Chapter 2

Curves, Lengths, Surfaces and Areas

Last revised 13 Oct 2010.

This chapter gives an overview of various methods for describing curves in 2-dimensional and 3-dimensional space, including parametrised curves, curves in polar coordinates, and some vector notation.

We then move on to formulae involving integration for the arc-length of curves in each of these cases, and for the area enclosed by curves in 2-dimensional polar coordinates.

Finally we extend this to defining parametrised surfaces in 3 dimensions (using 2 parameters), and the areas of these surfaces.

2.1 Parametrised curves

2.1.1 Parametrised curves: definition

(See Thomas 3.5)

By now you are familiar with expressing a curve on a 2-dimensional plane in Cartesian coordinates (x,y) as

$$y = f(x)\ ; \tag{2.1}$$

where $f(x)$ is a given function, and may be any combination of polynomials, trigonometric functions, exponential functions (called "elementary functions"), or more complicated functions. This form for a curve is called "explicit form" since a given f specifies exactly how to calculate y for any value of x. Clearly for a given function $f(x)$ we can draw a graph of this function by taking many values of x with suitably small steps, evaluating $f(x)$ at each of these so we have a "dot" at $(x,y = f(x))$ and then "joining the dots". If f is a continuous function, then there are no "gaps" in this curve.

This is straightforward, but we have the limitation that for each x the curve has a unique value of y (the converse is not true, i.e. choosing a particular value y_0 for y and solving the equation $f(x) = y_0$ may give none, one or many solutions known as "roots" for x) . So, a curve such as $y = f(x)$ can have "wiggles" in the y-direction but not in the x-direction.

In a few special cases we can have multiple values for y at a given x, e.g. for the familar case of the circle $x^2+y^2=a^2$ we can write $y=\pm\sqrt{a^2-x^2}$, and the $\pm$ term gives 2, 1 or 0 solutions for y depending whether $|x|<a$, $x=a$, or $|x|>|a|$ respectively ; but this quickly becomes excessively complicated for more general curves.

A second way to represent a curve in a plane is as the set of all points satisfying an equation such as $f(x,y)=0$ or $f(x,y)=c$, where c is a constant; here f depends on both x and y and may not be separable; this is called "implicit form". This has some advantages we will see later, for example choosing different values of c can give us a "family" of different curves from one function f; however, a clear disadvantage is that there is no easy way (in general) to calculate y at a given $x=x_0$; so sketching the curve (or programming a computer to sketch it) can be cumbersome, unless we can recognise the form of the solution from experience.

A third way to represent a curve in a 2-dimensional plane is via **parametrisation**: now we define **two** arbitrary functions $f(t),g(t)$ of a new real variable t, and we define our curve called C as the set of all points where

$$x=f(t),\quad y=g(t)\qquad \text{hence}\quad (x,y)=(f(t),g(t)) \tag{2.2}$$

for any value of t in a given **domain** (which may be finite or infinite). Here we call C the **parametric curve**, t is the **parameter**, and $x=f(t),y=g(t)$ are the **parametric equations** for the curve. These equations together with the defined domain of t constitute a complete definition called the **parametrisation** of the curve.

Given the above, it is clear that any value of t maps to a single point in the (x,y) plane; it is also fairly obvious that if the functions f,g are both continuous, then the resulting curve C is also continuous. (To prove this, pick a value t_0 giving a point on the curve x_0,y_0; then draw a tiny circle radius δ around x_0,y_0. If f,g are both continuous we can find some range of $t\pm\varepsilon$ for which the curve is contained inside the above circle, i.e. the curve has no "gaps"; if there were a finite gap in the curve, then either f or g must not be continuous, contradicting our assumption).

Using parametric form, we can express more complicated curves such as figure-eights, spirals and so on which can self-intersect and/or cross a given x value many times (including infinitely many), and we can "sketch the curve" by hand or by computer by just evaluating $f(t),g(t)$ at a sufficient number of points spaced in t and "joining the dots".

Note that here t is not necessarily "time", t is just a "label" so that each point on the curve is "labelled" with one value of t, or multiple values if the curve crosses itself at that point.

Clearly if we are given a curve $y=g(x)$, we can put that into parametric form by simply defining $f(t)=t$ in the above, so then $x=t$ and $y=g(t)=g(x)$; but the converse generally is not true, so the parametric form is more general.

Now for a few simple examples: a very simple example is a straight line, which is given by

$$x=x_0+at,\qquad y=y_0+bt\quad ; \tag{2.3}$$

and the domain $-\infty<t<\infty$. Here it is easy to see that this parametrises a straight line passing through the point (x_0,y_0) with direction vector (a,b) and slope b/a. (If $a\neq 0$, we can rearrange the $x-$equation to $t=(x-x_0)/a$, and then substitute that into the $y-$ equation to get $y=y_0+(b/a)(x-x_0)$.)

Note that many possible choices of x_0,y_0,a,b lead to the same straight line, only the mapping from t onto points on the line will change. If we want a line through (x_0,y_0) and (x_1,y_1) then we set $a=x_1-x_0,\ b=y_1-y_0$ in the above, and if we want our "curve" to be a finite straight line segment with endpoints (x_0,y_0) and (x_1,y_1), then we just specify that the domain of t is $0\le t\le 1$ above.

Another simple case is given by

$$x=a\cos t, y=a\sin t\quad ;$$

it is clear that this obeys $x^2+y^2=a^2$, so the curve C is a circle of radius a centred at the origin. Since the functions sin and cos are both periodic with period 2π, adding $n\times 2\pi$ to t (for any integer n) gets back to the same x,y. So if we let t run from $-\infty$ to $+\infty$, the resulting curve loops around the circle an infinite number of times. So, in this case it is more convenient to specify a finite domain for t, such as $0\le t<2\pi$, so over that range the curve goes round the circle exactly once. (Here we can choose any interval of length 2π, so e.g. $-\pi<t\le\pi$ works just as well).

We can generalise this to an ellipse by $x=a\cos t, y=b\sin t$; this is just a circle stretched by a factor b/a in the $y-$direction, so the semi-axes are a and b respectively. We can get an ellipse with mid-point at (x_0,y_0) with by $x=x_0+a\cos t, y=y_0+b\sin t$.

The above illustrates a convenient property: because if we know one parametric curve, we can produce a shifted copy of it just by adding x_0 and y_0 to the two functions; or we can stretch or squash it along the axes by multiplying our two functions by constants.

2.1.2 The cycloid

Another example of a curve which is easy to represent in parametric form is the **cycloid**, which can be expressed as

$$x=a(t-\sin t)\ , \qquad y=a(1-\cos t) \tag{2.4}$$

If we didn't have the at term in the $x-$equation above, it is easy to see this would be a circle of radius a centred at $(0,a)$; but the additional at term makes the circle's centre "roll along" in the x direction as t increases. It turns out that the above curve is the curve traced out by (for example) a pebble stuck to a bicycle's tyre as the tyre rolls along the ground without slipping, so we get a combination of the "axle" going along at constant rate and the pebble going in a circle round the moving axle. In the example above we have chosen things so the "ground" is the x-axis, the axle goes along the line $y=a$ and the point is at the origin at $t=0$.

This curve has applications in several real-world problems, and you can see above that it is quite simple to write in parametric form, but it is complicated in Cartesian coordinates (there is an expression in elementary functions for x in terms of y, but not the other way round).

There are generalised versions of this curve called the epicycloid and hypocycloid which are traced by a point on one circle rolling around a second circle (instead of along a straight line), and furthermore there are versions where the "point" is not on the circumference of the rolling circle; these are called **trochoids**. (You won't be expected to memorise these, but you might be given the equations as part of an exam question so it is worth knowing the general concept).

2.1.3 Lissajous figures

A curve parametrised by $x=a\cos k_1 t,\quad y=b\sin k_2 t$ where k_1,k_2 are constants (usually integers), is called a **Lissajous figure**. By considering what happens as t varies, we can see that both x and y oscillate between $\pm a$ and $\pm b$, so the curve must always lie inside a rectangle with corners at $(\pm a,\pm b)$; but now the curve oscillates at different rates in the x,y directions, and it can cross itself many times. If we choose $k_1=1,k_2=2$ it will turn out that we get a figure-of-eight. If k_1/k_2 is a simple fraction, it will turn out that the curve closes back on itself after a finite number of "wiggles" ; but if k_1/k_2 is irrational it can be shown that the curve gets arbitrarily close to every point in the above rectangle but never returns to exactly the same place; this sort of thing may be seen in some computer screensavers, where you have an icon wandering around the rectangular computer screen and it's helpful for the pattern not to repeat itself.

2.1.4 Parametric curves in 3 dimensions

We can easily extend the above parametric curves from 2 to 3 dimensions by defining a third function $h(t)$ for the $z-$coordinate, so that

$$x = f(t),\ y = g(t),\ z = h(t). \tag{2.5}$$

Clearly for each value of t we now get a point in 3-dimensional space, and the set of points $(f(t), g(t), h(t))$ defines a 1-dimensional curve which is continuous if f, g, h are all continuous; the basic principles are the same as in 2 dimensions.

A good example of this is the **helix**, where

$$x = a\cos t\ , \qquad y = a\sin t\ , \qquad z = bt \tag{2.6}$$

where a, b are constants. Here as t varies, the distance of the curve from the z-axis is $\sqrt{x^2+y^2} = a$ (constant), so the curve projected onto the x,y plane is a circle, but the z-value is increasing at a uniform rate, so we get a curve in 3 dimensions looking like the handrail of a spiral staircase winding around the z-axis. Each increase of 2π in t gives us one full "twist" around the $z-$axis.
Note: In everyday English, this may be called a spiral: however in maths terminology, the term spiral refers to various types of 2-dimensional plane curve, while a 3-dimensional curve as described above is properly called a helix).

Parametric curves may be expressed more compactly in vector notation as $\mathbf{r} = \mathbf{r}(t)$, but of course we still need to define the 3 functions for the 3 independent components of $\mathbf{r}$, so this doesn't change any of the results, it just makes the expressions more compact.

As an example, the parametric representation also makes it quite easy to express curves which aren't symmetric about the x,y,z axes: for example, if we choose any two fixed orthogonal unit vectors $\mathbf{u}, \mathbf{v}$, we can construct an ellipse with centroid at $\mathbf{c}$, semi-major axis a and semi-minor axis b respectively parallel to the two vectors $\mathbf{u}, \mathbf{v}$, by:

$$\mathbf{r}(t) = \mathbf{c} + (a\cos t)\mathbf{u} + (b\sin t)\mathbf{v} \quad ; \tag{2.7}$$

we can of course plug in the components to write x,y,z in terms of t, but then it will be a lot less clear geometrically.

2.2 Arc Length of a curve

Here we show how to calculate the arc-length of a curve between two given endpoints.

If we choose a point on the curve $\mathbf{r}(t)$, and a neighbouring point $\mathbf{r}(t+\delta t)$, then the vector difference of these is

$$\mathbf{r}(t+\delta t) - \mathbf{r}(t) \approx \frac{d\mathbf{r}}{dt}\,\delta t \quad ; \tag{2.8}$$

this is the vector separation between the two nearby points on the curve. Taking limits where δt tends to zero, and assuming that the derivative exists, the curve tends to an infinitesimal straight line segment, so we can define the infinitesimal **length** ds to be the modulus of the left-hand side above,

$$ds = |\mathbf{r}(t+dt) - \mathbf{r}(t)| = \left|\frac{d\mathbf{r}}{dt}\right| dt \tag{2.9}$$

$$= \left|\left(\frac{df}{dt}, \frac{dg}{dt}, \frac{dh}{dt}\right)\right| dt \tag{2.10}$$

Therefore, we have

$$\frac{ds}{dt} = \sqrt{\left(\frac{df}{dt}\right)^2 + \left(\frac{dg}{dt}\right)^2 + \left(\frac{dh}{dt}\right)^2} \tag{2.11}$$

$$\text{or } \frac{ds}{dt} = \sqrt{\left(\frac{dx}{dt}\right)^2 + \left(\frac{dy}{dt}\right)^2 + \left(\frac{dz}{dt}\right)^2} \tag{2.12}$$

(This is effectively just Pythagoras's theorem applied to an infinitesimal segment of the curve, which has the same length as a straight line joining its endpoints).

So, now we can define the **arc length** L of the parametric curve between two values t_1, t_2 by integrating the above, giving us

$$L = \int_{t_1}^{t_2} \sqrt{\left(\frac{df}{dt}\right)^2 + \left(\frac{dg}{dt}\right)^2 + \left(\frac{dh}{dt}\right)^2}\, dt \tag{2.13}$$

To be clear, L is the length of a virtual "piece of string" which exactly follows the curve between endpoints given by t_1 and t_2 at points $(f(t_1), g(t_1), h(t_1))$ and $(f(t_2), g(t_2), h(t_2))$, if the string was then "pulled out straight". The above L is **not** the straight-line distance between the endpoints which would be just $|\mathbf{r}(t_2) - \mathbf{r}(t_1)|$. Equation 2.13 remains valid even if some or all of the derivatives cross zero, as long as none of them become infinite or undefined. If our curve is in 2 dimensions x, y we just set $z = h(t) = 0$ and $dz/dt = 0$.

Note: In problems, you may be given a parametric curve, and the endpoints specified in terms of (x_1, y_1, z_1) and (x_2, y_2, z_2); in this case you will need to solve to find the values of t_1 and t_2 corresponding to the endpoints, before doing the integral above. For each endpoint you can solve whichever of the x, y, z equations is simplest to get t_1, t_2; then insert those t_1, t_2 into the other two equations to check.

Example 2.1. The parametric curve C is given by $x = t, y = t^2, z = \frac{2}{3}t^3$. Evaluate the arc-length L of the curve between points (0,0,0) and $(2, 4, \frac{16}{3})$.

Answer: The end-points have values $t_1 = 0$ and $t_2 = 2$ (solve the x equation for t, and check the other two equations give the desired point); the derivatives are $dx/dt = 1$, $dy/dt = 2t$, $dz/dt = 2t^2$. Therefore the required length is

$$L = \int_{t=0}^{2} \sqrt{1 + (2t)^2 + (2t^2)^2}\, dt = \int_0^2 \sqrt{1 + 4t^2 + 4t^4}\, dt = \int_0^2 1 + 2t^2\, dt = [t + \tfrac{2}{3}t^3]_0^2 = \tfrac{22}{3} \quad .$$

Equation 2.13 can easily be simplified to give us the arc-length of a curve in implicit form: i.e. if we are given a 2-dimensional curve given as $y = g(x)$, we can just define $f(t) = t$ so $t = x$, $y = g(t) = g(x)$ and $z = h(t) \equiv 0$; inserting this gives us the arc-length for the curve $y = g(x)$ between the endpoints at $(x_1, g(x_1))$ and $(x_2, g(x_2))$ as

$$L = \int_{x_1}^{x_2} \sqrt{1 + \left(\frac{dy}{dx}\right)^2}\, dx \tag{2.14}$$

Similarly, if we have a curve in 3 dimensions where any two of the coordinates are given as functions of the other one, e.g. $y = g(x), z = h(x)$, then we get

$$L = \int_{x_1}^{x_2} \sqrt{1 + \left(\frac{dy}{dx}\right)^2 + \left(\frac{dz}{dx}\right)^2}\, dx \quad . \tag{2.15}$$

for L the arc-length between end-points at x_1 and x_2.

(However, note that many common curves have arc-length integrals which are not soluble in elementary functions; an ellipse is a well-known case where the circumference is not elementary, but is given by a special function called an *elliptic integral*. But one-dimensional integrals are generally very fast to evaluate accurately with a computer, since only one loop is needed).

2.2.1 Tangent vector to a curve

Given a parametric curve in either 2 or 3 dimensions, we can clearly differentiate each of the component functions with respect to t ; assuming the functions are differentiable, this gives us a new vector

$$\frac{d\mathbf{r}}{dt} = \lim_{\delta t \to 0} \frac{\mathbf{r}(t+\delta t) - \mathbf{r}(t)}{\delta t} = \left(\frac{df}{dt}, \frac{dg}{dt}, \frac{dh}{dt} \right) \quad . \tag{2.16}$$

It is easy to see that this vector $d\mathbf{r}/dt$ is locally parallel to the curve at the selected point, as long as all the derivatives exist ; hence given a value of $t = t_0$, we can evaluate $\mathbf{r}_0 = \mathbf{r}(t_0)$ (which is the position vector of a point on the curve), and define $\mathbf{Q} = d\mathbf{r}/dt|_{t_0}$ which is a vector of direction tangent to the curve at the same point, so $\mathbf{Q}$ is a **tangent vector** to the parametric curve at the point $\mathbf{r}_0$. Thus, we can construct an equation for the **tangent line**

$$\mathbf{r} = \mathbf{r_0} + u\mathbf{Q} \quad , \quad -\infty < u < \infty \tag{2.17}$$

for any real u; here u is another parameter (giving position along the tangent line to our curve C at $\mathbf{r}_0$). If we write this out in components, the above is three linear equations giving x, y, z as linear functions of u, and if desired we can rearrange those to give two linear equations for the tangent line: e.g. if we want y and z in terms of x, we just rearrange the x equation to give u in terms of x, and substitute that into the y, z equations.

Note: in the above, we must evaluate $\mathbf{r}$ and $d\mathbf{r}/dt$ at the same point i.e. the same value of t, otherwise the result will not make sense. Also, if you are given the coordinates of $\mathbf{r}_0$ rather than t, you will have to find the value of t which gives you $\mathbf{r}(t) = \mathbf{r}_0$; you can pick whichever coordinate is the simplest to solve.

(**Warning:** equation 2.16 looks a bit like the equation for ∇f which we met earlier. However, it's actually very different because $f(\mathbf{r})$ was a scalar function of three variables x, y, z, while $\mathbf{r}(t)$ along a parametric curve is a vector-valued function of one variable t.)

2.3 Curves in polar coordinates

As we saw in the previous section, it is sometimes convenient if we are working with circles, ellipses or other closed curves to work in **plane polar coordinates**; here instead of the familiar x, y of Cartesian coordinates, we can label any point P in a plane by its distance r from a fixed origin O, and an angle θ between the line OP and the positive $x-$axis. By convention, θ is defined to increase "anticlockwise" (from $+x$ towards $+y$) so the positive $y-$axis has $\theta = +\frac{\pi}{2}$, the negative $x-$axis has $\theta = \pi$, and the negative $y-$axis has $\theta = \frac{3\pi}{2}$.

The conversion from (r, θ) to x, y is given by simple trigonometry:

$$x = r\cos\theta \qquad y = r\sin\theta \quad , \tag{2.18}$$

which is clearly unique. Rearranging these to give r, θ in terms of x, y, we have

$$r = \pm\sqrt{x^2 + y^2} \qquad \theta = \arctan(y/x) + (n\pi) \tag{2.19}$$

for some integer n. Note however that the conversion from (x, y) to (r, θ) is *non-unique*; one point in a plane has a unique (x, y) pair, and a particular pair (r, θ) map to a unique x, y; but one point (x, y) can be represented

by two possible values of r (positive and negative) and an infinite number of θ values differing by integer multiples of π i.e. different numbers of half-turns around the origin. So the point (r,θ) is actually the same point as $(-r,\,\theta+(2n+1)\pi)$ and $(+r,\,\theta+2n\pi)$.

In most cases of interest we will be taking the positive value of r and we'll take θ to lie in the interval $[0,2\pi]$, in which case the mapping is unique, but be aware of potential ambiguities with this.

Clearly one of the simplest curves in plane polar coordinates is

$$r = a \tag{2.20}$$

where a is a constant. Implicitly this also means $\theta =$ any value, so this is clearly a circle, centred on the origin, radius a. Likewise $\theta = b$ where b is a constant is a straight line through the origin at angle b.

More generally we can define curves as

$$r = f(\theta) \tag{2.21}$$

This is a convenient way to define certain types of curve; one example is

$$r = a + b\theta \tag{2.22}$$

which describes a spiral called an **Archimedes spiral**.

It is possible to express a straight line in polars, for example it is easy to show that the vertical line $x = b$ has the polar equation $r = b\sec\theta$. More generally for a line which has closest distance b to the origin at angle θ_0, we get $r = b\sec(\theta-\theta_0)$.

2.3.1 Conics in polar coordinates

A particularly useful family of curves in polar coordinates is given by

$$r(\theta) = \frac{\ell}{1+e\cos\theta} \tag{2.23}$$

where e is called the eccentricity and ℓ is the semi-latus rectum. It can be shown that this equation gives a **conic section** as we met before with the quadratic functions in x,y. Here $e = 0$ gives a circle (with radius ℓ), $0 < e < 1$ gives an ellipse, $e = 1$ gives a parabola and $e > 1$ gives a hyperbola, so by choice of e this expression can give any of the above conics ; and ℓ is just a scale factor giving the overall size.

Note: in the above representation Eq 2.23, the origin is one *focus* of the conic, the origin is not the centroid except if $e = 0$ (the circle). The form Eq. 2.23 is especially useful in astronomical orbit problems, since it will turn out that orbits of planets and comets around a central star have a solution of this type with the star at one focus (and nothing at the other focus). For ellipses, it is easy to show by plugging in $\theta = 0,\pi$ that the semi-major axis $a = \ell/(1-e^2)$. (Note we use ℓ rather than a in the above definition since ℓ is well-defined for all of the conics).

For any point given by eq 2.23, the distance to the origin is r and the distance d to a vertical line at $x = x_0$ is

$$d = x_0 - r\cos\theta = \frac{x_0 + x_0 e\cos\theta - \ell\cos\theta}{1+e\cos\theta} \qquad . \tag{2.24}$$

If we choose $x_0 = k = \ell/e$, this reduces to $d = r/e$ or $r = ed$, so our conic is the locus such that (distance from focus) = $e\times$ distance from the line $x = k$, called the **directrix** ; this works for any of the ellipse, parabola or hyperbola; though the directrix is "at infinity" for the circle $e = 0$.

Altogether there are (at least) four possible ways of defining the conic sections: one is the "plane slicing a cone" definition; the second is using the distances-from-foci properties; the third is via quadratic equations in Cartesian coordinates, and the fourth is as above. (It takes some straightforward but fairly long algebra to prove that all of these do actually end up with the same family of curves, which we won't repeat here (see e.g. Thomas Chapter 10)).

2.3.2 Arc length and area in plane polar coordinates

Given a curve in polar coordinates as $r = f(\theta)$, we can get the arc length in two ways: firstly we can put this into the parametric representation by $x = f(\theta)\cos\theta, y = f(\theta)\sin\theta$ where θ is the parameter (which behaves like t in the examples we saw before). If we differentiate the above, we have

$$\frac{dx}{d\theta} = -f(\theta)\sin\theta + \frac{df}{d\theta}\cos\theta \qquad \frac{dy}{d\theta} = f(\theta)\cos\theta + \frac{df}{d\theta}\sin\theta \tag{2.25}$$

Inserting these into equation 2.11 for the arc-length, we get

$$\frac{ds}{d\theta} = \sqrt{(f(\theta))^2 + \left(\frac{df}{d\theta}\right)^2}\,, \tag{2.26}$$

therefore the arc-length of the curve defined by $r = f(\theta)$ between endpoints given by $\theta = \theta_1$ and $\theta = \theta_2$ is

$$L = \int_{\theta_1}^{\theta_2} \sqrt{(f(\theta))^2 + \left(\frac{df}{d\theta}\right)^2}\, d\theta \tag{2.27}$$

(We can get the same result geometrically by drawing a segment of a curve from (r,θ) to $(r+\delta r, \theta+\delta\theta)$, also drawing the circular arc through (r,θ), and applying Pythagoras's theorem to the small triangle resulting).

Warning: the above is clearly different from Equation 2.14 which gave the arc-length for the case $y = g(x)$; comparing them, the second term looks the same, but the first term above is $f(\theta)^2$ instead of 1. The reason is that in polar coordinates, a small change of angle $\delta\theta$ shifts our point by a distance $r\,\delta\theta$ in the "circumferential" (around-the-origin) direction, **not** just $\delta\theta$. We will see a lot more of this sort of thing in later sections where we deal with 3-dimensional polar coordinates.

2.3.3 Area in polar coordinates

If we are given a curve $r = f(\theta)$, it is straightforward to evaluate the area of the sector bounded by two straight lines $\theta = a$, $\theta = b$ and the curve $r = f(\theta)$: by considering an interval from $(f(\theta_0),\theta_0)$ to $(f(\theta_0+\delta\theta),\theta_0+\delta\theta)$, if $\delta\theta$ is small this area approaches an isosceles triangle with long dimension $r = f(\theta)$ and width $r\,\delta\theta$, so the area is $\frac{1}{2}r^2\delta\theta$.

Thus, the area inside a curve defined in plane polar coordinates, between angles $\theta_1 \le \theta \le \theta_2$ is simply

$$A = \frac{1}{2}\int_{\theta_1}^{\theta_2} [r(\theta)]^2\, d\theta \qquad . \tag{2.28}$$

Warning: You need to beware of zero-crossings here: if $r(\theta)$ goes negative so the curve has several "petals", you need to be careful not to count the same petal twice at angles π apart. If this happens, it's advisable to sketch the curve, break the integral into suitable chunks where r does not cross zero, and add these up. If r is always non-negative, there are no problems.

Example 2.2. An easy example is the circle $r = a$: inserting $\theta_1 = 0, \theta_2 = 2\pi$ for the endpoints (as explained earlier) gives us

$$A = \tfrac{1}{2}\int_0^{2\pi} a^2 d\theta = \pi a^2. \tag{2.29}$$

giving us the familiar formula for the area of a circle.

Example 2.3. Another example is the **cardioid** defined by $r = a(1+\cos\theta)$; this curve has real-world applications since many microphones and radio antennas have a directional response given by a cardioid function. Inserting this into 2.28, and using the double-angle formula we easily get the area $\frac{3}{2}\pi a^2$.

2.4 Surfaces in 3 dimensions

In the previous section we looked at parametric representations of a curve in 2 or 3 dimensions, generally written as $(x,y,z) = (f(t),g(t),h(t))$; there we had 2 or 3 functions (one per coordinate) of **one** parameter t.

A further generalisation is to define a **surface** in 3-dimensional space; a plane is the simplest example, but in general we will deal with curved surfaces. We will see that it requires **two** parameters to describe an arbitrary surface, instead of the one parameter we had for a curve.

Now we'll call the parameters u and v, and as before we need 3 functions of these to define the 3 x,y,z coordinates of our surface in 3-dimensional space, so we get a **parametrised surface** as

$$x = f(u,v),\ y = g(u,v),\ z = h(u,v) \quad \text{i.e.} \quad \mathbf{r} = \mathbf{r}(u,v)\ . \tag{2.30}$$

If we pick a fixed value for v, say $v = v_0$, and allow u to vary, then we just have a one-dimensional curve (on the surface) as we vary u. If we now choose $v = v_0 + \Delta v$ and vary u again, we get another curve which is "close" to the first one if the functions are continuous, and we have a "ribbon" of surface bounded by the two curves. Repeating for lots of v_0's we see that we sweep out a 2-dimensional surface (call it S) in 3-dimensional space, as long as the curves for $v = v_0$ and $v = v_0 + \Delta v$ don't coincide. Technically we can define the partial derivatives

$$\frac{\partial \mathbf{r}}{\partial u} = \left(\frac{\partial f}{\partial u}, \frac{\partial g}{\partial u}, \frac{\partial h}{\partial u}\right)\ , \qquad \frac{\partial \mathbf{r}}{\partial v} = \left(\frac{\partial f}{\partial v}, \frac{\partial g}{\partial v}, \frac{\partial h}{\partial v}\right) \tag{2.31}$$

and as long as these two vectors are not parallel at any point, our locus of $\mathbf{r}(u,v)$ will in fact be a surface, not a line. We can see that both of the above two vectors are directions tangent to the surface at $\mathbf{r}(u,v)$. We can also take the vector product of these two,

$$\mathbf{N} = \frac{\partial \mathbf{r}}{\partial u} \times \frac{\partial \mathbf{r}}{\partial v} \qquad . \tag{2.32}$$

This cross-product $\mathbf{N}$ will be non-zero if the two partial derivatives above are both non-zero and not parallel. Assuming $\mathbf{N}$ is non-zero, it must be a **normal vector** to the surface S, because both the partial derivatives are parallel to the tangent plane to our surface at the point $\mathbf{r}(u,v)$, and $\mathbf{N}$ is perpendicular to both of them.

Thus, if we are given a surface $\mathbf{r}(u,v)$, and given a point on the surface defined by values (u_0,v_0), we have a clear procedure for finding the tangent plane to the surface at the corresponding point: we first evaluate the point in the surface $\mathbf{r}_0 = \mathbf{r}(u_0,v_0)$; next we evaluate the two partial derivatives at the same point, and take their cross-product $\mathbf{N}$ as above; thus in the usual vector notation for a plane through a given point normal to a given vector, the equation for the tangent plane is $(\mathbf{r} - \mathbf{r}_0).\mathbf{N} = 0$.

(Note: if instead we are given a point in the surface by defining its (x,y,z) space coordinates, we will first have to find the values of (u_0,v_0) which map onto that point *before* we evaluate the partial derivatives; in general that may not be simple to do, but it usually will be in the case of test questions.)

We can also use vectors to calculate the area of a parametric surface given by $\mathbf{r}(u,v)$: if we take the four points $\mathbf{r}(u,v)$, $\mathbf{r}(u+du,v)$, $\mathbf{r}(u,v+dv)$, $\mathbf{r}(u+du,v+dv)$, these define an infinitesimal parallelogram with sides $\frac{\partial \mathbf{r}}{\partial u}du$ and $\frac{\partial \mathbf{r}}{\partial v}dv$; as we saw from the definition of the vector product in Chapter 1, the area dA of this parallelogram is

$$dA = \left|\frac{\partial \mathbf{r}}{\partial u} \times \frac{\partial \mathbf{r}}{\partial v}\right| \, du\, dv \tag{2.33}$$

Thus, integrating the above with respect to both of u,v we get the **surface area** A of our parametric surface as

$$A = \int\int_D \left|\frac{\partial \mathbf{r}}{\partial u} \times \frac{\partial \mathbf{r}}{\partial v}\right| \, du\, dv \tag{2.34}$$

where the domain D of integration is the appropriate domain of u,v. (Note: if the surface is described geometrically, we will need to work out limits on u,v to cover the described surface).

In the special case where our surface is given as $z = h(x,y)$, we can just substitute $x = u$, $y = v$ into the above: then the two partial derivative vectors become $(1,0,\partial h/\partial x)$ and $(0,1,\partial h/\partial y)$, and the surface area becomes

$$A = \int\int_D \sqrt{\left(\frac{\partial h}{\partial x}\right)^2 + \left(\frac{\partial h}{\partial y}\right)^2 + 1}\, dx\, dy \tag{2.35}$$

where the integral is over some given domain in x,y.

Example 2.4. A good example of the above is the area of a sphere: a parametrisation of a sphere in 2 parameters (θ,ϕ) is

$$x = a\sin\theta\cos\phi, \quad y = a\sin\theta\sin\phi, \quad z = a\cos\theta \qquad 0 < \theta < \pi,\ 0 < \phi < 2\pi \tag{2.36}$$

(It is easy to show that this satisfies $x^2+y^2+z^2 = a^2$, so any (x,y,z) above does lie on the sphere. I'll state without proof that the limits given above define a unique mapping from a point on the sphere to θ,ϕ. We will meet this again later when we come to spherical polar coordinates).

Given this, evaluating the partial derivatives, we have $\partial\mathbf{r}/\partial\theta = (a\cos\theta\cos\phi, a\cos\theta\sin\phi, a\sin\theta)$, and $\partial\mathbf{r}/\partial\phi = (-a\sin\theta\sin\phi, a\sin\theta\cos\phi, 0)$. The cross product of these vectors is
$\mathbf{N} = (a^2\sin^2\theta\cos\phi, a^2\sin^2\theta\sin\phi, a^2\sin\theta\cos\theta)$ which is $a\sin\theta\mathbf{r}$, and has magnitude $a^2\sin\theta$. Then our surface area A of the sphere becomes

$$\begin{aligned} A &= \int_0^{2\pi}\left(\int_0^{\pi} a^2\sin\theta\, d\theta\right) d\phi \\ &= \int_0^{2\pi} [-a^2\cos\theta]_0^{\pi}\, d\phi \\ &= \int_0^{2\pi} 2a^2\, d\phi \\ &= 4\pi a^2 \qquad . \end{aligned}$$

2.4.1 Parametric forms of common surfaces

To conclude this chapter, I'll give some specific examples of parametric forms for common surfaces. These will turn out to be useful later, when we come to evaluate integrals over specified 2D surfaces in 3D space: the parametric form is usually the easiest way to do this.

A plane in 3D can be expressed in parametric form as

$$\mathbf{r}(u,v) = \mathbf{r}_0 + u\mathbf{a} + v\mathbf{b}$$

where $\mathbf{r}_0$ is a point in the plane, and $\mathbf{a}$, $\mathbf{b}$ are any two vectors parallel to the plane. Here if we take $-\infty < u,v < \infty$ we get the whole infinite plane.

If we want a finite parallelogram with one corner at $\mathbf{r}_0$ and two adjacent sides $\mathbf{a}$, $\mathbf{b}$, we can simply put limits $0 \le u \le 1,\ 0 \le v \le 1$ in the above. (A rectangle is a special case of this).

Finally, I'll repeat the parametric forms for a cylinder, sphere, ellipsoid and hyperboloid which we met briefly in Chapter 1.

	Implicit form	Parametric form		
Cylinder	$x^2+y^2=a^2$	$(x,y,z) = (a\cos\theta, a\sin\theta, z)$	Parameters θ, z	(2.37)
Sphere	$x^2+y^2+z^2=a^2$	$(a\sin\theta\cos\phi, a\sin\theta\sin\phi, a\cos\theta)$	Parameters θ, ϕ	(2.38)
Ellipsoid	$\frac{x^2}{a^2}+\frac{y^2}{b^2}+\frac{z^2}{c^2}=+1$	$(a\sin\theta\cos\phi, b\sin\theta\sin\phi, c\cos\theta)$	Parameters θ, ϕ	(2.39)
Hyperboloid	$\frac{x^2}{a^2}+\frac{y^2}{b^2}-\frac{z^2}{c^2}=1$	$(a\cos u, b\sin u\cosh v, c\sin u\sinh v)$	Parameters u, v	(2.40)

In the above, the choice of the names for the two parameters is slightly arbitrary, but follows common conventions.

It is easy to show that each parametric form above satisfies the implicit-form equation, just by substituting and using $\sin^2 A + \cos^2 A = 1$ or $\cosh^2 A - \sinh^2 A = 1$. It's not so obvious to see how to go the other way; but the parametric form for the sphere falls out naturally when we come to spherical polar coordinates; the ellipsoid is a simple "stretch" of the sphere; and the hyperboloid comes from replacing sin with sinh etc in the ellipsoid.