(MTH5109) GEOMETRY II: KNOTS AND SURFACES
LECTURE NOTES

DR. ARICK SHAO

1. Introduction to Curves and Surfaces

In this module, we are interested in studying the geometry of objects. According to our favourite source *Wikipedia* [9], geometry is the branch of mathematics concerned with studying the "shapes", "sizes", and "positions" of objects. Throughout this term, you will develop a better understanding of these concepts, as well as explore how these can be quantified and computed.

In your mathematical education up to this point—for instance, in calculus and linear algebra—you have considered flat, linear spaces such as the real line ($\mathbb{R}$), the plane ($\mathbb{R}^2$), and more general n-dimensional spaces ($\mathbb{R}^n$). This module will expand your outlook to "curved" objects.

1.1. Some Burning Questions.

Before launching into more involved mathematical discussions, we first address some common questions that you may have regarding this module, its contents and features, and how it fits into your overall maths education.

Question 1.1. *Why would we want to study curved objects?*

Curved objects are everywhere in our lives.

- If you throw a ball or shoot a missile into the air, then the trajectory of the ball or missile will not be linear, but rather a curve (due to gravity, for example).
- The surface of the earth is not flat, but rather like a sphere. To study this surface as a whole, we have to understand the effects of its curvature. This is important for many questions, such as determining the shortest flight path between two cities.
- According to Einstein's landmark theory of general relativity, the universe that we inhabit is not flat, but rather a 4-dimensional curved object ("spacetime"). Moreover, gravity itself is modelled by the shape and curvature of the spacetime.

These are only a small sample of motivations for having a firmer understanding of geometry.

However, since this is an elementary module, and since we have limited time, we will only discuss:

- 1-dimensional objects: curves.
- 2-dimensional objects: surfaces.

4-dimensional spacetimes and gravity will have to wait until another module; if you are interested in such things, you should consider the third-year module MTH6132: Relativity.

Question 1.2. *What maths will we use in this module? In other words, what should I know?*

This module is mainly concerned with the differential geometry of curves and surfaces. In particular, we look at objects that are "without jagged edges" and "vary smoothly enough" so that one can take derivatives, or linear approximations. In addition, to measure the sizes—e.g. length and area—of objects, we will need need to compute various integrals.

As a result, this module will assume that you have moderate familiarity with first-year calculus, for which differentiation and integration are two cornerstones.

- When studying 1-dimensional curves, we will make frequent use of single-variable calculus (mainly contents from MTH4100/4200: Calculus I).
- When studying 2-dimensional surfaces, we will require some knowledge of partial derivatives and double integrals (both of which you encountered in MTH4101/4201: Calculus II).

The simplest objects we can consider are lines and planes; these fall under the study of linear algebra. In particular, we will sometimes reference a bit of background on vectors and matrices:

- As we are working in one and two dimensions, the linear algebra background you need should have been covered in MTH4103/4203: Geometry I.
- Most of you are also currently learning this in more detail in MTH5112: Linear Algebra I.

Even for more complex curved objects, a useful tool in their analysis is linearisation—first determining and studying the linear object that best approximates it. Thus, we cannot really escape the need to understand linear algebra and its connections to geometry.

Finally, in order to construct common examples, we will make use of many elementary functions:

- Polynomials (e.g. $t^2 + 1$), and rational functions.
- Trigonometric functions (sin, cos).
- Exponential functions (exp) and logarithmic functions (ln).
- Rarely, we may encounter others, such as hyperbolic functions ($\sinh, \cosh$).

We will at times reference a few basic properties of these functions that you have learned before.

If you want to be optimally prepared for the material you will see in this module, it is recommended that you revise the material in calculus and linear algebra you have previously learned.

Question 1.3. *Help! What will I be expected to learn?*

The main focus of the module is on the interface between some mathematics you have already encountered—most notably, calculus and linear algebra—with concepts in geometry. For example, you will be expected to understand how notions such as derivatives, integrals, vectors, and matrices connect to the shapes and sizes of various objects you will encounter. This part of the knowledge is largely conceptual in nature; it is more about understanding the material in a critical way rather than memorising definitions, formulas, or algorithms.

As the module is also concerned with quantifying geometric properties, you will be expected to demonstrate that you are capable with performing various types of computations. Again, these computations will involve elements of calculus (e.g. derivatives and integrals) and linear algebra (e.g. vector and matrix algebra). In addition, you will be expected to graph various curves and surfaces, either in a plane or in space. While your hand-drawn figures need not be exquisite, you will need to produce convincing and largely accurate depictions of these objects.

Finally, though we will encounter a number of proofs in our discussions, they will not be a central focus of this module. In particular, you will not be asked to memorise and recite lengthy proofs of various results that you will encounter. On the other hand, you will be required to have a basic understanding of *why* these results are true.

1.2. **Some Informal Ideas.**

Here, we give a brief and informal introduction to the ideas we will explore and the questions we will answer in this module. This is to get you to begin engaging and thinking critically about basic concepts. More detailed discussions will be given in later chapters.

1.2.1. *Shape and Curvature.* Consider the three curves in the plane in Figure 1.1 below. You probably already have a very strong intuitive feeling that the curve C_1 is straight, while C_2 and C_3 are curved. But, how might you mathematically describe this?

Moreover, driving along a road, you feel a difference between a sharp turn along a very curved road and a smaller turn. Similarly, from the figure, you probably have a sense that C_3 is "more curved" than C_2. But, how would you capture and quantify this mathematically?

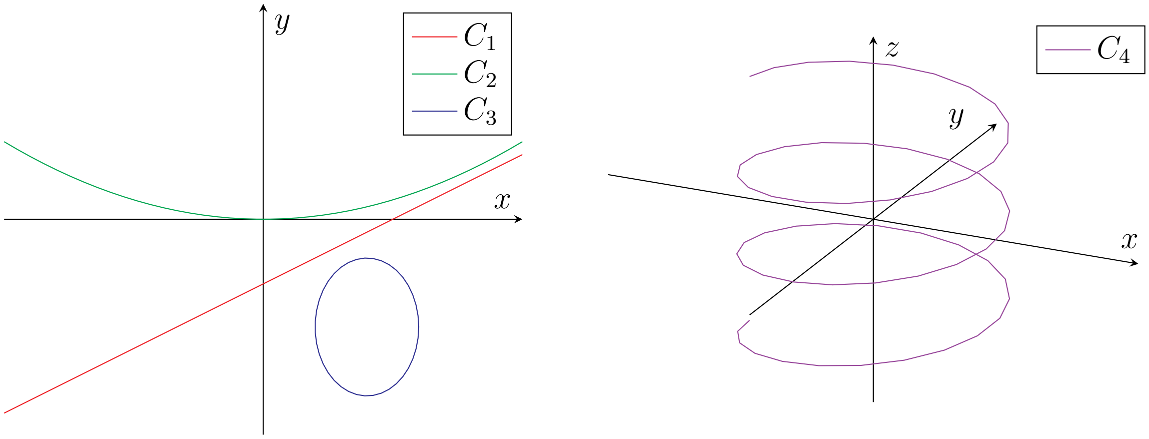


FIGURE 1.1. C_1, C_2, C_3 are curves on a plane, while C_4 is a curve in space.

Another question concerns the "direction" of curvature. For instance, continuing with the driving analogy, you can distinguish between turning left and turning right. Similarly, for C_2 and C_3, depending on which direction you traverse these curves, you would have an intuitive feeling of whether the curve is bending anticlockwise or clockwise. Moreover, if the curve is situated in 3-dimensional space (for example, see the helix C_4 in Figure 1.1), then there is an additional dimension of directions that the curve could bend. Again, the question of interest is how we can describe and quantify all of this mathematically.

Moving on to surfaces, the situation is similar but even more complicated. Consider, for instance, the surfaces in Figure 1.2. Again, you can distinguish that S_1 is flat, while S_2, S_3, and S_4 are curved. You can also tell when a surface is "very curved", as opposed to "slightly curved".

What is novel for these 2-dimensional surfaces, in contrast to 1-dimensional curves, is that a surface can bend in different ways along different directions. For instance, at any point of the spheres S_2 and S_3, the surfaces bend inward toward itself in the same way no matter which direction you go. However, for the "saddle" S_4, depending on which direction along the surface you look, it could be bending in opposite directions. The mathematical challenge here, then, is to somehow capture and quantify all this geometric information.

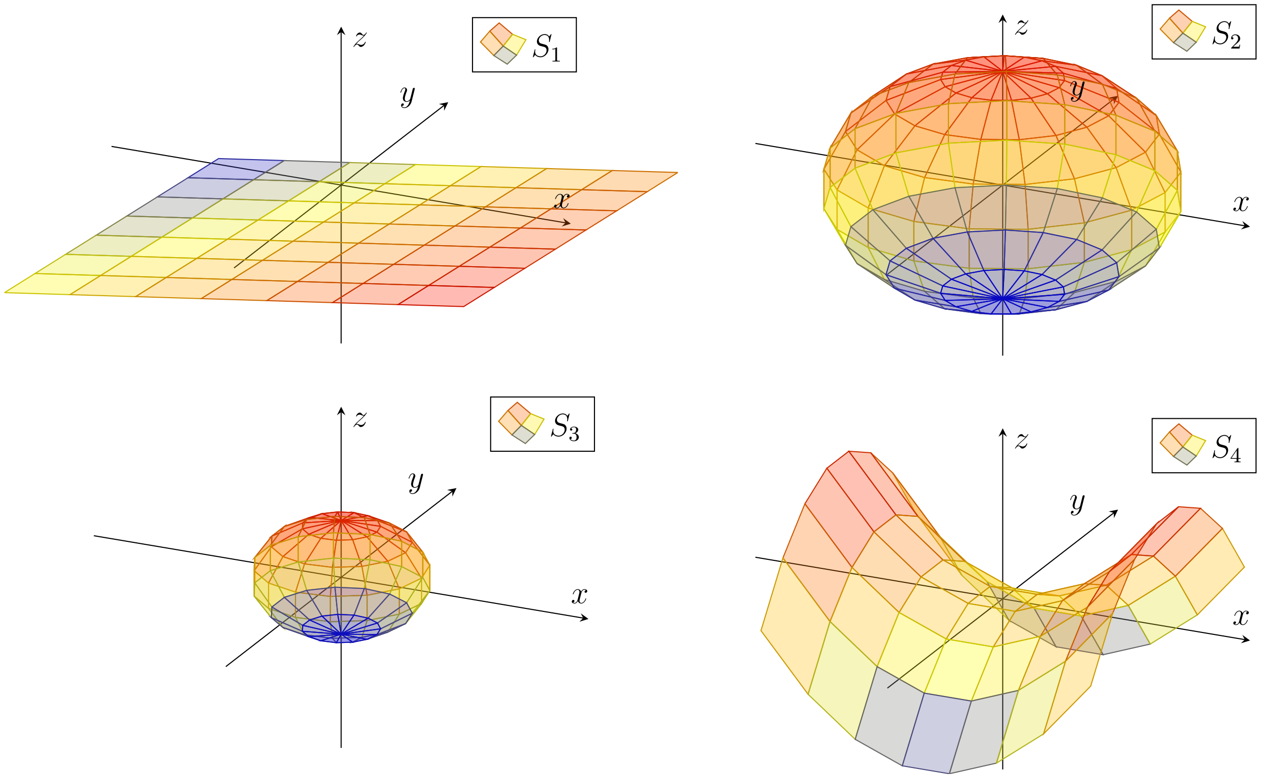


FIGURE 1.2. S_1, S_2, S_3, S_4 are surfaces in 3-dimensional space.

1.2.2. *Size, Length, and Area.* Another aspect of geometry is studying the sizes of objects. Consider the two circles in Figure 1.3. Your intuition already indicates that while they both have similar circular shape, they also have different sizes. You probably also have enough mathematical background to know that you can capture this by measuring their arc lengths (i.e., their circumferences).

There is an analogous sense of size for surfaces; for instance, the spheres S_2 and S_3 in Figure 1.2 have similar shapes but different sizes. This can be similarly captured by measuring their surface areas.

The mathematical goal, then, is to understand how we define and compute these arc lengths and surface areas. From calculus, you should know that lengths and areas are generally evaluated using integrals, and you should also know how to integrate along a line segment or a (flat) region in a plane. The questions of arc length and surface area will now force us to make sense of what it means to integrate along a curve or along a surface.

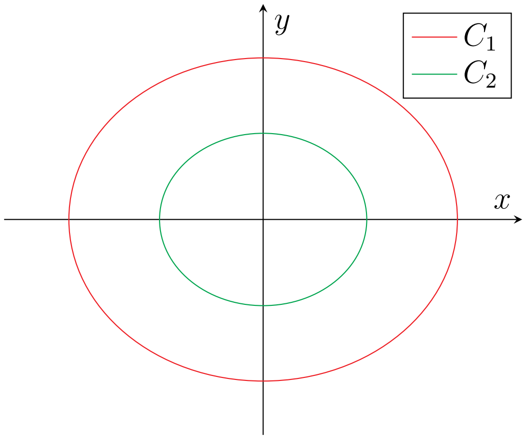


FIGURE 1.3. C_1 and C_2 are circles with different circumferences.

1.2.3. *Intrinsic and Extrinsic Geometry.* While we will not be particularly precise here, we will also touch upon the following classification of geometric properties:

- Extrinsic properties are those that depend on how an object is situated in a larger space.

- Intrinsic properties are of the object itself, regardless of how it is embedded in space.

For example, consider the unit circles in Figure 1.4. Clearly, these are different objects in terms of extrinsic geometric properties; after all, C_1 and C_2 lie in a plane, while C_3 lies in 3-dimensional space. Moreover, while C_1 and C_2 are in the same plane, they are situated at different locations. On the other hand, you probably have a sense that these three circles are somehow “the same”, regardless of what larger space they sit in or where in the space they sit. This intuition is behind the rough notion that the unit circles C_1, C_2, C_3 “have the same intrinsic geometry”.

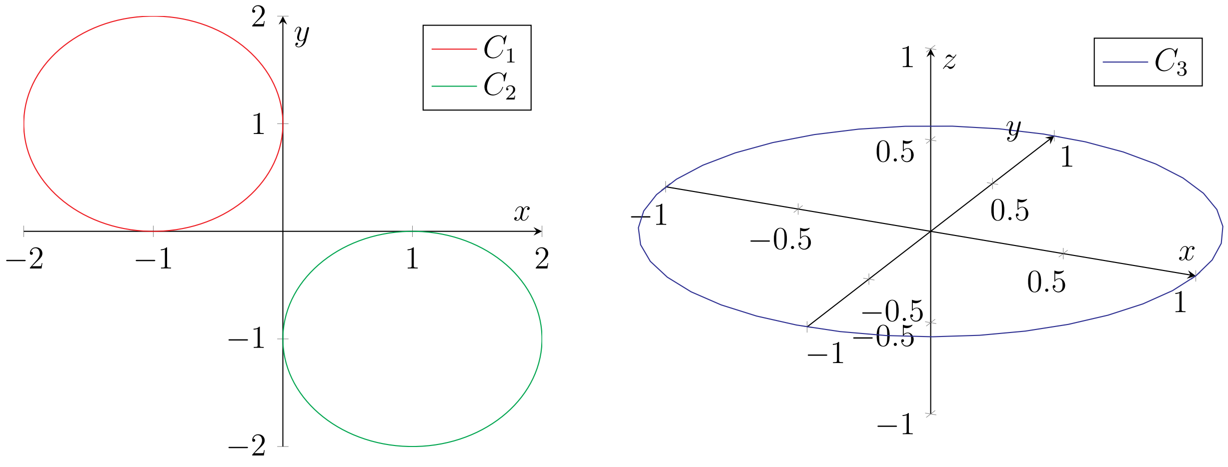


FIGURE 1.4. Three unit circles in different locations and settings.

One thought experiment you can try to further explore intrinsic geometry is to imagine if you are a bug living on one of these unit circles, with no knowledge of the larger dimensional space that the circle is in. As this imperceptive bug that knows only of the circle itself, you would not be able to distinguish between whether you were on C_1, C_2, or C_3.

Similarly, two copies of the sphere S_2 in Figure 1.2 in different positions would be extrinsically distinct, but also “the same” in a similar intrinsic sense.

For a more compelling example, consider the two curves in Figure 1.5. Clearly, they are extrinsically different, since they are embedded very differently in the plane. On the other hand, imagine again that you are the provincial bug. On either C_1 or C_2, you have exactly two options: move in one direction, or move in the other. That C_2 is bending and C_1 is not is purely a feature of how they are situated in the plane, and would not be evident to the bug that is unaware of the plane. As a result, *any infinite curve is intrinsically equivalent to a line!*

In contrast, consider a bug on a circle (e.g. C_1 in Figure 1.3) and a bug on a line (e.g. C_1 in Figure 1.5. A bug on

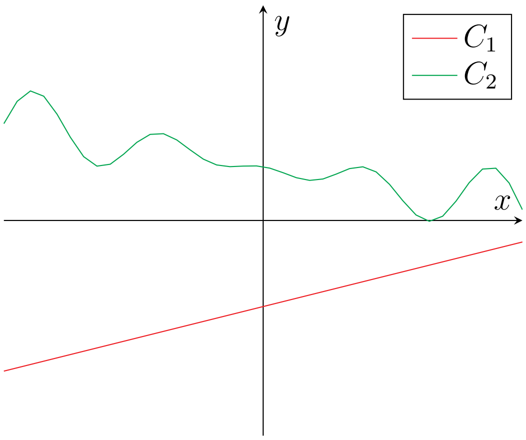


FIGURE 1.5. Two infinite curves that are extrinsically different but intrinsically the same.

the circle that walked in one direction long enough would return to where it started, whereas a bug on a line would not. Thus, one can argue that *a line and a circle are intrinsically different.*

Similarly, consider two circle with different radii, for instance, C_1 and C_2 in Figure 1.3. A bug would again be able to distinguish between them, since a bug on C_2 would need to travel less far to loop back to the starting point than a bug on C_1 would.

Exercise 1.1. *What about non-circular loops, such as that in Figure 1.6? What geometric properties of these loops are intrinsic? Can you classify all such loops based on their intrinsic geometry?*

For surfaces, the full story is far more complex, since the existence of an extra dimension allows for many more intrinsic properties. We will study some interesting ones later in the module. In contrast to curves, some aspects of how a surface is curved are in fact intrinsic!

Unfortunately, we will not have the time in this module to precisely define what we mean by "intrinsic" and "extrinsic". (These notions are formally captured by mathematical objects known as manifolds.) However, this discussion is meant to encourage you to think critically along these lines and to keep these notions in mind as we discuss various geometric properties in detail.

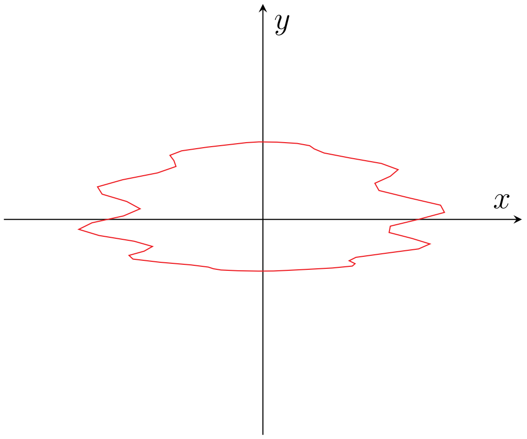


FIGURE 1.6. A non-circular loop.

1.2.4. *Topology.* Consider now the sphere S_2 in Figure 1.2 and the "egg" S_1 in Figure 1.7. Clearly, the sphere and the egg have different geometries (in fact, both extrinsically and intrinsically). On the other hand, there is an even weaker sense that these two surface are "the same". One way to think of this is that, if you visualise the objects as elastic, then you can stretch and compress the sphere in order to "deform" it into the egg.

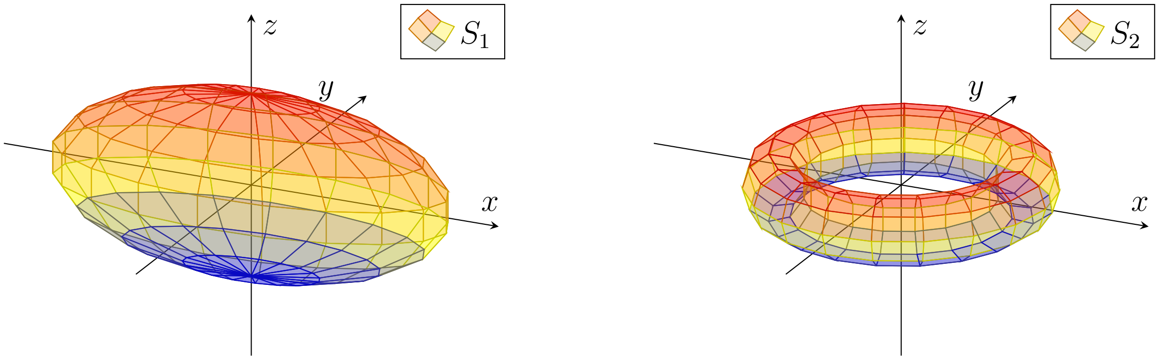


FIGURE 1.7. An ellipsoid (S_1) and a torus (S_2).

Consider in addition the "doughnut" S_2 in Figure 1.7. In this same sense of deformations, the doughnut is distinct from the sphere and the egg. Indeed, you cannot deform the sphere into the doughnut as you did into the egg; for this, you would have to be much more drastic by "poking your finger through the sphere" and puncturing a hole. However, you can "deform" this doughnut into the “coffee mug” on the left-hand side of Figure 1.8 (as clumsily indicated in that same figure). In this weaker sense, the doughnut and coffee mug are “the same”.

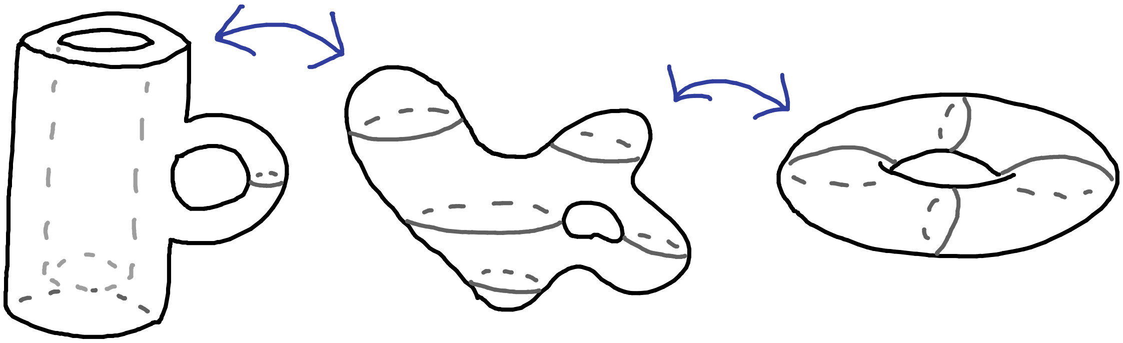


FIGURE 1.8. Deforming a coffee cup into a doughnut!

These intuitions are closely tied to an area of mathematics known as topology. While we will not have the time to formally delve into topology, we will in this module encounter some properties that are topological in nature. In other words, these properties do not depend on the shape or size of the object, but on other more basic features, such as how many holes a surface has.

One particular application of topology, in the one-dimensional setting, is to study and classify knots. In what ways can you tie a rope into a knot? When are two knots the same? Where topology enters is that intuitively, what matters here is “how the rope wraps around itself”, and not the exact shape of this wrapping. Time permitting, we will discuss some elementary aspects of this knot theory near the end of the module.

2. The Calculus of Curves

The first part of the module will focus on the differential geometry of curves. This chapter initiates this discussion by introducing some basic aspects of doing calculus on curves.

You probably already have a fairly solid intuition for what a curve should be. You may visualise a curve as a string tracing out some path. Or, perhaps you think of a curve as a trajectory traced out by a particle over time. As written, these intuitions are too vague for a careful mathematical study. Therefore, the first major question we will need to address is the following:

Question 2.1. *When we say "curve", what exactly do we mathematically mean?*

2.1. Parametric Curves.

Let us begin with the perspective of a trajectory of a moving particle. We consider a variable t, which we can think of as representing "time", and we let $\gamma(t)$ denote the position of the particle at a given time t. In other words, we view the trajectory as a function γ mapping a real number t to its position in $\mathbb{R}^2$ (a plane) or $\mathbb{R}^3$ (space). This is precisely the notion of parametric curves that you have probably encountered before:

Definition 2.1. *A parametric curve, or parametrisation, is a smooth function $\gamma : I \rightarrow \mathbb{R}^n$, where n is a positive integer, and where I is an open (finite or infinite) interval.*

By smooth, we mean that we can take as many *derivatives* of γ as we like (we will discuss differentiation in further detail below). In other words, γ contains no sudden instantaneous change of direction, speed, acceleration, and so on. We also note that the open interval I, which can be either finite or infinite, is allowed to be any of the forms

$$I = (a, b), \qquad I = (-\infty, b), \qquad I = (a, \infty), \qquad I = \mathbb{R}.$$

Moreover, since I is 1-dimensional, then the path traced out by γ is intuitively also "1-dimensional".

Also, if you wish to be more concrete, you can assume $n = 2$ (curve in a plane) or $n = 3$ (curve in space) in Definition 2.1. Conceptually, it really is not any different to consider curves embedded in higher dimensions ($n > 3$); it is just more difficult to visualise and draw.

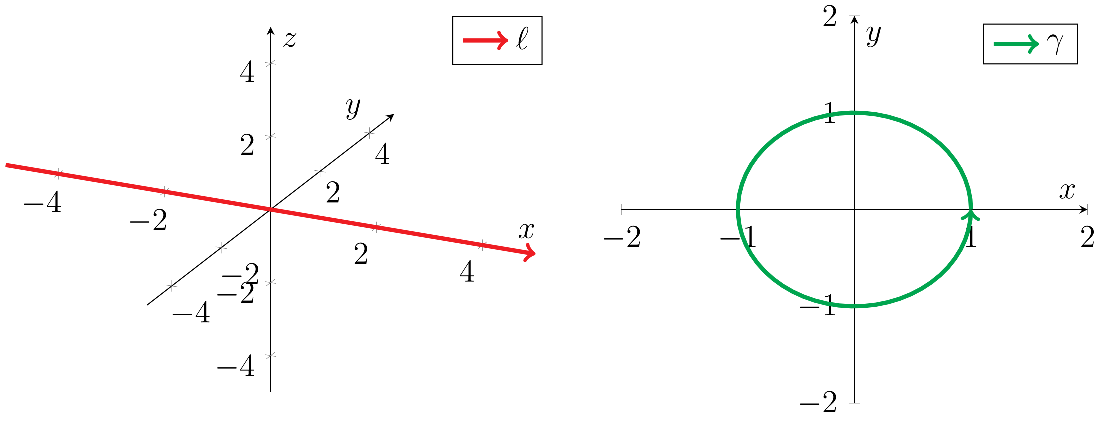


Figure 2.1. The parametric curves ℓ and γ from Example 2.1.

Example 2.1. *The following two functions are parametric curves:*

(1) x-*axis (line):*

$$\ell : \mathbb{R} \to \mathbb{R}^3, \qquad \ell(t) = (t, 0, 0).$$

(2) *Circle:*

$$\gamma : \mathbb{R} \to \mathbb{R}^2, \qquad \gamma(t) = (\cos t, \sin t).$$

Their graphs are given in Figure 2.1.

Now that we have our formal definition, we can begin to evaluate it critically: *do parametrised curves give an appropriate definition of curves for studying geometry?*

Example 2.2. *Consider the following two parametric curves:*

$$\gamma_1 : (0, \pi) \to \mathbb{R}^2, \qquad \gamma_1(t) = (\cos t, \sin t),$$
$$\gamma_2 : (-1, 1) \to \mathbb{R}^2, \qquad \gamma_2(t) = (-t, \sqrt{1 - t^2}).$$

Clearly γ_1, γ_2 are different functions, hence they represent different parametric curves. However, Figure 2.2 shows that they graph out exactly the same points, and in the same order.

If you think of γ_1, γ_2 as travelling particles, then this means that these two particles are traversing the same path at different speeds. From this perspective, it is sensible to think of γ_1 and γ_2 as distinct. This viewpoint would be relevant in various physics problems, where one investigates the motion in time of various particles.

On the other hand, here we are concerned with geometry, i.e. the shapes, sizes, and positions of objects. In terms of just these considerations, γ_1 and γ_2 are not any different at all, since both trace out the same points (and in the same order). Thus, the viewpoint of curves as particle trajectories falsely distinguishes two objects, γ_1 and γ_2, that should instead be "the same" for our geometric purposes.

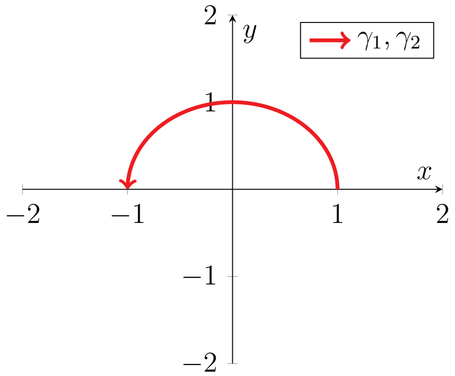


FIGURE 2.2. γ_1, γ_2 from Example 2.2 describe the same curve.

With these thoughts in mind, our current verdict on Question 2.1 is now summarised as follows:

- Any parametric curve, such as γ_1 or γ_2 in Example 2.2, defines a curve. On the other hand, we also want to view γ_1 and γ_2 as two different *parametrisations* of the *same* curve.
- More generally, any parametric curve describes a curve, while a single curve can be described by many different parametrisations.

Consequently, the idea is to characterise curves as parametric curves, with the caveat that two such parametric curves that trace out the same path are considered the same. Before we define this formally, however, we require some further background regarding differentiation.

2.1.1. *Derivatives.* Consider a parametric curve $\gamma : I \to \mathbb{R}^n$, which we again view as a position of particle over time, and fix a "time" $t_0 \in I$. To motivate the connection to differentiation, we ask, as one usually does in calculus, the following question:

Question 2.2. *What is the rate of change of γ at t_0?*

In physics, if γ represents the position of a particle over time, then its rate of change is called the velocity of the particle. As such, we often use the term “velocity” in place of “rate of change”.

Now, if $t \in I$ as well, then the vector $\gamma(t) - \gamma(t_0)$ can be viewed as the displacement, or the total change in position, between times t and t_0. This is visually represented in the first part of Figure 2.3 below; the green arrows indicate displacements over various time intervals. Moreover, $t - t_0$ is the total time elapsed during this displacement. In general, dividing the total change in position by the change in time yields the *rate of change* in position. Therefore, the quotient

$$\frac{\gamma(t) - \gamma(t_0)}{t - t_0}, \tag{2.1}$$

represents the *average* rate of change in γ (or average velocity of γ) over the time interval $[t_0, t]$.

Letting t become closer and closer to t_0, so that the elapsed time $t - t_0$ tends to 0, then (2.1) would become the *instantaneous* rate of change of γ at the single time t_0.

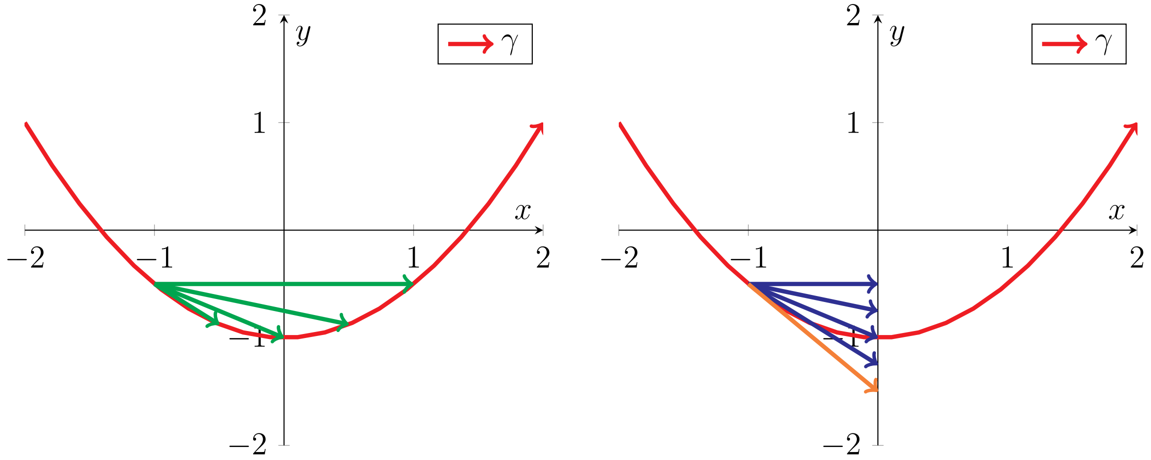


Figure 2.3. The parametric curve $\gamma(t) = \frac{1}{2}x^2 - 1$ is drawn in red above. The green arrows indicate the vectors $\gamma(t) - \gamma(-1)$ for various values of t (starting from $\gamma(-1)$), while the purple arrows represent $\frac{1}{t+1}[\gamma(t) - \gamma(-1)]$ for the same values of t. Finally, the orange arrow represents the limits of the blue arrows as $t \to -1$, which is precisely $\gamma'(-1)$.

As an example, for the parametric curve γ indicated in Figure 2.3, some average rates of change near time $t_0 = -1$ are given by blue arrows, while the instantaneous rate of change at $t_0 = -1$ is indicated by the orange arrow. Notice in particular that the blue arrows become progressively better approximations of the orange arrow as $t \to -1$.

This chain of reasoning has led to a familiar definition from calculus:

Definition 2.2. *For a parametric curve $\gamma : I \to \mathbb{R}^n$, we define the derivative of γ at $t_0 \in I$ to be*

$$\gamma'(t_0) = \lim_{t \to t_0} \frac{\gamma(t) - \gamma(t_0)}{t - t_0} \in \mathbb{R}^n. \tag{2.2}$$

Moreover, we also use the term derivative of γ to refer to the function $\gamma' : I \to \mathbb{R}^n$ mapping each $t_0 \in I$ to the vector $\gamma'(t_0)$ defined in equation (2.2).

In particular, from the discussions above, we see that the derivative γ' provides the solution to Question 2.2, since $\gamma'(t)$ precisely measures the velocity of γ at time t.

2.1.2. *Smoothness.* One basic calculus question that must be addressed is whether the limit (2.2) actually exists. More specifically, recall that a function is differentiable iff the limit that defines the derivative exists. In our current setting, a parametric curve γ being differentiable means that γ is "nice" enough so that the concept of instantaneous rate of change makes sense in the first place. Of course, this needs not always be the case:

Example 2.3. *Consider the function*

$$\lambda : \mathbb{R} \to \mathbb{R}^2, \qquad \lambda(t) = (t, |t|).$$

Note that this λ fails to be differentiable at $t = 0$. In particular, if one travels along λ, then there is an abrupt jump in the direction of λ at $t = 0$. Consequently, the (instantaneous) rate of change at time $t = 0$ fails to be well-defined; see Figure 2.4.

Note that if γ is differentiable, then Definition 2.2 is applicable to γ', so we can ask whether γ' is also differentiable. If so, then its derivative γ'' represents the rate of change of γ'. One can then repeat the process indefinitely for additional higher-order derivatives of γ.

Definition 2.3. *We say that $\lambda : I \to \mathbb{R}^n$ is smooth iff λ is infinitely differentiable, that is, any finite number of derivatives of λ is itself differentiable.*

Smooth parametric curves, for which one can always take as many derivatives as one desires, are in a sense the best behaved. Whenever γ is smooth, no derivative of γ of any order will have any discontinuities or "jagged edges" that prevent one measuring its rate of change.

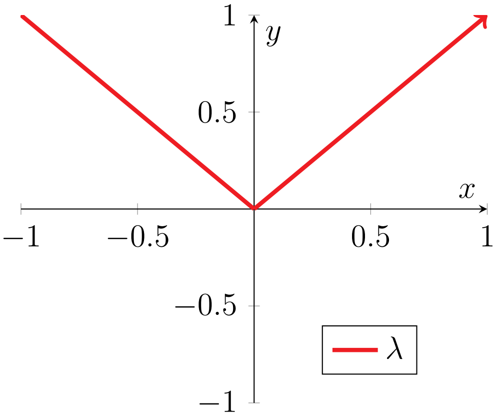


FIGURE 2.4. A plot of the function λ from Example 2.3.

Recall that *by Definition 2.1, all parametric curves that we consider in these notes are assumed to be smooth.* This is primarily done for convenience, as this module will only deal with objects that are differentiable. As a result, for the remainder of this module, we will no longer stress about differentiability of functions.

2.1.3. *Computing Derivatives.* While we have now defined and assigned some physical intuition to derivatives of parametric curves, we still have not discussed how to *compute* such derivatives.

Suppose, for concreteness, that $\gamma : I \to \mathbb{R}^2$, so that γ represents a trajectory in a plane. Since γ takes, for any $t \in I$, two real values, we can express γ componentwise as

$$\gamma(t) = (\gamma_1(t), \gamma_2(t)), \qquad \gamma_1, \gamma_2 : I \to \mathbb{R}. \tag{2.3}$$

In particular, the functions γ_1 and γ_2 represent the x-coordinate and y-coordinate of γ, respectively.

Now, recalling from linear algebra that both addition and scalar multiplication of vectors are defined to apply componentwise, we have, for any $t_0, t \in I$, that

$$\begin{aligned} \frac{\gamma(t) - \gamma(t_0)}{t - t_0} &= \frac{1}{t - t_0} (\gamma_1(t) - \gamma_1(t_0), \gamma_2(t) - \gamma_2(t_0)) \\ &= \left(\frac{\gamma_1(t) - \gamma_1(t_0)}{t - t_0}, \frac{\gamma_2(t) - \gamma_2(t_0)}{t - t_0} \right). \end{aligned}$$

Moreover, as you may remember from multivariable calculus, limits of vector-valued functions can also be applied componentwise. As a result, we obtain that

$$\lim_{t \to t_0} \frac{\gamma(t) - \gamma(t_0)}{t - t_0} = \left(\lim_{t \to t_0} \frac{\gamma_1(t) - \gamma_1(t_0)}{t - t_0}, \lim_{t \to t_0} \frac{\gamma_2(t) - \gamma_2(t_0)}{t - t_0} \right).$$

In summary, the derivative of γ is the vector containing the derivatives of γ_1 and γ_2:

Proposition 2.1. *Let $\gamma : I \to \mathbb{R}^2$ is a parametric curve, and suppose γ is expressed componentwise as in* (2.3)*. Then, given any $t_0 \in I$, we have that*

$$\gamma'(t_0) = (\gamma_1'(t_0), \gamma_2'(t_0)). \tag{2.4}$$

Now, both γ_1 and γ_2 are real-valued functions of one real variable, so you know in principle how to handle these from first-year calculus. In particular, you have a large number of tools and tricks at your disposal (e.g. power rule, product rule, chain rule).

Example 2.4. *Consider the parametric curve*

$$\gamma : \mathbb{R} \to \mathbb{R}^2, \qquad \gamma(t) = (\cos t, \sin t).$$

Let us first compute its derivative. From Proposition 2.1, we see that to compute $\gamma'(t)$, we need simply compute the derivatives of the components $\cos t$ and $\sin t$ of γ. Thus, recalling the formulas from first-year calculus for the trigonometric functions cos *and* sin*, we obtain*

$$\gamma'(t) = \left(\frac{d}{dt} \cos t, \frac{d}{dt} \sin t \right) = (-\sin t, \cos t).$$

To get a better visual sense of what is happening, we compute γ and γ' at various points:

t	0	$\frac{\pi}{4}$	$\frac{\pi}{2}$	$\frac{3\pi}{4}$	π
$\gamma(t)$	$(1,0)$	$(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}})$	$(0,1)$	$(-\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}})$	$(-1,0)$
$\gamma'(t)$	$(0,1)$	$(-\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}})$	$(-1,0)$	$(-\frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}})$	$(0,-1)$

The graph of γ, which maps out a unit circle, and the above values of γ' are drawn in the first plot in Figure 2.5. Here, the velocity $\gamma'(0) = (0,1)$ is drawn with its "tail" at $\gamma(0) = (1,0)$, which should help you see that $\gamma'(0)$ indeed indicates the direction that γ is heading in at the point $\gamma(0)$. The other values of γ and γ' in the above table are also graphed in a similar manner.

While all of the preceding discussion involved parametric curves lying in a plane, the fact that the ambient space is 2-dimensional really played no essential role. All of these results would apply equally well to parametric curves embedded in a space of any dimension. Thus, below we state the corresponding general formula for computing derivatives:

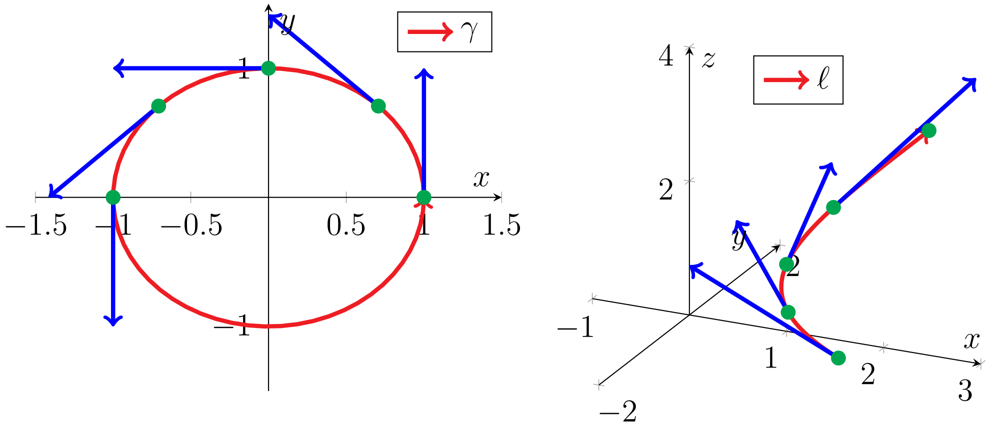


Figure 2.5. The left plot contains a graph of γ from Example 2.4, while the right plot contains a graph of γ from Example 2.5. Both plots contain some sample values of $\gamma(t)$ (in green) and $\gamma'(t)$ (in blue).

Theorem 2.2. *Let $\gamma : I \to \mathbb{R}^n$ be a parametric curve, and express γ as*

$$\gamma(t) = (\gamma_1(t), \gamma_2(t), \dots, \gamma_n(t)), \qquad \gamma_i : I \to \mathbb{R}.$$

Then, for any $t_0 \in I$,

$$\gamma'(t_0) = (\gamma_1'(t_0), \gamma_2'(t_0), \dots, \gamma_n'(t_0)). \tag{2.5}$$

Example 2.5. *Consider the parametric curve*

$$\gamma : (0,1) \to \mathbb{R}^3, \qquad \gamma(t) = (t^2 + 1, t, e^t).$$

Again, to find γ', we merely differentiate each component separately:

$$\gamma'(t) = \left(\frac{d}{dt}(t^2 + 1), \frac{d}{dt} t, \frac{d}{dt} e^t \right) = (2t, 1, e^t).$$

A graph of γ and some samples of $\gamma'(t)$ are provided in the second plot of Figure 2.5.

To draw such graphs yourself, one can in general attempt the following steps:

(1) *Compute $\gamma(t)$ for some points and plot them; for this particular example, these points are given by the green dots in the second plot in Figure 2.5. Once you have a large enough sample of points, you should have a reasonably good idea of what γ looks like.*
(2) *You can then connect the points you have sampled in a reasonable manner; for this particular example, this is indicated by the red path in the second plot of Figure 2.5. (If you are still not sure how to "connect the dots", you could also compute $\gamma'(t)$ at various points—the blue arrows in Figure 2.5—to see which direction γ is going.)*

2.1.4. *Speed and Direction.* Recall that given a vector

$$\mathbf{v} = (v_1, v_2, \dots, v_n) \in \mathbb{R}^n,$$

we can decompose it into two distinct pieces of information:

(1) The norm, or magnitude, of $\mathbf{v}$,

$$|\mathbf{v}| = \sqrt{v_1^2 + v_2^2 + \cdots + v_n^2}, \tag{2.6}$$

which we can interpret as the *length*, or *size*, of $\mathbf{v}$. Indeed, if we think of $\mathbf{v}$ as a directed line segment from the origin to the point in space given by coordinates $(v_1, \dots, v_n)$, then by the Pythagorean theorem, (2.6) is precisely the length of this segment.

(2) Whenever $\mathbf{v} \neq \mathbf{0}$, the vector

$$\frac{\mathbf{v}}{|\mathbf{v}|} \tag{2.7}$$

points in the same direction as $\mathbf{v}$ and has norm 1. In particular, (2.7) filters all the information about length contained in $\mathbf{v}$. Thus, we interpret this as the direction of $\mathbf{v}$.

Now, recall that if a parametric curve $\gamma : I \to \mathbb{R}^n$ models the motion of a particle over time, then its derivative $\gamma'(t)$ represents its velocity at time t. Following the discussion above, the velocity—how γ is changing—can be decomposed into two components:

(1) The norm $|\gamma'(t)|$, which measures the "size" of $\gamma'(t)$, can be interpreted as how fast γ is changing. For this reason, we often refer to $|\gamma'(t)|$ as the speed of γ at time t.

(2) The unit vector $|\gamma'(t)|^{-1}\gamma'(t)$, on the other hand, captures the direction of γ at time t.

Example 2.6. *Consider the parametric unit circle from Example 2.4 (see also Figure 2.5),*

$$\gamma : \mathbb{R} \to \mathbb{R}^2, \qquad \gamma(t) = (\cos t, \sin t).$$

We compute its speed for each $t \in \mathbb{R}$*:*

$$\gamma'(t) = (-\sin t, \cos t), \qquad |\gamma'(t)| = \sqrt{(-\sin t)^2 + (\cos t)^2} = 1.$$

Thus, if we view γ *as a particle trajectory (orbiting around the origin), then the above indicates that the particle represented by* γ *is always moving with constant unit speed.*

2.2. **What is a Curve?** Now that we have some basic working knowledge with parametric curves, we return to the task of nailing down exactly what a "curve" is in our context. Recall that the idea is to start with parametric curves and then identify those that map out the same path as representing the same curve. Here, we make this process more formal.

2.2.1. *Regular Parametrisations.* Our intent here is to study the *geometry* of curves through parametric curves (i.e. *parametrisations* of the curve). Thus, it makes sense to only consider parametrisations that adequately capture a curve's geometric contents. For instance, we wish to omit parametrisations that go back and forth over a single point of the trajectory multiple times. Such parametric curves would contain redundant geometric information whenever they pass by a point more than once. Even more importantly, if we wish to compute the total length of a curve, as we will later, then it is essential that we do not count any point more than once.

For instance, consider a parametric curve γ describing an object swinging along a pendulum, as indicated in Figure 2.6. Since the pendulum swings back and forth, γ would vacillate between going "forward" and "backward" along its trajectory. Consequently, if we wish to find the length of the arc of the pendulum, then measuring the total length of γ would produce the wrong answer, since γ traverses parts of this arc multiple times.

As a result, here we impose an additional condition that excludes the "undesirable" parametric curves that are unsuitable for studying geometric properties:

Definition 2.4. *A parametric curve $\gamma : I \to \mathbb{R}^n$ is called* <u>*regular*</u> *iff $|\gamma'(t)| \neq 0$ for every $t \in I$.*

Note that the condition in Definition 2.4 is in fact more stringent than merely ruling out revisiting points. We also exclude the scenario in which γ "stops" at some time t.

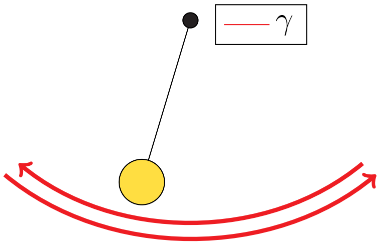


FIGURE 2.6. The red paths show part of the trajectory γ of the end of a pendulum.

Example 2.7. *The parametric unit circle from Example 2.4,*

$$\gamma : \mathbb{R} \to \mathbb{R}^2, \qquad \gamma(t) = (\cos t, \sin t),$$

is regular, since we computed in Example 2.4 that $|\gamma'(t)| = 1$ for any $t \in \mathbb{R}$.

Example 2.8. *Consider now the parametric curve*

$$\ell_* : \mathbb{R} \to \mathbb{R}^3, \qquad \gamma(t) = (t^3, 0, 0).$$

Note that ℓ_* parametrises the x-axis as in the first plot of Figure 2.1.

If we compute its speed, we see that

$$\ell_*'(t) = (3t^2, 0, 0), \qquad |\ell_*'(t)| = |3t^2| = 3t^2.$$

In particular, $|\ell_*'(0)| = 0$, so ℓ_* fails to be regular.

We can contrast ℓ_* with ℓ in Example 2.1, which maps out the same line. Since

$$|\ell'(t)| = |(1, 0, 0)| = 1 \neq 0,$$

then ℓ is a regular parametric curve. Thus, for the purposes of studying geometry, ℓ provides a more useful (and pleasant) parametrisation of the x-axis.

2.2.2. *Reparametrisations.* Recall now the two parametric curves from Example 2.2,

$$\gamma_1 : (0, \pi) \to \mathbb{R}^2, \qquad \gamma_1(t) = (\cos t, \sin t),$$
$$\gamma_2 : (-1, 1) \to \mathbb{R}^2, \qquad \gamma_2(t) = (-t, \sqrt{1 - t^2}).$$

By graphing out γ_1 and γ_2, as was done in Figure 2.2, we saw that both parametric curves trace out the upper half of the unit circle in the anticlockwise direction. We concluded from this that γ_1 and γ_2 should be viewed as parametrising the same curve.

We now want to explore this fact more explicitly. Let us suppose we are extra clever and magically know beforehand what has to be done: we adopt the change of variables

$$\tilde{t} = \phi(t) = -\cos(t). \tag{2.8}$$

Note that if the original parameter t lies within $(0, \pi)$, then $\tilde{t} = -\cos(t)$ takes values in $(-1, 1)$. Moreover, if one increases t, then $\tilde{t}$ also increases (though at a different rate). Thus, each value of

$t \in (0, \pi)$ is matched with exactly one value of $\tilde{t} \in (-1, 1)$. In more formal mathematical terms, ϕ is a strictly increasing bijection between the intervals $(0, \pi)$ and $(-1, 1)$.

There is a very particular reason we chose this particular change of variables. Let us apply γ_1 using the original parameter t and γ_2 with the new parameter $\tilde{t}$. Then, a short computation yields

$$\gamma_2(\tilde{t}) = \gamma_2(-\cos t) = \left(\cos t, \sqrt{1 - (-\cos t)^2}\right) = (\cos t, \sin t) = \gamma_1(t). \tag{2.9}$$

In other words, by switching t into $\tilde{t}$, we have transformed γ_1 into γ_2! Figure 2.7 highlights a few points of this half-circle, in terms of both t and $\tilde{t}$.

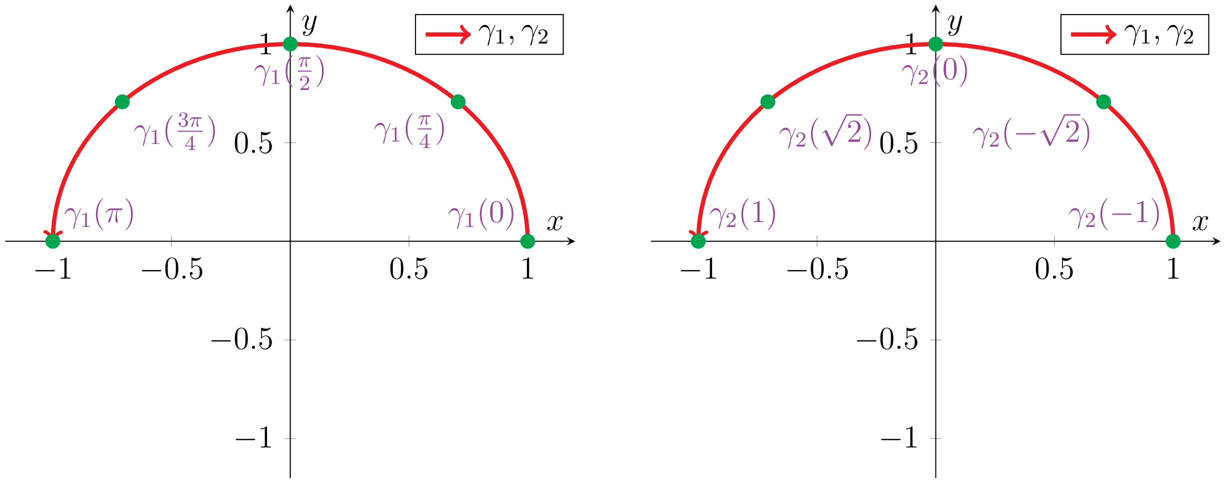


Figure 2.7. Two copies of the curve represented by both γ_1 and γ_2 (from Example 2.2). The same five points on this curve are labelled (in green) in two plots above, the first in terms of γ_1, and the second in terms of γ_2.

How might we interpret this? Recall that we could think of γ_1 and γ_2 as two different particles traversing the same path at different speeds. For any point P along this path, the particle γ_1 reaches P at some time t, while γ_2 reaches P at some other time $\tilde{t}$. The change of variables ϕ is simply the object that matches the time t for γ_1 to the time $\tilde{t}$ for γ_2. Indeed for the ϕ in (2.8), the identity (2.9) shows that $\gamma_2(\tilde{t})$ and $\gamma_1(t)$ represent the same point on the path.

Now, given two parametric curves γ and $\tilde{\gamma}$ tracing the same path, again by "matching the time variables" using an appropriate bijection ϕ (equivalently, a change of variables $\tilde{t} = \phi(t)$), we can identify how γ and $\tilde{\gamma}$ trace the same path. This motivates the following mathematical definition:

Definition 2.5. *Let $\gamma : I \to \mathbb{R}^n$ and $\tilde{\gamma} : \tilde{I} \to \mathbb{R}^n$ be parametric curves. We say that γ is a reparametrisation of $\tilde{\gamma}$ iff there exists a bijection $\phi : I \to \tilde{I}$ between I and $\tilde{I}$ such that*

- *Both ϕ and its inverse ϕ^{-1} are smooth.*
- *The following holds for all $t \in I$:*

$$\tilde{\gamma}(\phi(t)) = \gamma(t). \tag{2.10}$$

Moreover, in the above context, we refer to ϕ as the corresponding change of variables.

Students who are unfamiliar with formal mathematical language are often intimidated by the above definition, but there really is nothing to be afraid of. If we go back to the preceding intuition, and we think of $\phi(t)$ in (2.10) as the "transformed time" $\tilde{t}$, then (2.10) is simply a generalisation of the relation found in (2.9). Definition 2.5 merely extends (2.9) to general parametric curves.

Similar to Example 2.2, one can think of reparametrisations as two parametric curves describing the same curve. Indeed, Definition 2.5 can be reframed as stating that γ *and* $\tilde{\gamma}$ *describe the same curve whenever one has a "matching of times"* ($t \leftrightarrow \tilde{t}$) *that identifies the trajectories of* γ *and* $\tilde{\gamma}$.

Example 2.9. *Consider the two parametric lines (see also Figure 2.8)*

$$\ell : \mathbb{R} \to \mathbb{R}^2, \qquad \ell(t) = (t, 2t + 1, -t),$$
$$\tilde{\ell} : \mathbb{R} \to \mathbb{R}^2, \qquad \tilde{\ell}(\tilde{t}) = (2\tilde{t} - 1, 4\tilde{t} - 1, -2\tilde{t} + 1).$$

We claim that ℓ *is a reparametrisation of* $\tilde{\ell}$.

To show this, we must find the change of variables $t \to \tilde{t}$ *that identifies* ℓ *with* $\tilde{\ell}$. *The most straightforward way to do this is to solve* $\ell(t) = \tilde{\ell}(\tilde{t})$ *for a relation between* t *and* $\tilde{t}$*:*

$$(t, 2t + 1, -t) = \ell(t) = \tilde{\ell}(\tilde{t}) = (2\tilde{t} - 1, 4\tilde{t} - 1, -2\tilde{t} + 1).$$

By looking at each component of the above, we obtain a system of three equations:

$$t = 2\tilde{t} - 1, \qquad 2t + 1 = 4\tilde{t} - 1, \qquad -t = -2\tilde{t} + 1.$$

By doing a bit of algebra, we see that these equations have a simultaneous solution,

$$\tilde{t} = \frac{1}{2}(t + 1). \tag{2.11}$$

As a result, we have that $\ell(t) = \tilde{\ell}(\tilde{t})$ *if we relate* $\tilde{t}$ *and* t *by* (2.11). *Thus, if we define*

$$\phi : \mathbb{R} \to \mathbb{R}, \qquad \phi(t) = \frac{1}{2}(t + 1),$$

which is clearly a bijection between $\mathbb{R}$ *and itself, then the above implies*

$$\tilde{\ell}(\phi(t)) = \tilde{\ell}(\tilde{t}) = \ell(t),$$

that is, ℓ *is indeed a reparametrisation of* $\tilde{\ell}$.

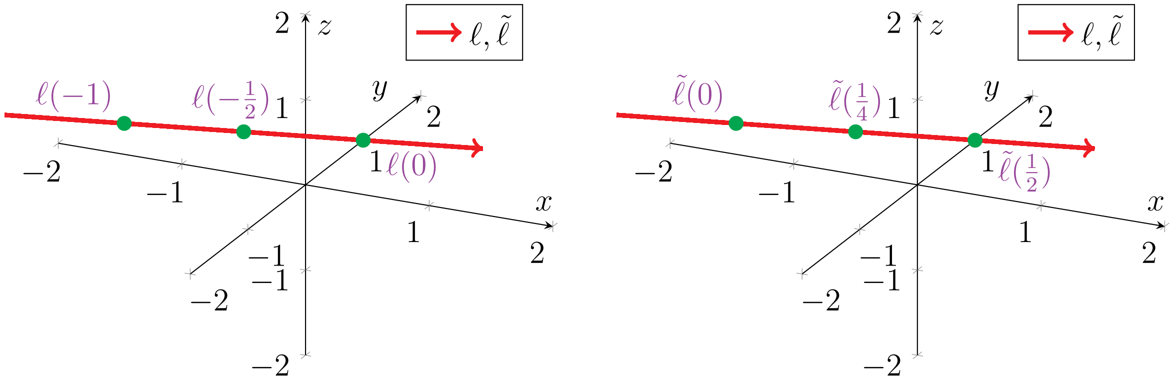


FIGURE 2.8. Two copies of the line represented by ℓ and $\tilde{\ell}$ from Example 2.9. Both plots contain three common points (indicated in green). The first plot is labelled in terms of ℓ_1, while the second is labelled in terms of ℓ_2.

Lastly, we remark that if γ is a reparametrisation of $\tilde{\gamma}$, in the sense of Definition 2.5, then $\tilde{\gamma}$ is also a reparametrisation of γ. Indeed, if there is a change of variables $\tilde{t} = \phi(t)$ satisfying (2.10), then the reverse transformation $t = \phi^{-1}(\tilde{t})$ yields, as desired,

$$\gamma(\phi^{-1}(\tilde{t})) = \tilde{\gamma}(\phi(\phi^{-1}(\tilde{t}))) = \tilde{\gamma}(\tilde{t}).$$

Intuitively, this should not be the least bit surprising, since any reasonable definition of "γ and $\tilde{\gamma}$ represent the same curve" ought to certainly imply that "$\tilde{\gamma}$ and γ represent the same curve".

2.2.3. *Curves Revisited.* We now have a formal notion of reparametrisation characterising when two parametric curves describe the same curve. Thus, we can finally define curves as being represented by parametric curves, with the caveat that reparametrisations represent the same curve. This can be captured in a mathematically precise way through the notion of equivalence classes.

Definition 2.6. *Consider the equivalence relation $\sim$ defined as follows: two parametric curves γ and $\tilde{\gamma}$ satisfy $\gamma \sim \tilde{\gamma}$ iff γ is a reparametrisation of $\tilde{\gamma}$. We then formally define a curve as an equivalence class of regular parametric curves under the equivalence relation $\sim$.*

Remark. From the preceding arguments, we already have that this relation $\sim$ is symmetric. It is also not hard to see that $\sim$ is reflexive and transitive, and hence an equivalence relation.

Note that Definition 2.6 only considers *regular* parametric curves (see Definition 2.4). This is for convenience, as there is no need to consider "undesirable" parametrisations in the first place.

Under Definition 2.6, a curve is formally the collection of parametric curves, all of which are reparametrisations of each other. Indeed, this awkward construction fulfills our intuitions:

- If γ is a parametric curve, then its equivalence class $[\gamma]$ is the curve represented by γ. Conversely, a curve C is represented by any parametric curve γ that is in C.
- On the other hand, if γ and $\tilde{\gamma}$ describe the same curve, then they are reparametrisations of each other. As a result, they are in the same equivalence class, and hence $[\gamma] = [\tilde{\gamma}]$ represents the common curve described by γ and $\tilde{\gamma}$.

Of course, no one in his or her right mind would actually communicate in terms of these equivalence classes when describing curves. In practice, one works instead with parametric curves, with the implicit understanding that the object of interest is the underlying equivalence class. The role of Definition 2.6 is to provide a precise mathematical characterisation of exactly what a curve is.

Remark. We note that the definition we use for curves here is rather unorthodox. For example, the term "curve" is often used elsewhere in the mathematical literature to refer to parametric curves. Furthermore, more advanced texts will often use more general (and more technical) terminology, such as manifolds and immersions. The use of Definition 2.6 reflects our intent to study geometric properties of curves and our desire to maintain an elementary exposition.

For other characterisations of curves throughout mathematics, see the *Wikipedia* article [8].

2.3. **Tangent Vectors and Lines.** The remainder of this chapter is dedicated toward exploring some basic geometric properties of curves. First, we must address the most basic question:

Question 2.3. *What exactly do we mean by a "geometric property" of a curve?*

Intuitively speaking, we can view a geometric property as some attribute that is possessed by every curve. To be more mathematically precise, we can formulate such a property as a *function* mapping each curve in $\mathbb{R}^n$ to some value. One basic example is arc length, which we can think as a function L mapping each curve C to the total length of C. (Spoiler: Arc length is indeed a geometric property, as we shall see later in this chapter.)

2.3.1. *Independence of Parametrisation.* In practice, we tend to study curves indirectly through parametric curves. Thus, it makes sense for us to first reframe this discussion in terms of parametric curves. To connect back to Question 2.3, we pose the following question:

Question 2.4. *Suppose we have a property $\mathcal{P}$ of parametric curves. When is this property $\mathcal{P}$ also a geometric property of the underlying curves represented by the parametric curves?*

Let γ and $\tilde{\gamma}$ be parametric curves, and suppose that γ and $\tilde{\gamma}$ describe the same underlying curve C (that is, they are reparametrisations of each other). Since γ and $\tilde{\gamma}$ are "the same" in the sense of curves, then for $\mathcal{P}$ to be a property of curves, we require that γ and $\tilde{\gamma}$ have the same $\mathcal{P}$-property. Moreover, this equality must be true for any parametrisations that represent the same curve. Thus, the key point is the following: *in order for $\mathcal{P}$ to be a geometric property of curves, we require that $\mathcal{P}$ is independent of how a curve is parametrised.*

To be less abstract, let us consider some simple examples of parametric curve properties:

Example 2.10. *Consider the parametric curves γ_1 and γ_2 from Example 2.2 and Figure 2.2,*

$$\gamma_1 : (0, \pi) \to \mathbb{R}^2, \qquad \gamma_1(t) = (\cos t, \sin t),$$
$$\gamma_2 : (-1, 1) \to \mathbb{R}^2, \qquad \gamma_2(t) = (-t, \sqrt{1 - t^2}),$$

which both represent the upper half of the unit circle C_+. Clearly, γ_1 and γ_2 have different domains:

$$\mathsf{D}(\gamma_1) = (0, \pi), \qquad \mathsf{D}(\gamma_2) = (-1, 1).$$

Thus, we conclude that "domain" is not a geometric property of curves, and of C_+ in particular, since it very much depends on how you parametrise the curve.

Example 2.11. *Remaining with the setting from Example 2.10, let us contrast the discussion there with the notion of arc length. Intuitively, it seems sensible to associate a length with C_+ itself, not with γ_1 or γ_2. Since C_+ is half the unit circle, we would expect its length to be*

$$\mathsf{L}(\mathsf{C}_+) = \frac{1}{2}(2\pi \cdot 1) = \pi.$$

Later in this chapter, we will learn how to compute $\mathsf{L}(\mathsf{C}_+)$ using γ_1 or γ_2, or any other parametrisation of C_+. We will then see that regardless of which parametrisation of C_+ we choose, the answer will always be π. This supports the claim that arc length is a geometric property of curves.

Next, let us return to yet another property of parametric curves: their *derivatives.* Using a familiar tool from calculus—the beloved chain rule—we can relate the deriatives of two parametric curves when they are reparametrisations of each other.

Theorem 2.3. *Let $\gamma : I \to \mathbb{R}^n$ and $\tilde{\gamma} : \tilde{I} \to \mathbb{R}^n$ be parametric curves. Suppose γ is a reparametrisation of $\tilde{\gamma}$, i.e. there is a change of variables $\phi : I \to \tilde{I}$ such that*

$$\tilde{\gamma}(\phi(t)) = \gamma(t).$$

Then, the following identity holds:

$$\gamma'(t) = \phi'(t) \cdot \tilde{\gamma}'(\phi(t)). \tag{2.12}$$

Proof. To derive (2.12), we first express γ and $\tilde{\gamma}$ in terms of their components,

$$\gamma(t) = (\gamma_1(t), \dots, \gamma_n(t)), \qquad \tilde{\gamma}(\tilde{t}) = (\tilde{\gamma}_1(\tilde{t}), \dots, \tilde{\gamma}_n(\tilde{t})).$$

Using Theorem 2.2, we can expand $\gamma'(t)$ in terms of these components as

$$\gamma'(t) = \frac{d}{dt}[\tilde{\gamma}(\phi(t))] = \left(\frac{d}{dt}[\tilde{\gamma}_1(\phi(t))], \dots, \frac{d}{dt}[\tilde{\gamma}_n(\phi(t))] \right).$$

Note that the right-hand side now contains derivatives of *real-valued* functions. As a result, we can apply the chain rule from first-year calculus to obtain (2.12):

$$\begin{aligned} \gamma'(t) &= (\phi'(t) \cdot \tilde{\gamma}_1'(\phi(t)), \dots, \phi'(t) \cdot \tilde{\gamma}_n'(\phi(t))) \\ &= \phi'(t) \cdot (\tilde{\gamma}_1'(\phi(t)), \dots, \tilde{\gamma}_n'(\phi(t))) \\ &= \phi'(t) \cdot \tilde{\gamma}'(\phi(t)). \end{aligned}$$ □

The first point to note is that the derivative is not a geometric property of curves, since (2.12) shows that a reparametrisation will generally change the derivative. On the other hand, (2.12) also shows that at each point, the derivative will only change by a rescaling (by the factor $\phi'(t)$).

To see how we might interpret Theorem 2.3, let us consider the curve C in red in Figure 2.9. The picture on the left shows a person walking slowly along C while humming his favourite tune, while the picture on the right shows a person running quickly along C because he is late to his next lecture. These two people can be modelled by two parametric curves γ and $\tilde{\gamma}$.

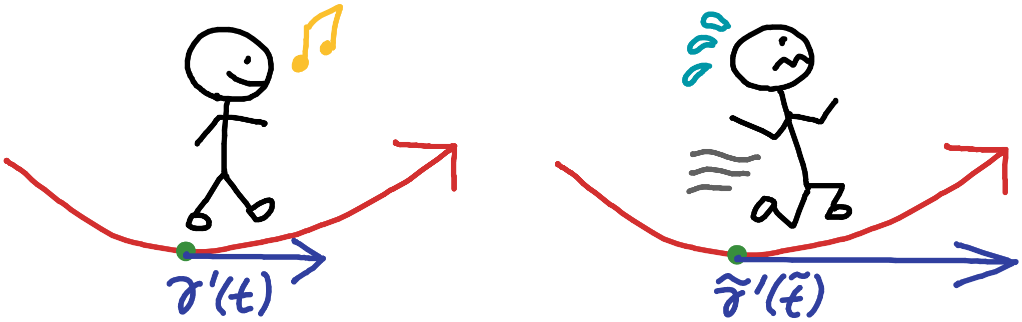


FIGURE 2.9. The pictures show one person walking and another person running along a common red curve. At a single point along the curve (in green), their velocities are indicated by blue arrows. Both people are heading in the same direction, but the runner has a higher speed.

Now, at a point of C, indicated by green dot on the figure, both the walking man (γ) and the running man ($\tilde{\gamma}$) are heading in the same direction, i.e. $\gamma'(t)$ and $\tilde{\gamma}'(\tilde{t})$ have the same direction.

However, the running man would be heading in this direction at a much faster speed, so the velocity $\tilde{\gamma}'(\tilde{t})$ of the running man would be larger than the velocity $\gamma'(t)$ of the walking man. These velocities are indicated by the blue arrows in Figure 2.9. This ratio of the speeds of γ and $\tilde{\gamma}$ is precisely what is captured by the factor $\phi'(t)$ in (2.12).

2.3.2. *Orientation.* The next topic of discussion is the orientation of curves. Though not a geometric property, orientation will play essential roles in some computations in the future. To introduce the concept, we begin with a very simple example:

Example 2.12. *Consider the parametric curves*

$$\ell_1 : \mathbb{R} \to \mathbb{R}^3, \qquad \ell_1(t) = (0, t, 0),$$
$$\ell_2 : \mathbb{R} \to \mathbb{R}^3, \qquad \ell_2(t) = (0, -t, 0),$$

whose graphs are illustrated in Figure 2.10.

Observe that ℓ_1 and ℓ_2 both describe the same curve, the y-axis. Indeed, ℓ_2 is a reparametrisation of ℓ_1, since $\ell_2(t) = \ell_1(\phi(t))$ for $\phi(t) = -t$. On the other hand, ℓ_1 and ℓ_2 traverse the y-axis in opposite directions—ℓ_1 in the increasing y-direction, and ℓ_2 in the decreasing y-direction.

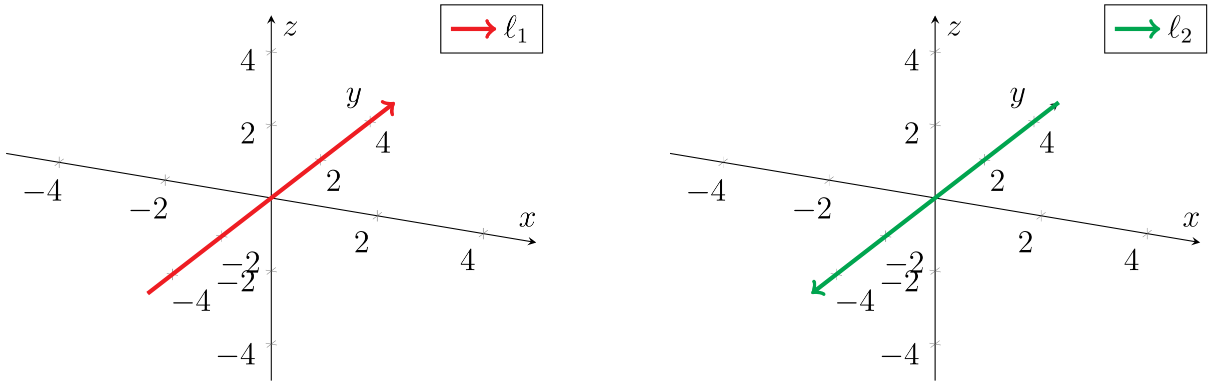


FIGURE 2.10. The two plots contain graphs of the oppositely oriented parametrisations ℓ_1 and ℓ_2 of the y-axis from Example 2.12.

Example 2.12 demonstrates two parametrisations of a curve having "opposite orientations". Intuitively, we can think of orientation as the choice of a direction that one travels along a curve. In particular, for any curve, there are exactly two orientations; in Example 2.12, the possible orientations are given by the directions of increasing or decreasing y-value.

When one chooses to work with a parametrisation γ of a curve, one also fixes an orientation, given by the direction γ traverses along the curve. In Example 2.12, if we use ℓ_2 to parametrise the y-axis, then we have chosen the direction in which the y-coordinate is decreasing.

In practice, orientation will usually be manifested in the signs of some computed quantities. One example we will soon encounter in this module is the signed curvature; whether this is positive or negative will depend on which orientation we choose. For this reason, orientation will be a persistent thorn in our side. In general, there is no reason to favour one orientation of a curve over the other (for instance, it is no more natural to run around in a circle clockwise than anticlockwise).

Consequently, we will see that many computations will carry a sign ambiguity resulting from this need to choose one of two equally valid orientations.

The following question addresses orientations from the point of view of parametric curves:

Question 2.5. *Suppose that we have two regular parametrisations γ, $\tilde{\gamma}$ of a curve. How do we detect whether they have the same or the opposite orientations?*

Since γ and $\tilde{\gamma}$ are reparametrisations of each other, there must be some change of variables ϕ such that $\tilde{\gamma}(\phi(t)) = \gamma(t)$. Letting $\tilde{t} = \phi(t)$, we observe that:

- If $\tilde{t} = \phi(t)$ is an increasing function of t, then γ and $\tilde{\gamma}$ are traversing in the same direction, and hence they must have the same orientation.
- If $\tilde{t} = \phi(t)$ is a decreasing function of t, then γ and $\tilde{\gamma}$ are traversing in opposite directions, and hence they define opposite orientations.

As you know from calculus, whether ϕ is increasing or decreasing can be determined by the sign of ϕ'. Thus, *γ and $\tilde{\gamma}$ have the same orientation if $\phi' > 0$, and opposite orientations if $\phi' < 0$.*

Another way to see this is through the identity (2.12):

- Whenever $\phi' > 0$, the velocities $\gamma'(t)$ and $\tilde{\gamma}'(\tilde{t})$ (at a common point of curve) point in the same direction, hence γ and $\tilde{\gamma}$ must have the same orientation.
- On the other hand, whenever $\phi' < 0$, the velocities $\gamma'(t)$ and $\tilde{\gamma}'(\tilde{t})$ point in the opposite directions, hence γ and $\tilde{\gamma}$ must have opposite orientations.

(Notice that Theorem 2.3 also ensures that as long as we remain with regular parametrisations, ϕ' can never vanish. Thus, ϕ' must always have constant, nonzero sign.)

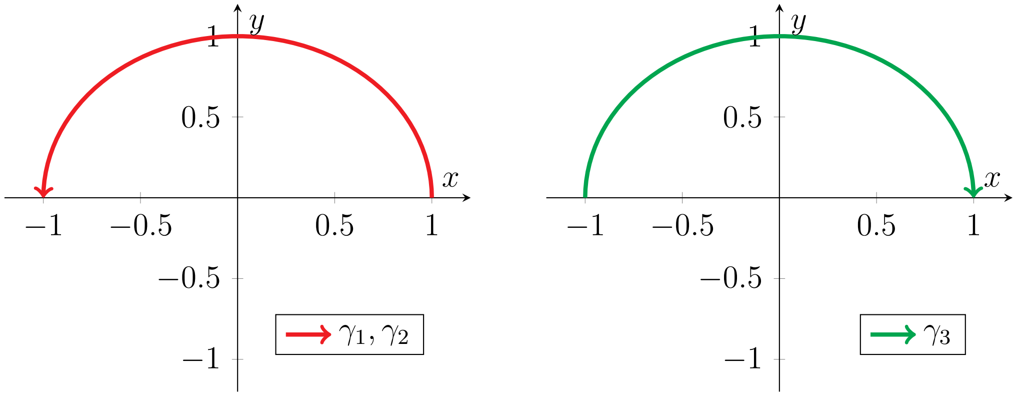


FIGURE 2.11. The parametric curves γ_1, γ_2, γ_3 from Example 2.13.

Example 2.13. *Consider the parametrisations of the upper half-circle from Example 2.2:*

$$\gamma_1 : (0, \pi) \to \mathbb{R}^2, \qquad \gamma_1(t) = (\cos t, \sin t),$$
$$\gamma_2 : (-1, 1) \to \mathbb{R}^2, \qquad \gamma_2(t) = (-t, \sqrt{1 - t^2}).$$

Recall that γ_1 is a reparametrisations of γ_2, with

$$\gamma_1(t) = \gamma_2(\phi(t)), \qquad \phi(t) = -\cos t.$$

As expected, γ_1 and γ_2 define the same (anticlockwise) orientation of the upper half-circle, since

$$\phi'(t) = \sin t > 0, \qquad t \in (0, \pi).$$

Consider also a third parametric curve:

$$\gamma_3 : (-1, 1) \to \mathbb{R}^2, \qquad \gamma_3(t) = (t, \sqrt{1 - t^2}),$$

Note that γ_3 is a reparmetrisation of γ_2, since

$$\gamma_3(t) = \gamma_2(\psi(t)), \qquad \psi(t) = -t.$$

Furthermore, γ_3 and γ_2 have opposite orientations, since

$$\psi'(t) = -1 < 0, \qquad t \in (-1, 1).$$

Indeed, by graphing γ_3 (see Figure 2.11), we see that it is traversing the circle clockwise.

There is a cheap method for reversing the orientation of a parametric curve γ, which we already implicitly demonstrated in Example 2.13. The idea is to simply adopt the change of variables $\tilde{t} = \phi(t) = -t$, which reverses the direction of travel.

Proposition 2.4. *Suppose $\gamma : (a, b) \to \mathbb{R}^n$ is a parametric curve, with $-\infty \leq a < b \leq \infty$. Then,*

$$\gamma^* : (-b, -a) \to \mathbb{R}^n, \qquad \gamma^*(t) = \gamma(-t)$$

defines a reparametrisation of γ that has the opposite orientation as γ.

Proof. γ^* is indeed a reparametrisation of γ, since

$$\gamma^*(t) = \gamma(\phi(t)), \qquad \phi(t) = -t, \qquad t \in (-b, -a).$$

Since $\phi'(t) = -1 < 0$, then $\tilde{\gamma}$ and γ have opposite orientations. □

Finally, a useful construct is the notion of an oriented curve, that is, "a curve along with a choice of orientation". While this informal description is sufficient as a working definition for future discussions, we could also formally describe this via the following definition:

Definition 2.7. *Consider the equivalence relation $\sim_o$ that is defined as follows: two parametric curves γ and $\tilde{\gamma}$ satisfy $\gamma \sim_o \tilde{\gamma}$ iff γ is a reparametrisation of $\tilde{\gamma}$, and γ and $\tilde{\gamma}$ have the same orientation. We can then formally define an oriented curve as an equivalence class of regular parametric curves under the equivalence relation $\sim_o$.*

2.3.3. *The Tangent Line.* Although the derivative is not a geometric property of curves, we can use it to generate a related geometric property. To do this, we define the following:

Definition 2.8. *Consider a regular parametric curve $\gamma : I \to \mathbb{R}^n$, as well as a point $\gamma(t)$ along this curve. We define the tangent line to γ at t to be the set*

$$\mathcal{T}_\gamma(t) = \{\gamma(t) + s \cdot \gamma'(t) \mid s \in \mathbb{R}\}. \tag{2.13}$$

To make geometric sense of the tangent line, we examine the formula $\gamma(t) + s \cdot \gamma'(t)$ in the right-hand side of (2.13). (Here, t is *fixed*, while s is the parameter that varies.) Observe that:

- Taking $s = 0$, we see that $\gamma(t)$ lies in $\mathcal{T}_\gamma(t)$.
- By varying s, we pick up points reached by moving from $\gamma(t)$ along the $\gamma'(t)$-direction.

Thus, we see that $\mathcal{T}_\gamma(t)$ *represents the line through the point* $\gamma(t)$ *in the direction given by* $\gamma'(t)$, *that is, in the direction in which the curve is moving.*

Furthermore, since $\mathcal{T}_\gamma(t)$ precisely captures the same direction of γ at t, we can also think of $\mathcal{T}_\gamma(t)$ as *the line that best approximates* γ *near* $\gamma(t)$. In particular, this reaffirms the interpretation of derivatives as the "best linear approximation" of a function.

To make this more clear, let us consider some concrete examples and figures.

Example 2.14. *Consider the parametric curve*

$$\gamma : \mathbb{R} \to \mathbb{R}^2, \qquad \gamma(t) = (\cos t, \sin t),$$

which represents the unit circle about the origin. Taking $t = \frac{\pi}{4}$, *we compute*

$$\gamma\left(\frac{\pi}{4}\right) = \left(\cos\frac{\pi}{4}, \sin\frac{\pi}{4}\right) = \left(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right),$$
$$\gamma'\left(\frac{\pi}{4}\right) = \left(-\sin\frac{\pi}{4}, \cos\frac{\pi}{4}\right) = \left(-\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right).$$

The tangent line to γ *at* $\frac{\pi}{4}$ *is then given by*

$$\begin{aligned}\mathcal{T}_\gamma\left(\frac{\pi}{4}\right) &= \left\{\left(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right) + s\left(-\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right) \mid s \in \mathbb{R}\right\} \\ &= \left\{\left(\frac{1}{\sqrt{2}} - \frac{1}{\sqrt{2}}\cdot s, \frac{1}{\sqrt{2}} + \frac{1}{\sqrt{2}}\cdot s\right) \mid s \in \mathbb{R}\right\}.\end{aligned}$$

See Figure 2.12 for a graphical representation: the curve represented by γ *is drawn in red, while* $\gamma(\frac{\pi}{4})$, $\gamma'(\frac{\pi}{4})$, *and* $\mathcal{T}_\gamma(\frac{\pi}{4})$ *are drawn in green, blue, and purple, respectively.*

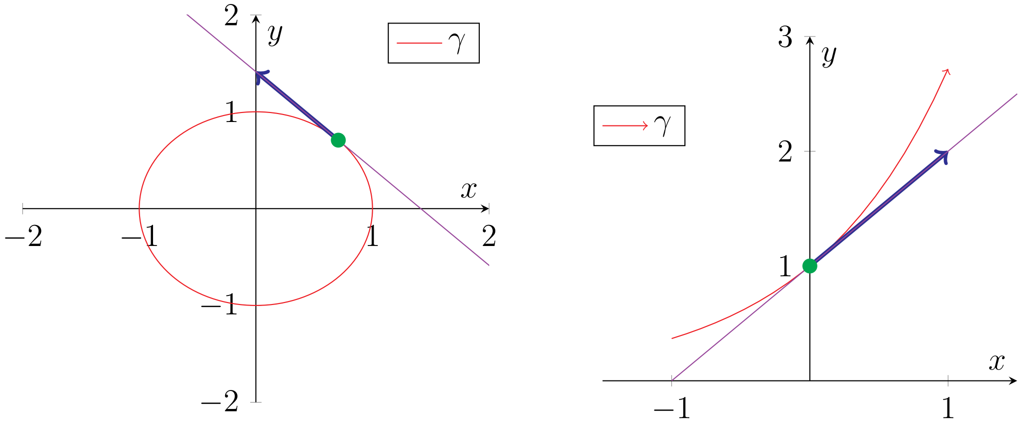


FIGURE 2.12. The left plot is the setup from Example 2.14—γ, $\gamma(\frac{\pi}{4})$, $\gamma'(\frac{\pi}{4})$, and $\mathcal{T}_\gamma(\frac{\pi}{4})$ are indicated in red, green, blue, and purple, respectively. The right plot contains the setup from Example 2.15, with $f(t) = e^t$—γ_f, $\gamma_f(0)$, $\gamma_f'(0)$, and $\mathcal{T}_{\gamma_f}(0)$ are indicated in red, green, blue, and purple, respectively.

Next, we consider a familiar setting from first-year calculus:

Example 2.15. *Let* $f : (-1, 1) \to \mathbb{R}$*, and consider the parametric curve*

$$\gamma_f : (-1, 1) \to \mathbb{R}^2, \qquad \gamma_f(t) = (t, f(t)),$$

which is precisely the <u>*graph*</u> *of the function* f*. For any* $t \in (-1, 1)$*, we have that*

$$\gamma_f'(t) = (1, f'(t)),$$

while the tangent line of the graph of f *at* t *is given by*

$$\mathcal{T}_{\gamma_f}(t) = \{(t, f(t)) + s(1, f'(t)) \mid s \in \mathbb{R}\}.$$

These objects are indicated in the second part of Figure 2.12.

Moreover, in the equation $(t, f(t)) + s(1, f'(t))$ *of the tangent line, note that if we increase* s *by* 1*, then the* x*-component increases by* 1*, while the* y*-component increases by* $f'(t)$*. Then, the slope of this tangent line through* $(t, f(t))$ *is given by this ratio of the changes in* y *and* x*:*

$$\textit{slope} = \frac{\Delta y}{\Delta x} = \frac{f'(t)}{1} = f'(t).$$

As expected, the slope of a tangent line through the graph of f *is given by* f'*.*

We previously hinted that the tangent line is a geometric property of curves, but why is this so? Suppose $\tilde{\gamma}$ is a reparametrisation of a parametric curve γ, and let t and $\tilde{t}$ be such that $\tilde{\gamma}(\tilde{t}) = \gamma(t)$, that is, we consider a common point along the trajectories of γ and $\tilde{\gamma}$.

Recall that Theorem 2.3 implies $\tilde{\gamma}'(\tilde{t})$ is a scalar multiple of $\gamma'(t)$. Thus, the set of all scalar multiples $s \cdot \tilde{\gamma}'(\tilde{t})$ of $\tilde{\gamma}'(t)$ is identical to the set of all scalar multiples $s \cdot \gamma'(t)$ of $\gamma'(t)$, and hence

$$\mathcal{T}_{\tilde{\gamma}}(\tilde{t}) = \{\tilde{\gamma}(\tilde{t}) + s \cdot \tilde{\gamma}'(\tilde{t}) \mid s \in \mathbb{R}\} = \{\gamma(t) + s \cdot \gamma'(t) \mid s \in \mathbb{R}\} = \mathcal{T}_{\gamma}(t).$$

In particular, the above shows that *the tangent line is independent of parametrisation*, hence it must be a geometric property of curves. Intuitively, this seems sensible, since the tangent line indicates the "two possible directions of the curve at a point"; this should depend on the shape and the positioning of the curve, not on how one moves along the curve.

2.3.4. *Tangent Vectors.* We conclude this section with a more conceptual discussion—we introduce a different (but equivalent) perspective for characterising and studying tangent lines.

For this, we must first go back to some basic intuitions behind vectors. When you first learned about vectors, you probably visualised them as "arrows" in the plane or in space. More specifically, you probably saw them as objects that "started from a point" and "pointed in some direction".

To describe these "arrows" formally, we require two pieces of information:

- The starting point $\mathbf{p} \in \mathbb{R}^n$ of the "arrow".
- The direction $\mathbf{v} \in \mathbb{R}^n$ in which the "arrow" is pointing.

Then, from this pair $\mathbf{v}$ and $\mathbf{p}$, we can craft the mathematical object $\mathbf{v}|_{\mathbf{p}}$ representing this "arrow" that begins at $\mathbf{p}$ and points in the direction $\mathbf{v}$.

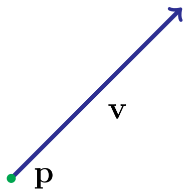


FIGURE 2.13. The "arrow" $\mathbf{v}|_{\mathbf{p}}$.

Remark. Note that this is closely related to the concept of <u>bound vectors</u>, which you encountered in <u>MTH4103/4203: Geometry I</u>, though we will use them in rather different ways. In particular, the "arrow" $\mathbf{v}|_{\mathbf{p}}$ above corresponds to the bound vector starting at $\mathbf{p}$ and terminating at $\mathbf{p}+\mathbf{v}$. Moreover, $\mathbf{v}$ is the free vector represented by this bound vector.

Observe that there is a natural *linear structure* associated with these "arrows", in the sense that we can do "vector-like things" with them. For example:

- Given an "arrow" and a scalar $c \in \mathbb{R}$, we can multiply them:

$$c \cdot \mathbf{v}|_{\mathbf{p}} = (c\mathbf{v})|_{\mathbf{p}}. \tag{2.14}$$

- Given two "arrows" *starting from the same point*, we can add them.

$$\mathbf{v}|_{\mathbf{p}} + \mathbf{w}|_{\mathbf{p}} = (\mathbf{v}+\mathbf{w})|_{\mathbf{p}}. \tag{2.15}$$

- We an also naturally measure the length of an "arrow":

$$|\mathbf{v}|_{\mathbf{p}}| = |\mathbf{v}|. \tag{2.16}$$

Figure 2.14 demonstrates the geometric interpretations associated with (2.14) and (2.15).

Remark. Formally, one would describe these "arrows" an ordered pair $(\mathbf{v}, \mathbf{p}) \in \mathbb{R}^n \times \mathbb{R}^n$. However, we remain with the notation $\mathbf{v}|_{\mathbf{p}}$, as it is more suggestive of the intuitions we wish to capture.

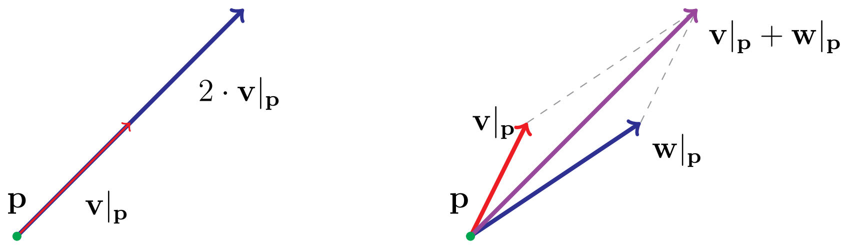

$2 \cdot \mathbf{v}|_{\mathbf{p}}$
$\mathbf{p}$
$\mathbf{v}|_{\mathbf{p}}$
$\mathbf{v}|_{\mathbf{p}} + \mathbf{w}|_{\mathbf{p}}$
$\mathbf{v}|_{\mathbf{p}}$
$\mathbf{w}|_{\mathbf{p}}$
$\mathbf{p}$

FIGURE 2.14. The above show the geometric meanings of (2.14) and (2.15).

Now, how do these "arrows" relate to studying the geometry of curves? Consider a parametric curve $\gamma : I \to \mathbb{R}^n$, which models the trajectory of a particle. While we already argued that $\gamma'(t)$ represents the velocity of the particle at time t, conceptually speaking, this does not yet capture all the information we have. Indeed, it makes more conceptual sense to model the velocity as an "arrow" beginning at $\gamma(t)$ and pointing in the direction of $\gamma'(t)$, that is, the object $\gamma'(t)|_{\gamma(t)}$.

Definition 2.9. *Given a parametric curve $\gamma : I \to \mathbb{R}^n$, we define its tangent vector at $t \in I$ to be*

$$\gamma'(t)|_{\gamma(t)}. \tag{2.17}$$

Example 2.16. *In Figures 2.5, 2.9, and 2.12, the blue arrows that were drawn to indicate the velocity were actually the tangent vectors ($\gamma'(t)|_{\gamma(t)}$), since they also had starting points.*

Next, notice that the two pieces of information in the tangent vector (2.17)—$\gamma'(t)$ and $\gamma(t)$—are precisely the same information describing the tangent line $T_\gamma(t)$ from Definition 2.8. As a result, we can alternately characterise the tangent line as:

Definition 2.10. *Assuming the setting of Definition 2.8, we define the tangent line to γ at t as*

$$T_\gamma(t) = \{s \cdot \gamma'(t)|_{\gamma(t)} \mid s \in \mathbb{R}\}. \tag{2.18}$$

Definition 2.10 gives a slightly different view of the tangent line. Here, we characterise it as the set of all "arrows" beginning at the point $\gamma(t)$ on the curve and pointing in a manner tangent to γ. Of course, this is formally equivalent to the previous Definition 2.8. The main conceptual difference is that Definition 2.10 highlights the linear structure inherent to the tangent line.

Suppose $\tilde{\gamma}$ is a reparametrisation of γ. By Theorem 2.3, we see that at the corresponding point $\tilde{\gamma}(\tilde{t})$, the tangent vector of $\tilde{\gamma}$ is but a scalar multiple of that of γ, that is,

$$\tilde{\gamma}'(\tilde{t})|_{\tilde{\gamma}(\tilde{t})} = \tilde{\gamma}'(\phi(t))|_{\tilde{\gamma}(\phi(t))} = \frac{1}{\phi'(t)} \cdot \gamma'(t)|_{\gamma(t)},$$

In other words, when a curve is reparametrised, its corresponding tangent vector is transformed via a scalar multiplication of the form (2.14). Since we can adapt reparametrisations—via clever enough choices of ϕ in Theorem 2.3—to achieve any value we want for $\phi'(t)$, this leads to a fully geometric characterisation of the tangent line of Definition 2.10:

Theorem 2.5. *For a curve C, its tangent line at a given point $\mathbf{p}$ is the set of all possible tangent vectors at $\mathbf{p}$ from all possible parametrisations of C.*

In the case of curves, the one-dimensional structure of the tangent line is simple enough that there is little direct benefit to introducing all this new machinery. However, this alternative way of thinking will be more important in higher dimensions, in particular when we study surfaces.

2.4. **Curve Integrals.**

We now discuss another question that is closely tied to calculus:

Question 2.6. *Given a finite curve, how do we define and compute its length?*

We answer this question using tools from single-variable calculus, and we show that this length defines a geometric property of curves. In addition, this discussion will suggest a notion of "integration along a curve", which we treat at the end of the section.

2.4.1. *Arc Length.*

Let us first tackle Question 2.6. If the curve in question is a line segment ℓ, say from a point $\mathbf{p}$ to another point $\mathbf{q}$, then what the length of the segment is should be clear. Indeed, by treating $\mathbf{p}$ and $\mathbf{q}$ as "position vectors", this segment can be represented by the difference $\mathbf{q} - \mathbf{p}$. The length of ℓ then corresponds to the norm of this difference

$$L(\ell) = |\mathbf{q} - \mathbf{p}|.$$

Now, for general curves, we adopt a common strategy from integral calculus: we *approximate the curve as a finite number of line segments.* We consider a parametric curve $\gamma : (a, b) \to \mathbb{R}^n$, and we choose a finite number of points along γ:

$$\gamma(t_0), \gamma(t_1), \dots, \gamma(t_N), \qquad a < t_0 < t_1 < \cdots < t_N < b.$$

See the plots in Figure 2.15 for an example of this situation; the curve represented by γ is drawn in red, while the sample points along γ are indicated in green.

We now approximate γ by "connecting the dots", that is, by taking the collection Γ of line segments made by connecting each sample point $\gamma(t_{i-1})$ with its successive point $\gamma(t_i)$. In Figure 2.15, this is indicated by the blue line segments. The total length of Γ would then give an approximation of the length of γ. Since the segment connecting $\gamma(t_{i-1})$ and $\gamma(t_i)$ has length $|\gamma(t_i)-\gamma(t_{i-1})|$, the total length of Γ can obtained by adding these lengths of all the individual segments:

$$L(\Gamma) = \sum_{i=1}^{N} |\gamma(t_i) - \gamma(t_{i-1})| = \sum_{i=1}^{N} \frac{|\gamma(t_i) - \gamma(t_{i-1})|}{|t_i - t_{i-1}|} \cdot \Delta t_i, \qquad \Delta t_i = t_i - t_{i-1}. \tag{2.19}$$

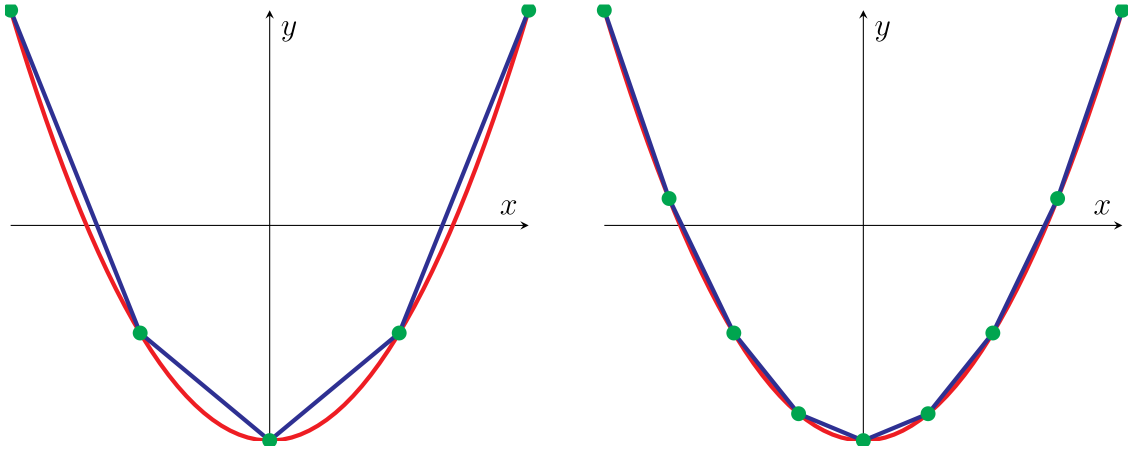


FIGURE 2.15. To approximate the length of the red parabolic curve, we pick a finite sample of points (in green) on the curve, and we compute the total length of the line segments (in blue) connecting the sampled points. In terms of (2.19), we took $N = 4$ on the left plot, and $N = 8$ on the right.

Of course, $L(\Gamma)$ does not give the actual the length of γ, only an approximate value. If we are not satisfied with this approximation, we can try again with a larger sampling of points, i.e. we increase the value of N. By choosing a larger number of points along γ that are closer to each other, the resulting line segments will be a better approximation of γ than before; for example, compare the left and right plots in Figure 2.15. For this new Γ, we can once again calculate its length as in (2.19), obtaining an improved approximation of the length of γ.

Now, to obtain the *exact* length of γ, the idea is to use an "infinitely good" approximation. To do this, we let the number of points we sample (N) tend to infinity, and we let the distance between successive t-parameters in our sampling tend to 0. Note that as $t_i - t_{i-1} \to 0$, the ratio

$$\frac{\gamma(t_i) - \gamma(t_{i-1})}{t_i - t_{i-1}}$$

approaches the derivative $\gamma'(t_{i-1})$. While we do not have the background for discussing this rigorously, the intuitive idea is that in taking the above limits, we obtain that

$$\lim_{\substack{N\to\infty \\ \Delta t \to 0}} L(\Gamma) = \lim_{\substack{N\to\infty \\ \Delta t \to 0}} \sum_{i=1}^{N} \frac{|\gamma(t_i) - \gamma(t_{i-1})|}{|t_i - t_{i-1}|} \cdot \Delta t_i = \int_a^b |\gamma'(t)| dt.$$

Remark. Those who interested in more rigorous developments of integration theory should consult material from real analysis, or from the module MTH5105: Differential and Integral Analysis.

In summary, the above process motivates our next definition:

Definition 2.11. *Given a regular parametric curve $\gamma : (a, b) \to \mathbb{R}^n$, we define its arc length by*

$$L(\gamma) = \int_a^b |\gamma'(t)| dt. \tag{2.20}$$

Let us test our brand new definition on some basic examples:

Example 2.17. *Fix $x_0, y_0 \in \mathbb{R}$, and consider the parametric line segment*

$$\ell : (0, 1) \to \mathbb{R}^2, \qquad \ell(t) = (x_0 t, y_0 t).$$

This is the line segment between the origin ($t = 0$) and the point (x_0, y_0) ($t = 1$).

To compute its length, we first compute its speed:

$$\ell'(t) = (x_0, y_0), \qquad |\ell'(t)| = \sqrt{x_0^2 + y_0^2}.$$

Since x_0, y_0 are constants, then the arc length is given by

$$L(\ell) = \int_0^1 |\ell'(t)| dt = \sqrt{x_0^2 + y_0^2} \int_0^1 dt = \sqrt{x_0^2 + y_0^2}.$$

Note that this is exactly the length you would expect from the Pythagorean theorem.

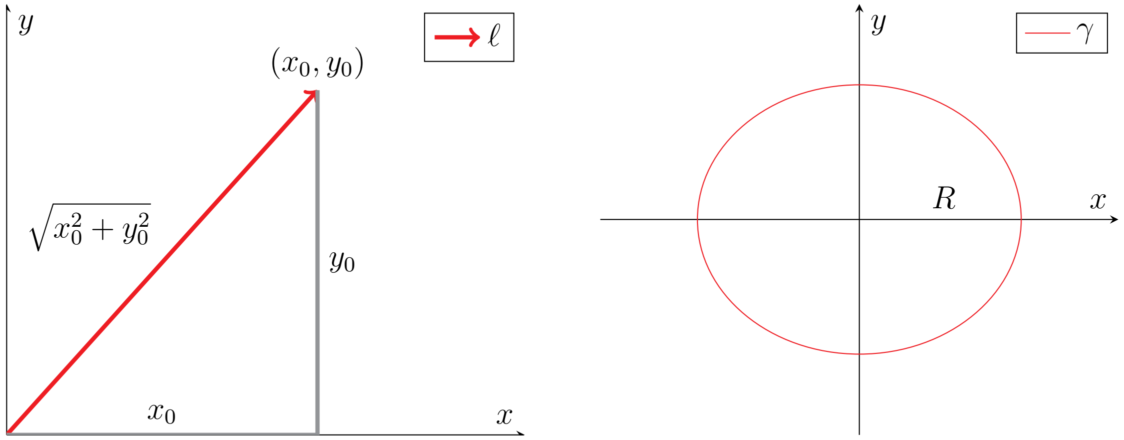


FIGURE 2.16. The settings from Examples 2.17 and 2.18, respectively.

Next, you have probably for years learned and memorised that the circumference of a circle of radius $R > 0$ is $2\pi R$. However, you may not been taught *why* this formula holds. Now that we have properly defined arc length, we can use it to justify this formula:

Example 2.18. *Fix $R > 0$, and consider the parametric curve*

$$\gamma : (0, 2\pi) \to \mathbb{R}^2, \qquad \gamma(t) = (R \cos t, R \sin t).$$

Note that γ parametrises (one revolution of) a circle of radius R about the origin.

A direct computation yields the speed of γ:

$$\gamma'(t) = (-R \sin t, R \cos t), \qquad |\gamma'(t)| = \sqrt{R^2 \sin^2 t + R^2 \cos^2 t} = R.$$

Consequently, using (2.20)*, we obtain the familiar formula:*

$$\mathsf{L}(\gamma) = \int_0^{2\pi} |\gamma'(t)| dt = R \int_0^{2\pi} dt = 2\pi R.$$

We had previously advertised that arc length is a geometric property of curves. Let us now explore this in detail, beginning with a return to an example from the beginning of this chapter.

Example 2.19. *Let us first return to the parametric curves from Example 2.2 (see Figure 2.2),*

$$\gamma_1 : (0, \pi) \to \mathbb{R}^2, \qquad \gamma_1(t) = (\cos t, \sin t),$$
$$\gamma_2 : (-1, 1) \to \mathbb{R}^2, \qquad \gamma_2(t) = (-t, \sqrt{1-t^2}),$$

which trace out the upper half C_+ *of the unit circle. We previously indicated in Example 2.11 that the length of* C_+ *should be* π *regardless of how* C_+ *is parametrised. Let us now test this assertion.*

The length of γ_1 *can be computed like in Example 2.18. We first calculate*

$$|\gamma_1'(t)| = |(-\sin t, \cos t)| = 1.$$

Integrating the above over the domain $(0, \pi)$ *of* γ_1 *yields its arc length:*

$$\mathsf{L}(\gamma_1) = \int_0^{\pi} |\gamma_1'(t)| dt = \int_0^{\pi} dt = \pi.$$

The computation for γ_2 *is a bit tricker, but the process is similar. First, observe that*

$$\gamma_2'(t) = \left(-1, \frac{1}{2} \cdot \frac{-2t}{\sqrt{1-t^2}}\right) = \left(-1, -\frac{t}{\sqrt{1-t^2}}\right),$$
$$|\gamma_2'(t)| = \sqrt{1 + \frac{t^2}{1-t^2}} = \sqrt{\frac{(1-t^2)+t^2}{1-t^2}} = \frac{1}{\sqrt{1-t^2}}.$$

To integrate $|\gamma_2'(t)|$*, we* ~~*Google the answer*~~ *apply a clever trigonometric substitution,*

$$t = -\cos u, \qquad dt = \sin u \cdot du,$$

and hence obtain

$$\mathsf{L}(\gamma_2) = \int_{-1}^{1} \frac{1}{\sqrt{1-t^2}} dt = \int_0^{\pi} \frac{1}{\sqrt{1-\cos^2 u}} \cdot \sin u \cdot du = \int_0^{\pi} du = \pi.$$

Indeed, we obtain the same arc length π *as for* γ_1*.*

You may have noticed that the trigonometric substitution $t = -\cos u$ in Example 2.19 was the same as the change of variables (2.8) that linked γ_1 and γ_2. This is, of course, no accident, and it plays a major role in establishing general parametrisation-independence for the arc length:

Theorem 2.6. *If* $\gamma : (a, b) \to \mathbb{R}^n$ *and* $\tilde{\gamma} : (\tilde{a}, \tilde{b}) \to \mathbb{R}^n$ *are regular parametrisations of the same finite curve, then* $\mathsf{L}(\gamma) = \mathsf{L}(\tilde{\gamma})$*. In other words, arc length is a geometric property of curves.*

Proof. Since γ is reparametrisation of $\tilde{\gamma}$, there is a change of variables $\phi : (a, b) \to (\tilde{a}, \tilde{b})$ with

$$\gamma(t) = \tilde{\gamma}(\phi(t)), \qquad t \in (a, b).$$

By Theorem 2.3, we have

$$|\gamma'(t)| = |\phi'(t)||\tilde{\gamma}'(\phi(t))|$$

Thus, using Definition 2.11 and the above yields

$$L(\gamma) = \int_a^b |\gamma'(t)| dt = \int_a^b |\tilde{\gamma}'(\phi(t))||\phi'(t)| dt.$$

We now apply the substitution $\tilde{t} = \phi(t)$ and $d\tilde{t} = \phi'(t) \cdot dt$. (Some care has to be taken regarding orientation; in particular, the orientation is reversed when $\phi' < 0$.) From this, we see that

$$L(\gamma) = \begin{cases} \int_{\tilde{a}}^{\tilde{b}} |\tilde{\gamma}'(\tilde{t})| d\tilde{t} & \phi' > 0 \\ \int_{\tilde{b}}^{\tilde{a}} |\tilde{\gamma}'(\tilde{t})| (-d\tilde{t}) & \phi' < 0 \end{cases} = \int_{\tilde{a}}^{\tilde{b}} |\tilde{\gamma}'(\tilde{t})| d\tilde{t} = L(\tilde{\gamma}). \qquad \square$$

One consequence of Theorem 2.6 is that it makes sense to speak of the arc length of a curve C—that is, $L(C)$—rather than just the arc length of a parametric curve.

2.4.2. *Unit Speed Parametrisations.* Recall that for any curve, one can find (infinitely) many different parametrisations of it. When studying the geometry of curves, this can be positively viewed as the freedom to choose a parametrisation that best suits us. One such candidate parametrisation, which also has close ties to the geometric properties, is the following:

Definition 2.12. *A parametric curve $\gamma : I \to \mathbb{R}^n$ is called a unit-speed parametrisation iff*

$$(2.21) \qquad |\gamma'(s)| = 1, \qquad s \in I.$$

If γ represents the trajectory of a particle, then γ being a unit-speed parametrisation means that the particle is travelling along its path at a constant, unit speed. One simple example of this is the standard polar parametrisation of the unit circle:

Example 2.20. *Let γ be the parametric curve*

$$\gamma : \mathbb{R} \to \mathbb{R}^n, \qquad \gamma(s) = (\cos s, \sin s),$$

which represents the unit circle. Recall from Example 2.6 that

$$\gamma'(s) = (-\sin s, \cos s), \qquad |\gamma'(s)| = 1.$$

Thus, γ is also a unit-speed parametrisation of the unit circle.

Remark. As a general convention, we will use "s" to refer to unit-speed parameters.

We had previously, in Definition 2.11, discussed how to compute the arc length of a parametric curve $\gamma : (a, b) \to \mathbb{R}^n$. Similarly, if we wanted to compute the arc length of only a portion of this curve, say from $\gamma(t_0)$ to $\gamma(t_1)$ (with $t_0 < t_1$), then we would integrate

$$(2.22) \qquad L(\gamma; t_0, t_1) = \int_{t_0}^{t_1} |\gamma'(t)| dt.$$

Note that the integrand is the same as in (2.20), while the limits t_0 and t_1 of integration mean that we are precisely capturing the portion of γ with parameter t between t_0 and t_1.

Now, the formula for (2.22) becomes considerably simpler for unit-speed parametrisations:

Proposition 2.7. *If the parametric curve $\gamma : (a, b) \to \mathbb{R}^n$ is a unit-speed parametrisation, then*

$$L(\gamma; s_0, s_1) = s_1 - s_0, \qquad a \leq s_0 < s_1 \leq b. \tag{2.23}$$

Proof. Since γ has unit speed, (2.22) immediately yields

$$L(\gamma; s_0, s_1) = \int_{s_0}^{s_1} |\gamma'(s)| ds = \int_{s_0}^{s_1} ds = s_1 - s_0. \qquad \square$$

Observe that Proposition 2.7 states that for a unit-speed parametrisation, its parameter s corresponds precisely to the length travelled along the curve. As a result of this, unit-speed parametrisations are also often called arc length parametrisations.

Intuitively, this result should not be surprising. For instance, Proposition 2.7 can be interpreted as follows: if you move along a path (γ) at a speed of 1 metre per second ($|\gamma'(s)| = 1$), then after 12 seconds ($s_1 - s_0 = 12$), you have travelled exactly 12 metres ($L(\gamma; s_1, s_0) = 12$).

Example 2.21. *For the unit-speed parametrisation of the unit circle in Example 2.20,*

$$\gamma : \mathbb{R} \to \mathbb{R}^n, \qquad \gamma(s) = (\cos s, \sin s),$$

the parameter s corresponds to the polar angle (in radians) of the point on the circle. Thus, if you start on the point of the circle at angle θ_0 and move anticlockwise along the circle to the point with angle θ_1 (see Figure 2.17), then the total distance you travelled is

$$L(\gamma; \theta_0, \theta_1) = \theta_1 - \theta_0.$$

In particular, if you walked around the entire circle once ($\theta_1 - \theta_0 = 2\pi$), then the total distance you travelled would be 2π, i.e. the circumference of the circle.

We have now discussed the meaning of unit-speed parametrisations and their connections to arc length. The next, more practical question, is the following:

Question 2.7. *Given a curve, how can we find a unit speed parametrisation of it?*

The answer, roughly, is to reverse the logic we have already discussed. Recall that for a unit-speed parametrisation, the parameter itself corresponds to the distance travelled. The strategy, then, is to instead define the parameter by the length travelled and then show that this has unit speed. In other words, if after s seconds, for any s, you have travelled s metres, then you must have been moving at a speed of 1 metre per second all along.

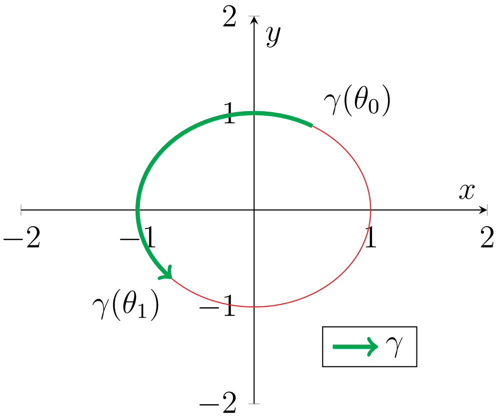


FIGURE 2.17. The green curve is an arc of the unit circle, ranging from angle $\theta_0 = \frac{\pi}{3}$ to $\theta_1 = \frac{5\pi}{4}$.

Theorem 2.8. *Let $\gamma : (a, b) \to \mathbb{R}^n$ be a regular parametric curve. Then, there exists a reparametrisation $\tilde{\gamma}$ of γ such that $\tilde{\gamma}$ has unit speed. Furthermore, $\tilde{\gamma}$ can be defined as*

$$\tilde{\gamma} : (0, L(\gamma)) \to \mathbb{R}^n, \qquad \tilde{\gamma}(s) = \gamma(t), \qquad s = \phi(t) = L(\gamma; a, t). \tag{2.24}$$

Proof. Let us define ϕ as in (2.24). Recalling (2.22) and the fundamental theorem of calculus,

$$\phi'(t) = \frac{d}{dt} L(\gamma; a, t) = \frac{d}{dt} \int_a^t |\gamma'(\tau)| d\tau = |\gamma'(t)|.$$

Since γ is regular, the above implies that ϕ' is always positive, and hence ϕ always has a (smooth, bijective) inverse ϕ^{-1}. We can now define our reparametrisation $\tilde{\gamma}$ by

$$\tilde{\gamma} : (0, L(\gamma)) \to \mathbb{R}^n, \qquad \tilde{\gamma}(s) = \gamma(\phi^{-1}(s)).$$

(Note that $0 < s < L(\gamma)$ corresponds exactly to $a < t < b$.)

Observe that the above is precisely (2.24). Thus, it remains only to show that $\tilde{\gamma}$ is a unit speed parametrisation. For this, we start with (2.24) and recall Theorem 2.3:

$$|\tilde{\gamma}'(s)| = |\tilde{\gamma}'(\phi(t))| = \frac{1}{|\phi'(t)|} |\gamma'(t)| = \frac{|\gamma'(t)|}{|\gamma'(t)|} = 1. \qquad \square$$

Example 2.22. *Consider the parametric line segment (see the first part of Figure 2.18)*

$$\ell : (0, 1) \to \mathbb{R}^3, \qquad \ell(t) = (2t, -t, 4t).$$

We now use Theorem 2.8 to find a unit-speed reparametrisation of ℓ.

First, we calculate the speed of ℓ:

$$\ell'(t) = (2, -1, 4), \qquad |\ell'(t)| = \sqrt{2^2 + 1^2 + 4^2} = \sqrt{21}.$$

With $|\ell'|$, we can compute the "distance travelled thus far" on ℓ:

$$L(\ell; 0, t) = \int_0^t |\ell'(\tau)| d\tau = \sqrt{21} \int_0^t d\tau = \sqrt{21} \cdot t.$$

By Theorem 2.8, we obtain the unit-speed parameter by

$$s = \phi(t) = L(\ell; 0, t) = \sqrt{21} \cdot t, \qquad t = \frac{1}{\sqrt{21}} \cdot s.$$

Thus, the unit-speed reparametrisation is given by

$$\tilde{\ell} : (0, L(\ell)) = (0, \sqrt{21}) \to \mathbb{R}^3, \qquad \tilde{\ell}(s) = \ell(t) = \ell\left(\frac{1}{\sqrt{21}} \cdot s\right) = \frac{1}{\sqrt{21}}(2s, -s, 4s).$$

Example 2.23. *Let γ denote the logarithmic spiral,*

$$\gamma : (0, \infty) \to \mathbb{R}^2, \qquad \gamma(t) = (\cos t \cdot e^t, \sin t \cdot e^t).$$

A plot of γ is given in the second part of Figure 2.18.

Again, we begin by computing the speed of γ:

$$\gamma'(t) = (\cos t \cdot e^t - \sin t \cdot e^t, \sin t \cdot e^t + \cos t \cdot e^t) = e^t(\cos t - \sin t, \sin t + \cos t),$$
$$|\gamma'(t)| = e^t \sqrt{(\cos t - \sin t)^2 + (\sin t + \cos t)^2} = e^t \sqrt{2\cos^2 t + 2\sin^2 t} = \sqrt{2} \cdot e^t.$$

The "distance travelled thus far" on γ is then

$$L(\gamma; 0, t) = \int_0^t |\gamma'(\tau)| d\tau = \sqrt{2} \int_0^t e^\tau d\tau = \sqrt{2}(e^t - 1).$$

We now apply Theorem 2.8 to find the unit-speed parameter:

$$s = \phi(t) = \sqrt{2}(e^t - 1), \qquad e^t = \frac{1}{\sqrt{2}} \cdot s + 1, \qquad t = \ln\left(\frac{1}{\sqrt{2}} \cdot s + 1\right).$$

Thus, the unit-speed reparametrisation is given by

$$\begin{aligned} \tilde{\gamma} &: (0, L(\gamma; 0, \infty)) = (0, \infty) \to \mathbb{R}^n, \\ \tilde{\gamma}(s) &= \gamma(t) \\ &= \gamma\left(\ln\left(\frac{s}{\sqrt{2}} + 1\right)\right) \\ &= \left(\frac{s}{\sqrt{2}} + 1\right)\left(\cos\ln\left(\frac{s}{\sqrt{2}} + 1\right), \sin\ln\left(\frac{s}{\sqrt{2}} + 1\right)\right). \end{aligned}$$

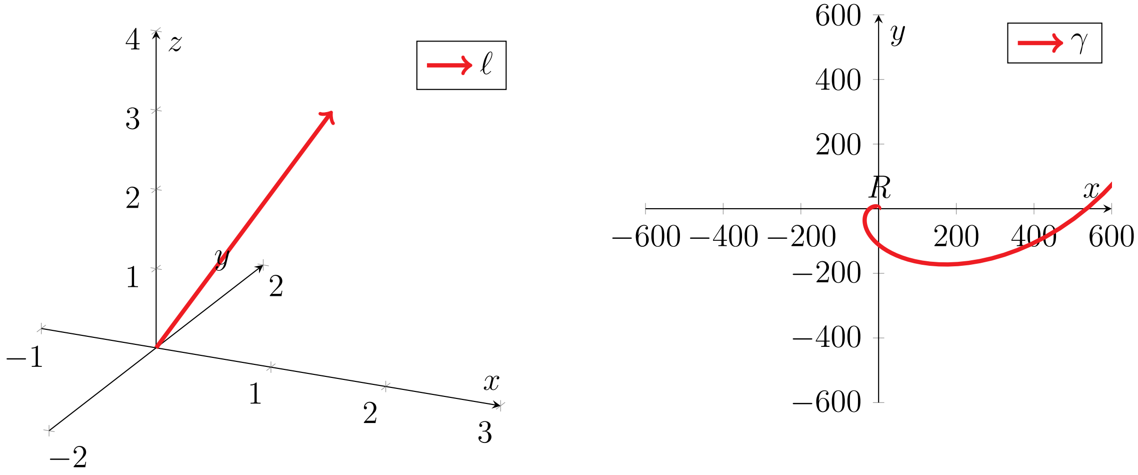


FIGURE 2.18. The left plot contains the segment ℓ from Example 2.22, while the right plot contains part of the logarithmic spiral γ from Example 2.23.

Unit-speed parametrisations are natural to consider, since they tie the parameter to a geometric property—the arc length. On the other hand, unit-speed parametrisations can in general be impractical to compute and usually cannot be evaluated explicitly in terms of elementary functions. (In fact, examples such as the spiral in Example 2.23 were very carefully chosen so that the resulting arc length integral could be fully expanded.) As a result, unit-speed parametrisations are usually more useful for theoretical purposes than for practical computations.

2.4.3. *Path Integrals.* As you know from first year calculus, in flat settings, (single) integrals are closely tied to the notion of length. More specifically, given an interval (a, b) in $\mathbb{R}$, the integral

$$\int_a^b 1 \cdot dx = b - a.$$

gives the length of this interval. More generally, given a function $f : (a, b) \to \mathbb{R}$, we can interpret

$$\int_a^b f(x) dx$$

as a "weighted length" of the interval (a, b), where the "weight" is given by the integrand f.

The term "weighted length" is left deliberately vague, as it can have many different meanings in various contexts. One example that you have likely encountered in calculus is the following:

Example 2.24. *Suppose $f : (a, b) \to \mathbb{R}$ is everywhere positive, and suppose we want to compute the area of the region R lying between the graph of f and the x-axis, i.e. "the area under the curve":*

$$R = \{(x, y) \in \mathbb{R}^2 \mid x \in (a, b),\ 0 < y < f(x)\}.$$

(See Figure 2.19 for a graphical representation.) In calculus, you learned that the area of R is

$$A(R) = \int_a^b f(x) dx.$$

One way to interpret this area integral is as a "weighted length". In other words, we can think of this area as a "length" of the interval (a, b), except that at each $x \in (a, b)$, we assign the height $f(x)$ of f as a weight. As a result, points at which f is higher count more toward this "weighted length" than points at which f is lower, as we would expect from intuition.

Such "weighted lengths" also appear in physics. One simple example is as follows:

Example 2.25. *Let the interval (a, b) represent a rod. Consider a function $f : (a, b) \to \mathbb{R}$, where $f(x)$ represents the mass density (i.e. mass per unit length) of the rod at a position x.*

Suppose we wish to compute the total mass of the rod. The idea is once again to take a "weighted sum" of the mass distributed throughout the rod, which is given by

$$m = \int_a^b f(x) dx.$$

In particular, places along the rod where the mass density f is higher would "count more" toward the mass than places where the mass density is lower.

It is now natural to ask whether these intuitions for intervals could be further extended to curves. In particular, can we define a notion of integration on curves, which:

(1) Is tied to its arc length, and
(2) Generalises the usual integral from calculus?

To explore this, let us fumble around a bit.

Let C be a curve in $\mathbb{R}^n$, and let $\gamma : (a, b) \to \mathbb{R}^n$ be a parametrisation of C. Like for intervals, we should expect the "integral of 1 over C" (which we have yet to define) to yield its arc length:

$$\int_C ds = L(C) = \int_a^b |\gamma'(t)| dt. \tag{2.25}$$

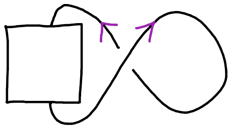


FIGURE 2.19. The setting from Example 2.24; the region R is in green.

Now, for a real-valued function F defined on C, we wish to define its integral over C,

$$\int_C F ds,$$

to represent a "weighted arc length". Looking at the right-hand side of (2.25), it makes sense to insert F into the integrand there, which fits with the intuition of F being a weight.

Moreover, since $|\gamma'(t)|$ represents the speed of γ *at* $\gamma(t)$, it would be sensible that the weight F is also being applied *at the same point* $\gamma(t)$, that is,

$$\int_C F ds = \int_a^b F(\gamma(t))|\gamma'(t)|dt.$$

In fact, this is the formal definition we will use for the integration over C.

Definition 2.13. *Let C be a curve in $\mathbb{R}^n$, and let F be a real-valued function that is defined on the image of C. Then, we define the path integral (or line integral) of F over the curve C by*

$$\int_C F ds = \int_a^b F(\gamma(t))|\gamma'(t)|dt. \tag{2.26}$$

where $\gamma : (a, b) \to \mathbb{R}^n$ is a parametrisation of C.

Before continuing our theoretical discussion, let us first compute a simple example:

Example 2.26. *Let C_+ be the upper half of the unit circle (see Example 2.2), and let*

$$F : \mathbb{R}^2 \to \mathbb{R}, \qquad F(x, y) = y.$$

Let us now evaluate the path integral

$$\int_{C_+} F ds = \int_{C_+} y ds.$$

According to Definition 2.13, the first step is to choose a parametrisation of C_+. Here, we choose

$$\gamma_1 : (0, \pi) \to \mathbb{R}^2, \qquad \gamma_1(t) = (\cos t, \sin t).$$

Recall that the speed of γ_1 is

$$|\gamma_1'(t)| = |(-\sin t, \cos t)| = 1,$$

while F, evaluated at $\gamma_1(t)$, is

$$F(\gamma_1(t)) = F(\cos t, \sin t) = \sin t.$$

Thus, from Definition 2.13, we compute

$$\int_{C_+} F ds = \int_0^\pi F(\gamma_1(t))|\gamma_1'(t)|dt = \int_0^\pi \sin t \cdot 1 \cdot dt = -\cos t|_0^\pi = 2.$$

In defining the path integral as in Definition 2.13, we have missed one very crucial detail. In (2.26), the path integral is expressed in terms of some chosen parametrisation γ of the given curve C. However, if we were to take another parametrisation $\tilde{\gamma}$ of C, we could hypothetically obtain a different value for the integral. If this was the case, then our definition no longer makes sense, since it would allow for multiple values associated to our integral.

The following theorem saves the day by showing that this dangerous scenario cannot hold. In other words, the path integral, as defined in (2.26), is independent of parametrisation.

Theorem 2.9. *Assume the setting of Definition 2.13. Then, the right-hand side of* (2.26) *does not depend on the chosen parametrisation* γ *of* C*.*

Proof. The proof is a straightforward generalisation of the proof of Theorem 2.6 (for the independence of parametrisation of the arc length). Suppose $\tilde{\gamma} : (\tilde{a}, \tilde{b}) \to \mathbb{R}^n$ is a reparametrisation of γ, and let $\phi : (a, b) \to (\tilde{a}, \tilde{b})$ be the change of variables satisfying

$$\gamma(t) = \tilde{\gamma}(\phi(t)), \qquad t \in (a, b).$$

Since Theorem 2.3 implies that $|\gamma'(t)| = |\phi'(t)||\tilde{\gamma}'(\phi(t))|$, then

$$\int_a^b F(\gamma(t))|\gamma'(t)|dt = \int_a^b F(\tilde{\gamma}(\Phi(t)))|\tilde{\gamma}'(\phi(t))||\phi'(t)|dt.$$

Apply the substitution $\tilde{t} = \phi(t)$ and $d\tilde{t} = \phi'(t) \cdot dt$, we see that

$$\begin{aligned} \int_a^b F(\gamma(t))|\gamma'(t)|dt &= \begin{cases} \int_{\tilde{a}}^{\tilde{b}} F(\tilde{\gamma}(\tilde{t}))|\tilde{\gamma}'(\tilde{t})|d\tilde{t} & \phi' > 0 \\ \int_{\tilde{b}}^{\tilde{a}} F(\tilde{\gamma}(\tilde{t}))|\tilde{\gamma}'(\tilde{t})|(-d\tilde{t}) & \phi' < 0 \end{cases} \\ &= \int_{\tilde{a}}^{\tilde{b}} F(\tilde{\gamma}(\tilde{t}))|\tilde{\gamma}'(\tilde{t})|d\tilde{t}, \end{aligned}$$

which is the desired parametrisation-independence. □

In particular, since path integrals are generally independent of parametrisations, then any path integral of a fixed function produces a geometric property of curves.

Example 2.27. *Returning to the upper half circle* C_+ *from Example 2.26, we now try computing the same path integral using another parametrisation of* C_+*,*

$$\gamma_2 : (-1, 1) \to \mathbb{R}^2, \qquad \gamma_2(t) = (-t, \sqrt{1 - t^2}).$$

In Example 2.19, we computed the speed of γ_2*:*

$$|\gamma_2'(t)| = \frac{1}{\sqrt{1 - t^2}}.$$

Evaluating F *at* $\gamma_2(t)$*, we obtain*

$$F(\gamma_2(t)) = F(-t, \sqrt{1 - t^2}) = \sqrt{1 - t^2}.$$

Thus, Definition 2.13 applied to γ_2 *yields*

$$\int_{C_+} F ds = \int_{-1}^1 F(\gamma_2(t))|\gamma_2'(t)|dt = \int_{-1}^1 \sqrt{1 - t^2} \cdot \frac{1}{\sqrt{1 - t^2}} \cdot dt = \int_{-1}^1 dt = 2.$$

In particular, we obtain the same answer as in Example 2.26.

You may now ask: *which parametrisation of a curve should you choose when computing a path integral?* The good news is that since you get the same answer regardless of your choice, the best course of action is to use the parametrisation that results in the simplest computations. In other words, you have the freedom to be as lazy as possible.

Example 2.28. *Next, we compute the path integral*

$$\int_L (y+z)\,ds,$$

where L *is the curve described by the parametric line from Example 2.22,*

$$\ell : (0,1) \to \mathbb{R}^3, \qquad \ell(t) = (2t, -t, 4t).$$

In Example 2.22, we already showed that

$$|\ell'(t)| = \sqrt{21}.$$

Moreover, for the function $F(x,y,z) = y+z$*, we have*

$$F(\ell(t)) = F(2t, -t, 4t) = -t + 4t = 3t.$$

Consequently, by Definition 2.13,

$$\int_L (y+z)\,ds = \int_0^1 F(\ell(t))|\ell'(t)|\,dt = \sqrt{21}\int_0^1 (3t)\,dt = \frac{3\sqrt{21}}{2}.$$

How might one interpret a path integral along a curve? Again, the answers are as diverse as for the usual integrals on intervals. Indeed, essentially any interpretation for standard integrals has a direct generalisation to path integrals on curves:

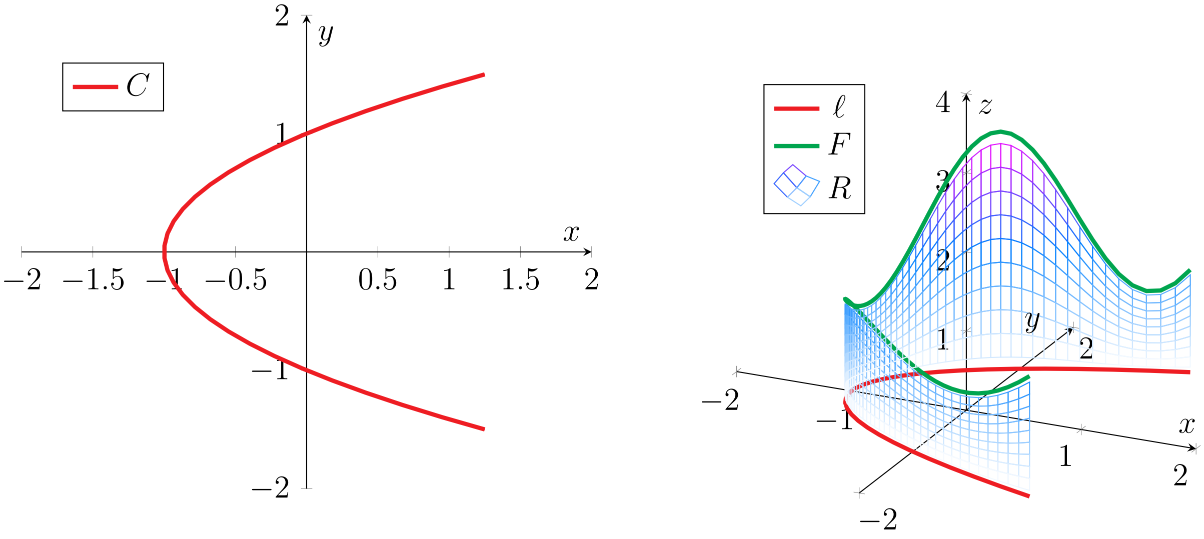


FIGURE 2.20. The setting from Example 2.30. The left plot shows the curve C in $\mathbb{R}^2$, while the right plot shows the region R in 3-dimensional space that is lying between C and the graph of F.

Example 2.29. *Suppose the rod from Example 2.25 is now curved; let us model this rod as a curve* C*. If* F *represents the mass density along the rod, then its total mass would be*

$$\int_C F\,ds.$$

Example 2.30. *Let* C *be a curve in the plane* $\mathbb{R}^2$ *(see the first part of Figure 2.20), and let* F *be a positive function on* C*. Similar to Example 2.24, suppose we wish to compute "the area under the graph of* F*", that is, the area of the region (see the second part of Figure 2.20)*

$$R = \{(x, y, z) \in \mathbb{R}^3 \mid (x, y) \in C,\ 0 < z < F(x, y)\}.$$

between C *and the graph of* F*. Then, the area of* R *can be expressed as*

$$A(R) = \int_C F ds.$$

3. How is a Curve Curved?

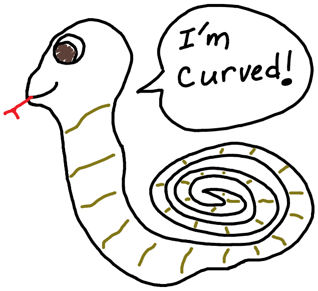
In Chapter 2, we explored how curves can be described, in particular through parametrisations. We also studied some geometric properties, such as tangent lines and arc length, and we discussed their connections to differential and integral calculus. In this chapter, we continue studying the geometry of curves by addressing the following fundamental question:

Question 3.1. *How can we describe how "curved" a curve is? How can we quantify this curvature?*

We will open this chapter by discussing some general ideas around curvature and how it could be measured. In subsequent sections, we will explore in more detail the special cases of plane curves (i.e. curves in $\mathbb{R}^2$) and space curves (i.e. curves in $\mathbb{R}^3$).

3.1. General Ideas.

To answer Question 3.1, we begin with some intuitions. If asked to think of something curved, then you may visualise a piece of string or rope that is bent, or you may picture a snake that is wound up in a coil. In more abstract settings, you might then think of curvature as a curve that is also, in some sense, "bending".



If you have a more negative outlook, then you might also think of "curved" or "bent" as the opposite of what a straight line would be. The defining feature of straight lines is that no matter where you are along the line, it is always maintains the same direction. In contrast, we can then characterise "curving" or "bending" as the change in the direction of a curve. The more that the direction is changing, the more "curved" a curve is said to be.

Luckily for us, "direction" is a concept that we have defined quite precisely. Therefore, with the ideas presented above, we are now in a position to derive a geometric definition of curvature.

3.1.1. *Defining Curvature.*

Let γ be a regular parametric curve. Recall that $\gamma'(t)$ represents the velocity, or the rate of change of the position, of γ at parameter t. Moreover, if we filter out the length of $\gamma'(t)$, that is, we consider the unit vector $|\gamma'(t)|^{-1}\gamma'(t)$, then we are left only with the *direction* of γ at t. Since our idea is to characterise curvature as the change of direction, this suggests that we should measure the derivative of the direction:

$$\left| \frac{d}{dt} \left[\frac{\gamma'(t)}{|\gamma'(t)|} \right] \right| . \tag{3.1}$$

While this seems like a reasonable definition for curvature, there is one subtle but important issue. Since curvature should capture a geometric property of curves, we require that a satisfactory definition of curvature must be independent of parametrisations. However, the outer derivative $\frac{d}{dt}$ in the equation (3.1) seems to violate this criterion. (On the other hand, the direction $|\gamma'(t)|^{-1}\gamma'(t)$ is parametrisation-independent up to orientation—can you see why?)

The idea now is to replace this $\frac{d}{dt}$ by a derivative that is more geometric in nature. For this, we recall that arc length is a geometric property. Thus, it would be more sensible to differentiate with respect to an arc length parameter s. In other words, we attempt to define the curvature as

$$\left| \frac{d}{ds} \left[\frac{\gamma'(t)}{|\gamma'(t)|} \right] \right| = \left| \frac{dt}{ds} \cdot \frac{d}{dt} \left[\frac{\gamma'(t)}{|\gamma'(t)|} \right] \right| . \tag{3.2}$$

Now, we are almost in good shape (partial pun not intended); we only need to rewrite this $\frac{d}{ds}$ in terms of something that we can work with. From the relations in Theorem 2.8 between the arc length parameter s and an arbitrary parameter t, we (rather informally) derive that

$$\frac{ds}{dt} = \frac{d}{dt}[L(\gamma; t_0, t)] = |\gamma'(t)|, \qquad \frac{dt}{ds} = \frac{1}{|\gamma'(t)|},$$

where t_0 represents any chosen "starting" parameter of γ. Substituting the above formula for $\frac{dt}{ds}$ into (3.2), we arrive at the formal definition of curvature:

Definition 3.1. *Let $\gamma : I \rightarrow \mathbb{R}^n$ be a regular parametric curve, and let $t \in I$. We define the* <u>*curvature*</u> *of γ at the point $\gamma(t)$ (or alternatively, at the parameter t) as*

$$\kappa|_{\gamma(t)} = \frac{1}{|\gamma'(t)|} \left| \frac{d}{dt} \left[\frac{\gamma'(t)}{|\gamma'(t)|} \right] \right| . \tag{3.3}$$

Next, we establish that Definition 3.1 is indeed independent of parametrisation, i.e. that curvature is a geometric property. Intuitively, the main idea is that both the direction $|\gamma'(t)|^{-1}\gamma'(t)$ and the outer derivative $\frac{d}{ds} = |\gamma'(t)|^{-1} \frac{d}{dt}$ in (3.3) depend only on the orientation, through a factor of ± 1. The absolute values in (3.3) then do away with this factor.

Theorem 3.1. *Assuming the setting of Definition 3.1, then the right hand side of* (3.3) *is independent of parametrisation. More specifically, if $\tilde{\gamma} : \tilde{I} \rightarrow \mathbb{R}^n$ is a reparametrisation of γ, and if $\phi : I \rightarrow \tilde{I}$ is the change of variables satisfying $\gamma(t) = \tilde{\gamma}(\phi(t))$ for all $t \in I$, then*

$$\frac{1}{|\gamma'(t)|} \left| \frac{d}{dt} \left[\frac{\gamma'(t)}{|\gamma'(t)|} \right] \right| = \frac{1}{|\tilde{\gamma}'(\tilde{t})|} \left| \frac{d}{d\tilde{t}} \left[\frac{\tilde{\gamma}'(\tilde{t})}{|\tilde{\gamma}'(\tilde{t})|} \right] \right| , \qquad \tilde{t} = \phi(t). \tag{3.4}$$

Proof. Using the relations from Theorem 2.3, we have that

$$\tilde{\gamma}'(\tilde{t}) = \tilde{\gamma}'(\phi(t)) = \frac{1}{\phi'(t)} \cdot \gamma'(t), \qquad |\tilde{\gamma}'(\tilde{t})| = \frac{|\gamma'(t)|}{|\phi'(t)|}.$$

Furthermore, observe that

$$\frac{d}{d\tilde{t}} = \frac{dt}{d\tilde{t}} \cdot \frac{d}{dt} = \frac{1}{\phi'(t)} \cdot \frac{d}{dt}.$$

Combining the above and noting that

$$\frac{|\phi'(t)|}{\phi'(t)} = \pm 1,$$

we obtain the desired identity (3.4):

$$\begin{aligned} \frac{1}{|\tilde{\gamma}'(\tilde{t})|} \left| \frac{d}{d\tilde{t}} \left[\frac{\tilde{\gamma}'(\tilde{t})}{|\tilde{\gamma}'(\tilde{t})|} \right] \right| &= \frac{|\phi'(t)|}{|\gamma'(t)|} \left| \frac{1}{\phi'(t)} \cdot \frac{d}{dt} \left[\frac{|\phi'(t)|}{\phi'(t)} \cdot \frac{\gamma'(t)}{|\gamma'(t)|} \right] \right| \\ &= \frac{1}{|\gamma'(t)|} \left| \frac{d}{dt} \left[\frac{\gamma'(t)}{|\gamma'(t)|} \right] \right| . \end{aligned}$$

□

One particular consequence of Theorem 3.1 is that it makes sense to speak of the curvature of a curve, not only of a parametric curve. Furthermore, when computing the curvature, we have the freedom to choose a pleasant parametrisation that makes the calculations as simple as possible.

Finally, we note that for unit-speed parametrisations, the formula (3.3) simplifies considerably:

Corollary 3.2. *If $\gamma : I \to \mathbb{R}^n$ is a unit-speed parametrisation, then*

$$\kappa|_{\gamma(s)} = |\gamma''(s)|, \qquad s \in I. \tag{3.5}$$

Proof. For this, we simply apply (3.3) and use that $|\gamma'(s)| = 1$:

$$\kappa|_{\gamma(s)} = \frac{1}{|\gamma'(s)|} \left| \frac{d}{ds} \left[\frac{\gamma'(s)}{|\gamma'(s)|} \right] \right| = \left| \frac{d}{ds} [\gamma'(s)] \right| = |\gamma''(s)|. \qquad \square$$

3.1.2. *First Examples.* In Definition 3.1, we put forward a definition of curvature for curves. But, is this definition reasonable? Thus far, our arguments leading up to this definition seem to indicate an affirmative answer. Here, we test Definition 3.1 further by applying it to some basic examples in order to check whether we obtain the expected answers.

Example 3.1. *Consider an arbitrary parametric line in $\mathbb{R}^n$,*

$$\ell : \mathbb{R} \to \mathbb{R}^n, \qquad \ell(t) = t\mathbf{v} + \mathbf{p},$$

where $\mathbf{v}, \mathbf{p} \in \mathbb{R}^n$ are fixed vectors. Observe that ℓ describes precisely the line passing through the point $\mathbf{p}$ (when $t = 0$) and moving in the direction of $\mathbf{v}$ (which is equal to $\ell'(t)$).

To compute the curvature of ℓ, we first calculate

$$\ell'(t) = \mathbf{v}, \qquad |\ell'(t)| = |\mathbf{v}|, \qquad \frac{\ell'(t)}{|\ell'(t)|} = \frac{\mathbf{v}}{|\mathbf{v}|}.$$

Since all three of the above are constant, we see from Definition 3.1 that

$$\kappa|_{\ell(t)} = \frac{1}{|\ell'(t)|} \left| \frac{d}{dt} \left[\frac{\ell'(t)}{|\ell'(t)|} \right] \right| = \frac{1}{|\mathbf{v}|} \left| \frac{d}{dt} \left(\frac{\mathbf{v}}{|\mathbf{v}|} \right) \right| = 0.$$

Example 3.1 shows that, as expected, a straight line has zero curvature. Intuitively, we expect an even stronger statement—we also expect the converse to also hold. Indeed, if a curve has no curvature, then it should be straight, hence a line. We prove precisely this in the following theorem:

Theorem 3.3. *Let C be a curve in $\mathbb{R}^n$. Then, C is a line if and only if C has zero curvature.*

Proof. Example 3.1 already showed that a line has zero curvature, hence we need only show the converse. Suppose C has zero curvature; our goal is to show that C is a line.

By Theorem 2.8, we can find a unit-speed parametrisation $\tilde{\gamma}$ of C. Corollary 3.2 then implies

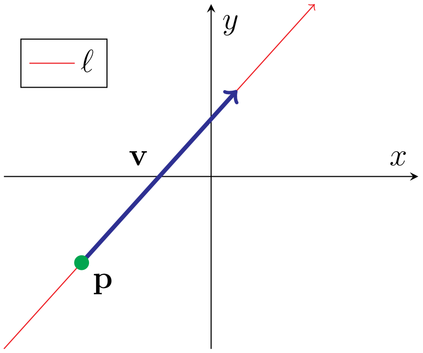


FIGURE 3.1. An instance of ℓ (in red) from Example 3.1. Here, $\mathbf{p}$ is drawn in green, while $\mathbf{v}$ is shown in blue.

$$0 = \kappa|_{\tilde{\gamma}(s)} = |\tilde{\gamma}''(s)|, \qquad \tilde{\gamma}''(s) = \mathbf{0}.$$

To solve for $\tilde{\gamma}$, we integrate the above vector equation twice (component-wise) to obtain

$$\tilde{\gamma}'(s) = \mathbf{v}, \qquad \tilde{\gamma}(s) = s\mathbf{v} + \mathbf{p},$$

for some fixed vectors $\mathbf{v}, \mathbf{p} \in \mathbb{R}^n$. Since this $\tilde{\gamma}$ is linear, C must indeed be a line. □

Next, we investigate another fundamental shape: circles.

Example 3.2. *Let C denote the circle in $\mathbb{R}^2$ of radius $R > 0$ about the origin (see Figure 2.16). One way to parametrise C is through the parametric curve*

$$\gamma : \mathbb{R} \to \mathbb{R}^2, \qquad \gamma(t) = (R\cos t, R\sin t).$$

To compute the curvature of C, we observe that

$$\gamma'(t) = (-R\sin t, R\cos t), \qquad |\gamma'(t)| = R, \qquad \frac{\gamma'(t)}{|\gamma'(t)|} = (-\sin t, \cos t).$$

Applying Definition 3.1 to the above yields

$$\kappa|_{\gamma(t)} = \frac{1}{|\gamma'(t)|}\left|\frac{d}{dt}\left[\frac{\gamma'(t)}{|\gamma'(t)|}\right]\right| = \frac{1}{R}\left|\frac{d}{dt}(-\sin t, \cos t)\right| = \frac{1}{R}|(\cos t, \sin t)| = \frac{1}{R}.$$

Thus, we obtain that the circle C has constant curvature equal to the inverse of its radius.

Let us understand this computation more intuitively. First, note that any point of the circle is curved in the same way as any other point; this accounts for the fact that circles have constant curvature. We next ask what accounts for the dependence of the curvature on the radius.

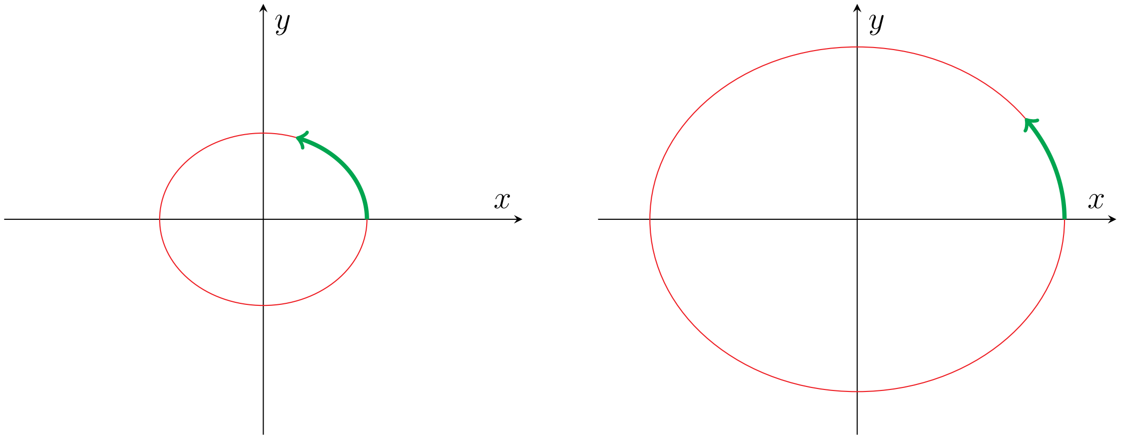


FIGURE 3.2. The plots show two circles of different radii (in red). On each circle, the green arc has length 1. Note that on the larger circle (on the right), the green arc turns less than on the smaller circle.

The key observation comes from the derivative

$$\frac{1}{|\gamma'(t)|} \cdot \frac{d}{dt} = \frac{d}{ds}$$

in the definition (3.3) of curvature. Recall that this represents the *rate of change per unit length.* As the radius of a circle increases, then an arc of unit length accounts for a smaller portion of the circle; see Figure 3.2. In other words, as the radius increases, then a particle traveling along the circle will have turned less after travelling a unit length.

Thus, it makes intuitive sense for larger circles to have smaller curvature at each point. More specifically, one can also see that the amount that this particle has turned after travelling a unit length will be inversely proportional to the radius of the circle. This suggests that κ should be inversely proportional to R, which confirms the answer we obtained from Example 3.2.

3.1.3. *An Expanded Formula.* We conclude this section by deriving an expanded formula for the curvature. Before doing this, however, we first revise some basic facts from vector algebra.

Definition 3.2. *Given two vectors*

$$\mathbf{v} = (v_1, \dots, v_n) \in \mathbb{R}^n, \qquad \mathbf{w} = (w_1, \dots, w_n) \in \mathbb{R}^n,$$

we define their dot product by

$$\mathbf{v} \cdot \mathbf{w} = v_1 w_1 + v_2 w_2 + \cdots + v_n w_n \in \mathbb{R}. \tag{3.6}$$

Also, for two "arrows" $\mathbf{v}|_\mathbf{p}$ and $\mathbf{w}|_\mathbf{p}$ from a common point $\mathbf{p} \in \mathbb{R}^n$, we can similarly define

$$\mathbf{v}|_\mathbf{p} \cdot \mathbf{w}|_\mathbf{p} = \mathbf{v} \cdot \mathbf{w}. \tag{3.7}$$

Remark. Note that while $\mathbf{v}$ and $\mathbf{w}$ are vectors, their dot product $\mathbf{v} \cdot \mathbf{w}$ is a scalar.

While equation (3.6) is usually the most convenient for computations, another formula for dot products is more illuminating for obtaining geometric information:

Proposition 3.4. *For $\mathbf{v}, \mathbf{w}, \mathbf{p}$ as in Definition 3.2, we have that*

$$\mathbf{v}|_\mathbf{p} \cdot \mathbf{w}|_\mathbf{p} = \mathbf{v} \cdot \mathbf{w} = |\mathbf{v}||\mathbf{w}| \cos \theta. \tag{3.8}$$

where θ is the angle made between the "arrows" $\mathbf{v}|_\mathbf{p}$ and $\mathbf{w}|_\mathbf{p}$. In addition,

$$|\mathbf{v}|^2 = \mathbf{v} \cdot \mathbf{v}, \qquad |\mathbf{v}|_\mathbf{p}|^2 = \mathbf{v}|_\mathbf{p} \cdot \mathbf{v}|_\mathbf{p}. \tag{3.9}$$

See Figure 3.3 for a simple diagram describing (3.8). This formula indicates that the dot product contains a mixture of two pieces of geometric information:

1. The lengths of the vectors involved.
2. The angle between the vectors involved.

In the case of (3.9), since there is only one vector involved (hence the angle θ vanishes), the dot product $\mathbf{v} \cdot \mathbf{v}$ carries information only about the length of $\mathbf{v}$.

Next, since parametric curves take values in $\mathbb{R}^n$, we can also consider dot products involving them, which will have similar geometric meaning as was mentioned above. For the moment, we will only require the following identity, which can be proved using the product rule from basic calculus.

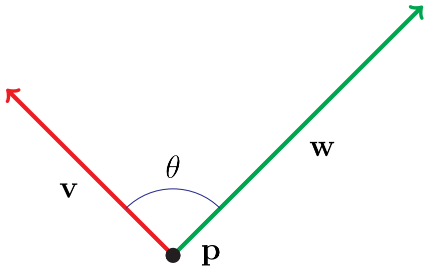


FIGURE 3.3. This diagram demonstrates the setup for (3.8).

Proposition 3.5. *Let* I *be an open interval, and let* $\lambda : I \to \mathbb{R}^n$ *and* $\alpha : I \to \mathbb{R}^n$ *be smooth. Then,*

$$\frac{d}{dt}[\lambda(t) \cdot \alpha(t)] = \lambda'(t) \cdot \alpha(t) + \lambda(t) \cdot \alpha'(t), \qquad t \in I. \tag{3.10}$$

With the background from above, we can now derive a general formula for the curvature:

Proposition 3.6. *Let* $\gamma : I \to \mathbb{R}^n$ *be a regular parametric curve, and let* $t \in I$*. Then,*

$$\frac{d}{dt}\left[\frac{\gamma'(t)}{|\gamma'(t)|}\right] = \frac{[\gamma'(t) \cdot \gamma'(t)]\gamma''(t) - [\gamma'(t) \cdot \gamma''(t)]\gamma'(t)}{|\gamma'(t)|^3}, \tag{3.11}$$
$$\kappa|_{\gamma(t)} = \frac{|[\gamma'(t) \cdot \gamma'(t)]\gamma''(t) - [\gamma'(t) \cdot \gamma''(t)]\gamma'(t)|}{|\gamma'(t)|^4}.$$

Remark. Note that the quantity $[\gamma'(t) \cdot \gamma'(t)]\gamma''(t) - [\gamma'(t) \cdot \gamma''(t)]\gamma'(t)$ in the numerator of the right-hand side of (3.11) is a vector. For example, in the first term of the above, the scalar $\gamma'(t) \cdot \gamma'(t) \in \mathbb{R}$ is multiplied by the vector $\gamma''(t) \in \mathbb{R}^n$; the product is then also a vector.

Proof. First, we apply (3.9) and the power rule to obtain

$$\frac{d}{dt}|\gamma'(t)| = \frac{d}{dt}\sqrt{\gamma'(t) \cdot \gamma'(t)} = \frac{1}{2\sqrt{\gamma'(t) \cdot \gamma'(t)}} \cdot \frac{d}{dt}[\gamma'(t) \cdot \gamma'(t)].$$

The derivative of the dot product can then be expanded using (3.10):

$$\frac{d}{dt}|\gamma'(t)| = \frac{2[\gamma'(t) \cdot \gamma''(t)]}{2\sqrt{\gamma'(t) \cdot \gamma'(t)}} = \frac{\gamma'(t) \cdot \gamma''(t)}{|\gamma'(t)|}. \tag{3.12}$$

Now, from the product and power rules, we obtain

$$\begin{aligned} \frac{d}{dt}\left[\frac{\gamma'(t)}{|\gamma'(t)|}\right] &= \frac{\gamma''(t)}{|\gamma'(t)|} - \gamma'(t)\frac{d}{dt}\left[\frac{1}{|\gamma'(t)|}\right] \\ &= \frac{\gamma''(t)}{|\gamma'(t)|} - \frac{\gamma'(t)\frac{d}{dt}|\gamma'(t)|}{|\gamma'(t)|^2}. \end{aligned}$$

Applying (3.12) and (3.9) to the above now yields

$$\begin{aligned} \frac{d}{dt}\left[\frac{\gamma'(t)}{|\gamma'(t)|}\right] &= \frac{\gamma''(t)}{|\gamma'(t)|} - \frac{[\gamma'(t) \cdot \gamma''(t)]\gamma'(t)}{|\gamma'(t)|^3} \\ &= \frac{[\gamma'(t) \cdot \gamma'(t)]\gamma''(t) - [\gamma'(t) \cdot \gamma''(t)]\gamma'(t)}{|\gamma'(t)|^3}, \end{aligned}$$

which is the first part of (3.11). Finally, applying Definition 3.1, we obtain

$$\kappa|_{\gamma(t)} = \frac{1}{|\gamma'(t)|}\left|\frac{d}{dt}\left[\frac{\gamma'(t)}{|\gamma'(t)|}\right]\right| = \frac{|[\gamma'(t) \cdot \gamma'(t)]\gamma''(t) - [\gamma'(t) \cdot \gamma''(t)]\gamma'(t)|}{|\gamma'(t)|^4}.$$ □

Remark. Our only use for the identity (3.11) is to assist in deriving specialised formulas in the cases of plane and space curves. We will not require (3.11) for any other purposes.

3.2. **Plane Curves.** In this section, we continue our study of curvature, but we specialise to the case of curves on a plane. In other words, we assume throughout that $n = 2$.

3.2.1. *The Curvature Formula.* The first step is to return to the rather unpleasant formula (3.11). We show below that for plane curves, the formula simplifies considerably.

Theorem 3.7. *Let* $\gamma : I \to \mathbb{R}^2$ *be a regular parametric curve, and express* γ *as*

$$\gamma(t) = (x(t), y(t)), \qquad x, y : I \to \mathbb{R}.$$

Then, for any $t \in I$*,*

$$\kappa|_{\gamma(t)} = \frac{|x'(t)y''(t) - y'(t)x''(t)|}{|\gamma'(t)|^3}. \tag{3.13}$$

Proof. For convenience, we drop all instances of "t" from the notations below. Expanding the second formula of (3.11) in terms of x and y, we have that

$$\begin{aligned} \kappa|_{\gamma(t)} &= \frac{|(\gamma' \cdot \gamma')\gamma'' - (\gamma' \cdot \gamma'')\gamma'|}{|\gamma'|^4} \\ &= \frac{|(x'x' + y'y')(x'', y'') - (x'x'' + y'y'')(x', y')|}{|\gamma'|^4} \\ &= \frac{|(y'y'x'' - y'y''x', x'x'y'' - x'x''y')|}{|\gamma'(t)|^4} \\ &= \frac{|(x'y'' - y'x'')(-y', x')|}{|\gamma'|^4} \\ &= \frac{|x'y'' - y'x''| \cdot |(-y', x')|}{|\gamma'|^4}. \end{aligned}$$

Since

$$|(-y', x')| = \sqrt{(y')^2 + (x')^2} = |(x', y')| = |\gamma'|,$$

the desired formula (3.13) follows immediately. □

To make remembering (3.13) easier, we recall the following from linear algebra:

Definition 3.3. *For a* 2×2 *matrix, we define its determinant by*

$$\det \begin{bmatrix} a & b \\ c & d \end{bmatrix} = ad - bc. \tag{3.14}$$

Then, using (3.14), we can rewrite (3.13) as:

$$\kappa|_{\gamma(t)} = \frac{1}{|\gamma'(t)|^3} \left| \det \begin{bmatrix} x'(t) & y'(t) \\ x''(t) & y''(t) \end{bmatrix} \right|. \tag{3.15}$$

Example 3.3. *The curve representing the parabola* $y = x^2$ *can parametrised by*

$$\gamma : \mathbb{R} \to \mathbb{R}^2, \qquad \gamma(t) = (t, t^2).$$

(See the left plot in Figure 3.4.) Let us compute its curvature at each point.

First, we note that

$$\gamma'(t) = (x'(t), y'(t)) = (1, 2t), \qquad \gamma''(t) = (x''(t), y''(t)) = (0, 2),$$

as well as that

$$|\gamma'(t)| = \sqrt{1 + 4t^2}.$$

Now, we simply substitute these values into (3.13)*:*

$$\kappa|_{\gamma(t)} = \frac{|x'(t)y''(t) - y'(t)x''(t)|}{|\gamma'(t)|^3} = \frac{|1 \cdot 2 - 2t \cdot 0|}{(1+4t^2)^{\frac{3}{2}}} = \frac{2}{(1+4t^2)^{\frac{3}{2}}}.$$

Example 3.4. *The "right half" of the hyperbola $x^2 - y^2 = 1$ can be parametrised by*

$$\gamma : \mathbb{R} \to \mathbb{R}^2, \qquad \gamma(t) = (\cosh t, \sinh t).$$

(See the right plot in Figure 3.4.) Again, we compute its curvature at each point.

The first step is to compute its derivatives

$$\gamma'(t) = (\sinh t, \cosh t), \qquad \gamma''(t) = (\cosh t, \sinh t), \qquad |\gamma'(t)| = \sqrt{\cosh^2 t + \sinh^2 t}.$$

Substituting these values into (3.13) *and recalling that*

$$\cosh^2 t - \sinh^2 t = 1,$$

we obtain the curvature of γ:

$$\kappa|_{\gamma(t)} = \frac{|\sinh t \sinh t - \cosh t \cosh t|}{(\cosh^2 t + \sinh^2 t)^{\frac{3}{2}}} = \frac{1}{(\cosh^2 t + \sinh^2 t)^{\frac{3}{2}}}.$$

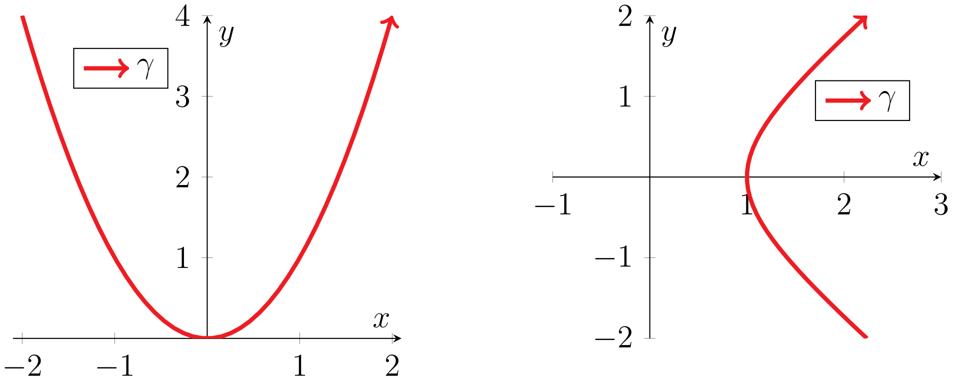


Figure 3.4. The curves from Examples 3.3 and 3.4, respectively.

3.2.2. *Signed Curvature.* The primary reason plane curves are simpler than curves in higher-dimensional spaces is that there are fewer directions in which a plane curve can turn. For example, when you are driving a car, you can only turn it in two ways: left (anticlockwise) or right (clockwise). Similarly, in the abstract setting, an oriented parametric plane curve can potentially turn in only two ways at any point: either anticlockwise or clockwise.

Thus, we can ask whether there is a simple way to capture not only *how much a plane curve is turning* (given by its curvature κ), but also *which way it is turning*. Since this new information is binary in nature (anticlockwise vs. clockwise), and since the curvature is always nonnegative, we can conveniently encode the direction of turning as a *sign* (i.e. positive or negative) attached to the curvature. This leads us to the following definition:

Definition 3.4. *Let $\gamma : I \to \mathbb{R}^2$ be a regular parametric plane curve, and let $t \in I$. We define the* <u>*signed curvature*</u> *of γ at the point $\gamma(t)$ (or at the parameter t), denoted $\kappa_s|_{\gamma(t)}$, to be:*

- $+\kappa|_{\gamma(t)}$, *if* γ *is turning anticlockwise at* $\gamma(t)$.
- $-\kappa|_{\gamma(t)}$, *if* γ *is turning clockwise at* $\gamma(t)$.
- 0, *if* γ *is not turning at* $\gamma(t)$ *(that is,* $\kappa|_{\gamma(t)} = 0$*).*

In summary, we associate the (unsigned) curvature κ with a positive sign when γ is turning anticlockwise, and we attach instead a negative sign when γ is turning clockwise.

Example 3.5. *The parametrised parabola* γ *in Example 3.3 is turning anticlockwise at each of its points; see the left plot in Figure 3.4. As a result, its signed curvature is given by*

$$\kappa_s|_{\gamma(t)} = +\kappa|_{\gamma(t)} = \frac{2}{(1+4t^2)^{\frac{3}{2}}}.$$

Example 3.6. *On the other hand, the parametrised half-hyperbola* γ *in Example 3.4 is always turning clockwise; see the right plot in Figure 3.4. Thus, its signed curvature is*

$$\kappa_s|_{\gamma(t)} = -\kappa|_{\gamma(t)} = \frac{-1}{(\cosh^2 t + \sinh^2 t)^{\frac{3}{2}}}.$$

To obtain the signed curvature using Definition 3.4, we would simply stare at the graph of a parametric curve to see whether it is turning clockwise or anticlockwise, and we would adjust the sign of the curvature accordingly. We can now ask whether there is a more computational method for obtaining the signed curvature. An affirmative answer is provided in the following theorem:

Theorem 3.8. *Let* γ *be as in Definition 3.4. Then, for any* $t \in I$,

$$\kappa_s|_{\gamma(t)} = \frac{x'(t)y''(t) - y'(t)x''(t)}{|\gamma'(t)|^3}. \tag{3.16}$$

Proof. We begin with the first formula in (3.11), and we expand γ' in terms of x' and y':

$$\begin{aligned} \frac{d}{dt}\left[\frac{\gamma'(t)}{|\gamma'(t)|}\right] &= \frac{(x'x' + y'y')(x'', y'') - (x'x'' + y'y'')(x', y')}{|\gamma'|^3} \\ &= \frac{(x'y'' - y'x'')(-y', x')}{|\gamma'|^3}. \end{aligned}$$

Now, the left-hand side of the above indicates how the direction of γ is changing, i.e. which direction γ is turning. Moreover, the vector $(-y', x')$ on the right-hand side is simply the velocity $\gamma' = (x', y')$ rotated 90 degrees anticlockwise. Thus, we see that:

- Whenever $x'y'' - y'x''$ is positive, γ is turning anticlockwise (toward $(-y', x')$).
- Whenever $x'y'' - y'x''$ is negative, γ is turning clockwise (away from $(-y', x')$).
- Whenever $x'y'' - y'x'' = 0$, then $\kappa = 0$ by (3.13).

With the above in mind, we see that if we take the formula (3.13) for the curvature, and we remove the absolute values around the numerator $x'y'' - y'x''$ and consider instead (3.16), then we capture precisely the direction that γ is turning in the sign of (3.16). □

Example 3.7. *Consider the cardioid (see Figure 3.5), which can be parametrised as*

$$\gamma : (0, 2\pi) \to \mathbb{R}^2, \qquad \gamma(t) = ((1 - \cos t)\cos t, (1 - \cos t)\sin t).$$

Suppose, for the time being, that a graph of γ was not readily available. Using (3.16)*, we can still compute its signed curvature directly and hence determine its direction of turning.*

First, we compute

$$\begin{aligned}\gamma'(t) &= (-\sin t + 2\cos t \sin t, \cos t - \cos^2 t + \sin^2 t) \\ &= (-\sin t + \sin(2t), \cos t - \cos(2t)), \\ \gamma''(t) &= (-\cos t + 2\cos(2t), -\sin t + 2\sin(2t)), \\ |\gamma'(t)| &= \sqrt{[-\sin t + \sin(2t)]^2 + [\cos t - \cos(2t)]^2} \\ &= \sqrt{2 - 2\sin t \sin(2t) - 2\cos t \cos(2t)} \\ &= \sqrt{2 - 2\cos t},\end{aligned}$$

where in the above, we used the angle addition formulas [11]

$$\sin(t+u) = \sin t \cos u + \cos t \sin u, \qquad \cos(t+u) = \cos t \cos u - \sin t \sin u.$$

Then, by Theorem 3.8 and the above, we obtain

$$\begin{aligned}\kappa_s|_{\gamma(t)} &= \frac{[-\sin t + \sin(2t)][-\sin t + 2\sin(2t)] - [\cos t - \cos(2t)][-\cos t + 2\cos(2t)]}{[2 + 2(\sin t + \cos t)\sin(2t)]^{\frac{3}{2}}} \\ &= \frac{3 - 3\sin t \sin(2t) - 3\cos t \cos(2t)}{(2 - 2\cos t)^{\frac{3}{2}}}.\end{aligned}$$

Applying the above angle addition formula and simplying, we conclude that

$$\kappa_s|_{\gamma(t)} = \frac{3 - 3\cos t}{(2 - 2\cos t)^{\frac{3}{2}}} = \frac{3}{2\sqrt{2}\sqrt{1 - \cos t}}.$$

Finally, recalling the half-angle formula [11]

$$\sin\frac{t}{2} = \sqrt{\frac{1 - \cos t}{2}}, \qquad 0 \leq t < 2\pi,$$

the formula for the signed curvature simplifies to

$$\kappa_s|_{\gamma(t)} = \frac{3}{4\sin\frac{t}{2}}.$$

In particular, note that κ_s is always positive, hence γ is rotating anticlockwise.

Recall that the curvature is a geometric property of curves, in that its value is independent of how a curve is parametrised. We now ask the analogous question for the signed curvature.

Question 3.2. *Is the signed curvature a geometric property of curves, like the (unsigned) curvature?*

That the answer to Question 3.2 is negative can be demonstrated through the subsequent example:

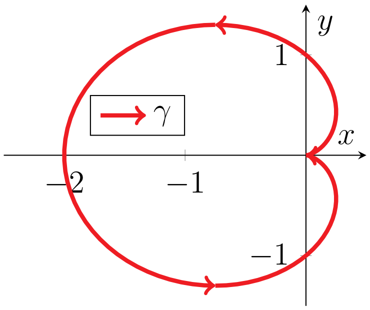


FIGURE 3.5. The parametric cardioid of Example 3.7.

Example 3.8. *Consider the following two parametrisations of the parabola* $y = x^2$*:*

$$\gamma_1 : \mathbb{R} \to \mathbb{R}^2, \qquad \gamma_1(t) = (t, t^2),$$
$$\gamma_2 : \mathbb{R} \to \mathbb{R}^2, \qquad \gamma_2(t) = (-t, t^2).$$

Notice that γ_1 *is precisely the parametric curve from Example 3.3, while* γ_2 *is the parametric curve obtained by reversing the orientation of* γ_1 *(see Proposition 2.4).*

Now, we already saw from Example 3.5 that

$$\kappa_s|_{\gamma_1(t)} = \frac{2}{(1+4t^2)^{\frac{3}{2}}} > 0.$$

In particular, this reflects that the graph of γ_1 *(see the first plot in Figure 3.6) is turning anticlockwise. On the other hand, since* γ_2 *is turning clockwise (see the second plot of Figure 3.6), then we expect* $\kappa_s|_{\gamma_2(t)}$ *to be everywhere negative. Indeed, another computation (which we omit here) yields*

$$\kappa_s|_{\gamma_2(t)} = \frac{-2}{(1+4t^2)^{\frac{3}{2}}} < 0.$$

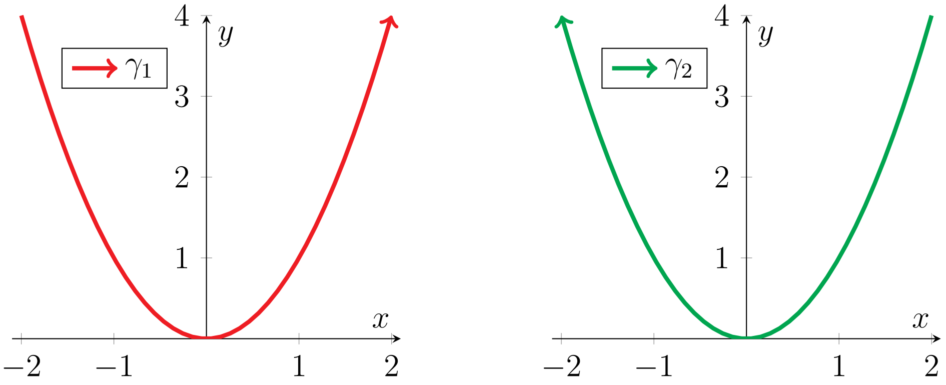


FIGURE 3.6. The parametrised parabolas γ_1 and γ_2 from Example 3.8.

A similar argument can be made for general plane curves. Indeed, if one reverses the orientation of a curve, then the reversed curve will turn in the opposite direction as the original, just as in Figure 3.6. This has the effect of negating the value of the signed curvature.

All this can be summarised more precisely in the following theorem:

Theorem 3.9. *Let* $\gamma : I \to \mathbb{R}^2$ *be a regular parametric curve, and let* $\tilde{\gamma} : \tilde{I} \to \mathbb{R}^2$ *be a reparametrisation of* γ*, with* ϕ *being the change of variables such that* $\gamma(t) = \tilde{\gamma}(\phi(t))$ *for all* $t \in I$*. Then:*

- *If* γ *and* $\tilde{\gamma}$ *have the same orientation, then* $\kappa_s|_{\tilde{\gamma}(\phi(t))} = \kappa_s|_{\gamma(t)}$.
- *If* γ *and* $\tilde{\gamma}$ *have the opposite orientations, then* $\kappa_s|_{\tilde{\gamma}(\phi(t))} = -\kappa_s|_{\gamma(t)}$.

In summary, the signed curvature fails to be a geometric property of curves, since it is changed by reparametrisations that reverse the orientation. However, Theorem 3.9 shows that it is a property of *oriented curves*, i.e. signed curvature is independent of *orientation-preserving* reparametrisations.

3.2.3. *Angular Displacement.* Previously, we discussed how the signed curvature represents the rate of change of the direction of an oriented curve, with additional information on which way the curve is turning. One could equivalently think of this information as describing the rate of change of the *angle* the curve is facing (say, with respect to the positive x-axis).

Intuitively, if we were to "add up" this change in the angle along the entire curve, then we expect to recover the total change in angle for the curve. Of course, you might then ask what exactly it means to "add up" along the curve. Fortunately, we have defined precisely this notion earlier, namely, the *path integral* along the curve.

To study this more closely, consider a regular parametric curve $\gamma : (a, b) \to \mathbb{R}^2$. Furthermore, for each $t \in (a, b)$, we let $\theta(t)$ denote the angle that γ relative to the positive x-axis. More specifically, $\theta(t)$ is the polar angle between the tangent vector $\gamma'(t)|_{\gamma(t)}$ and $(1, 0)|_{\gamma(t)}$. See Figure 3.7 for a graphical representation.

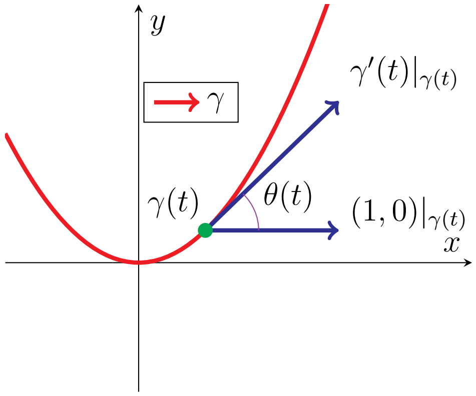


FIGURE 3.7. This plot shows how $\theta(t)$ is defined from $\gamma'(t)$.

We now examine how θ changes with respect to t. In particular, we make an explicit connection between the angle and the signed curvature:

Proposition 3.10. *Let γ and θ be as above. Then, the following holds for any $t \in I$:*

$$\theta'(t) = \kappa_s|_{\gamma(t)} |\gamma'(t)|. \tag{3.17}$$

Proof. From the proof of 3.8, writing $\gamma(t) = (x(t), y(t))$, we had computed that

$$\frac{d}{dt}\left[\frac{(x'(t), y'(t))}{|\gamma'(t)|}\right] = \frac{d}{dt}\left[\frac{\gamma'(t)}{|\gamma'(t)|}\right] = \frac{(x'y'' - y'x'')(-y', x')}{|\gamma'|^3}.$$

Using (3.16) and decomposing the into components, the above becomes

$$\frac{d}{dt}\left[\frac{x'(t)}{|\gamma'(t)|}\right] = \kappa_s|_{\gamma(t)} \cdot (-y'(t)), \qquad \frac{d}{dt}\left[\frac{y'(t)}{|\gamma'(t)|}\right] = \kappa_s|_{\gamma(t)} \cdot x'(t).$$

Now, using a bit of trigonometry, we see that

$$\cos\theta(t) = \frac{x'(t)}{|\gamma'(t)|}, \qquad \sin\theta(t) = \frac{y'(t)}{|\gamma'(t)|}.$$

Substituting this into the two differential equations yields

$$-\sin\theta(t) \cdot \theta'(t) = \frac{d}{dt}[\cos\theta(t)] = -\sin\theta(t) \cdot \kappa_s|_{\gamma(t)}|\gamma'(t)|,$$
$$\cos\theta(t) \cdot \theta'(t) = \frac{d}{dt}[\sin\theta(t)] = \cos\theta(t) \cdot \kappa_s|_{\gamma(t)}|\gamma'(t)|.$$

Since either $\sin\theta(t)$ or $\cos\theta(t)$ is nonzero for any t, we can divide one of the above two equations by $\sin t$ or $\cos t$. This results in the desired formula (3.17). □

Now that we have obtained how the angle of a parametric curve changes, we can integrate this formula (3.17) and apply the fundamental theorem of calculus:

$$\int_a^b \kappa_s|_{\gamma(t)}|\gamma'(t)|dt = \int_a^b \theta'(t)dt = \theta(b) - \theta(a). \tag{3.18}$$

Note that:

- The right-hand side of (3.18) gives the *total change in angle* of γ.
- Recalling the definition of path integrals (Definition 2.13), we see that the left-hand side of (3.18) is the integral of the signed curvature κ_s along the curve represented by γ.

Combining these observations yields the following result:

Theorem 3.11. *If* C *is an oriented curve, and if* $\Delta\theta$ *is the total angular displacement in* C*, then*

$$\Delta\theta = \int_C \kappa_s ds. \tag{3.19}$$

Moreover, if $\gamma : (a, b) \to \mathbb{R}^2$ *is a parametrisation of* C *(with the correct orientation), then* (3.19)*, written in terms of* γ*, is simply the previous formula* (3.18)*.*

Remark. To be precise, the path integral in (3.19) is not entirely like that in Definition 2.13. In particular, if C is self-intersecting, then the integrand κ_s is multiple-valued at the intersection points of C. However, this will not cause any problems in practice, since (3.18) is itself well-defined.

Example 3.9. *Fix* $R > 0$*, and consider part of the parametric circle from Example 3.2,*

$$\gamma : (a, b) \to \mathbb{R}^2, \qquad \gamma(t) = (R\cos t, R\sin t).$$

Recall that we have computed that

$$|\gamma'(t)| = R, \qquad \kappa|_{\gamma(t)} = \frac{1}{R}.$$

Since γ *is always turning anticlockwise,* $\kappa_s|_{\gamma(t)} = \frac{1}{R}$ *as well.*

Applying (3.18) *or* (3.17)*, we obtain that*

$$\theta(b) - \theta(a) = \int_a^b \kappa|_{\gamma(t)}|\gamma'(t)|dt = \int_a^b dt = b - a.$$

In other words, if one traverses the segment of the unit circle given by γ*, then one will have turned anticlockwise by exactly an angle of* $b - a$*.*

In particular, if $a = 0$ *and* $b = 2\pi$*, corresponding to exactly one anticlockwise rotation around the circle, then the total angular displacement of* γ *is exactly* $2\pi - 0 = 2\pi$*, which is sensible, since* γ *has made exactly one anticlockwise rotation about the origin.*

More generally, if we set $a = 0$ *and* $b = 2\pi m$*, then this corresponds to* γ *making* m *anticlockwise revolutions around the circle. As expected, the total angular displacement is* $b - a = 2\pi m$*.*

3.2.4. *The Winding Number.* Next, we can transform Example 3.9 into a rather general observation. For this, we consider oriented curves that are periodic:

Definition 3.5. *A parametric curve $\gamma : \mathbb{R} \to \mathbb{R}^n$ is closed iff there exists $b > 0$ such that*

$$\gamma(t + b) = \gamma(t), \qquad t \in \mathbb{R}. \tag{3.20}$$

Moreover, when γ is closed, the smallest b such that (3.20) holds is called the period of γ.

In short, a closed parametric curve is one that starts repeating itself after t has increased by b. For such a closed γ, if one begins at $\gamma(0)$ and travels along the curve, then one returns to $\gamma(0)$ when $t = b$, and this motion repeats itself afterwards.

A simple graphical example of a closed parametric curve γ in a plane is depicted in Figure 3.8. Let b be the period of γ, and suppose the green dot on the x-axis corresponds to $\gamma(0)$. As t increases, the point $\gamma(t)$ would move anticlockwise along the red curve. One returns once again to the green dot at $\gamma(b)$, and the parametrisation continues in a periodic fashion.

Moreover, if t is decreasing instead, then one would move clockwise, returning to the green dot at $\gamma(-b)$.

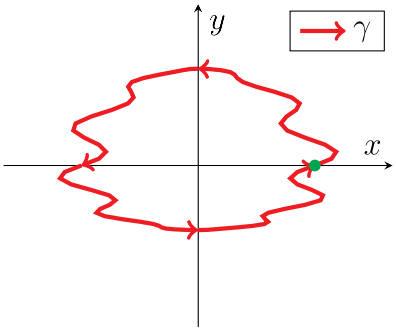


FIGURE 3.8. An example of a closed curve.

Example 3.10. *Consider the usual parametric unit circle*

$$\gamma : \mathbb{R} \to \mathbb{R}^2, \qquad \gamma(t) = (\cos t, \sin t).$$

Recall that if we start from $t = 0$, then γ begins from $\gamma(0) = (1, 0)$ and traverses the unit circle anticlockwise. This continues until $t = 2\pi$, when γ returns to its starting point $(1, 0)$. After this, γ starts repeating itself. Thus, γ is closed, with period 2π.

A more intuitive way to model a closed parametric curve is as a smooth loop, that is:

- The starting and terminal points of the parametric curve are the same.
- The starting and terminal points are joined in a smooth manner.

However, we elect to remain with Definition 3.5, as it is slightly easier to use in practice.

Definition 3.6. *Let $\gamma : \mathbb{R} \to \mathbb{R}^2$ be a closed parametric curve, with period $b > 0$. We then define the winding number of γ to be given by the formula*

$$N(\gamma) = \frac{1}{2\pi} \int_0^b \kappa_s|_{\gamma(t)} |\gamma'(t)| dt = \frac{1}{2\pi} \int_C \kappa_s ds, \tag{3.21}$$

where C is the curve represented by γ.

Let us see what this winding number means. By Theorem 3.11, we have that

$$N(\gamma) = \frac{1}{2\pi} \int_0^b \kappa_s|_{\gamma(t)} |\gamma'(t)| dt = \frac{1}{2\pi} [\theta(b) - \theta(0)], \tag{3.22}$$

where θ is the polar angle of γ, as before. Now, since γ has period b, then $\gamma'(b)$ and $\gamma'(0)$ must be the same and hence have the same angle. It then follows that

$$\theta(b) - \theta(0) = 2\pi m, \qquad m \in \mathbb{Z}.$$

Now, what does this integer m represent? To answer this, we recall that $\theta(b) - \theta(0)$ represents the total angle displacement along one period of γ. Observe then that:

- If γ revolves anticlockwise exactly once, then this total angle displacement is 2π.
- If γ revolves clockwise exactly once, then this total angle displacement is -2π.
- If γ revolves anticlockwise m times, then this total angle displacement is $2\pi m$.

In other words, if we divide this total angular displacement $\theta(b) - \theta(0)$ by 2π, we obtain the number of times, m, that γ has revolved anticlockwise. Combining this with (3.22) yields:

Theorem 3.12. *Assume the setting of Definition 3.6. Then, the winding number* $N(\gamma)$ *is precisely the number of times* γ *revolves anticlockwise in one period.*

Let us look at a couple simple examples:

Example 3.11. *Consider the closed parametric curve from Example 3.10,*

$$\gamma : \mathbb{R} \to \mathbb{R}^2, \qquad \gamma(t) = (\cos t, \sin t).$$

Recall that γ *has period* 2π*, and*

$$|\gamma'(t)| = 1, \qquad \kappa_s|_{\gamma(t)} = 1.$$

Thus, the winding number of γ *is given by*

$$N(\gamma) = \frac{1}{2\pi} \int_0^{2\pi} 1 \cdot 1 \cdot dt = 1.$$

This matches the intuition from Theorem 3.12, since γ *indeed revolves precisely once anticlockwise about the origin within one period, say* $0 < t < 2\pi$*.*

Example 3.12. *Consider next the "squashed trefoil knot", parametrised as*

$$\gamma : \mathbb{R} \to \mathbb{R}^2, \qquad \gamma(t) = (\sin t + 2\sin(2t), \cos t - 2\cos(2t)).$$

A graph of this is given in Figure 3.9; if we inspect this graph, we see that in one period, γ *makes exactly two anticlockwise revolutions. Therefore, by Theorem 3.12, its winding number is*

$$N(\gamma) = +2.$$

We can also try to compute $N(\gamma)$ *via its definition as an integral. In this direction, we obtain*

$$\int_0^{2\pi} \kappa_s|_{\gamma(t)} |\gamma'(t)| dt = \int_0^{2\pi} \frac{31 + 4\cos(3t)}{17 + 8\cos(3t)} dt,$$

after some irritating computations for γ'*,* $|\gamma'|$*,* γ''*, and* κ_s*. Now, you may find the integral on the right-hand side quite tricky to evaluate using the usual calculus tricks. On the other hand, by counting the revolutions that* γ *makes and applying Theorem 3.12, we quite easily obtain*

$$\int_0^{2\pi} \frac{31 + 4\cos(3t)}{17 + 8\cos(3t)} dt = 2\pi \cdot N(\gamma) = 4\pi.$$

Observe that since the signed curvature κ_s and the path integral are properties of oriented curves and curves, respectively, then Definition 3.6 shows that the winding number is a property of oriented loops. However, the characterisation of the winding number in Theorem 3.12 leads to an even bolder claim: that the winding number is a *topological* property.

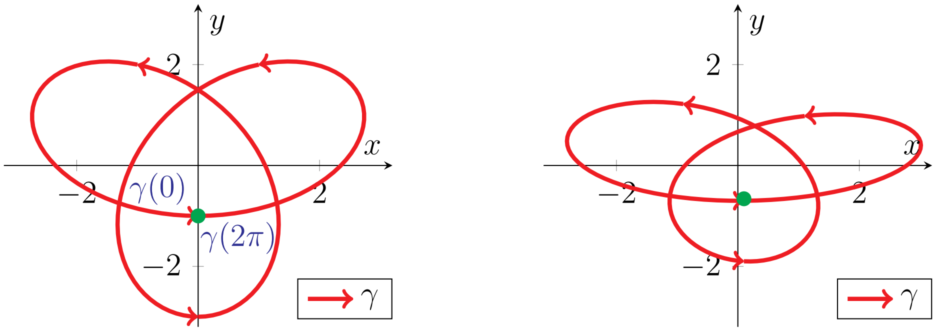


FIGURE 3.9. The left plot contains the "squashed trefoil" from Example 3.12, while the loop in the right plot is a slight deformation of that on the left. Note that both loops have the same winding number $+2$.

Recall from our discussion in Chapter 1 that topological properties are roughly those that are preserved by "deformations". In the current setting, suppose we have a closed parametric curve γ, whose winding number is $N(\gamma) = m$, so that γ revolves m times anticlockwise in one period. If we then deformed γ by a small amount by "pulling" or "compressing" it (see the second plot in Figure 3.9), then the deformed loop will still revolve m times anticlockwise. In this way, the winding number is *invariant* under deformations and hence is a topological property.

One should take a bit of time to reflect on how surprising this is at first glance. On one hand, the winding number is defined in very geometric terms, in particular via the signed curvature; to compute these quantities, one must take numerous derivatives. On the other hand, Theorem 3.12 indicates that the winding number is in fact *independent* of these geometric considerations, i.e. of the shape and size of the loop. Thus, in effect, the winding number provides one hidden connection between the areas of differential geometry and topology.

3.3. **Space Curves.**

In the previous section, we studied the special case of curves in the plane. We now consider the next simplest case—curves in 3-dimensional space.

One of the main features of (oriented) plane curves was that at any point, they can only turn in two ways: anticlockwise and clockwise. For space curves, however, this is no longer the case. In fact, there is now an additional dimension in which any curve could potentially turn. This means that to capture the shapes of space curves, one needs a different set of information than before.

3.3.1. *The Curvature Formula.*

Let us first start by deriving a more refined formula for the curvature (from Definition 3.1) which is specialised to space curves. Before we do this, however, we first revise an additional bit of vector algebra that is specific to three dimensions.

Definition 3.7. *Given two vectors*

$$\mathbf{v} = (v_1, v_2, v_3) \in \mathbb{R}^3, \qquad \mathbf{w} = (w_1, w_2, w_3) \in \mathbb{R}^3,$$

we define its cross product to be the vector

$$\mathbf{v} \times \mathbf{w} = (v_2 w_3 - v_3 w_2, v_3 w_1 - v_1 w_3, v_1 w_2 - v_2 w_1) \in \mathbb{R}^3. \tag{3.23}$$

Moreover, for two "arrows" $\mathbf{v}|_{\mathbf{p}}$ *and* $\mathbf{w}|_{\mathbf{p}}$ *from a common point* $\mathbf{p} \in \mathbb{R}^3$*, we can define*

$$\mathbf{v}|_{\mathbf{p}} \times \mathbf{w}|_{\mathbf{p}} = (\mathbf{v} \times \mathbf{w})|_{\mathbf{p}}. \tag{3.24}$$

Remark. If you are comfortable with 3×3-matrices and their determinants, then one easy way to remember formula (3.23) is to express it heuristically in the form

$$\begin{aligned} \mathbf{v} \times \mathbf{w} &= \det \begin{bmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ v_1 & v_2 & v_3 \\ w_1 & w_2 & w_3 \end{bmatrix} \\ &= \det \begin{bmatrix} v_1 & v_2 \\ w_1 & w_2 \end{bmatrix} \mathbf{i} - \det \begin{bmatrix} v_1 & v_3 \\ w_1 & w_3 \end{bmatrix} \mathbf{j} + \det \begin{bmatrix} v_2 & v_3 \\ w_2 & w_3 \end{bmatrix} \mathbf{k}, \end{aligned} \tag{3.25}$$

where

$$\mathbf{i} = (1, 0, 0), \qquad \mathbf{j} = (0, 1, 0), \qquad \mathbf{k} = (0, 0, 1).$$

To see why cross-products are useful, we recall the following geometric characterisation of it:

Proposition 3.13. *For* $\mathbf{v}, \mathbf{w}, \mathbf{p}$ *as in Definition 3.7, the product* $\mathbf{v}|_p \times \mathbf{w}|_p$ *is such that:*

- *If* $\mathbf{v}|_p$ *and* $\mathbf{w}|_p$ *lie on the same line, then* $\mathbf{v}|_p \times \mathbf{w}|_p = 0$.
- *Otherwise,* $\mathbf{v}|_p \times \mathbf{w}|_p$ *points in the direction that is perpendicular to both* $\mathbf{v}|_p$ *and* $\mathbf{w}|_p$ *and satisfies the right-hand rule. Furthermore, the norm of* $\mathbf{v}|_p \times \mathbf{w}|_p$ *satisfies*

$$|\mathbf{v}|_p \times \mathbf{w}|_p| = |\mathbf{v}||\mathbf{w}| \sin \theta, \tag{3.26}$$

where θ *is the angle made between the "arrows"* $\mathbf{v}|_{\mathbf{p}}$ *and* $\mathbf{w}|_{\mathbf{p}}$.

In particular, if you already have vectors which span two of the three dimensions of the space, then the cross product provides a simple, computable way to generate the third dimension.

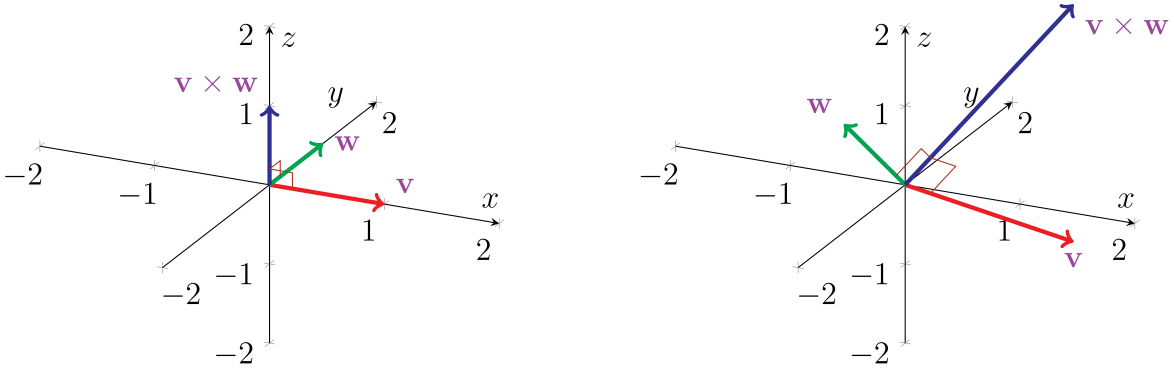


FIGURE 3.10. The diagrams demonstrate cross products of vectors, represented here as arrows beginning at the origin.

Figure 3.10 contains some graphical examples of cross products, with all vectors represented as arrows from the origin. In particular, the left and right plots depict, respectively,

$$\mathbf{v} = (1, 0, 0), \qquad \mathbf{w} = (0, 1, 0), \qquad \mathbf{v} \times \mathbf{w} = (0, 0, 1),$$
$$\mathbf{v} = (1, 1, -1), \qquad \mathbf{w} = (-1, 1, 0), \qquad \mathbf{v} \times \mathbf{w} = (1, 1, 2).$$

Remark. Keep in mind that *cross products are specific to* 3*-dimensional space!* Although there do exist generalisations of cross products to other dimensions (see, for example, Hodge duals), these are not in the form of a vector-valued product of two vectors.

We recall one additional algebraic property involving cross products:

Proposition 3.14. *Given any* $\mathbf{a}, \mathbf{b}, \mathbf{c} \in \mathbb{R}^3$*, the following identity holds:*

$$\mathbf{a} \times (\mathbf{b} \times \mathbf{c}) = (\mathbf{a} \cdot \mathbf{c})\mathbf{b} - (\mathbf{a} \cdot \mathbf{b})\mathbf{c}. \tag{3.27}$$

Proof. Check it yourself! :) □

We can now state our refined curvature formula for space curves:

Theorem 3.15. *Let* $\gamma : I \to \mathbb{R}^3$ *be a regular parametric space curve. Then, for any* $t \in I$*,*

$$\kappa|_{\gamma(t)} = \frac{|\gamma'(t) \times \gamma''(t)|}{|\gamma'(t)|^3}. \tag{3.28}$$

Proof. We begin once again with (3.11) and now recall (3.27):

$$\begin{aligned} \kappa|_{\gamma(t)} &= \frac{|[\gamma'(t) \cdot \gamma'(t)]\gamma''(t) - [\gamma'(t) \cdot \gamma''(t)]\gamma'(t)|}{|\gamma'(t)|^4} \\ &= \frac{|\gamma'(t) \times [\gamma''(t) \times \gamma'(t)]|}{|\gamma'(t)|^4}. \end{aligned}$$

By Proposition 3.13, we see that $\gamma'(t)$ is perpendicular to $\gamma''(t) \times \gamma'(t)$, and hence

$$\kappa|_{\gamma(t)} = \frac{|\gamma'(t)| \cdot |\gamma''(t) \times \gamma'(t)|}{|\gamma'(t)|^4} = \frac{|\gamma''(t) \times \gamma'(t)|}{|\gamma'(t)|^3},$$

where we applied (3.26) in the first step. □

Example 3.13. *Consider the* helix *(see Figure 3.11), parametrised by*

$$\gamma : \mathbb{R} \to \mathbb{R}^3, \qquad \gamma(t) = (\cos t, \sin t, t).$$

Let us compute its curvature using Theorem 3.15.

First, direct computations show that

$$\gamma'(t) = (-\sin t, \cos t, 1), \qquad \gamma''(t) = (-\cos t, -\sin t, 0), \qquad |\gamma'(t)| = \sqrt{2}.$$

Moreover, we compute the required cross-product in (3.28)*:*

$$\begin{aligned} \gamma'(t) \times \gamma''(t) &= (\cos t \cdot 0 + \sin t \cdot 1, -1 \cdot \cos t + \sin t \cdot 0, \sin t \cdot \sin t + \cos t \cdot \cos t) \\ &= (\sin t, -\cos t, 1), \\ |\gamma'(t) \times \gamma''(t)| &= \sqrt{2}. \end{aligned}$$

Therefore, applying Theorem 3.15, we see that

$$\kappa|_{\gamma(t)} = \frac{|\gamma'(t) \times \gamma''(t)|}{|\gamma'(t)|^3} = \frac{1}{2}.$$

Remark. In the case of Example 3.13, it is probably easier to apply Definition 3.1 directly:

$$\kappa|_{\gamma(t)} = \frac{1}{|\gamma'(t)|}\left|\frac{d}{dt}\left[\frac{\gamma'(t)}{|\gamma'(t)|}\right]\right| = \frac{1}{2}\left|\frac{d}{dt}(-\sin t, \cos t, 1)\right| = \frac{1}{2}|(-\cos t, -\sin t, 0)| = \frac{1}{2}.$$

Of course, this will not always be the case (for instance, the next example).

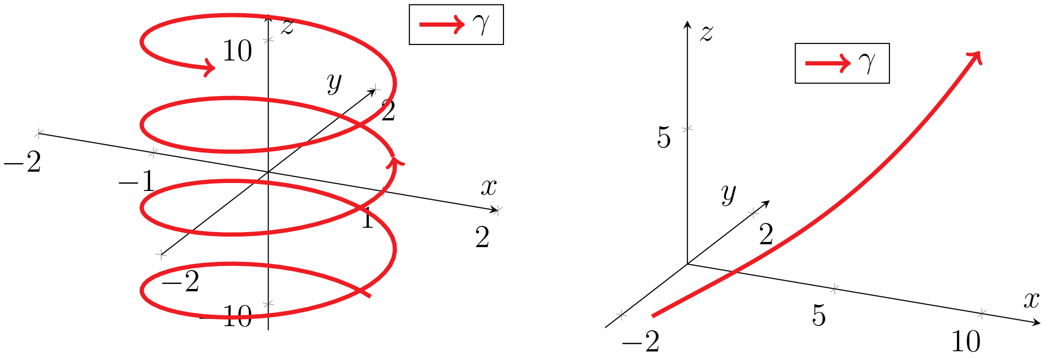


Figure 3.11. The left plot is the helix from Example 3.13, while the right plot is the parametric curve γ from Example 3.14.

Example 3.14. *Consider the parametric curve (see Figure 3.11)*

$$\gamma : (0, \infty) \to \mathbb{R}^3, \qquad \gamma(t) = (2\sqrt{2}e^t, 2t, e^{2t}).$$

To compute its curvature, we first calculate

$$\begin{aligned}\gamma'(t) &= (2\sqrt{2}e^t, 2, 2e^{2t}),\\ \gamma''(t) &= (2\sqrt{2}e^t, 0, 4e^{2t}),\\ |\gamma'(t)| &= \sqrt{8e^{2t} + 4 + 4e^{4t}} = 2(1 + e^{2t}).\end{aligned}$$

Furthermore, for the requisite cross product, we have

$$\begin{aligned}\gamma'(t) \times \gamma''(t) &= (8e^{2t} - 0, 4\sqrt{2}e^{3t} - 8\sqrt{2}e^{3t}, 0 - 4\sqrt{2}e^t)\\ &= 4\sqrt{2}e^t(\sqrt{2}e^t, -e^{2t}, -1),\\ |\gamma'(t) \times \gamma''(t)| &= 4\sqrt{2}e^t\sqrt{2e^{2t} + e^{4t} + 1}\\ &= 4\sqrt{2}e^t(1 + e^{2t}).\end{aligned}$$

Thus, by Theorem 3.15, we obtain

$$\kappa|_{\gamma(t)} = \frac{|\gamma'(t) \times \gamma''(t)|}{|\gamma'(t)|^3} = \frac{4\sqrt{2}e^t(1 + e^{2t})}{8(1 + e^{2t})^3} = \frac{e^t}{\sqrt{2}(1 + e^{2t})^2}.$$

3.3.2. *Torsion.* Up to this point, we have characterised the shape of a space curve, represented by a regular parametric curve $\gamma : I \to \mathbb{R}^3$, in two directions:

- The direction γ is heading, represented by γ'.
- The direction γ is bending, represented by γ''.

Furthermore, recall the curvature κ of γ measures the magnitude of the bending of γ.

Let us assume for now that $\kappa|_{\gamma(t)} \neq 0$. By (3.28), it follows that

$$\gamma'(t) \times \gamma''(t) \neq 0,$$

and Proposition 3.13 implies $\gamma'(t)$ and $\gamma''(t)$ are pointing in different directions. As a result of this, $\gamma'(t)|_{\gamma(t)}$ and $\gamma''(t)|_{\gamma(t)}$ together span a plane through $\gamma(t)$, called the osculating plane. This setup is demonstrated by the two plots in Figure 3.12.

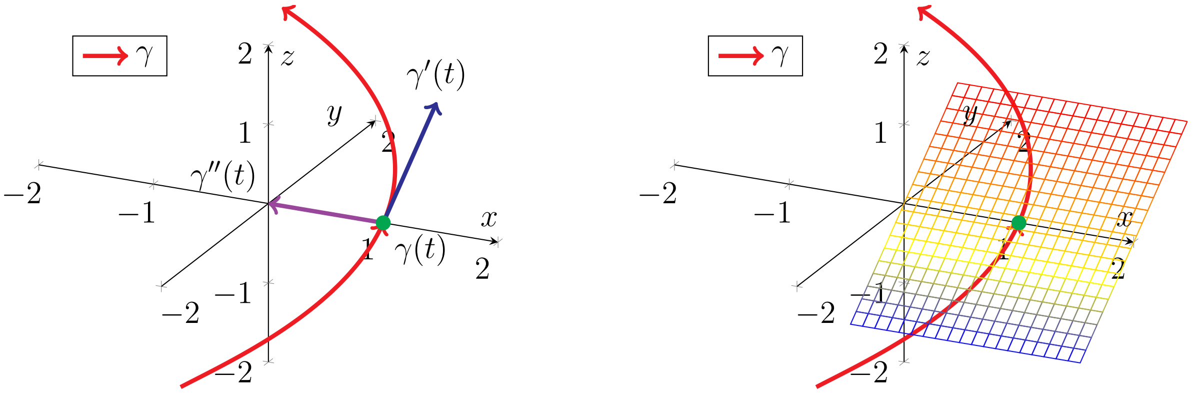


FIGURE 3.12. The left plot shows a parametric space curve γ (in red), with $\gamma'(t)|_{\gamma(t)}$ and $\gamma''(t)|_{\gamma(t)}$ indicated at a single point. The right plot shows the same γ along with its osculating plane at the same point.

Observe that in $\mathbb{R}^3$, we have an additional direction that is normal to this osculating plane. Thus, for space curves, there is one more piece of information that we must examine: the torsion of γ, that is, the tendency of γ to move away from the osculating plane.

Question 3.3. *How can we quantify this torsion of γ? Moreover, can we make this quantity a geometric property of curves, that is, can we make it independent of parametrisation?*

From Proposition 3.13, the remaining direction normal to the osculating plane is represented by

$$\gamma^\times(t)|_{\gamma(t)} = \gamma'(t)|_{\gamma(t)} \times \gamma''(t)|_{\gamma(t)} \neq 0.$$

To capture the motion of γ in the $\gamma^\times(t)$-direction, we must look at the $\gamma^\times(t)$-component of the next order of γ: its *third* derivative (in particular, since $\gamma^\times(t)$ is perpendicular to $\gamma'(t)$ and $\gamma''(t)$, neither $\gamma'(t)$ nor $\gamma''(t)$ has a $\gamma^\times(t)$-component). More specifically, we must measure the quantity

$$\gamma^\times(t) \cdot \gamma'''(t) = [\gamma'(t) \times \gamma''(t)] \cdot \gamma'''(t). \tag{3.29}$$

However, we cannot use (3.29) as the definition of torsion. This is because we also want the torsion to be independent of parametrisation, and (3.29) will fail to have this property. Consequently, we further adjust (3.29) to end up with the following definition of torsion.

Definition 3.8. *Let $\gamma : I \to \mathbb{R}^3$ be a regular parametric space curve, fix $t \in I$, and suppose that $\kappa|_{\gamma(t)} \neq 0$. We then define the torsion of γ at $\gamma(t)$ (or at the parameter t) to be*

$$\tau|_{\gamma(t)} = \frac{[\gamma'(t) \times \gamma''(t)] \cdot \gamma'''(t)}{|\gamma'(t) \times \gamma''(t)|^2}. \tag{3.30}$$

As we will see below, the factor $|\gamma'(t) \times \gamma''(t)|^{-2}$ is precisely what is needed so that the quantity is independent of parametrisation. We will prove this further below in Theorem 3.16. For the moment, however, let us first consider a couple examples.

Example 3.15. *Let us return to the helix from Example 3.13,*

$$\gamma : \mathbb{R} \to \mathbb{R}^3, \qquad \gamma(t) = (\cos t, \sin t, t).$$

Recall that γ has nonzero curvature at each point, so that its torsion is everywhere well-defined.

From Example 3.13, we have that

$$\gamma'(t) = (-\sin t, \cos t, 1), \qquad \gamma''(t) = (-\cos t, -\sin t, 0), \qquad \gamma'''(t) = (\sin t, -\cos t, 0).$$

Recall also from Example 3.13 that

$$\gamma'(t) \times \gamma''(t) = (\sin t, -\cos t, 1), \qquad |\gamma'(t) \times \gamma''(t)| = \sqrt{2}.$$

Combining the above with Definition 3.8, we calculate that

$$\tau|_{\gamma(t)} = \frac{(\sin t, -\cos t, 1) \cdot (\sin t, -\cos t, 0)}{2} = \frac{1}{2}.$$

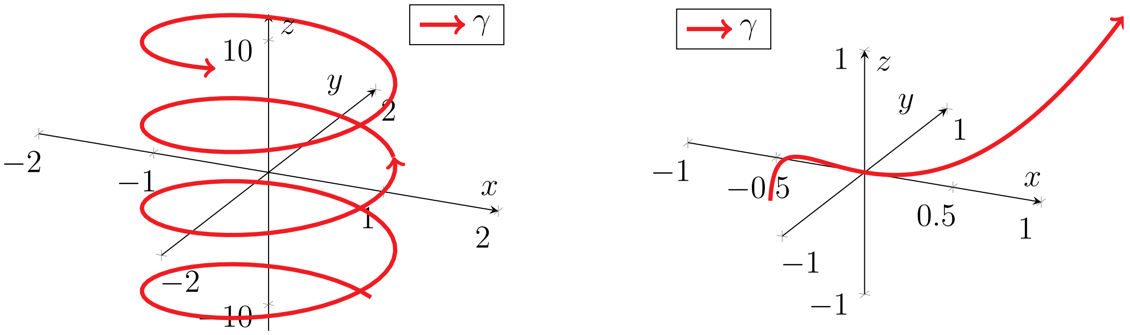


FIGURE 3.13. The left plot is the helix from Example 3.15, while the right plot is the parametric curve γ from Example 3.16.

Example 3.16. *Consider now the parametric curve*

$$\gamma : \mathbb{R} \to \mathbb{R}^3, \qquad \gamma(t) = (t, t^2, t^3).$$

Again, let us compute the torsion of γ.

First, computing its derivatives, we see that

$$\gamma'(t) = (1, 2t, 3t^2), \qquad \gamma''(t) = (0, 2, 6t), \qquad \gamma'''(t) = (0, 0, 6).$$

Furthermore, taking a cross product we see that

$$\gamma'(t) \times \gamma''(t) = (6t^2, -6t, 2), \qquad |\gamma'(t) \times \gamma''(t)|^2 = 4 + 36t^2 + 36t^4.$$

Finally, applying Definition 3.8 and the above yields

$$\tau|_{\gamma(t)} = \frac{(6t^2, -6t, 2) \cdot (0, 0, 6)}{36t^4 + 36t^2 + 4} = \frac{3}{9t^4 + 9t^2 + 1}.$$

Finally, we conclude by confirming that the torsion is indeed a geometric property of curves:

Theorem 3.16. *Assuming the setting of Definition 3.8, then the right hand side of* (3.30) *is independent of parametrisation. More specifically, if* $\tilde{\gamma} : \tilde{I} \to \mathbb{R}^n$ *be a reparametrisation of* γ*, and if* $\phi : I \to \tilde{I}$ *is the change of variables satisfying* $\gamma(t) = \tilde{\gamma}(\phi(t))$ *for all* $t \in I$*, then*

$$\frac{[\gamma'(t) \times \gamma''(t)] \cdot \gamma'''(t)}{|\gamma'(t) \times \gamma''(t)|^2} = \frac{[\tilde{\gamma}'(\tilde{t}) \times \tilde{\gamma}''(\tilde{t})] \cdot \tilde{\gamma}'''(\tilde{t})}{|\tilde{\gamma}'(\tilde{t}) \times \tilde{\gamma}''(\tilde{t})|^2}, \qquad \tilde{t} = \phi(t). \tag{3.31}$$

Proof. By the chain rule (see also Theorem 2.3), we have that

$$\begin{aligned}
\gamma'(t) &= \frac{d}{dt}[\tilde{\gamma}(\phi(t))] = \phi'(t)\tilde{\gamma}'(\tilde{t}), \\
\gamma''(t) &= \frac{d^2}{dt^2}[\tilde{\gamma}(\phi(t))] = [\phi'(t)]^2\tilde{\gamma}''(\tilde{t}) + \mathcal{A}(t)\tilde{\gamma}'(\tilde{t}), \\
\gamma'''(t) &= \frac{d^3}{dt^3}[\tilde{\gamma}(\phi(t))] = [\phi'(t)]^3\tilde{\gamma}'''(\tilde{t}) + \mathcal{B}(t)\tilde{\gamma}''(\tilde{t}) + \mathcal{C}(t)\tilde{\gamma}'(\tilde{t}),
\end{aligned}$$

where $\mathcal{A}(t)$, $\mathcal{B}(t)$, and $\mathcal{C}(t)$ are some functions of t (however, we need not know their exact forms here). Now, since $\tilde{\gamma}'(\tilde{t}) \times \tilde{\gamma}'(\tilde{t}) = 0$, a quick computation yields that

$$\begin{aligned}
\gamma'(t) \times \gamma''(t) &= \phi'(t)\tilde{\gamma}'(\tilde{t}) \times [\phi'(t)]^2\tilde{\gamma}''(\tilde{t}) + \phi'(t)\tilde{\gamma}'(\tilde{t}) \times \mathcal{A}(t)\tilde{\gamma}'(\tilde{t}) \\
&= [\phi'(t)]^3[\tilde{\gamma}'(\tilde{t}) \times \tilde{\gamma}''(\tilde{t})].
\end{aligned}$$

Furthermore, since $\tilde{\gamma}'(\tilde{t}) \times \tilde{\gamma}''(\tilde{t})$ is perpendicular to both $\tilde{\gamma}'(\tilde{t})$ and $\tilde{\gamma}''(\tilde{t})$, we obtain

$$\begin{aligned}
[\gamma'(t) \times \gamma''(t)] \times \gamma'''(t) &= [\phi'(t)]^3[\tilde{\gamma}'(\tilde{t}) \times \tilde{\gamma}''(\tilde{t})] \cdot [\phi'(t)]^3\tilde{\gamma}'''(\tilde{t}) \\
&\quad + [\phi'(t)]^3[\tilde{\gamma}'(\tilde{t}) \times \tilde{\gamma}''(\tilde{t})] \cdot \mathcal{B}(t)\tilde{\gamma}''(\tilde{t}) \\
&\quad + [\phi'(t)]^3[\tilde{\gamma}'(\tilde{t}) \times \tilde{\gamma}''(\tilde{t})] \cdot \mathcal{C}(t)\tilde{\gamma}'(\tilde{t}) \\
&= [\Phi'(t)]^6[\tilde{\gamma}'(\tilde{t}) \times \tilde{\gamma}''(\tilde{t})] \cdot \tilde{\gamma}'''(\tilde{t}).
\end{aligned}$$

Combining the above results in the desired equality (3.31):

$$\begin{aligned}
\frac{[\gamma'(t) \times \gamma''(t)] \cdot \gamma'''(t)}{|\gamma'(t) \times \gamma''(t)|^2} &= \frac{[\Phi'(t)]^6[\tilde{\gamma}'(\tilde{t}) \times \tilde{\gamma}''(\tilde{t})] \cdot \tilde{\gamma}'''(\tilde{t})}{[|\Phi'(t)|^3|\tilde{\gamma}'(\tilde{t}) \times \tilde{\gamma}''(\tilde{t})|]^2} \\
&= \frac{[\tilde{\gamma}'(\tilde{t}) \times \tilde{\gamma}''(\tilde{t})] \cdot \tilde{\gamma}'''(\tilde{t})}{|\tilde{\gamma}'(\tilde{t}) \times \tilde{\gamma}''(\tilde{t})|^2}.
\end{aligned}$$

□

3.3.3. *The Geometry of Space Curves.* Here, we complete our study of space curves by engaging in a somewhat informal discussion on how curvature and torsion affect geometry. The informal nature is partly by necessity, since some of the material requires additional background in *differential equations* (see the module MTH5123: Differential Equations) to explore fully.

We begin by first thinking about the case when the torsion vanishes. Since torsion measures the tendency of the curve to veer away from its osculating plane, zero torsion should imply that the curve stays on this plane. Moreover, as the curve is not twisting away from its osculating plane, this plane itself should remain the same along the curve. Consequently, a reasonable guess would be that any curve with vanishing torsion must itself remain within some plane.

The following theorem confirms our educated guess:

Theorem 3.17. *Let* C *be a space curve with non-vanishing curvature. Then,* C *has zero torsion everywhere if and only if* C *lies in some plane* $P \subseteq \mathbb{R}^3$.

The proof uses a fair amount of calculus and linear algebra (though not beyond the background required for this module), hence we only sketch it here without all the details.

Proof. Throughout, we let $\gamma : I \to \mathbb{R}^3$ be a regular parametrisation of C. Suppose first that C lies on a plane P. Then, at any $t \in I$, we have that $\gamma'(t)|_{\gamma(t)}$, $\gamma''(t)|_{\gamma(t)}$, and $\gamma'''(t)|_{\gamma(t)}$ all lie along P. By Proposition 3.13, the cross product $\gamma'(t)|_{\gamma(t)} \times \gamma''(t)|_{\gamma(t)}$ is normal to P, and hence

$$\tau|_{\gamma(t)} = \frac{[\gamma'(t) \times \gamma''(t)] \cdot \gamma'''(t)}{|\gamma'(t) \times \gamma''(t)|^2} = 0.$$

Assume next that C has zero torsion everywhere. First, we claim that $\gamma' \times \gamma''$ is always pointing in the same direction. Assuming this claim for the moment, it then follows that γ' and γ'', which are perpendicular to $\gamma' \times \gamma''$, must always lie within a plane. Since γ' lies within a plane, then γ can only be moving in a 2-dimensional set of directions, hence γ lies in a plane P as well.

It remains to prove the aforementioned claim. For this, we compute that

$$\begin{aligned}\frac{d}{dt}[\gamma'(t) \times \gamma''(t)] \cdot \gamma'(t) &= [\gamma''(t) \times \gamma''(t)] \cdot \gamma'(t) + [\gamma'(t) \times \gamma'''(t)] \cdot \gamma'(t) \\ &= [\gamma'(t) \times \gamma'''(t)] \cdot \gamma'(t) \\ &= 0, \\ \frac{d}{dt}[\gamma'(t) \times \gamma''(t)] \cdot \gamma''(t) &= [\gamma'(t) \times \gamma'''(t)] \cdot \gamma''(t) \\ &= -[\gamma'(t) \times \gamma''(t)] \cdot \gamma'''(t) \\ &= -\tau|_{\gamma(t)} |\gamma'(t) \times \gamma''(t)|^2 \\ &= 0.\end{aligned}$$

In other words, $\frac{d}{dt}[\gamma'(t) \times \gamma''(t)]$ has no component in the $\gamma'(t)$- and $\gamma''(t)$-directions, hence it must be pointing purely in the direction normal to $\gamma'(t)$ and $\gamma''(t)$, i.e. $\gamma'(t) \times \gamma''(t)$. Thus, we conclude that $\gamma'(t) \times \gamma''(t)$ only changes along the $[\gamma'(t) \times \gamma''(t)]$-direction, that is, $\gamma'(t) \times \gamma''(t)$ is always pointing in the same direction. This proves the claim and hence the theorem. □

Example 3.17. *Consider the parametric curve (see Figure 3.14)*

$$\gamma : \mathbb{R} \to \mathbb{R}^3, \qquad \gamma(t) = (t, t^2, t^2).$$

Computing its derivatives, we obtain

$$\gamma'(t) = (1, 2t, 2t), \qquad \gamma''(t) = (0, 2, 2), \qquad \gamma'''(t) = (0, 0, 0).$$

Since γ''' *vanishes, then the torsion vanishes as well:*

$$\tau|_{\gamma(t)} = \frac{[\gamma'(t) \times \gamma''(t)] \cdot \gamma'''(t)}{|\gamma'(t) \times \gamma''(t)|^2} = 0.$$

Alternatively, one could observe that γ *lies in the plane* $y = z$. *(Indeed, the* y*- and* z*-components of* $\gamma(t)$ *are always identical.) Thus, Theorem 3.17 also yields that* γ *has zero torsion.*

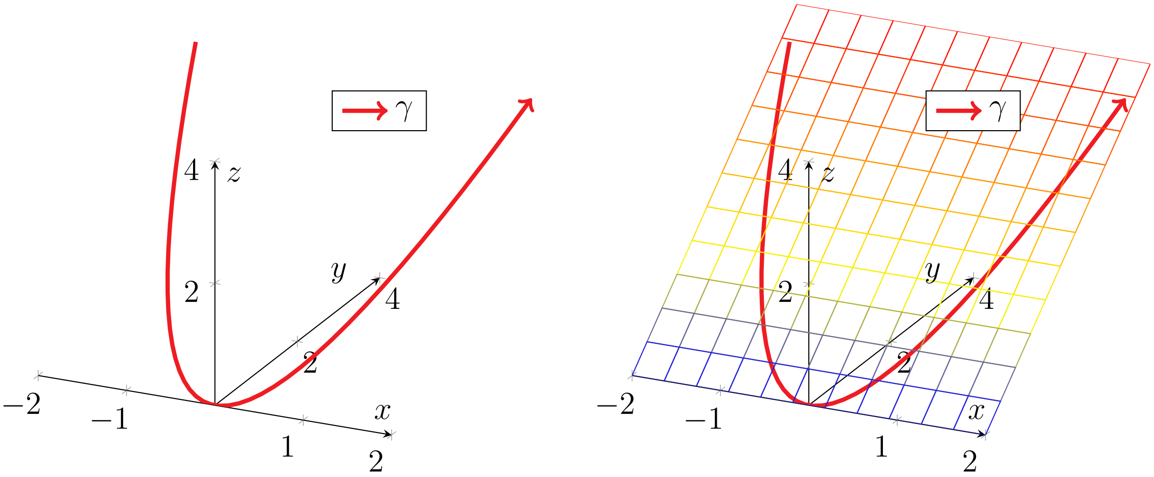


FIGURE 3.14. The parametric curve γ from Example 3.17. The left plot contains only γ, while the right plot shows γ is contained in the plane $y = z$.

We now expand the discussion to general torsions. The rough intuition is that the curvature describes the bending of a curve on the osculating plane, while the torsion describes its tendency to move in the direction normal to this plane. Since this exhausts all three dimensions, one might expect that the shape of the curve should be determined by its curvature and torsion. This idea is described more precisely in the statement of the following theorem:

Theorem 3.18. *Let C be a space curve with non-vanishing curvature. If we know*

- *the curvature and torsion of C everywhere,*
- *the position of a single point of C,*
- *the direction of C at that (same) point, and*
- *the direction that C is bending at that (same) point,*

then we know exactly what C is, that is, we can reconstruct the entire curve C.

The key idea of the proof is to solve a system of (ordinary) differential equations for C. In particular, basic results from differential equations imply that given the information from Theorem 3.18, there is a unique solution to the system that determines precisely what C is. (For details, see discussions on Frenet–Serret formulas, for example in previous years' lecture notes [4, 6].)

Another way to interpret Theorem 3.18 is as follows: *if we know the curvature and torsion of C, then we know C itself, excepting the initial position, direction, and direction of bending of C.* To demonstrate this more concretely, we consider the helix H from Examples 3.13 and 3.15; a graph is given again in the first plot of Figure 3.15. Recall we computed in those examples that the curvature and torsion of H are constant: $\frac{1}{2}$ at every point.

Now, suppose we shift H to a different position, as in the second plot of Figure 3.15. Since the shifted helix has the same shape as H, another round of computations will show, unsurprisingly, that the shifted helix will have curvature and torsion $\frac{1}{2}$ at every point. In other words, given only the curvature and torsion, we cannot distinguish two helices beginning from different locations.

Similarly, rather than shifting H, we rotate H so that it is initially moving in a different direction, as in the third plot of Figure 3.15. Again, rotations preserve the shape of the helix, so the rotated helix will again have curvature and torsion $\frac{1}{2}$. Thus, knowing only the curvature and torsion does not allow one to distinguish between H and the rotated helix.

Finally, even if two helices begin from the same position and with the same direction, they will still retain the same shape—and hence curvature and torsion—if they initially bend in different ways. This can be visualised by "twisting" H about a single point; see Figure 3.15.

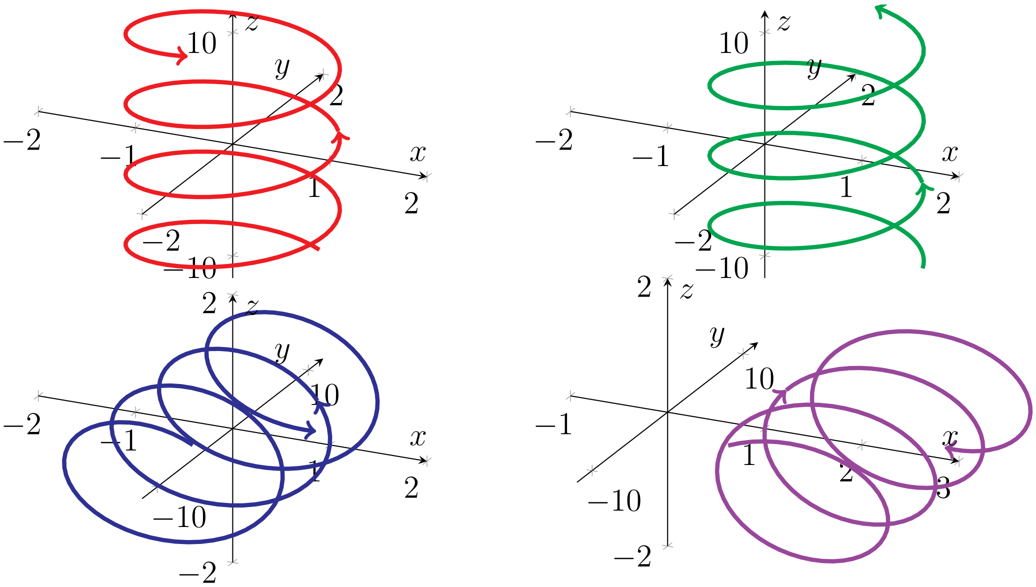


FIGURE 3.15. The top left plot contains the helix H from Examples 3.13 and 3.15. The top right, bottom left, and bottom right plots depict H shifted to a different position, H rotated about $(1, 0, 0)$, and H "twisted" about $(1, 0, 0)$—so that only the direction of bending at $(1, 0, 0)$ is altered.

Thus, to conclude, we see that for space curves that are bending (i.e. with nonzero curvature), the information within Theorem 3.18 suffices to fully determine the geometry of the curve.

3.4. **Notes on Intrinsic Geometry.** We conclude our discussion of curve geometry by revisiting the notion of intrinsic geometry that was informally introduced in Chapter 1. Though we remain informal in our approach, we now take a more detailed look at the following fundamental question:

Question 3.4. *Given a curve* C*, what are the possible intrinsic geometries that* C *could have? In other words, can we classify all the possible intrinsic geometries of curves?*

Again, we do not have the space or the technical background to formally define what we mean by intrinsic geometry. However, as an intuitive guide, we resort to the thought experiment of a bug who resides on a curve without awareness of any larger space in which the curve is embedded:

- The bug is aware of its "intrinsic velocity", that is, it is aware of (a) which of the two ways along the curve it is going, and (b) how fast it is going.
- (On the other hand, the bug does not identify either of the two directions as being special. In formal terminology, we do not fix an orientation for the curve.)
- In particular, by traversing the curve and integrating its speed over time (yes, the bug is capable of doing integrals), the bug can measure the distance it has travelled.

- In its quest to gather information about the curve, the bug is allowed to wander around the curve for all of eternity. Moreover, the bug remembers where it has been and how it has travelled in the past, and it is aware if it returns to a point it has previously visited.

3.4.1. *Curve Types.* Let us begin by listing the "types" of curves that we can come across:

(1) Curves extending indefinitely in both directions.
(2) Curves that terminate along one direction but extend indefinitely in the other direction.
(3) Curves with endpoints in both directions.
(4) Closed curves, i.e. loops.

An example of each of the four types is drawn in Figure 3.16. Note that by how we defined our bug, *it is capable of distinguishing among these four configurations and determining which of the four it is living in.* Indeed, the bug needs only venture out arbitrarily far in both directions.

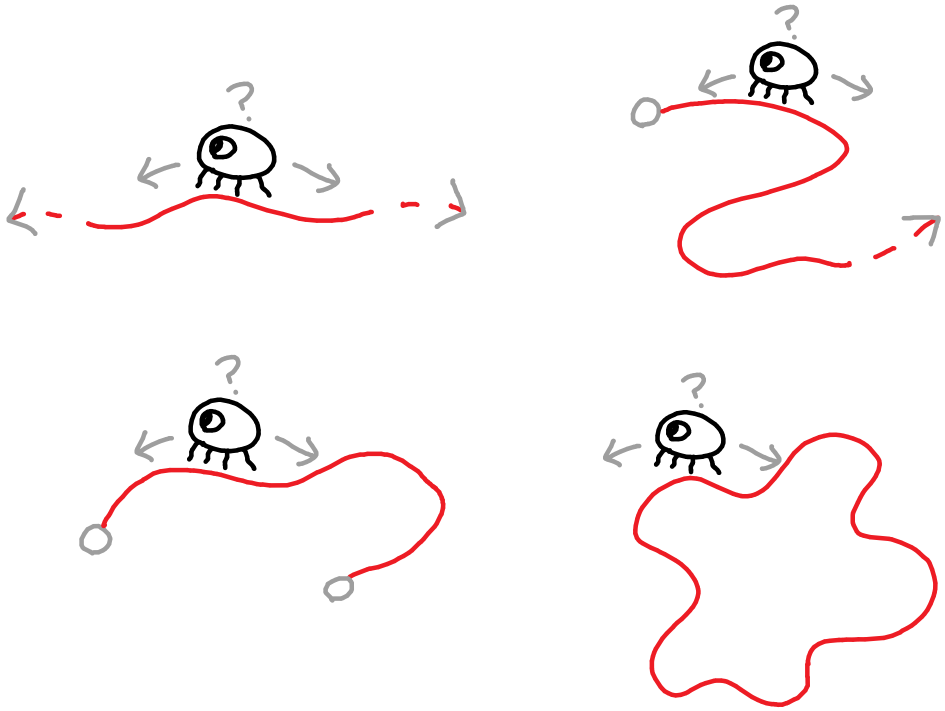


FIGURE 3.16. The different "types" of curves: (1) infinite in both directions, (2) infinite in one direction, (3) finite in both directions, and (4) closed.

More specifically, the bug will eventually determine whether the curve is "type (2)" or "type (3)" if it hits an endpoint on either or both ends. Similarly, if the bug ventures far enough and returns to a previously visited point, then the curve is "type (4)". If the bug never hits an endpoint or travels in a loop, then the curve must be "type (1)".

Note that the bug will need an eternity to determine for certain whether a curve is "type (1)" or "type (2)". But in our hypothetical situation here, this is allowed.

3.4.2. *Analysis of Curve Types.* Now that we have distinguished four types of curves above, let us further analyse each type and see what kinds of intrinsic geometries are possible.

Consider first curves of “type (1)” that extend infinitely in both directions. Suppose C_1 and C_2 are any two such curves. Then, a bug on either curve can crawl indefinitely in either direction without hitting a dead end or returning to its starting point. Consequently, as long as they are unaware of how C_1 and C_2 are situated in a larger space, these two bugs would have exactly the same experience. Our bug argument thus suggests that C_1 and C_2 have the same intrinsic geometry; more generally, *any two curves of “type (1)” have the same intrinsic geometry.*

A similar argument can be made for “type (2)” curves, which extend infinitely in one direction but terminate finitely in the other. If C_1 and C_2 are two such curves, then a bug on either curve will eventually hit a dead end in one direction, but can crawl indefinitely in the other direction. Thus, again, bugs on C_1 and C_2 will have the exact same experience. (Again, since we do not highlight one direction over the other in either curve, it does not matter which direction of C_1 or C_2 contains the dead end.) Thus, *any two “type (2)” curves have the same intrinsic geometry.*

Now, consider “type (3)” curves. A bug exploring such a curve will encounter dead ends in both directions. It will then be able to measure the distance between the two dead ends. In particular, bugs on two such curves C_1 and C_2 will have identical experiences only if they both measure the same length between endpoints. In other words, *two “type (3)” curves have the same intrinsic geometry if and only if they have the same arc length.*

Finally, the case of “type (4)” curves is similar to the preceding case. Once the bug discovers that it is on a closed curve, it can measure the distance required to travel once around the loop. Again, the bug can distinguish between two closed curves with different lengths—*two “type (4)” curves have the same intrinsic geometry if and only if they have the same length (per cycle).*

3.4.3. *Conclusions.* Thus, putting together the results of our rather informal thought experiment with bugs, we conclude that the intrinsic geometry of curves is determined by the following factors:

- The “type” of curve: (1), (2), (3), or (4).
- The length of the curve (only in (3) or (4), where the length is finite).

More specifically, all the possible intrinsic geometries are listed below:

- “Type (1)” curve.
- “Type (2)” curve.
- “Type (3)” curve of length L, for every $\mathsf{L} > 0$.
- “Type (4)” curve of length L, for every $\mathsf{L} > 0$.

In particular, there are very few properties that distinguish intrinsic geometry. Indeed, many of the properties of curves that we have studied—such as curvature and torsion—are *extrinsic* in nature, that is, these relate only to how the curve is embedded in a larger space.

Exercise 3.1. *How does the answer to Question 3.4 change if we consider oriented curves instead?*

4. An Introduction to Knot Theory

In previous chapters, we were concerned with studying *geometry* of curves. We were mainly concerned with geometric properties—roughly, those which depended only on the *shape*, *size*, and *position* of a curve. In particular, we concluded by studying, and quantifying, the shape of space curves, that is, curves lying in $\mathbb{R}^3$. For this chapter, we continue working with space curves, but we now ask questions that are of a very different nature:

Question 4.1. *In what ways can we tie a rope into a knot? How do we tell different knots apart?*

These are some foundational questions for a field of mathematics known as knot theory. In this chapter, we give a brief introduction to some basic elements of this theory.

4.1. What is a Knot?

Before we can begin to address Question 4.1 in a mathematical manner, we must first make all of these notions precise:

Question 4.2. *How do we mathematically define, or model, a knot?*

Before we get to the actual definition, which may seem quite technical at first glance, we first explore the various considerations which lead to our eventual definition.

4.1.1. *Knots as Topological Objects.* If you think in terms of the real world, then you would most likely envision a knot as a rope or string wrapped around itself. Mathematically, such a rope could be modelled as curved one-dimensional objects—the *curves* described in Chapters 2 and 3.

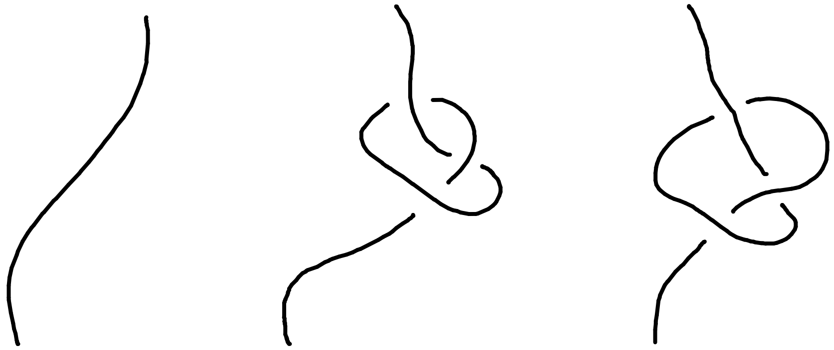


Figure 4.1. The drawings depict three curves; the leftmost curve is "untied", while the remaining two curves are "tied into a knot".

For example, the curves in Figure 4.1 could be thought of as representing knots:

- The left curve represents an untied rope—a trivial or untied knot.
- The other two curves represent ropes which contain a simple knot.

However, take a closer look at the middle and right curves. Since they are positioned and curved differently, they clearly have different geometric properties. On the other hand, you would still intuitively consider them to represent the same knot, since they are "tied the same way".

More specifically, by "pulling on various parts of the middle curve", we can deform it into the right curve without altering the fundamental structure of the knot (which we have yet to define). Because knot are intuitively impervious to such deformations, we conclude that *the structure of a knot is a topological, and not geometric, consideration.*

4.1.2. *Knots as Simple Closed Curves.* Returning again to the middle curve in Figure 4.1, we are confronted with another issue. If we are allowed to "deform" curves, then we can also pull one end of the curve through the knotted portion and hence untie the knot; see Figure 4.2. This is an undesirable feature, as we have now deformed a nontrivial knot into a trivial knot, even though we intuitively view these two knots as being clearly different.

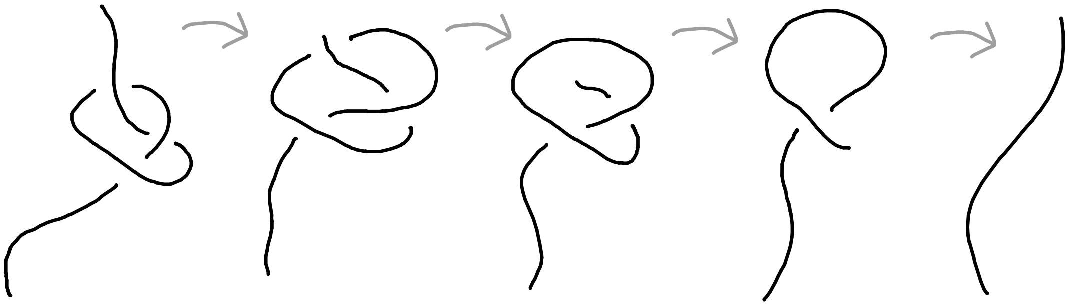


FIGURE 4.2. Diagrams showing one "untying a knotted curve" by pulling one end of the curve through the knot.

One very convenient way to prevent this issue is to fuse the two ends of the rope together, that is, we consider loops rather than curves as ends. Figure 4.3 shows the two knots from Figure 4.1, recast as loops rather than open-ended curves. Now, no deformation of the loop on the right can "untie" its knot and leave us with the loop on the left.

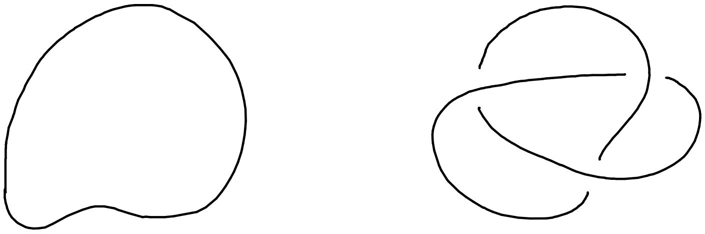


FIGURE 4.3. The diagrams depict the leftmost and middle curves in Figure 4.1, but with both ends of the curves fused together.

The preceding point is only true if the "rope" is not allowed to pass through itself, that is, *the loop must not be self-intersecting.* This is captured by the subsequent definition:

Definition 4.1. *A* <u>*simple closed curve*</u> *is a closed curve that does not intersect itself. (A closed curve can be defined as one that is generated by a closed parametric curve; see Definition 3.5.)*

4.1.3. *Knots as Space Curves.* There is yet another fundamental point that has thus far been undiscussed: the dimension n of the ambient space in which the curves are lying.

Suppose first that $n = 2$, that is, we are dealing with plane curves. Here, we immediately run into problems due to the small number of dimensions. To be more specific, we cannot wrap a plane curve around itself (i.e. "tie it into a knot") without having the curve intersect itself. In other words, a simple closed curve in $\mathbb{R}^2$ must necessarily be untied. Therefore, we could not possibly build an interesting theory of knots with plane curves.

Suppose, on the other hand, that $n > 3$, and picture any knot that you could tie with a rope (again, with the ends of the rope fused together). Then, although this is difficult to visualise, what you could do is to take a nontrivial knot, such as the one in the right diagram in Figure 4.3, and untie the knot by "deforming it along the extra fourth dimension". Thus, when $n > 3$, any simple closed curve in $\mathbb{R}^n$ must also necessarily be untied. From the above arguments, we see that to have an intuitive and interesting theory of knots, we must consider only space curves.

4.1.4. *The Formal Definition.* Combining all the intuitions from before, we conclude the following: we should *view knots as simple closed space curves, with the caveat that two such loops that can be deformed into each other (without self-intersections) are considered the same knot.*

It remains only to turn the above into a formal definition for knots. Here, the strategy is the similar to that for formally defining curves (see Definition 2.6). The first step is to characterise when one can "deform" one loop into another, so that these loops represent the same knot:

Definition 4.2. *Two simple closed curves C_1 and C_2 in $\mathbb{R}^3$ are knot-equivalent iff there exists*

$$\Phi : [0, 1] \times \mathbb{R}^3 \to \mathbb{R}^3, \qquad \Phi(s, p) = \Phi_s(p), \tag{4.1}$$

with Φ everywhere continuous, such that:

- *Each $\Phi_s : \mathbb{R}^3 \to \mathbb{R}^3$ is a bijection between $\mathbb{R}^3$ and itself.*
- *Φ_0 maps C_1 to itself.*
- *Φ_1 maps C_1 to C_2.*

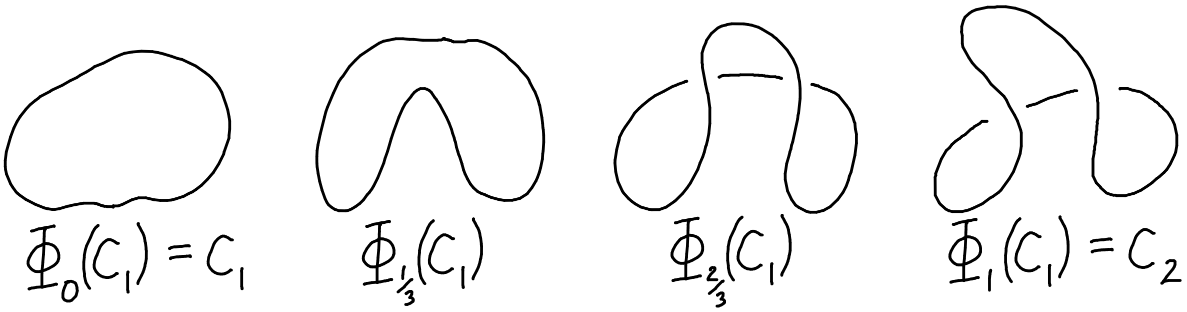


Figure 4.4. An example of a deformation Ψ, in the sense of Definition 4.2. The curves C_1 and C_2 in the above illustrations are hence knot-equivalent.

Intuitively, we can view the Φ_s's as deformations of the space $\mathbb{R}^3$, with Φ_s deforming space more and more as s increases. At the beginning, Φ_0, we have only C_1; however, after deforming and reaching Φ_1, we have transformed C_1 to C_2. That all the Φ_s's are bijections means that the curves will never pass through themselves throughout this deformation process. For a demonstration of this, see Figure 4.4, in which this is done with an untied loop.

Finally, it is not difficult to see that knot-equivalence defines an equivalence relation on the set of all simple closed curves in $\mathbb{R}^3$. Thus, we can characterise knots as follows:

Definition 4.3. *Consider the following equivalence relation $\sim$: two simple closed space curves C_1 and C_2 satisfy $C_1 \sim C_2$ iff they are knot-equivalent. We then formally define a knot as an equivalence class of simple closed space curves under the relation $\sim$.*

Remark. As an analogy, simple closed curves play the same role in the above definition of knots as parametric curves did in the definition of curves (Definition 2.6):

- Each simple closed curve represents a knot.
- A knot can be represented by many simple closed curves.

Like for curves, it is infeasible, and quite absurd, to always speak of knots as these equivalence classes. In the next section, we discuss how knots are represented in practice.

4.2. **Representation of Knots.** Suppose I wished to describe to you a certain knot, which might possibly be exceptionally complex. What I would need, then, is a efficient and accurate way to communicate the structure of the knot to you. This leads to the following question:

Question 4.3. *How can we represent knots in a way that is accurate, intuitive, and easy to draw?*

4.2.1. *Knot Diagrams.* Recall that in previous chapters, we studied curves through parametric curves, with the implicit understanding that the objects of interest are the underlying curves themselves. Indeed, parametric curves could be defined directly and hence were easier to use.

For knots, the idea is similar. For instance, we could study knots through the simple closed curves (or their parametrisations) that represent them. However, as you have likely experienced by now, space curves can be difficult to draw, since the drawing surface has only two dimensions.

Fortunately, with regards to knots, it is not so important to capture the third dimension so thoroughly, since we can deform a curve along the third dimension without altering the knot structure. We take advantage of this to make the task of drawing knots much easier.

The idea is to "flatten" a simple closed curve onto a 2-dimensional plane, that is, we project a space curve into a plane curve. Of course, for a nontrivial knot, this would cause the projected loop to self-intersect. However, because of the freedom to deform curves, at these self-intersections, the only additional information we require is whether the curve is going over or under itself. These considerations lead us to the notion of knot diagrams:

Definition 4.4. *A knot diagram is a projection (i.e. a "flattening") of a simple closed space curve onto $\mathbb{R}^2$, along with the following extra information at each self-intersection (called a crossing):*

- *The portion of the curve that is on top (the overcrossing) is drawn as usual.*
- *The portion of the curve that is on the bottom (the undercrossing) is drawn with a break.*

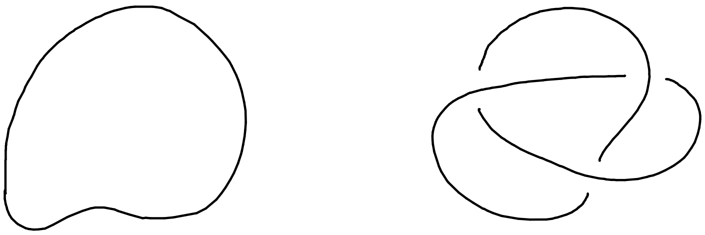


FIGURE 4.5. The left picture is a knot diagram for the unknot, while the right picture is a knot diagram for the trefoil.

Example 4.1. *Consider the left knot diagram in Figure 4.5, depicting a rope that is completely untied. The knot represented by this diagram is called the unknot, or the trivial knot.*

Example 4.2. *In contrast, the right diagram in Figure 4.5 is not trivial (though this is also not trivial to prove). This diagram represents one of the simplest nontrivial knots, called the trefoil.*

Remark. Note that since we are free to deform curves, we can always assume in a knot diagram that every crossing will involve only two curve segments.

For completeness, we extend the notion of knot equivalence to knot diagrams:

Definition 4.5. *Two knot diagrams* K_1 *and* K_2 *are said to be knot-equivalent iff they could be generated from two knot-equivalent simple closed curves.*

4.2.2. *The Reidemeister Theorem.* As we mentioned earlier, if we were to deform a simple closed curve, then as long as we avoid self-intersections, the knot structure that is represented remains unchanged. Moreover, as knot diagrams are "flattenings" of these simple closed curves onto a plane, then similar deformations of knot diagrams will also not change the knot being represented.

What are some simple examples of such knot diagram deformations? Three such instances, often called *Reidemeister moves*, are described in Figure 4.6 below:

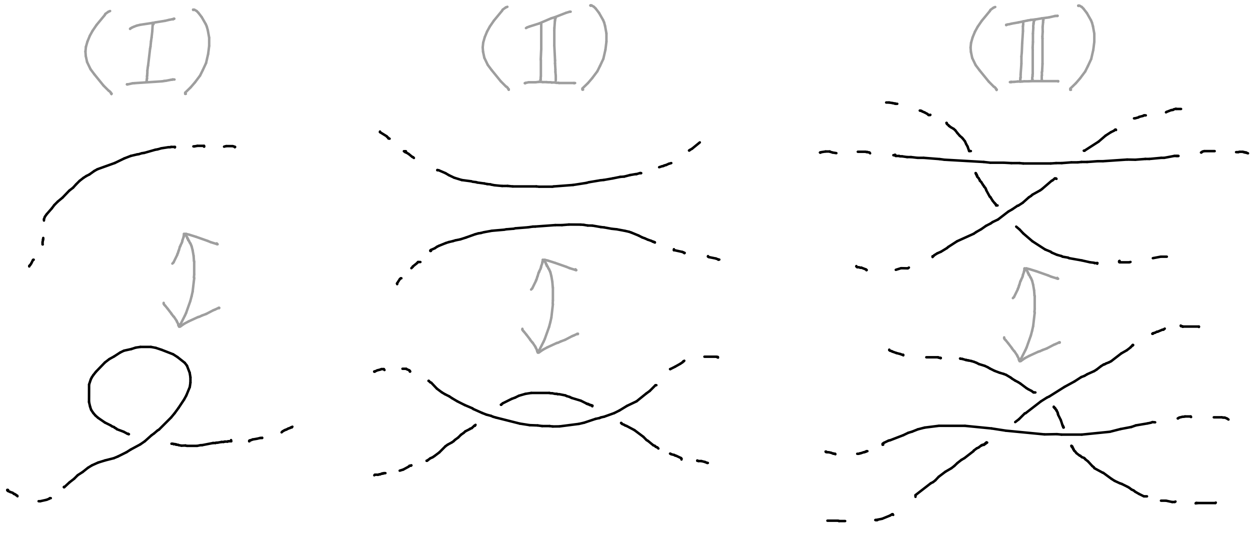


FIGURE 4.6. The three Reidemeister moves: (I), (II), and (III), respectively.

To clarify, the depicted curve segments in Figure 4.6 represent parts of a knot diagram. The dotted lines indicate that what is shown is connected to some other part of the same knot diagram that is (a) not shown, and also (b) not altered by the deformation.

Definition 4.6. *The Reidemeister moves are (somewhat informally) defined as follows:*

(I) *One either "twists" a segment of the curve to form a "loop" with an extra crossing, or one removes such a crossing by "untwisting" the loop.*

(II) *One takes one curve segment that is lying over or under another segment and pulls them both apart, or one takes two segments and places one over or under the other.*

(III) *One takes a curve segment that is lying over or under an "X-shape" formed by two other segments in the knot diagram, and one moves the first segment "across the X".*

Example 4.3. *Figure 4.7 below shows three different knot diagrams being transformed to the unknot diagram from Example 4.1 through a sequence of Reidemeister moves.*

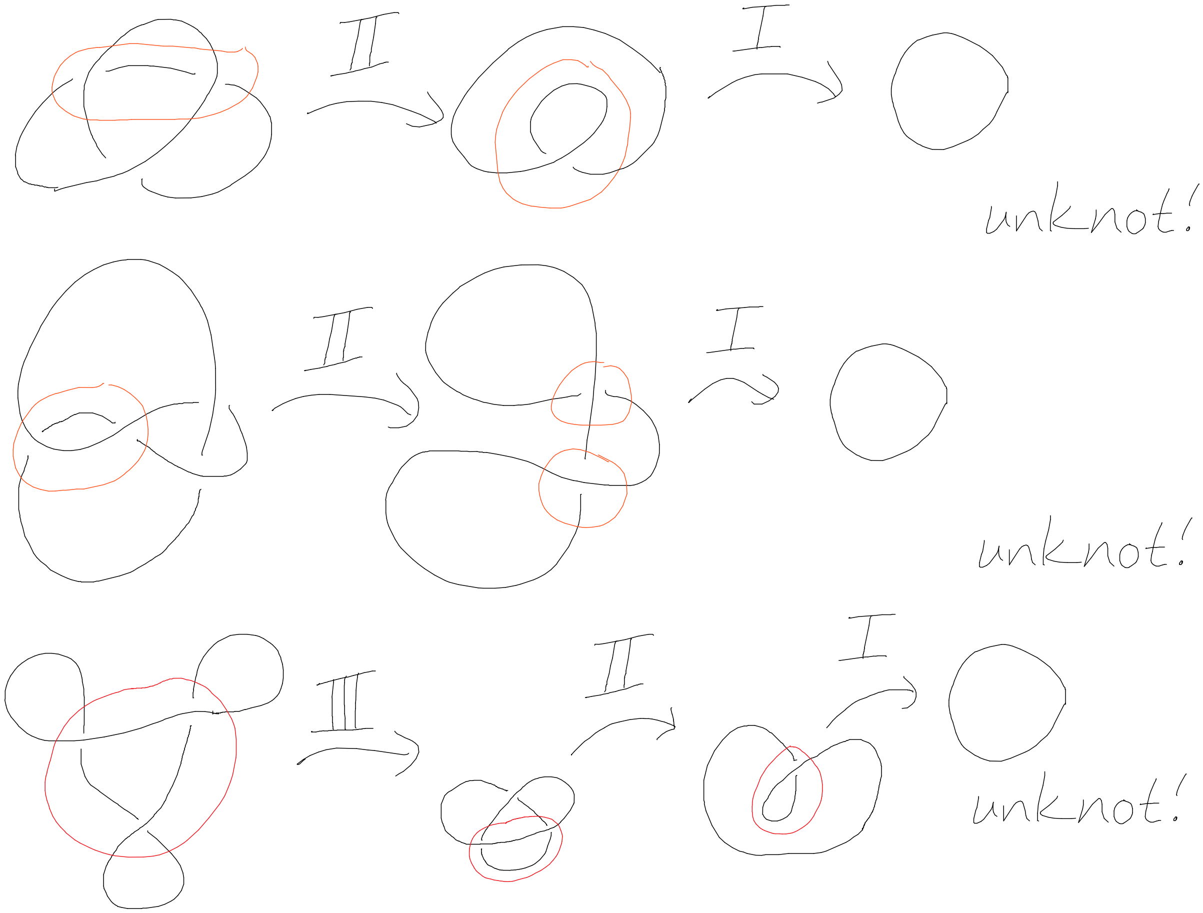


FIGURE 4.7. Three separate knot diagrams being transformed into the standard unknot diagram, via a sequence of Reidemeister moves. The parts of the diagrams where the moves are applied are circled in red.

While the Reidemeister moves should seem simple and intuitive enough, you may be wondering why mathematicians bothered to name them after anyone. The reason is that in fact, *all deformations of knot diagrams can be described using these Reidemeister moves.* This amazing fact was independently discovered in 1927 by Kurt Reidemeister (German mathematician, 1893-1971) and by James Alexander (American mathematician, 1888-1971) with his student Garland Briggs.

Theorem 4.1. *Two knot diagrams are knot-equivalent if and only if one can be transformed into the other via a finite sequence of Reidemeister moves.*

To see this in action, consider again Figure 4.7. Each of the three knot diagrams on the left-hand side is transformed into the standard unknot diagram of Example 4.1 using Reidemeister moves. As a result, each of the diagrams on the left-hand side also represents the unknot.

4.3. **Knot Invariants.** In the preceding sections, we motivated and then precisely defined what a mathematical knot is, and we described how these knots can represented as diagrams. However, this is only the beginning of the story, as our real objective is to study the structures of knots.

Suppose you are given two ropes, each tied into a knot. Then, a basic question would be for you to see whether these two knots are the same. Since abstract knots are usually graphically expressed via knot diagrams, the corresponding mathematical question is the following:

Question 4.4. *Given two knot diagrams, can we discern whether they represent the same the knot? If so, then is there an algorithm to efficiently compute the answer?*

Although Question 4.4 is a fundamental question of knot theory, it is also one in which mathematicians do not know the answer in general. Even now, a significant amount of research in knot theory is directed toward answering this question.

Moreover, we need not be nearly as ambitious as in Question 4.4. Even the following unknotting problem, which seems far simpler in comparison, does not yet have a fully satisfactory answer:

Question 4.5. *Given a knot diagram, can we discern whether it represents the unknot? Again, if so, then is there an algorithm with which we can efficiently compute the answer?*

Indeed, while there are algorithms for solving the unknotting problem, it is unknown whether there is one that can be executed in a tractable amount of time.

The above discussions suggest that it is very difficult—in fact, usually too daunting—to study Questions 4.4 and 4.5 directly. Instead, mathematicians often approach these problems more indirectly, by instead studying various *properties* of knots.

4.3.1. *Properties of Knots.* We must now clarify what we mean by properties of knots, which are often called knot invariants. In fact, we can approach this in the same way that we previously defined geometric properties of curves. At its most basic level, we can formally view knot properties as functions mapping each knot to a value representing its attribute:

Definition 4.7. *A knot invariant is a function mapping each knot to some value.*

While Definition 4.7 is simple, it is also not particularly useful, since we rarely work with knots directly. Rather, since we usually work with knot diagrams in practice, it is more useful to think first of properties of knot diagrams. More specifically, we first consider a *knot diagram property*, i.e. a function mapping any knot diagram to a value denoting its attribute.

Example 4.4. *One simple knot diagram property is the following: given a knot diagram D, we let $n_{\mathrm{cr}}(\mathsf{D})$ denote the total number of crossings in D.*

In particular, for the three knot diagrams on the left-hand sides of Figure 4.7, their total number of crossings are, from top to bottom, 3, 4, and 3.

While Example 4.4 defines a perfectly valid property of knot diagrams, we should also ask whether this is a property of knots. The difference between these two notions is the fact that *many different knot diagrams can represent the same knot.* For instance, all the knot diagrams in Figure 4.7 represent the same unknot. On the other hand, the numbers of crossings in all the diagrams in Figure 4.7 vary between 0 and 4, inclusive.

Thus, Example 4.4 would not directly assign a property to knots. For instance, would we assign the unknot the value 0, 1, 2, 3, or 4? The property n_{cr} of knot diagrams fails to answer this.

Thinking more generally now, to properly define a property of knots, we must avoid the above ambiguity. Thus, *for a knot diagram property to also define a knot invariant, we must also check that any two knot-equivalent diagrams map to the same value.*

Now, checking this is, in general, a dauntingly unrealistic task. Indeed, any knot has infinitely many possible diagrams that represent it, and we would still have to check the above property *for all such possible diagrams.* However, we can further simplify our task by recalling the *Reidemeister theorem*, which states that all knot-equivalences can be expressed in terms of three basic types of deformations. Thus, we need not consider all possible knot-equivalences, only individual Reidemeister moves that make up the building blocks of knot-equivalences.

In other words, we can characterise a *knot invariant as a property of knot diagrams, such that its value does not change when one deforms a knot diagram via any of the three Reidemeister moves.* We will further expand on this idea when we discuss basic examples.

Now, why would knot invariants be useful tools with regards to studying Question 4.4? To answer this, suppose we have a property P of knot diagrams that we also know to be a knot invariant. Suppose also that we have two diagrams D_1 and D_2 such that $P(D_1) \neq P(D_2)$. Then, we automatically know that D_1 *and* D_2 *must represent different knots* (since otherwise, the knot invariance property would dictate that $P(D_1) = P(D_2)$. Consequently, *we can use knot invariants in this way to establish that two diagrams represent different knots.*

In the remainder of this section, we describe a few simple knot invariants.

4.3.2. *Crossing Number.* While Example 4.4 fails to be a knot invariant, we can, after a slight modification, define a knot invariant based on counting crossings.

Definition 4.8. *Given a knot K, we define its crossing number to be the minimum possible value of $n_{\text{cr}}(D)$ for any knot diagram D representing K. In other words, the crossing number of K is the minimum number of possible crossings that are required in order to draw K as a knot diagram.*

Note the difference between Definition 4.8 and Example 4.4. In particular, the crossing number is indeed a knot invariant (that is, independent of how a knot is represented), since it is obtained by minimising over *all possible knot diagram representations* of a knot.

Example 4.5. *The crossing numbers of some simple knots are given below:*

(1) *The unknot has crossing number 0.*
(2) *The trefoil has crossing number 3.*
(3) *The figure-8 knot has crossing number 4.*

Knot diagrams for each of the above that achieve the crossing number are drawn in Figure 4.8.

While the crossing number is quite simple conceptually as well as fairly informative, it is also not a particularly useful quantity. For example, to find the crossing number of the trefoil (suppose

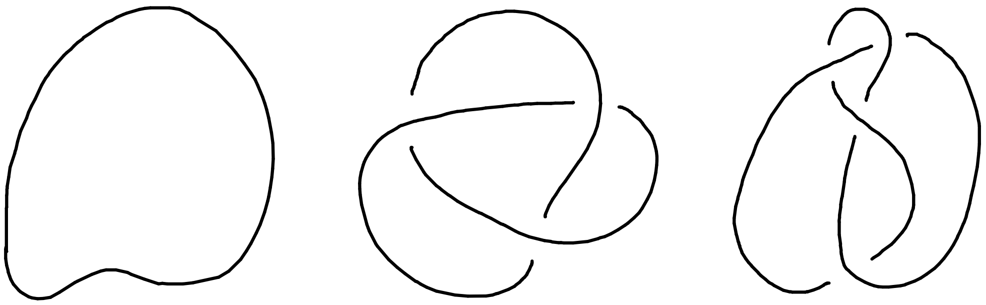


FIGURE 4.8. Knot diagrams for the unknot (left), trefoil (middle), and figure-8 knot (right). Each of these diagrams achieves the corresponding crossing number indicated in Example 4.5.

the answer had not already been spoiled), one would have to consider *all possible knot diagrams* representing the trefoil. More specifically, while it is believable that the middle diagram in Figure 4.8 contains the least possible number of crossings for any digrammatic representation of the trefoil, it is considerably more difficult to *prove* that this is the case. Again, this would involve considering all possible cases; for more complicated knots, this task would be far more difficult.

4.3.3. *Tricolourability.* The next knot invariant that we will study involves studying whether knot diagrams can be coloured in a particular way. Let us begin from the perspective of knot diagrams.

First, we can view a knot diagram as a collection of curve segments, with the endpoints of each segment being an undercrossing in the diagram. Moreover, observe that at each such crossing, there are three segments involved:

(1) The overcrossing represents one of the segments.
(2) The undercrossing has the remaining two segments, which are separated by the overcrossing.

This is demonstrated in Figure 4.9, which contains a knot diagram for the figure-8 with the segments coloured.

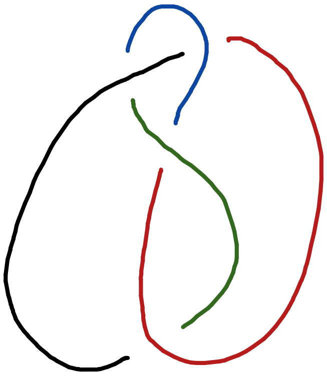


FIGURE 4.9. A knot diagram for the figure-8 knot, with each of its segments given a different colour.

Now, one can imagine taking a knot diagram and assigning a colour to each of its curve segments. In particular, let us consider assigning one of three colours—say, red, green, and blue—to these segments. The following property of knot diagrams concerns whether one can be coloured red, green, and blue in a certain manner:

Definition 4.9. *A knot diagram* D *is* <u>*tricolourable*</u> *iff it is possible to colour its curve segments in three different colours (for instance, red, green, and blue) such that the following are satisfied:*

(1) *All three colours are used somewhere in* D*.*
(2) *At each crossing of* D*, one of the following holds:*
    - *All three segments there have the same colour.*
    - *All three segments there have different colours.*

Possible colourings of a single crossing are demonstrated in Figure 4.10 below:

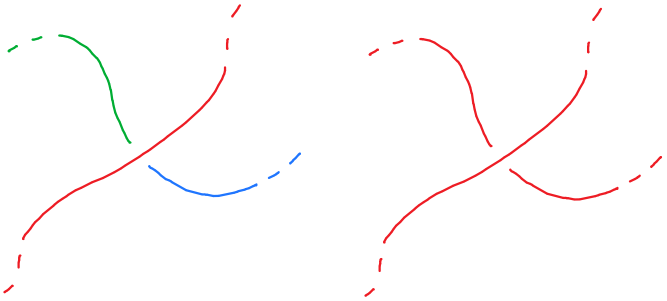


FIGURE 4.10. These illustrations demonstrate single crossings within a 3-coloured knot diagram. In the left crossing, three different colours are used, while in the right crossing, the same colour (red) is used.

Tricolourability, or lack thereof, is most easily demonstrated via examples:

Example 4.6. *The standard unknot diagram (see the left diagram in Figure 4.11) contains only one curve segment, hence this can only be coloured using a single colour. Thus, this diagram fails condition (1) of Definition 4.9 and hence fails to be tricolourable.*

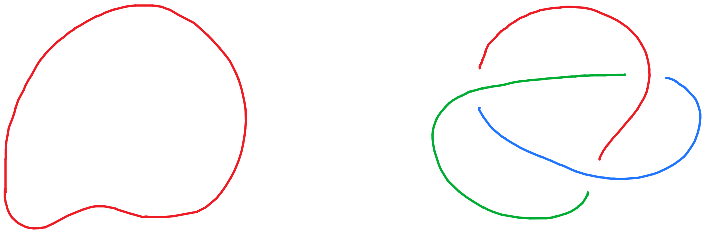


FIGURE 4.11. The left diagram shows that only one colour can be used for the standard unknot diagram, which is not tricolourable. The right diagram shows a valid 3-colouring for the standard trefoil diagram.

Example 4.7. *On the other hand, the usual trefoil knot diagram is tricolourable. One valid* 3*-colouring can be found in the right diagram in Figure 4.11.*

So far, we could only speak about the tricolourability of knot diagrams. However, its usefulness stems from the fact that it is also a knot invariant, which is shown in the following theorem:

Theorem 4.2. *Tricolourability is a knot invariant.*

Sketch of proof. Since any two knot-equivalent diagrams are related via a sequence of Reidemeister moves (by the Reidemeister Theorem), then we need only show that tricolourability, or lack thereof, is preserved by each of the three types of Reidemeister moves. In other words, for each Reidemeister move, we show if the knot diagram has a 3-colouring before the move, then there is a way to 3-colour the new diagram after the move. For this, the main idea is to show that one can recolour the crossings that have been changed by the Reidemeister move, without changing any colours elsewhere in the portions of the diagram that remain unchanged.

To demonstrate this, we show examples of such recolourings in Figure 4.12; in fact, all recolourings through Reidemeister moves are equivalent to one of the cases shown in Figure 4.12. □

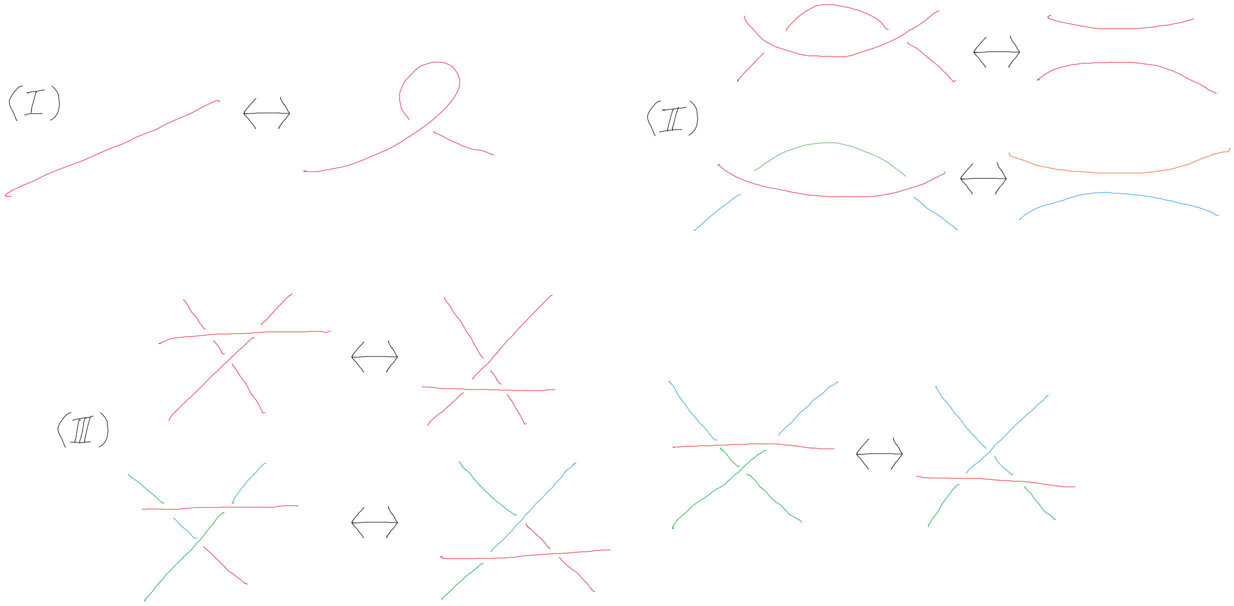


Figure 4.12. These illustrations show various cases in the proof of Theorem 4.2. More specifically, these show how applying Reidemeister moves to a knot diagram does not change whether it is tricolourable.

As a result of Theorem 4.2, it makes sense to talk about knots being tricolourable, and not just knot diagrams. We can now update our knowledge from Examples 4.6 and 4.7.

Example 4.8. *By Examples 4.6 and 4.7:*

- *The unknot is not tricolourable.*
- *The trefoil is tricolourable.*

*Moreover, a careful case-by-case analysis shows that the figure-*8 *knot is not tricolourable.*

In particular, the above observations prove that:

- *The unknot and the trefoil are different knots.*
- *The trefoil and the figure-*8 *are different knots.*

Observe that in order to check whether a knot, such as the trefoil, is tricolourable, *we need only check this for a single diagram that represents it.* Indeed, Theorem 4.2 guarantees that the result will be the same for any other knot-equivalent diagram.

4.3.4. *Chirality.* Next, we study whether a knot has a certain "mirror symmetry". As in the case of tricolourability, we begin by making precise sense of this at the level of knot diagrams.

Definition 4.10. *Given a knot diagram* D*, we define its mirror image to be the knot diagram obtained from* D *by inverting all of its crossings (that is, each overcrossing becomes an undercrossing, while each undercrossing becomes an overcrossing).*

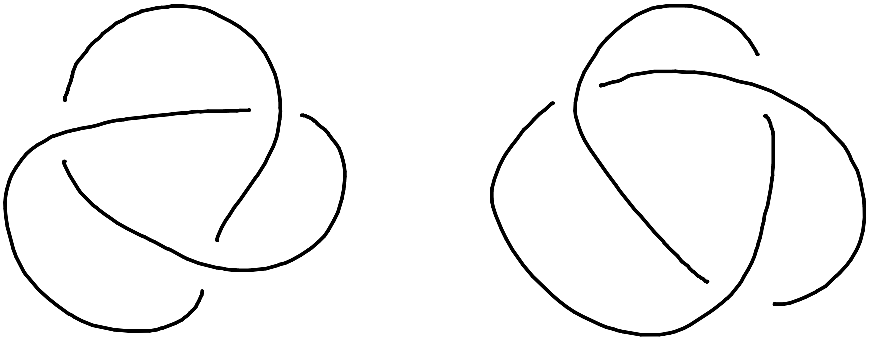


Figure 4.13. The usual trefoil diagram (left), and its mirror image (right).

Example 4.9. *The trefoil diagram and its mirror image are drawn in Figure 4.13.*

The notion of chirality refers to the *absence* of a mirror image symmetry.

Definition 4.11. *A knot diagram* D *is called chiral iff its mirror image is not knot-equivalent to* D *itself. A knot diagram that is not chiral is said to be achiral.*

Although chirality is defined in terms of knot diagrams, it is, like tricolourabilty, actually a property of knots. This is formally established in the subsequent theorem. Therefore, it makes sense to say that a knot, rather than knot diagram, is chiral or achiral.

Theorem 4.3. *Chirality is a knot invariant.*

Sketch of proof. By the Reidemeister theorem, it suffices to show that if we apply any of the Reidemeister moves to a knot diagram, then this transformation does not change whether the diagram is chiral or not. The main observation behind obtaining this is the following: if two knot diagrams D_1 and D_2 are related by a type k Reidemeister move, then their mirror images D_1', D_2' (respectively) are also related by a type k Reidemeister move.

Suppose now that D_1 is achiral, so that D_1 is knot-equivalent to D_1'. Then, for D_2 as above (related to D_1 via a Reidemeister move), we have that:

- D_2 is knot-equivalent to D_1 (via a Reidemeister move).
- D_1 is knot-equivalent to D_1' (by the assumed achirality).
- D_1' is knot-equivalent to D_2' (by the previous observation).

As a result, D_2 is knot-equivalent to D_2', and hence D_2 is also achiral.

Furthermore, by symmetry, if D_2 is achiral, then so is D_1. Consequently, if D_1 is chiral, then so is D_2 (since otherwise, D_1 must be achiral as well), completing the proof of the theorem. □

As a result of Theorem 4.3, to check whether a knot is chiral, we need only consider the question for a single knot diagram that represents this knot.

Example 4.10. *Observe that the mirror image of the standard unknot diagram (i.e. the left diagram of Figure 4.3) is itself. Therefore, the unknot is achiral.*

Example 4.11. *The figure-8 knot is also achiral. Figure 4.14 sketches a sequence of deformations taking the standard figure-8 knot diagram to its mirror image.*

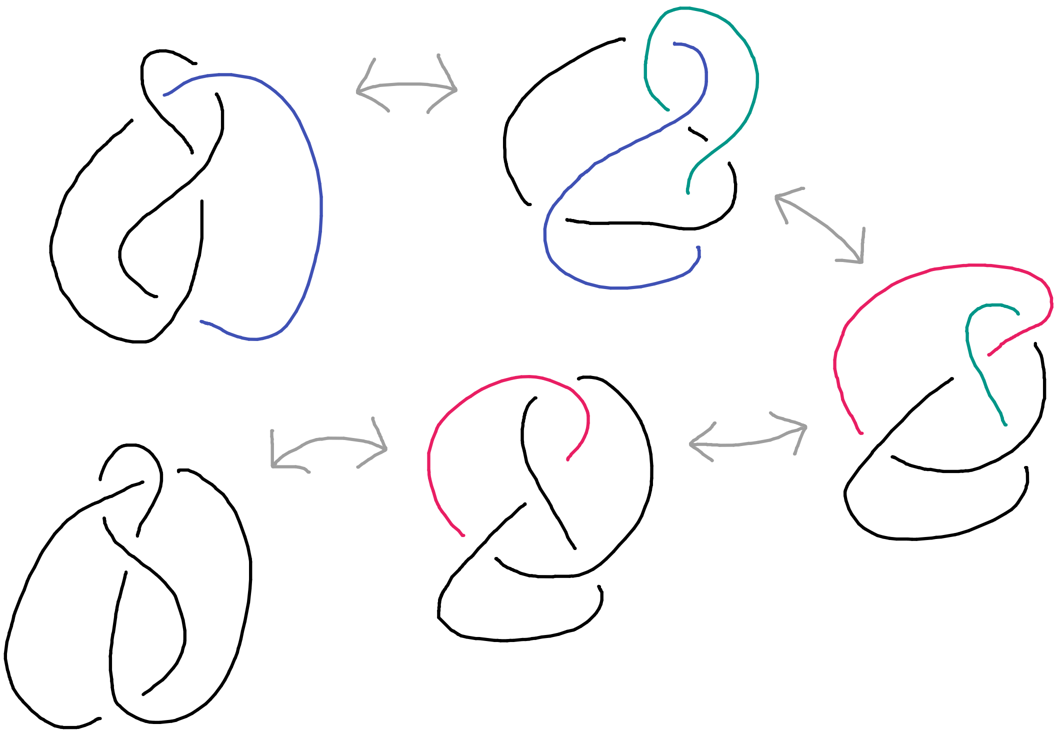


FIGURE 4.14. These sketches show how the standard figure-8 knot diagram can be deformed to its mirror image (proving that the figure-8 is achiral). The colours show which segments are moved in each step.

Example 4.12. *On the other hand, one can prove that the trefoil is chiral. (We defer this proof for later in this chapter, once we have developed more tools; see Example 4.21.)*

In other words, the reverse trefoil (represented by the right diagram in Figure 4.13) has a knot structure that is distinct from that of the trefoil (the left diagram in Figure 4.13).

4.4. **The Jones Polynomial.** In the previous section, we discussed various properties of knots—*knot invariants*—as ways to study knots themselves. One "holy grail" of this area of investigation is to discern when two given knot diagrams represent the same knot, that is, to answer Questions 4.4 and 4.5. Knot invariants provided a partial answer to this question, in that two knot diagrams which possess different values for a knot invariant must represent distinct knots.

With this mind, we can then ask: *Is a given knot invariant effective at telling knots apart?* In order for the answer to be positive, a knot invariant should have two desirable properties:

(1) A "good" knot invariant should distinguish among as many different knots as possible.
(2) A "good" knot invariant should be relatively easy to compute.

With these two considerations in mind, let us first see how the knot invariants we have studied thus far fail to satisfy one of these two criteria.

First, recall that tricolourability maps each knot into only two different values: "yes" (tricolourable) or "no" (not tricolourable). As a result, using tricolourability, one can only distinguish between a knot that is tricolourable (e.g. the trefoil knot) and another that is not (e.g. the unknot). In particular, it cannot distinguish between two tricolourable knots or two non-tricolourable knots.

For example, one cannot use tricolourability to distinguish between the unknot and the figure-8 (both fail to be tricolourable), even though they are clearly quite different.

Similarly, chirality maps only into two different values, hence it is also not a particularly effective tool for discerning knots. For instance, both the unknot and the figure-8 are achiral, hence chirality again cannot be used to distinguish these as two different knots.

On the other hand, the crossing number does a better job at classifying knots, since it can map to any non-negative integer (in particular, it does distinguish between the unknot and the figure-8). However, the serious disadvantage of the crossing number is that it is in general tremendously difficult to compute. To find the crossing number of a knot K, we have to consider every possible knot diagram representation of K and count its number of crossings. As this is computationally intractable, this essentially rules out the crossing number as a practical tool.

The goal of this section, then, is to discuss another knot invariant that fares far better in both criteria (1) and (2): the *Jones polynomial*. In contrast to previous invariants, the Jones polynomial maps each knot to a *polynomial*, rather than a number or a yes/no answer. As a result, it can take many more distinct values and can hence tell many more knots apart. Furthermore, as we will see, the Jones polynomial can be computed via a recursive process. Although this is usually quite time-consuming, this does provide an effective algorithm that allows a person or a computer to find the Jones polynomial of any given knot diagram.

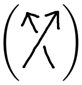
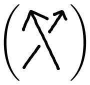
However, before we can define the Jones polynomial itself, we must first complete several preliminary steps. In particular, we must first discuss some quantities that fail to be knot invariants.

4.4.1. *Writhe.* The first "not-invariant" (pun intended, with apologies) we discuss is the *writhe*, which roughly measures how much the arcs in a knot diagram are "coiled around each other".

Definition 4.12. *Given a knot diagram D, we define its* <u>*writhe*</u> $\mathsf{W}(\mathsf{D})$ *as follows:*

(1) *First, we assign an orientation to D, that is, we choose a direction of travel along the closed curve representing the diagram (i.e. an orientation of this curve).*

(2) *For each crossing of D, we associate with it a* <u>*signature*</u>*, or* <u>*sign*</u>*, using the following rule:*

(4.2) sgn () $= +1$, sgn () $= -1$.

More specifically, we rotate the oriented diagram so that the two portions of curve in the crossing are heading from the bottom-left to the top-right and from the bottom-right to the top-left. This crossing has positive signature if the bottom-left to top-right branch is on top, and negative signature if the bottom-right to top-left branch is on top.

(3) *Finally, the* <u>*writhe*</u> *of D is simply the sum of all the signatures of its crossings:*

$$\mathsf{W}(\mathsf{D}) = \sum_{\textit{crossings } \mathsf{P} \textit{ in } \mathsf{D}} \operatorname{sgn}(\mathsf{P}). \tag{4.3}$$

While the formal definition may seem complicated, it is demonstrated more easily via example:

Example 4.13. *Let us compute the writhe of the standard trefoil knot diagram T, drawn on the left illustration in Figure 4.15. The first step is to set an orientation for T; our chosen direction of travel is indicated by the purple arrows in the middle diagram of Figure 4.15.*

Now, for each crossing of this oriented diagram, we compute its signature according to the rule given in (4.2)*. By rotating the diagram as needed, we see that all three crossings have signature* $+1$*; see the right diagram of Figure 4.15. Finally, by* (4.3)*, the writhe of* T *is hence given by*

$$\mathsf{W}(\mathsf{T}) = +1 + 1 + 1 = +3.$$

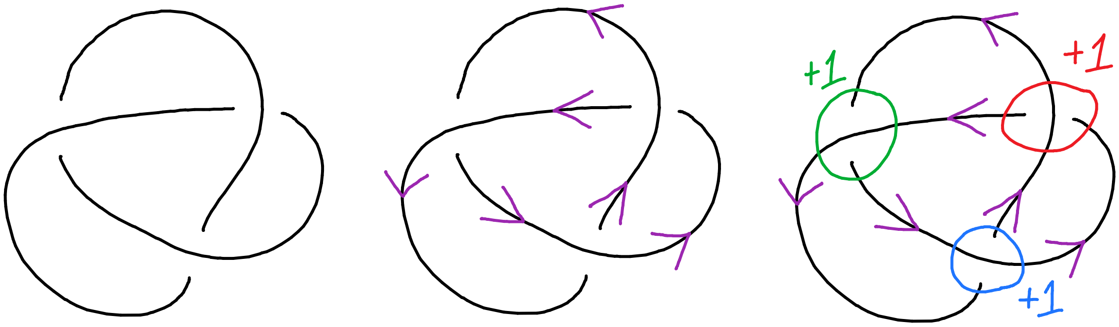


FIGURE 4.15. The left drawing is the standard trefoil knot diagram T from Example 4.13. In the middle drawing, we have assigned an orientation to T, which is indicated by the purple arrows. In the right drawing, the signature of each crossing has been added to the oriented diagram.

One point to note is that even though we had to choose an orientation of a given knot diagram in order to compute its writhe, *the value of the writhe that we ultimately compute is actually independent of our choice of orientation.* To see this, we observe that if we were to switch to the opposite orientation, then at each crossing, both branches would have their directions reversed. In particular, this would not change how the crossing looks with respect to (4.2) (although we would, however, have to turn the diagram upside-down to see this).

For completeness, we now provide two more basic examples:

Example 4.14. *Next, consider the standard unknot diagram* D *from the left drawing of Figure 4.5, to which we can assign an orientation. Since* D *contains no crossings, we have that*

$$\mathsf{W}(\mathsf{D}) = 0.$$

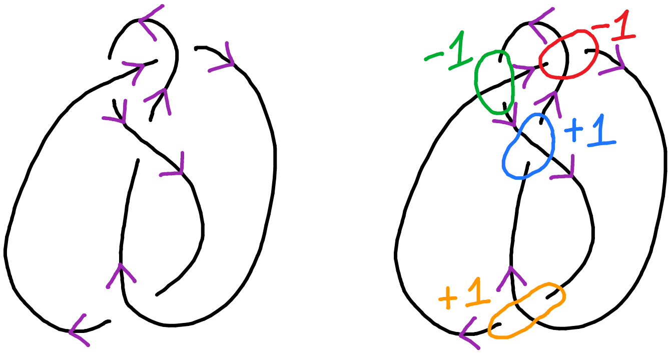


FIGURE 4.16. The left diagram contains the figure-8 knot diagram D from Example 4.15, along with a chosen orientation (in purple). The signature of each crossing of D is indicated in the right diagram.

Example 4.15. *Consider the standard figure-8 knot diagram* D *from the rightmost drawing of Figure 4.8. An orientation is indicated in the left drawing of Figure 4.16, while the signature of each crossing in the oriented diagram is indicated in the right drawing. Consequently, the writhe is*

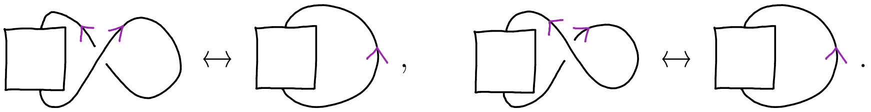
$$W(\mathsf{D}) = -1 - 1 + 1 + 1 = 0.$$

Previously, we had already mentioned that the writhe fails to be a knot invariant; we now study this in further detail. Recall from the Reidemeister theorem (Theorem 4.1) that any knot equivalence can be decomposed into a finite sequence of Reidemeister moves. As a result, it suffices to study how the writhe is impacted by each Reidemeister move.

Let us begin with type I Reidemeister moves. To describe these generally, let us first add one bit of notation for convenience. In our knot diagrams, we will use a rectangular box to represent a part of the diagram that remains unchanged over the operations depicted.

To be more specific, type I moves can be abstractly described using these "black boxes" by

(4.4)



(For future reference, we also chose an orientation for each of the diagrams in (4.4).) The rectangular boxes in (4.4) could represent something very simple, e.g. one curve segment without any crossings, or something excessively complicated, consisting of millions of arcs and crossings. However, what is assumed in drawings such as (4.4) is that whatever is in the "black box" remains unchanged between the left-hand and right-hand sides of the "↔".

Returning to the writhe, we see that in the first transformation of (4.4), the left-hand side has one additional crossing compared to the right-hand side. Moreover, using the given orientation in (4.4), we see that this crossing has signature $+1$. As a result, we conclude that *the first operation of* (4.4) (from left to right) *has the effect of decreasing the writhe by* 1.

A similar analysis of the second operation in (4.4) shows that the extra crossing on the left-hand side now has signature -1. Consequently, *this operation* (again from left to right) *has the effect of increasing the writhe by* 1. These results are summarised in Figure 4.17 below.

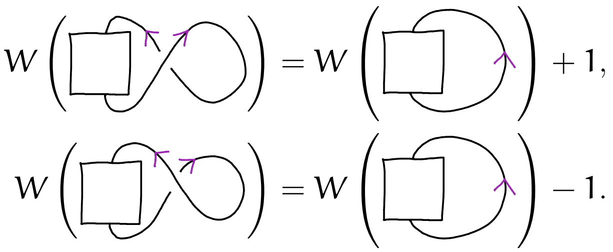
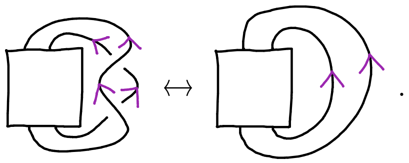


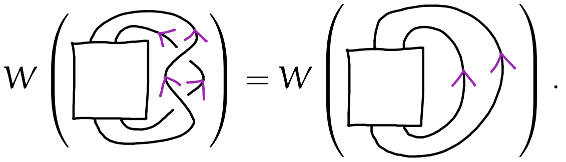
FIGURE 4.17. The equations above demonstrate how the writhe of a knot diagram is impacted by type I Reidemeister moves.

Next, let us consider type II Reidemeister moves. This can be approached in the same manner as we did for type I moves. First, we observe that by using the "black box" notation again, we can describe one case of such type II moves graphically as

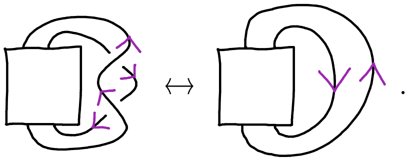


Note that aside from the "black box", the left-hand side has two additional crossings not present in the right-hand side. Moreover, the extra top and bottom crossings have signatures $+1$ and -1, respectively. As a result, we conclude from Definition 4.12 that

(4.5)

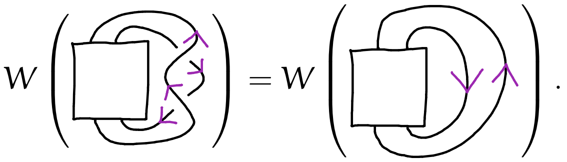


The above, however, does not quite complete the analysis of type II moves. Indeed, depending on what happens inside the "black box", it may happen that one of the two arcs protruding from the black box is forced to have the opposite orientation, that is,



Although this is in principle a different case than before, it can still be handled in exactly the same manner. In particular, on the left-hand side of the above, the extra top and bottom crossings now have signatures -1 and $+1$, respectively, hence we again conclude that

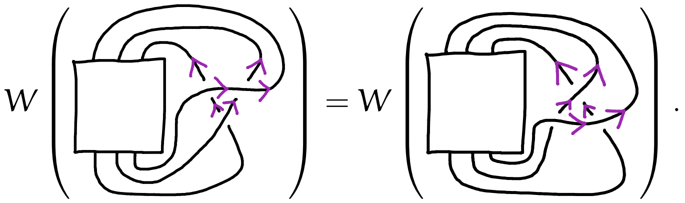
(4.6)



In other words, in every case, *applying a type II Reidemeister move does not alter the writhe.*

It remains only to consider type III Reidemeister moves. This is once again handled in the same manner as before, except there are now many more cases to consider. (In particular, since this deals with alterations to three branches of the diagram, there are more distinct cases of orientation configurations than in the preceding type II setting.) To keep the discussion reasonably brief, let us consider only one of the several cases of a type III move below, in equation (4.7):

(4.7)



By checking the signatures of all crossings on both sides, we see that (4.7) indeed holds. In fact, similar computations would also show that no type III move alters the writhe.

Finally, combining our observations from Figure 4.17 and (4.5)–(4.7), we conclude:

Theorem 4.4. *The writhe of a knot diagram is not altered when a type II or type III Reidemeister move is applied to the diagram. On the other hand, type I Reidemeister moves alter the writhes of knot diagrams in the manners indicated in Figure 4.17.*

4.4.2. *Links.* Before we can define the second preliminary quantity needed in order to define the Jones polynomial, we must first briefly expand our viewpoint beyond knots to objects called *links.* Recall that a knot models a piece of rope (with its ends glued to each other) that can be tied around itself in various ways. *Links*, on the other hand, *represent one or more pieces of rope which can be tied around itself as well as around each other.*

Before giving the formal definitions, we first provide some simple examples:

- The simplest instance of a link that is not a knot is the (two-component) unlink, given by two unknots that are separate from each other. See the first drawing in Figure 4.18.
- Just having two copies of the unknot is not particularly interesting. A more substantial example is given in the second drawing of Figure 4.18 and is known as the Hopf link. This again consists of two unknots, but they are now tied around each other.
- A more elaborate example, again crafted from unknots and resembling an Olympic rings logo gone horribly wrong, is given in the last drawing of Figure 4.18.

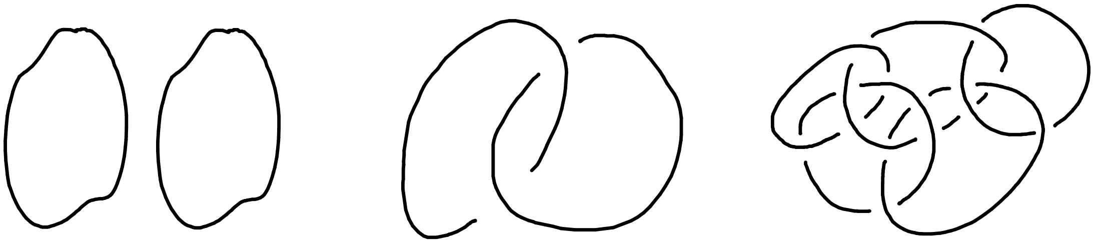


FIGURE 4.18. The drawings represent some simple examples of links. The link on the left contains two separated unknots. The middle diagram is the Hopf link, comprising of two unknots linked to each other. The link in the last diagram is crafted by tying five unknots together.

Regarding the formal definitions of links and their associated objects, this discussion proceeds in a manner analogous to knots. For completeness, we sketch the process below.

First, whereas knots are represented by a simple closed curve, links must be represented by a finite collection of disjoint simple closed curves. Like for knots, two such collections of curves are *link-equivalent* whenever one collection can be "continuously deformed" into the other.

Definition 4.13. *Let* L_1 *and* L_2 *be two collections of* m *disjoint simple closed space curves:*

$$\mathsf{L}_1 = (\mathsf{L}_{1,1}, \mathsf{L}_{1,2}, \dots, \mathsf{L}_{1,m}), \qquad \mathsf{L}_2 = (\mathsf{L}_{2,1}, \mathsf{L}_{2,2}, \dots, \mathsf{L}_{2,m}).$$

We say that L_1 *and* L_2 *are* link-equivalent *iff there exists a continuous*

$$\Phi : [0,1] \times \mathbb{R}^3 \to \mathbb{R}^3, \qquad \Phi(s,p) = \Phi_s(p),$$

such that:

- *Each* $\Phi_s : \mathbb{R}^3 \to \mathbb{R}^3$ *is a bijection between* $\mathbb{R}^3$ *and itself.*

- Φ_0 *maps each* $L_{1,i}$*, where* $1 \leq i \leq m$*, to itself.*
- Φ_1 *maps each* $L_{1,i}$ *to* $L_{2,i}$*, where* $1 \leq i \leq m$*.*

Here, the map Φ in Definition 4.13 plays the same role as the Φ in Definition 4.2. With this notion of link equivalence in hand, we can now formally define links:

Definition 4.14. *Consider the following equivalence relation* $\sim$*: two collections* L_1 *and* L_2 *of disjoint simple closed space curves satisfy* $L_1 \sim L_2$ *iff they are link-equivalent. Then, we define a link as an equivalence class of collections of curves under the above relation* $\sim$*.*

The next step is to construct an efficient method for graphically representing links. This leads us to the notion of *link diagrams*, which are direct analogues of knot diagrams:

Definition 4.15. *A link diagram is a projection of a collection of disjoint simple closed space curve onto* $\mathbb{R}^2$*, along with the following extra information at each crossing:*

- *The curve segment that is on top (the overcrossing) is drawn as usual.*
- *The curve segment that is on the bottom (the undercrossing) is drawn with a break.*

Moreover, two link diagrams D_1 *and* D_2 *are said to be link-equivalent iff they could be generated from two link-equivalent collections of simple closed curves.*

Example 4.16. *In particular, the drawings in Figure 4.18 are link diagrams, in the sense of Definition 4.15. Moreover, none of the three diagrams are link-equivalent.*

Given a link diagram, we can once again deform it without changing the underlying link structure. As was the case for knots, we can once again decompose such deformations into their most basic components. In fact, these are once again the *Reidemeister moves* described in Definition 4.6 and demonstrated in Figure 4.6. The only difference is that in the setting of links, these transformations can also happen at crossings made from two different curves of the diagram.

Perhaps unsurprisingly, a direct analogue of the Reidemeister theorem holds for links:

Theorem 4.5. *Two link diagrams are link-equivalent if and only if one can be transformed into the other via a finite sequence of Reidemeister moves.*

4.4.3. *The Kauffman Bracket.* We now discuss the second property of knot diagrams needed in order to define the Jones polynomial: the *Kauffman bracket.* In contrast to the quantities we have studied so far, the *Kauffman bracket takes polynomials as its values.* While this may seem strange at first, its practical effect is that we now have a much larger variety of values of assign to knot digrams. This will ultimately give us far more flexibility for distinguishing different knots.

We now give the formal definition of the Kauffman bracket. This is recursive by nature, in that the Kauffman bracket of a diagram is defined in terms of Kauffman brackets of simpler diagrams. One annoying but important point is that this recursive definition forces us to define Kauffman brackets for all link diagrams (see Definition 4.15), and not just knot diagrams. This is the reason for our excursion into links in the preceding discussions.

Definition 4.16. *We let* $B(D, x)$ *denote the Kauffman bracket of a link diagram* D*, with* x *and* x^{-1} *being the variables in the resulting polynomial.* B *is then defined via the following rules:*

(1) *Unknot: For the standard unknot diagram, we define*

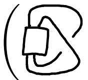
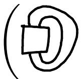
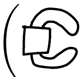
$$\mathrm{B}\left(\bigcirc, x\right) = 1. \tag{4.8}$$

(2) *Separation: Given an arbitrary link diagram* $\square$*, we have*

$$\mathrm{B}\left(\square\ \bigcirc, x\right) = -(x^2 + x^{-2}) \cdot \mathrm{B}\left(\square, x\right), \tag{4.9}$$

where $\bigcirc$ *again denotes the standard unknot diagram.*

(3) *Disentangling: The following rule holds:*

(4.10) $\mathrm{B}\Big($  $, x\Big) = x \cdot \mathrm{B}\Big($  $, x\Big) + x^{-1} \cdot \mathrm{B}\Big($  $, x\Big).$

In (4.10), the box denotes a portion of a link diagram (the same in all three pictures), with the remaining curve segments being arcs that are connected to arcs within the box.

Several clarifications regarding Definition 4.16 are in order. As mentioned before, the main idea is define the Kauffman bracket of a diagram in terms of simpler diagrams, with less crossings or components. Rule (1) provides the starting point of this recursion—for the simplest possible unknot diagram, we just define its Kauffman bracket to be the constant 1.

The role of rule (2) is to reduce the number of components in a link diagram. More specifically, given a link diagram $\square$ along with an unknot diagram $\bigcirc$ that is completely separated from $\square$, we can remove the $\bigcirc$ from the Kauffman bracket, at the cost of an extra factor of $-(x^2 + x^{-2})$. In particular, this reduces the number of components in the diagram by 1.

We still need a way to reduce the number of crossings in a diagram. This is accomplished through rule (3). The left-hand side of (4.10) contains the Kauffman bracket of a nontrivial diagram, with a single crossing highlighted. The right-hand side of (4.10) expresses this in terms of brackets of two other diagrams that have precisely this highlighted crossing removed. In particular, both diagrams here have one less crossing than the diagram in the left-hand side.

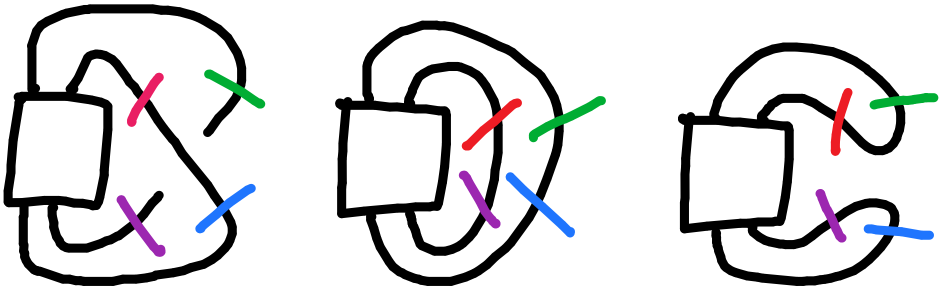


FIGURE 4.19. Link diagrams with coloured marks, demonstrating how a crossing is "cut" when applying the disentangling rule (3) in Definition 4.16.

Let us now look more closely more closely at how a crossing is removed in rule (3). We begin with the diagram on the left-hand side of (4.10). To aid in the explanation, let us colour the four ends of the crossing as in the left drawing in Figure 4.19:

- First, we align the overcrossing so that it goes toward the top-left and bottom-right corners. The top-left end is then coloured red, while the bottom-right end is coloured blue.
- Similarly, we align the undercrossing to go in the top-right and bottom-left directions. The top-right end is coloured green, while the bottom-right end is coloured purple.

To obtain the two diagrams on the right-hand side of (4.10), the idea is to imagine taking a pair of scissors, "cutting" the two arcs at the crossing, and then gluing the severed strings together:

- For the first diagram in the right-hand side of (4.10), we glue the red end to the purple end and the green end to the blue end. This is shown in the middle diagram in Figure 4.19.
- For the second diagram in the right-hand side of (4.10), we glue the red end to the green end and the blue end to the purple end. This is shown in the right diagram in Figure 4.19.

The next task is to derive formulas detailing how the Kauffman bracket is changed when a Reidemeister move is applied. These formulas will simplify a number of computations later on, and they will be essential in the definition of the upcoming Jones polynomial. These computations will also serve as basic examples for demonstrating how Kauffman brackets are computed.

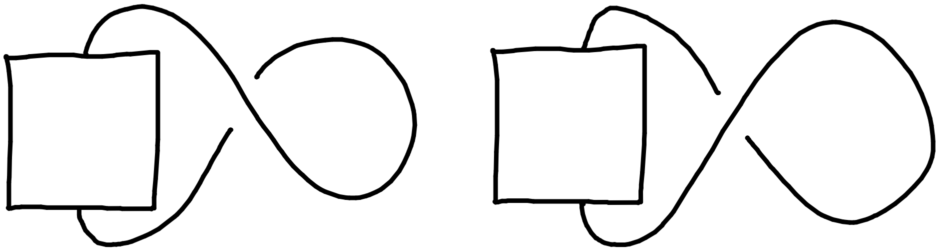
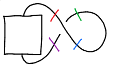
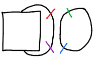
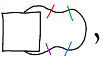


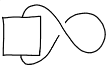
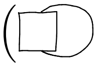
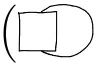
FIGURE 4.20. Two abstract link diagrams, each with a "loop".

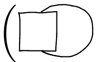
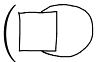
Let us begin with type I moves. First, consider the left link diagram in Figure 4.20. As before, the $\square$ denotes a portion of the diagram that will remain unchanged. Note the one curve segment attached to $\square$, with a "twist" that crosses itself. The specific problem, then, is to understand how the Kauffman bracket is affected by undoing this twist.

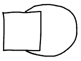
The first step is to apply rule (3) to this crossing. Note that the overcrossing is already aligned toward the upper-left and lower-right corners, so we can simply colour the four ends as in Figure 4.19. Applying (4.10) with the colouring scheme from Figure 4.19 yields

$\mathsf{B}\Big($  $, x\Big) = x \cdot \mathsf{B}\Big($  $, x\Big) + x^{-1} \cdot \mathsf{B}\Big($  $, x\Big).$

The first diagram on the right-hand side contains a separated unknot, which we can then eliminate using rule (2). Indeed, applying (4.9) along with some algebra results in the following:

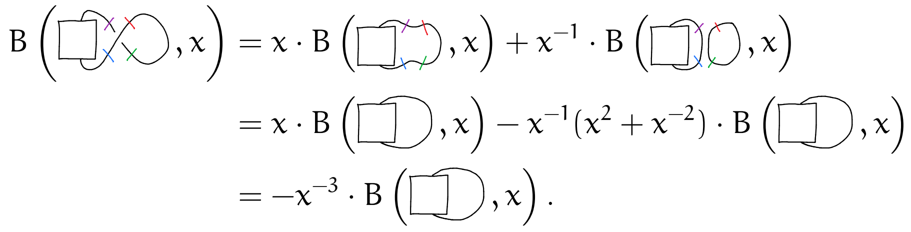
(4.11) $\mathsf{B}\Big($  $, x\Big) = -x(x^2 + x^{-2}) \cdot \mathsf{B}\Big($  $, x\Big) + x^{-1} \cdot \mathsf{B}\Big($  $, x\Big)$

$= (-x - x^{-1}) \cdot \mathsf{B}\Big($  $, x\Big) + x^{-1} \cdot \mathsf{B}\Big($  $, x\Big)$

$= -x^3 \cdot \mathsf{B}\Big($  $, x\Big).$

Similarly, consider the diagram on the right in Figure 4.20, which is similar but has with the opposite “twisting”. Again, we apply rule (3) and then (2) to eliminate the loop:

(4.12)

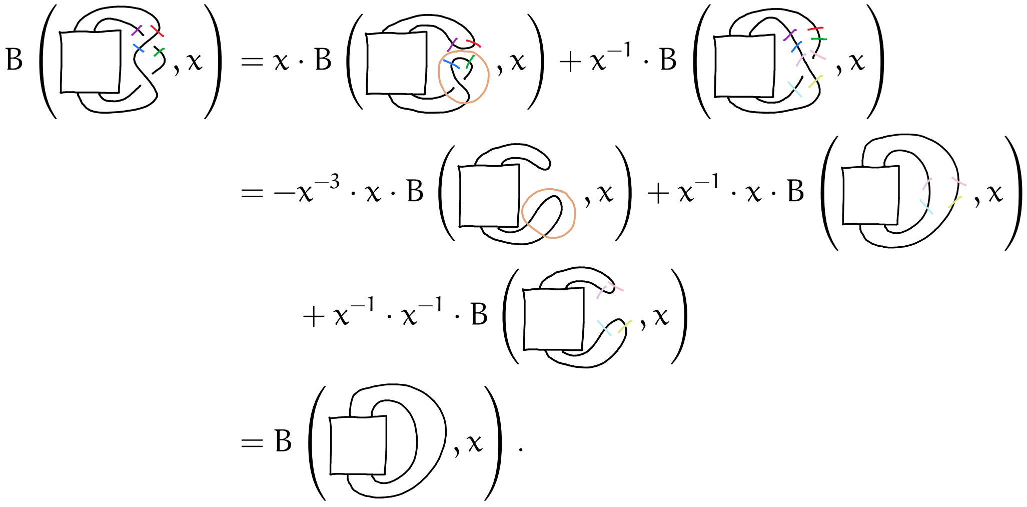


$$\begin{aligned} B\left(\quad , x\right) &= x \cdot B\left(\quad , x\right) + x^{-1} \cdot B\left(\quad , x\right) \\ &= x \cdot B\left(\quad , x\right) - x^{-1}(x^2 + x^{-2}) \cdot B\left(\quad , x\right) \\ &= -x^{-3} \cdot B\left(\quad , x\right). \end{aligned}$$

The only caveat is that in order to colour the four ends and apply rule (3), we have to first rotate the diagram 90 degrees, so that the overcrossing is aligned toward the top-left and bottom-right corners. As a result of this, the four ends in (4.12) are now coloured differently than in (4.11).

We now consider the setting of type II moves. One case of this is given below:

(4.13)



$$\begin{aligned} B\left(\quad , x\right) &= x \cdot B\left(\quad , x\right) + x^{-1} \cdot B\left(\quad , x\right) \\ &= -x^{-3} \cdot x \cdot B\left(\quad , x\right) + x^{-1} \cdot x \cdot B\left(\quad , x\right) \\ &\quad + x^{-1} \cdot x^{-1} \cdot B\left(\quad , x\right) \\ &= B\left(\quad , x\right). \end{aligned}$$

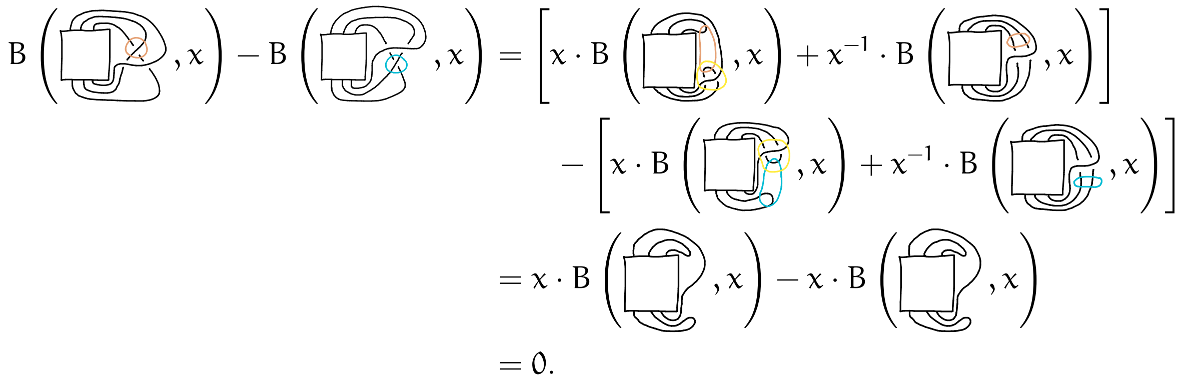
A more detailed explanation of the above steps is listed below:

- In the first equality, we applied rule (3) to the top crossing. For convenience, we added the same colouring scheme that was detailed in Figure 4.19.
- In the second equality, we applied the above formula (4.12) for type I Reidemeister moves to the first term to the area circled in beige. (Note that by rotating the diagram, we have the situation in (4.12), and not (4.11).) We also applied rule (3) to the last term; here, we adopted a different colouring scheme in order to distinguish from the previous step.
- For the last equality, we noticed that the first and last terms cancel.

In particular, observe that the diagrams in the left-hand and right-hand sides of (4.13) indeed differ by a type II Reidemeister move. The computations in (4.13) then show that *this type II move does not change the value of the Kauffman bracket.* Similar calculations can be applied to other examples of type II moves, with the same results; for brevity, we will not show these here. Thus, we conclude that all type II moves leave the Kauffman bracket unchanged.

In fact, the same conclusion holds for type III Reidemeister moves as well, as one can similarly compute that *type III moves do not alter the Kauffman bracket.* While it is rather painstaking to handle all the cases, one example is given in the subsequent computation:

(4.14)


B(, x) − B(, x) = [x · B(, x) + x^{−1} · B(, x)]
− [x · B(, x) + x^{−1} · B(, x)]
= x · B(, x) − x · B(, x)
= 0.

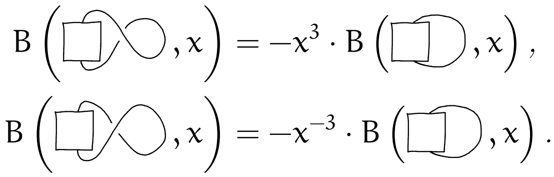
- In the first step, rule (3) is applied to each of the circled crossings on the left-hand side.
- In the second step, we handle the first and third terms by recalling that type II moves do not change the Kauffman bracket (for clarity, the affected areas are circled in yellow), and we note that the second and fourth terms cancel.

Again, other cases of type III moves are treated analogously, with the same result.

The above developments can now be summarised in the following theorem:

Theorem 4.6. *The Kauffman bracket of a link diagram is not altered when a type II or type III Reidemeister move is applied to the diagram. On the other hand, type I Reidemeister moves alter the Kauffman brackets of link diagrams in the following manners:*

(4.15)

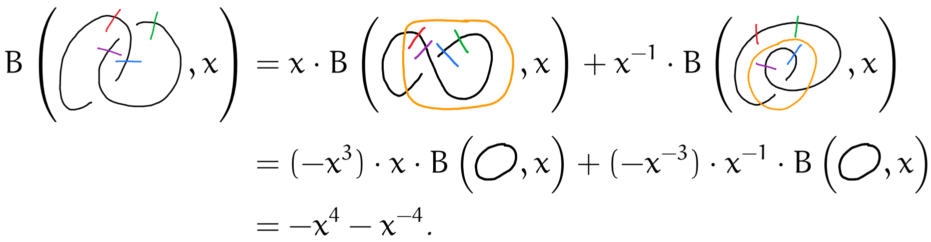


One confusing aspect of applying (4.15) is determining whether one obtains $-x^{+3}$ or $-x^{-3}$. This depends on how the given loop is oriented, and can be systematised via the following rules:

- The factor is $-x^{+3}$ whenever one traverses anticlockwise along the loop from the overcrossing to the undercrossing. This is the case for the first formula in (4.15).
- On the other hand, the factor is $-x^{-3}$ whenever one traverses clockwise along the loop from the overcrossing to the undercrossing. This is the case for the second formula in (4.15).

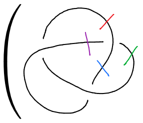
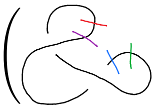
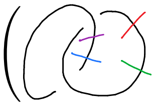
Finally, we compute the Kauffman bracket for two simple, concrete link diagrams:

Example 4.17. *First, we find the Kauffman bracket of the Hopf link, i.e. the middle diagram in Figure 4.18. For this, we apply rule (3), followed by the formulas* (4.15) *for type I moves:*

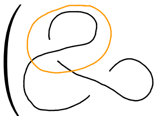
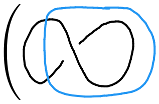
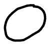

B(, x) = x · B(, x) + x^{−1} · B(, x)
= (−x^3) · x · B(, x) + (−x^{−3}) · x^{−1} · B(, x)
= −x^4 − x^{−4}.

In the first step, where rule (3) is applied, we coloured the four ends of the top crossing in the same manner as in Figure 4.19. In the second step, we applied both formulas in (4.15) *to the loops circled in orange. Note that in the first term, the first loop moves anticlockwise from the overcrossing to the undercrossing, resulting in a factor of* $-x^3$*. On the other hand, the loop in the second term goes clockwise from the overcrossing to the undercrossing, resulting in a factor of* $-x^{-3}$*. Finally, note that in the last step, we applied rule (1).*

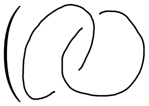
Example 4.18. *Next, we consider the standard trefoil diagram (that is, the second diagram in Figure 4.5). Applying rule (3) (again, the colourings are included for convenience), we obtain*

$B\Big($  $, x\Big) = x \cdot B\Big($  $, x\Big) + x^{-1} \cdot B\Big($  $, x\Big).$

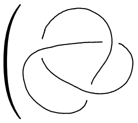
The first term on the right-hand side can be treated by applying (4.15) *twice:*

$x \cdot B\Big($  $, x\Big) = -x^4 \cdot B\Big($  $, x\Big) = x^7 \cdot B\Big($  $, x\Big) = x^7.$

The remaining term is the Hopf link, and we simply recall the result from Example 4.17:

$x^{-1} \cdot B\Big($  $, x\Big) = x^{-1}(-x^4 - x^{-4}) = -x^3 - x^{-5}.$

Finally, combining all of the above yields

$B\Big($  $, x\Big) = x^7 - x^3 - x^{-5}.$

4.4.4. *The Jones Polynomial.* We are now ready to define the Jones polynomial using the writhe and the Kauffman bracket. Note, however, that although the Jones polynomial is intended to be a knot invariant, both the writhe and the Kauffman bracket lack this property. Thus, we must first ask how a knot invariant could be generated out of these two non-invariant quantities.

The key observation is that both the writhe and the Kauffman bracket fail to be knot invariants in very specific ways. In particular, recall from Theorems 4.4 and 4.6 that:

- Both the writhe and the Kauffman bracket are invariant under type II and III moves.
- Neither the writhe nor the Kauffman bracket are invariant under type I moves.

The idea, then, is to combine the writhe and the Kauffman bracket in a way such that the change in the writhe via a type I move (see Figure 4.17) is exactly offset by the corresponding change in the Kauffman bracket (see (4.15)). With this in mind, we now define the following:

Definition 4.17. *Given a knot diagram* D*, we define its* *Jones polynomial* *to be*

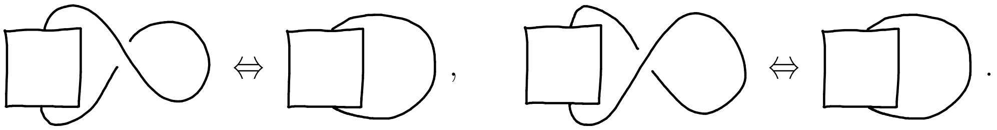
$$J(D, t) = \left(-t^{\frac{1}{4}}\right)^{3 \cdot W(D)} B\left(D, t^{\frac{1}{4}}\right). \tag{4.16}$$

In particular, since type I moves increase or decrease the writhe by 1, then in the right-hand side of (4.16), such a change in the writhe yields an extra factor of $(-t^{1/4})^{\pm 3}$. This precisely matches the change in $\mathsf{B}(\mathsf{D}, \mathsf{t}^{1/4})$ by the same type I move, hence the formula for $\mathsf{J}(\mathsf{D}, \mathsf{t})$ should remain unchanged overall. We show this rigorously in the following theorem:

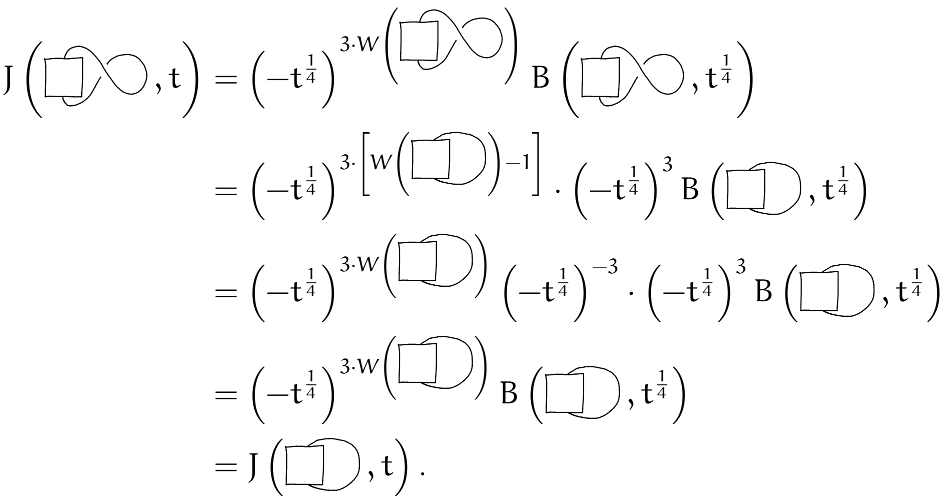
Theorem 4.7. *The Jones polynomial is a knot invariant.*

Proof. By the Reidemeister theorem (Theorem 4.1), we need only show that (4.16) does not change under any Reidemeister move. Moreover, since both the writhe and Kauffman bracket remain unchanged by type II and type III moves (by Theorems 4.4 and 4.6), the same also holds for (4.16). Thus, it remains only to show that (4.16) does not change under type I moves.

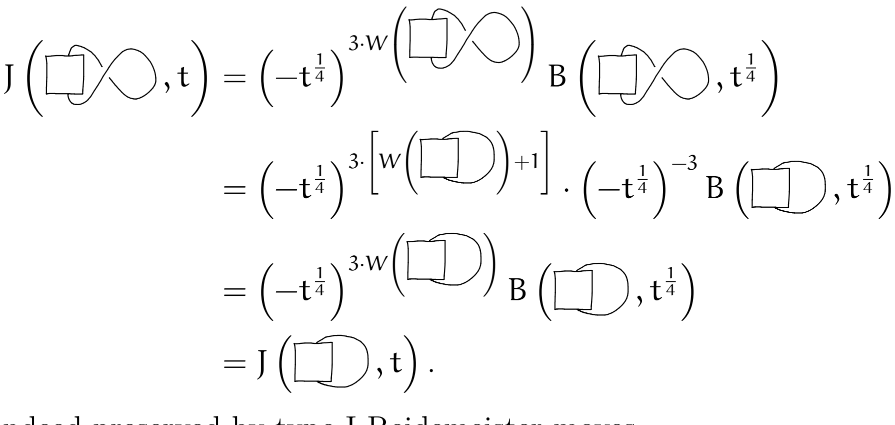
For this, we proceed abstractly and compute how J changes under the transformations



Applying the formulas in Figure 4.17 and (4.15), we see that



Similarly, for the opposite twisting, we have



Thus, (4.16) is indeed preserved by type I Reidemeister moves. □

Thanks to Theorem 4.7, we can now make unambiguous sense of the Jones polynomial of any knot, and not just of any knot diagram. This can be done in the obvious way:

Definition 4.18. *Given any knot* K*, we define its* <u>*Jones polynomial*</u> $\mathsf{J}(\mathsf{K}, \mathsf{t})$ *to simply be* $\mathsf{J}(\mathsf{D}, \mathsf{t})$*, where* D *denotes any knot diagram that represents* K*.*

Yet another aspect of the formula (4.16) that deserves explanation is the presence of the symbol $t^{1/4}$. This is, in fact, purely aesthetic. In practice, when one computes the Jones polynomial of knots, with x in the place of $t^{1/4}$, then one generally obtains powers of x that are multiples of 4. Thus, replacing x by $t^{1/4}$ has the effect of doing away with these extra 4's, leaving us with smaller but still integer exponents. We will see this in action in the examples below.

Example 4.19. *Let us first consider the unknot, which can be represented by the left diagram in Figure 4.5. Using this particular diagram, we first note that its writhe is*

$$W\left(\bigcirc\right) = 0,$$

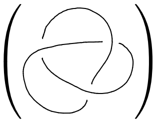
since it has no crossings. Furthermore, by rule (1) of Definition 4.16,

$$B\left(\bigcirc, x\right) = 1.$$

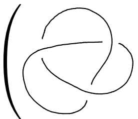
As a result, by Definition 4.17, we conclude that

$$J\left(\bigcirc, t\right) = \left(-t^{\frac{1}{4}}\right)^{3\cdot W\left(\bigcirc\right)} \cdot B\left(\bigcirc, x\right) = \left(-t^{\frac{1}{4}}\right)^{0} \cdot 1 = 1.$$

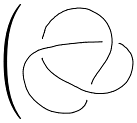
Example 4.20. *For a nontrivial example, we consider the trefoil knot, which can be represented by the right diagram in Figure 4.5. Using this diagram, we recall from Example 4.13 that*

$W\Bigg($  $\Bigg) = 3,$

and from Example 4.18 that

$B\Bigg($  $, x\Bigg) = x^7 - x^3 - x^{-5}.$

Thus, by Definition 4.17, we have

$J\Bigg($  $, t\Bigg)$

$$\begin{aligned} &= \left(-t^{\frac{1}{4}}\right)^{3\cdot 3}\left[\left(t^{\frac{1}{4}}\right)^{7} - \left(t^{\frac{1}{4}}\right)^{3} - \left(t^{\frac{1}{4}}\right)^{-5}\right] \\ &= \left[-\left(t^{\frac{1}{4}}\right)^{16} + \left(t^{\frac{1}{4}}\right)^{12} + \left(t^{\frac{1}{4}}\right)^{4}\right] \\ &= -t^4 + t^3 + t. \end{aligned}$$

In particular, the process of computing the Jones polynomial is a matter of combining the computations we have done before. As shown in Examples 4.19 and 4.20, one computes the writhe and the Kauffman bracket as before and then combines them using the formula in (4.16).

Finally, for a bit of historical perspective, the Jones polynomial was discovered in 1984 (with the results published in 1986) by Vaughan Jones (New Zealand and American mathematician, born 1952). This was the first new polynomial-valued knot invariant discovered in over sixty years, since the *Alexander polynomial* in 1923. For his work, Jones was awarded the *Fields Medal*—often nicknamed the "Nobel prize of mathematics"—in 1990.

4.4.5. *Some Loose Ends.* In this subsection, we conclude our brief introduction of knot theory by tying up some final loose ends (last knot pun, I promise).

The first piece of unfinished business goes back to our discussion on chirality. In Example 4.12, we claimed that the trefoil is chiral but avoided giving a proof. This was due to the difficulty in directly showing that the trefoil and reverse trefoil (see Figure 4.13) represent distinct knots. Below, we use the Jones polynomial to give a very simple proof of this fact.

The key observation is the following property of the Jones polynomial:

Proposition 4.8. *Let* K *be a knot, and let* $\tilde{\mathsf{K}}$ *be its mirror image. Then,*

$$J(\tilde{K}, t) = J(K, t^{-1}). \tag{4.17}$$

Proof sketch. We claim that the following two identities hold:

$$W(\tilde{K}) = -W(K), \tag{4.18}$$

$$B(\tilde{K}, x) = B(K, x^{-1}). \tag{4.19}$$

Assuming (4.18) and (4.19) for the moment, we conclude that

$$\begin{aligned} J(\tilde{K}, t) &= \left(-t^{\frac{1}{4}}\right)^{3\cdot W(\tilde{K})} B\left(\tilde{K}, t^{\frac{1}{4}}\right) \\ &= \left(-t^{\frac{1}{4}}\right)^{-3\cdot W(K)} B\left(K, \left(t^{\frac{1}{4}}\right)^{-1}\right) \\ &= \left[-(t^{-1})^{\frac{1}{4}}\right]^{3\cdot W(K)} B\left(K, (t^{-1})^{\frac{1}{4}}\right) \\ &= J(K, t^{-1}), \end{aligned}$$

which completes the proof of (4.17). Thus, it remains only to show (4.18) and (4.19).

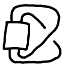
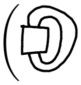
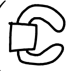
First, for (4.18), the main point is to recall that the mirror image $\tilde{K}$ is obtained by taking each crossing of K and interchanging the overcrossing and undercrossing (see Definition 4.10). This has the effect of negating the signature of each crossing (e.g. if a crossing of K has signature -1, then the corresponding crossing of $\tilde{K}$ has signature $+1$). Since the writhe is defined to be the sum of the signatures of all crossings, then the identity (4.18) follows.

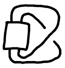
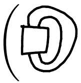
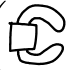
Next, for (4.19), the main observations are that:

- Rules (1) and (2) in Definition 4.16 are unaffected when x is replaced by x^{-1}:

$$B\left(\bigcirc, x^{-1}\right) = 1, \qquad B\left(\square\ \bigcirc, x^{-1}\right) = -(x^2 + x^{-2}) \cdot B\left(\square, x^{-1}\right). \tag{4.20}$$

- For rule (3), if the crossing in the left-hand side of (4.10) is reversed, then the two brackets on the right-hand side of (4.10) are interchanged:

(4.21) $B\Big($  $, x\Big) = x^{-1} \cdot B\Big($  $, x\Big) + x \cdot B\Big($  $, x\Big),$

$B\Big($  $, x^{-1}\Big) = x \cdot B\Big($  $, x^{-1}\Big) + x^{-1} \cdot B\Big($  $, x^{-1}\Big).$

A formal proof of (4.19) now relies on a combination of (4.20), (4.21), and an induction argument based on the recursive structure of Definition 4.16. However, to keep the discussion in these notes relatively simple and brief, we skip the technical details here. □

Proposition 4.8 can then be connected to chirality as follows:

Corollary 4.9. *Let* K *is a knot. If*

$$J(K, t) \neq J(K, t^{-1}), \tag{4.22}$$

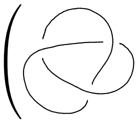
then K *must be chiral.*

Proof. Assume (4.22) holds, and suppose, for an eventual contradiction, that K is achiral, that is, K and its mirror image $\tilde{K}$ are the same knot. Since the Jones polynomial is a knot-invariant by Theorem 4.7, the mirror property of Proposition 4.8 implies that

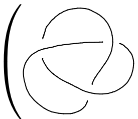
$$J(K, t) = J(\tilde{K}, t) = J(K, t^{-1}),$$

which contradicts our assumption (4.22). As a result, K is chiral. □

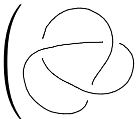
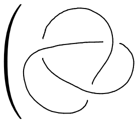
Example 4.21. *Let us return to the trefoil. Recall from Example 4.20 that*

$J\Bigg($  $, t\Bigg) = -t^4 + t^3 + t.$

Moreover, substituting in t^{-1} *in the place of* t *yields*

$J\Bigg($  $, t^{-1}\Bigg) = -t^{-4} + t^{-3} + t^{-1}.$

In particular, the above formulas show that

$J\Bigg($  $, t^{-1}\Bigg) \neq J\Bigg($  $, t\Bigg),$

thus Corollary 4.9 implies that the trefoil is chiral.

Remark. While Corollary 4.9 provides an effective test for whether a knot is chiral, *it cannot be similarly used to test whether a knot is achiral.* In particular, it is possible that a knot K satisfies the identity $J(K, t) = J(K, t^{-1})$, but K is nonetheless chiral.

Finally, let us return to our original problem (see Question 4.4) of *whether two knot diagrams represent the same knot.* The Jones polynomial, as a knot invariant, provides a partial answer to this question in the usual way: *if two knots* K_1 *and* K_2 *satisfy* $J(K_1, t) \neq J(K_2, t)$*, then* K_1 *and* K_2 *must be different knots.* The main contrast with our previous knot invariants—crossing number, tricolourability, chirality—is that the Jones polynomial takes far more possible values and hence does a better job of distinguishing knots, while still remaining (somewhat) computable.

However, the above is still a bit vague, and one can demand more specific answers—for instance, *just how good is the Jones polynomial at distinguishing knots*? One way to begin exploring this question is to just compute the Jones polynomial for all the simplest knots. With enough time and patience, one can obtain the following through brute force calculations:

(1) The Jones polynomial completely distinguishes between all knots with crossing number strictly less than 10. In other words, if K_1 and K_2 are two distinct knots with crossing numbers strictly less than 10, then $J(K_1, t) \neq J(K_2, t)$.
(2) On the other hand, the Jones polynomial fails to completely distinguish between all knots with crossing number at most 10. In other words, there do exist two distinct knots K_1' and K_2', with crossing number at most 10, such that $J(K_1', t) = J(K_2, t')$.

As a result, for the simplest knots—at least, those with less than 10 crossings—the Jones polynomial will definitively answer Question 4.4. However, for more complicated knots—with at least 10 crossings—the Jones polynomials will only give imperfect "educated guesses" to Question 4.4.

One can also do a similar accounting for the simpler Question 4.5: *can we determine whether a given knot is the unknot*? In particular, one can again ask if the Jones polynomial provides a definitive answer to this question. To be more specific, we ask the following:

Question 4.6. *If K is a knot, and $J(K, t) = 1$, then must K be the unknot? In other words, is the unknot the only knot with Jones polynomial equal to 1?*

Interestingly, *the answer to Question 4.6 is not known.* Thus far, researchers have computed the Jones polynomial for a wide variety of knots, simple and complex, and have not found any nontrivial knot with Jones polynomial 1. However, there could potentially be an even more complicated knot K out there with $J(K, t) = 1$. To this point, no one knows whether this is true or not.

In closing, the material discussed in this chapter is only a brief introduction to knot theory, which today is a vibrant area of mathematical research. Since the Jones polynomial in the 1980s, researchers have uncovered even more general invariants in order to better study knots and links. Examples that are actively used in recent research include the *HOMFLY polynomial* and the *Khovanov homology.* If you are interested in studying mathematical knots in further detail, there is certainly much more fascinating material to pursue! One place to start is the textbook [3].

5. Surfaces and Parametrisations

In the latter portion of this course, we move on from one-dimensional curves and turn our attention toward 2-dimensional surfaces. Like for curves, you probably have several intuitive ideas regarding what the notion of a surface should be. In this chapter, we discuss these in greater detail, and we explore how we can achieve a precise mathematical definition of surfaces. We also connect the geometry of surfaces to ideas in calculus and linear algebra.

5.1. Parametric Surfaces.

Before dealing with formal definitions, let us first list some real world contexts in which you have encountered the word "surface":

- Surface of a table.
- Surface of the earth.
- Surface of the water.

You have likely also seen texts refer to the following mathematical objects as "surfaces":

Example 5.1. *Consider the following subsets of $\mathbb{R}^3$:*

1. *The yz-plane, given by the equation $x = 0$.*
2. *The unit sphere, given by the equation $x^2 + y^2 + z^2 = 1$.*
3. *The paraboloid, given by the equation $z = x^2 + y^2$.*

Plots of these sets can be found in Figure 5.1.

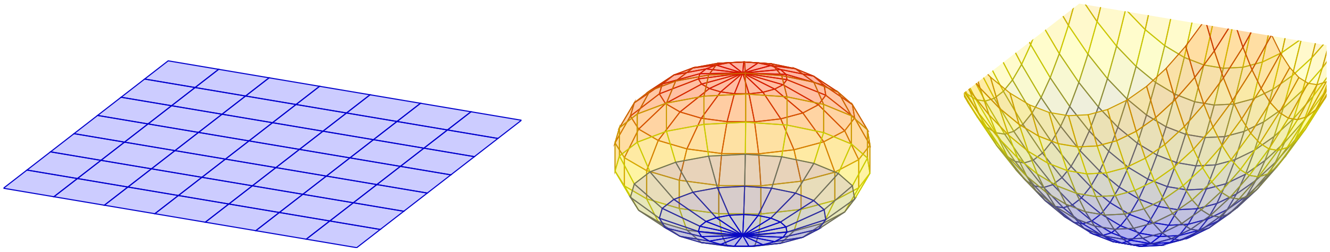


Figure 5.1. The surfaces (1), (2), (3), respectively, from Example 5.1.

What do these objects have in common? Your intuitions probably already indicate that each of the three sets in Example 5.1, with its corresponding diagram in Figure 5.1, is "2-dimensional" in nature. You also probably feel the same about the real world examples further above.

Thus, we can think of this dimensional property (which we have yet to define) as the common attribute linking all the above objects. Just as we intuitively characterised curves as 1-dimensional geometric objects, we can similarly *characterise surfaces as 2-dimensional geometric objects.* Of course, this immediately leads to the obvious question:

Question 5.1. *What exactly do we mean by "2-dimensional geometric objects"?*

Below, we explore this question in detail, beginning with the notion of parametric surfaces. We conclude this discussion in the next section with a formal definition of surfaces.

5.1.1. *Surface Parametrisations.* Let us begin by considering a function $f : \mathbb{R}^2 \to \mathbb{R}$ of two variables. To graph f, the usual convention is to represent it as the equation

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

that is, the xy-plane represents the domain of f, and the z-coordinate captures the value of f.

Example 5.2. *For example, the paraboloid in part (3) of Example 5.1 is such a graph, with*

$$f(x, y) = x^2 + y^2.$$

To be even more explicit, we can describe all the points on the graph of f as

$$G(f) = \{(x, y, f(x, y)) \mid x, y \in \mathbb{R}\}.$$

Here, x and y both vary freely over all real values, while each pair $(x, y) \in \mathbb{R}^2$ is associated with the z-value $f(x, y)$. Note that $G(f)$ can also be represented as the image of the function

$$\sigma : \mathbb{R}^2 \to \mathbb{R}^3, \qquad \sigma(u, v) = (u, v, f(u, v)). \tag{5.1}$$

Intuitively, one would view the graph of f as a "surface", i.e. a 2-dimensional object. From the representation σ, we begin to see why. Indeed, the 2-dimensional nature of σ is captured by its dependence on two variables, u and v. If we vary the parameters u and v over all the real numbers, then $\sigma(u, v)$ will trace out precisely all the points on the graph of f. In other words, σ shows that the graph of f is *parametrised* by two real variables.

Example 5.3. *The following are examples of graphs represented in the form* (5.1)*:*

$$\sigma_1 : \mathbb{R}^2 \to \mathbb{R}^3, \qquad \sigma_1(u, v) = (u, v, u - v),$$
$$\sigma_2 : \mathbb{R}^2 \to \mathbb{R}^3, \qquad \sigma_2(u, v) = (u, v, u^2 - v^2).$$

Plots for σ_1 and σ_2 are drawn below in Figure 5.2.

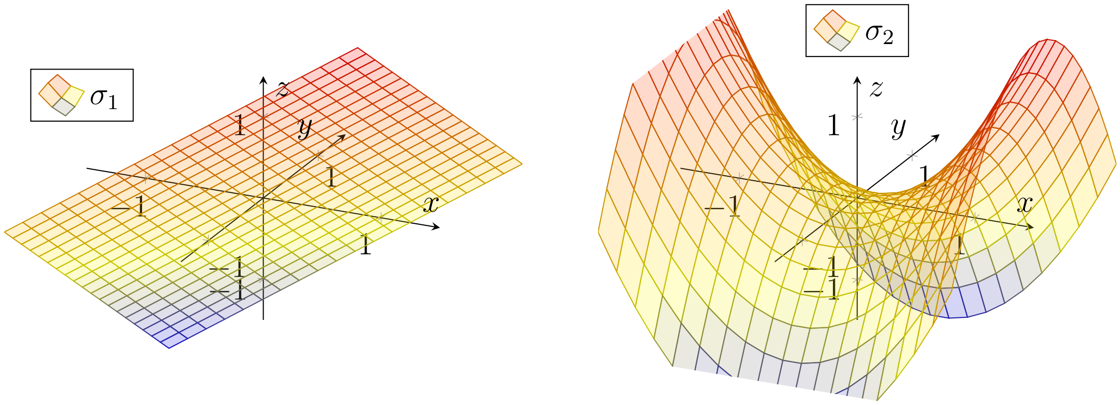


FIGURE 5.2. Plots for σ_1 (left) and σ_2 (right) in Example 5.3.

Now, suppose we switch the roles of the x, y, and z coordinates, and define

$$\sigma_y : \mathbb{R}^2 \to \mathbb{R}^3, \qquad \sigma_y(u, v) = (u, f(u, v), v),$$

$$\sigma_x : \mathbb{R}^2 \to \mathbb{R}^3, \qquad \sigma_x(u, v) = (f(u, v), u, v).$$

Neither σ_y nor σ_x needs be the graph of a function in the conventional sense (with the z-coordinate capturing the value of f). However, looking at Examples of such σ_y and σ_x below in Figure 5.4, your intuition would likely still indicate that they are "surfaces". Like for (5.1), the images of σ_y and σ_y are still parametrised by two variables, hence they retain their 2-dimensional nature.

Example 5.4. *Consider the following functions:*

$$\sigma_y : \mathbb{R}^2 \to \mathbb{R}^3, \qquad \sigma_y(u, v) = (u, u^2 + v^2, v),$$
$$\sigma_x : \mathbb{R}^2 \to \mathbb{R}^3, \qquad \sigma_x(u, v) = (u^2 + v^2, u, v).$$

Plots for σ_y and σ_x are given in Figure 5.3.

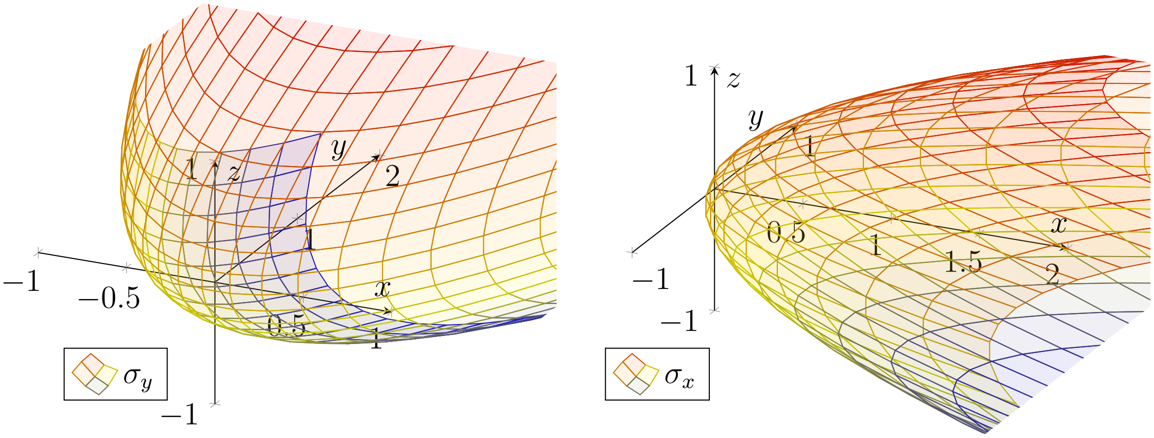


FIGURE 5.3. Plots for σ_y (left) and σ_x (right) in Example 5.4.

Now, the main feature in Examples 5.3 and 5.4 is that those objects are described using two parameters. In particular, to construct objects that seem "surface-like", there is no need for the parameters (say, u and v) to directly represent two of the Cartesian components x, y, z. Thus, we can expand our possibilities by considering more general maps of the form

$$\sigma : \mathbb{R}^2 \to \mathbb{R}^3, \qquad \sigma(u, v) = (\sigma_1(u, v), \sigma_2(u, v), \sigma_3(u, v)).$$

Example 5.5. *The <u>cylinder</u> $x^2 + y^2 = 1$ can described as the image of*

$$\sigma : \mathbb{R}^2 \to \mathbb{R}^3, \qquad \sigma(u, v) = (\cos u, \sin u, v).$$

A plot for this σ can be found in Figure 5.4.

If you have not had much experience with drawing 2-dimensional objects, then you may be wondering how this σ can be plotted. One simple strategy is to hold one of the two parameters, say u, constant. Then, by varying v, one obtains a curve that can be plotted using our previous experiences with graphing curves; in this case, these are vertical lines in the z-direction, whose position depends on the fixed value of u. By sampling sufficiently many values for u, one plots enough vertical lines so that the overall shape of the figure can be reasonably guessed.

Alternatively, if we fix v instead and vary u, then we obtain unit circles about the z-axis, whose z-heights depend on the chosen fixed value for v. Again, with enough values sampled for v, we can piece together these circles into the cylinder in Figure 5.4.

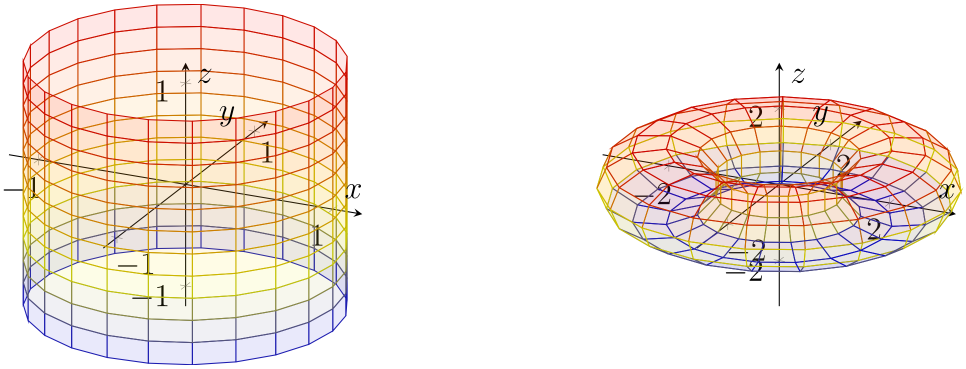


FIGURE 5.4. Plots for the functions in Examples 5.5 (left) and 5.6 (right).

Example 5.6. *One representation of the torus (the formal name for the doughnut-shaped object you saw in Figure 1.7) is given by the image of the map*

$$\sigma : \mathbb{R}^2 \to \mathbb{R}^3, \qquad \sigma(u, v) = ((2 + \cos u)\cos v, (2 + \cos u)\sin v, \sin u).$$

A plot for this σ is also found in Figure 5.4.

All our examples thus far (and also almost all the examples in the remainder of these notes) feature surfaces embedded in 3-dimensional space. However, like in our discussions for curves, a portion of the theory that we will study will also apply to surfaces embedded in higher dimensions. In other words, we can also consider functions of the form

$$\sigma : \mathbb{R}^2 \to \mathbb{R}^n, \qquad \sigma(u, v) = (\sigma_1(u, v), \sigma_2(u, v), \dots, \sigma_n(u, v)).$$

We will state which parts of the theory are special to surfaces in $\mathbb{R}^3$ and which apply more generally. However, if you prefer more concrete objects, you are safe to always assume $n = 3$.

Furthermore, it would be too constraining to require that these functions σ that we have been considering always be defined on all of $\mathbb{R}^2$. For example, one natural way to describe the upper hemisphere is via a function defined on the unit disk:

Example 5.7. *Consider now the upper (unit) hemisphere in $\mathbb{R}^3$, which can be described by*

$$x^2 + y^2 + z^2 = 1, \qquad z > 0.$$

Moreover, if we define B_0 to be the open unit disk in $\mathbb{R}^2$,

$$B_0 = \{(u, v) \in \mathbb{R}^2 \mid u^2 + v^2 < 1\},$$

then this upper hemisphere can also be characterised as the image of

$$\sigma : B_0 \to \mathbb{R}^3, \qquad \sigma(u, v) = (u, v, \sqrt{1 - u^2 - v^2}).$$

For a plot of this σ, see Figure 5.5.

Example 5.7 shows it is useful to consider maps whose domains are not all of $\mathbb{R}^2$. The natural question to ask, then, is what kinds of domains should be permitted.

Recall that for parametric curves, the 1-dimensional analogues for these σ, the domains we allowed were *open intervals*. Such intervals are convenient as domains because of two of their defining properties:

- They are "open", meaning roughly that they have no "boundary points", or "endpoints".
- They are "connected", i.e. given any two points on this domain, the line segment connecting the two points also lies in this domain.

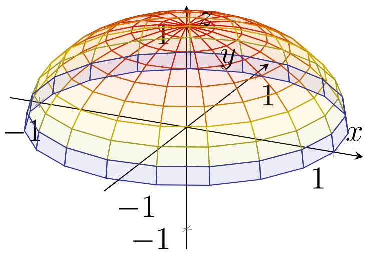


FIGURE 5.5. Plot for the function σ from Example 5.7.

The idea is that for our surface-parametrising functions σ, we wish to allow domains with corresponding properties of being "open" and "connected". Of course, we must first determine what these terms would mean in the 2-dimensional context. These are, in fact, fundamental questions in the field of mathematics known as topology. To not distract from our present discussion, here we will give only a very brief discussion of these questions.

First, we wish to characterise "openness" by a domain lacking boundary points. Consider now a subset $A \subseteq \mathbb{R}^2$. One way to describe a boundary point $\mathbf{p}$ of A is as follows: if we stand on this point $\mathbf{p}$, then there is some direction from $\mathbf{p}$ such that if we take even the smallest step in that direction, then we would no longer be in A. Therefore, we can describe $\mathbf{q}$ *not* being a boundary point of A by the property that if we were to take a small enough step from $\mathbf{q}$ in *any* direction, then we would still remain in A. This is formally captured in the following definition:

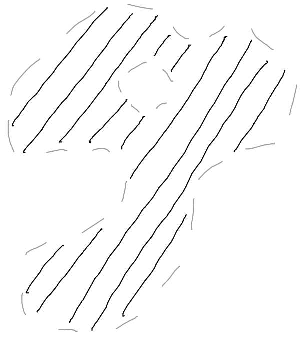


FIGURE 5.6. The shaded "blob" is a connected and open subset of $\mathbb{R}^2$.

Definition 5.1. *A subset $U \subseteq \mathbb{R}^2$ is said to open iff for any $\mathbf{q} \in U$, there is some open rectangle $R = (a, b) \times (c, d)$ such that $\mathbf{q} \in R$ and $R \subseteq U$. (In other words, U has no boundary points.)*

Example 5.8. *The following subsets of $\mathbb{R}^2$ are open:*

(1) *Any open rectangle $(a, b) \times (c, d)$.*
(2) *The unit disk, $B_0 = \{(u, v) \in \mathbb{R}^2 \mid u^2 + v^2 < 1\}$.*
(3) *Any disk $B(\mathbf{q}, r) = \{\mathbf{p} \in \mathbb{R}^2 \mid |\mathbf{p} - \mathbf{q}| < r\}$.*
(4) *Informally, any "blob shape without boundary"; see Figure 5.6 for an example.*

The second notion, "connectedness", is easier to define in this context. The only point to keep in mind is that on a 2-dimensional domain, one can connect two points in many more ways than just a line segment. As a result, we extend our previous characterisation in one dimension as follows:

Definition 5.2. *An open subset $U \subseteq \mathbb{R}^2$ is said to be connected iff for any two points $\mathbf{p}, \mathbf{q} \in U$, there is a curve, lying entirely in U, that connects $\mathbf{p}$ to $\mathbf{q}$.*

Remark. For more general discussions on open and connected sets, see [2, 7, 12].

After all these considerations, we finally arrive at our definition of parametrised surfaces:

Definition 5.3. *A parametric surface (also called a parametrisation in some contexts) is a smooth function $\sigma : U \to \mathbb{R}^n$, where U is a connected open subset of $\mathbb{R}^2$.*

In particular, all the objects in Examples 5.1-5.7 are parametric surfaces.

Remark. The term "smooth" in Definition 5.3 means that we can take as many *derivatives* of σ as we like; see the subsequent subsection for discussions on derivatives.

5.1.2. *Partial Derivatives.* Recall that for a parametric curve $\gamma : I \to \mathbb{R}^n$, which one can view as a particle travelling along a trajectory, its derivative γ' captures its velocity, or its rate of change of its position. Similarly, for a parametric surface $\sigma : U \to \mathbb{R}^n$, we can also consider its derivatives, although the specifics of this situation will be a bit different.

Since σ is a function of two variables, one would not differentiate it in the same manner as γ. Instead, a natural notion to consider is that of a partial derivative from multivariable calculus—by fixing one of the two variables, we can then view σ as a function of the single remaining variable and hence differentiate it as before. This leads us to the following definition:

Definition 5.4. *Let $\sigma : U \to \mathbb{R}^n$ be a parametric surface, and fix $(u_0, v_0) \in U$. We define the partial derivative of σ at (u_0, v_0) with respect to u and v, respectively, by*

$$\begin{aligned} \partial_u \sigma(u_0, v_0) &= \lim_{u \to u_0} \frac{\sigma(u, v_0) - \sigma(u_0, v_0)}{u - u_0}, \\ \partial_v \sigma(u_0, v_0) &= \lim_{v \to v_0} \frac{\sigma(u_0, v) - \sigma(u_0, v_0)}{v - v_0}. \end{aligned} \tag{5.2}$$

How can we interpret these partial derivatives? Consider the parametric curve in $\mathbb{R}^n$ given by

$$\gamma(u) = \sigma(u, v_0),$$

that is, we fix v_0 and vary only u in σ. In particular, γ is a parametric curve that is on lying on the image of σ. Now, if we take a derivative of γ at u_0, we then obtain

$$\gamma'(u_0) = \lim_{u \to u_0} \frac{\gamma(u) - \gamma(u_0)}{u - u_0} = \lim_{u \to u_0} \frac{\sigma(u, v_0) - \sigma(u, v_0)}{u - u_0} = \partial_u \sigma(u_0, v_0).$$

In other words, *we can view the partial derivative $\partial_u \sigma$ as the velocities of the family of curves on σ obtained by holding v constant at various values.*

Similarly, the parametric curve on σ given by

$$\lambda(v) = \sigma(u_0, v)$$

has its derivative given by

$$\lambda'(v_0) = \lim_{v \to v_0} \frac{\lambda(v) - \lambda(v_0)}{v - v_0} = \lim_{v \to v_0} \frac{\sigma(u_0, v) - \sigma(u_0, v_0)}{v - v_0} = \partial_v \sigma(u_0, v_0).$$

Thus, *we can view $\partial_v \sigma$ as the velocities of curves on σ obtained by holding u constant.* For a graphical demonstration of these curves γ and λ, see Figure 5.7 below.

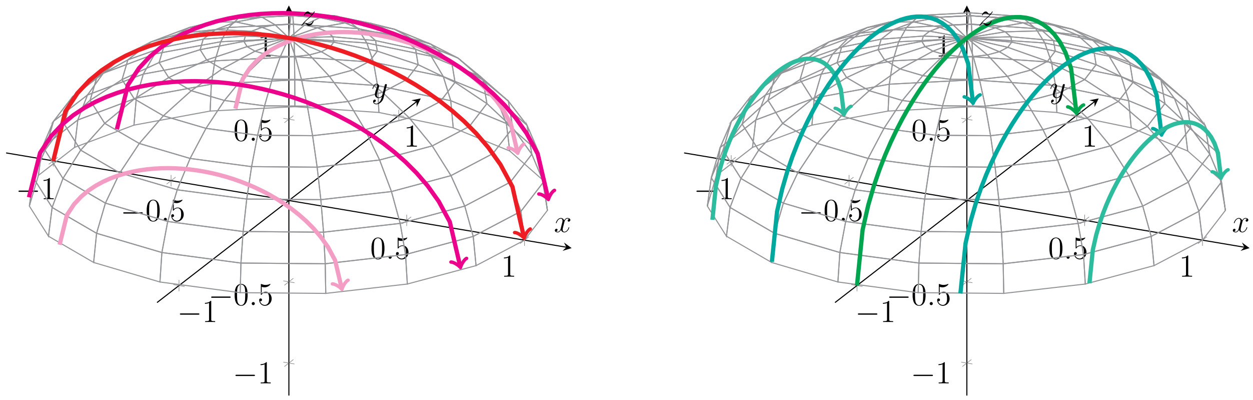


FIGURE 5.7. Two copies of the upper hemisphere from Example 5.7. In the left diagram, some curves γ, with v held constant, are drawn; in the right diagram, some curves λ, with u held constant, are drawn.

Example 5.9. *Let σ be the parametrised upper hemisphere from Example 5.7 and Figure 5.5,*

$$\sigma : B_0 \to \mathbb{R}^3, \qquad \sigma(u, v) = (u, v, \sqrt{1 - u^2 - v^2}),$$

where B_0 is the unit disk about the origin in $\mathbb{R}^2$. In terms of the above language, we then have

$$\gamma(u) = \sigma(u, v_0) = \left(u, v_0, \sqrt{(1 - v_0^2) - u^2} \right),$$

$$\lambda(v) = \sigma(u_0, v) = \left(u_0, v, \sqrt{(1 - u_0^2) - v^2} \right).$$

Various instances of γ's and λ's are plotted in Figure 5.7. In particular, the $\partial_u \sigma$'s and $\partial_v \sigma$'s are given precisely as the velocities of these red and green curves, respectively.

Now that we have an intuitive idea of what these partial derivatives represent, let us discuss how these can be computed. Here, the situation is completely analogous to that for calculating derivatives of parametric curves: we can differentiate σ componentwise.

Theorem 5.1. *Let $\sigma : U \to \mathbb{R}^n$ be a parametric surface, and let*

$$\sigma(u, v) = (\sigma_1(u, v), \sigma_2(u, v), \dots, \sigma_n(u, v)) \in \mathbb{R}^n,$$

where the real-valued $\sigma_k : U \to \mathbb{R}^n$ represents the k-th component of σ. Then, for any $(u_0, v_0) \in U$,

$$\begin{aligned} \partial_u \sigma(u_0, v_0) &= (\partial_u \sigma_1(u_0, v_0), \dots, \partial_u \sigma_n(u_0, v_0)), \\ \partial_v \sigma(u_0, v_0) &= (\partial_v \sigma_1(u_0, v_0), \dots, \partial_v \sigma_n(u_0, v_0)), \end{aligned} \tag{5.3}$$

where the symbols ∂_u and ∂_v in the right-hand sides of equation (5.3) denote the standard partial derivatives of real-valued functions in multivariable calculus.

Proof. The proof is like that of Theorem 2.2. To demonstrate, let us take $\partial_u \sigma$ when $n = 3$ for simplicity. (The remaining cases can be proved similarly.) Since limits are taken componentwise,

$$\partial_u \sigma(u_0, v_0) = \lim_{u \to u_0} \frac{\sigma(u, v_0) - \sigma(u_0, v_0)}{u - u_0}$$

$$= \lim_{u \to u_0} \left(\frac{\sigma_1(u, v_0) - \sigma_1(u_0, v_0)}{u - u_0}, \frac{\sigma_2(u, v_0) - \sigma_2(u_0, v_0)}{u - u_0}, \frac{\sigma_3(u, v_0) - \sigma_3(u_0, v_0)}{u - u_0} \right)$$
$$= (\partial_u \sigma_1(u_0, v_0), \partial_u \sigma_2(u_0, v_0), \partial_u \sigma_3(u_0, v_0)).$$

In the last equality, we used the definition of standard partial derivatives. □

In other words, to take a partial derivative of σ, we need only take partial derivatives of its individual components σ_k. The latter task is a basic tenet of multivariable and vector calculus; most of you will have already encountered this in MTH4101/4201: Calculus II.

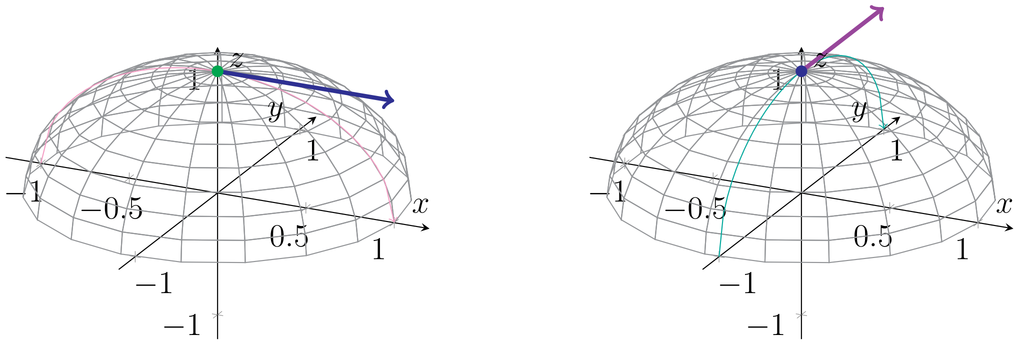


Figure 5.8. The plots contain the setup from Example 5.10. The north pole $\sigma(0,0)$ is marked in each plot. The left diagram depicts $\partial_u \sigma(0,0)$ (blue) as the derivative of $\gamma(u) = \sigma(u, 0)$ (pink), while the right diagram depicts $\partial_v \sigma(0,0)$ (purple) as the derivative of $\lambda(v) = \sigma(0, v)$ (teal).

Example 5.10. *Let us return to the* σ *from Example 5.9,*

$$\sigma : B_0 \to \mathbb{R}^3, \qquad \sigma(u, v) = (u, v, \sqrt{1 - u^2 - v^2}).$$

Let us now compute both of its partial derivatives.

To compute $\partial_u \sigma$*, we apply Proposition 5.1, which tells us that we need only hold* v *constant and then differentiate each component of* σ *with respect to* u *separately. From this, we obtain*

$$\partial_u \sigma(u, v) = \left(1, 0, \frac{1}{2} \cdot \frac{\partial_u (1 - u^2 - v^2)}{\sqrt{1 - u^2 - v^2}} \right) = \left(1, 0, -\frac{u}{\sqrt{1 - u^2 - v^2}} \right).$$

(Note that we applied both the power rule and the chain rule in the above line.) Next, holding u *constant and differentiating with respect to* v*, we similarly compute that*

$$\partial_v \sigma(u, v) = \left(0, 1, \frac{1}{2} \cdot \frac{\partial_v (1 - u^2 - v^2)}{2\sqrt{1 - u^2 - v^2}} \right) = \left(0, 1, -\frac{v}{\sqrt{1 - u^2 - v^2}} \right).$$

In particular, at $(u_0, v_0) = (0, 0)$*, which corresponds to the north pole*

$$\sigma(u_0, v_0) = \sigma(0, 0) = (0, 0, 1),$$

we have that

$$\partial_u \sigma(0, 0) = (1, 0, 0), \qquad \partial_v \sigma(0, 0) = (0, 1, 0).$$

These vectors are plotted in Figure 5.8.

Example 5.11. *Consider the (parametric) torus from Example 5.6,*

$$\sigma : \mathbb{R}^2 \to \mathbb{R}^3, \qquad \sigma(u,v) = ((2+\cos u)\cos v, (2+\cos u)\sin v, \sin u).$$

To compute $\partial_u\sigma$ and $\partial_v\sigma$, we again take the corresponding partial derivatives of each component on the right-hand side. Fixing v and differentiating with respect to u, we have

$$\partial_u\sigma(u,v) = (-\sin u\cos v, -\sin u\sin v, \cos u).$$

Similarly, fixing u and differntiating with respect to v yields

$$\partial_v\sigma(u,v) = (-(2+\cos u)\sin v, (2+\cos u)\cos v, 0).$$

Remark. Other constructions in calculus, such as directional derivatives and gradients, also have analogues in this setting. However, for simplicity, we will develop our theory in terms of partial derivatives, as they represent the most elementary objects of analysis.

5.1.3. *Tangent Directions.* Recall that for a parametric curve $\gamma : I \to \mathbb{R}^n$, the tangent vector $\gamma'(t)|_{\gamma(t)}$ represents the direction and speed that γ is going at parameter t. Furthermore, when γ is regular, $\gamma'(t)|_{\gamma(t)}$ generates the full tangent line $T_\gamma(t)$ at $\gamma(t)$; see Definition 2.10.

Remark. Recall that tangent vectors were defined in Definition 2.9, while general "arrows"—vectors based at a point—were discussed earlier in that same section. From here forward, we will use these concepts regularly, as they are quite useful for intuition. Thus, the reader may want to revise their knowledge and understanding of these concepts.

We can then ask what the analogous structure is for parametric surfaces:

Question 5.2. *Given a parametric surface, how do we describe the directions tangent to it? Moreover, what structures do these sets of directions have?*

To investigate this, we let $\sigma : U \to \mathbb{R}^n$ denote a parametric surface. Furthermore, we fix a point $(u_0, v_0) \in U$, so that $\mathbf{p}_0 = \sigma(u_0, v_0) \in \mathbb{R}^n$ represents a point on this parametric surface.

Recall from the preceding subsection that

$$\gamma(u) = \sigma(u, v_0), \qquad \lambda(v) = \sigma(u_0, v)$$

define parametric curves lying in the image of σ. Moreover, at $u = u_0$, the tangent vector of γ is

$$\gamma'(u_0)|_{\gamma(u_0)} = \partial_u\sigma(u_0, v_0)|_{\sigma(u_0,v_0)} = \partial_u\sigma(u_0, v_0)|_{\mathbf{p}_0}.$$

Since the above is tangent to γ, which is lying in σ, it follows that the "arrow" $\partial_u\sigma(u_0, v_0)|_{\mathbf{p}_0}$ is tangent to σ at the point $\mathbf{p}_0$ as well. Similarly, at $v = v_0$, the tangent vector of λ is

$$\lambda'(v_0)|_{\lambda(u_0)} = \partial_v\sigma(u_0, v_0)|_{\sigma(u_0,v_0)} = \partial_v\sigma(u_0, v_0)|_{\mathbf{p}_0},$$

hence $\partial_v\sigma(u_0, v_0)|_{\mathbf{p}_0}$ is also tangent to σ at the point $\mathbf{p}_0$.

Combining the above observations, we conclude that $\partial_u\sigma(u_0, v_0)|_{\mathbf{p}_0}$ *and* $\partial_v\sigma(u_0, v_0)|_{\mathbf{p}_0}$ *represent two directions which are tangent to σ at $\mathbf{p}_0 = \sigma(u_0, v_0)$.*

Example 5.12. *Let us go back to the upper hemisphere from Examples 5.9 and 5.10,*

$$\sigma : B_0 \to \mathbb{R}^3, \qquad \sigma(u, v) = (u, v, \sqrt{1 - u^2 - v^2}).$$

We now retrace the above discussion for this σ*, with*

$$(u_0, v_0) = (0, 0), \qquad \mathbf{p}_0 = \sigma(u_0, v_0) = (0, 0, 1).$$

Recalling the computations we already made in Example 5.10, we see that two "arrows" which are tangent to σ *at the point* $\mathbf{p}_0 = (0, 0, 1)$ *are given by*

$$\partial_u \sigma(0, 0)|_{\mathbf{p}_0} = (1, 0, 0)|_{(0,0,1)}, \qquad \partial_v \sigma(0, 0)|_{\mathbf{p}_0} = (0, 1, 0)|_{(0,1,0)}.$$

Within Figure 5.8, these are represented by the blue ($\partial_u \sigma$*) and purple (*$\partial_v \sigma$*) arrows, respectively.*

Returning once again to the general case, we note that while $\partial_u \sigma(u_0, v_0)|_{\mathbf{p}_0}$ and $\partial_v \sigma(u_0, v_0)|_{\mathbf{p}_0}$ potentially yield us two tangent directions to σ at $\mathbf{p}_0$, this does not yet tell us what other tangent directions remain to be found. Recall that the goal from Question 5.2 is to find *all* tangent directions to σ at $\mathbf{p}_0$. For this, we will have to take a more general outlook.

The idea is now to take another parametric curve α lying on the image of σ that passes through this point $\mathbf{p}_0$. By measuring its derivative α' at $\mathbf{p}_0$, we capture yet another such tangent direction of σ. Therefore, by doing this for *every possible* α, we will obtain all tangent directions.

Now, since $\alpha(t)$ lies in the image of σ, we can write, for any t,

$$\alpha(t) = \sigma(u(t), v(t))$$

for some real-valued functions u and v. In particular, these functions u and v contain the *components* of α, with respect to the σ-parametrisation. See Figure 5.9 for a graphical representation; the plane curve $t \mapsto (u(t), v(t))$ is graphed in red in the left diagram (the uv-plane), while the corresponding curve α is plotted in the right diagram (the image of σ).

Suppose, for convenience, that $u(0) = u_0$ and $v(0) = v_0$, so that

$$\alpha(0) = \sigma(u_0, v_0) = \mathbf{p}_0.$$

Then, measuring $\alpha'(0)|_{\alpha(0)}$ produces a tangent direction to σ at $\mathbf{p}_0$. To find $\alpha'(0)$, we first observe

$$\alpha'(0) = \frac{d}{dt}[\sigma(u(t), v(t))]\bigg|_{t=0}.$$

For the right-hand side, we recall the (multivariable) chain rule, which yields

$$\begin{aligned} \alpha'(0) &= \partial_u \sigma(u(t), v(t)) \cdot u'(t)|_{t=0} + \partial_v \sigma(u(t), v(t)) \cdot v'(t)|_{t=0} \\ &= u'(0) \cdot \partial_u \sigma(u_0, v_0) + v'(0) \cdot \partial_v \sigma(u_0, v_0), \end{aligned}$$

and hence

$$\alpha'(0)|_{\alpha(0)} = u'(0) \cdot \partial_u \sigma(u_0, v_0)|_{\mathbf{p}_0} + v'(0) \cdot \partial_v \sigma(u_0, v_0)|_{\mathbf{p}_0}. \tag{5.4}$$

In Figure 5.9, the quantities $\alpha'(0)|_{\alpha(0)}$, $\partial_u \sigma(u_0, v_0)|_{\mathbf{p}_0}$, and $\partial_v \sigma(u_0, v_0)|_{\mathbf{p}_0}$ are drawn in the right diagram in orange, blue, and purple, respectively. The corresponding objects in the uv-plane are drawn in the left diagram using the same colours.

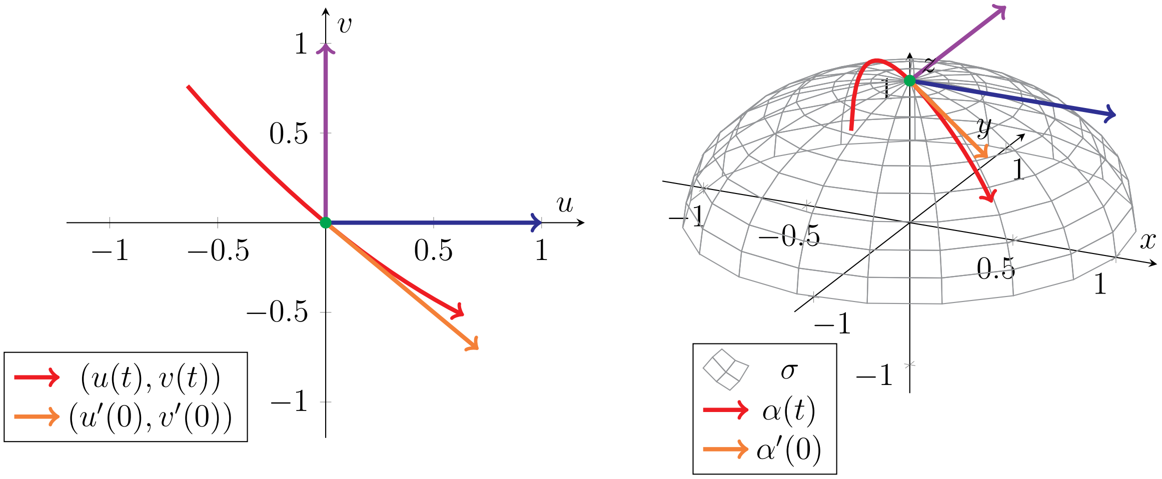


FIGURE 5.9. Again, the upper hemisphere from Examples 5.9, 5.10, and 5.12. The right diagram shows a parametric curve α passing through the north pole $\alpha(0) = (0, 0, 1)$. The vectors $\alpha'(0)|_{(0,0,1)}$ (orange), $\partial_u \sigma(0, 0)|_{(0,0,1)}$ (blue), and $\partial_v \sigma(0, 0)|_{(0,0,1)}$ (green) are also shown. The left diagram shows the corresponding picture in the uv-plane; for instance, the parametric curve $(u(t), v(t))$ is shown in red. Objects with a given colour in the left diagram map, via σ, to the object with the same colour in the right.

Let us now parse the relation (5.4) to understand what we have done:

- On the left-hand side is the tangent vector $\alpha'(0)|_{\mathbf{p}_0}$, which is tangent to σ.
- On the right-hand side is a *linear combination* of the two "special" tangent directions $\partial_u \sigma(u_0, v_0)|_{\mathbf{p}_0}$ and $\partial_v \sigma(u_0, v_0)|_{\mathbf{p}_0}$ that we found. The quantities $u'(0)$ and $v'(0)$ are two constants which are determined by the details of α.

Furthermore, the above computation applies to *any arbitrary* α that goes through $\mathbf{p}_0$, hence *any tangent direction* to σ at $\mathbf{p}_0$ can be described using (5.4).

As a result, (5.4) can now be interpreted as follows: *any tangent direction to* σ *at* $\mathbf{p}_0 = \sigma(u_0, v_0)$ *can be captured as a linear combination of the tangent vectors* $\partial_u \sigma(u_0, v_0)|_{\mathbf{p}_0}$ *and* $\partial_v \sigma(u_0, v_0)|_{\mathbf{p}_0}$.

From linear algebra, you should know that if you take all linear combinations of two vectors, i.e.

$$S = \{a\mathbf{v} + b\mathbf{w} \mid a, b \in \mathbb{R}\}, \qquad \mathbf{v}, \mathbf{w} \in \mathbb{R}^n,$$

then this resulting collection S will generally be a 2-*dimensional (vector) space*, that is, a *plane*. This argument finally leads us to the following definition:

Definition 5.5. *Let* $\sigma : U \to \mathbb{R}^n$ *be a parametric surface, and let* $(u_0, v_0) \in U$*. Then, the* <u>*tangent plane*</u> *of* σ *at* $\mathbf{p}_0 = \sigma(u_0, v_0)$ *(or alternatively, at parameter* (u_0, v_0)*) is defined to be*

$$\begin{aligned} T_\sigma(u_0, v_0) &= \{a \cdot \partial_u \sigma(u_0, v_0)|_{\mathbf{p}_0} + b \cdot \partial_v \sigma(u_0, v_0)|_{\mathbf{p}_0} \mid a, b \in \mathbb{R}\} \\ &= \operatorname{span}\{\partial_u \sigma(u_0, v_0)|_{\mathbf{p}_0}, \partial_v \sigma(u_0, v_0)|_{\mathbf{p}_0}\}. \end{aligned} \tag{5.5}$$

An element $\mathbf{v}|_{\mathbf{p}_0} \in T_\sigma(u_0, v_0)$ *is called a* <u>*tangent vector*</u> *of* σ *at* $\mathbf{p}_0$.

Example 5.13. *In the case of the upper hemisphere,*

$$\sigma : B_0 \to \mathbb{R}^3, \qquad \sigma(u, v) = (u, v, \sqrt{1 - u^2 - v^2}),$$

we computed in Exercise 5.10 that

$$\partial_u \sigma(0, 0)|_{(0,0,1)} = (1, 0, 0), \qquad \partial_v \sigma(0, 0)|_{(0,0,1)} = (0, 1, 0).$$

Thus, by Definition 5.5, the tangent plane of σ *at* $(0, 0, 1)$ *is*

$$\begin{aligned} T_\sigma(0, 0) &= \{a \cdot \partial_u \sigma(u_0, v_0)|_{(0,0,1)} + b \cdot \partial_v \sigma(u_0, v_0)|_{(0,0,1)} \mid a, b \in \mathbb{R}\} \\ &= \{a \cdot (1, 0, 0)|_{(0,0,1)} + b \cdot (0, 1, 0)|_{(0,0,1)} \mid a, b \in \mathbb{R}\} \\ &= \{(a, b, 0)|_{(0,0,1)} \mid a, b \in \mathbb{R}\}. \end{aligned}$$

$T_\sigma(0, 0)$ *and some tangent vectors there are drawn in Figure 5.10 (on the left).*

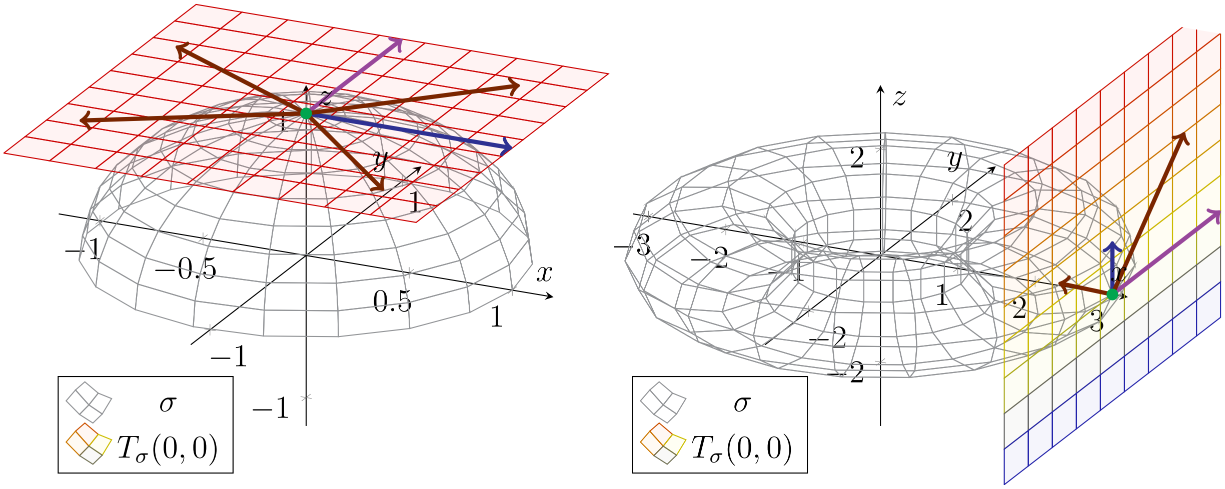


FIGURE 5.10. The left diagram shows the tangent plane $T_0\sigma(0, 0)$ from Example 5.13. The tangent vectors $(1, 0, 0)|_{(0,0,1)}$ and $(0, 1, 0)|_{(0,0,1)}$ are drawn in blue and purple, as usual; other examples of tangent vectors are depicted in brown. The right diagram shows the corresponding tangent plane $T_\sigma(0, 0)$ from Example 5.15 below, along with some tangent vectors.

Finally, like for the tangent line of a curve, we can also equivalently describe the tangent plane as a set of points. In the case of Definition 5.5, this would be the plane passing through $\mathbf{p}_0$ which contains the directions $\partial_u \sigma(u_0, v_0)|_{\mathbf{p}_0}$ and $\partial_v \sigma(u_0, v_0)|_{\mathbf{p}_0}$. This can be described as follows:

Definition 5.6. *Let* σ *and* (u_0, v_0) *be as before, in Definition 5.5. Then, we can also define the* *tangent plane* *of* σ *at* $\mathbf{p}_0 = \sigma(u_0, v_0)$ *to be the set*

$$\mathcal{T}_\sigma(u_0, v_0) = \{\mathbf{p}_0 + a \cdot \partial_u \sigma(u_0, v_0) + b \cdot \partial_v \sigma(u_0, v_0) \mid a, b \in \mathbb{R}\}. \tag{5.6}$$

Example 5.14. *Continuing from Example 5.13, the tangent plane of* σ *at* $(0, 0, 1)$, *described in terms of Definition 5.6 as a set of points, is then*

$$\mathcal{T}_\sigma(0, 0) = \{(0, 0, 1) + a \cdot (1, 0, 0) + b \cdot (0, 1, 0) \mid a, b \in \mathbb{R}\}$$

$$= \{(a, b, 1) \mid a, b \in \mathbb{R}\}.$$

Visually, this is also described through the left plot in Figure 5.10.

Example 5.15. *Let us return to the torus from Example 5.6,*

$$\sigma : \mathbb{R}^2 \to \mathbb{R}^3, \qquad \sigma(u, v) = ((2 + \cos u) \cos v, (2 + \cos u) \sin v, \sin u).$$

For the sake of concreteness, we now take

$$(u_0, v_0) = (0, 0), \qquad \mathbf{p}_0 = \sigma(u_0, v_0) = (3, 0, 0).$$

Recall from Example 5.11 that

$$\partial_u \sigma(u, v) = (-\sin u \cos v, -\sin u \sin v, \cos u),$$
$$\partial_v \sigma(u, v) = (-(2 + \cos u) \sin v, (2 + \cos u) \cos v, 0).$$

In particular, at $(u_0, v_0) = (0, 0)$*, we have*

$$\partial_u \sigma(0, 0)|_{\mathbf{p}_0} = (0, 0, 1)|_{(3,0,0)}, \qquad \partial_v \sigma(0, 0)|_{\mathbf{p}_0} = (0, 3, 0)|_{(3,0,0)}.$$

Thus, the tangent plane of σ *at* $\mathbf{p}_0 = (3, 0, 0)$ *(according to both Definitions 5.5 and 5.6), is*

$$T_\sigma(0, 0) = \{a(0, 0, 1)|_{(3,0,0)} + b(0, 3, 0)|_{(3,0,0)} \mid a, b \in \mathbb{R}\} = \{(0, a, b)|_{(3,0,0)} \mid a, b \in \mathbb{R}\},$$
$$\mathcal{T}_\sigma(0, 0) = \{(3, 0, 0) + a(0, 0, 1) + b(0, 3, 0) \mid a, b \in \mathbb{R}\} = \{(3, a, b) \mid a, b \in \mathbb{R}\}.$$

A plot of this is given in the right plot of Figure 5.10.

Remark. The characterisation of tangent planes in Definition 5.6 (as a set of points) might seem more intuitive at first. However, Definition 5.5 (in terms of tangent vectors) captures the vector space structure of the tangent plane. This will be important in some discussions later.

5.1.4. *Regular Parametrisations.* Before we continue our discussion on surfaces and how they are formally defined, let us first pause and take note of how the things we have constructed thus far can go wrong. Recall that the purpose of defining parametric surfaces was as a step toward defining surfaces, that is, “2-dimensional geometric objects”. However, there are several ways in which a parametric surface, as defined in Definition 5.3, intuitively fails to be 2-dimensional in nature.

Let us begin with the worst:

Example 5.16. *Consider the parametric surface*

$$\sigma : \mathbb{R}^2 \to \mathbb{R}^3, \qquad \sigma(u, v) = (1, 2, 0).$$

Here, the image of σ *is only a single point,* $(1, 2, 0)$*. Although* σ *has a* 2*-dimensional description by being a function of two variables, the resulting object (one point) parametrised by* σ *is* 0*-dimensional.*

The degenerate nature of σ *can also be captured through its tangent planes. Indeed, since*

$$\partial_u \sigma(u, v) = (0, 0, 0), \qquad \partial_v \sigma(u, v) = (0, 0, 0),$$

then the tangent plane satisfies the following:

- *In terms of tangent vectors,* $T_\sigma(u, v)$ *contains only the zero vector,* $\mathbf{0}|_{(1,2,0)}$*.*

- *As a set of points, $\mathcal{T}_\sigma(u,v)$ contains only a single point, $(1,2,0)$.*

However, the situation needs not be so bad for a parametric surface to fail at being 2-dimensional. The next example considers one which degenerates in only one dimension.

Example 5.17. *Consider next the parametric surface*

$$\sigma : \mathbb{R}^2 \to \mathbb{R}^3, \qquad \sigma(u,v) = (\cos u, \sin u, 1).$$

Here, the image of σ is merely a (1-dimensional) circle (see the right plot in Figure 5.11).

At the derivative level, since

$$\partial_u\sigma(u,v) = (-\sin u, \cos u, 0), \qquad \partial_v\sigma(u,v) = (0,0,0),$$

the tangent plane is also merely a line:

$$T_\sigma(u,v) = \{a \cdot (-\sin u, \cos u, 0)|_{(\cos u, \sin u, 1)} \mid a \in \mathbb{R}\},$$
$$\mathcal{T}_\sigma(u,v) = \{(\cos u, \sin u, 1) + a \cdot (-\sin u, \cos u, 0) \mid a \in \mathbb{R}\}.$$

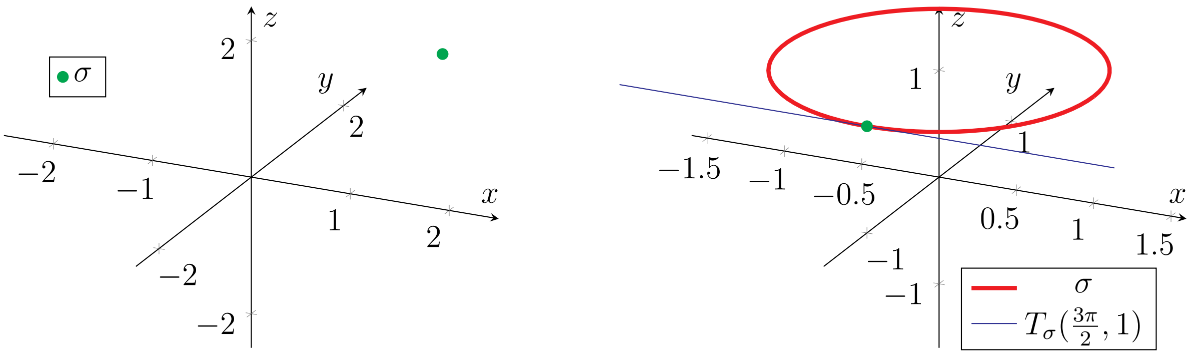


FIGURE 5.11. The left diagram contains the parametric surface σ (in green) from Example 5.16. The right diagram contains the parametric surface σ (in red) from Example 5.16; the point $\sigma(\frac{3\pi}{2},1)$ is indicated in green, while the tangent "plane" $T_\sigma(\frac{3\pi}{2},1)$ is drawn in blue.

In both Examples 5.16 and 5.17, the parametric surfaces were exceedingly pathological, in that these σ depended only on zero and one out of the two variables, respectively. However, even if we were to exclude these exceptional cases, we can still find more benign situations where a parametrisation is undesirable. To demonstrate, we consider the following example:

Example 5.18. *Consider the parametric surface*

$$\sigma : \mathbb{R} \times \left(-\frac{\pi}{2}, \frac{\pi}{2}\right) \to \mathbb{R}^2, \qquad \sigma(u,v) = (\cos u \sin v, \sin u \sin v, \cos v).$$

Note that the partial derivatives of σ are given by

$$\partial_u\sigma(u,v) = (-\sin u \sin v, \cos u \sin v, 0), \qquad \partial_v\sigma(u,v) = (\cos u \cos v, \sin u \cos v, -\sin v).$$

We claim that this σ parametrises exactly the same upper hemisphere as in previous examples (see, for instance, Example 5.9). To see this, observe that the above formulas that define σ represent standard spherical coordinates. More specifically, observe that:

- *u represents the polar angle in the xy-plane.*
- *v represents the angle from the positive z-axis (i.e. the azimuthal angle).*

Thus, in contrast to Examples 5.16 and 5.17, this σ represents an honest 2-dimensional object.

Note that the parameter $(u, v) = (0, 0)$ corresponds to the north pole:

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

Furthermore, evaluating the partial derivatives of σ at $(0, 0)$, we have

$$\partial_u\sigma(0, 0) = (0, 0, 0), \qquad \partial_v\sigma(0, 0) = (1, 0, 0).$$

Thus, the corresponding tangent plane of σ at the north pole is

$$T_\sigma(0, 0) = \{b \cdot (1, 0, 0)|_{(0,0,1)} \mid b \in \mathbb{R}\},$$

which represents a 1-dimensional line, not a 2-dimensional plane.

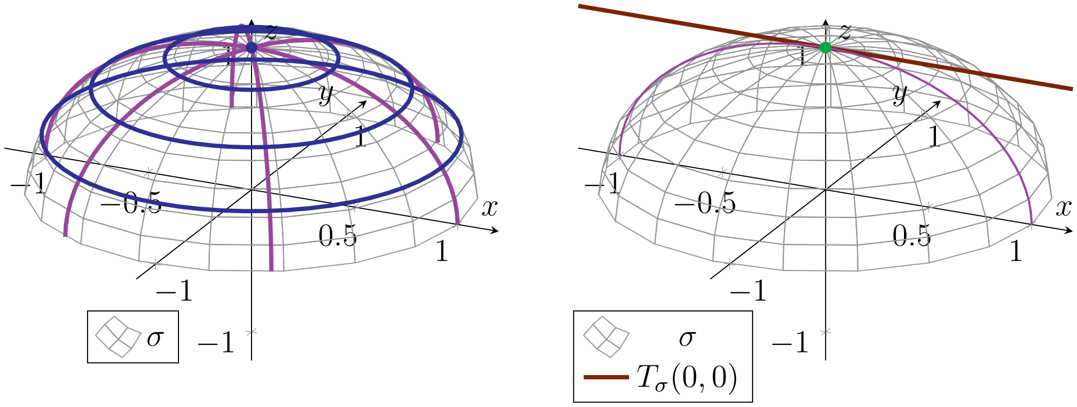


Figure 5.12. Both plots contain the upper hemisphere σ from Example 5.18. In the left plot, the blue curves are generated by holding v constant and varying u, while the purple curves are generated by holding u constant and varying v. In the right plot, the point $\sigma(0, 0) = (0, 0, 1)$ is indicated in green, and $T_\sigma(0, 0)$ is depicted in brown.

Although the image of σ from Example 5.18 is 2-dimensional, σ fails to generate a reasonable tangent plane at the north pole. To see what went wrong, let us fix $v = 0$ and vary u:

$$\lambda(u) = \sigma(u, 0) = (0, 0, 1).$$

In other words, when $v = 0$, varying u does not change the value of σ. This results in the loss of the "u-dimension" from the tangent plane $T_\sigma(0, 0)$.

One way to interpret the situation in Example 5.18 is that this σ is an undesirable way to parametrise the upper hemisphere. The tangent plane $T_\sigma(0, 0)$, representing the directions one can go along this hemisphere while at the north pole, is a fundamental geometric feature. Thus, σ is deficient in that it fails to adequately capture this geometric feature at the north pole.

Consequently, just as we excluded undesirable parametric curves through the notion of regular parametrisations (see Definition 2.4), we also wish to accomplish an analogous task for parametric surfaces. Thinking from the other direction, we can ask: what does it take for a parametric surface to “not be deficient”? Since the aforementioned deficiency arose from the loss of a dimension in the tangent plane, it seems reasonable to associate “not deficient” with the tangent plane spanning two full dimensions. This leads us to make the following definition:

Definition 5.7. *A parametric surface $\sigma : U \to \mathbb{R}^n$ is* <u>*regular*</u> *iff for any $(u, v) \in U$, the vectors $\partial_u \sigma(u, v)$ and $\partial_v \sigma(u, v)$ are linearly independent, that is, they point in different directions.*

To bring the discussion back to tangent planes, we exploit the connection between Definition 5.7 and linear algebra. In particular, σ is regular if and only if $\partial_u \sigma(u, v)$ and $\partial_v \sigma(u, v)$ span a 2-dimensional vector space. Since the tangent plane $T_\sigma(u, v)$ is simply the linear span of the tangent vectors $\partial_u \sigma(u, v)|_{\sigma(u,v)}$ and $\partial_v \sigma(u, v)|_{\sigma(u,v)}$, then $T_\sigma(u, v)$ is also a 2-dimensional vector space.

From the above argument, we now conclude the following:

Theorem 5.2. *A parametric surface $\sigma : U \to \mathbb{R}^n$ is regular if and only if for any (u, v), the tangent plane $T_\sigma(u, v)$ is a 2-dimensional vector space. Furthermore, when σ is regular, then $\partial_u \sigma(u, v)|_{\sigma(u,v)}$ and $\partial_v \sigma(u, v)|_{\sigma(u,v)}$ form a basis for the vector space $T_\sigma(u, v)$ (for any $(u, v) \in U$).*

Example 5.19. *Consider our original parametrisation of the upper hemisphere,*

$$\sigma : B_0 \to \mathbb{R}^3, \qquad \sigma(u, v) = (u, v, \sqrt{1 - u^2 - v^2}).$$

In Example 5.10, we already computed the partial derivatives of σ:

$$\partial_u \sigma(u, v) = \left(1, 0, \frac{1}{2} \cdot \frac{\partial_u (1 - u^2 - v^2)}{\sqrt{1 - u^2 - v^2}}\right) = \left(1, 0, -\frac{u}{\sqrt{1 - u^2 - v^2}}\right),$$
$$\partial_v \sigma(u, v) = \left(0, 1, \frac{1}{2} \cdot \frac{\partial_v (1 - u^2 - v^2)}{2\sqrt{1 - u^2 - v^2}}\right) = \left(0, 1, -\frac{v}{\sqrt{1 - u^2 - v^2}}\right).$$

Note that for any $(u, v) \in B_0$, the vector $\partial_u \sigma(u, v)$ always has a nonzero x-component and a zero y-component, while $\partial_v \sigma(u, v)$ always has a nonzero y-component and a zero x-component. Therefore, $\partial_u \sigma(u, v)$ and $\partial_v \sigma(u, v)$ always point in different directions and hence are linearly independent. By Definition 5.7, we conclude that σ is regular.

In Example 5.19, it was relatively easy to see by inspection that $\partial_u \sigma$ and $\partial_v \sigma$ are everywhere linearly independent. For other parametric surfaces, this inspection may not be so simple. Thus, we may ask whether there is a simpler, more computational method for checking whether a parametric surface is regular. It turns out that when $n = 3$, a convenient method does exist:

Theorem 5.3. *A parametric surface $\sigma : U \to \mathbb{R}^3$ in 3-dimensional space is regular if and only if*

$$|\partial_u \sigma(u, v) \times \partial_v \sigma(u, v)| \neq 0 \tag{5.7}$$

for any parameters $(u, v) \in U$.

Proof. Recall from Proposition 3.13 that

$$|\partial_u \sigma(u, v) \times \partial_v \sigma(u, v)| = |\partial_u \sigma(u, v)||\partial_v \sigma(u, v)| \sin \theta,$$

where θ is the angle made between $\partial_u \sigma(u, v)|_{\sigma(u,v)}$ and $\partial_v \sigma(u, v)|_{\sigma(u,v)}$. Thus, (5.7) holds if and only if both $\partial_u \sigma(u, v)$ and $\partial_v \sigma(u, v)$ are nonzero and $\sin \theta \neq 0$. Furthermore, $\sin \theta \neq 0$ is equivalent to $\partial_u\sigma(u,v)$ and $\partial_v\sigma(u,v)$ pointing in different directions. Combining all the above, we conclude that (5.7) is indeed equivalent to σ being regular. □

Let us conclude by applying Theorem 5.3 to the torus:

Example 5.20. *Consider the torus from Example 5.6 (see also Figure 5.4),*

$$\sigma : \mathbb{R}^2 \to \mathbb{R}^3, \qquad \sigma(u,v) = ((2+\cos u)\cos v, (2+\cos u)\sin v, \sin u).$$

Recall from Example 5.11 that

$$\begin{aligned}\partial_u\sigma(u,v) &= (-\sin u\cos v, -\sin u\sin v, \cos u),\\ \partial_v\sigma(u,v) &= (-(2+\cos u)\sin v, (2+\cos u)\cos v, 0).\end{aligned}$$

Taking a cross product of the above yields

$$\partial_u\sigma(u,v) \times \partial_v\sigma(u,v) = -(2+\cos u)(\cos u\cos v, \cos u\sin v, \sin u),$$

for which the norm is

$$\begin{aligned}|\partial_u\sigma(u,v) \times \partial_v\sigma(u,v)| &= (2+\cos u)(\cos^2 u\cos^2 v + \cos^2 u\sin^2 v + \sin^2 u)\\ &= (2+\cos u)(\cos^2 u + \sin^2 u)\\ &= 2+\cos u.\end{aligned}$$

Since this is everywhere nonzero, Theorem 5.3 implies that σ is regular.

Remark. The notion of regularity for parametric curves can also be characterised in terms of the tangent line. In particular, a parametric curve γ is regular if and only if all of its tangent lines $T_\gamma(t)$ are one-dimensional vector spaces, that is, they are actual lines.

5.2. **What is a Surface?** In the previous section, we discussed how 2-dimensional geometric objects can be constructed using parametric surfaces. In this section, we use what we have learned to finally give an answer to Question 5.1—*what exactly is a surface*?

5.2.1. *Multiple Parametrisations.* We already discussed the importance of parametric surfaces being regular (see Definition 5.7), as this ensures that they remain "fully 2-dimensional". However, there are other ways that parametric surfaces can fall short of our intuitive requirements.

Example 5.21. *Consider the parametric surface given by*

$$\sigma : \mathbb{R} \times (-1,1) \to \mathbb{R}^3, \qquad \sigma(u,v) = (u^2-1, u^3-u, v).$$

Taking partial derivatives of σ, we see that

$$\partial_u\sigma(u,v) = (2u, 3u^2-1, 0), \qquad \partial_v\sigma(u,v) = (0,0,1).$$

In particular,

$$\partial_u\sigma(u,v) \times \partial_v\sigma(u,v) = (3u^2-1, -2u, 0),$$

which can never vanish entirely for any $u \in \mathbb{R}$. Thus, Theorem 5.3 shows that σ is regular.

Now, a plot of σ can be found in Figure 5.13. Observe that for any $-1 < v < 1$,

$$\sigma(-1, v) = (0, 0, v) = \sigma(1, v).$$

In other words, σ passes through itself at the line segment

$$\mathcal{L} = \{(0, 0, v) \mid -1 < v < 1\}.$$

That the image of σ in Example 5.21 intersects itself clashes with our intuitions on what a surface should be. As a result, here *we choose to rule out this particular scenario.* In the context of a parametric surface σ, we must make sure that every point in its image is mapped to only once. In other words, we must *assume that σ is injective*, or one-to-one.

Remark. This should be contrasted with our study of curves in previous chapters, where self-intersections were allowed. That we allow curves to self-intersect and forbid surfaces to do the same (at least in this module) is purely for cultural reasons, to align with our basic intuitions for these words.

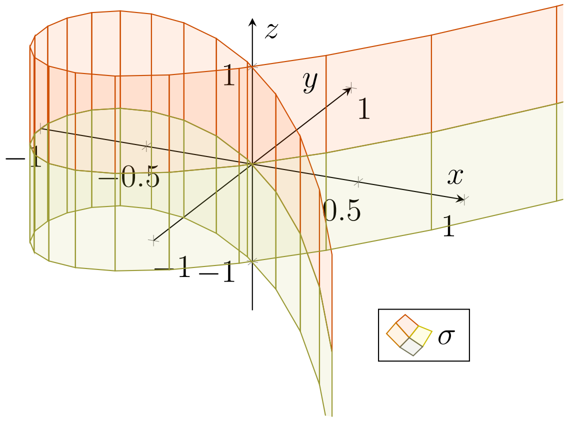


FIGURE 5.13. The parametric surface σ from Example 5.21. In particular, σ passes through itself when $x = y = 0$.

However, by restricting ourselves to injective parametric surfaces, we create new problems:

Example 5.22. *Consider the cylinder $x^2 + y^2 = 1$ from Example 5.5, represented by*

$$\sigma : \mathbb{R}^2 \to \mathbb{R}^3, \qquad \sigma(u, v) = (\cos u, \sin u, v).$$

A plot of σ can be in Figure 5.4. The partial derivatives of σ are given by

$$\partial_u \sigma(u, v) = (-\sin u, \cos u, 0), \qquad \partial_v \sigma(u, v) = (0, 0, 1).$$

In particular, σ is regular by Theorem 5.3, since for any $(u, v) \in \mathbb{R}^2$,

$$|\partial_u \sigma(u, v) \times \partial_v \sigma(u, v)| = |(\cos u, \sin u, 0)| = 1 > 0.$$

Note, however, that σ fails to be injective—for example,

$$\sigma(0, v) = (1, 0, v) = \sigma(2\pi, v), \qquad v \in \mathbb{R}.$$

Indeed, σ repeats itself once its u-parameter has increased or decreased by 2π. Thus, for injectivity, we must restrict its u-values to an (open) interval of length at most 2π; for instance, we can take

$$\sigma_1 : (0, 2\pi) \times \mathbb{R} \to \mathbb{R}^3, \qquad \sigma_1(u, v) = (\cos u, \sin u, v).$$

Unfortunately, σ_1 creates a new shortcoming, because now the points

$$(1, 0, v) = \sigma(0, v), \qquad v \in \mathbb{R}$$

are missing from the image of σ_1. *In other words,* σ_1 *only represents the cylinder with a vertical line removed. This problem persists no matter how we restrict the* u*-values of* σ*, as long as we demand that our parametric surfaces are injective.*

In fact, *it is impossible to cover all of the cylinder* $x^2+y^2=1$ *using only one injective parametric surface*. To overcome this obstacle, we *describe this cylinder using multiple parametric surfaces*:

Example 5.23. *Consider now the following pair of parametric surfaces:*

$$\sigma_1 : (0, 2\pi) \times \mathbb{R} \to \mathbb{R}^3, \qquad \sigma_1(u, v) = (\cos u, \sin u, v),$$
$$\sigma_2 : (-\pi, \pi) \times \mathbb{R} \to \mathbb{R}^3, \qquad \sigma_2(u, v) = (\cos u, \sin u, v).$$

Note both σ_1 *and* σ_2 *are regular and injective parametric surfaces. Also, both* σ_1 *and* σ_2 *cover all of the cylinder* $x^2 + y^2 = 1$ *except for a single vertical line; see Figure 5.14 below.*

The main point, though, is that σ_1 *and* σ_2 *exclude different vertical lines. Thus, the images of* σ_1 *and* σ_2 *together cover precisely the full cylinder* $x^2 + y^2 = 1$.

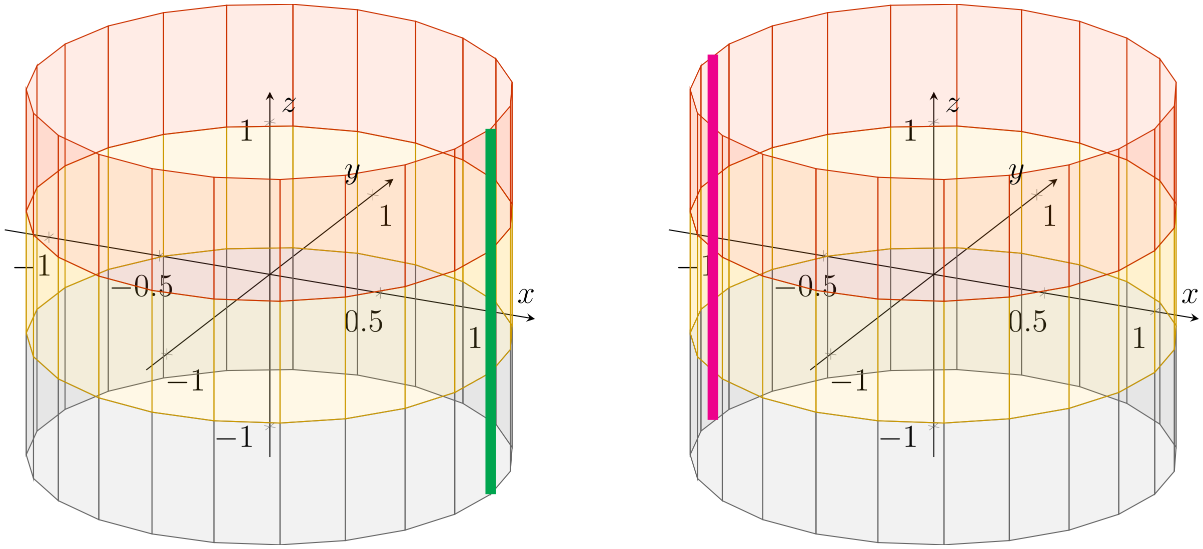


FIGURE 5.14. The left plot shows the image of σ_1, which covers all of the cylinder $x^2+y^2=1$ except for the green line segment. The right plot shows the image of σ_2, covering all of the cylinder except for the magenta segment.

Remark. In Example 5.22, one may instead consider letting u lie in the interval $[0, 2\pi)$. However, by doing so, the corresponding domain, $[0, 2\pi) \times \mathbb{R}$, would then fail to be an open subset, which creates even bigger problems in the long run. In particular, allowing non-open domains in general would allow for many undesirable objects that violate our intuitions regarding surfaces.

Recall that although the cylinder $x^2+y^2=1$ cannot be covered by a single injective parametric surface, the σ from Example 5.22 that does cover this cylinder is still perfectly regular. Next, we consider an example of a situation that is considerably worse.

Definition 5.8. *Let $\mathbb{S}^2$ denote the* <u>*unit sphere about the origin*</u> *in $\mathbb{R}^3$:*

$$\mathbb{S}^2 = \{(x, y, z) \in \mathbb{R}^3 \mid x^2 + y^2 + z^2 = 1\}.$$

A plot of the sphere is given in the left diagram in Figure 5.15.

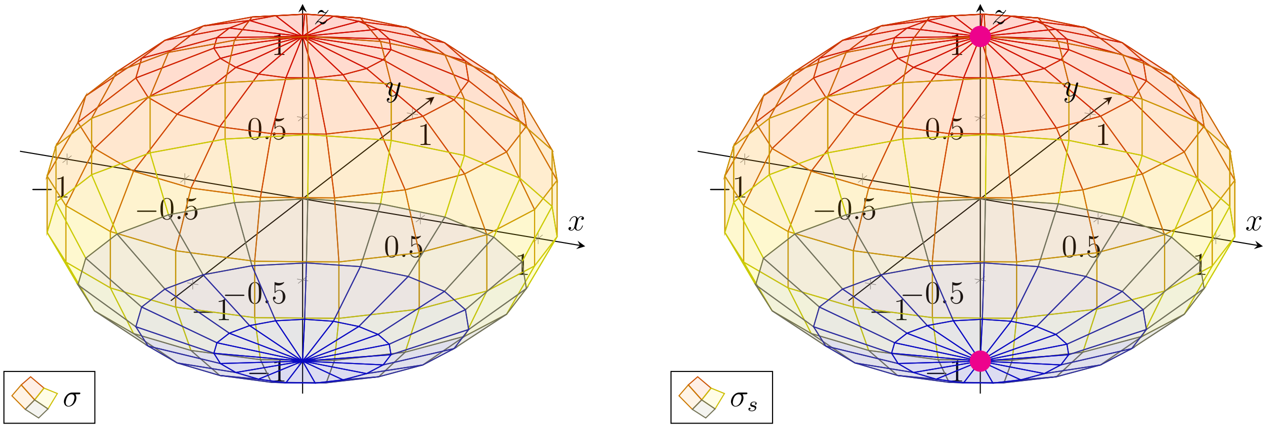


Figure 5.15. The left plot shows the unit sphere $x^2 + y^2 + z^2 = 1$, i.e. the image of σ from Example 5.24. The right plot shows the image of σ_s from Example 5.24: the unit sphere except for the poles $(0, 0, \pm 1)$ (in magenta).

Example 5.24. *We can cover all of $\mathbb{S}^2$ using the* <u>*spherial coordinates*</u> *from Example 5.18:*

$$\sigma : \mathbb{R}^2 \to \mathbb{R}^3, \qquad \sigma(u, v) = (\cos u \sin v, \sin u \sin v, \cos v).$$

(That this σ covers $\mathbb{S}^2$ follows from the usual study of spherical coordinates from multivariable calculus.) Computing the partial derivatives of σ yields

$$\begin{aligned} \partial_u \sigma(u, v) &= (-\sin u \sin v, \cos u \sin v, 0), \\ \partial_v \sigma(u, v) &= (\cos u \cos v, \sin u \cos v, -\sin v). \end{aligned}$$

Taking their cross product, we obtain that

$$\begin{aligned} \partial_u \sigma(u, v) \times \partial_v \sigma(u, v) &= (-\cos u \sin^2 v, -\sin u \sin^2 v, -\sin v \cos v) \\ &= -\sin v (\cos u \sin v, \sin u \sin v, \cos v), \\ |\partial_u \sigma(u, v) \times \partial_v \sigma(u, v)| &= |\sin v| \cdot |(\cos u \sin v, \sin u \sin v, \cos v)| \\ &= |\sin v| \sqrt{\cos^2 u \sin^2 v + \sin^2 u \sin^2 v + \cos^2 v} \\ &= |\sin v|. \end{aligned}$$

In particular, $|\partial_u \sigma(u, v) \times \partial_v \sigma(u, v)|$ vanishes whenever $\sin v = 0$—for instance, at $v = 0$ and $v = \pi$. Therefore, by Theorem 5.3, we see that σ fails to be regular.

For regularity, we must restrict σ away from where $\sin v$ vanishes. For example, we can restrict its v-values to the interval $(0, \pi)$. In other words, the following parametric surface is regular:

$$\sigma_s : \mathbb{R} \times (0, \pi) \to \mathbb{R}^3, \qquad \sigma_s(u, v) = (\cos u \sin v, \sin u \sin v, \cos v).$$

Observe that σ_s covers all of $\mathbb{S}^2$ except for the north pole $(0, 0, 1)$ (which corresponds to $v = 0$) and the south pole $(0, 0, -1)$ (which corresponds to $v = \pi$).

More generally, one can actually show that *it is impossible to find a single regular parametric surface, injective or otherwise, that covers all of $\mathbb{S}^2$*. This shows an even more pressing need to use multiple parametric surfaces when devising the formal definition of surfaces.

Below, we show one way to cover $\mathbb{S}^2$ with both regular and injective parametric surfaces.

Example 5.25. *Recall from Example 5.7 that the parametric surface*

$$\sigma_{z,+} : B_0 \rightarrow \mathbb{R}^3, \qquad \sigma_{z,+}(u, v) = (u, v, \sqrt{1 - u^2 - v^2}),$$

where B_0 is the open unit disk about the origin in $\mathbb{R}^2$, covers the upper hemisphere of $\mathbb{S}^2$. If we negate the z-component, that is, we define the parametric surface

$$\sigma_{z,-} : B_0 \rightarrow \mathbb{R}^3, \qquad \sigma_{z,+}(u, v) = (u, v, -\sqrt{1 - u^2 - v^2}),$$

then this $\sigma_{z,-}$ covers instead the lower hemisphere of $\mathbb{S}^2$. Observe that $\sigma_{z,+}$ and $\sigma_{z,-}$ combined cover all of $\mathbb{S}^2$ except for the equator—the points of $\mathbb{S}^2$ satisfying $z = 0$.

To cover the missing equator, we can swap coordinates and consider

$$\sigma_{y,+} : B_0 \rightarrow \mathbb{R}^3, \qquad \sigma_{y,+}(u, v) = (u, \sqrt{1 - u^2 - v^2}, v),$$
$$\sigma_{y,-} : B_0 \rightarrow \mathbb{R}^3, \qquad \sigma_{y,+}(u, v) = (u, -\sqrt{1 - u^2 - v^2}, v),$$

which cover the hemispheres given by $y > 0$ and $y < 0$. Then, the four parametric surfaces $\sigma_{z,\pm}$, $\sigma_{y,\pm}$ combined cover all the points of $\mathbb{S}^2$ except for the poles $(\pm 1, 0, 0)$ (corresponding to $y = z = 0$).

To take care of the last two points, we consider the hemispheres with respect to x:

$$\sigma_{x,+} : B_0 \rightarrow \mathbb{R}^3, \qquad \sigma_{x,+}(u, v) = (\sqrt{1 - u^2 - v^2}, u, v),$$
$$\sigma_{x,-} : B_0 \rightarrow \mathbb{R}^3, \qquad \sigma_{x,+}(u, v) = (-\sqrt{1 - u^2 - v^2}, u, v),$$

Then, the six parametric surfaces $\sigma_{z,\pm}$, $\sigma_{y,\pm}$, $\sigma_{x,\pm}$ together cover all of $\mathbb{S}^2$; see Figure 5.16 below.

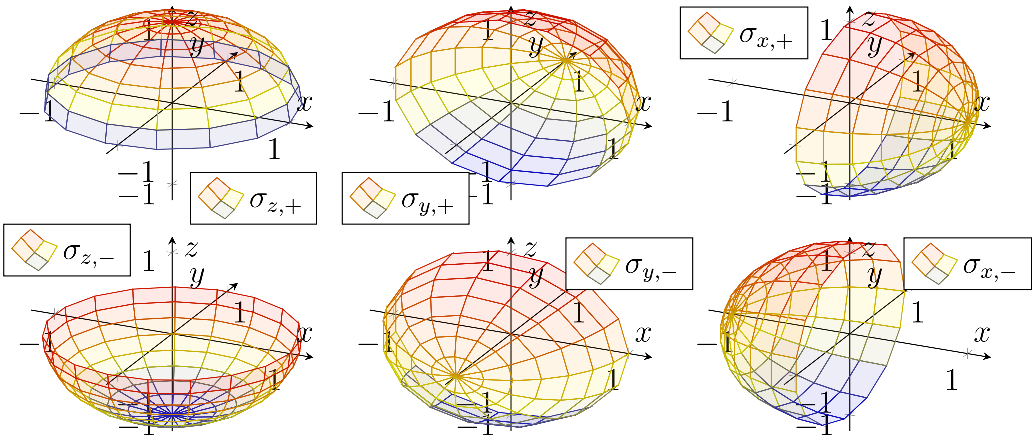


Figure 5.16. The six injective and regular parametric surfaces from Example 5.25: $\sigma_{z,\pm}$ (left), $\sigma_{y,\pm}$ (middle), and $\sigma_{x,\pm}$ (right).

A similar "patching" of $\mathbb{S}^2$ can be made using spherical coordinates:

Example 5.26. *Let us return to spherical coordinates on* $\mathbb{S}^2$*—namely, the parametric surface* σ_s *from Example 5.24. Like in the case of the cylinder (see Example 5.23), in order to make* σ_s *injective, we must restrict the polar angle* u *to a interval of length at most* 2π*.*

More specifically, if we define

$$\sigma_{s,z} : (0, 2\pi) \times (0, \pi) \rightarrow \mathbb{R}^3, \qquad \sigma_{s,z}(u, v) = (\cos u \sin v, \sin u \sin v, \cos v),$$

then this $\sigma_{s,z}$ *covers all of* $\mathbb{S}^2$ *except for a single half circle (see the left drawing in Figure 5.17). To cover this missing half circle, we must supplement* $\sigma_{s,z}$ *with other parametric surfaces. For example, this can be accomplished by swapping coordinates:*

$$\sigma_{s,y} : (0, 2\pi) \times (0, \pi) \rightarrow \mathbb{R}^3, \qquad \sigma_{s,y}(u, v) = (\sin u \sin v, \cos v, \cos u \sin v),$$
$$\sigma_{s,x} : (0, 2\pi) \times (0, \pi) \rightarrow \mathbb{R}^3, \qquad \sigma_{s,x}(u, v) = (\cos v, \cos u \sin v, \sin u \sin v).$$

Although each of $\sigma_{s,y}$ *and* $\sigma_{s,x}$ *also misses a half circle (see the middle and right diagrams in Figure 5.17), the three parametrisations* $\sigma_{s,z}$*,* $\sigma_{s,y}$*, and* $\sigma_{s,x}$ *together cover all of* $\mathbb{S}^2$*.*

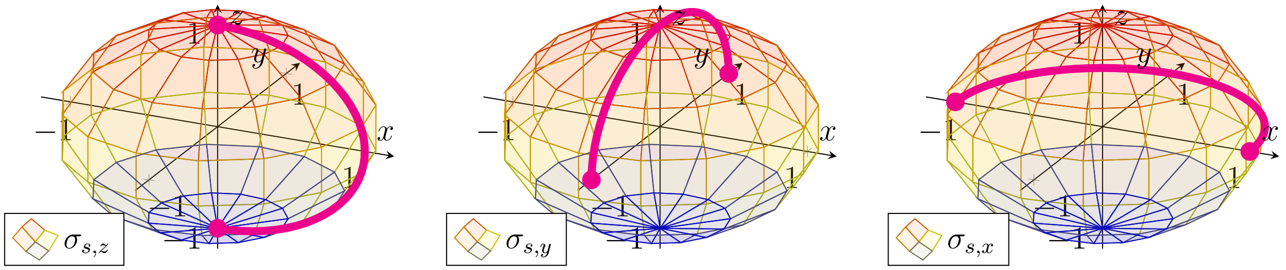


FIGURE 5.17. The parametric surfaces $\sigma_{s,z}$ (left), $\sigma_{s,y}$ (right), $\sigma_{s,x}$ (right) from Example 5.26. Each of these parametrisations covers all of $\mathbb{S}^2$ except for a single half circle, drawn in magenta in each diagram.

5.2.2. *A Precise Definition.* From the preceding discussions, we come across the conclusion that general 2-dimensional objects should be constructed using one or more regular and injective parametric surfaces. With this idea in mind, we now concoct a precise definition of surfaces.

We begin by first generalising the previous definition of open sets (see Definition 5.1):

Definition 5.9. *A subset* $V \subseteq \mathbb{R}^n$ *is* <u>*open*</u> *iff for any* $\mathbf{p} \in V$*, there is some open "hypercube"*

$$\mathcal{R} = (a_1, b_1) \times (a_2, b_2) \times \cdots \times (a_n, b_n)$$

such that $\mathbf{p} \in \mathcal{R}$ *and* $\mathcal{R} \subseteq V$*.*

The intuition for open subsets of $\mathbb{R}^n$ is exactly the same as that in Definition 5.1 for open subsets of $\mathbb{R}^2$. If one were to parachute onto a point $\mathbf{q}$ of an open set $V \subseteq \mathbb{R}^n$, then one can take a small enough step in any direction from $\mathbf{q}$ and still remain in this set V.

In practice, we will use open sets in the following way: given a point $\mathbf{p} \in \mathbb{R}^n$, an open subset $V \subseteq \mathbb{R}^n$ containing $\mathbf{p}$ can be considered as the set of points that are "sufficiently near" $\mathbf{p}$.

Definition 5.10. *A subset $S \subseteq \mathbb{R}^n$ is called a* <u>*surface*</u> *iff for any point $\mathbf{p} \in S$, there is*

1. *An open subset $V \subseteq \mathbb{R}^n$ such that $\mathbf{p} \in V$, and*
2. *A regular and injective parametric surface $\sigma : U \to S$,*

such that σ is a bijection between U and $S \cap V$.

In short, Definition 5.10 states that every point $\mathbf{p}$ of S is covered by some regular and injective parametric surface σ whose image lies in S. In other words, S is constructed by "patching together" various (regular and injective) parametric surfaces. This is precisely the idea that was demonstrated in Examples 5.23 and 5.25 for the cylinder and sphere, respectively.

On the other hand, what is new and mysterious in Definition 5.10 is the open subset V and the role it plays. Basically, this has to do with ensuring that S does not "go through itself", such as in Example 5.21. We defer more detailed demonstrations of this point until further below.

We now demonstrate Definition 5.10 through some concrete examples:

Example 5.27. *Consider the cylinder*

$$\mathcal{C} = \{(x, y, z) \in \mathbb{R}^3 \mid x^2 + y^2 = 1\}$$

that was studied in Example 5.22. We now formally show that $\mathcal{C}$ is a surface.

Consider the subsets

$$V_1 = \{(x, y, z) \in \mathbb{R}^3 \mid x < 1\}, \qquad V_2 = \{(x, y, z) \in \mathbb{R}^3 \mid x > -1\}.$$

It is not too hard to see that both V_1 and V_2 are open subsets of $\mathbb{R}^3$. Furthermore, observe:

- *$\mathcal{C} \cap V_1$ is precisely $\mathcal{C}$ except for the vertical line $\mathcal{L}_1 = \{(1, 0, v) \mid v \in \mathbb{R}\}$.*
- *$\mathcal{C} \cap V_2$ is precisely $\mathcal{C}$ except for the vertical line $\mathcal{L}_2 = \{(-1, 0, v) \mid v \in \mathbb{R}\}$.*

Recall also the regular, injective parametric surfaces σ_1 and σ_2 from Example 5.23. In particular,

- *σ_1 is a bijection between $U_1 = (0, 2\pi) \times \mathbb{R}$ and $\mathcal{C} \cap V_1$.*
- *σ_2 is a bijection between $U_2 = (-\pi, \pi) \times \mathbb{R}$ and $\mathcal{C} \cap V_2$.*

With the above setup in place, consider now any point $\mathbf{p} = (x_0, y_0, z_0) \in \mathcal{C}$.

- *If $\mathbf{p} \notin \mathcal{L}_1$ (i.e. $x_0 \neq 1$), then V_1 is an open subset of $\mathbb{R}^3$ containing $\mathbf{p}$, and σ_1 is a regular and injective parametric surface that is also a bijection between U_1 and $\mathcal{C} \cap V_1$.*
- *Otherwise, $\mathbf{p} \notin \mathcal{L}_2$ (i.e. $x_0 \neq -1$), hence V_2 is an open subset of $\mathbb{R}^3$ containing $\mathbf{p}$. Also, σ_2 is a regular and injective parametric surface that is a bijection between U_2 and $\mathcal{C} \cap V_2$.*

Thus, we see that the conditions of Definition 5.10 are always satisfied.

In essence, Example 5.27 states, in considerable detail, that $\mathcal{C}$ is constructed by patching the two regular and injective parametric surfaces σ_1 and σ_2 together.

Example 5.28. *Next, consider the unit sphere $\mathbb{S}^2$ from Definition 5.8. Define the subsets*

$$V_{z,+} = \{(x, y, z) \in \mathbb{R}^3 \mid z > 0\}, \qquad V_{z,-} = \{(x, y, z) \in \mathbb{R}^3 \mid z < 0\},$$
$$V_{y,+} = \{(x, y, z) \in \mathbb{R}^3 \mid y > 0\}, \qquad V_{y,-} = \{(x, y, z) \in \mathbb{R}^3 \mid y < 0\},$$
$$V_{x,+} = \{(x, y, z) \in \mathbb{R}^3 \mid x > 0\}, \qquad V_{x,-} = \{(x, y, z) \in \mathbb{R}^3 \mid x < 0\},$$

which one can show to all be open. Furthermore, recalling the (regular and injective) parametric surfaces from Example 5.25, we observe that $\sigma_{z,\pm}$ is a bijection between B_0 and $\mathbb{S}^2 \cap V_{z,\pm}$. Analogous relations hold between $\sigma_{y,\pm}$, $\sigma_{x,\pm}$ and $\mathbb{S}^2 \cap V_{y,\pm}$, $\mathbb{S}^2 \cap V_{x,\pm}$, respectively.

Consider now a point $\mathbf{p} = (x_0, y_0, z_0) \in \mathbb{S}^2$. Since $x_0^2 + y_0^2 + z_0^2 = 1$, then one of the coordinates x_0, y_0, z_0 must be nonzero, and thus $\mathbf{p}$ lies in one of the sets $V_{z,\pm}$, $V_{y,\pm}$, $V_{x,\pm}$. Thus, for this $\mathbf{p}$, one of the parametric surfaces $\sigma_{z,\pm}$, $\sigma_{y,\pm}$, $\sigma_{x,\pm}$ will satisfy the conditions dictated in Definition 5.10.

Consequently, $\mathbb{S}^2$ is indeed a surface. Furthermore, the above shows that $\mathbb{S}^2$ can be constructed by patching the parametric surfaces $\sigma_{z,\pm}$, $\sigma_{y,\pm}$, and $\sigma_{z,\pm}$ together.

It is instructive to compare and contrast Definition 5.10 for surfaces with the corresponding Definition 2.6 for curves. The main similarity is that both curves and surfaces are built using parametrisations—parametric curves and surfaces. In particular, these parametrisations are the concrete objects on which we can do calculus. As such, we use them to get at the geometric properties of curves and surfaces that are of interest to us.

One simple, but striking, difference is that curves required only one parametrisation to define, while a surface many require many parametrisations. Indeed, for a surface S, a single parametric surface may only describe a *local* region of S. To describe the *global* geometry of S, one would need to patch multiple parametric surfaces together. This is one demonstration of how moving to higher dimensions often leads to significantly more complex situations.

Finally, we turn our attention toward the open set V in Definition 5.10, which had no analogue in the definition of curves. As mentioned before, this is in place to prevent S from having self-intersections. Rather than provide a general discussion (which extends a bit beyond the scope of this module), we informally demonstrate this through the following example:

Example 5.29. *Let S denote the image of the regular (but not injective) parametric surface,*

$$\sigma : \mathbb{R} \times (-1, 1) \to \mathbb{R}^3, \qquad \sigma(u, v) = (u^2 - 1, u^3 - u, v),$$

from Example 5.21. Consider, for instance, the point $\mathbf{p} = (0, 0, 0) \in S$. In particular, $\mathbf{p}$ lies exactly on the self-intersection of two different "strips" of S, since

$$\mathbf{p} = \sigma(1, 0) = \sigma(-1, 0).$$

This is drawn in the left diagram in Figure 5.18.

Consider now any open subset $V \subseteq \mathbb{R}^n$ that contains $\mathbf{p}$. Since V is open, V must extend beyond $\mathbf{p}$ in every direction, i.e. V "has width in every dimension". In particular, V must contain parts of both strips of S that cross each other; see the right plot in Figure 5.18.

Suppose there is a regular and injective parametric surface σ that maps to $S \cap V$, as in Definition 5.10. Then, σ maps onto a region of S containing both strips where they intersect; see the right picture in Figure 5.18. However, this is not possible because of the smoothness of σ; where the two strips intersect, one can only turn from one strip to the other in a jagged and non-smooth way.

(Even aside from smoothness, such a σ is impossible due to topological considerations. However, we avoid discussing this here, as it extends beyond the scope of this module.)

As a result, Definition 5.10 would be violated at this intersection point $\mathbf{p}$.

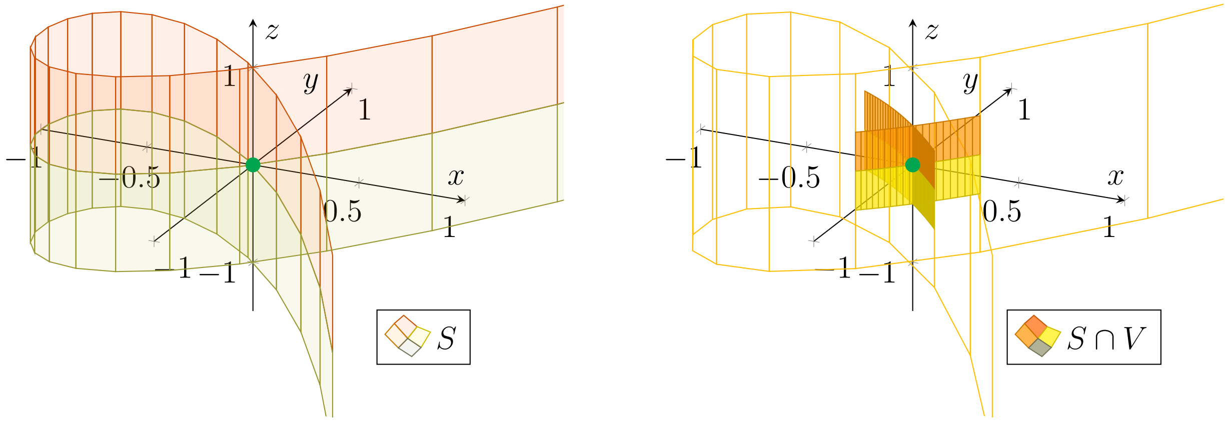


FIGURE 5.18. The left plot contains S from Examples 5.21 and 5.29, with the point $\mathbf{p} = (0, 0, 0) \in S$ indicated in green. In the right plot, S appears as a transparent mesh; the shaded region depicts an example of a typical region $S \cap V$ within S, where $V \subseteq \mathbb{R}^3$ is open.

Remark. We could also have used a 1-dimensional analogue of Definition 5.10 to define curves. Doing this would have generated a corresponding notion of "curves that do not self-intersect".

Remark. Also, one advantage of Definition 5.10 is that it can be directly generalised in order to define higher-dimensional geometric objects (called manifolds), beyond curves and surfaces. These are, unfortunately, well beyond the scope of the current module.

In practice, we will often be a bit lax and describe surfaces using regular parametric surfaces, without formally checking all the conditions of Definition 5.10. The implicit understanding is that such surfaces will have no self-intersections, and one can carefully check all the conditions of Definition 5.10 in a manner similar to Examples 5.27 and 5.28. Furthermore, we will sometimes use non-injective parametric surfaces, with the implicit understanding that these can be split into injective parametrisations, as was done in Examples 5.23 and 5.25.

In light of the above, we make the following definition to simplify future writing:

Definition 5.11. *Let $S \subseteq \mathbb{R}^n$ be a surface. We refer to any regular parametric surface $\sigma : U \to S$ mapping into S (injective or not) as a parametrisation of S.*

For completeness, we conclude the present discussion with some additional examples:

Example 5.30. *Let $f : \mathbb{R}^2 \to \mathbb{R}$ be any smooth function of two variables, and let*

$$G_f = \{(u, v, f(u, v)) \mid u, v \in \mathbb{R}\},$$

i.e. the graph of f. We claim that G_f is a surface.

Clearly, $V = \mathbb{R}^3$ is an open subset of $\mathbb{R}^3$. Then, the parametric surface

$$\sigma : \mathbb{R}^2 \to \mathbb{R}^3, \qquad \sigma(u, v) = (u, v, f(u, v))$$

is injective, and it is a bijection between $\mathbb{R}^2$ and $G_f \cap V = G_f$. Furthermore, σ is regular, since

$$\partial_u \sigma(u, v) = (1, 0, \partial_u f(u, v)),$$

$$\partial_v \sigma(u,v) = (0, 1, \partial_v f(u,v)),$$
$$\partial_u \sigma(u,v) \times \partial_v \sigma(u,v) = (-\partial_u f(u,v), -\partial_v f(u,v), 1) \neq 0.$$

As a result, for any $\mathbf{p} \in G_f$, the above σ is a parametrisation of S which covers $\mathbf{p}$. Thus, Definition 5.10 is satisfied, and G_f is indeed a surface (defined from the single parametrisation σ).

Example 5.31. *Next, recall the torus T^2 from Example 5.6, generated from the parametric surface*

$$\sigma : \mathbb{R}^2 \to \mathbb{R}^3, \qquad \sigma(u,v) = ((2 + \cos u)\cos v, (2 + \cos u)\sin v, \sin u).$$

(A plot of T^2 is found in Figure 5.4.) From the computations in Example 5.20, we had determined that σ is regular. On the other hand, σ fails to be injective, since cos *and* sin *are periodic.*

To decompose σ into injective parametrisations of T^2, we proceed as for the cylinder in Example 5.23—we avoid including one full period of cos *and* sin*. For this, we require four parametrisations:*

$$\begin{aligned} \sigma_1 &: (0, 2\pi) \times (0, 2\pi) \to \mathbb{R}^3, & \sigma_1(u,v) &= ((2 + \cos u)\cos v, (2 + \cos u)\sin v, \sin u), \\ \sigma_2 &: (0, 2\pi) \times (-\pi, \pi) \to \mathbb{R}^3, & \sigma_2(u,v) &= ((2 + \cos u)\cos v, (2 + \cos u)\sin v, \sin u), \\ \sigma_3 &: (-\pi, \pi) \times (0, 2\pi) \to \mathbb{R}^3, & \sigma_3(u,v) &= ((2 + \cos u)\cos v, (2 + \cos u)\sin v, \sin u), \\ \sigma_4 &: (-\pi, \pi) \times (-\pi, \pi) \to \mathbb{R}^3, & \sigma_4(u,v) &= ((2 + \cos u)\cos v, (2 + \cos u)\sin v, \sin u). \end{aligned}$$

Furthermore, with a bit more care, one can also observe that σ_1, σ_2, σ_3, and σ_4 are sufficient to show that T^2 is a surface, in the sense of Definition 5.10.

5.2.3. *The Ambient Dimension.* Thus far, all of our concrete examples have involved surfaces lying in 3-dimensional space. This will also be the case for nearly all of the remainder of this module. As a result, we take a bit of time here to consider the following question:

Question 5.3. *Are there surfaces that cannot be embedded in $\mathbb{R}^3$? In other words, if we only consider surfaces lying in $\mathbb{R}^3$, then will we be missing any surfaces?*

Unfortunately, the answer to Question 5.3 is *yes.* One interesting and notorious example is the Klein bottle, named after Felix Klein (German mathematician, 1849–1925).

One way to approach the Klein bottle is as follows. Consider a rectangle, as in Figure 5.19. If we glue two opposite edges of this rectangle together, as in the left diagrams in Figure 5.19, then the resulting surface is a (finite) cylinder. Furthermore, if we were to also glue together the remaining two edges of this rectangle (in other words, we glue together the ends of the cylinder), then the resulting shape is a torus. See the middle drawings of Figure 5.19 for a demonstration.

Now, rather than gluing the last two edges as before, we try *gluing these edges with opposite orientations.* In other words, with regards to Figure 5.19, we glue the top-left corner to the lower-right corner, and the top-right to the lower-left. While this can be done at the abstract level, you could not physically do this with a piece of paper. Indeed, you could not bring the edges of the paper together in this way without passing the paper through itself.

However, such a gluing would be possible *if we had an extra dimension to move about.* If the piece of paper was sitting in $\mathbb{R}^4$ instead, then we could "slide it along this extra dimension" to perform the second gluing of its edges without it passing through itself. The Klein bottle is then the geometric object one obtains as a result of this strange gluing.

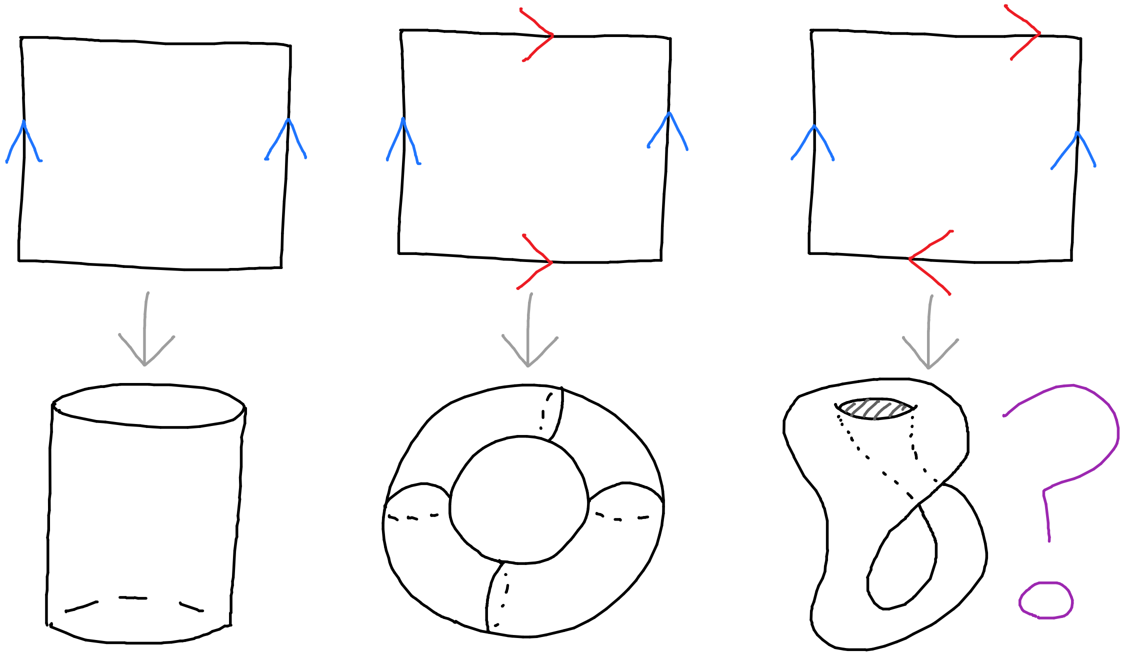


FIGURE 5.19. The left drawings show a rectangle with two edges glued together, yielding a cylinder. The middle drawings show a rectangle with both pairs of edges glued together, which results in a torus. Lastly, in the drawings on the right, the second pair of edges are glued with opposite orientations; the result is the famous Klein bottle.

Example 5.32. *There are many ways that one can describe the Klein bottle as a surface in $\mathbb{R}^4$. One description, through a (regular, non-injective) parametric surface, is given below:*

$$\sigma : \mathbb{R}^2 \to \mathbb{R}^4, \qquad \sigma(u, v) = \left((2 + \cos u) \cos v, (2 + \cos u) \sin v, \sin u \cos \frac{v}{2}, \sin u \sin \frac{v}{2} \right) .$$

With appropriate care, the image of σ can be shown to satisfy Definition 5.10.

Now, while there do exist surfaces (such as the Klein bottle) that cannot be embedded in $\mathbb{R}^3$, one can prove, in rather surprising contrast, that *every surface can be embedded in* $\mathbb{R}^4$. Thus, if we studied surfaces in $\mathbb{R}^4$ (and hence lost much of our ability to visualise them), we would have essentially studied every possible surface. This result is a special case of the famous Whitney embedding theorem, named after Hassler Whitney (American mathematician, 1907–1989).

On the other hand, while the Klein bottle could not be embedded in $\mathbb{R}^3$ as a surface, it is possible to represent it as a parametric surface in $\mathbb{R}^3$, *if we allow for self-intersections.* Of course, this would fail to be a surface in the sense of Definition 5.10.

Example 5.33. *In $\mathbb{R}^3$, one can describe the Klein bottle parametrically as*

$$\sigma : \mathbb{R}^2 \to \mathbb{R}^3, \qquad \sigma(u, v) = (x(u, v), y(u, v), z(u, v)),$$

where

$$
\begin{aligned}
x(u, v) &= \left[6 + \cos\frac{u}{2}\sin v - \sin\frac{u}{2}\sin(2v)\right]\cos u, \\
y(u, v) &= \left[6 + \cos\frac{u}{2}\sin v - \sin\frac{u}{2}\sin(2v)\right]\sin u, \\
z(u, v) &= \sin\frac{u}{2}\sin v + \cos\frac{u}{2}\sin(2v).
\end{aligned}
$$

This is graphed in Figure 5.20.

Other representations of the Klein bottle can be found on the Wikipedia page [10].

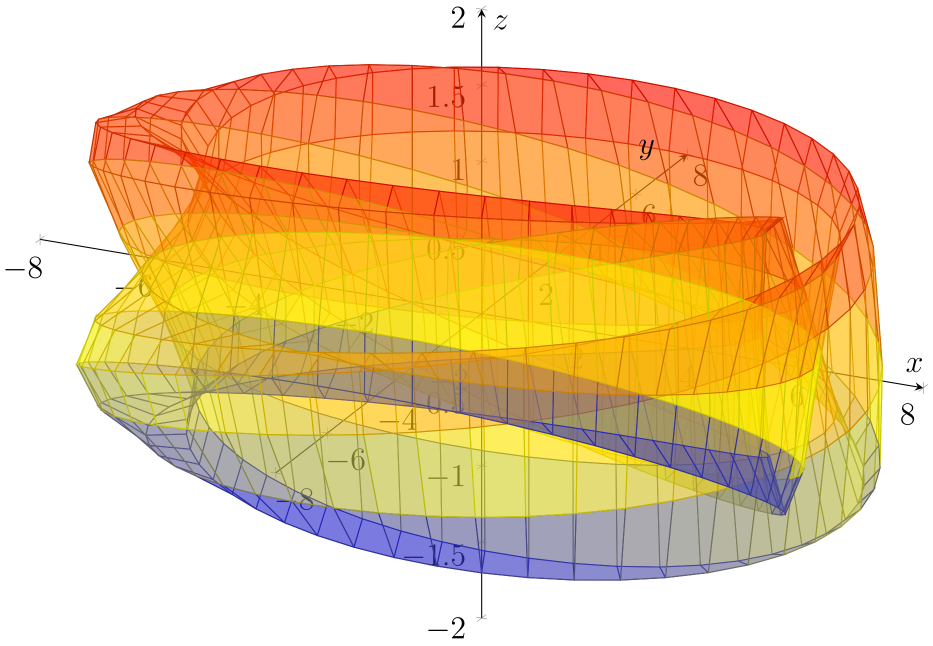


FIGURE 5.20. The parametric Klein bottle σ from Example 5.33.

Finally, we mention the following modification of the Whitney embedding theorem: *if we also allowed for self-intersections, then any surface can be placed in* $\mathbb{R}^3$. (This is a special case of the Whitney immersion theorem.) In other words, if we wished instead to study surfaces with self-intersections, like we did for curves, then we need only consider such objects as subsets of $\mathbb{R}^3$.

5.3. **The Tangent Plane.** Now that we have formally defined surfaces, our next aim is to study their geometric properties. Similar to our previous discussions for curves, we are using the term "geometric property" to refer to some attribute that is possessed by every surface, i.e. a function mapping every surface to some value. At a less formal level, geometric properties refer to those that depend only on the geometry (i.e. shape, size, and position) of a surface.

Also, like for curves, we rarely work with a surface directly; rather, we probe for information about surfaces through more concrete objects—their parametrisations. One example of this is our definition of *tangent planes* associated with a parametric surface (see Definition 5.5). It is then pertinent to ask whether this information is a geometric property of the surface, or merely a property of the parametric surface. This is precisely the same question that we faced when studying curves: *is a given property of parametric surfaces independent of parametrisation?*

To be more specific, suppose we have a surface S. Assume also that we are interested in some geometric property $\mathcal{P}$ of S, which we probe using parametrisations of S. if we were to measure $\mathcal{P}$ using two different parametrisations σ_1 and σ_2 of S, then we should obtain the same value. This is because $\mathcal{P}$ is a property of S itself, while σ_1 and σ_2 are merely local representations of S.

Below, we initiate the study of geometric properties of surfaces by discussing tangent planes. In the remainder of this section, we discuss some related properties, such as orientation.

5.3.1. *Independence of Parametrisation.* Before making any formal statements, let us first motivate our discussion through a simple example involving the sphere $\mathbb{S}^2$:

Example 5.34. *Consider the sphere $\mathbb{S}^2$ (see Definition 5.8) and the point $\mathbf{p} = (-1, 0, 0) \in \mathbb{S}^2$. As noted in Example 5.25, one parametrisation of $\mathbb{S}^2$ that covers $\mathbf{p}$ is*

$$\sigma_{x,-} : B_0 \to \mathbb{R}^3, \qquad \sigma_{x,-}(u, v) = (-\sqrt{1 - u^2 - v^2}, u, v).$$

Also, from Example 5.24, we obtain another (non-injective) parametrisation of $\mathbb{S}^2$ containing $\mathbf{p}$:

$$\sigma_s : \mathbb{R} \times (0, \pi) \to \mathbb{R}^3, \qquad \sigma_s(u, v) = (\cos u \sin v, \sin u \sin v, \cos v),$$

Let us compare the tangent planes at $\mathbf{p}$ obtained through $\sigma_{x,-}$ and σ_s.

For $\sigma_{x,-}$, we first note that $\mathbf{p} = \sigma_{x,-}(0, 0)$. Taking partial derivatives yields

$$\partial_u \sigma_{x,-}(u, v) = \left(\frac{u}{\sqrt{1 - u^2 - v^2}}, 1, 0 \right), \qquad \partial_v \sigma_{x,-}(u, v) = \left(\frac{v}{\sqrt{1 - u^2 - v^2}}, 0, 1 \right),$$
$$\partial_u \sigma_{x,-}(0, 0) = (0, 1, 0), \qquad \partial_v \sigma_{x,-}(0, 0) = (0, 0, 1).$$

Recalling Definition 5.5 for the tangent plane, we see that

$$\begin{aligned} T_{\sigma_{x,-}}(0, 0) &= \{a \cdot \partial_u \sigma_{x,-}(0, 0)|_{\mathbf{p}} + b \cdot \partial_v \sigma_{x,-}(0, 0)|_{\mathbf{p}} \mid a, b \in \mathbb{R}\} \\ &= \{a \cdot (0, 1, 0)|_{(-1,0,0)} + b \cdot (0, 0, 1)|_{(-1,0,0)} \mid a, b \in \mathbb{R}\} \\ &= \{(0, a, b)|_{(-1,0,0)} \mid a, b \in \mathbb{R}\}. \end{aligned}$$

Similarly, for σ_s, we have that $\mathbf{p} = \sigma_s(\pi, \frac{\pi}{2})$. From Example 5.24, we have that

$$\partial_u \sigma_s(u, v) = (-\sin u \sin v, \cos u \sin v, 0), \qquad \partial_v \sigma_s(u, v) = (\cos u \cos v, \sin u \cos v, -\sin v),$$
$$\partial_u \sigma_s \left(\pi, \frac{\pi}{2} \right) = (0, -1, 0), \qquad \partial_v \sigma_s \left(\pi, \frac{\pi}{2} \right) = (0, 0, -1).$$

As a result, we obtain

$$\begin{aligned} T_{\sigma_s} \left(\pi, \frac{\pi}{2} \right) &= \{a \cdot (0, -1, 0)|_{\mathbf{p}} + b \cdot (0, 0, -1)|_{\mathbf{p}} \mid a, b \in \mathbb{R}\} \\ &= \{(0, -a, -b)|_{(-1,0,0)} \mid a, b \in \mathbb{R}\}. \end{aligned}$$

Finally, comparing the above two computations, we see that

$$T_{\sigma_{x,-}}(0, 0) = T_{\sigma_s} \left(\pi, \frac{\pi}{2} \right),$$

that is, the tangent planes at $\mathbf{p}$ computed via $\sigma_{x,-}$ and σ_s are the same.

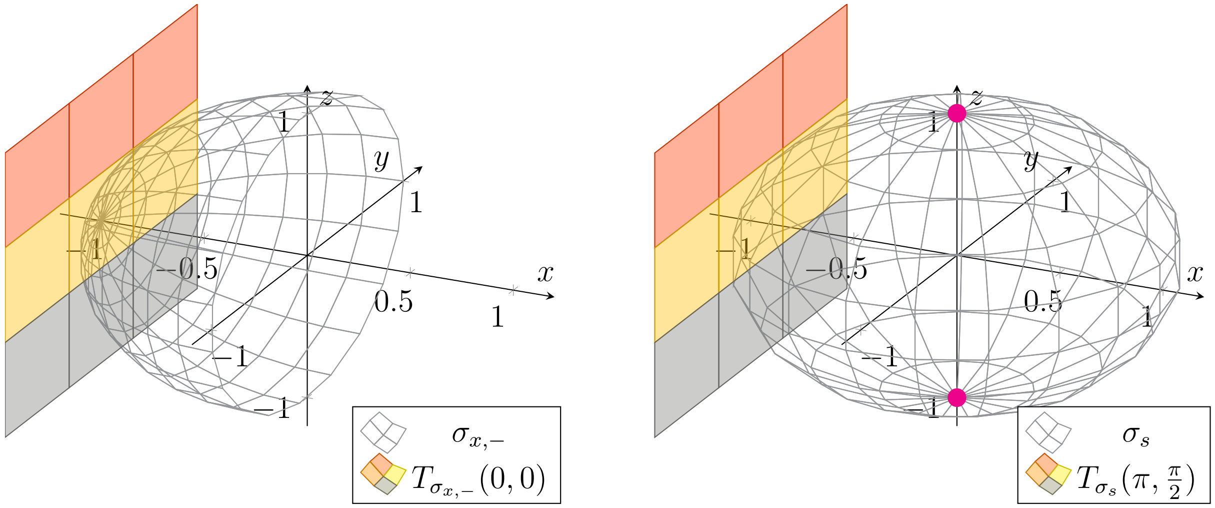


FIGURE 5.21. These figures show the two parametric tangent planes computed in Example 5.34. The left diagram depicts $\sigma_{x,-}$ and $T_{\sigma_{x,-}}(0,0)$, while the right diagram depicts σ_s and $T_{\sigma_s}(\pi, \frac{\pi}{2})$.

Example 5.34 provides some initial evidence that the tangent plane at a point $\mathbf{p} \in \mathbb{S}^2$ is in fact a geometric property of $\mathbb{S}^2$. To confirm that this is indeed the case, we will have to check that the tangent planes match for any point $\mathbf{p} \in \mathbb{S}^2$ and any pair of parametrisations of $\mathbb{S}^2$ covering $\mathbf{p}$.

In fact, we will be far more ambitious and show that we can do this for *any* surface, not just $\mathbb{S}^2$:

Theorem 5.4. *Let $S \subseteq \mathbb{R}^n$ be a surface, and let $\mathbf{p} \in S$. Also, let*

$$\sigma : U \to S, \qquad \tilde{\sigma} : \tilde{U} \to S$$

be two parametrisations of S (in the sense of Definition 5.11) such that

$$\sigma(u_0, v_0) = \tilde{\sigma}(\tilde{u}_0, \tilde{v}_0) = \mathbf{p}, \qquad (u_0, v_0) \in U, \quad (\tilde{u}_0, \tilde{v}_0) \in \tilde{U}.$$

(See Figure 5.22 for a graphical depiction of this setting.) Then,

$$T_\sigma(u_0, v_0) = T_{\tilde{\sigma}}(\tilde{u}_0, \tilde{v}_0),$$

that is, the tangent planes with respect to σ and $\tilde{\sigma}$ coincide at $\mathbf{p}$.

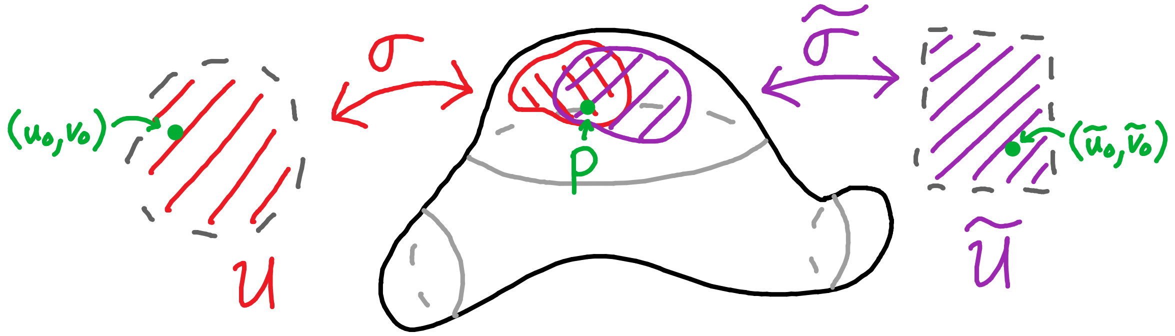


FIGURE 5.22. The setting of Theorem 5.4. This shows two parametrisations σ and $\tilde{\sigma}$ of the surface S (in black) which cover a common point $\mathbf{p} \in S$.

Proof. The idea of the proof is similar to the analogous property for tangent lines. First, we note that we can reduce the domains of σ and $\tilde{\sigma}$ in a way so that both parametrisations are injective but still map to $\mathbf{p}$. Note that at the points where the images of σ of $\tilde{\sigma}$ overlap, we can write

$$\sigma(u, v) = \tilde{\sigma}(\Phi(u, v)), \qquad \Phi(u, v) = \tilde{\sigma}^{-1}(\sigma(u, v)).$$

Here, Φ can be viewed as the *change of variables* between $(u, v) \in U$ and $(\tilde{u}, \tilde{v}) \in \tilde{U}$. (Note that $\tilde{\sigma}^{-1}$ makes sense, since $\tilde{\sigma}$ is assumed to be injective.) Furthermore, since Φ is an $\mathbb{R}^2$-valued function, we can expand Φ in terms of its components:

$$\Phi(u, v) = (\tilde{u}(u, v), \tilde{v}(u, v)).$$

Next, we use the multivariable chain rule to write $\partial_u \sigma$ in terms of partial derivatives of $\tilde{\sigma}$:

$$\begin{aligned} \partial_u \sigma(u, v) &= \partial_u[\tilde{\sigma}(\tilde{u}(u, v), \tilde{v}(u, v))] \\ &= \partial_{\tilde{u}} \tilde{\sigma}(\tilde{u}(u, v), \tilde{v}(u, v)) \cdot \partial_u \tilde{u}(u, v) + \partial_{\tilde{v}} \tilde{\sigma}(\tilde{u}(u, v), \tilde{v}(u, v)) \cdot \partial_u \tilde{v}(u, v). \end{aligned}$$

Similarly, for $\partial_v \sigma$, we have

$$\partial_v \sigma(u, v) = \partial_{\tilde{u}} \tilde{\sigma}(\tilde{u}(u, v), \tilde{v}(u, v)) \cdot \partial_v \tilde{u}(u, v) + \partial_{\tilde{v}} \tilde{\sigma}(\tilde{u}(u, v), \tilde{v}(u, v)) \cdot \partial_v \tilde{v}(u, v).$$

We now apply the above formulas at $(u, v) = (u_0, v_0)$ (corresponding to the point $\mathbf{p}$). Note that

$$\Phi(u_0, v_0) = (\tilde{u}(u_0, v_0), \tilde{v}(u_0, v_0)) = (\tilde{u}_0, \tilde{v}_0),$$

since $(\tilde{u}_0, \tilde{v}_0)$ are the $\tilde{\sigma}$-coordinates associated with $\mathbf{p}$. Thus, we obtain that

$$\begin{aligned} \partial_u \sigma(u_0, v_0) &= \partial_u \tilde{u}(u_0, v_0) \partial_{\tilde{u}} \tilde{\sigma}(\tilde{u}_0, \tilde{v}_0) + \partial_u \tilde{v}(u_0, v_0) \partial_{\tilde{v}} \tilde{\sigma}(\tilde{u}_0, \tilde{v}_0), \\ \partial_v \sigma(u_0, v_0) &= \partial_v \tilde{u}(u_0, v_0) \partial_{\tilde{u}} \tilde{\sigma}(\tilde{u}_0, \tilde{v}_0) + \partial_v \tilde{v}(u_0, v_0) \partial_{\tilde{v}} \tilde{\sigma}(\tilde{u}_0, \tilde{v}_0). \end{aligned}$$

As a result, by Definition 5.5, we see that

$$\begin{aligned} T_\sigma(u_0, v_0) &= \{a \cdot \partial_u \sigma(u_0, v_0)|_{\sigma(u_0, v_0)} + b \cdot \partial_v \sigma(u_0, v_0)|_{\sigma(u_0, v_0)} \mid a, b \in \mathbb{R}\} \\ &= \{a \cdot \partial_u \sigma(u_0, v_0)|_{\mathbf{p}} + b \cdot \partial_v \sigma(u_0, v_0)|_{\mathbf{p}} \mid a, b \in \mathbb{R}\} \\ &= \{[a \cdot \partial_u \tilde{u}(u_0, v_0) + b \cdot \partial_v \tilde{u}(u_0, v_0)] \cdot \partial_{\tilde{u}} \tilde{\sigma}(\tilde{u}_0, \tilde{v}_0)|_{\mathbf{p}} \\ &\qquad + [a \cdot \partial_u \tilde{v}(u_0, v_0) + b \cdot \partial_v \tilde{v}(u_0, v_0)] \cdot \partial_{\tilde{v}} \tilde{\sigma}(\tilde{u}_0, \tilde{v}_0)|_{\mathbf{p}} \mid a, b \in \mathbb{R}\}. \end{aligned}$$

Since the quantities

$$a \cdot \partial_u \tilde{u}(u_0, v_0) + b \cdot \partial_v \tilde{u}(u_0, v_0), \qquad a \cdot \partial_u \tilde{v}(u_0, v_0) + b \cdot \partial_v \tilde{v}(u_0, v_0)$$

are simply real numbers, it follows that

$$T_\sigma(u_0, v_0) \subseteq \{\tilde{a} \cdot \partial_{\tilde{u}} \tilde{\sigma}(\tilde{u}_0, \tilde{v}_0) + \tilde{b} \cdot \partial_{\tilde{v}} \tilde{\sigma}(\tilde{u}_0, \tilde{v}_0) \mid \tilde{a}, \tilde{b} \in \mathbb{R}\} = T_{\tilde{\sigma}}(\tilde{u}_0, \tilde{v}_0).$$

Finally, by symmetry, we can repeat all of the above work, but with the roles of σ and $\tilde{\sigma}$ interchanged; this results in the opposite subset relation

$$T_{\tilde{\sigma}}(\tilde{u}_0, \tilde{v}_0) \subseteq T_\sigma(u_0, v_0).$$

From this, we conclude that $T_\sigma(u_0, v_0)$ and $T_{\tilde{\sigma}}(\tilde{u}_0, \tilde{v}_0)$ are the same, as desired. □

In light of Theorem 5.4, which states that tangent planes are independent of parametrisation, we can now make sense of tangent planes of a surface.

Definition 5.12. *Given a surface $S \subseteq \mathbb{R}^n$ and $\mathbf{p} \in S$, we define the tangent plane to S at $\mathbf{p}$ by*

$$T_{\mathbf{p}}S = T_\sigma(u_0, v_0),$$

where $\sigma : U \to S$ is any parametrisation of S (in the sense of Definition 5.11), and where

$$\sigma(u_0, v_0) = \mathbf{p}, \qquad (u_0, v_0) \in U.$$

Moreover, an element of $T_{\mathbf{p}}S$ is called a tangent vector of S at $\mathbf{p}$.

Example 5.35. *Recall the cylinder from Example 5.27,*

$$\mathcal{C} = \{(x, y, z) \in \mathbb{R}^3 \mid x^2 + y^2 = 1\},$$

and consider the point $\mathbf{p} = (0, 1, 5) \in \mathcal{C}$. Let us now find the tangent plane $T_{\mathbf{p}}\mathcal{C}$.

From Definition 5.12, we see that we must find a parametrisation of $\mathcal{C}$ which covers $\mathbf{p}$. The simplest way is to take the parametrisation σ from Example 5.22, which satisfies

$$\mathbf{p} = (0, 1, 5) = \sigma\left(\frac{\pi}{2}, 5\right).$$

Furthermore, we can compute that

$$\partial_u \sigma\left(\frac{\pi}{2}, 5\right) = (-1, 0, 0), \qquad \partial_v \sigma\left(\frac{\pi}{2}, 5\right) = (0, 0, 1).$$

As a result, by Definition 5.12 again, we have that

$$\begin{aligned} T_{\mathbf{p}}\mathcal{C} &= T_\sigma\left(\frac{\pi}{2}, 5\right) \\ &= \left\{ a \cdot \partial_u \sigma\left(\frac{\pi}{2}, 5\right)\Big|_{\mathbf{p}} + b \cdot \partial_v \sigma\left(\frac{\pi}{2}, 5\right)\Big|_{\mathbf{p}} \mid a, b \in \mathbb{R} \right\} \\ &= \{a \cdot (-1, 0, 0)|_{(0,1,5)} + b \cdot (0, 0, 1)|_{(0,1,5)} \mid a, b \in \mathbb{R}\}. \end{aligned}$$

A plot of $\mathcal{C}$ and $T_{\mathbf{p}}\mathcal{C}$ is found in Figure 5.23.

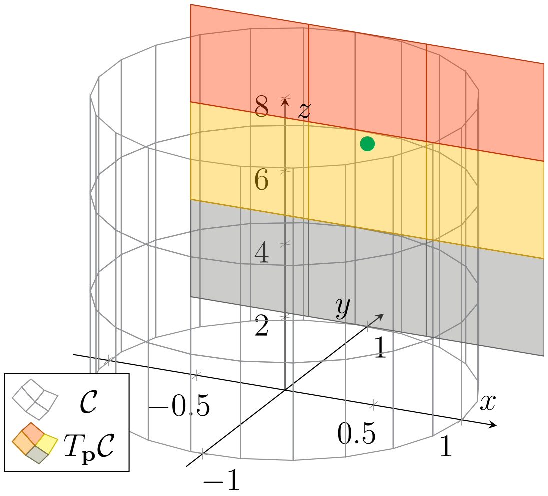


FIGURE 5.23. The cylinder $\mathcal{C}$ and the tangent plane $T_{\mathbf{p}}\mathcal{C}$ from Example 5.35. The point $\mathbf{p}$ is drawn in green.

We conclude our discussions here by connecting our results with linear algebra. Suppose again that we are in the setting of Theorem 5.4 (see also Figure 5.22 for a visual representation).

Now, from the discussion following Definition 5.7, we have that $T_{\mathbf{p}}S = T_\sigma(u_0, v_0)$ is a 2-dimensional vector space. Moreover, the "arrows" $\partial_u \sigma(u_0, v_0)|_{\mathbf{p}}$ and $\partial_v \sigma(u_0, v_0)|_{\mathbf{p}}$ form a basis for $T_{\mathbf{p}}S$. A similar argument holds for $\tilde{\sigma}$ as well, hence $\partial_{\tilde{u}} \tilde{\sigma}(\tilde{u}_0, \tilde{v}_0)|_{\mathbf{p}}$ and $\partial_{\tilde{v}} \tilde{\sigma}(\tilde{u}_0, \tilde{v}_0)|_{\mathbf{p}}$ is another basis for $T_p S = T_{\tilde{\sigma}}(\tilde{u}_0, \tilde{v}_0)$.

In other words, *for any parametrisation* (e.g. σ or $\tilde{\sigma}$) *of S which covers $\mathbf{p}$, we have a basis of the 2-dimensional vector space $T_{\mathbf{p}}S$ associated with this* *parametrisation.* Furthermore, *when we change our parametrisation*, say from σ to $\tilde{\sigma}$, then *this is observed at the level of the tangent plane as a change of basis of* T_pS.

To restate this more intuitively, we can think of a parametrisation of S as one person's particular perspective (or, in physics terms, *frame of reference*) of S. Thus, even though S is a single unchanging object, there are (infinitely) many different parametrisations of S, which correspond to the many different ways one can possibly view S.

In addition, one can also think of a basis of T_pS as a particular perspective (or frame of reference) of this tangent plane. Thus, by fixing any parametrisation, one also selects a particular way of viewing T_pS. When one changes parametrisations, the resulting change of basis of T_pS can hence be interpreted as a linear change of perspective for T_pS.

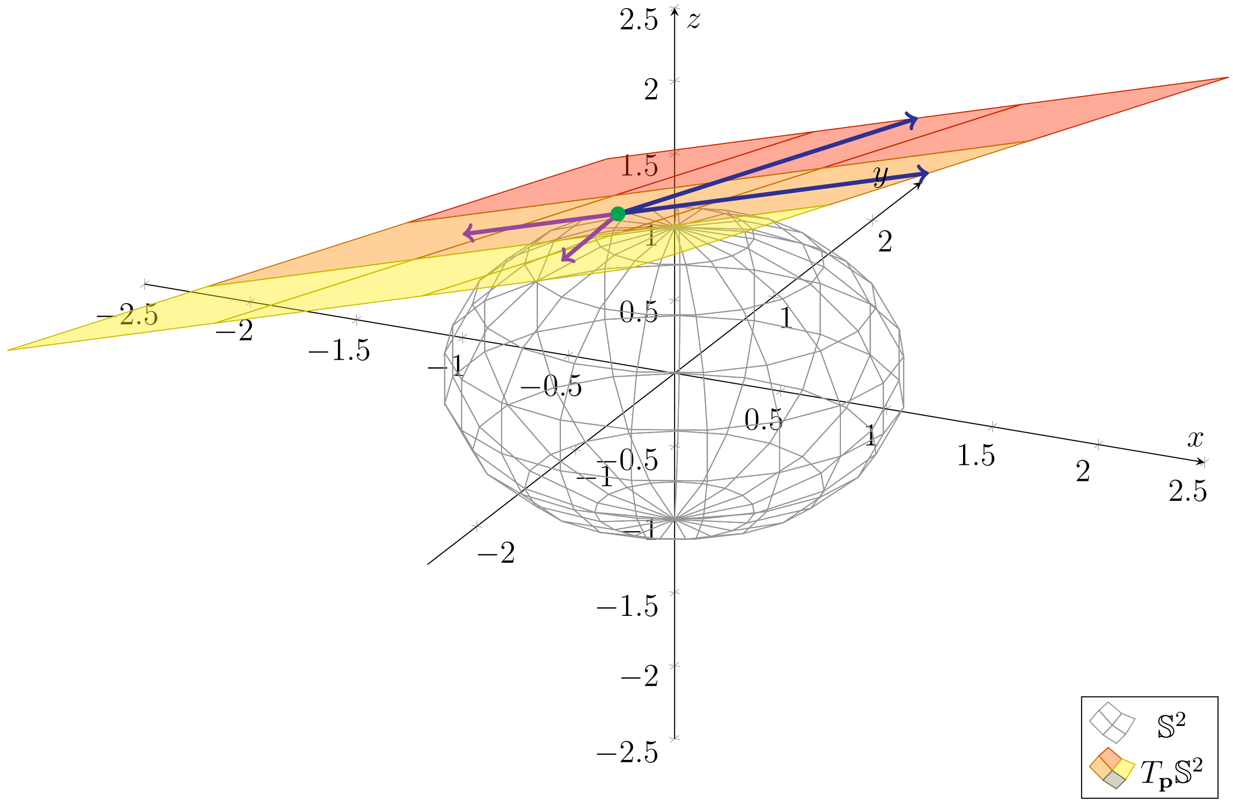


FIGURE 5.24. This plot shows $\mathbb{S}^2$ and $T_\mathbf{p}\mathbb{S}^2$ from Example 5.36. The basis vectors of $T_\mathbf{p}\mathbb{S}^2$ with respect to $\sigma_{x,-}$ are drawn in blue, while the basis vectors of $T_\mathbf{p}\mathbb{S}^2$ with respect to σ_s are drawn in purple.

Example 5.36. *Let us return to the sphere* $\mathbb{S}^2$*. Consider the tangent plane* $T_\mathbf{p}\mathbb{S}^2$*, where*

$$\mathbf{p} = \left(-\frac{1}{2}, \frac{1}{2}, \frac{1}{\sqrt{2}} \right) .$$

Note that $\mathbf{p}$ *is covered by the parametrisations* $\sigma_{x,-}$ *and* σ_s *(see Examples 5.34), and*

$$\mathbf{p} = \sigma_{x,-}\left(\frac{1}{2}, \frac{1}{\sqrt{2}} \right) = \sigma_s\left(\frac{3\pi}{4}, \frac{\pi}{4} \right) .$$

We can then compute the associated partial derivatives:

$$\partial_u \sigma_{x,-}(0,0) = (1,1,0), \qquad \partial_v \sigma_{x,-}(0,0) = (\sqrt{2},0,1),$$
$$\partial_u \sigma_s \left(\pi, \frac{\pi}{2}\right) = \left(-\frac{1}{2}, -\frac{1}{2}, 0\right), \qquad \partial_v \sigma_s \left(\pi, \frac{\pi}{2}\right) = \left(-\frac{1}{2}, \frac{1}{2}, -\frac{1}{\sqrt{2}}\right).$$

Then, the basis of $T_\mathbf{p}\mathbb{S}^2$ *associated with* $\sigma_{x,-}$ *is*

$$\{(1,1,0)|_\mathbf{p}, (\sqrt{2},0,1)|_\mathbf{p}\},$$

while the basis of $T_\mathbf{p}\mathbb{S}^2$ *associated with* σ_s *is*

$$\left\{ \left(-\frac{1}{2}, -\frac{1}{2}, 0\right)\bigg|_\mathbf{p}, \left(-\frac{1}{2}, \frac{1}{2}, -\frac{1}{\sqrt{2}}\right)\bigg|_\mathbf{p} \right\}.$$

These bases are graphically depicted in Figure 5.24.

5.3.2. *Normal Vectors.* Consider now a surface $S \subseteq \mathbb{R}^3$ lying in 3-dimensional space, as well as a point $\mathbf{p} \in S$. Since $T_\mathbf{p}S$ is a 2-dimensional plane sitting within a 3-dimensional space, then there is one remaining dimension that is perpendicular to $T_\mathbf{p}S$.

It is often useful to pick out a normalised *unit* vector from this remaining dimension:

Definition 5.13. *Let* $S \subseteq \mathbb{R}^3$ *be a surface, and let* $\mathbf{p} \in S$. *The "arrow"* $N|_\mathbf{p}$ *from* $\mathbf{p}$ *is said to be a unit normal to* S *at* $\mathbf{p}$ *iff* $N|_\mathbf{p}$ *is perpendicular to* $T_\mathbf{p}S$ *and has unit length (i.e.* $|N|_\mathbf{p}| = 1$*).*

Example 5.37. *Consider the sphere* $\mathbb{S}^2 \subseteq \mathbb{R}^3$ *and the point* $\mathbf{p} = (0,0,1) \in \mathbb{S}^2$. *Since* $\mathbf{p}$ *is the north pole—the "top" of the sphere—the tangent plane* $T_\mathbf{p}\mathbb{S}^2$ *is simply a copy of the* xy*-plane on* $\mathbf{p}$*:*

$$T_\mathbf{p}\mathbb{S}^2 = \{a \cdot (1,0,0)|_\mathbf{p} + b \cdot (0,1,0)|_\mathbf{p} \mid a, b \in \mathbb{R}\} = \{(a,b,0)|_\mathbf{p} \mid a, b \in \mathbb{R}\}.$$

Thus, the z*-direction is perpendicular to* $T_\mathbf{p}\mathbb{S}^2$.

In particular, the unit vectors in this z*-direction are* $(0,0,1)|_\mathbf{p}$ *and* $(0,0,-1)|_\mathbf{p}$. *Therefore, by Definition 5.13, both of these are unit normals to* $\mathbb{S}^2$ *at* $\mathbf{p} = (0,0,1)$.

See Figure 5.25 for a visual representation of this setting.

We begin with some general observations:

Proposition 5.5. *Let* $S \subseteq \mathbb{R}^3$ *be a surface. Then, there are exactly two unit normals to* S *at any* $\mathbf{p} \in S$.

Also, if $N|_\mathbf{p}$ *is one unit normal to* S *at* $\mathbf{p} \in S$*, then* $-N|_\mathbf{p}$ *is the other unit normal to* S *at* $\mathbf{p}$.

Proof. That there are exactly two unit normals is a consequence of the following observation: the vectors that are perpendicular to the tangent plane form a 1-dimensional line, from which there are exactly two elements with magnitude 1. For the remaining statement, we simply note that if $N|_\mathbf{p}$ satisfies the conditions of Definition 5.13, then $-N|_\mathbf{p}$ does as well. □

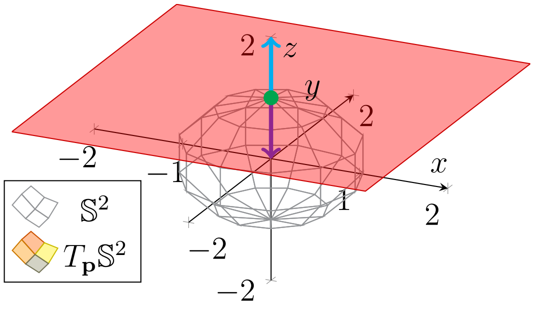


FIGURE 5.25. The setting from Example 5.37. The point $\mathbf{p}$ is drawn in green, while the unit normals $(0,0,\pm 1)|_\mathbf{p}$ to $\mathbb{S}^2$ at $\mathbf{p}$ are in blue and purple.

Now that we have precisely defined what unit normals are, we can ask the following:

Question 5.4. *How do we generally compute unit normals to surfaces?*

Like for tangent planes (see Definition 5.12), the idea is to use parametrisations in order to compute the unit normals. The result is summarised in the following theorem:

Theorem 5.6. *Let $S \subseteq \mathbb{R}^3$ be a surface. Suppose $\sigma : U \to S$ is a parametrisation of S, and let*

$$\mathbf{p} = \sigma(u_0, v_0), \qquad (u_0, v_0) \in U.$$

Then, the unit normals to S at $\mathbf{p}$ are precisely given by

$$\pm \left[\frac{\partial_u \sigma(u_0, v_0) \times \partial_v \sigma(u_0, v_0)}{|\partial_u \sigma(u_0, v_0) \times \partial_v \sigma(u_0, v_0)|} \right] \bigg|_{\mathbf{p}} .$$

Proof. Recall that the "arrows" $\partial_u \sigma(u_0, v_0)|_{\mathbf{p}}$ and $\partial_v \sigma(u_0, v_0)|_{\mathbf{p}}$ satisfy

$$\partial_u \sigma(u_0, v_0)|_{\mathbf{p}}, \partial_v \sigma(u_0, v_0)|_{\mathbf{p}} \in T_{\sigma}(u_0, v_0) = T_{\mathbf{p}} S,$$

that is, they point in directions tangent to S. Furthermore, since σ is regular by definition, then $\partial_u \sigma(u_0, v_0)|_{\mathbf{p}}$ and $\partial_v \sigma(u_0, v_0)|_{\mathbf{p}}$ point in different directions and hence span $T_{\sigma}(u_0, v_0) = T_{\mathbf{p}} S$.

As a result, by Proposition 3.13, the cross product

$$\partial_u \sigma(u_0, v_0)|_{\mathbf{p}} \times \partial_v \sigma(u_0, v_0)|_{\mathbf{p}} = [\partial_u \sigma(u_0, v_0) \times \partial_v \sigma(u_0, v_0)]|_{\mathbf{p}}$$

is nonzero and is perpendicular to both of its factors, and hence to $T_{\mathbf{p}} S$. To obtain the *unit* normals, we simply divide the above by its norm and recall Proposition 5.5. □

Example 5.38. *Let us return to the sphere $\mathbb{S}^2$. Consider a point $\mathbf{p} = (x_0, y_0, z_0) \in \mathbb{S}^2$, with $z_0 > 0$ (i.e. with $\mathbf{p}$ lying on the upper hemisphere). Recall that*

$$\mathbf{p} = \sigma_{z,+}(x_0, y_0),$$

where $\sigma_{z,+}$ was defined in Example 5.25:

$$\sigma_{z,+} : B_0 \to \mathbb{R}^3, \qquad \sigma_{z,+}(u, v) = (u, v, \sqrt{1 - u^2 - v^2}).$$

Recall also from Example 5.10 that the partial derivatives of $\sigma_{z,+}$ satisfy

$$\partial_u \sigma_{z,+}(u, v) = \left(1, 0, -\frac{u}{\sqrt{1 - u^2 - v^2}} \right), \qquad \partial_v \sigma_{z,+}(u, v) = \left(0, 1, -\frac{v}{\sqrt{1 - u^2 - v^2}} \right).$$

Taking their cross product, we see that

$$\partial_u \sigma_{z,+}(u, v) \times \partial_v \sigma_{z,+}(u, v) = \left(\frac{u}{\sqrt{1 - u^2 - v^2}}, \frac{v}{\sqrt{1 - u^2 - v^2}}, 1 \right).$$

Moreover, its norm satisfies

$$|\partial_u \sigma_{z,+}(u, v) \times \partial_v \sigma_{z,+}(u, v)| = \sqrt{\frac{u^2}{1 - u^2 - v^2} + \frac{v^2}{1 - u^2 - v^2} + 1} = \frac{1}{\sqrt{1 - u^2 - v^2}}.$$

From the above, we then obtain

$$\frac{\partial_u \sigma_{z,+}(u, v) \times \partial_v \sigma_{z,+}(u, v)}{|\partial_u \sigma_{z,+}(u, v) \times \partial_v \sigma_{z,+}(u, v)|} = (u, v, \sqrt{1 - u^2 - v^2}) = \sigma_{z,+}(u, v).$$

Thus, by Theorem 5.6, the unit normals to $\mathbb{S}^2$ *at* $\mathbf{p}$ *are given by*

$$\pm \left[\frac{\partial_u \sigma_{z,+}(x_0, y_0) \times \partial_v \sigma_{z,+}(x_0, y_0)}{|\partial_u \sigma_{z,+}(x_0, y_0) \times \partial_v \sigma_{z,+}(x_0, y_0)|} \right] \Bigg|_{\mathbf{p}} = \pm \sigma_{z,+}(x_0, y_0)|_{\mathbf{p}} = \pm \mathbf{p}|_{\mathbf{p}}.$$

In particular, the direction of the unit normals is the same as the position of the point where it is computed. To convince yourself that this is sensible, see the left plot of Figure 5.26.

In particular, at the north pole, $\mathbf{p} = (0, 0, 1)$*, the unit normals are*

$$\pm \mathbf{p}|_{\mathbf{p}} = (0, 0, \pm 1)|_{(0,0,1)},$$

which matches the expected answer from Example 5.37.

Example 5.39. *Let us now approach the sphere* $\mathbb{S}^2$ *using spherical coordinates,*

$$\sigma_s : \mathbb{R} \times (0, \pi) \to \mathbb{S}^2, \qquad \sigma_s(u, v) = (\cos u \sin v, \sin u \sin v, \cos v).$$

Recall from Example 5.24 that σ_s *is regular, and that*

$$\begin{aligned} \partial_u \sigma_s(u, v) &= (-\sin u \sin v, \cos u \sin v, 0), \\ \partial_v \sigma_s(u, v) &= (\cos u \cos v, \sin u \cos v, -\sin v), \\ \partial_u \sigma(u, v) \times \partial_v \sigma(u, v) &= (-\cos u \sin^2 v, -\sin u \sin^2 v, -\sin v \cos v), \\ |\partial_u \sigma(u, v) \times \partial_v \sigma(u, v)| &= \sin v. \end{aligned}$$

Now, given any point $\mathbf{p} = \sigma_s(u_0, v_0)$ *on the image of* σ_s*, we compute*

$$\frac{\partial_u \sigma_s(u_0, v_0) \times \partial_v \sigma_s(u_0, v_0)}{|\partial_u \sigma_s(u_0, v_0) \times \partial_v \sigma_s(u_0, v_0)|} = -(\cos u_0 \sin v_0, \sin u_0 \sin v_0, \cos v_0) = -\mathbf{p}.$$

By Theorem 5.6, we conclude that the unit normals to $\mathbb{S}^2$ *at* $\mathbf{p}$ *are given by* $\mp \mathbf{p}|_{\mathbf{p}}$*.*

In particular, combining this with Example 5.38, we see, as expected, that when $\mathbf{p}$ *lies in the images of both* $\sigma_{z,+}$ *and* σ_s*, the unit normals computed using* $\sigma_{z,+}$ *and* σ_s *coincide.*

Furthermore, by doing similar calculations with other parametrisations that cover $\mathbb{S}^2$, we can obtain the following: *at any point* $\mathbf{p} \in \mathbb{S}^2$*, the unit normals to* $\mathbb{S}^2$ *at* $\mathbf{p}$ *are precisely* $\pm \mathbf{p}|_{\mathbf{p}}$*.*

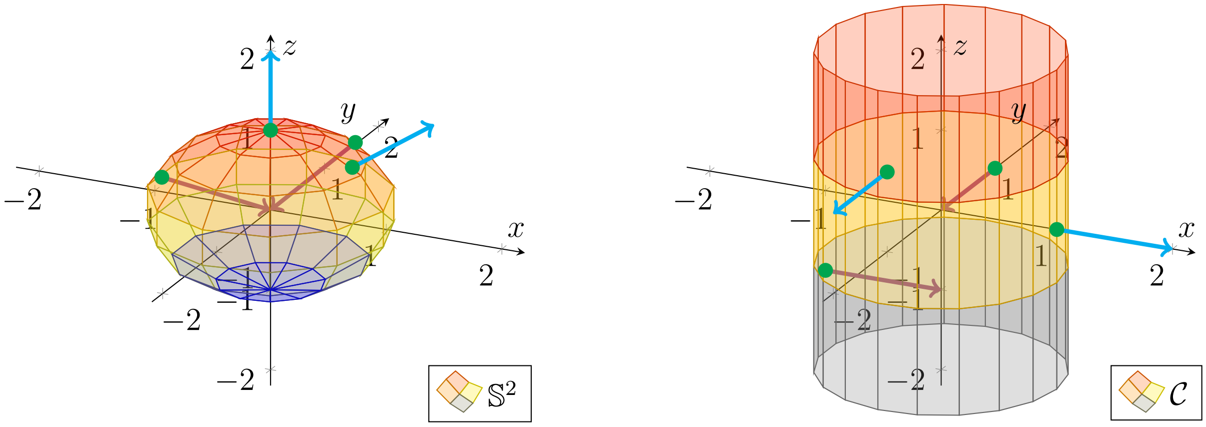


FIGURE 5.26. Plots of $\mathbb{S}^2$ (left) and cylinder $x^2 + y^2 = 1$ (right). Some unit normals of each surface are drawn in blue and purple.

Observe that in Example 5.38, taking the cross product $\partial_u \sigma_{z,+} \times \partial_v \sigma_{z,+}$ selects the unit normal $+\mathbf{p}|_{\mathbf{p}}$, which points *outwards* from the sphere. On the other hand, the cross product $\partial_u \sigma_s \times \partial_v \sigma_s$ in Example 5.39 selects the *inward-pointing* normal $-\mathbf{p}|_{\mathbf{p}}$. This is a general phenomenon applicable to all surfaces and parametrisations; which of the two normals is selected by the cross product in Theorem 5.6 depends on the particular parametrisation that is used.

Example 5.40. *Consider next the cylinder from Example 5.22,*

$$\mathcal{C} = \{(x, y, z) \in \mathbb{R}^3 \mid x^2 + y^2 = 1\},$$

which is covered by the regular parametrisation

$$\sigma : \mathbb{R}^2 \to \mathbb{R}^3, \qquad \sigma(u, v) = (\cos u, \sin u, v).$$

Recall from Example 5.22 that

$$\begin{aligned} \partial_u \sigma(u, v) &= (-\sin u, \cos u, 0), & \partial_v \sigma(u, v) &= (0, 0, 1), \\ \partial_u \sigma(u, v) \times \partial_v \sigma(u, v) &= (\cos u, \sin u, 0), & |\partial_u \sigma(u, v) \times \partial_v \sigma(u, v)| &= 1. \end{aligned}$$

Thus, for any $\mathbf{p} = \sigma(u_0, v_0) \in \mathcal{C}$, the unit normals to $\mathcal{C}$ at $\mathbf{p}$ are

$$\pm \left[\frac{\partial_u \sigma(u_0, v_0) \times \partial_v \sigma(u_0, v_0)}{|\partial_u \sigma(u_0, v_0) \times \partial_v \sigma(u_0, v_0)|} \right] \bigg|_{\mathbf{p}} = \pm (\cos u_0, \sin u_0, 0)|_{\mathbf{p}}.$$

This setting is drawn in the right plot in Figure 5.26.

Finally, we relate unit normals to another intuitive property of tangent planes—that they are "two-sided". Roughly speaking, we can think of a plane as a flat (infinite) piece of paper. Just as a piece of paper has a "front" and a "back" side, a plane, in an abstract sense, has the same.

More specifically, for a surface $S \subseteq \mathbb{R}^3$, we can think of a tangent plane $T_{\mathbf{p}}S$ as partitioning the space $\mathbb{R}^3$ into two "halves". Then, we can view one "side" of $T_{\mathbf{p}}S$ as facing one of the halves of $\mathbb{R}^3$, and the other "side" of $T_{\mathbf{p}}S$ as facing the other half; see the left picture of Figure 5.27.

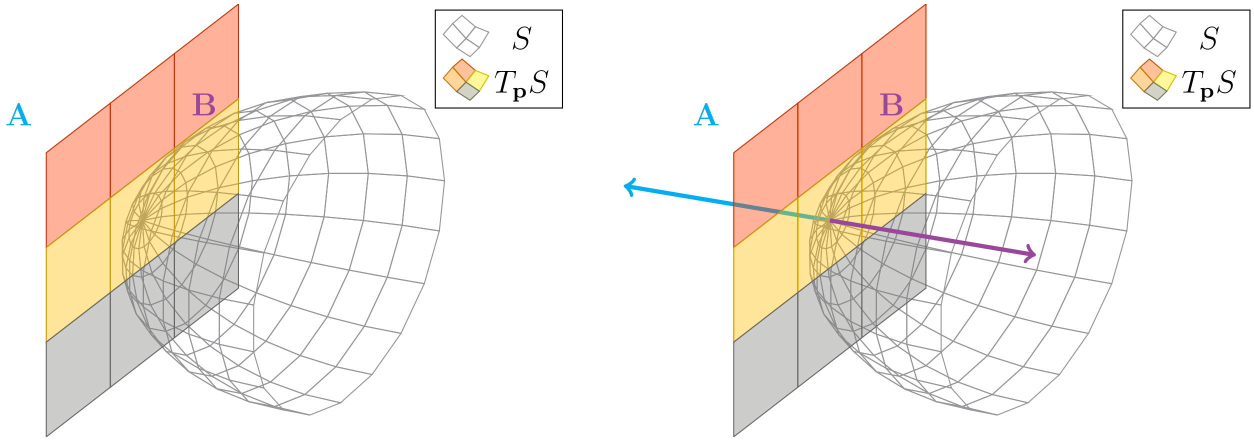


FIGURE 5.27. The left diagram shows the two sides (labelled "A" and "B") of a tangent plane $T_{\mathbf{p}}S$. In the right diagram, two unit normals (in blue and purple) extending from $T_{\mathbf{p}}S$ are added; the blue normal represents side "A", while the purple normal represents side "B".

Another precise way to capture these "sides" of $\mathsf{T}_{\mathbf{p}}\mathsf{S}$ is through unit normals. In particular, we can simply associate a unit normal $\mathsf{N}|_{\mathbf{p}}$ with the side of $\mathsf{T}_{\mathbf{p}}\mathsf{S}$ facing the direction that $\mathbf{N}|_{\mathbf{p}}$ is pointing. This is demonstrated in the right picture in Figure 5.27.

Now, since we can view $\mathsf{T}_{\mathbf{p}}\mathsf{S}$ as extending from this point $\mathbf{p}$, then we can also *view a unit normal of* S *at* $\mathbf{p}$ *as choosing a side of* S *at* $\mathbf{p}$. In other words, a surface at each point has two sides; one of the two sides can be "selected" by choosing one of the two unit normals at $\mathbf{p}$.

Example 5.41. *Consider the sphere* $\mathbb{S}^2$*. Recall from Examples 5.38 and 5.39, as well as the surrounding discussions, that at each* $\mathbf{p} \in \mathbb{S}^2$*, the unit normals to* $\mathbb{S}^2$ *at* $\mathbf{p}$ *are precisely* $\pm\mathbf{p}|_{\mathbf{p}}$.

Observe that the unit normal $\mathbf{p}|_{\mathbf{p}}$ *points outward from the sphere; see the blue arrows in the left plot of Figure 5.26. Thus,* $\mathbf{p}|_{\mathbf{p}}$ *represents the "outward-facing" side of* $\mathbb{S}^2$ *at* $\mathbf{p}$.

Similarly, the other unit normal $-\mathbf{p}|_{\mathbf{p}}$ *points inward from the sphere; see the purple arrows in the left plot of Figure 5.26. As a result,* $-\mathbf{p}|_{\mathbf{p}}$ *represents the "inward-facing" side of* $\mathbb{S}^2$ *at* $\mathbf{p}$.

Remark. In fact, even for surfaces in higher dimensions, i.e. $\mathsf{S} \subseteq \mathbb{R}^n$, we can still make sense of its tangent planes being "two-sided", and hence of S being "two-sided" at a point. However, this property is trickier to capture mathematically, hence we avoid discussing this here.

5.3.3. *Orientation.* Consider again a surface $\mathsf{S} \subseteq \mathbb{R}^3$. We have already discussed how S is "two-sided" at every $\mathbf{p} \in \mathsf{S}$, and we noted that a particular side can be represented by a unit normal of S at $\mathbf{p}$. Observe that the above concerns the *local* geometry of S, in that it deals only with a small portion of S. Next, we pose a similar question involving the *global* geometry of S.

Question 5.5. *Is* S *as a whole two-sided?*

The first step is to make Question 5.5 more precise. Since S being two-sided at $\mathbf{p}$ is manifested by the choice of a unit normal at $\mathbf{p}$, then the natural course of action is to associate global two-sidedness of S with a similarly global choice of unit normals at all points of S. However, this by itself is not quite enough. In particular, we need to ensure that given any two points $\mathbf{p}, \mathbf{q} \in \mathsf{S}$, the unit normals of S chosen at $\mathbf{p}$ and $\mathbf{q}$ "represent the same face of S" (though we have yet to precisely define what this means). These considerations motivate the following definition:

Definition 5.14. *A surface* $\mathsf{S} \subseteq \mathbb{R}^3$ *is* <u>*orientable*</u> *iff one can choose a unit normal* $\mathsf{N}|_{\mathbf{p}}$ *of* S *at every* $\mathbf{p} \in \mathsf{S}$ *in a way such that the direction of* $\mathsf{N}|_{\mathbf{p}}$ *varies smoothly with* $\mathbf{p}$.

In the case that S *is orientable, then a smoothly varying choice of unit normals* $\mathsf{N}|_{\mathbf{p}}$ *for each* $\mathsf{p} \in \mathsf{S}$*, as described above, is called an* <u>*orientation*</u> *of* S.

Some explanations regarding Definition 5.14 are in order. First, as mentioned before, the choice of unit normals corresponds to a choice of one side of S at each of its points. In addition, for the assumption that the normals $\mathsf{N}|_{\mathbf{p}}$ vary smoothly with $\mathbf{p}$, the intuition is that *this prevents us from "suddenly jumping" from one side of* S *to the other as one considers different points.* Thus, when S is orientable, we are allowed to make a consistent choice of one side of S everywhere.

In other words, we can interpret S *being orientable as* S *being globally two-sided.* Furthermore, *an orientation of* S—defined as a choice of smoothly varying unit normals—corresponds to a choice of one of the two sides of S (whenever two such sides exist).

Example 5.42. *Observe that the sphere $\mathbb{S}^2$ is orientable. To see this, we first recall (from Examples 5.38 and 5.39) that at each $\mathbf{p} \in \mathbb{S}^2$, the unit normals of $\mathbb{S}^2$ at $\mathbf{p}$ are precisely $\pm \mathbf{p}|_{\mathbf{p}}$. Then, one smooth choice of unit normals on $\mathbb{S}^2$ is to associate each $\mathbf{p} \in \mathbb{S}^2$ with $+\mathbf{p}|_{\mathbf{p}}$.*

Moreover, recall that the normals $\mathbf{p}|_{\mathbf{p}}$ point outward from the sphere; see the left plot in Figure 5.28. Thus, the orientation given by the $+\mathbf{p}|_{\mathbf{p}}$'s represents the outward-facing side of $\mathbb{S}^2$.

On the other hand, had we chosen the $-\mathbf{p}|_{\mathbf{p}}$'s instead (see the right plot in Figure 5.28), then we would have captured the opposite orientation, representing the inward-facing side of $\mathbb{S}^2$.

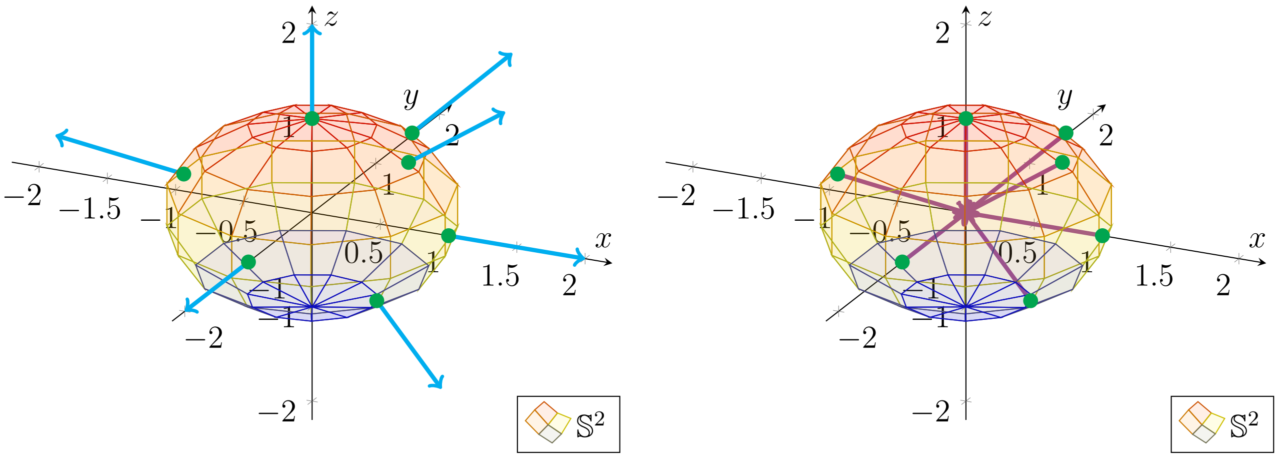


FIGURE 5.28. The left diagram shows $\mathbb{S}^2$ with the orientation representing its outward-facing side; some of the associated unit normals are drawn in blue. The right diagram shows the opposite inward-facing orientation of $\mathbb{S}^2$, with some associated unit normals drawn in purple.

Next, we consider an example of a *non-orientable* surface: the Möbius strip. One simple way to view the Möbius strip is through the gluing construction we previous applied to the Klein bottle (see Figure 5.19). More specificaly, we begin with a rectangle, and we glue two opposite edges together "in reverse order"; this is demonstrated in Figure 5.29. To try this yourself, take a thin strip of paper, twist it halfway, and then glue the two ends together.

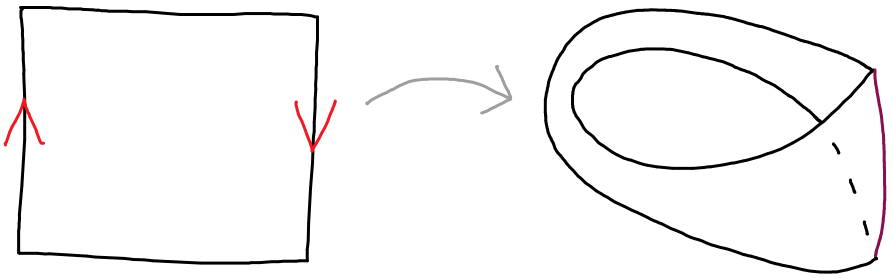


FIGURE 5.29. A demonstration of how the Möbius strip can be realised by gluing together two opposite edges of a rectangle.

Intuitively, the Möbius strip is not orientable since it has only one side rather than two. Indeed, although any small portion of the strip is two sided, at the global scale, the two sides blend into each other whenever one goes a full revolution around the strip.

Example 5.43. *One explicit parametrisation of the Möbius strip is as follows:*

$$\sigma : (-1,1) \times \mathbb{R} \to \mathbb{R}^3, \qquad \sigma(u,v) = \left(\left(1 - \frac{u}{2} \sin \frac{v}{2} \right) \cos v, \left(1 - \frac{u}{2} \sin \frac{v}{2} \right) \sin v, \frac{u}{2} \cos \frac{v}{2} \right).$$

See Figure 5.30 for a plot of σ.

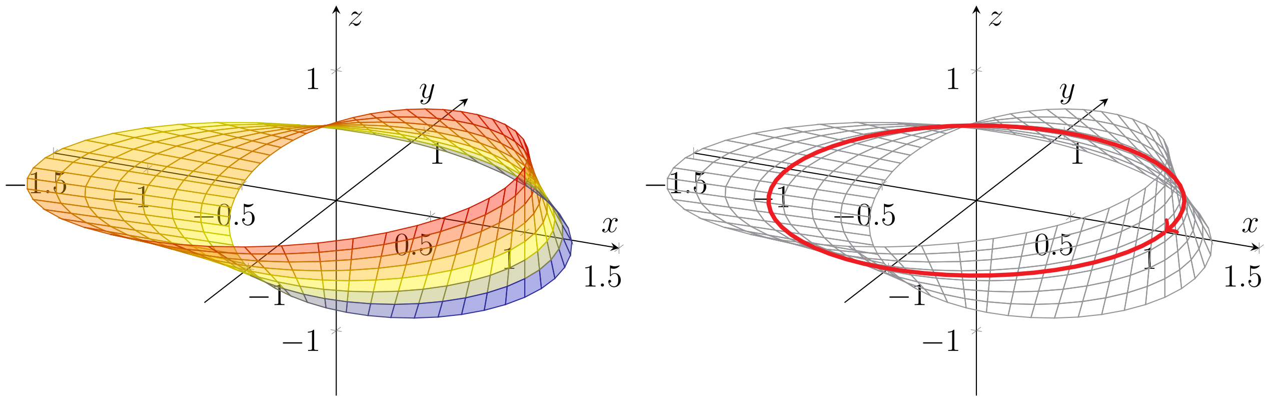
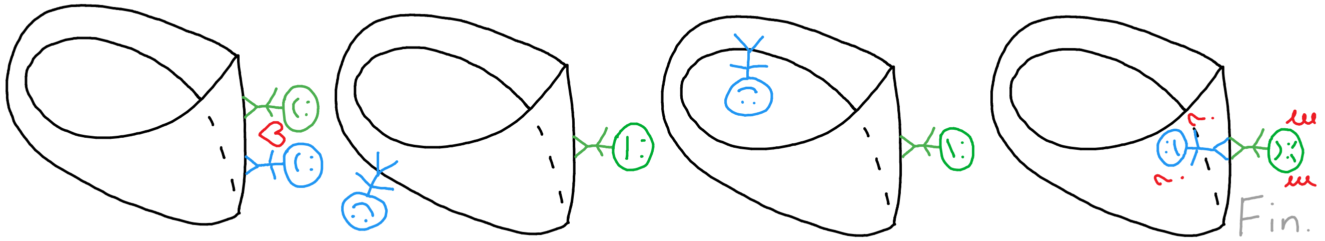


FIGURE 5.30. The left plot contains the graph of the parametric Möbius strip σ from Example 5.43. The right plot demonstrates why the Möbius strip fails to be orientable—in particular, any choice of unit normals along the strip must fail to be continuous at some point.

We can also give a heuristic argument in the context of Definition 5.14 for why the Möbius strip fails to be orientable. Consider the strip M in the right picture in Figure 5.30, and fix, for example, the outward-pointing unit normal $N|_{\mathbf{p}}$ (in blue) at the point $\mathbf{p} \in M$ (in green). Let us now try to construct from $N|_{\mathbf{p}}$ a smooth choice of unit normals along all of M.

As we traverse along the red curve on M, the corresponding unit normals are determined by the requirement that they vary smoothly; these normals are drawn in purple. The problem arises once we travel a full lap around M along the red curve. In particular, once we return to $\mathbf{p}$, the unit normal that we are forced to have is the opposite of the one we began with. Thus, our attempt to construct a smooth global choice of unit normals must necessarily fail.



The above heuristic argument can be converted to a formal proof using the parametric representation σ from Example 5.43 and the unit normals constructed using Theorem 5.6.

Remark. One can also define orientability and orientation for general surfaces $S \subseteq \mathbb{R}^n$. However, this is trickier to capture mathematically, hence we do not discuss this in this module.

For example, the Klein bottle (see, for instance, Example 5.32) fails to be orientable.

6. The Geometry of Surfaces

Thus far, we have given a precise definition of surfaces. Moreover, we have defined the tangent plane to a surface $S \subseteq \mathbb{R}^n$ at any point $\mathbf{p} \in S$, representing the directions and speeds one can go along S at $\mathbf{p}$. In this chapter, we apply these concepts to explore the geometry of surfaces.

In particular, at the end of this chapter (and these notes), we discuss some landmark results in surface geometry, in particular the *theorema egregium* and the *Gauss-Bonnet theorem*.

6.1. **The First Fundamental Form.**

We begin with the question of "size": *how large is a surface S, or part of S?* One critical ingredient for measuring size is the ability to determine the lengths of vectors. In addition, in 2-dimensions, we now have infinitely many directions to contend with. This leads to another question: *given two directions, what is the angle between them?*

Recall that the issues of measuring length and angle are fundamentally tied to the *dot product*. Since our concern lies within a surface S, this motivates the following definition:

Definition 6.1. *Given a surface $S \subseteq \mathbb{R}^n$ and a point $\mathbf{p} \in S$, we define the first fundamental form of S at $\mathbf{p}$ to be the dot product on the tangent plane $T_{\mathbf{p}}S$.*

To put it simply, the first fundamental form at $\mathbf{p}$ provides us with the means to measure lengths and angles in S, at least at $\mathbf{p}$. The main objective of this section, then, is to show how this information can be used to compute the "sizes", or areas, of surfaces.

Remark. The term "first fundamental form" is mostly historical. While it is often used in introductory courses in surface geometry, it is rarely used in more modern contexts. In contemporary differential geometry, the term most often used is induced metric, or metric; see [1].

6.1.1. *Parametric Representations.*

While the definition of the first fundamental form is simple enough, our aim is to use it to compute properties of surfaces. Since we usually do not work with a surface S directly, but rather through parametrisations of S, it will be useful for us to express the first fundamental form in terms of such a parametrisation $\sigma : U \to S$.

Given this σ, let us also fix a point $\mathbf{p} = \sigma(u_0, v_0)$, for some $(u_0, v_0) \in U$, as well as two tangent vectors $\mathbf{w}|_{\mathbf{p}}, \mathbf{z}|_{\mathbf{p}} \in T_{\mathbf{p}}S$. Our objective, then, is to express the dot product

$$\mathbf{w}|_{\mathbf{p}} \cdot \mathbf{z}|_{\mathbf{p}} = \mathbf{w} \cdot \mathbf{z} \tag{6.1}$$

in terms of σ. For this, let us also recall that the tangent vectors $\partial_u \sigma(u_0, v_0)|_{\mathbf{p}}$ and $\partial_v \sigma(u_0, v_0)|_{\mathbf{p}}$ form a basis of $T_{\mathbf{p}}S$. As a result, we can express $\mathbf{w}|_{\mathbf{p}}$ and $\mathbf{z}|_{\mathbf{p}}$ as

$$\mathbf{w}|_{\mathbf{p}} = w_u \cdot \partial_u \sigma(u_0, v_0)|_{\mathbf{p}} + w_v \cdot \partial_v \sigma(u_0, v_0)|_{\mathbf{p}},$$
$$\mathbf{z}|_{\mathbf{p}} = z_u \cdot \partial_u \sigma(u_0, v_0)|_{\mathbf{p}} + z_v \cdot \partial_v \sigma(u_0, v_0)|_{\mathbf{p}}.$$

In other words, w_u and w_v are the *components* of $\mathbf{w}|_{\mathbf{p}}$ with respect to the above basis obtained from σ. The same statement can also be made for z_u, z_v, and $\mathbf{z}|_{\mathbf{p}}$.

Combining the above, we can now expand (6.1) as

$$\mathbf{w}|_{\mathbf{p}} \cdot \mathbf{z}|_{\mathbf{p}} = [w_u \cdot \partial_u \sigma(u_0, v_0)|_{\mathbf{p}} + w_v \cdot \partial_v \sigma(u_0, v_0)|_{\mathbf{p}}] \cdot [z_u \cdot \partial_u \sigma(u_0, v_0)|_{\mathbf{p}} + z_v \cdot \partial_v \sigma(u_0, v_0)|_{\mathbf{p}}]$$

$$= w_u z_u [\partial_u \sigma(u_0, v_0) \cdot \partial_u \sigma(u_0, v_0)] + w_u z_v [\partial_u \sigma(u_0, v_0) \cdot \partial_v \sigma(u_0, v_0)]$$
$$+ w_v z_u [\partial_v \sigma(u_0, v_0) \cdot \partial_u \sigma(u_0, v_0)] + w_v z_v [\partial_v \sigma(u_0, v_0) \cdot \partial_v \sigma(u_0, v_0)],$$

where the quantities inside the brackets denote *dot products* of partial derivatives of σ. Rearranging a bit, we can then write the above as a *matrix product*,

$$\mathbf{w}|_\mathbf{p} \cdot \mathbf{z}|_\mathbf{p} = \begin{bmatrix} w_u & w_v \end{bmatrix} \begin{bmatrix} \partial_u \sigma(u_0, v_0) \cdot \partial_u \sigma(u_0, v_0) & \partial_u \sigma(u_0, v_0) \cdot \partial_v \sigma(u_0, v_0) \\ \partial_v \sigma(u_0, v_0) \cdot \partial_u \sigma(u_0, v_0) & \partial_v \sigma(u_0, v_0) \cdot \partial_v \sigma(u_0, v_0) \end{bmatrix} \begin{bmatrix} z_u \\ z_v \end{bmatrix} .$$

In other words, (6.1) is expressed as a product of matrices obtained from the components w_u, w_v, z_u, z_v of $\mathbf{w}|_\mathbf{p}$ and $\mathbf{z}|_\mathbf{p}$, as well as another matrix comprised of partial derivatives of σ. In particular, this latter matrix contains precisely the contents of the dot product on $T_\mathbf{p} S$.

The preceding derivation inspires the following definition:

Definition 6.2. *Assume the setting of Definition 6.1, and let* $\sigma : U \to S$ *be a parametrisation of* S*. We define the* <u>*first fundamental form*</u> *of* S *with respect to* σ *to be the matrix-valued function*

$$F^I_\sigma(u, v) = \begin{bmatrix} \partial_u \sigma(u, v) \cdot \partial_u \sigma(u, v) & \partial_u \sigma(u, v) \cdot \partial_v \sigma(u, v) \\ \partial_v \sigma(u, v) \cdot \partial_u \sigma(u, v) & \partial_v \sigma(u, v) \cdot \partial_v \sigma(u, v) \end{bmatrix} , \qquad (u, v) \in U.$$

In other words, in order to compute the dot product on $T_\mathbf{p} S$ from the point of view of a parametrisation σ of S, the main step is to calculate the quantity F^I_σ.

Example 6.1. *Recall the cylinder* $\mathcal{C}$ *from Example 5.22, for which a parametrisation is given by*

$$\sigma : \mathbb{R}^2 \to \mathbb{R}^3, \qquad \sigma(u, v) = (\cos u, \sin u, v).$$

Let us compute the first fundamental form of $\mathcal{C}$ *with respect to* σ.

First, recall that in Example 5.22, we also computed

$$\partial_u \sigma(u, v) = (-\sin u, \cos u, 0), \qquad \partial_v \sigma(u, v) = (0, 0, 1).$$

Taking dot products, we then see that

$$\partial_u \sigma(u, v) \cdot \partial_u \sigma(u, v) = 1, \qquad \partial_u \sigma(u, v) \cdot \partial_v \sigma(u, v) = 0,$$
$$\partial_v \sigma(u, v) \cdot \partial_u \sigma(u, v) = 0, \qquad \partial_v \sigma(u, v) \cdot \partial_v \sigma(u, v) = 1.$$

Thus, by Definition 6.2, the first fundamental form of $\mathcal{C}$ *with respect to* σ *is*

$$F^I_\sigma(u, v) = \begin{bmatrix} \partial_u \sigma(u, v) \cdot \partial_u \sigma(u, v) & \partial_u \sigma(u, v) \cdot \partial_v \sigma(u, v) \\ \partial_v \sigma(u, v) \cdot \partial_u \sigma(u, v) & \partial_v \sigma(u, v) \cdot \partial_v \sigma(u, v) \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} ,$$

that is, the identity matrix at every point.

In particular, the above implies that both $\partial_u \sigma$ *and* $\partial_v \sigma$ *have unit length everywhere, and that* $\partial_u \sigma$ *and* $\partial_v \sigma$ *are everywhere perpendicular to each other.*

Example 6.2. *Next, consider the sphere* $\mathbb{S}^2$*, with the (regular) parametrisation*

$$\sigma_s : \mathbb{R} \times (0, \pi) \to \mathbb{R}^3, \qquad \sigma_s(u, v) = (\cos u \sin v, \sin u \sin v, \cos v),$$

from Example 5.24. Recall that

$$\partial_u \sigma_s(u,v) = (-\sin u \sin v, \cos u \sin v, 0),$$
$$\partial_v \sigma_s(u,v) = (\cos u \cos v, \sin u \cos v, -\sin v).$$

Taking dot products, we obtain

$$\partial_u \sigma_s(u,v) \cdot \partial_u \sigma_s(u,v) = \sin^2 v, \qquad \partial_u \sigma_s(u,v) \cdot \partial_v \sigma_s(u,v) = 0,$$
$$\partial_v \sigma_s(u,v) \cdot \partial_u \sigma_s(u,v) = 0, \qquad \partial_v \sigma_s(u,v) \cdot \partial_v \sigma_s(u,v) = 1.$$

Recalling Definition 6.2 then yields

$$F^I_{\sigma_s}(u,v) = \begin{bmatrix} \sin^2 v & 0 \\ 0 & 1 \end{bmatrix}.$$

In particular, the above $\partial_v \sigma_s$ always has unit length, while the length of $\partial_u \sigma_s$ varies with the v-coordinate. Furthermore, $\partial_u \sigma_s$ and $\partial_v \sigma_s$ are everywhere perpendicular to each other.

Example 6.3. *Consider next the paraboloid*

$$S = \{(x, y, x^2 + y^2) \mid x, y \in \mathbb{R}\},$$

which can also be described as the image of the parametric surface

$$\sigma : \mathbb{R}^2 \to \mathbb{R}^3, \qquad \sigma(u,v) = (u, v, u^2 + v^2).$$

Note that its partial derivatives are given by

$$\partial_u \sigma(u,v) = (1, 0, 2u), \qquad \partial_v \sigma(u,v) = (0, 1, 2v).$$

Taking cross products, we see that

$$|\partial_u \sigma(u,v) \times \partial_v \sigma(u,v)| = |(-2u, -2v, 1)| = \sqrt{1 + 4u^2 + 4v^2}.$$

Since the above is always nonzero, Theorem 5.3 implies that σ is regular and hence is a parametrisation of S. Next, taking dot products, we obtain that

$$\partial_u \sigma(u,v) \cdot \partial_u \sigma(u,v) = 1 + 4u^2, \qquad \partial_u \sigma(u,v) \cdot \partial_v \sigma(u,v) = 4uv,$$
$$\partial_v \sigma(u,v) \cdot \partial_u \sigma(u,v) = 4uv, \qquad \partial_v \sigma(u,v) \cdot \partial_v \sigma(u,v) = 1 + 4v^2.$$

Thus, the first fundamental form with respect to σ is

$$F^I_{\sigma}(u,v) = \begin{bmatrix} 1 + 4u^2 & 4uv \\ 4uv & 1 + 4v^2 \end{bmatrix}.$$

Finally, let us return to a general surface $S \subseteq \mathbb{R}^n$. Suppose now that we have two parametrisations $\sigma : U \to S$ and $\tilde{\sigma} : \tilde{U} \to S$ of S. Suppose in addition that

$$\mathbf{p} = \sigma(u_0, v_0) = \tilde{\sigma}(\tilde{u}_0, \tilde{v}_0),$$

as indicated in Figure 6.1 below. We can then ask the following:

Question 6.1. *How are the matrices F^I_{σ} and $F^I_{\tilde{\sigma}}$ related to each other?*

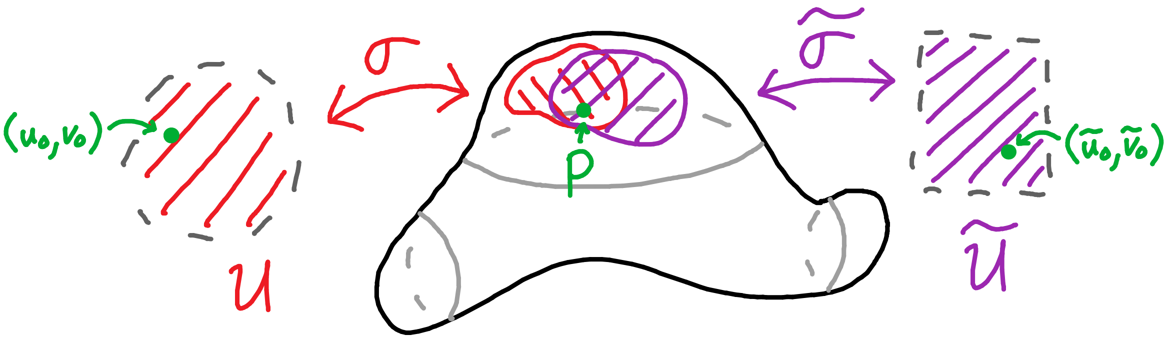


FIGURE 6.1. A reproduction of Figure 5.22.

Here, we briefly sketch the answer to Question 6.1. As in the proof of Theorem 5.4, we consider the change of variables between $(u, v) \in U$ and $(\tilde{u}, \tilde{v}) \in \tilde{U}$:

$$(\tilde{u}(u, v), \tilde{v}(u, v)) = \Phi(u, v) = \tilde{\sigma}^{-1}(\sigma(u, v)).$$

Then, using the (multivariable) chain rule, we can derive (see the proof of Theorem 5.4)

$$\begin{bmatrix} \partial_u \sigma \\ \partial_v \sigma \end{bmatrix} \Bigg|_{(u_0,v_0)} = J_{\sigma,\tilde{\sigma}}(u_0, v_0)^T \begin{bmatrix} \partial_{\tilde{u}} \tilde{\sigma} \\ \partial_{\tilde{v}} \tilde{\sigma} \end{bmatrix} \Bigg|_{(\tilde{u}_0,\tilde{v}_0)} , \qquad J_{\sigma,\tilde{\sigma}}(u_0, v_0) = \begin{bmatrix} \partial_u \tilde{u} & \partial_v \tilde{u} \\ \partial_u \tilde{v} & \partial_v \tilde{v} \end{bmatrix} \Bigg|_{(u_0,v_0)} ,$$

where the "T" denotes the *transpose* of a matrix.

Substituting this into Definition 6.2 and applying another extensive computation yields

$$\begin{aligned} F^{\sigma}_{I}(u_0, v_0) &= \begin{bmatrix} \partial_u \sigma \cdot \partial_u \sigma & \partial_u \sigma \cdot \partial_v \sigma \\ \partial_v \sigma \cdot \partial_u \sigma & \partial_v \sigma \cdot \partial_v \sigma \end{bmatrix} \Bigg|_{(u_0,v_0)} \\ &= J_{\sigma,\tilde{\sigma}}(u_0, v_0)^T \begin{bmatrix} \partial_{\tilde{u}} \tilde{\sigma} \cdot \partial_{\tilde{u}} \tilde{\sigma} & \partial_{\tilde{u}} \tilde{\sigma} \cdot \partial_{\tilde{v}} \tilde{\sigma} \\ \partial_{\tilde{v}} \tilde{\sigma} \cdot \partial_{\tilde{u}} \tilde{\sigma} & \partial_{\tilde{v}} \tilde{\sigma} \cdot \partial_{\tilde{v}} \tilde{\sigma} \end{bmatrix} \Bigg|_{(\tilde{u}_0,\tilde{v}_0)} J_{\sigma,\tilde{\sigma}}(u_0, v_0). \end{aligned}$$

Thus, we conclude that *at* $\mathbf{p}$, *the matrices* $F^{I}_{\sigma}(u_0, v_0)$ *and* $F^{I}_{\tilde{\sigma}}(\tilde{u}_0, \tilde{v}_0)$ *are related by the formula*

$$F^{\sigma}_{I}(u_0, v_0) = J_{\sigma,\tilde{\sigma}}(u_0, v_0)^T F^{I}_{\tilde{\sigma}}(\tilde{u}_0, \tilde{v}_0) J_{\sigma,\tilde{\sigma}}(u_0, v_0). \tag{6.2}$$

Remark. Observe that the matrix $J_{\sigma,\tilde{\sigma}}$ is precisely the Jacobian, from multivariable calculus, associated with the change of variables $(u, v) \leftrightarrow (\tilde{u}, \tilde{v})$. In particular, recall that this Jacobian is an important quantity in the *change of variables formula* for double integrals.

6.1.2. *Surface Area.* The next task is to connect the first fundamental form to the question of measuring the "size" of a surface. For curves, this meant computing its arc length. For surfaces, the corresponding question would be to compute the *surface area.*

Recall that to compute the arc length of a curve, we approximated the curve as line segments joining chosen sample points along the curve. Measuring the lengths of these line segments yields an approximation of the arc length of the curve. By taking more sample points along the curve, and hence a larger number of line segments, the approximate length becomes closer to the actual length. Finally, the arc length is obtained by taking a "limit" as the number of line segments approaches infinity. A graphical demonstration was given in Figure 2.15.

Now, in the context of surfaces, the idea is similar; we approximate the surface by "tiling" it using parallelograms, for which we already know how to compute the area. (If you do not know or remember this, a brief summary is given in the following page.) Taking the total areas of the parallelograms yields an approximation of the surface area, which can then be improved by refining the tiling and increasing the number of parallelograms. Finally, taking a "limit" as the number of such parallelograms tends toward infinity results in an integral formula for the surface area.

To be more specific, let us fix a surface S, as well as an injective parametrisation $\sigma : U \to S$. A "tiling" of (a portion of) S can then be achieved using σ. For this, we partition its domain U into a rectangular grid, with each rectangle having length Δu and height Δv. The parametrisation σ then maps each of these rectangles to a "curved rectangle" on S; see Figure 6.2.

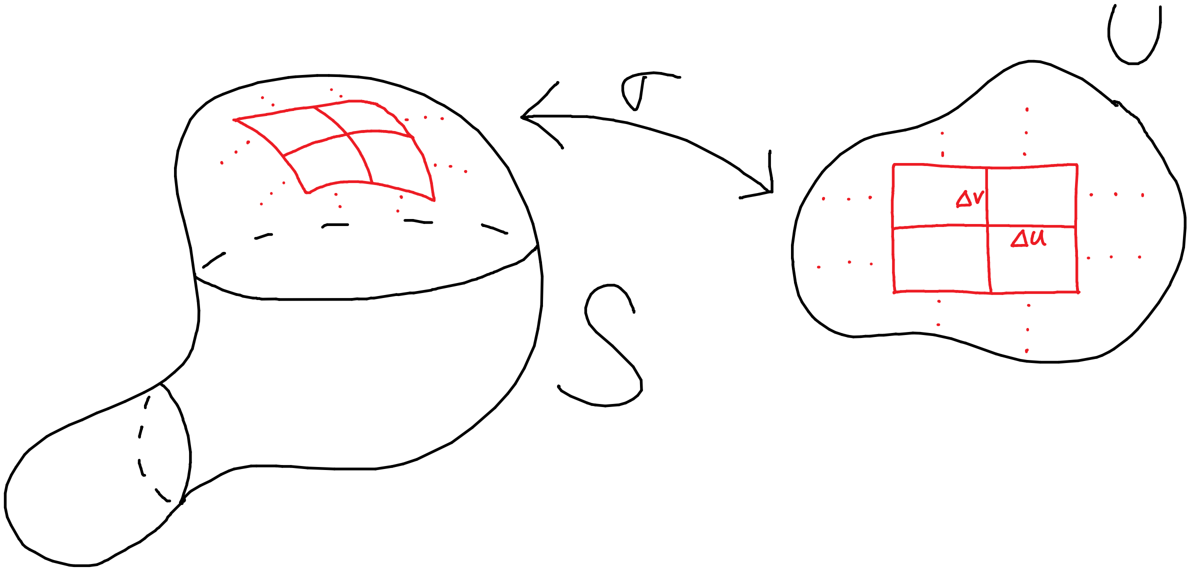


FIGURE 6.2. This shows how a surface S can be "tiled" using a parametrisation σ. One splits U is into rectangles and then maps them through σ.

In particular, to measure the surface area of the image $\sigma(U)$ of σ, we must find and sum up the areas of all the "curved rectangles" $\mathcal{R}$ comprising $\sigma(U)$:

$$\mathcal{A}(\sigma(U)) = \sum_{\text{Curved rectangles } \mathcal{R} \text{ in } \sigma(U)} \mathcal{A}(\mathcal{R}).$$

Of course, we generally cannot measure the area of a "curved rectangle" $\mathcal{R}$ exactly, just like we cannot find the exact area of S. However, we can approximate the area of $\mathcal{R}$ by replacing it with a parallelogram. In particular, the four corners of $\mathcal{R}$ are given by

$$\sigma(u_0, v_0), \qquad \sigma(u_0 + \Delta u, v_0), \qquad \sigma(u_0, v_0 + \Delta v), \qquad \sigma(u_0 + \Delta u, v_0 + \Delta v).$$

Thus, we can approximate $\mathcal{R}$ by the parallelgram $\mathcal{P}_{\mathcal{R}}$ with sides given by the vectors

$$\mathbf{a}|_{\sigma(u_0,v_0)} = [\sigma(u_0 + \Delta u, v_0) - \sigma(u_0, v_0)]|_{\sigma(u_0,v_0)},$$
$$\mathbf{b}|_{\sigma(u_0,v_0)} = [\sigma(u_0, v_0 + \Delta v) - \sigma(u_0, v_0)]|_{\sigma(u_0,v_0)}.$$

This process is shown in the left drawing of Figure 6.3.

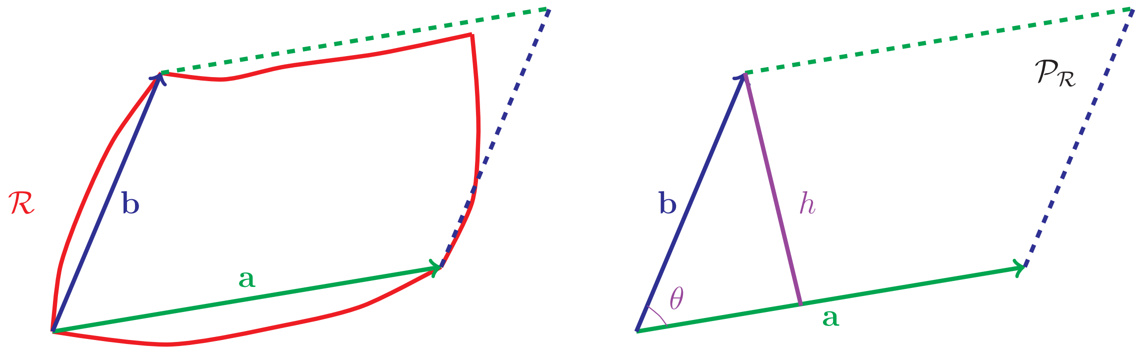


Figure 6.3. The left diagram is a crude drawing of how a curved rectangle $\mathcal{R}$ is approximated by a parallelogram $\mathcal{P}_{\mathcal{R}}$. The boundary of $\mathcal{R}$ is drawn in red, while the vectors $\mathbf{a}$ and $\mathbf{b}$ defining $\mathcal{P}_{\mathcal{R}}$ are drawn in green and blue, respectively. The right diagram describes the setup for computing the area of $\mathcal{P}_{\mathcal{R}}$ in the proof of Proposition 6.1.

To approximate the area of $\mathcal{R}$, we determine the area of $\mathcal{P}_{\mathcal{R}}$:

Proposition 6.1. *The area of the parallelogram $\mathcal{P}_{\mathcal{R}}$ defined above is*

$$\mathcal{A}(\mathcal{P}_{\mathcal{R}}) = \sqrt{(\mathbf{a} \cdot \mathbf{a})(\mathbf{b} \cdot \mathbf{b}) - (\mathbf{a} \cdot \mathbf{b})^2}.$$

When $n = 3$, that is, the surface is embedded in 3-dimensional space, we also have

$$\mathcal{A}(\mathcal{P}_{\mathcal{R}}) = |\mathbf{a} \times \mathbf{b}|.$$

Proof. Let θ denote the angle between the "arrows" $\mathbf{a}|_{\sigma(u_0,v_0)}$ and $\mathbf{b}|_{\sigma(u_0,v_0)}$. If we take $\mathbf{b}|_{\sigma(u_0,v_0)}$ to represent the base of $\mathcal{P}_{\mathcal{R}}$, then the corresponding height of $\mathcal{P}_{\mathcal{R}}$, indicated in purple in the right drawing in Figure 6.3, has length $h = |\mathbf{a}| \sin \theta$. As a result, we obtain

$$\mathcal{A}(\mathcal{P}_{\mathcal{R}}) = |\mathbf{b}| h = |\mathbf{a}||\mathbf{b}| \sin \theta.$$

When $n = 3$, Proposition 3.13 immediately implies, as desired,

$$\mathcal{A}(\mathcal{P}_{\mathcal{R}}) = |\mathbf{a} \times \mathbf{b}|.$$

In general, we expand further, while recalling Proposition 3.4:

$$\begin{aligned} [\mathcal{A}(\mathcal{P}_{\mathcal{R}})]^2 &= |\mathbf{a}|^2 |\mathbf{b}|^2 \sin^2 \theta \\ &= |\mathbf{a}|^2 |\mathbf{b}|^2 - |\mathbf{a}|^2 |\mathbf{b}|^2 \cos^2 \theta \\ &= (\mathbf{a} \cdot \mathbf{a})(\mathbf{b} \cdot \mathbf{b}) - (\mathbf{a} \cdot \mathbf{b}). \end{aligned}$$

This proves both identities in the statement of the proposition. □

We can now use Proposition 6.1 to approximate the area of $\sigma(U)$. Defining

$$\begin{aligned} \mathbf{a}_* &= \frac{\mathbf{a}}{\Delta u} = \frac{\sigma(u_0 + \Delta u, v_0) - \sigma(u_0, v_0)}{\Delta u}, \\ \mathbf{b}_* &= \frac{\mathbf{b}}{\Delta v} = \frac{\sigma(u_0, v_0 + \Delta v) - \sigma(u_0, v_0)}{\Delta v}, \end{aligned}$$

we then sum the areas of all the $\mathcal{P}_{\mathcal{R}}$'s:

$$\begin{aligned}\mathcal{A}(\sigma(U)) &\approx \sum_{\text{Curved rectangles } \mathcal{R}} \mathcal{A}(\mathcal{P}_{\mathcal{R}}) \\ &= \sum_{\mathcal{R}} \sqrt{(\mathbf{a}\cdot\mathbf{a})(\mathbf{b}\cdot\mathbf{b}) - (\mathbf{a}\cdot\mathbf{b})^2} \\ &= \sum_{\mathcal{R}} \Delta u \Delta v \sqrt{(\mathbf{a}_*\cdot\mathbf{a}_*)(\mathbf{b}_*\cdot\mathbf{b}_*) - (\mathbf{a}_*\cdot\mathbf{b}_*)^2}.\end{aligned}$$

Like in the case of curves, the above approximation can be refined by taking a smaller rectangular grid in U, i.e. by decreasing Δu and Δv. To generate the actual area of $\sigma(U)$, we need an "infinitely good" approximation, that is, we take $\Delta u, \Delta v \searrow 0$. In this case, we have that

$$\mathbf{a}_* \to \partial_u \sigma(u_0, v_0), \qquad \mathbf{b}_* \to \partial_v \sigma(u_0, v_0).$$

Moreover, the above summation becomes an integral. (Again, we omit the discussion of why this limit holds formally, as this is beyond the scope of this module.)

Combining all the above, we (at least informally) see that

$$\begin{aligned}\mathcal{A}(\mathcal{R}) &= \text{“}\lim_{\Delta u, \Delta v \searrow 0}\text{"} \sum_{\mathcal{R}} \Delta u \Delta v \sqrt{(\mathbf{a}_*\cdot\mathbf{a}_*)(\mathbf{b}_*\cdot\mathbf{b}_*) - (\mathbf{a}_*\cdot\mathbf{b}_*)^2} \\ &= \iint_U \sqrt{(\partial_u\sigma\cdot\partial_u\sigma)(\partial_v\sigma\cdot\partial_v\sigma) - (\partial_u\sigma\cdot\partial_v\sigma)^2}|_{(u,v)} \, du dv \\ &= \iint_U \sqrt{\det F_I^\sigma(u,v)} \, du dv.\end{aligned}$$

This leads us to the following definition:

Definition 6.3. *Let $S \subseteq \mathbb{R}^n$ be a surface, and let $\sigma : U \to S$ be an injective parametrisation of S. We then define the surface area of the portion $\sigma(U)$ of S to be*

$$\mathcal{A}(\sigma(U)) = \iint_U \sqrt{\det F_I^\sigma(u,v)} \, du dv.$$

Moreover, if we also perform the above summation and limit arguments using the second identity in Proposition 6.1, then we obtain the following area formula for surfaces in $\mathbb{R}^3$:

Theorem 6.2. *Assume the setting of Definition 6.3, and suppose also that $n = 3$. Then,*

$$\mathcal{A}(\sigma(U)) = \iint_U |\partial_u \sigma(u,v) \times \partial_v \sigma(u,v)| du dv.$$

Proof. Using the second identity in Proposition 6.1 instead, we see that

$$\mathcal{A}(\sigma(U)) \approx \sum_{\mathcal{R}} |\mathbf{a}\times\mathbf{b}| = \sum_{\mathcal{R}} \Delta u \Delta v |\mathbf{a}_* \times \mathbf{b}_*|.$$

The result again follows by taking the limits $\Delta u, \Delta v \searrow 0$. □

Example 6.4. *Fix a "height" $h > 0$, and consider the finite cylinder*

$$\mathcal{C}_h = \{(x,y,z) \in \mathbb{R}^3 \mid x^2 + y^2 = 1, \, 0 < z < h\}.$$

(This can be shown to be a surface.) We now compute the surface area of $\mathcal{C}_h$.

Consider the following parametrisation of $\mathcal{C}_h$:

$$\sigma : (0, 2\pi) \times (0, h) \to \mathcal{C}_h, \qquad \sigma(u, v) = (\cos u, \sin u, v).$$

By examining the graph of σ, we can see that the image of σ is all of $\mathcal{C}_h$ except for a vertical line; see the left plot in Figure 6.4. For this σ, we can then compute

$$\partial_u \sigma(u, v) = (-\sin u, \cos u, 0), \qquad \partial_v \sigma(u, v) = (0, 0, 1),$$

and it follows from Definition 6.2 that (the computation is identical to that of Example 6.1)

$$F^I_\sigma(u, v) = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

Now, let $\mathcal{I}$ denote the image of σ. Then, by Definition 6.3, its area is

$$\begin{aligned} \mathcal{A}(\mathcal{I}) &= \iint_{(0,2\pi)\times(0,h)} \sqrt{\det F^I_\sigma(u, v)}\, du dv \\ &= \int_0^{2\pi} du \int_0^h dv \\ &= 2\pi h, \end{aligned}$$

where we applied Fubini's theorem from multivariable calculus to split the double integral into iterated single integrals. Moreover, recall that $\mathcal{I}$ is all of $\mathcal{C}_h$ except for a 1-dimensional vertical line, which has area 0. (We take this fact for granted here, as formally demonstrating this is unfortunately beyond the scope of this module.) Consequently, the area of $\mathcal{C}_h$ is

$$\mathcal{A}(\mathcal{C}_h) = \mathcal{A}(\mathcal{I}) = 2\pi h.$$

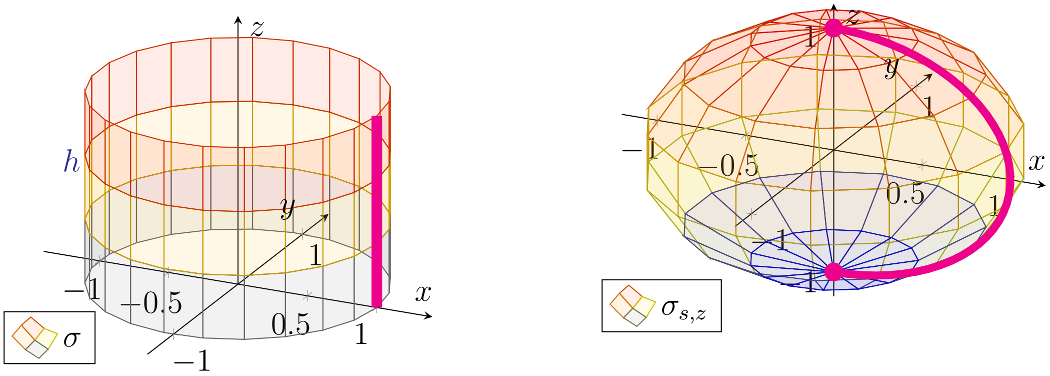


Figure 6.4. The left plot shows the image of the parametrisation σ from Example 6.4, while the right plot shows the image of σ_s from Example 6.5.

Next, you have probably had to learn earlier in your education that the surface area of a sphere of radius R is $4\pi R^2$. Below, we show how this formula is derived when $R = 1$. (Moreover, the computations here can be easily adapted to spheres with any radius.)

Example 6.5. *We now compute the surface area of $\mathbb{S}^2$. For this, we consider the parametrisation*

$$\sigma_{s,z} : (0, 2\pi) \times (0, \pi) \to \mathbb{R}^3, \qquad \sigma_{s,z}(u, v) = (\cos u \sin v, \sin u \sin v, \cos v),$$

introduced in Example 5.26. In particular, Example 5.26 showed that $\sigma_{s,z}$ *is regular and injective, and* $\sigma_{s,z}$ *covers all of* $\mathbb{S}^2$ *except for a closed arc; see the right plot in Figure 6.4.*

From computations in (5.24)*, we see that*

$$|\partial_u \sigma_{s,z}(u, v) \times \partial_v \sigma_{s,z}(u, v)| = \sin v.$$

Then, letting $\mathcal{I}$ *denote the image of* $\sigma_{s,z}$*, we see from Theorem 6.2 that*

$$\begin{aligned} \mathcal{A}(\mathcal{I}) &= \iint_{(0,2\pi)\times(0,\pi)} |\partial_u \sigma_{s,z}(u, v) \times \partial_v \sigma_{s,z}(u, v)| du dv \\ &= \int_0^{2\pi} du \int_0^{\pi} \sin v \, dv \\ &= 2\pi \int_0^{\pi} \sin v \, dv \\ &= 4\pi. \end{aligned}$$

Since $\mathbb{S}^2$ *differs from* $\mathcal{I}$ *only by a one-dimensional set, which has zero area, we conclude that*

$$\mathcal{A}(\mathbb{S}^2) = \mathcal{A}(\mathcal{I}) = 4\pi.$$

This result can also be obtained using Definition 6.3. The computations in Example 6.2 imply

$$F^I_{\sigma_{s,z}}(u, v) = \begin{bmatrix} \sin^2 v & 0 \\ 0 & 1 \end{bmatrix}, \qquad \det F^I_{\sigma_{s,z}} = \sin^2 v.$$

Thus, applying Definition 6.3 yields

$$\mathcal{A}(\mathbb{S}^2) = \mathcal{A}(\mathcal{I}) = \iint_{(0,2\pi)\times(0,\pi)} \sqrt{\sin^2 v} \, du dv = \int_0^{2\pi} du \int_0^{\pi} \sin v \, dv = 4\pi.$$

Note now that Definition 6.3 only defined the area of subsets of S which are the image of a single injective parametrisation of S. Furthermore, in Examples 6.4 and 6.5, the whole surface S only differed from the image of an injective parametrisation by a one-dimensional subset that has zero area. Thus, in practice, Definition 6.3 was sufficient to compute the full surface area in those examples. However, this may not necessarily be the case for other surfaces.

Question 6.2. *For a general surface* $S \subseteq \mathbb{R}^n$*, how do we compute the surface area of all of* S*?*

In full generality, the issue is that S may not be covered (or almost covered) by a single parametrisation. The idea, then, is that we have to take several different parametrisations of S which, when combined, cover all of S. For each of these parametrisations, we measure the surface area of its image. The total surface area of S can then be recovered by summing all the areas of the parametrisations. This process is encapsulated in the following informal statement:

> Let $\sigma_i : U_i \to S$, where $1 \leq i \leq M$, be parametrisations of S such that (i) the images $\sigma_i(U_i)$ are pairwise disjoint, and (ii) the union of all the images $\sigma_i(U_i)$ is all of S, except possibly a subset with zero area. Then, the total surface area of S can be defined as
>
> $$\mathcal{A}(S) = \sum_{i=1}^{M} \mathcal{A}(\sigma_i(U_i)).$$

Unfortunately, a fully formal description of the above process lies a bit beyond the scope of this module. Therefore, for all the examples we will consider within this module, the surface area will computable from just a single injective parametrisation.

6.1.3. *Surface Integrals.* Recall that in our study of curves, we generalised our formula for the arc length of curves to that of weighted arc lengths; this was the definition of *path integrals.* Similarly, we can extend our notion of surface area, given in Definition 6.3, to that of a *weighted* surface area.

Definition 6.4. *Let* $S \subseteq \mathbb{R}^n$ *be a surface, let* $\sigma : U \to S$ *be an injective parametrisation of* S*, and let* $G : S \to \mathbb{R}$*. We then define the* *<u>surface integral</u>* *of* G *over* $\sigma(U)$ *to be*

$$\iint_{\sigma(U)} G \, dA = \iint_U G(\sigma(u, v)) \sqrt{\det F_I^\sigma(u, v)} \, du dv.$$

Furthermore, the surface area of G *over all of* S *is defined as*

$$\iint_S G \, dA = \sum_{i=1}^M \iint_{\sigma_i(U_i)} G \, dA,$$

where $\sigma_i : U_i \to S$*,* $1 \leq i \leq M$*, are parametrisations of* S *whose images are pairwise disjoint and cover all of* S*, except possibly a subset with zero area.*

Remark. Like for the surface area in the preceding discussion, we will only consider in these notes surface integrals which are computable using only a single parametrisation.

Remark. When $n = 3$, the surface area could also be computed using the formula

$$\iint_{\sigma(U)} G \, dA = \iint_U G(\sigma(u, v)) |\partial_u \sigma(u, v) \times \partial_v \sigma(u, v)| du dv.$$

Remark. Like for path integrals, the surface area is a special case of the surface integral:

$$\mathcal{A}(S) = \iint_S 1 \, dA.$$

First, we note that the only difference between Definitions 6.3 and 6.4 is the presence of an additional function G, which represents the weight associated with our weighted surface area. In particular, for each $\mathbf{p} \in S$, the value $G(\mathbf{p})$ represents the weight applied at $\mathbf{p}$.

Observe that this weight G is inserted into the surface integral in the same manner that as was previously done in the definition of path integrals (see Definition 2.13). The only subtlety here is regarding where the weight G is applied. The main point to keep in mind is that the value

$$\sqrt{\det F_I^\sigma(u, v)} \, du dv$$

represents the area of an infinitesimal parallelogram *at the point* $\sigma(u, v) \in S$. Therefore, the appropriate weight to associate with this value in the integrand is $G(\sigma(u, v))$, that is, at $\sigma(u, v)$.

For a physics example, suppose we have an electrically charged unit sphere, which we model by $\mathbb{S}^2$. Suppose also that the charge is distributed unevenly along the sphere. We let $G(\mathbf{p})$ denote the charge density at any point $\mathbf{p} \in \mathbb{S}^2$, which can vary as $\mathbf{p}$ changes. ($G(\mathbf{p})$ can also be either positive or negative, depending on the sign of the charge.) Then, to compute *total charge* of the sphere, we would integrate the charge density G along $\mathbb{S}^2$:

$$\text{Total charge} = \iint_{\mathbb{S}^2} \mathsf{G}\, d\mathsf{A}.$$

In other words, we can view the total charge abstractly as a weighted area of $\mathbb{S}^2$.

Example 6.6. *Let us return to the sphere $\mathbb{S}^2$, and let us compute*

$$\iint_{\mathbb{S}^2} z^2\, d\mathsf{A}.$$

In particular, the weight G is given by

$$\mathsf{G}(x, y, z) = z^2.$$

Like in Example 6.5, we compute the integral using the parametrisation

$$\sigma_{s,z} : (0, 2\pi) \times (0, \pi) \to \mathbb{R}^3, \qquad \sigma_{s,z}(u, v) = (\cos u \sin v, \sin u \sin v, \cos v),$$

which we recall satisfies

$$\sqrt{\det \mathsf{F}^{\mathrm{I}}_{\sigma_{s,z}}} = \sin v.$$

Thus, letting $\mathcal{I}$ denote the image of $\sigma_{s,z}$, we obtain from Definition 6.4 that

$$\begin{aligned} \iint_{\mathbb{S}^2} z^2\, d\mathsf{A} &= \iint_{\mathcal{I}} \mathsf{G}\, d\mathsf{A} \\ &= \iint_{(0,2\pi)\times(0,\pi)} \mathsf{G}(\cos u \sin v, \sin u \sin v, \cos v) \sin v\, du dv \\ &= \int_0^{2\pi} du \int_0^{\pi} \cos^2 v \sin v\, dv \\ &= 2\pi \left[-\frac{1}{3} \cos^3 v \right]_0^{\pi} \\ &= \frac{4\pi}{3}. \end{aligned}$$

Example 6.7. *Consider the surface S defined as the image of the parametric surface*

$$\sigma : (0, 1) \times (0, 1) \to \mathbb{R}^3, \qquad \sigma(u, v) = (u, v, u^2 + v^2).$$

Clearly, σ is injective. It is also regular, since

$$|\partial_u \sigma(u, v) \times \partial_v \sigma(u, v)| = |(1, 0, 2u) \times (0, 1, 2v)| = \sqrt{1 + 4u^2 + 4v^2},$$

which is everywhere strictly positive.

Let us now compute the surface integral

$$\iint_{\mathsf{S}} \sqrt{1 + 2x^2 + 2y^2 + 2z}\, d\mathsf{A}.$$

Here, the integrand is the function defined

$$\mathsf{G}(x, y, z) = \sqrt{1 + 2x^2 + 2y^2 + 2z}.$$

Since S is the image of σ, then Definition 6.4 gives

$$\begin{aligned}\iint_S \sqrt{1+2x^2+2y^2+z}\,dA &= \iint_{(0,1)\times(0,1)} G(\sigma(u,v))|\partial_u\sigma(u,v)\times\partial_v\sigma(u,v)|dudv\\ &= \int_0^1\int_0^1 G(u,v,u^2+v^2)\sqrt{1+4u^2+4v^2}\,dudv\\ &= \int_0^1\int_0^1 \sqrt{1+2u^2+2v^2+2(u^2+v^2)}\sqrt{1+4u^2+4v^2}\,dvdu\\ &= \int_0^1\int_0^1 (1+4u^2+4v^2)dvdu.\end{aligned}$$

Evaluating the above double integral results in the answer:

$$\begin{aligned}\iint_S \sqrt{1+2x^2+2y^2+z}\,dA &= 1+4\int_0^1 u^2du+4\int_0^1 v^2dv\\ &= 1+\frac{4}{3}+\frac{4}{3}\\ &= \frac{11}{3}.\end{aligned}$$

There is still one important unfinished business in our discussion of surface integrals, and by extension, surface area. More specifically, surface integrals over a surface S were defined using any choice of injective parametrisations of S. However, we have not addressed whether the integral is independent of the parametrisations that were used to compute it. In other words, *we have not determined whether surface integrals (and also the surface area) are geometric properties of the surface* S. The subsequent theorem ties up this final loose end:

Theorem 6.3. *Let $S\subseteq\mathbb{R}^n$ be a surface, and let $G: S\to\mathbb{R}$. Moreover, suppose $\sigma: U\to S$ and $\tilde\sigma:\tilde U\to S$ are injective parametrisations of S, with $\sigma(U)=\tilde\sigma(\tilde U)$. Then,*

$$\iint_U G(\sigma(u,v))\sqrt{\det F^I_\sigma(u,v)}\,dudv = \iint_{\tilde U} G(\tilde\sigma(\tilde u,\tilde v))\sqrt{\det F^I_{\tilde\sigma}(\tilde u,\tilde v)}\,d\tilde u d\tilde v.$$

Sketch of proof. By the formula (6.2) relating F^I_σ and $F^I_{\tilde\sigma}$, we obtain that

$$\begin{aligned}\sqrt{\det F^I_\sigma(u,v)} &= \sqrt{\det J_{\sigma,\tilde\sigma}(u,v)^T\det F^I_{\tilde\sigma}(\tilde u,\tilde v)\det J_{\sigma,\tilde\sigma}(u,v)}\\ &= \sqrt{\det J_{\sigma,\tilde\sigma}(u,v)\det F^I_{\tilde\sigma}(\tilde u,\tilde v)\det J_{\sigma,\tilde\sigma}(u,v)}\\ &= \det J_{\sigma,\tilde\sigma}(u,v)\sqrt{\det F^I_{\tilde\sigma}(\tilde u,\tilde v)},\end{aligned}$$

where $(\tilde u,\tilde v)=\tilde\sigma^{-1}(\sigma(u,v))$. Then, by the change of variables formula for double integrals,

$$\begin{aligned}\iint_U G(\sigma(u,v))\sqrt{\det F^I_\sigma(u,v)}\,dudv &= \iint_U G(\sigma(u,v))\sqrt{\det F^I_{\tilde\sigma}(\tilde u,\tilde v)}\det J_{\sigma,\tilde\sigma}(u,v)\,dudv\\ &= \iint_{\tilde U} G(\tilde\sigma(\tilde u,\tilde v))\sqrt{\det F^I_{\tilde\sigma}(\tilde u,\tilde v)}\,d\tilde u d\tilde v.\end{aligned}$$

(In particular, $\det J_{\sigma,\tilde\sigma}(u,v)$ is precisely the factor needed in the change of variables formula.) □

6.2. **Curvature.** The previous section explored how one would measure the "size", or *area*, of a surface. Now, we study another fundamental aspect of geometry: the "shape" of a surface.

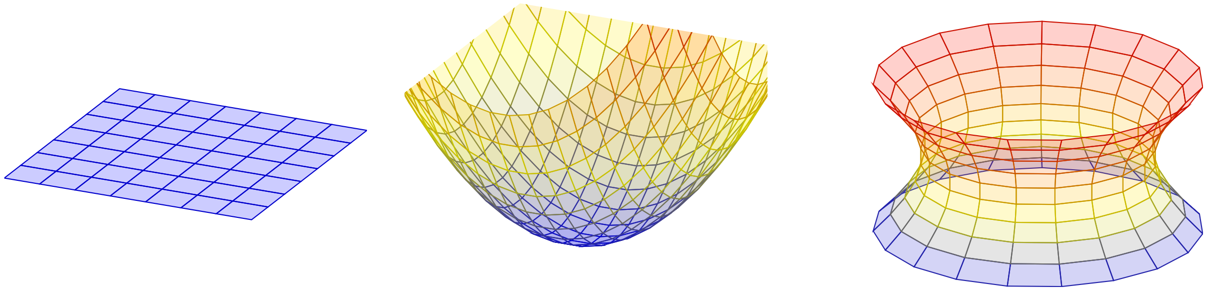


Figure 6.5. From left to right, a plane, paraboloid, and hyperboloid. Note that the plane is "flat", while the paraboloid and hyperboloid are "curved".

Consider a plane, a paraboloid, and a hyperboloid, as in Figure 6.5. Intuitively, we know that the plane is "flat", as it is not "curving" at any of its points. In contrast, our intuitions indicate that both the paraboloid and hyperboloid are "curved", in fact at all of their points. As we previously did for curves, we now ask how to make these intuitive ideas mathematically precise:

Question 6.3. *How do we make the notion of "curvature" of a surface mathematically precise? Moreover, how would we quantify and compute this "curvature"?*

The first observation is that Question 6.3 is considerably more complex than the corresponding question for curves, since there are many more directions to consider along a 2-dimensional surface. In particular, a surface can curve differently at various points and in various directions. Thus, a considerable amount of information is required to fully describe how a surface is curved.

For instance, the paraboloid "curves in the same way" (i.e. "upward", or away from its lower face) at each of its points and in all directions. However, the paraboloid can be "more curved" in some directions than others. In contrast, the hyperboloid bends differently in different directions; if one fixes a face of this hyperboloid, then some directions will curve toward this face, while others will curve away from it. In the upcoming discussions, we will discuss how to precisely characterise this information, and as a result address Question 6.3.

Remark. While one can make sense of curvature for surfaces in any dimension, for simplicity, we will only consider surfaces in $\mathbb{R}^3$ here. In particular, in $\mathbb{R}^3$, we can take advantage of the notion of unit normals to a surface, which are relatively simple to describe and compute.

6.2.1. *The Second Fundamental Form.* Let us return to the "flat" plane and "curved" paraboloid from Figure 6.3, and let us see we how we can capture this geometric difference. As mentioned in the preceding remark, one way to approach this is to consider their unit normals, a notion we previously discussed in Definition 5.13 and Theorem 5.5. Figure 6.6 depicts a sampling of what the unit normals look like for both the plane and the paraboloid.

Example 6.8. *Let* $a, b, c, d \in \mathbb{R}$ *such that* $(a, b, c) \neq (0, 0, 0)$. *In other words, the set*

$$P = \{(x, y, z) \in \mathbb{R}^3 \mid ax + by + cz = d\},$$

describes a general plane. We claim that the unit normal to P *at any* $\mathbf{p} \in \mathsf{P}$ *is*

$$N|_{\mathbf{p}} = \pm \left. \frac{(a,b,c)}{\sqrt{a^2+b^2+c^2}} \right|_{\mathbf{p}} .$$

To see this, let us first assume that $c \neq 0$*, so that* P *is covered by a single parametrisation*

$$\sigma_z : \mathbb{R}^2 \to \mathbb{R}^3, \qquad \sigma_z(u,v) = \left(u, v, \frac{1}{c}(d - au - bv) \right) .$$

(Essentially, we set u *and* v *to be* x *and* y*, respectively.) A direct computation shows that*

$$\partial_u \sigma_z(u,v) \times \partial_v \sigma_z(u,v) = \left(1, 0, -\frac{a}{c} \right) \times \left(0, 1, -\frac{b}{c} \right) = \left(\frac{a}{c}, \frac{b}{c}, 1 \right) ,$$

and applying Theorem 5.5 proves the claim:

$$N|_{\sigma(u,v)} = \pm \left. \frac{\partial_u \sigma_z(u,v) \times \partial_v \sigma_z(u,v)}{|\partial_u \sigma_z(u,v) \times \partial_v \sigma_z(u,v)|} \right|_{\sigma(u,v)} = \pm \left. \frac{(a,b,c)}{\sqrt{a^2+b^2+c^2}} \right|_{\sigma(u,v)} .$$

Finally, the computations in the remaining cases $b \neq 0$ *or* $a \neq 0$ *are analogous to the above.*

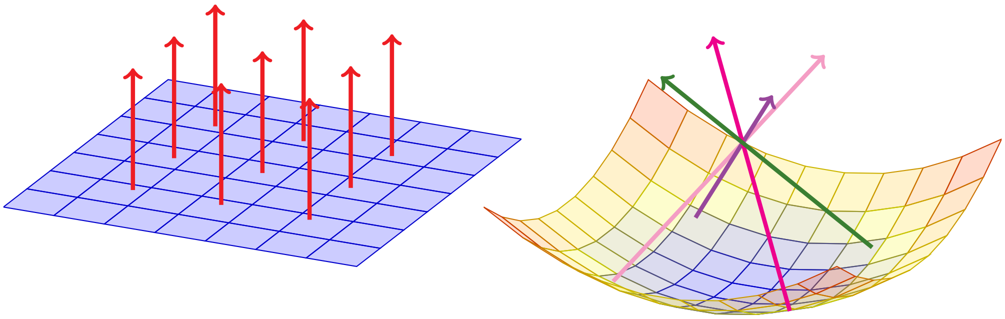


FIGURE 6.6. The plots depict the plane and parabola from Figure 6.5, with some unit normals drawn on each surface. In particular, the unit normals of the plane all have the same direction, while the unit normals of the parabola change directions depending on the position.

In particular, Example 6.8 shows that *for any plane, its unit normals have the same direction everywhere*. However, this property fails to hold for the "curved" parabola. Because of its curvature, the unit normals to the parabola point in different directions at different points; see the right diagram in Figure 6.6. The same could be said for the hyperboloid in Figure 6.5.

One way to interpret the above is to contend that it is the curvature of a surface that causes its unit normals to change directions as one moves along the surface. This lead us to the main idea for describing surface curvature: *the curvature of a surface can be characterised by how the directions of its unit normals change with the position*. In other words, to capture the curvature of a surface, we should measure the *derivative* of its unit normals.

Although this sounds simple in theory, this derivative contains many pieces of information:

(1) One can differentiate the unit normals in many different directions along S.
(2) The unit normal is a vector-valued function with three components.

Thus, to adequately characterise the curvature, we will need to unpack all of this information. As usual, to convert the above into digestible quantities, we work with parametrisations.

To pursue this more precisely, let us consider a surface $S \subseteq \mathbb{R}^3$. In addition, we fix a parametrisation $\sigma : U \to S$ of S. By Theorem 5.6, we can use σ to make a particular choice of unit normals:

$$N_\sigma(u, v) = +\frac{\partial_u \sigma(u, v) \times \partial_v \sigma(u, v)}{|\partial_u \sigma(u, v) \times \partial_v \sigma(u, v)|}.$$

As a result, our aim is now to look at derivatives of N_σ, in various directions along S.

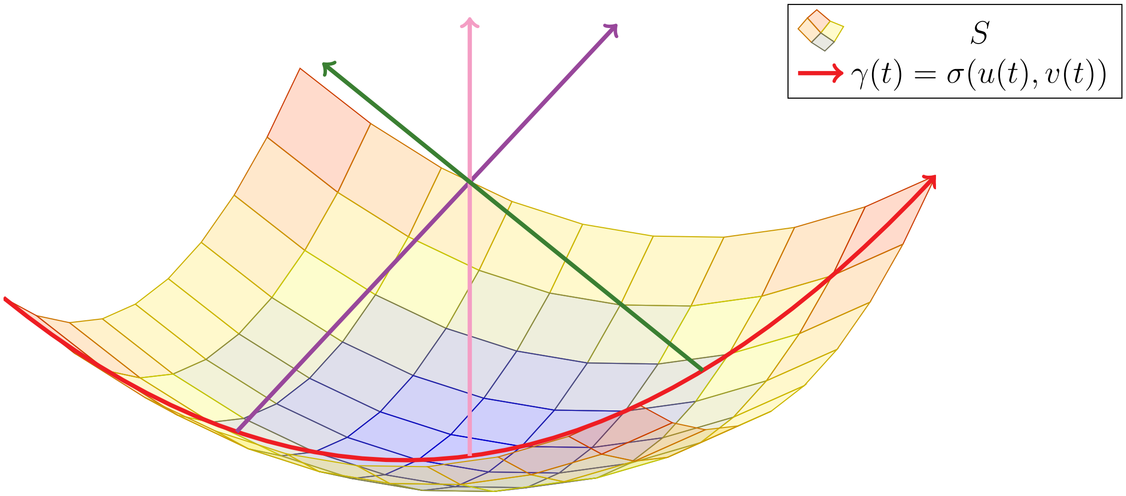


FIGURE 6.7. This diagram demonstrates how the derivative of N_σ is measured. One chooses a curve γ along S (drawn in red), and differentiates N_σ along γ. Some samples of N_σ are drawn as arrows in the diagram.

To capture a direction along S with which to differentiate N_σ, we consider an arbitrary curve

$$\gamma(t) = \sigma(u(t), v(t))$$

in the image of σ. In particular, the values $u(t)$ and $v(t)$ represent the coordinates of $\gamma(t)$, with respect to the parametrisation σ. The derivative of N_σ in the γ-direction is then given by

$$\frac{d}{dt}[N_\sigma(u(t), v(t))]. \tag{6.3}$$

This setup is illustrated in Figure 6.7.

Now, although (6.3) is a vector in $\mathbb{R}^3$, the information that it carries is in fact constrained to S:

Proposition 6.4. *Assuming the above setup, we have that*

$$\frac{d}{dt}[N_\sigma(u(t), v(t))]\bigg|_{\sigma(u(t),v(t))} \in T_\sigma(u(t), v(t)) = T_{\sigma(u(t),v(t))}S, \tag{6.4}$$

that is, $\frac{d}{dt}[N_\sigma(u(t), v(t))]$ *is everywhere tangent to* S*.*

Proof. Since $N_\sigma(u(t), v(t))$ has unit (and constant) length, then by Proposition 3.5,

$$0 = \frac{1}{2}\frac{d}{dt}[|N_\sigma(u(t), v(t))|^2] = N_\sigma(u(t), v(t)) \cdot \frac{d}{dt}[N_\sigma(u(t), v(t))].$$

The above equality implies that $\frac{d}{dt}[N_\sigma(u(t), v(t))]$ has no component in the N_σ-direction, i.e. the direction normal to S. In other words, $\frac{d}{dt}[N_\sigma(u(t), v(t))]$ must be tangent to S and hence must lie in the corresponding tangent plane; this is precisely (6.4). □

Now, Proposition 6.4 establishes that the derivatives of N_σ lie on the tangent planes of S, which are spanned by the tangent vectors $\partial_u\sigma(u(t), v(t))|_{\sigma(u(t),v(t))}$ and $\partial_v\sigma(u(t), v(t))|_{\sigma(u(t),v(t))}$. Thus, to capture all the information in the derivative of N_σ, we need only compute the *dot products*

$$\frac{d}{dt}[N_\sigma(u(t), v(t))] \cdot \partial_u\sigma(u(t), v(t)), \qquad \frac{d}{dt}[N_\sigma(u(t), v(t))] \cdot \partial_v\sigma(u(t), v(t)),$$

that is, *the magnitudes of the* u *and* v*-components of* $\frac{d}{dt}[N_\sigma(u(t), v(t))]$.

Now, since $N_\sigma(u(t), v(t))$ is perpendicular to $\partial_u\sigma(u(t), v(t))$ (the former is normal to S, while the latter is tangent to S), then applying the product and chain rules yields

$$\begin{aligned}\frac{d}{dt}[N_\sigma(u(t), v(t))] \cdot \partial_u\sigma(u(t), v(t)) &= \frac{d}{dt}[N_\sigma(u(t), v(t)) \cdot \partial_u\sigma(u(t), v(t))] \\ &\quad - N_\sigma(u(t), v(t)) \cdot \frac{d}{dt}[\partial_u\sigma(u(t), v(t))] \\ &= -[N_\sigma(u(t), v(t)) \cdot \partial^2_{uu}\sigma(u(t), v(t))]u'(t) \\ &\quad - [N_\sigma(u(t), v(t)) \cdot \partial^2_{uv}\sigma(u(t), v(t))]v'(t).\end{aligned}$$

A similar computation also yields

$$\begin{aligned}\frac{d}{dt}[N_\sigma(u(t), v(t))] \cdot \partial_v\sigma(u(t), v(t)) &= -[N_\sigma(u(t), v(t)) \cdot \partial^2_{vu}\sigma(u(t), v(t))]u'(t) \\ &\quad - [N_\sigma(u(t), v(t)) \cdot \partial^2_{vv}\sigma(u(t), v(t))]v'(t).\end{aligned}$$

The above two equations can be concisely summarised in matrix form:

$$\begin{bmatrix} \frac{d}{dt}[N_\sigma(u(t), v(t))] \cdot \partial_u\sigma(u(t), v(t)) \\ \frac{d}{dt}[N_\sigma(u(t), v(t))] \cdot \partial_v\sigma(u(t), v(t)) \end{bmatrix} = - \begin{bmatrix} \partial_{uu}\sigma \cdot N_\sigma & \partial_{uv}\sigma \cdot N_\sigma \\ \partial_{vu}\sigma \cdot N_\sigma & \partial_{vv}\sigma \cdot N_\sigma \end{bmatrix}\Bigg|_{(u(t),v(t))} \begin{bmatrix} u'(t) \\ v'(t) \end{bmatrix}. \tag{6.5}$$

This above identity (6.5) motivates the following definition:

Definition 6.5. *Consider a surface* $S \subseteq \mathbb{R}^3$*, and a parametrisation* $\sigma : U \to S$ *of* S*. We define the* <u>*second fundamental form*</u> *of* S *with respect to* σ *to be the matrix-valued function*

$$F^{II}_\sigma(u, v) = \begin{bmatrix} \partial_{uu}\sigma(u, v) \cdot N_\sigma(u, v) & \partial_{uv}\sigma(u, v) \cdot N_\sigma(u, v) \\ \partial_{vu}\sigma(u, v) \cdot N_\sigma(u, v) & \partial_{vv}\sigma(u, v) \cdot N_\sigma(u, v) \end{bmatrix}, \qquad (u, v) \in U,$$

where N_σ *is the unit normal to* S *given by*

$$N_\sigma(u, v) = +\frac{\partial_u\sigma(u, v) \times \partial_v\sigma(u, v)}{|\partial_u\sigma(u, v) \times \partial_v\sigma(u, v)|}.$$

Recall that $u'(t)$ and $v'(t)$ in (6.5) represent the components of $\gamma'(t)$, where the curve γ represented the direction in which we differentiated N_σ. Consequently, (6.5) and Definition 6.5 indicate that if we know the second fundamental form F^{II}_σ, then we can recover all components of the derivative of N_σ. Before we accomplish this fully, and hence complete our answer of Question 6.3, we first present some simple examples of Definition 6.5 in action:

Example 6.9. *Consider the cylinder from Examples 5.5 and 5.22,*

$$\mathcal{C} = \{(x, y, z) \in \mathbb{R}^3 \mid x^2 + y^2 = 1\},$$

which is fully covered by the parametrisation

$$\sigma : \mathbb{R}^2 \to \mathbb{R}^3, \qquad \sigma(u, v) = (\cos u, \sin u, v).$$

Recall from Example 5.40 that

$$N_\sigma(u, v) = \frac{\partial_u \sigma(u, v) \times \partial_v \sigma(u, v)}{|\partial_u \sigma(u, v) \times \partial_v \sigma(u, v)|} = (\cos u, \sin u, 0).$$

Furthermore, computing second derivatives of σ yields

$$\begin{aligned} \partial_{uu}\sigma(u, v) &= (-\cos u, -\sin u, 0), \qquad & \partial_{uv}\sigma(u, v) &= (0, 0, 0), \\ \partial_{vu}\sigma(u, v) &= (0, 0, 0), \qquad & \partial_{vv}\sigma(u, v) &= (0, 0, 0). \end{aligned}$$

Thus, taking dot products and applying Definition 6.5, we have for any $(u, v) \in \mathbb{R}^2$ that

$$F^{II}_\sigma(u, v) = \begin{bmatrix} \partial_{uu}\sigma(u, v) \cdot N_\sigma(u, v) & \partial_{uv}\sigma(u, v) \cdot N_\sigma(u, v) \\ \partial_{vu}\sigma(u, v) \cdot N_\sigma(u, v) & \partial_{vv}\sigma(u, v) \cdot N_\sigma(u, v) \end{bmatrix} = \begin{bmatrix} -1 & 0 \\ 0 & 0 \end{bmatrix}.$$

Example 6.10. *Next, consider the sphere $\mathbb{S}^2$, with the spherical coordinate parametrisation*

$$\sigma_s : \mathbb{R} \times (0, \pi) \to \mathbb{R}^3, \qquad \sigma_s(u, v) = (\cos u \sin v, \sin u \sin v, \cos v).$$

Recall from Example 5.39 that

$$N_\sigma(u, v) = \frac{\partial_u \sigma_s(u, v) \times \partial_v \sigma_s(u, v)}{|\partial_u \sigma_s(u, v) \times \partial_v \sigma_s(u, v)|} = -(\cos u \sin v, \sin u \sin v, \cos v).$$

Furthermore, taking second partial derivatives, we see that

$$\begin{aligned} \partial_{uu}\sigma(u, v) &= (-\cos u \sin v, -\sin u \sin v, 0), \\ \partial_{uv}\sigma(u, v) &= (-\sin u \cos v, \cos u \cos v, 0), \\ \partial_{vu}\sigma(u, v) &= (-\sin u \cos v, \cos u \cos v, 0), \\ \partial_{vv}\sigma(u, v) &= (-\cos u \sin v, -\sin u \sin v, -\cos v). \end{aligned}$$

As a result, by Definition 6.5, we obtain

$$F^{II}_\sigma(u, v) = \left. \begin{bmatrix} \partial_{uu}\sigma \cdot N_\sigma & \partial_{uv}\sigma \cdot N_\sigma \\ \partial_{vu}\sigma \cdot N_\sigma & \partial_{vv}\sigma \cdot N_\sigma \end{bmatrix} \right|_{(u,v)} = \begin{bmatrix} \sin^2 v & 0 \\ 0 & 1 \end{bmatrix}.$$

6.2.2. *The Weingarten Matrix.* Let us return to the setting of the preceding discussion, with

$$S \subseteq \mathbb{R}^3, \qquad \sigma : U \to S, \qquad \gamma(t) = \sigma(u(t), v(t))$$

as before. In the preceding discussion, we had derived the identity (6.5):

$$\begin{bmatrix} \frac{d}{dt}[N_\sigma(u(t), v(t))] \cdot \partial_u \sigma(u(t), v(t)) \\ \frac{d}{dt}[N_\sigma(u(t), v(t))] \cdot \partial_v \sigma(u(t), v(t)) \end{bmatrix} = -F^{II}_\sigma(u(t), v(t)) \begin{bmatrix} u'(t) \\ v'(t) \end{bmatrix}.$$

We now use this formula to complete our description of the derivative of N_σ.

Recall that Proposition 6.4 implied that the derivative

$$\frac{d}{dt}[N_\sigma(u(t), v(t))]$$

is everywhere tangent to S. Since the tangent vectors

$$\partial_u \sigma(u(t), v(t))|_{\sigma(u(t),v(t))}, \qquad \partial_v \sigma(u(t), v(t))|_{\sigma(u(t),v(t))}$$

form a basis of $T_{\sigma(u(t),v(t))}S$, then we can write the above as a linear combination,

$$\begin{aligned} \frac{d}{dt}[N_\sigma(u(t), v(t))]|_{\sigma(u(t),v(t))} &= a(t) \cdot \partial_u \sigma(u(t), v(t))|_{\sigma(u(t),v(t))} \\ &\quad + b(t) \cdot \partial_v \sigma(u(t), v(t))|_{\sigma(u(t),v(t))}. \end{aligned} \tag{6.6}$$

In particular, $a(t)$ and $b(t)$ represent the *components* of the left-hand side of (6.6), with respect to the above basis of $T_{\sigma(u(t),v(t))}S$. Our aim now is to find $a(t)$ and $b(t)$.

Taking dot products of (6.6) with the partial derivatives of σ, we obtain

$$\begin{aligned} \frac{d}{dt}[N_\sigma(u(t), v(t))] \cdot \partial_u \sigma(u(t), v(t)) &= a(t)[\partial_u \sigma(u(t), v(t)) \cdot \partial_u \sigma(u(t), v(t))] \\ &\quad + b(t)[\partial_v \sigma(u(t), v(t)) \cdot \partial_u \sigma(u(t), v(t))], \\ \frac{d}{dt}[N_\sigma(u(t), v(t))] \cdot \partial_v \sigma(u(t), v(t)) &= a(t)[\partial_u \sigma(u(t), v(t)) \cdot \partial_v \sigma(u(t), v(t))] \\ &\quad + b(t)[\partial_v \sigma(u(t), v(t)) \cdot \partial_v \sigma(u(t), v(t))]. \end{aligned}$$

Compacting into matrix form, the above now becomes

$$\begin{aligned} \begin{bmatrix} \frac{d}{dt}[N_\sigma(u(t), v(t))] \cdot \partial_u \sigma(u(t), v(t)) \\ \frac{d}{dt}[N_\sigma(u(t), v(t))] \cdot \partial_v \sigma(u(t), v(t)) \end{bmatrix} &= \left. \begin{bmatrix} \partial_u \sigma \cdot \partial_u \sigma & \partial_u \sigma \cdot \partial_v \sigma \\ \partial_v \sigma \cdot \partial_u \sigma & \partial_v \sigma \cdot \partial_v \sigma \end{bmatrix} \right|_{(u(t),v(t))} \begin{bmatrix} a(t) \\ b(t) \end{bmatrix} \\ &= F^I_\sigma(u(t), v(t)) \begin{bmatrix} a(t) \\ b(t) \end{bmatrix}. \end{aligned} \tag{6.7}$$

Finally, combining (6.5) and (6.7) results in the identity

$$F^I_\sigma(u(t), v(t)) \begin{bmatrix} a(t) \\ b(t) \end{bmatrix} = -F^{II}_\sigma(u(t), v(t)) \begin{bmatrix} u'(t) \\ v'(t) \end{bmatrix},$$

and rearranging the above yields

$$\begin{bmatrix} a(t) \\ b(t) \end{bmatrix} = -F^I_\sigma(u(t), v(t))^{-1} F^{II}_\sigma(u(t), v(t)) \begin{bmatrix} u'(t) \\ v'(t) \end{bmatrix}. \tag{6.8}$$

Let us now interpret (6.8). First, on the right-hand side, the quantities $u'(t)$ and $v'(t)$ represent the components of $\gamma'(t)$, with respect to the σ-coordinates. Thus, $u'(t)$ and $v'(t)$ represent the direction in which we are differentiating N_σ. Now, on the left-hand side, the quantities $a(t)$ and $b(t)$ represent the components of the derivative of N_σ, in this γ-direction.

Consequently, *the matrix* $-F^I_\sigma(u(t), v(t))^{-1} F^{II}_\sigma(u(t), v(t))$ *maps the direction in which we choose to differentiate* N_σ *with the value of the resulting derivative.* In particular, this matrix contains all the information about the derivative of the unit normals of S (and hence, the curvature of S).

These considerations motivate the following definition:

Definition 6.6. *Consider a surface $S \subseteq \mathbb{R}^3$, and a parametrisation $\sigma : U \to S$ of S. We define the Weingarten matrix of S with respect to σ to be the matrix-valued function*

$$W_\sigma(u,v) = F^I_\sigma(u,v)^{-1} F^{II}_\sigma(u,v), \qquad (u,v) \in U.$$

We now compute the Weingarten matrix of some basic surfaces and parametrisations:

Example 6.11. *Let us return to the cylinder $\mathcal{C}$ from Example 6.9, with parametrisation*

$$\sigma : \mathbb{R}^2 \to \mathbb{R}^3, \qquad \sigma(u,v) = (\cos u, \sin u, v).$$

Recall from Examples 6.1 and 6.9 that

$$F^I_\sigma(u,v) = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \qquad F^{II}_\sigma(u,v) = \begin{bmatrix} -1 & 0 \\ 0 & 0 \end{bmatrix}.$$

Then, by Definition 6.6, the Weingarten matrix with respect to σ is

$$W_\sigma(u,v) = F^I_\sigma(u,v)^{-1} F^{II}_\sigma(u,v) = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} -1 & 0 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} -1 & 0 \\ 0 & 0 \end{bmatrix}.$$

Example 6.12. *Next, for the sphere $\mathbb{S}^2$, with parametrisation*

$$\sigma_s : \mathbb{R} \times (0,\pi) \to \mathbb{R}^3, \qquad \sigma_s(u,v) = (\cos u \sin v, \sin u \sin v, \cos v).$$

we have from Examples 6.2 and 6.10 that

$$F^I_{\sigma_s}(u,v) = \begin{bmatrix} \sin^2 v & 0 \\ 0 & 1 \end{bmatrix}, \qquad F^{II}_{\sigma_s}(u,v) = \begin{bmatrix} \sin^2 v & 0 \\ 0 & 1 \end{bmatrix}.$$

Thus, by Definition 6.6, we have that

$$W_{\sigma_s}(u,v) = \begin{bmatrix} \sin^2 v & 0 \\ 0 & 1 \end{bmatrix}^{-1} \begin{bmatrix} \sin^2 v & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

Example 6.13. *Consider now the paraboloid described as the image of the parametric surface*

$$\sigma : \mathbb{R}^2 \to \mathbb{R}^3, \qquad \sigma(u,v) = (u, v, u^2 + v^2).$$

Recall from Example 6.3 that its first fundamental form with respect to σ is

$$F^I_\sigma(u,v) = \begin{bmatrix} 1+4u^2 & 4uv \\ 4uv & 1+4v^2 \end{bmatrix}.$$

Moreover, the inverse of $F^I_\sigma(u,v)$ is given by

$$F^I_\sigma(u,v)^{-1} = \frac{1}{\det F^I_\sigma(u,v)} \begin{bmatrix} 1+4v^2 & -4uv \\ -4uv & 1+4u^2 \end{bmatrix} = \frac{1}{1+4u^2+4v^2} \begin{bmatrix} 1+4v^2 & -4uv \\ -4uv & 1+4u^2 \end{bmatrix}.$$

Next, to compute the second fundamental form, we first note that

$$N_\sigma(u,v) = \frac{\partial_u \sigma(u,v) \times \partial_v(u,v)}{|\partial_u \sigma(u,v) \times \partial_v \sigma(u,v)|} = \frac{1}{\sqrt{1+4u^2+4v^2}}(-2u, -2v, 1),$$

while the second derivatives of σ *satisfy*

$$\partial_{uu}\sigma(u,v) = (0,0,2), \qquad \partial_{uv}\sigma(u,v) = (0,0,0),$$
$$\partial_{vu}\sigma(u,v) = (0,0,0), \qquad \partial_{vv}\sigma(u,v) = (0,0,2).$$

Thus, by Definition 6.5,

$$\mathrm{F}^{\mathrm{II}}_\sigma(u,v) = \frac{1}{\sqrt{1+4u^2+4v^2}} \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}.$$

Finally, combining all the above with Definition 6.6, we obtain

$$\begin{aligned} W_\sigma(u,v) &= \frac{1}{1+4u^2+4v^2} \begin{bmatrix} 1+4v^2 & -4uv \\ -4uv & 1+4u^2 \end{bmatrix} \cdot \frac{1}{\sqrt{1+4u^2+4v^2}} \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix} \\ &= \frac{2}{(1+4u^2+4v^2)^{\frac{3}{2}}} \begin{bmatrix} 1+4v^2 & -4uv \\ -4uv & 1+4u^2 \end{bmatrix}. \end{aligned}$$

Finally, let us make sense of the Weingarten matrix, which we recall mapped the direction in which we differentiated N_σ to the resulting derivative of N_σ, from a more conceptual point of view. Consider the following definition, which we will only treat here in an informal manner:

Definition 6.7. *Given a surface* $S \subseteq \mathbb{R}^3$ *and* $\mathbf{p} \in S$*, we define the* <u>*shape operator*</u> *of* S *at* $\mathbf{p}$ *to be the linear transformation* $\mathcal{S} : T_\mathbf{p}S \to T_\mathbf{p}S$*, which maps a direction* $z|_\mathbf{p} \in T_\mathbf{p}S$ *at* $\mathbf{p}$ *along* S *to the derivative of the unit normal in the* z*-direction at* $\mathbf{p}$ *(which lies in* $T_\mathbf{p}S$ *by Proposition 6.4).*

Remark. That the shape operator is linear is a basic fact about directional derivatives from calculus and analysis. However, we will not discuss this formally here.

Now, by choosing a parametrisation σ of S that contains $\mathbf{p}$, we in effect fixed a basis of $T_\mathbf{p}S$. Then, the above discussions, in particular (6.8), show that $-W_\sigma$ is precisely the matrix representation of the shape operator $\mathcal{S}$ from Definition 6.7, with respect to this basis of $T_\mathbf{p}S$.

6.2.3. *Principal Curvatures.* The preceding discussion consolidated all the information contained in the derivative of the unit normal into the *Weingarten matrix* of Definition 6.6. Here, we extract additional scalar quantities from the Weingarten matrix, representing aspects of curvature.

First, we recall the following definitions from linear algebra:

Definition 6.8. *Let* A *be a* 2×2 *real-valued matrix. Then,* $\lambda \in \mathbb{R}$ *is an* <u>*eigenvalue*</u> *of* A *iff there exists a nonzero* $\mathbf{v} \in \mathbb{R}^2$ *such that, when represented as a column vector,*

$$A\mathbf{v} = \lambda\mathbf{v}.$$

Furthermore, we refer to $\mathbf{v}$ *as a corresponding* <u>*eigenvector*</u> *of* A*.*

In particular, we consider eigenvalues and eigenvectors of the Weingarten matrix:

Definition 6.9. *Assume the setting of Definition 6.6, and fix in addition* $\mathbf{p} = \sigma(u,v) \in S$*.*

- *The (two) eigenvalues of* $W_\sigma(u,v)$ *are called* <u>*principal curvatures*</u> *of* S *at* $\mathbf{p}$ *(with respect to* σ*). These will be denoted in the notes as* $\kappa_1|_\mathbf{p}$ *and* $\kappa_2|_\mathbf{p}$*.*

- *Eigenvectors of $W_\sigma(u,v)$ are called principal directions of S at $\mathbf{p}$ (with respect to σ).*

Example 6.14. *Consider the cylinder $\mathcal{C}$ from Example 6.11, with parametrisation*

$$\sigma : \mathbb{R}^2 \to \mathbb{R}^3, \qquad \sigma(u,v) = (\cos u, \sin u, v).$$

Recall also from Example 6.11 that the Weingarten matrix with respect to σ is

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

Since $W_\sigma(u,v)$ is diagonal, its eigenvalues are simply its diagonal elements: -1 and 0. Thus, by Definition 6.9, the principal curvatures of $\mathcal{C}$ (with respect to σ) at any $\mathbf{p} \in \mathcal{C}$ are

$$\kappa_1|_{\mathbf{p}} = -1, \qquad \kappa_2|_{\mathbf{p}} = 0.$$

Furthermore, note that the column vectors

$$\mathbf{v}_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \qquad \mathbf{v}_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

satisfy the relations

$$W_\sigma(u,v)\mathbf{v}_1 = \begin{bmatrix} -1 \\ 0 \end{bmatrix} = -1 \cdot \mathbf{v}_1, \qquad W_\sigma(u,v)\mathbf{v}_2 = \begin{bmatrix} 0 \\ 0 \end{bmatrix} = 0 \cdot \mathbf{v}_2.$$

Thus, $\mathbf{v}_1$, $\mathbf{v}_2$ represent principal directions of $\mathcal{C}$ (with respect to σ) at any $\mathbf{p} \in \mathcal{C}$.

Let us now explore what the principal curvatures and directions mean. Throughout, we will assume the setting of Definition 6.9. In other words, we let $S \subseteq \mathbb{R}^3$ be a surface, and we let $\sigma : U \to S$ be a parametrisation of S. In addition, we fix a point $\mathbf{p} = \sigma(u,v) \in S$.

Given any $\mathbf{v} = (v_1, v_2) \in \mathbb{R}^2$, we can think of it as a direction along S at $\mathbf{p}$:

$$\mathbf{v} \mapsto v_1 \cdot \partial_u \sigma(u,v)|_{\mathbf{p}} + v_2 \cdot \partial_v \sigma(u,v)|_{\mathbf{p}} = T_{\mathbf{p}} S.$$

Moreover, let $\mathbf{w} = (w_1, w_2) \in \mathbb{R}^2$ be such that

$$\mathbf{w} = W_\sigma(u,v)\mathbf{v},$$

where $\mathbf{v}$ and $\mathbf{w}$ are treated as column vectors. Then, $\mathbf{w}$ can also be viewed as a direction along S at $\mathbf{p}$, representing the derivative $-d_{\mathbf{v}} N_\sigma$ of the unit normal N_σ in the $(-\mathbf{v})$-direction at $\mathbf{p}$.

In general, $\mathbf{w} = -d_{\mathbf{v}} N_\sigma$ will not point in the same direction as $\mathbf{v}$. However, the exceptions to this are precisely along the principal directions, i.e. the eigenvectors of $W_\sigma(u,v)$. In particular, if

$$W_\sigma(u,v)\mathbf{v} = \mathbf{w} = \kappa \mathbf{v},$$

that is, κ is a principal curvature of S at $\mathbf{p}$ and $\mathbf{v}$ a corresponding principal direction, then:

- When $\kappa > 0$, then along the $\mathbf{v}$-direction, the unit normal N_σ is precisely bending away from this $\mathbf{v}$-direction. Thus, S is bending toward N_σ in the $\mathbf{v}$-direction from $\mathbf{p}$.
- When $\kappa < 0$, then along the $\mathbf{v}$-direction, the unit normal N_σ is precisely bending toward this $\mathbf{v}$-direction. Thus, S is bending away from N_σ in the $\mathbf{v}$-direction from $\mathbf{p}$.

- When $\kappa = 0$, then S is not bending at $\mathbf{p}$ along the $\mathbf{v}$-direction.

See Figure 6.8 below for a visual demonstration of each case.

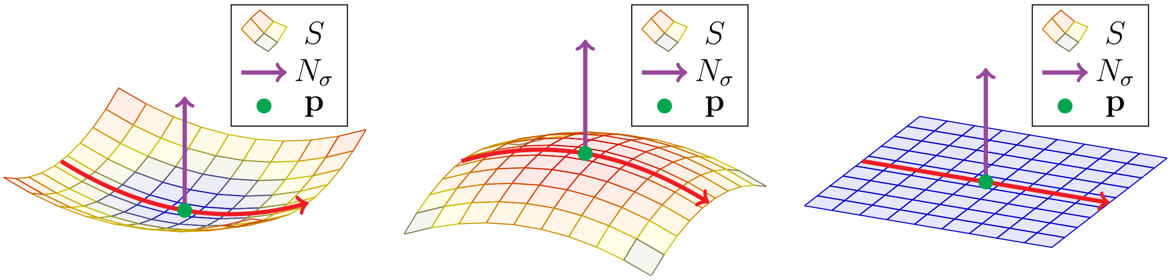


FIGURE 6.8. The plots show parts of three surfaces S. On each plot, the purple arrow denotes a unit normal $N_\sigma(u, v)$ on the surface, while the red curve indicates a principal direction along S through $\mathbf{p} = \sigma(u, v)$. The left, middle, and right plots demonstrate cases in which the corresponding (σ-)principal curvature is positive, negative, and zero, respectively.

Let us now the consider some basic examples of surfaces:

Example 6.15. *Let us return to the setting from Example 6.14. Recall from Example 6.9 that*

$$N_\sigma(u, v) = (\cos u, \sin u, 0),$$

which always points outward from $\mathcal{C}$. Using the notations from Example 6.14:

- *The principal direction $\mathbf{v}_1$ corresponds to the direction $\partial_u \sigma(u, v)|_{\mathbf{p}}$, that is, along the circles within $\mathcal{C}$ where the z-coordinate is constant. Along these curves, one is bending away from the unit normal N_σ. This direction corresponds to the negative principal curvature $\kappa_1 = -1$.*
- *The principal direction $\mathbf{v}_2$ corresponds to the direction $\partial_v \sigma(u, v)|_{\mathbf{p}}$, along the vertical lines of $\mathcal{C}$. Along these lines, $\mathcal{C}$ is not bending, corresponding to the principal curvature $\kappa_2 = 0$.*

See the left plot of Figure 6.9 for a graphical representation of the above.

Example 6.16. *For the sphere $\mathbb{S}^2$, with parametrisation*

$$\sigma_s : \mathbb{R} \times (0, \pi) \to \mathbb{R}^3, \qquad \sigma_s(u, v) = (\cos u \sin v, \sin u \sin v, \cos v).$$

we recall from Examples 6.10 and 6.12 that

$$N_{\sigma_s}(u, v) = -(\cos u \sin v, \sin u \sin v, \cos v), \qquad W_{\sigma_s}(u, v) = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

In particular, note that the unit normal N_{σ_s} always points inward toward the origin.

Again, for any point $\mathbf{p} = \sigma_s(u, v) \in \mathbb{S}^2$, since $W_{\sigma_s}(u, v)$ is diagonal, then its eigenvalues, and hence the principal curvatures at $\mathbf{p}$, are simply its diagonal elements:

$$\kappa_1|_{\mathbf{p}} = \kappa_2|_{\mathbf{p}} = 1.$$

Furthermore, since $W_{\sigma_s}(u, v)$ is the identity matrix, it follows that any nonzero $\mathbf{v} \in \mathbb{R}^2$ is a principal direction at $\mathbf{p}$, corresponding to the above principal curvatures.

More intuitively, note that in any direction along $\mathbb{S}^2$ *from* $\mathbf{p}$*, the sphere is bending inward toward (the inward-pointing)* N_σ*. This reflects the fact that both principal curvatures are positive. See the right plot of Figure 6.9 for a graphical representation.*

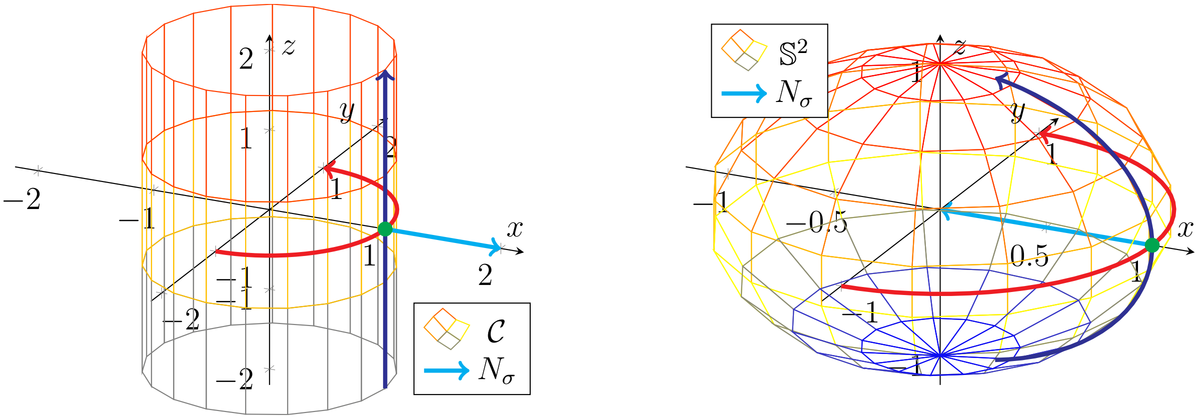


FIGURE 6.9. The left plot contains the cylinder $\mathcal{C}$ from Example 6.15; the red curve represents the principal direction associated with $\kappa_1 = -1$, while the blue curve represents the principal direction from $\kappa_2 = 0$. The right plot contains the sphere $\mathbb{S}^2$ from Example 6.16; note that all directions bend toward the inward-pointing unit normal (in cyan).

Example 6.17. *Next, consider the* <u>*one-sheeted hyperboloid*</u>*,*

$$\mathcal{H} = \{(x, y, z) \in \mathbb{R}^3 \mid x^2 + y^2 - z^2 = 1\},$$

a surface which can be covered by a single parametrisation

$$\sigma : \mathbb{R}^2 \to \mathbb{R}^3, \qquad \sigma(u, v) = (\cosh v \cos u, \cosh v \sin u, \sinh v).$$

A graph of σ *is given in Figure 6.10.*

With respect to σ*, one can then compute that*

$$N_\sigma = \frac{1}{\sqrt{\cosh^2 v + \sinh^2 v}} (\cosh v \cos u, \cosh v \sin u, -\sinh v),$$

which always points outward from $\mathcal{H}$*. Furthermore, another computation yields*

$$\begin{aligned} W_\sigma(u, v) &= \begin{bmatrix} -(\cosh^2 v + \sinh^2 v)^{-\frac{1}{2}} & 0 \\ 0 & (\cosh^2 v + \sinh^2 v)^{-\frac{3}{2}} \end{bmatrix} \\ &= \begin{bmatrix} -(1 + 2\sinh^2 v)^{-\frac{1}{2}} & 0 \\ 0 & (1 + 2\sinh^2 v)^{-\frac{3}{2}} \end{bmatrix} . \end{aligned}$$

Thus, at any $\mathbf{p} = \sigma(u, v) \in \mathcal{H}$*, the principal curvatures are given by*

$$\kappa_1|_{\mathbf{p}} = -(1 + 2\sin^2 v)^{-\frac{1}{2}}, \qquad \kappa_2|_{\mathbf{p}} = (1 + 2\sin^2 v)^{-\frac{3}{2}}.$$

In particular, at each point of $\mathcal{H}$*, there is one positive and one negative principal curvature.*

To see this more intuitively, note that:

- *In the "vertical" direction, the hyperboloid is bending toward the outward-pointing unit normal N_σ, which captures the positive principal curvature κ_2.*
- *In the "horizontal" direction, the hyperboloid is bending away from the outward-pointing unit normal N_σ, reflecting the negative principal curvature κ_1.*

This is demonstrated graphically in Figure 6.10.

Observe the principal curvatures were always obtained using the Weingarten matrix, *with respect to a chosen parametrisation* σ. In order for these quantities to be true geometric properties of the surface S, they must also be independent of the chosen σ. The following theorem shows that this is "almost true", with the only deficiency being an a sign ambiguity:

Theorem 6.5. *Let $S \subseteq \mathbb{R}^3$ be a surface, and let $\mathbf{p} \in S$. Then, the principal curvatures $\kappa_1|_\mathbf{p}$ and $\kappa_2|_\mathbf{p}$ at $\mathbf{p}$ are independent of the parametrisation σ of S used to define them, with the exception of a possible change in sign of both $\kappa_1|_\mathbf{p}$ and $\kappa_2|_\mathbf{p}$.*

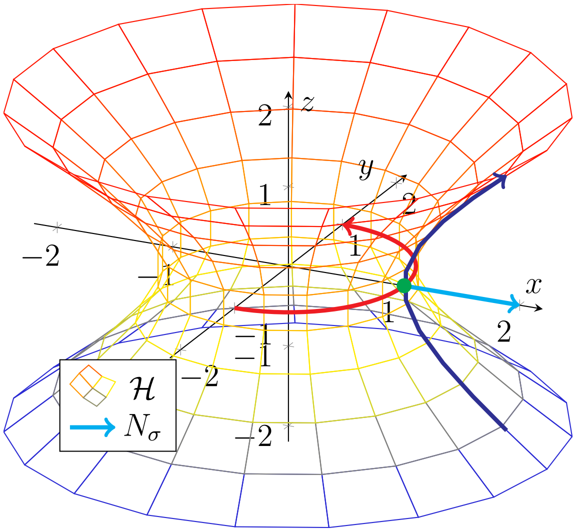


Figure 6.10. The hyperboloid $\mathcal{H}$ discussed in Example 6.17. The red and blue curves represent the principal directions associated with $\kappa_1 < 0$ and $\kappa_2 > 0$, respectively.

Rather than proving Theorem 6.5 formally, which requires a bit of background that is beyond the scope of this module, let us instead give an informal explanation.

The first main idea comes from the discussion below Definition 6.7, i.e. that W_σ is the matrix associated with the *shape operator*, a linear transformation of the tangent space $T_\mathbf{p}S$, with respect to the basis of $T_\mathbf{p}S$ obtained from σ. A change of parametrisation, say from σ to $\tilde{\sigma}$, results in a change of basis of $T_\mathbf{p}S$. Therefore, the matrices $W_\sigma(u, v)$ and $W_{\tilde{\sigma}}(\tilde{u}, \tilde{v})$, where $\sigma(u, v) = \mathbf{p} = \tilde{\sigma}(\tilde{u}, \tilde{v})$, are related through this same change of basis. Now, recall from linear algebra that the eigenvalues of a matrix are preserved by such changes of bases. Thus, we expect that the principal curvatures of S at $\mathbf{p}$ should not change.

To see where the sign ambiguity comes from, we recall that for all our computations and intuitions, we fixed the unit normal N_σ generated from our parametrisation σ. It is possible that a different parametrisation $\tilde{\sigma}$ would select the opposite normal. In this case, the second fundamental form, the Weingarten matrix, and hence the principal curvatures would all have the opposite sign, corresponding to this flip in sign of the unit normal.

A more intuitive way to see this change in sign is the following: *if a surface is bending toward one unit normal, then it is bending away from the opposite unit normal, and vice versa.* Thus, the signs of the principal curvatures depend crucially on which of the two unit normals to S is selected.

Example 6.18. *If we choose a different parametrsiation of $\mathbb{S}^2$, say,*

$$\sigma_* : (0, \pi) \times \mathbb{R} \to \mathbb{R}^3, \qquad \sigma_*(u, v) = \sigma_s(v, u) = (\cos v \sin u, \sin v \sin u, \cos u),$$

then the corresponding unit normals $N_{\sigma_*}(u,v)$ *will point outward. Moreover, we can compute*

$$W_{\sigma_*}(u,v) = \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix}, \qquad \kappa_1|_{\mathbf{p}} = \kappa_2|_{\mathbf{p}} = -1,$$

for any point $\mathbf{p} = \sigma_*(u,v)$. *Intuitively speaking, since we have now selected the outward-pointing unit normal to* $\mathbb{S}^2$, *the sphere now bends in every direction away from this unit normal. This explains the negative principal curvatures that were obtained with respect to* σ_*.

Remark. Finally, one can also view the principal curvatures $\kappa_1|_{\mathbf{p}}$ and $\kappa_2|_{\mathbf{p}}$ as the "maximum" and "minimum" amounts of curvature at $\mathbf{p}$. Moreover, the corresponding principal directions are those along which the "maximum" and "minimum" curvatures are achieved. For brevity, we omit describing in these notes the precise sense that the above statements hold.

6.2.4. *Mean and Gauss Curvatures.* Next, we consider some additional geometric quantities that can be defined from the principal curvatures of Definition 6.9.

Definition 6.10. *Assume the setting of Definition 6.9. In particular, let* $\mathbf{p} = \sigma(u,v) \in S$, *and let* $\kappa_1|_{\mathbf{p}}$ *and* $\kappa_2|_{\mathbf{p}}$ *be the principal curvatures of* S *at* $\mathbf{p}$, *with respect to* σ.

- *The mean curvature of* S *at* $\mathbf{p}$ *(with respect to* σ*) is defined*
$$H|_{\mathbf{p}} = \frac{1}{2}(\kappa_1|_{\mathbf{p}} + \kappa_2|_{\mathbf{p}}).$$
- *The Gauss curvature of* S *at* $\mathbf{p}$ *is defined*
$$\mathcal{K}|_{\mathbf{p}} = \kappa_1|_{\mathbf{p}} \cdot \kappa_2|_{\mathbf{p}}.$$

We first apply Definition 6.10 to some elementary examples:

Example 6.19. *Consider the cylinder* $\mathcal{C}$ *from Example 6.14, with parametrisation*

$$\sigma : \mathbb{R}^2 \to \mathbb{R}^3, \qquad \sigma(u,v) = (\cos u, \sin u, v).$$

Using the computations from Example 6.14, we obtain for any $\mathbf{p} \in \mathcal{C}$ *that*

$$\begin{aligned} H|_{\mathbf{p}} &= \frac{1}{2}(\kappa_1|_{\mathbf{p}} + \kappa_2|_{\mathbf{p}}) = \frac{1}{2}(-1+0) = -\frac{1}{2}, \\ \mathcal{K}|_{\mathbf{p}} &= \kappa_1|_{\mathbf{p}} \cdot \kappa_2|_{\mathbf{p}} = -1 \cdot 0 = 0. \end{aligned}$$

(Here, the mean curvature is defined with respect to σ.*)*

Example 6.20. *For the sphere* $\mathbb{S}^2$, *with parametrisation*

$$\sigma_s : \mathbb{R} \times (0,\pi) \to \mathbb{R}^3, \qquad \sigma_s(u,v) = (\cos u \sin v, \sin u \sin v, \cos v),$$

we can apply the computations from Example 6.16 to obtain, for any $\mathbf{p} = \sigma_s(u,v) \in \mathbb{S}^2$,

$$H|_{\mathbf{p}} = \frac{1}{2}(\kappa_1|_{\mathbf{p}} + \kappa_2|_{\mathbf{p}}) = 1, \qquad \mathcal{K}|_{\mathbf{p}} = \kappa_1|_{\mathbf{p}} \cdot \kappa_2|_{\mathbf{p}} = 1.$$

Example 6.21. *Next, consider the hyperboloid* $\mathcal{H}$ *from Example 6.17, with parametrisation*

$$\sigma : \mathbb{R}^2 \to \mathbb{R}^3, \qquad \sigma(u,v) = (\cosh v \cos u, \cosh v \sin u, \sinh v).$$

From the computations in Example 6.17, we obtain for any $\mathbf{p} = \sigma(u, v) \in \mathcal{H}$ *that*

$$H|_{\mathbf{p}} = \frac{1}{2}[-(1 + 2\sinh^2 v)^{-\frac{1}{2}} + (1 + 2\sinh^2 v)^{-\frac{3}{2}}] = -\frac{\sinh^2 v}{(1 + 2\sinh^2 v)^{-\frac{3}{2}}},$$

$$\mathcal{K}|_{\mathbf{p}} = -(1 + 2\sinh^2 v)^{-\frac{1}{2}}(1 + 2\sinh^2 v)^{-\frac{3}{2}} = -(1 + 2\sinh^2 v)^{-2}.$$

Now, since Theorem 6.5 implies that the principal curvatures are independent of parametrisation except for a sign ambiguity, the same then holds for the mean curvature. On the other hand, observe that when the principal curvatures $\kappa_1|_{\mathbf{p}}$ and $\kappa_2|_{\mathbf{p}}$ are replaced by $-\kappa_1|_{\mathbf{p}}$ and $-\kappa_2|_{\mathbf{p}}$, the Gauss curvature—i.e. the product of the principal curvatures—does not change. As a result, the Gauss curvature is fully independent of the chosen parametrisation.

These statements are more precisely summarised in the following theorem:

Theorem 6.6. *Let* $S \subseteq \mathbb{R}^3$ *be a surface, and let* $\mathbf{p} \in S$.

- *The mean curvature* $H|_{\mathbf{p}}$ *is a geometric property of* S, *except for a sign ambiguity.*
- *The Gauss curvature* $\mathcal{K}|_{\mathbf{p}}$ *is a geometric property of* S.

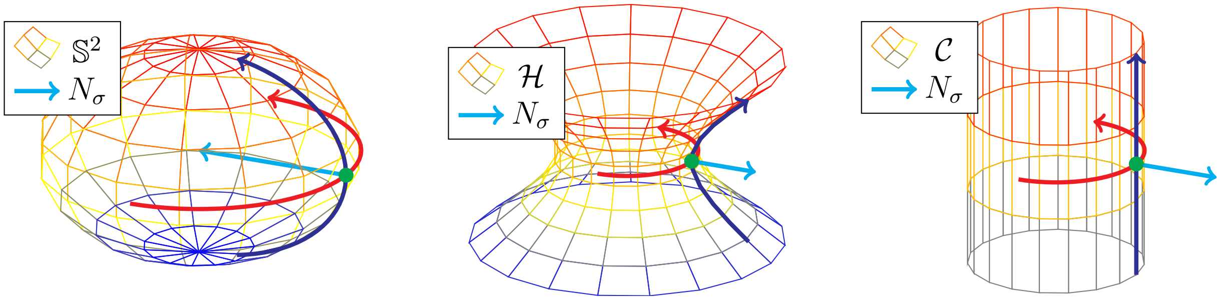


FIGURE 6.11. The left, middle, and right plots show surfaces with positive, negative, and zero Gauss curvature, respectively, at the point indicated in green. The red and blue curves indicate principal directions; these bend in the same way in the left plot, but in opposite ways in the right plot.

Let us look a bit closer at how the sign of the Gauss curvature can be interpreted. Suppose first that $\mathcal{K}|_{\mathbf{p}} > 0$. Then, the corresponding principal curvatures $\kappa_1|_{\mathbf{p}}$ and $\kappa_2|_{\mathbf{p}}$ (with respect to any parametrisation σ covering $\mathbf{p}$) must have the same sign. As a result, S must bend in the same way along any direction from $\mathbf{p}$—either toward or away from the chosen unit normal at $\mathbf{p}$.

One example of this is the sphere $\mathbb{S}^2$ (see Examples 6.16 and 6.20). Indeed, one can observe that $\mathbb{S}^2$ is, along any direction, always bending toward the inward-pointing unit normal.

Remark. Note that whether the surface bends toward or away from the chosen unit normal is irrelevant, as this can be reversed by choosing the opposite unit normal.

On the other hand, if $\mathcal{K}|_{\mathbf{p}} < 0$, then the principal curvatures $\kappa_1|_{\mathbf{p}}$ and $\kappa_2|_{\mathbf{p}}$ must have opposite signs. This implies that S must bend one way along some directions, and the other way along other directions. An example of this is the hyperboloid $\mathcal{H}$ from Examples 6.17 and 6.21. Here, one bends toward the outward-pointing unit normal in the "vertical" directions, and away from the outward-pointing unit normal in the "horizontal" directions.

Finally, when $\mathcal{K}|_{\mathbf{p}} = 0$, then one of the principal curvatures must vanish. In this case, there must be some (principal) direction along which S is not bending at $\mathbf{p}$. One simple case of this is the cylinder $\mathcal{C}$ from Examples 6.14 and 6.19; indeed, recall that $\mathcal{C}$ does not bend along the z-direction.

All three cases described above are depicted in Figure 6.11.

Yet another application of the mean and Gauss curvatures is that they can often be easier to compute than the principal curvatures. In the examples we have considered so far, the Weingarten matrices have been diagonal, so the principal curvatures could be easily obtained by simply reading off the diagonal elements. On the other hand, if $\mathcal{W}_\sigma(u, v)$ fails to be diagonal, then one would need to go through the process of diagonalising $\mathcal{W}_\sigma(u, v)$, as one learned in linear algebra.

In contrast, the following theorem shows that one can employ some linear algebraic "tricks" in order to compute the mean and Gauss curvatures without first diagonalising $\mathcal{W}_\sigma(u, v)$.

Definition 6.11. *Given a 2×2 real-valued matrix,*

$$A = \begin{bmatrix} a & b \\ c & d \end{bmatrix} ,$$

we define the trace of A to be the sum of its diagonal elements:

$$\operatorname{tr} A = a + d.$$

Theorem 6.7. *Let $S \subseteq \mathbb{R}^3$ be a surface, and let $\sigma : U \to S$ be a parametrisation of S. In addition, we fix $\mathbf{p} = \sigma(u, v) \in S$. Then, the mean (with respect to σ) and Gauss curvatures of S at $\mathbf{p}$ satisfy*

$$H|_{\mathbf{p}} = \frac{1}{2} \operatorname{tr} \mathcal{W}_\sigma(u, v), \qquad \mathcal{K}|_{\mathbf{p}} = \det \mathcal{W}_\sigma(u, v).$$

Proof. By diagonalising $\mathcal{W}_\sigma(u, v)$ (that is, by changing to a basis of eigenvectors of $\mathcal{W}_\sigma(u, v)$), then $\mathcal{W}_\sigma(u, v)$ correspondingly transforms to the diagonal matrix

$$D = \begin{bmatrix} \kappa_1|_{\mathbf{p}} & 0 \\ 0 & \kappa_2|_{\mathbf{p}} \end{bmatrix} .$$

Furthermore, note that

$$H|_{\mathbf{p}} = \frac{1}{2} \operatorname{tr} D, \qquad \mathcal{K}|_{\mathbf{p}} = \det D.$$

Now, standard linear algebra results state that both the trace and the determinant of a matrix remain unchanged after a change of basis transformation. Consequently, $\mathcal{W}_\sigma(u, v)$ and D have the same trace and determinant, and our desired formulas follow. □

Furthermore, one can use the mean and Gauss curvatures to recover the principal curvatures, circumventing the matrix diagonalisation process altogether:

Theorem 6.8. *Assume the setting of Theorem 6.7. Then,*

$$\kappa_1|_{\mathbf{p}}, \kappa_2|_{\mathbf{p}} = H|_{\mathbf{p}} \pm \sqrt{(H|_{\mathbf{p}})^2 - \mathcal{K}|_{\mathbf{p}}}.$$

where the principal and mean curvatures at $\mathbf{p}$ are defined with respect to σ.

Proof. This is proved via some clever algebraic tricks. First, by Definition 6.10, we have that

$$2H|_{\mathbf{p}} = \kappa_1|_{\mathbf{p}} + \kappa_2|_{\mathbf{p}}, \qquad \mathcal{K}|_{\mathbf{p}} = \kappa_1|_{\mathbf{p}} \cdot \kappa_2|_{\mathbf{p}}.$$

From this, we can compute

$$(\kappa_2|_{\mathbf{p}} - \kappa_1|_{\mathbf{p}})^2 = (\kappa_2|_{\mathbf{p}} + \kappa_1|_{\mathbf{p}})^2 - 4\kappa_1|_{\mathbf{p}} \cdot \kappa_2|_{\mathbf{p}} = 4(H|_{\mathbf{p}})^2 - 4\mathcal{K}|_{\mathbf{p}}.$$

As a result,

$$\begin{aligned}\kappa_1|_{\mathbf{p}}, \kappa_2|_{\mathbf{p}} &= \frac{1}{2}(\kappa_2|_{\mathbf{p}} + \kappa_1|_{\mathbf{p}}) \pm \frac{1}{2}(\kappa_2|_{\mathbf{p}} - \kappa_1|_{\mathbf{p}}) \\ &= H|_{\mathbf{p}} \pm \frac{1}{2}\sqrt{4(H|_{\mathbf{p}})^2 - 4\mathcal{K}|_{\mathbf{p}}} \\ &= H|_{\mathbf{p}} \pm \sqrt{(H|_{\mathbf{p}})^2 - \mathcal{K}|_{\mathbf{p}}}.\end{aligned}$$

□

Example 6.22. *Consider the paraboloid $\mathcal{P}$ described as the image of the parametric surface*

$$\sigma : \mathbb{R}^2 \to \mathbb{R}^3, \qquad \sigma(u, v) = (u, v, u^2 + v^2).$$

Recall from Example 6.13 that

$$\mathcal{W}_\sigma(u, v) = \frac{2}{(1 + 4u^2 + 4v^2)^{\frac{3}{2}}} \begin{bmatrix} 1 + 4v^2 & -4uv \\ -4uv & 1 + 4u^2 \end{bmatrix}.$$

Now, fix a point $\mathbf{p} = \sigma(u, v) \in \mathcal{P}$. By Theorem 6.7, we can compute

$$\begin{aligned} H|_{\mathbf{p}} &= \frac{1}{2} \operatorname{tr} \mathcal{W}_\sigma(u, v) = \frac{2 + 4u^2 + 4v^2}{(1 + 4u^2 + 4v^2)^{\frac{3}{2}}}, \\ \mathcal{K}|_{\mathbf{p}} &= \det \mathcal{W}_\sigma(u, v) = \frac{4}{(1 + 4u^2 + 4v^2)^2}. \end{aligned}$$

Then, by applying Theorem 6.8, we recover the principal curvatures (with respect to σ):

$$\kappa_1|_{\mathbf{p}} = \frac{2}{(1 + 4u^2 + 4v^2)^{\frac{1}{2}}}, \qquad \kappa_2|_{\mathbf{p}} = \frac{2}{(1 + 4u^2 + 4v^2)^{\frac{3}{2}}}.$$

6.3. **Some Landmark Theorems.** We conclude these notes by discussing two landmark theorems in the geometry of surfaces: the *theorema egregium* and the *Gauss–Bonnet theorem*. Both results involve the Gauss curvature and are closely connected to the notion of intrinsic geometry. As a result, we begin this section with an informal discussion on the intrinsic geometry of surfaces.

6.3.1. *Notes on Intrinsic Geometry.* Recall from the very beginning of these notes, in Chapter 1, that the intrinsic geometry of a surface refers to geometric properties "of the surface itself", regardless of how it is embedded in some larger space. A basic example is to consider the spheres

$$\begin{aligned} \mathcal{C}_1 &= \{(x, y, z) \in \mathbb{R}^3 \mid x^2 + y^2 + z^2 = 1\}, \\ \mathcal{C}_2 &= \{(x, y, z) \in \mathbb{R}^3 \mid (x-1)^2 + (y-1)^2 + (z-1)^2 = 1\}, \end{aligned} \tag{6.9}$$

both of which are spheres of unit radius; see Figure 6.12. $\mathcal{C}_1$ and $\mathcal{C}_2$ are different surfaces from an extrinsic point of view, since they are situated at different points in 3-dimensional space. However, $\mathcal{C}_1$ and $\mathcal{C}_2$ can also be considered “the same” in an intrinsic sense, as both are unit spheres.

As was mentioned before, one thought experiment for informally exploring this notion of intrinsic properties was to imagine a bug living on a surface, without any awareness of the larger space in which the surface is lying. In the case of the spheres in (6.9), such a bug living on $\mathcal{C}_1$ would have exactly the same experience as another bug living on $\mathcal{C}_2$. Without any knowledge of the ambient space $\mathbb{R}^3$ in which the spheres lie, the bug would not be able to distinguish between $\mathcal{C}_1$ and $\mathcal{C}_2$, lending credence to the idea that $\mathcal{C}_1$ and $\mathcal{C}_2$ have the same intrinsic geometry.

On the other hand, consider another sphere of radius 2 (see again Figure 6.12):

$$\mathcal{C}_3 = \{(x, y, z) \in \mathbb{R}^3 \mid x^2 + y^2 + z^2 = 4\}. \tag{6.10}$$

Might $\mathcal{C}_3$ be considered intrinsically different from $\mathcal{C}_1$ and $\mathcal{C}_2$? Resorting to our thought experiment, we see that a bug living on the larger $\mathcal{C}_3$ would have to crawl further in any direction before returning to its starting point than a bug living on $\mathcal{C}_1$. Therefore, in contrast to the preceding situation, our hypothetical bug would be able to tell $\mathcal{C}_1$ and $\mathcal{C}_3$ apart.

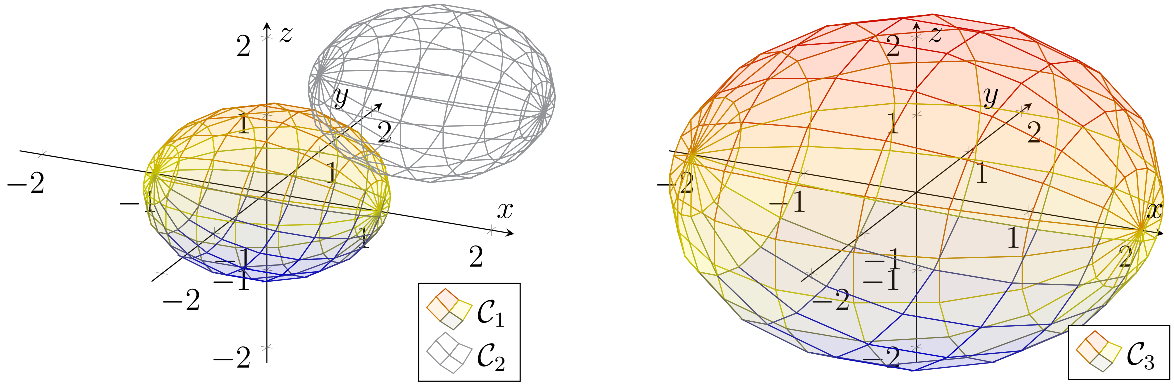
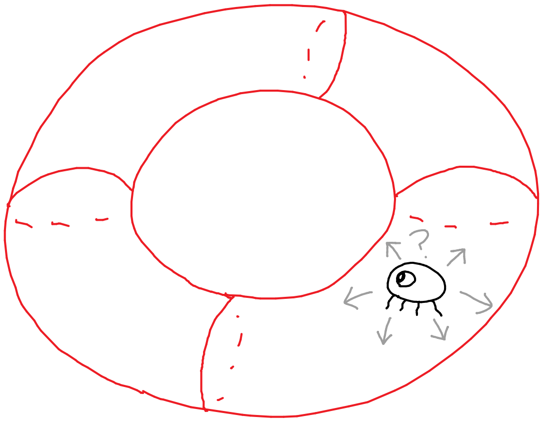
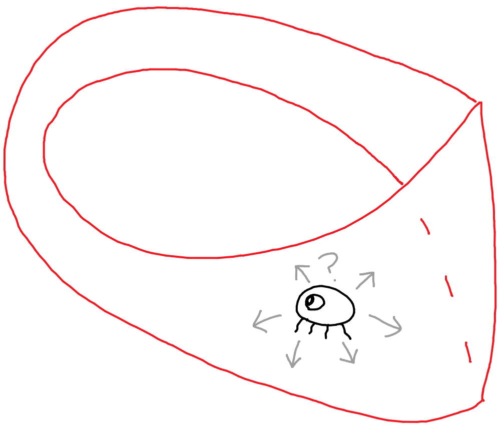


FIGURE 6.12. The left diagram contains two unit spheres $\mathcal{C}_1$ and $\mathcal{C}_2$ from (6.9), situated about different points. The right diagram contains the sphere $\mathcal{C}_3$ from (6.10), which has a larger radius. Here, $\mathcal{C}_1$ and $\mathcal{C}_2$ are “intrinsically the same”, while $\mathcal{C}_3$ is “intrinsically different” from $\mathcal{C}_1$ and $\mathcal{C}_2$.

As before, we have neither the space nor the technical background here to formally make sense of intrinsic geometry. However, like we did in our discussion of curves in Section 3.4, let us try to be a bit more precise about our bug situation by adopting the following rules:

- The bug is aware of its velocity, that is, it has knowledge of (a) which direction along the surface it is going, and (b) how fast it is going in that direction.
- In contrast to curves, there are now many more directions along a surface for the bug to crawl. Here, our hypothetical bug knows when, and how much, it is changing directions.
- Again, the bug is allowed to wander along the surface for all of eternity. Moreover, the bug has excellent memory and can recall where it has been and how it has moved in the past. In particular, it knows whether it returns to a point it has previously visited.





While the above points are informal, let us see if we can connect them to the formal concepts we have learned in the past two chapters. First, recall that for a surface $S \subseteq \mathbb{R}^n$ and a point $\mathbf{p} \in S$, the velocities along S at $\mathbf{p}$—the directions and speeds that one could go along S at $\mathbf{p}$—are represented by the tangent vectors. Since the intrinsic bug is aware of these velocities, we would expect these tangent vectors to be intrinsic. Moreover, the tangent plane $T_{\mathbf{p}}S$, which is simply the set of all tangent vectors at $\mathbf{p}$, would also be considered intrinsic by this line of reasoning.

Now, you may (and should) find this suspicious, since $T_{\mathbf{p}}S$ is defined as a plane lying $\mathbb{R}^n$ (see Definition 5.6) or as a set of "arrows" in $\mathbb{R}^n$ (see Definition 5.5), both of which depend on the ambient space $\mathbb{R}^n$. Thus, this suggests that the tangent plane ought to be extrinsic rather than intrinsic. However, it turns out that one can in fact make sense of tangent vectors and tangent planes of S in a way that is *independent* of the ambient space. (This is one of the first, and also most confusing, definitions that one would learn in an advanced differential geometry module; to keep things relatively simple, we will not discuss this definition in these notes.) This establishes *both tangent planes and vectors as intrinsic properties of surfaces.*

Next, the bug is also capable of determining its speed, which corresponds to measuring the length of a tangent vector. The bug can also measure changes in directions, which can be represented as angles between tangent vectors. These two pieces of information—lengths and angles—precisely comprise the dot product of tangent vectors. As a result, we conclude that *the first fundametal form* (i.e. the dot product on tangent planes) *is also an intrinsic property of surfaces.*

In particular, the bug can crawl around a surface and compute its first fundamental form at various points. By integrating the determinant of the first fundamental form, as in Definition 6.3, the bug can measure the area of the surface (by the way, the bug received an $A+$ in calculus). This suggests that surface area is yet another intrinsic property of surfaces.

Moreover, going back to our rules for what our bug can perceive, we see that these correspond rather precisely to the information contained in the tangent planes and first fundamental form of a surface. Thus, mathematically speaking, we can go out on a limb and try to *formally define* what we mean by the intrinsic geometry of a surface as follows:

> *The intrinsic geometric properties of a surface are defined to be precisely those that can be characterised using only the points of the surface itself, the various tangent planes of the surface, and the first fundamental form of the surface.*

The branch of differential geometry arising from this viewpoint is known as *Riemannian geometry*, which originated from Bernhard Riemann (German mathematician, 1826–1866) in the 1860s.

We end this discussion by considering some other properties of surfaces that we have studied:

- The unit normals of a surface S are defined as arrows based on S but pointing away from S. To "point away from S", one requires S to live in some larger space. Thus, we see that the unit normal is an extrinsic property of S depending on the ambient setting.
- The second fundamental form of a surface $\mathsf{S} \subseteq \mathbb{R}^3$, which measures how its unit normal changes as one moves along S, is also an extrinsic property.
- Recall that the "two-sidedness" of a surface S was defined through its unit normals. As a result, one might expect orientability (see Definition 5.14) to be an extrinsic property. However, one can characterise orientability using only the tangent planes and the first fundamental form (again, for brevity, we avoid discussing any details in these notes). As a result, orientability is fact an intrinsic property of surfaces.
- Imagine a bug crawling along a surface without changing its velocity, that is, without altering either its speed or direction. These trajectories of non-accelerating bugs are known as the *geodesics* of a surface. One can in fact show that geodesics can be constructed using only the first fundamental form and hence are intrinsic properties of surfaces.

In Section 3.4, we described all the possible intrinsic curve geometries. However, the situation for surfaces is far most complicated, as we will see in the subsequent discussions, and we will not be able to achieve such a simple characterisation of all surface geometries.

6.3.2. *The Theorema Egregium.* Previously, we argued that the unit normals and the second fundamental form of a surface $\mathsf{S} \subseteq \mathbb{R}^3$ are extrinsic properties. Following this chain of reasoning:

- The Weingarten matrix, defined from the first and second fundamental forms (see the formula in Definition 6.6), is also an extrinsic property of surfaces.
- Since the principal curvatures are simply the eigenvalues of the Weingarten matrix, then the principal curvatures are also extrinsic properties.
- The mean curvature (see Definition 6.10) can also be shown to be extrinsic.

Now, since the Gauss curvature is simply the product of the principal curvatures, one would expect it to also be an extrinsic property of surfaces. What is miraculous, however, is that the Gauss curvature is in fact an intrinsic property. This is the statement of the theorema egregium, established by Carl Friedrich Gauss (German mathematician, 1777–1855) in the 1820s:

Theorem 6.9. *The Gauss curvature is an intrinsic geometric property of surfaces.*

The *theorema egregium*, which is Latin for "remarkable theorem", stands as one of the crowning achievements in differential geometry. The remarkable fact is that this Gauss curvature, defined using a combination of extrinsic quantities, actually contains information that is more fundamental and intrinsic in nature, that is, it is completely independent of how a surface is situated in a larger space. Indeed, through some extensive computations, one can show that the Gauss curvature can be expressed only in terms of the first fundamental form and its derivatives; see, for instance, [5].

Let us discuss some implications of the theorema egregium. First, a bit of terminology: we say *isometric* to mean that two surfaces possess the same intrinsic geometry. More specifically:

Definition 6.12. *Let S_1 and S_2 be two surfaces. We say that S_1 and S_2 are isometric iff they "have the same tangent planes and first fundamental form".*

Of course, Definition 6.12 is merely an informal statement. There is a formal, rigorous definition of isometry (see [1]), however we will not discuss the details in these notes. Note that in light of Definition 6.12, the *theorema egregium* can be restated as follows:

Corollary 6.10. *Two surfaces that are isometric have the same Gauss curvature.*

Although our understanding of isometry is informal, we can still discuss its interpretations. Recall that the tangent planes and the first fundamental form describe directions, lengths, and angles along a surface. Thus, surfaces S_1 and S_2 that are isometric have the same information regarding directions, lengths, and angles. One way to think of this is to imagine that S_1 can be transformed into S_2, and vice versa, in a manner that preserves distances and angles; in other words, S_1 can be transformed into S_2 without "stretching" the surface along any direction.

Example 6.23. *The spheres $\mathcal{C}_1$ and $\mathcal{C}_2$ from (6.9) are isometric. In fact, $\mathcal{C}_1$ can be transformed into $\mathcal{C}_2$ by simply translating (i.e. shifting) its center from the origin to the point $(1, 1, 1)$. Note this transformation is isometric, as it does not alter any distances and angles along the sphere.*

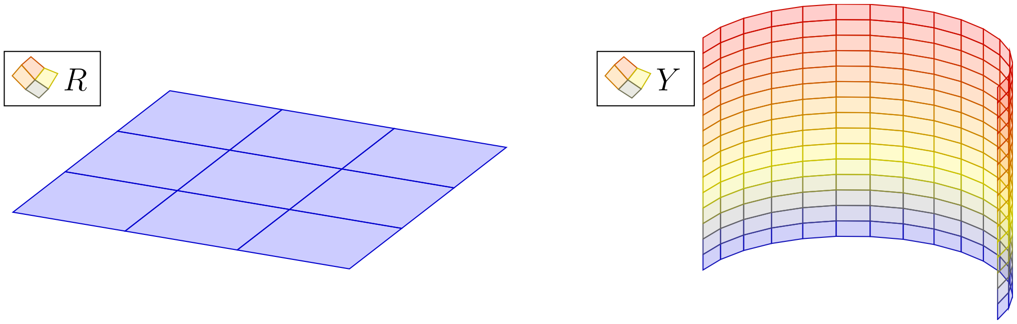


Figure 6.13. The surfaces R (left) and Y (right) from Example 6.24.

Example 6.24. *For a more interesting example, consider a rectangle and a half-cylinder:*

$$R = \{(x, y, z) \in \mathbb{R}^3 \mid 0 < x < \pi,\ 0 < y < \pi,\ z = 0\},$$
$$Y = \{(x, y, z) \in \mathbb{R}^3 \mid x^2 + y^2 = 1,\ y > 0,\ 0 < z < \pi\}.$$

See Figure 6.13 for graphical depictions of R and Y, and observe the obvious: R looks flat, while Y looks curved. However, one can actually argue that R and Y are isometric.

Informally, one can imagine R as being a flat piece of paper; one can then pick up this paper and bend it into the half cylinder Y. Although this transformation curves the piece of paper along one direction, it does not change any distances or angles, i.e. it does not stretch this piece of paper along any direction. As a result, we see that R and Y are intrinsically the same.

On a more formal level, we can show that, with respect to certain parametrisations, the first fundamental forms on both R *and* Y *are simply the identity matrix; see Example 6.1.*

In addition, note that R*, being flat, has zero Gauss curvature. Similarly, recall from Example 6.19 that the cylinder, and hence its subset* Y*, also has zero Gauss curvature. (Recall that this is due to* Y *not being curved in the* z*-direction.) Consequently, in this case, the theorema egregium is affirmed: the isometric surfaces* R *and* Y *indeed have the same Gauss curvature.*

For a real-world connection to the *theorema egregium*, let us think of a map of the world projected onto a flat surface. Recall that some continent on this map (usually Antarctica) is always much bigger than it actually is; in other words, *some part of this map is distorted*. Why is this so? Why doesn't someone draw a map that does not distort anywhere?

The reason lies in the fact that the surface of the globe—the contents of the map—is actually (approximately) a sphere. Recall from Example 6.20 that the sphere has positive Gauss curvature. The surface of the map, in contrast, is flat and hence has zero Gauss curvature.

Now, if one could transplant the sphere onto a flat piece of paper without stretching, then the theorema egregium implies that the paper and the sphere must have the same Gauss curvature. However, the above shows this is not the case. Hence, we conclude that *any mapping from a globe's surface onto a flat piece of paper must stretch and distort the surface somewhere.*

The second real-world connection we discuss comes in the form of a life hack. Suppose you are eating a slice of pizza. To keep your hands clean, you of course grab the pizza by the crust on the edge. However, if the crust is soft, then the pizza slice will bend downward toward the centre, and you will be very sad as all the topping falls off; see the left drawing in Figure 6.14.

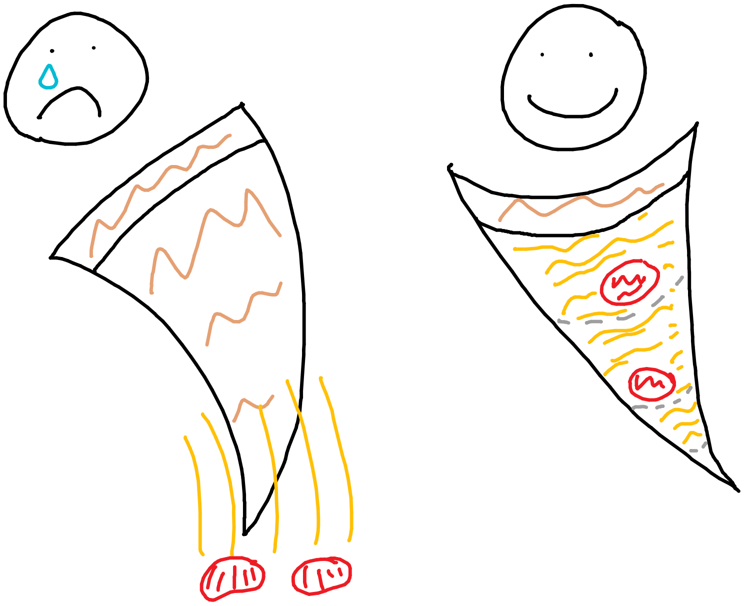


FIGURE 6.14. The left picture shows the "wrong" way to eat a slice of pizza, in which all the scrumptious topping falls off. The right picture shows how to eat the slice of pizza "correctly", so that the topping stays on.

In this situation, the life hack is as follows: *when holding the pizza, curl it upwards a bit in the angular direction*, as in the right drawing in Figure 6.14. With the pizza slightly bent in the angular direction, it will no longer bend downward toward the centre. Thus, the topping will remain safely on the pizza, and you remain happy with your meal.

So why does this work? The idea is that an unbent pizza slice lies flat and hence has zero Gauss curvature. Assuming the crust does not stretch, then by the *theorema egregium*, any bending of the pizza must preserve the vanishing Gauss curvature. Here, the pizza can bend downward toward the centre, as in the left drawing in Figure 6.14, without changing the Gauss curvature, since the pizza remains unbent in the remaining angular direction; see Figure 6.15.

On the other hand, if we hold the pizza slice as in the right drawing in Figure 6.14, then it is already bent in the angular direction. In this case, one cannot further bend the pizza radially toward the centre without changing the Gauss curvature (since if this is done, then the pizza slice is bent in both principal directions). In summary, thanks to the theorema egregium, by preemptively bending in one direction, the pizza slice now becomes rigid in the other direction.

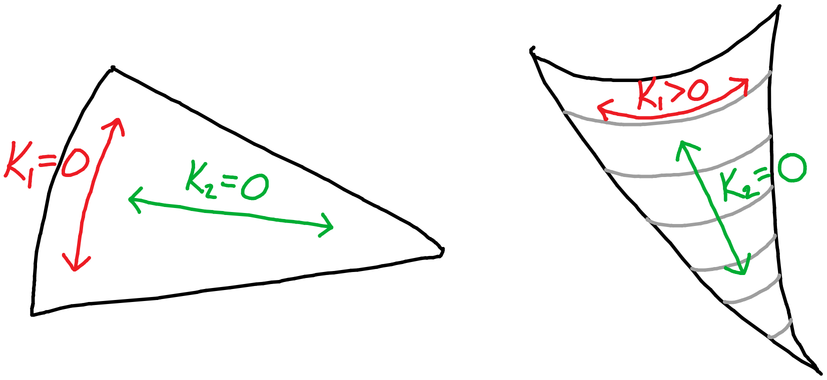


Figure 6.15. In the left diagram, the pizza slice is flat in both the radial (green) and angular (red) directions; here, the pizza can be bent in either one of the two directions and still maintain zero Gauss curvature. In the right diagram, the pizza is flat in the radial (green) direction but already bent in the angular (red) direction; here, the pizza can no longer be bent in the radial direction without changing the Gauss curvature.

Remark. One can do a similar but less messy demonstration of this principle using a piece of paper in the place of a greasy pizza slice. For the same reason as before, bending the paper along one dimension again prevents the paper from being bent along the remaining dimension.

Finally, recall from Section 3.4 that curves could only have four basic "types" of intrinsic geometries: (1) lines, (2) rays, (3) finite intervals (of various lengths), and (4) loops (of various lengths). In particular, while a curve could be curved in a wide variety of manners, all of this information is in fact purely extrinsic in nature. In other words, how a curve is curved is completely dependent on how this curve is situated within some larger space.

The *theorema egregium* implies, in contrast, that the intrinsic geometry of surfaces is far more complicated. While the principal and mean curvatures are extrinsic and depend on how a surface is embedded in a larger space, a very special part of this curvature information—the Gauss curvature—is in fact intrinsic. Thus, unlike curves, a surface can be "more fundamentally" curved in a manner that is divorced from how it sits in the ambient space. Simple examples we have already discussed include the sphere (Example 6.20) and the hyperboloid (Example 6.21).

6.3.3. *The Gauss–Bonnet Theorem.* For a surface $\mathsf{S} \subseteq \mathbb{R}^3$, the *theorema egregium* shows that its Gauss curvature $\mathcal{K}$ contains information about its intrinsic geometry. Now, since $\mathcal{K}$ is a real-valued function on S, we can make sense of the *surface integral* of the Gauss curvature $\mathcal{K}$ over S. Another miraculous fact, on top of the *theorema egregium*, is that this integral of $\mathcal{K}$ contains information that is even more "fundamental" than the intrinsic geometry of S.

This is the topic of the other landmark result of this section, the *Gauss–Bonnet theorem*, which was proved independently by Carl Friedrich Gauss (of the *theorema egregium*) and by Pierre Bonnet (French mathematician, 1819–1892) in the early-to-mid 1800s. Here, we will discuss the simplest of many versions of the theorem. For more general statements of the Gauss–Bonnet theorem, see previous years' lecture notes or a textbook on differential geometry.

Before, giving a formal statement of the theorem, we must first set some terminology.

Definition 6.13. *A surface $\mathsf{S} \subseteq \mathbb{R}^3$ is called closed iff the following hold:*

- *S is bounded, that is, S is contained in a finite region of $\mathbb{R}^3$.*
- *S is boundaryless, that is, S "has no edge, where it suddenly terminates".*

Rather than giving formal definitions of *bounded* and *boundaryless*, we keep with the informal nature of our discussions and instead demonstrate these by example:

Example 6.25. *The sphere $\mathbb{S}^2$ (see Definition 5.8) is bounded, since it lies within the finite region*

$$\{(x, y, z) \in \mathbb{R}^3 \mid x^2 + y^2 + z^2 \leq 1\},$$

i.e. the closed unit ball about the origin. On the other hand, the hyperboloid from Example 6.17 is not bounded, since there is a portion of it that "goes off toward infinity"; see Figure 6.10.

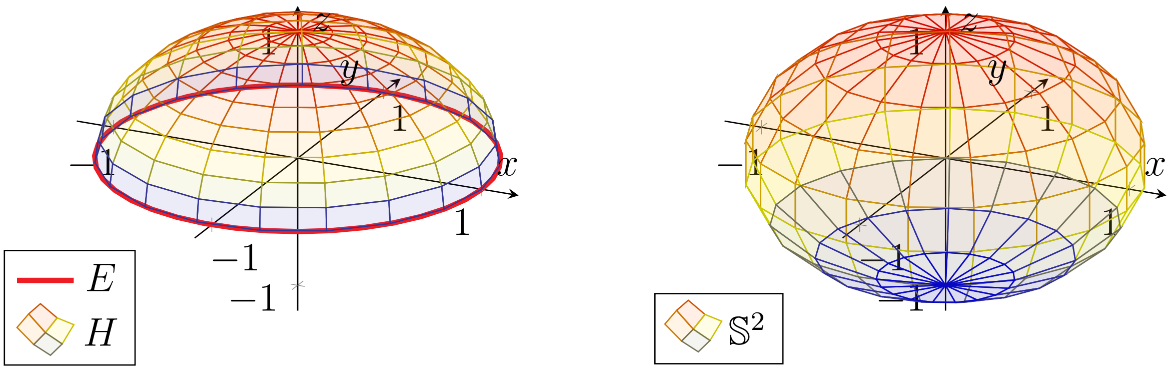


FIGURE 6.16. The left drawing is the hemisphere H from Example 6.26, with its boundary E drawn as a red curve. The right drawing, on the other hand, is the sphere $\mathbb{S}^2$, which is boundaryless.

Example 6.26. *The hemisphere from Example 5.7, given by*

$$\mathsf{H} = \{(x, y, z) \mid x^2 + y^2 + z^2 = 1,\ z > 0\},$$

is not boundaryless. In particular, its boundary is given by the equator

$$\mathsf{E} = \{(x, y, z) \mid x^2 + y^2 = 1,\ z = 0\}.$$

This is drawn in the left diagram of Figure 6.16, with the boundary E *in red.*

On the other hand, the sphere $\mathbb{S}^2$ *is boundaryless, as it lacks a similar boundary set at which* $\mathbb{S}^2$ *suddenly "terminates"; see the right drawing in Figure 6.16.*

Definition 6.14. *The genus of a surface* $\mathsf{S} \subseteq \mathbb{R}^3$ *is the "number of holes in* S*".*

Again, for brevity, the genus is better demonstrated informally through examples:

Example 6.27. *Consider the three surfaces drawn in Figure 6.17, which have, from left to right, zero "holes", one "hole", and two "holes". Thus, by Definition 6.14, the genus of each of the surfaces in Figure 6.17 are—again from left to right—*0*,* 1*, and* 2*.*

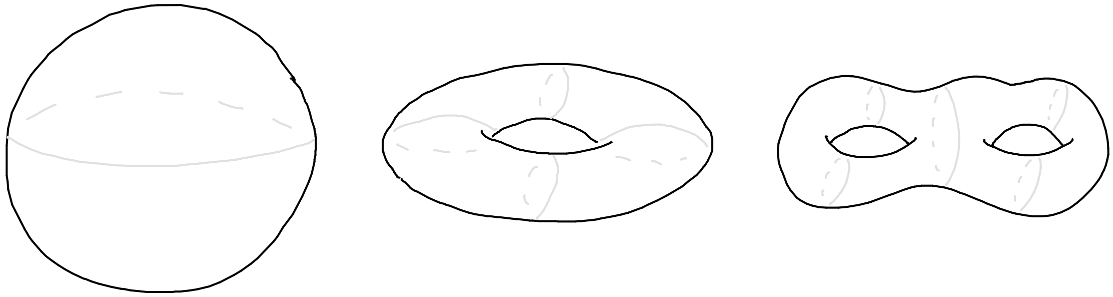


FIGURE 6.17. The above surfaces have genus 0, 1, and 2, respectively.

For more formal discussions of all the above terminology, the reader is referred to [1, 5]. With this terminology in place, we can now state the Gauss–Bonnet theorem:

Theorem 6.11. *Let* $\mathsf{S} \subseteq \mathbb{R}^3$ *be a closed surface, and let* $\mathcal{K}$ *denote its Gauss curvature. Then,*

$$\iint_{\mathsf{S}} \mathcal{K} \, dA = 4\pi(1 - g_{\mathsf{S}}),$$

where g_{S} *denotes the genus of* S*.*

The proof of the Gauss–Bonnet theorem relies on finding special identities for the Gauss curvature (this is closely connected to the *theorema egregium* and the intrinsic description of the Gauss curvature) and on applying *Green's theorem* from vector calculus to integrals on local parametrisations of S. See the lecture notes [4] or the textbooks [1, 5] for details.

Example 6.28. *Consider the sphere in the left drawing* S_1 *of Figure 6.18. From the brief discussion in Example 6.27, we know that* S_1 *which has genus zero. Thus, by the Gauss–Bonnet theorem,*

$$\iint_{\mathsf{S}_1} \mathcal{K} \, dA = 4\pi(1 - 0) = 4\pi.$$

In case you are not yet convinced, this integral can also be computed directly. Suppose S_1 *is the unit sphere* $\mathbb{S}^2$*, so that* $\mathcal{K} \equiv 1$ *by Example 6.20. As a result,*

$$\iint_{\mathsf{S}_1} \mathcal{K} \, dA = \iint_{\mathbb{S}^2} dA$$

is simply the surface area of $\mathbb{S}^2$*, which, by the computations in Example 6.5, is indeed* 4π*.*

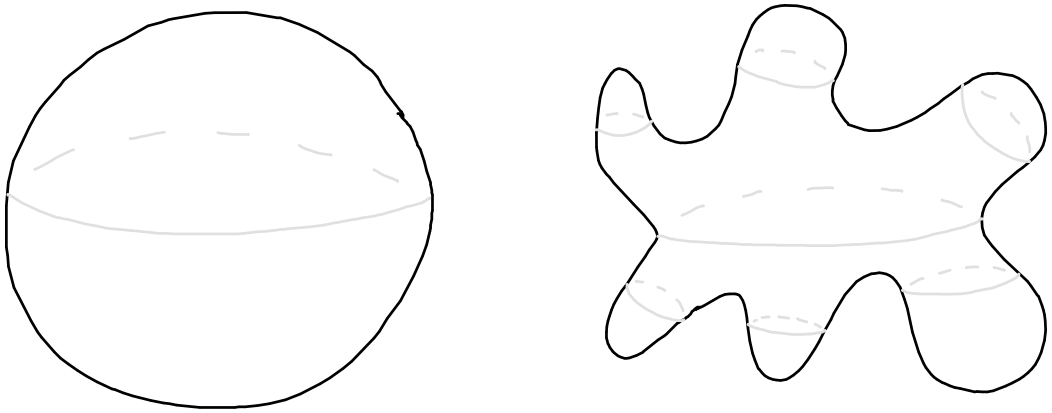


FIGURE 6.18. The surfaces S_1 and S_2 in Examples 6.28 and 6.29.

Example 6.29. *Next, observe that the surface* S_2 *in the right drawing of Figure 6.18 also has no holes, and hence also has genus* 0*. Therefore, the Gauss–Bonnet theorem again implies*

$$\iint_{S_2} \mathcal{K}\, dA = 4\pi(1-0) = 4\pi.$$

Of course, S_2 *has vastly different shape than* S_1*, with* S_1 *and* S_2 *having very different values for their Gauss curvatures. (In particular, the Gauss curvature of* S_2 *is negative in many places.) What is interesting, though, is that the surface integrals over* S_1 *and* S_2 *of these rather different Gauss curvatures are equal! As a result, while the Gauss curvature is negative at some points of* S_2*, it must also be extra positive at other points in order to compensate for the negativity.*

In other words, these surface integrals of the Gauss curvature over S_1 *and* S_2 *do not see the specific shapes of these surfaces. Indeed, the Gauss–Bonnet theorem states that the only information that is captured by this is the number of "holes" (that is, zero) in* S_1 *and* S_2*.*

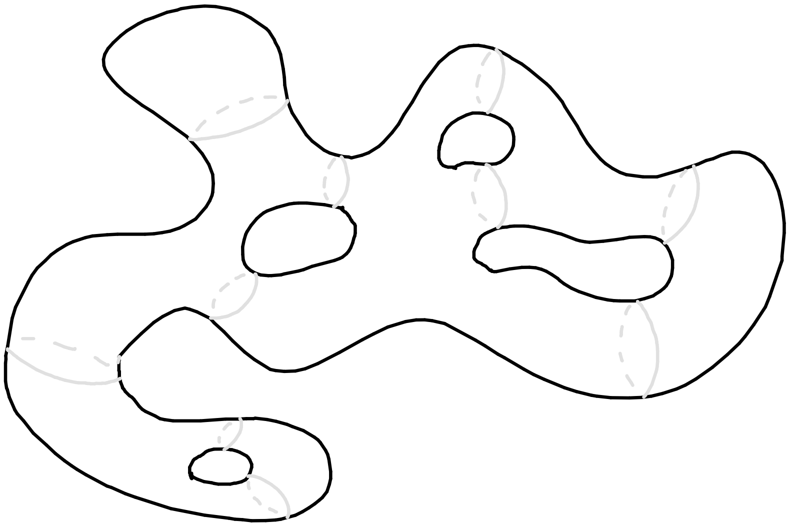


FIGURE 6.19. The surface S_3 discussed in Examples 6.30.

Example 6.30. *Finally, the surface* S_3 *in Figure 6.19 has genus* 4*, thus*

$$\iint_{S_3} \mathcal{K}\, dA = 4\pi(1-4) = -12\pi .$$

Again, no matter how complex the shape of S_3 *may be, as long as it contains four "holes", then the above integral must necessarily evaluate to* -12π*.*

Now, if we take a surface and we deform (i.e. stretch or compress) it just a tiny bit, then this operation will not change the genus of the surface. For example, this was the case in Examples 6.28 and 6.29, when S_1 was "deformed" into S_2. Thus, the Gauss–Bonnet theorem states that the surface integral of the Gauss curvature is in fact independent of these deformations. In other words, this integral is more than just an intrinsic geometric property of surfaces; it describes an even more fundamental *topological* property, i.e. the number of "holes" the surface contains.

This is the remarkable aspect of the Gauss–Bonnet theorem. Here, one is dealing with the Gauss curvature and surface integrals, both of which are (intrinsic) geometric properties of surfaces. However, this special combination of these geometric quantities—the surface integral of the Gauss curvature—yields instead a purely topological property. Indeed, the Gauss–Bonnet theorem highlights a deep connection between the geometry and the topology of surfaces.

Furthermore, observe that around each "hole" of a surface, for instance one of the four found in S_3 from Example 6.30, the Gauss curvature is negative. (This can be argued from the intuitions described within and following Figure 6.11.) More specifically, the Gauss–Bonnet theorem states that each "hole" contributes precisely -4π to the total Gauss curvature of the entire surface.

Finally, the Gauss–Bonnet theorem can be compared another familiar quantity: the winding number of plane curves. Recall that the winding number, defined using geometric quantities (the signed curvature and path integrals) in Definition 3.6, also describes a topological property that is not changed by deformations: the number of times the curve revolves anticlockwise. Thus, one can think of the Gauss–Bonnet theorem as a kind of analogue of winding numbers for surfaces.

Notes and Acknowledgments

The present notes are based on notes [6] from an earlier iteration of *MTH5109* that I taught in *Autumn 2016*. (Moreover, the *Autumn 2016 MTH5109* lecture notes were themselves heavily based on the *Autumn 2015 MTH5109* lecture notes [4] by *Prof. Shahn Majid*.)

For the current notes, I added several elaborations on basic concepts, including:

- An introductory section on the main ideas and aims of the module.
- More discussion on what curves and surfaces formally are.
- More detailed interpretations of various definitions and results.
- More emphasis on tangent lines and planes.
- A more systematic development of integrals on curves and surfaces.
- More details on unit normals and orientations of surfaces.
- Additional examples of curves and surfaces, along with computer-generated plots.
- Some informal discussions on intrinsic versus extrinsic geometry.

To compensate for this additional material, some topics from previous years were removed:

- The Frenet–Serret formulas.
- The theory of curves in surfaces: geodesic and normal curvatures, Euler's theorem.
- Geodesic curves.
- More general versions of the Gauss–Bonnet theorem (i.e. for surfaces with boundaries).
- Proofs of the Gauss–Bonnet theorems.

While these notes were written with the intention of being entirely self-contained, the interested reader may also be interested in the following texts from the official module reading list:

- Christian Bär, *Elementary Differential Geometry*, 2010, [1].
- Andrew Pressley, *Elementary Differential Geometry*, 2010, [5].
- W. B. Raymond Lickorish, *An Introduction to Knot Theory*, 1997, [3].

These texts contain material related to and beyond what is covered by these notes.

References

1. C. Bär, Elementary Differential Geometry, Cambridge University Press, 2010.
2. K. R. Davidson and A. P. Donsig, Real Analysis and Applications: Theory in Practice, Springer, 2010.
3. W. B. R. Lickorish, An Introduction to Knot Theory, Springer, 1997.
4. S. Majid, MTH5109 (Geometry II) Lecture Notes, 2015–2016, Queen Mary University of London.
5. A. Pressley, Elementary Differential Geometry, Springer, 2010.
6. A. Shao, MTH5109 (Geometry II) Lecture Notes, 2016–2017, Queen Mary University of London.
7. Wikipedia: Connected space,
   `http://en.wikipedia.org/wiki/Connected_space`.
8. Wikipedia: Curve,
   `http://en.wikipedia.org/wiki/Curve`.
9. Wikipedia: Geometry,
   `http://en.wikipedia.org/wiki/Geometry`.
10. Wikipedia: Klein bottle,
   `http://en.wikipedia.org/wiki/Klein_bottle`.

11. Wikipedia: List of trigonometric identities,
`http://en.wikipedia.org/wiki/List_of_trigonometric_identities`.

12. Wikipedia: Open set,
`http://en.wikipedia.org/wiki/Open_set`.

School of Mathematical Sciences, Queen Mary University of London, London E1 4NS, United Kingdom
Email address: `a.shao@qmul.ac.uk`