_1}\right]^3,(u_2-u_1)^2\left[t+\frac{u_1}{u_2-u_1}\right]^2,(u_2-u_1)\left[t+\frac{u_1}{u_2-u_1}\right],1\right]\\&\quad\times\mathbf{NPN}^T\begin{bmatrix}(w_2-w_1)^3\left[v+\frac{w_1}{w_2-w_1}\right]^3\\(w_2-w_1)^2\left[v+\frac{w_1}{w_2-w_1}\right]^2\\(w_2-w_1)\left[v+\frac{w_1}{w_2-w_1}\right]\\1\end{bmatrix}\end{aligned}$$

$$
\begin{aligned}
&= [t^3, t^2, t, 1] \begin{bmatrix} (u_2-u_1)^3 & 0 & 0 & 0 \\ 3u_1(u_2-u_1)^2 & (u_2-u_1)^2 & 0 & 0 \\ 3u_1^2(u_2-u_1) & 2u_1(u_2-u_1) & u_2-u_1 & 0 \\ u_1^3 & u_1^2 & u_1 & 0 \end{bmatrix} \qquad (1.32) \\
&\times \mathbf{NPN}^T \begin{bmatrix} (w_2-w_1)^3 & 3w_1(w_2-w_1)^2 & 3w_1^2(w_2-w_1) & w_1^3 \\ 0 & (w_2-w_1)^2 & 2w_1(w_2-w_1) & w_1^2 \\ 0 & 0 & w_2-w_1 & w_1 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} v^3 \\ v^2 \\ v \\ 1 \end{bmatrix} \\
&= [t^3, t^2, t, 1]\mathbf{LNPN}^T\mathbf{R}[v^3, v^2, v, 1]^T \\
&= [t^3, t^2, t, 1]\mathbf{NQN}^T[v^3, v^2, v, 1]^T,
\end{aligned}
$$

where the new points $\mathbf{Q}$ are related to the original points by $\mathbf{Q} = \mathbf{N}^{-1}\mathbf{LNPN}^T\mathbf{R}(\mathbf{N}^T)^{-1}$.

To illustrate the application of matrices $\mathbf{L}$ and $\mathbf{R}$ of Equation (1.32), we apply them to the special case $u_1 = 0$, $u_2 = 1/2$, $w_1 = 1/2$, and $w_2 = 1$ to isolate the gray rectangle of Figure 1.23. The resulting matrices are

$$
\mathbf{L} = \begin{pmatrix} 1/8 & 0 & 0 & 0 \\ 0 & 1/4 & 0 & 0 \\ 0 & 0 & 1/2 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \quad \mathbf{R} = \begin{pmatrix} 1/8 & 3/8 & 3/8 & 1/8 \\ 0 & 1/4 & 1/2 & 1/4 \\ 0 & 0 & 1/2 & 1/2 \\ 0 & 0 & 0 & 1 \end{pmatrix}.
$$

These should be compared with matrices $\mathbf{L}$ and $\mathbf{R}$ of Equations (1.12) and (1.14), respectively.

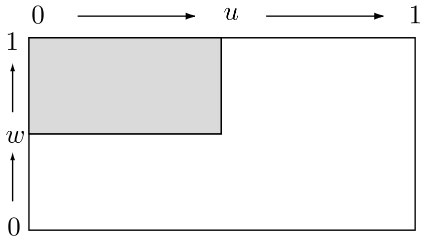


Figure 1.23: A Rectangle on a Surface Patch.

1.13 Surface Normals

The main aim of computer graphics is to display real-looking, solid surfaces. This is done by applying a shading algorithm to every pixel on the surface. Such algorithms may be very complex, but the main task of shading is to compute the amount of light reflected from every surface point. This requires the calculation of the normal to the surface at every point. The normal is the vector that's perpendicular to the surface at the point. It can be defined in two ways:

1. We imagine a flat plane touching the surface at the point (this is called the *osculating plane*). The normal is the vector that's perpendicular to this plane.

2. We calculate two tangent vectors to the surface at the point. The normal is the vector that's perpendicular to both tangents.

The following shows how to calculate the normal vectors for various types of surfaces.

- The normal to the implicit surface $F(x,y,z)=0$ at point (x_0,y_0,z_0) is the vector

$$\left(\frac{\partial F(x_0,y_0,z_0)}{\partial x}, \frac{\partial F(x_0,y_0,z_0)}{\partial y}, \frac{\partial F(x_0,y_0,z_0)}{\partial z}\right).$$

Example: The ellipsoid $x^2/a^2+y^2/b^2+z^2/c^2-1=0$. A partial derivative would be, for example, $\partial f/\partial x = 2x/a^2$, so the normal is

$$\left(\frac{2x}{a^2}, \frac{2y}{b^2}, \frac{2z}{c^2}\right) \quad \text{which is in the same direction as} \quad \left(\frac{x}{a^2}, \frac{y}{b^2}, \frac{z}{c^2}\right).$$

For example, the normal at point $(0,0,-c)$ is $(0,0,-c/c^2)=(0,0,-1/c)$. This is a vector in the direction $(0,0,-1)$.

◇ **Exercise 1.30:** What is the normal to the explicit surface $z=f(x,y)$ at point (x_0,y_0)?

> No money, no job, no rent. Hey, I'm back to normal.
> —Mickey Rourke (as Henry Chinaski) in *Barfly*, 1987.

- The normal to the parametric surface $\mathbf{P}(u,w)$ is calculated in two steps. In step 1, the two tangent vectors $\mathbf{U}=\partial\mathbf{P}(u,w)/\partial u$ and $\mathbf{V}=\partial\mathbf{P}(u,w)/\partial w$ are calculated. In step 2, the normal is calculated as their cross-product $\mathbf{U}\times\mathbf{V}$ (Equation (1.5), page 7).

- The normal to a polygon in a polygonal surface (Section 2.2) can be calculated as shown for an implicit surface. The (implicit) plane equation is $F(x,y,z)=Ax+By+Cz+D=0$, so the normal is $\left(\frac{\partial F}{\partial x}, \frac{\partial F}{\partial y}, \frac{\partial F}{\partial z}\right)$, which is simply (A,B,C). Another way of calculating the normal, especially suited for triangles, is to find two vectors on the surface and calculate their cross-product. Two suitable vectors are $\mathbf{U}=\mathbf{P}_1-\mathbf{P}_2$ and $\mathbf{V}=\mathbf{P}_1-\mathbf{P}_3$, where $\mathbf{P}_1$, $\mathbf{P}_2$, and $\mathbf{P}_3$ are the triangle's corners. Their cross product is

$$\mathbf{U}\times\mathbf{V}=(U_yV_z-U_zV_y, U_zV_x-U_xV_z, U_xV_y-U_yV_x).$$

Example: A polygon with vertices $(1,1,-1)$, $(1,1,1)$ $(1,-1,1)$, and $(1,-1,-1)$. All the vertices have $x=1$, so they are on the $x=1$ plane, which means that the normal should be a vector in the x direction. The calculation is straightforward:

$$\begin{aligned}\mathbf{U}&=(1,1,1)-(1,1,-1)=(0,0,2),\\ \mathbf{V}&=(1,-1,1)-(1,1,-1)=(0,-2,2),\\ \mathbf{U}\times\mathbf{V}&=(0-(-4),0-0,0-0)=(4,0,0).\end{aligned}$$

This is a vector in the right direction.

◇ **Exercise 1.31:** What will happen if we calculate **U** as $(1,1,-1)-(1,1,1)$?

◇ **Exercise 1.32:** Find the normal to the pyramid face of Equation (Ans.4).

◇ **Exercise 1.33:** Find the normal to the cone of Equation (Ans.3).

◇ **Exercise 1.34:** Construct a cylinder as a sweep surface (Chapter 9) and find its normal vector. Assume that the cylinder is swept when the line from $(-a,0,R)$ to $(a,0,R)$ is rotated 360° about the x axis.

> John's leaning against the window, probably trying to figure out what parametric equation generated the petals on that eight-foot-tall, carnivorous plant. He turns around to be introduced. "John Cantrell."
> "Harvard Li. Didn't you get my e-mail?"
> Harvard Li! Now Randy is starting to remember this guy. Founder of Harvard Computer Company, a medium-sized PC clone manufacturer in Taiwan.
>
> Neal Stephenson, *Cryptonomicon* (2002)