2.7 Curves and Surfaces

The geometry of curves, and especially surfaces, plays a central role in graphics, and here, we review the basics of curves and surfaces in 2D and 3D space.

2.7.1 2D Implicit Curves

Intuitively, a *curve* is a set of points that can be drawn on a piece of paper without lifting the pen. A common way to describe a curve is using an *implicit equation*. An implicit equation in two dimensions has the form

$$f(x, y) = 0.$$

The function $f(x, y)$ returns a real value. Points (x, y) where this value is zero are on the curve, and points where the value is nonzero are not on the curve. For example, let's say that $f(x, y)$ is

$$f(x, y) = (x - x_c)^2 + (y - y_c)^2 - r^2, \tag{2.9}$$

where (x_c, y_c) is a 2D point and r is a nonzero real number. If we take $f(x, y) = 0$, the points where this equality holds are on the circle with center (x_c, y_c) and radius r. The reason that this is called an "implicit" equation is that the points (x, y) on the curve cannot be immediately calculated from the equation and instead must be determined by solving the equation. Thus, the points on the curve are not generated by the equation *explicitly*, but they are buried somewhere *implicitly* in the equation.

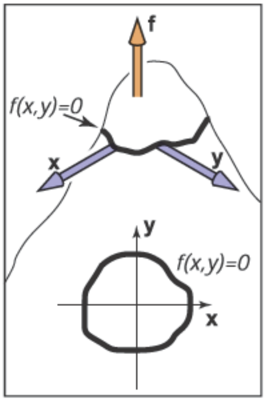

Figure 2.24. An implicit function $f(x,y) = 0$ can be thought of as a height field where f is the height (top). A path where the height is zero is the implicit curve (bottom).

It is interesting to note that f does have values for all (x, y). We can think of f as a terrain, with sea level at $f = 0$ (Figure 2.24). The shore is the implicit curve. The value of f is the altitude. Another thing to note is that the curve partitions space into regions where $f > 0$, $f < 0$, and $f = 0$. So you evaluate f to decide whether a point is "inside" a curve. Note that $f(x, y) = c$ is a curve for any constant c, and $c = 0$ is just used as a convention. For example, if $f(x, y) = x^2 + y^2 - 1$, varying c just gives a variety of circles centered at the origin (Figure 2.25).

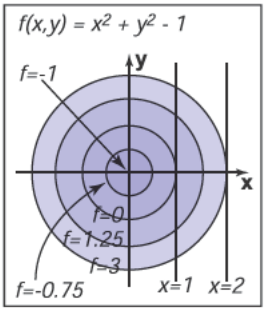

Figure 2.25. An implicit function defines a curve for any constant value, with zero being the usual convention.

We can compress our notation using vectors. If we have $\mathbf{c} = (x_c, y_c)$ and $\mathbf{p} = (x, y)$, then our circle with center $\mathbf{c}$ and radius r is defined by those position vectors that satisfy

$$(\mathbf{p} - \mathbf{c}) \cdot (\mathbf{p} - \mathbf{c}) - r^2 = 0.$$

This equation, if expanded algebraically, will yield Equation (2.9), but it is easier to see that this is an equation for a circle by "reading" the equation geometrically. It reads, "points $\mathbf{p}$ on the circle have the following property: the vector from $\mathbf{c}$ to $\mathbf{p}$ when dotted with itself has value r^2." Because a vector dotted with itself is just its own length squared, we could also read the equation as, "points $\mathbf{p}$ on the circle have the following property: the vector from $\mathbf{c}$ to $\mathbf{p}$ has squared length r^2."

Even better, is to observe that the squared length is just the squared distance from $\mathbf{c}$ to $\mathbf{p}$, which suggests the equivalent form

$$\|\mathbf{p} - \mathbf{c}\|^2 - r^2 = 0,$$

and, of course, this suggests

$$\|\mathbf{p} - \mathbf{c}\| - r = 0.$$

The above could be read "the points $\mathbf{p}$ on the circle are those a distance r from the center point $\mathbf{c}$," which is as good a definition of circle as any. This illustrates that the vector form of an equation often suggests more geometry and intuition than the equivalent full-blown Cartesian form with x and y. For this reason, it is usually advisable to use vector forms when possible. In addition, you can support a vector class in your code; the code is cleaner when vector forms are used. The vector-oriented equations are also less error prone in implementation: once you implement and debug vector types in your code, the cut-and-paste errors involving x, y, and z will go away. It takes a little while to get used to vectors in these equations, but once you get the hang of it, the payoff is large.

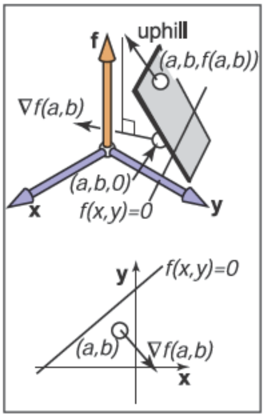

Figure 2.26. A surface height = $f(x,y)$ is locally planar near $(x,y) = (a,b)$. The gradient is a projection of the uphill direction onto the height = 0 plane.

2.7.2 The 2D Gradient

If we think of the function $f(x, y)$ as a height field with height $= f(x, y)$, the *gradient* vector points in the direction of maximum upslope, i.e., straight uphill. The gradient vector $\nabla f(x, y)$ is given by

$$\nabla f(x, y) = \left(\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y} \right).$$

The gradient vector evaluated at a point on the implicit curve $f(x, y) = 0$ is perpendicular to the *tangent* vector of the curve at that point. This perpendicular vector is usually called the *normal vector* to the curve. In addition, since the gradient points uphill, it indicates the direction of the $f(x, y) > 0$ region.

In the context of height fields, the geometric meaning of partial derivatives and gradients is more visible than usual. Suppose that near the point (a, b), $f(x, y)$ is a plane (Figure 2.26). There is a specific uphill and downhill direction. At right angles to this direction is a direction that is level with respect to the plane. Any intersection between the plane and the $f(x, y) = 0$ plane will be in the direction that is level. Thus, the uphill/downhill directions will be perpendicular to the line of intersection $f(x, y) = 0$. To see why the partial derivative has something to do with this, we need to visualize its geometric meaning. Recall that the conventional derivative of a 1D function $y = g(x)$ is

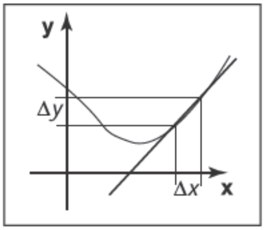

Figure 2.27. The derivative of a 1D function measures the slope of the line tangent to the curve.

$$\frac{dy}{dx} \equiv \lim_{\Delta x \to 0} \frac{\Delta y}{\Delta x} = \lim_{\Delta x \to 0} \frac{g(x + \Delta x) - g(x)}{\Delta x}. \tag{2.10}$$

This measures the *slope* of the *tangent* line to g (Figure 2.27).

The partial derivative is a generalization of the 1D derivative. For a 2D function $f(x, y)$, we can't take the same limit for x as in Equation (2.10), because f can change in many ways for a given change in x. However, if we hold y constant, we can define an analog of the derivative, called the *partial derivative* (Figure 2.28):

$$\frac{\partial f}{\partial x} \equiv \lim_{\Delta x \to 0} \frac{f(x + \Delta x, y) - f(x, y)}{\Delta x}.$$

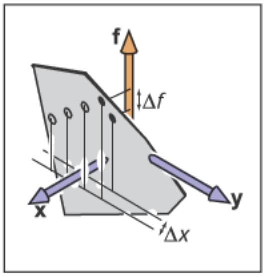

Figure 2.28. The partial derivative of a function f with respect to x must hold y constant to have a unique value, as shown by the dark point. The hollow points show other values of f that do not hold y constant.

Why is it that the partial derivatives with respect to x and y are the components of the gradient vector? Again, there is more obvious insight in the geometry than in the algebra. In Figure 2.29, we see the vector **a** travels along a path where **f** does not change. Note that this is again at a small enough scale that the surface height $(x, y) = f(x, y)$ can be considered locally planar. From the figure, we see that the vector $\mathbf{a} = (\Delta x, \Delta y)$.

Because the uphill direction is perpendicular to **a**, we know the dot product is equal to zero:

$$(\nabla f) \cdot \mathbf{a} \equiv (x_\nabla, y_\nabla) \cdot (x_a, y_a) = x_\nabla \Delta x + y_\nabla \Delta y = 0. \qquad (2.11)$$

We also know that the change in f in the direction (x_a, y_a) equals zero:

$$\Delta f = \frac{\partial f}{\partial x}\Delta x + \frac{\partial f}{\partial y}\Delta y \equiv \frac{\partial f}{\partial x}x_a + \frac{\partial f}{\partial y}y_a = 0.$$

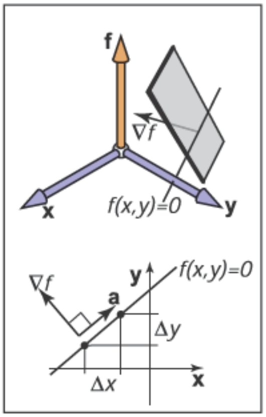

Figure 2.29. The vector **a** points in a direction where f has no change and is thus perpendicular to the gradient vector ∇f.

Given any vectors (x, y) and (x', y') that are perpendicular, we know that the angle between them is 90 degrees, and thus, their dot product equals zero (recall that the dot product is proportional to the cosine of the angle between the two vectors). Thus, we have $xx' + yy' = 0$. Given (x, y), it is easy to construct valid vectors whose dot product with (x, y) equals zero, the two most obvious being $(y, -x)$ and $(-y, x)$; you can verify that these vectors give the desired zero dot product with (x, y). A generalization of this observation is that (x, y) is perpendicular to $k(y, -x)$ where k is any nonzero constant. This implies that

$$(x_a, y_a) = k\left(\frac{\partial f}{\partial y}, -\frac{\partial f}{\partial x}\right). \qquad (2.12)$$

Combining Equations (2.11) and (2.12) gives

$$(x_\nabla, y_\nabla) = k'\left(\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}\right),$$

where k' is any nonzero constant. By definition, "uphill" implies a positive change in f, so we would like $k' > 0$, and $k' = 1$ is a perfectly good convention.

As an example of the gradient, consider the implicit circle $x^2 + y^2 - 1 = 0$ with gradient vector $(2x, 2y)$, indicating that the outside of the circle is the positive region for the function $f(x, y) = x^2 + y^2 - 1$. Note that the length of the gradient vector can be different depending on the multiplier in the implicit equation. For example, the unit circle can be described by $Ax^2 + Ay^2 - A = 0$ for any nonzero A. The gradient for this curve is $(2Ax, 2Ay)$. This will be normal (perpendicular) to the circle, but will have a length determined by A. For $A > 0$, the normal will point outward from the circle, and for $A < 0$, it will point inward. This switch from outward to inward is as it should be, since the positive region switches inside the circle. In terms of the height-field view, $h = Ax^2 + Ay^2 - A$, and the circle is at zero altitude. For $A > 0$, the circle encloses a depression, and for $A < 0$, the circle encloses a bump. As A becomes more negative, the bump increases in height, but the $h = 0$ circle doesn't change. The direction of maximum uphill doesn't change, but the slope increases. The length of the gradient reflects this change in degree of the slope. So intuitively, you can think of the gradient's direction as pointing uphill and its magnitude as measuring how uphill the slope is.

Implicit 2D Lines

The familiar "slope-intercept" form of the line is

$$y = mx + b. \tag{2.13}$$

This can be converted easily to implicit form (Figure 2.30):

$$y - mx - b = 0. \tag{2.14}$$

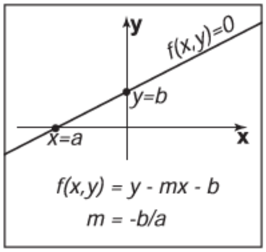

Figure 2.30. A 2D line can be described by the equation $y - mx - b = 0$.

Here, m is the "slope" (ratio of rise to run), and b is the y value where the line crosses the y-axis, usually called the *y-intercept*. The line also partitions the 2D plane, but here "inside" and "outside" might be more intuitively called "over" and "under."

Because we can multiply an implicit equation by any constant without changing the points where it is zero, $kf(x, y) = 0$ is the same curve for any nonzero k. This allows several implicit forms for the same line, for example,

$$2y - 2mx - 2b = 0.$$

One reason the slope-intercept form is sometimes awkward is that it can't represent some lines such as $x = 0$ because m would have to be infinite. For this reason, a more general form is often useful:

$$Ax + By + C = 0, \tag{2.15}$$

for real numbers A, B, C.

Suppose we know two points on the line, (x_0, y_0) and (x_1, y_1). What A, B, and C describe the line through these two points? Because these points lie on the line, they must both satisfy Equation (2.15):

$$Ax_0 + By_0 + C = 0,$$
$$Ax_1 + By_1 + C = 0.$$

Unfortunately, we have two equations and *three* unknowns: A, B, and C. This problem arises because of the arbitrary multiplier we can have with an implicit equation. We could set $C = 1$ for convenience:

$$Ax + By + 1 = 0,$$

but we have a similar problem to the infinite slope case in slope-intercept form: lines through the origin would need to have $A(0) + B(0) + 1 = 0$, which is a contradiction. For example, the equation for a 45–° line through the origin can be written $x - y = 0$, or equally well $y - x = 0$, or even $17y - 17x = 0$, but it cannot be written in the form $Ax + By + 1 = 0$.

Whenever we have such pesky algebraic problems, we try to solve the problems using geometric intuition as a guide. One tool we have, as discussed in Section 2.7.2, is the gradient. For the line $Ax + By + C = 0$, the gradient vector is (A, B). This vector is perpendicular to the line (Figure 2.31), and points to the side of the line where $Ax + By + C$ is positive. Given two points on the line (x_0, y_0) and (x_1, y_1), we know that the vector between them points in the same direction as the line. This vector is just $(x_1 - x_0, y_1 - y_0)$, and because it is parallel to the line, it must also be perpendicular to the gradient vector (A, B). Recall that there are an infinite number of (A, B, C) that describe the line because of the arbitrary scaling property of implicits. We want any one of the valid (A, B, C).

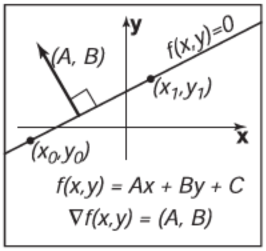

Figure 2.31. The gradient vector (A, B) is perpendicular to the implicit line $Ax + By + C = 0$.

We can start with any (A, B) perpendicular to $(x_1 - x_0, y_1 - y_0)$. Such a vector is just $(A, B) = (y_0 - y_1, x_1 - x_0)$ by the same reasoning as in Section 2.7.2. This means that the equation of the line through (x_0, y_0) and (x_1, y_1) is

$$(y_0 - y_1)x + (x_1 - x_0)y + C = 0. \qquad (2.16)$$

Now we just need to find C. Because (x_0, y_0) and (x_1, y_1) are on the line, they must satisfy Equation (2.16). We can plug either value in and solve for C. Doing this for (x_0, y_0) yields $C = x_0y_1 - x_1y_0$, and thus, the full equation for the line is

$$(y_0 - y_1)x + (x_1 - x_0)y + x_0y_1 - x_1y_0 = 0. \qquad (2.17)$$

Again, this is one of infinitely many valid implicit equations for the line through two points, but this form has no division operation and thus no numerically degenerate cases for points with finite Cartesian coordinates. A nice thing about

Equation (2.17) is that we can always convert to the slope-intercept form (when it exists) by moving the non-y terms to the right-hand side of the equation and dividing by the multiplier of the y term:

$$y = \frac{y_1 - y_0}{x_1 - x_0}x + \frac{x_1 y_0 - x_0 y_1}{x_1 - x_0}.$$

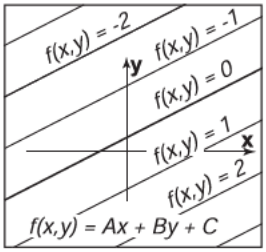

Figure 2.32. The value of the implicit function $f(x,y) = Ax + By + C$ is a constant times the signed distance from $Ax + By + C = 0$.

An interesting property of the implicit line equation is that it can be used to find the signed distance from a point to the line. The value of $Ax + By + C$ is proportional to the distance from the line (Figure 2.32). As shown in Figure 2.33, the distance from a point to the line is the length of the vector $k(A, B)$, which is

$$\text{distance} = k\sqrt{A^2 + B^2}. \tag{2.18}$$

For the point $(x, y) + k(A, B)$, the value of $f(x, y) = Ax + By + C$ is

$$\begin{aligned} f(x + kA, y + kB) &= Ax + kA^2 + By + kB^2 + C \\ &= k(A^2 + B^2). \end{aligned} \tag{2.19}$$

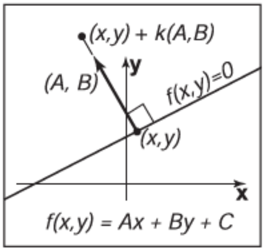

Figure 2.33. The vector $k(A,B)$ connects a point (x,y) on the line closest to a point not on the line. The distance is proportional to k.

The simplification in that equation is a result of the fact that we know (x, y) is on the line, so $Ax + By + C = 0$. From Equations (2.18) and (2.19), we can see that the signed distance from line $Ax + By + C = 0$ to a point (a, b) is

$$\text{Distance} = \frac{f(a, b)}{\sqrt{A^2 + B^2}}.$$

Here, "signed distance" means that its magnitude (absolute value) is the geometric distance, but on one side of the line, distances are positive and on the other, they are negative. You can choose between the equally valid representations $f(x, y) = 0$ and $-f(x, y) = 0$ if your problem has some reason to prefer a particular side being positive. Note that if (A, B) is a unit vector, then $f(a, b)$ is the signed distance. We can multiply Equation (2.17) by a constant that ensures that (A, B) is a unit vector:

$$\begin{aligned} f(x, y) = {} & \frac{y_0 - y_1}{\sqrt{(x_1 - x_0)^2 + (y_0 - y_1)^2}}x + \frac{x_1 - x_0}{\sqrt{(x_1 - x_0)^2 + (y_0 - y_1)^2}}y \\ & + \frac{x_0 y_1 - x_1 y_0}{\sqrt{(x_1 - x_0)^2 + (y_0 - y_1)^2}} = 0. \end{aligned} \tag{2.20}$$

Note that evaluating $f(x, y)$ in Equation (2.20) directly gives the signed distance, but it does require a square root to set up the equation. Implicit lines will turn out to be very useful for triangle rasterization (Section 9.1.2). Other forms for 2D lines are discussed in Chapter 13.

Implicit Quadric Curves

In the previous section, we saw that a linear function $f(x, y)$ gives rise to an implicit line $f(x, y) = 0$. If f is instead a quadratic function of x and y, with the general form

$$Ax^2 + Bxy + Cy^2 + Dx + Ey + F = 0,$$

the resulting implicit curve is called a quadric. Two-dimensional quadric curves include ellipses and hyperbolas, as well as the special cases of parabolas, circles, and lines.

Examples of quadric curves include the circle with center (x_c, y_c) and radius r,

$$(x - x_c)^2 + (y - y_c)^2 - r^2 = 0,$$

and axis-aligned ellipses of the form

$$\frac{(x - x_c)^2}{a^2} + \frac{(y - y_c)^2}{b^2} - 1 = 0,$$

where (x_c, y_c) is the center of the ellipse, and a and b are the minor and major semi-axes (Figure 2.34).

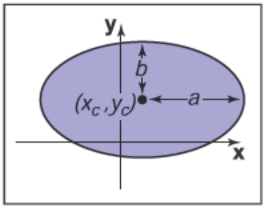

Figure 2.34. The ellipse with center (x_c, y_c) and semi-axes of length a and b.

Try setting $a = b = r$ in the ellipse equation and compare to the circle equation.

2.7.3 3D Implicit Surfaces

Just as implicit equations can be used to define curves in 2D, they can be used to define surfaces in 3D. As in 2D, implicit equations *implicitly* define a set of points that are on the surface:

$$f(x, y, z) = 0.$$

Any point (x, y, z) that is on the surface results in zero when given as an argument to f. Any point not on the surface results in some number other than zero. You can check whether a point is on the surface by evaluating f, or you can check which side of the surface the point lies on by looking at the sign of f, but you cannot always explicitly construct points on the surface. Using vector notation, we will write such functions of $\mathbf{p} = (x, y, z)$ as

$$f(\mathbf{p}) = 0.$$

2.7.4 Surface Normal to an Implicit Surface

A surface normal (which is needed for lighting computations, among other things) is a vector perpendicular to the surface. Each point on the surface may have a different normal vector. In the same way that the gradient provides a normal to an implicit curve in 2D, the surface normal at a point $\mathbf{p}$ on an implicit surface is given by the gradient of the implicit function

$$\mathbf{n} = \nabla f(\mathbf{p}) = \left(\frac{\partial f(\mathbf{p})}{\partial x}, \frac{\partial f(\mathbf{p})}{\partial y}, \frac{\partial f(\mathbf{p})}{\partial z} \right).$$

The reasoning is the same as for the 2D case: the gradient points in the direction of fastest increase in f, which is perpendicular to all directions tangent to the surface, in which f remains constant. The gradient vector points toward the side of the surface where $f(\mathbf{p}) > 0$, which we may think of as "into" the surface or "out from" the surface in a given context. If the particular form of f creates inward-facing gradients, and outward-facing gradients are desired, the surface $-f(\mathbf{p}) = 0$ is the same as surface $f(\mathbf{p}) = 0$ but has directionally reversed gradients, i.e., $-\nabla f(\mathbf{p}) = \nabla(-f(\mathbf{p}))$.

2.7.5 Implicit Planes

As an example, consider the infinite plane through point $\mathbf{a}$ with surface normal $\mathbf{n}$. The implicit equation to describe this plane is given by

$$(\mathbf{p} - \mathbf{a}) \cdot \mathbf{n} = 0. \tag{2.21}$$

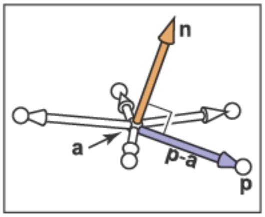

Figure 2.35. Any of the points $\mathbf{p}$ shown are in the plane with normal vector $\mathbf{n}$ that includes point $\mathbf{a}$ if Equation (2.21) is satisfied.

Note that $\mathbf{a}$ and $\mathbf{n}$ are known quantities. The point $\mathbf{p}$ is any unknown point that satisfies the equation. In geometric terms this equation says "the vector from $\mathbf{a}$ to $\mathbf{p}$ is perpendicular to the plane normal." If $\mathbf{p}$ were not in the plane, then $(\mathbf{p} - \mathbf{a})$ would not make a right angle with $\mathbf{n}$ (Figure 2.35).

Sometimes, we want the implicit equation for a plane through points $\mathbf{a}$, $\mathbf{b}$, and $\mathbf{c}$. The normal to this plane can be found by taking the cross product of any two vectors in the plane. One such cross product is

$$\mathbf{n} = (\mathbf{b} - \mathbf{a}) \times (\mathbf{c} - \mathbf{a}).$$

This allows us to write the implicit plane equation:

$$(\mathbf{p} - \mathbf{a}) \cdot ((\mathbf{b} - \mathbf{a}) \times (\mathbf{c} - \mathbf{a})) = 0. \tag{2.22}$$

A geometric way to read this equation is that the volume of the parallelepiped defined by $\mathbf{p} - \mathbf{a}$, $\mathbf{b} - \mathbf{a}$, and $\mathbf{c} - \mathbf{a}$ is zero; i.e., they are coplanar. This can only be true if $\mathbf{p}$ is in the same plane as $\mathbf{a}$, $\mathbf{b}$, and $\mathbf{c}$. The full-blown Cartesian representation for this is given by the determinant (this is discussed in more detail in Section 6.3):

$$\begin{vmatrix} x - x_a & y - y_a & z - z_a \\ x_b - x_a & y_b - y_a & z_b - z_a \\ x_c - x_a & y_c - y_a & z_c - z_a \end{vmatrix} = 0. \qquad (2.23)$$

The determinant can be expanded (see Section 6.3 for the mechanics of expanding determinants) to the bloated form with many terms.

Equations (2.22) and (2.23) are equivalent, and comparing them is instructive. Equation (2.22) is easy to interpret geometrically and will yield efficient code. In addition, it is relatively easy to avoid a typographic error that compiles into incorrect code if it takes advantage of debugged cross and dot product code. Equation (2.23) is also easy to interpret geometrically and will be efficient provided an efficient 3 × 3 determinant function is implemented. It is also easy to implement without a typo if a function *determinant* $(\mathbf{a}, \mathbf{b}, \mathbf{c})$ is available. It will be especially easy for others to read your code if you rename the *determinant* function *volume*. So both Equations (2.22) and (2.23) map well into code. The full expansion of either equation into x-, y-, and z-components is likely to generate typos. Such typos are likely to compile and, thus, to be especially pesky. This is an excellent example of clean math generating clean code and bloated math generating bloated code.

3D Quadric Surfaces

Just as quadratic polynomials in two variables define quadric curves in 2D, quadratic polynomials in x, y, and z define *quadric surfaces* in 3D. For instance, a sphere can be written as

$$f(\mathbf{p}) = (\mathbf{p} - \mathbf{c})^2 - r^2 = 0,$$

and an axis-aligned ellipsoid may be written as

$$f(\mathbf{p}) = \frac{(x - x_c)^2}{a^2} + \frac{(y - y_c)^2}{b^2} + \frac{(z - z_c)^2}{c^2} - 1 = 0.$$

3D Curves from Implicit Surfaces

One might hope that an implicit 3D curve could be created with the form $f(\mathbf{p}) = 0$. However, all such curves are just degenerate surfaces and are rarely useful in practice. A 3D curve can be constructed from the intersection of two simultaneous implicit equations:

$$f(\mathbf{p}) = 0,$$
$$g(\mathbf{p}) = 0.$$

For example, a 3D line can be formed from the intersection of two implicit planes. Typically, it is more convenient to use parametric curves instead; they are discussed in the following sections.

2.7.6 2D Parametric Curves

A *parametric* curve is controlled by a single *parameter* that can be considered a sort of index that moves continuously along the curve. Such curves have the form

$$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} g(t) \\ h(t) \end{bmatrix}.$$

Here, (x, y) is a point on the curve, and t is the parameter that influences the curve. For a given t, there will be some point determined by the functions g and h. For continuous g and h, a small change in t will yield a small change in x and y. Thus, as t continuously changes, points are swept out in a continuous curve. This is a nice feature because we can use the parameter t to explicitly construct points on the curve. Often, we can write a parametric curve in vector form,

$$\mathbf{p} = f(t),$$

where f is a vector-valued function, $f : \mathbb{R} \mapsto \mathbb{R}^2$. Such vector functions can generate very clean code, so they should be used when possible.

We can think of the curve with a position as a function of time. The curve can go anywhere and could loop and cross itself. We can also think of the curve as having a velocity at any point. For example, the point $\mathbf{p}(t)$ is traveling slowly near $t = -2$ and quickly between $t = 2$ and $t = 3$. This type of "moving point" vocabulary is often used when discussing parametric curves even when the curve is not describing a moving point.

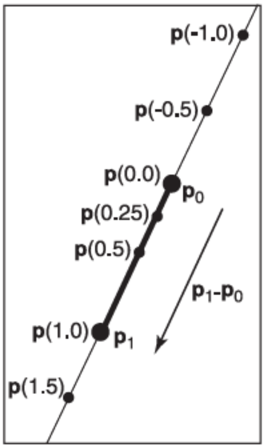

Figure 2.36. A 2D parametric line through $\mathbf{p}_0$ and $\mathbf{p}_1$. The line segment defined by $t \in [0,1]$ is shown in bold.

2D Parametric Lines

A parametric line in 2D that passes through points $\mathbf{p}_0 = (x_0, y_0)$ and $\mathbf{p}_1 = (x_1, y_1)$ can be written as

$$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} x_0 + t(x_1 - x_0) \\ y_0 + t(y_1 - y_0) \end{bmatrix}.$$

Because the formulas for x and y have such similar structure, we can use the vector form for $\mathbf{p} = (x, y)$ (Figure 2.36):

$$\mathbf{p}(t) = \mathbf{p}_0 + t(\mathbf{p}_1 - \mathbf{p}_0).$$

You can read this in geometric form as "start at point $\mathbf{p}_0$ and go some distance toward $\mathbf{p}_1$ determined by the parameter t." A nice feature of this form is that $\mathbf{p}(0) = \mathbf{p}_0$ and $\mathbf{p}(1) = \mathbf{p}_1$. Since the point changes linearly with t, the value of t between $\mathbf{p}_0$ and $\mathbf{p}_1$ measures the fractional distance between the points. Points with $t < 0$ are to the "far" side of $\mathbf{p}_0$, and points with $t > 1$ are to the "far" side of $\mathbf{p}_1$.

Parametric lines can also be described as just a point $\mathbf{o}$ and a vector $\mathbf{d}$:

$$\mathbf{p}(t) = \mathbf{o} + t(\mathbf{d}).$$

When the vector $\mathbf{d}$ has unit length, the line is *arc-length parameterized*. This means t is an exact measure of distance along the line. Any parametric curve can be arc-length parameterized, which is obviously a very convenient form, but not all can be converted analytically.

2D Parametric Circles

A circle with center (x_c, y_c) and radius r has a parametric form:

$$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} x_c + r\cos\phi \\ y_c + r\sin\phi \end{bmatrix}.$$

To ensure that there is a unique parameter ϕ for every point on the curve, we can restrict its domain: $\phi \in [0, 2\pi)$ or $\phi \in (-\pi, \pi]$ or any other half-open interval of length 2π.

An axis-aligned ellipse can be constructed by scaling the x and y parametric equations separately:

$$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} x_c + a\cos\phi \\ y_c + b\sin\phi \end{bmatrix}.$$

2.7.7 3D Parametric Curves

A 3D parametric curve operates much like a 2D parametric curve:

$$\begin{aligned} x &= f(t), \\ y &= g(t), \\ z &= h(t). \end{aligned}$$

For example, a spiral around the z-axis is written as

$$\begin{aligned} x &= \cos t, \\ y &= \sin t, \\ z &= t. \end{aligned}$$

As with 2D curves, the functions f, g, and h are defined on a domain $D \subset \mathbb{R}$ if we want to control where the curve starts and ends. In vector form, we can write

$$\begin{bmatrix} x \\ y \\ z \end{bmatrix} = \mathbf{p}(t).$$

The parametric curve is the range of $\mathbf{p}$: $\mathbb{R} \to \mathbb{R}^3$.

In this chapter, we only discuss 3D parametric lines in detail. General 3D parametric curves are discussed more extensively in Chapter 15.

3D Parametric Lines

A 3D parametric line can be written as a straightforward extension of the 2D parametric line, e.g.,

$$\begin{aligned} x &= 2 + 7t, \\ y &= 1 + 2t, \\ z &= 3 - 5t. \end{aligned}$$

This is cumbersome and does not translate well to code variables, so we will write it in vector form:

$$\mathbf{p} = \mathbf{o} + t\mathbf{d},$$

where, for this example, $\mathbf{o}$ and $\mathbf{d}$ are given by

$$\begin{aligned} \mathbf{o} &= (2, 1, \quad 3), \\ \mathbf{d} &= (7, 2, -5). \end{aligned}$$

Note that this is very similar to the 2D case. The way to visualize this is to imagine that the line passes through $\mathbf{o}$ and is parallel to $\mathbf{d}$. Given any value of t, you get some point $\mathbf{p}(t)$ on the line. For example, at $t = 2$, $p(t) = (2, 1, 3) + 2(7, 2, -5) = (16, 5, -7)$. This general concept is the same as for two dimensions (Figure 2.36).

As in 2D, a *line segment* can be described by a 3D parametric line and an interval $t \in [t_a, t_b]$. The line segment between two points $\mathbf{a}$ and $\mathbf{b}$ is given by $\mathbf{p}(t) = \mathbf{a} + t(\mathbf{b} - \mathbf{a})$ with $t \in [0, 1]$. Here, $\mathbf{p}(0) = \mathbf{a}$, $\mathbf{p}(1) = \mathbf{b}$, and $\mathbf{p}(0.5) = (\mathbf{a} + \mathbf{b})/2$, the midpoint between $\mathbf{a}$ and $\mathbf{b}$.

A *ray*, or *half-line*, is a 3D parametric line with a half-open interval, usually $[0, \infty)$. From now on, we will refer to all lines, line segments, and rays as "rays." This is sloppy, but corresponds to common usage and makes the discussion simpler.

2.7.8 3D Parametric Surfaces

The parametric approach can be used to define surfaces in 3D space in much the same way we define curves, except that there are two parameters to address the two-dimensional area of the surface. These surfaces have the form

$$\begin{aligned} x &= f(u, v), \\ y &= g(u, v), \\ z &= h(u, v). \end{aligned}$$

or, in vector form,

$$\begin{bmatrix} x \\ y \\ z \end{bmatrix} = \mathbf{p}(u, v).$$

The parametric surface is the range of the function $\mathbf{p}$: $\mathbb{R}^2 \to \mathbb{R}^3$.

Example 1 For example, a point on the surface of the Earth can be described by the two parameters longitude and latitude. If we define the origin to be at the center of the Earth, and let r be the radius of the Earth, then a spherical coordinate system centered at the origin (Figure 2.37), lets us derive the parametric equations

$$\begin{aligned} x &= r \cos\phi \sin\theta, \\ y &= r \sin\phi \sin\theta, \\ z &= r \cos\theta. \end{aligned} \tag{2.24}$$

Pretend for the sake of argument that the Earth is exactly spherical.

Ideally, we'd like to write this in vector form, but it isn't feasible for this particular parametric form.

We would also like to be able to find the (θ, ϕ) for a given (x, y, z). If we assume that $\phi \in (-\pi, \pi]$, this is easy to do using the *atan2* function from Equation (2.2):

$$\begin{aligned} \theta &= \operatorname{acos}(z/\sqrt{x^2 + y^2 + z^2}), \\ \phi &= \operatorname{atan2}(y, x). \end{aligned} \tag{2.25}$$

The θ and ϕ here may or may not seem reversed depending on your background; the use of these symbols varies across disciplines. In this book we will always assume the meaning of θ and ϕ used in Equation (2.24) and depicted in Figure 2.37.

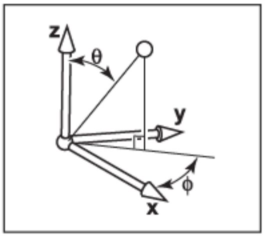

Figure 2.37. The geometry for spherical coordinates.

With implicit surfaces, the derivative of the function f gave us the surface normal. With parametric surfaces, the derivatives of $\mathbf{p}$ also give information about the surface geometry.

Consider the function $\mathbf{q}(t) = \mathbf{p}(t, v_0)$. This function defines a parametric curve obtained by varying u while holding v fixed at the value v_0. This curve, called an *isoparametric curve* (or sometimes "isoparm" for short), lies in the surface. The derivative of $\mathbf{q}$ gives a vector tangent to the curve, and since the curve lies in the surface, the vector $\mathbf{q}'$ also lies in the surface. Since it was obtained by varying one argument of $\mathbf{p}$, the vector $\mathbf{q}'$ is the partial derivative of $\mathbf{p}$ with respect to u, which we'll denote $\mathbf{p}_u$. A similar argument shows that the partial derivative $\mathbf{p}_v$ gives the tangent to the isoparametric curves for constant u, which is a second tangent vector to the surface.

The derivative of $\mathbf{p}$, then, gives two tangent vectors at any point on the surface. The normal to the surface may be found by taking the cross product of these vectors: since both are tangent to the surface, their cross product, which is perpendicular to both tangents, is normal to the surface. The right-hand rule for cross products provides a way to decide which side is the front, or outside, of the surface; we will use the convention that the vector

$$\mathbf{n} = \mathbf{p}_u \times \mathbf{p}_v$$

points toward the outside of the surface.

2.7.9 Summary of Curves and Surfaces

Implicit curves in 2D or surfaces in 3D are defined by scalar-valued functions of two or three variables, $f : \mathbb{R}^2 \to \mathbb{R}$ or $f : \mathbb{R}^3 \to \mathbb{R}$, and the surface consists of all points where the function is zero:

$$S = \{\mathbf{p} \,|\, f(\mathbf{p}) = 0\}.$$

Parametric curves in 2D or 3D are defined by vector-valued functions of one variable, $\mathbf{p} : D \subset \mathbb{R} \to \mathbb{R}^2$ or $\mathbf{p} : D \subset \mathbb{R} \to \mathbb{R}^3$, and the curve is swept out as t varies over all of D:

$$S = \{\mathbf{p}(t) \,|\, t \in D\}.$$

Parametric surfaces in 3D are defined by vector-valued functions of two variables, $\mathbf{p} : D \subset \mathbb{R}^2 \to \mathbb{R}^3$, and the surface consists of the images of all points (u, v) in the domain:

$$S = \{\mathbf{p}(u, v) \,|\, (u, v) \in D\}.$$

For implicit curves and surfaces, the normal vector is given by the derivative of f (the gradient), and the tangent vector (for a curve) or vectors (for a surface) can be derived from the normal by constructing a basis.

For parametric curves and surfaces, the derivative of $\mathbf{p}$ gives the tangent vector (for a curve) or vectors (for a surface), and the normal vector can be derived from the tangents by constructing a basis.