MTH5113: INTRODUCTION TO DIFFERENTIAL GEOMETRY LECTURE NOTES (2021–2022)

DR. ARICK SHAO

1. Introduction to Curves and Surfaces

In this module, we will study *geometry*. According to our favourite source *Wikipedia* [11], geometry is the branch of mathematics that deals with the *shapes* and *sizes* of objects. Throughout this term, you will develop a better understanding of these concepts, as well as explore how they can be quantified and computed.

In your mathematical education—for instance, in calculus and linear algebra—you have considered flat, linear spaces such as the real line ($\mathbb{R}$), the Euclidean plane ($\mathbb{R}^2$), and higher-dimensional Euclidean spaces ($\mathbb{R}^n$). This module will expand your outlook to *curved* settings. For example, how is the study of the surface of a ball similar to, and different from, the study of a flat piece of paper?

Another aim of this module, as suggested by its title, is to apply ideas and tools from calculus and linear algebra to study these geometric questions. In addition, we will use the geometric insights we gain to obtain new understandings of familiar concepts, such as vectors, differentiation, and integration.

1.1. Frequently Asked Questions.

Before launching into more serious discussions, let us first address some common questions you may have regarding this module.

Question 1.1. Why would I want to study curved objects? 😒

Curved objects can be found everywhere in our lives:

- If you throw a ball or shoot a rocket into the air, then the trajectory of the ball or rocket will be curved rather than linear. Curves have also historically been used in astronomy to model the motions of celestial bodies.
- The surface of the earth is not flat, but like a sphere. To study this surface as a whole, we have to understand the effects of its curvature. Historically, this has played important roles in navigation and astronomy; a very concrete example is 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 (called a "spacetime"). Moreover, gravity itself is modelled by the shape and curvature of this spacetime. In fact, a careful understanding of this curvature is necessary for the GPS systems in our phones to work correctly!

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

As this is an introductory module, we will unfortunately have to restrict our discussions here to one-dimensional (*curves*) and two-dimensional (*surfaces*) objects. If you are interested in four-dimensional spacetimes and gravity, then you should consider the third-year module *MTH6132: Relativity*.

Question 1.2. Sounds interesting. But, what maths will I need to know? 😮

This module is mainly concerned with the *differential geometry* of curves and surfaces. In particular, we look at objects that "vary smoothly enough" so that we can take *derivatives*, or linear approximations, of them. Moreover, to measure the sizes—e.g. length and area—of objects, we will need to compute various *integrals*.

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

- In our study of curves, we will make frequent use of single-variable calculus (mainly, contents from *MTH4*00: Calculus I*).
- For surfaces, we will require knowledge of partial derivatives and double/triple integrals (which you saw in *MTH4*01: Calculus II*).

The simplest geometric objects that we can consider are lines and planes; these fall under the study of *linear algebra*. For this task, we will sometimes refer to background knowledge in *MTH4*15: Vectors and Matrices*, as well as either *MTH5112: Linear Algebra I* or *MTH5212: Applied Linear Algebra*.

Even for more complex curved objects, a useful tool in their analysis is *linearisation*—to first study the linear objects that best approximate them. Thus, we cannot really escape the need to understand linear algebra and its connections to geometry.

Finally, in order to construct a diverse collection of examples of geometric objects, we will make use of many elementary functions:

- *Polynomials* (e.g. $t^2 + 1$), and *rational* functions (quotients of polynomials).
- *Trigonometric* (sin, cos) and *hyperbolic* (sinh, cosh) functions.

- *Exponential* (exp) and *logarithmic* (ln) functions.

We will, at times, reference a few basic properties of these functions.

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

Question 1.3. Help! What will I be expected to do? 😬

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

As the module is also concerned with quantifying geometric properties, you will be expected to demonstrate that you are capable of performing various types of computations. Again, these computations will involve elements of calculus (e.g. derivatives and integrals) and linear algebra (e.g. vector computations).

You should also gain a visual understanding of these geometric concepts. As a result, you will be asked to graph various curves and surfaces, on a plane or in space.

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 need to have a general understanding of why results are true.

Question 1.4. I can do it! Now, what will I actually learn in the module? 😄

In the remainder of this chapter, we give a brief and informal outline of the main themes to be discussed throughout this module.

1.2. **Vectors and Calculus, Revisited.** The first part of the module is geared toward revising various aspects of vector algebra and calculus. This provides an opportunity for you to revisit details that you have forgotten and to fill in gaps in your existing knowledge. However, rather than simply repeating material you have already learned (boring!), we will also introduce new geometric perspectives, thereby developing a different and more refined understanding of old ideas.

The module begins by revisiting the notion of *vectors* and their role in geometry. As you know well, vectors are often visually depicted as arrows. We will develop this

simple intuition formally, via the notion of *tangent vectors*. Later, these objects will play key roles in quantifying various geometric properties of curves and surfaces.

From here, we move into the study of *vector-valued functions* and *vector fields* (see Figure 1.1). These are used to represent a variety of physical phenomena, such as fluid flow, gravitational fields, and the populations of competing species in an ecosystem. Our objective is to better understand how such quantities are depicted visually, as well as to remind ourselves of how to perform various computations with these objects.

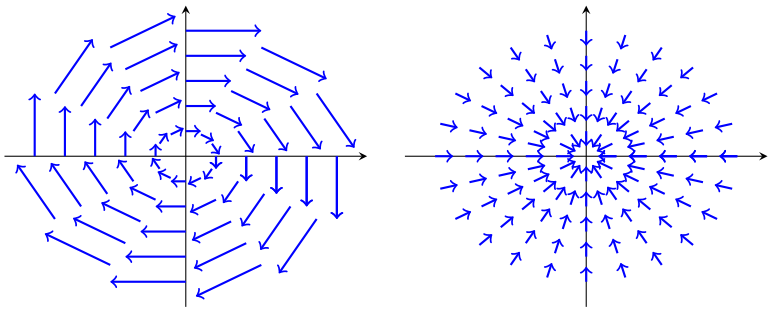


FIGURE 1.1. The two graphics depict vector fields on a plane. The left vector field describes clockwise rotation around the origin, while the right vector field describes attraction toward the origin.

The main focus of this discussion will be on calculus. First, we recall the standard differential operations in vector calculus (e.g. *derivatives* and *partial derivatives*). We extend these concepts to vector-valued functions, and we revise how these derivatives can be computed. We then turn our attention toward the other side of calculus—*integration*. In particular, we discuss the view of integrals as "weighted" lengths, areas, and volumes; we then recall some basic tricks for computing them.

1.3. **The Geometry of Curves.** With revisions behind us, we next move into the core of the module: studying the geometry of curves and surfaces. The first half of this effort focuses on 1-dimensional geometric objects: *curves*.

The first step is to make precise sense of what a curve is, by formulating a rigorous definition of curves. We then discuss some familiar ways to describe curves, for instance, parametrically or as level sets of functions. One fundamental idea here is *independence of parametrisation*—although one can view a curve from different perspectives, all these viewpoints are of the same underlying object.

We will also explore various geometric properties of curves. As a basic example, consider the curves C_1, C_2, C_3 in Figure 1.2. Your intuition likely tells you C_1 is

"straight", while C_2 and C_3 are "curved". You probably also sense C_3 is "more curved" than C_2. But, how might you capture and quantify this mathematically?

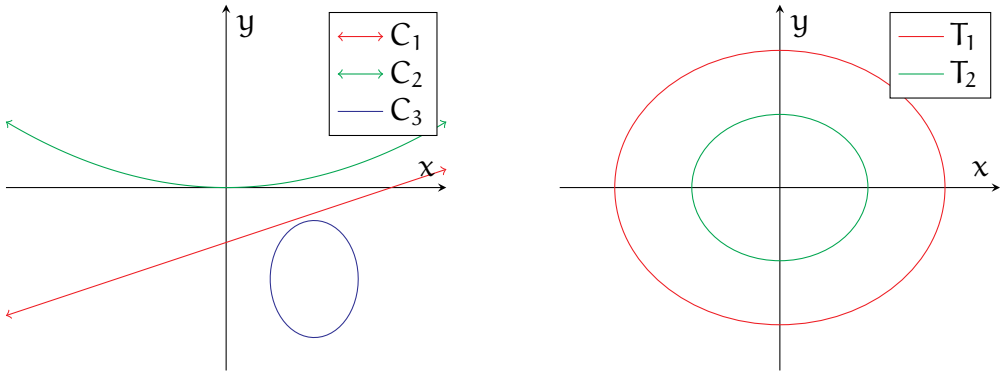


FIGURE 1.2. In the left graphic, C_1 is "straight" while C_2 and C_3 are "curved". On the right, T_1 and T_2 are circles with different circumferences.

Another aspect of geometry is studying the sizes of objects. Consider the circles T_1 and T_2 in Figure 1.2. Intuitively, T_1 and T_2 have similar shapes but different "sizes". You probably also have enough background to know that this can be captured by measuring their arc lengths (i.e. circumferences).

From calculus, you know that lengths are generally evaluated using integrals, and you should be familiar with integrals along an interval of the real line. The question of computing arc lengths will force us to make sense of what it means to integrate instead along curves, leading us to the notion of *curve* or *path integrals*. As we will see, curve integrals are a powerful abstraction with a variety of applications; one example from classical mechanics is the total work done by a force.

1.4. **The Geometry of Surfaces.** The other half of our foundations concerns the geometry of surfaces. This is largely analogous to the study of curves, but now the objects are 2-dimensional, which brings forth a new batch of complications.

Again, our analysis begins with precisely defining what a surface is, as well as with devising various methods for describing surfaces (e.g. parametrically or as level sets of functions). We then move on to study various geometric properties of surfaces.

Considering the surfaces in Figure 1.3, 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" (for instance, S_3 and S_2). What is novel, in contrast to curves, is that a surface can bend in different ways along different directions. For

instance, both spheres S_2 and S_3 always bend inward toward itself, while the "saddle" S_4 bends both toward and away from itself, depending on which direction you look.

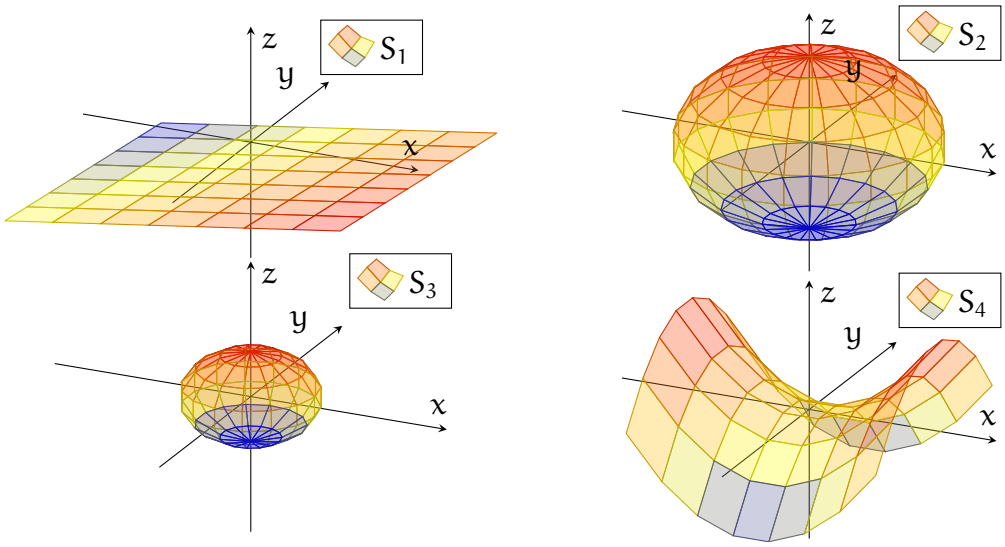


FIGURE 1.3. S_1, S_2, S_3, S_4 are surfaces in 3-dimensional space. S_1 is "flat", while the remaining surfaces are "curved".

Yet another interesting example of a geometric property of surfaces is *orientability*—whether a surface is "two-sided". Most familiar examples, including all the surfaces found in Figure 1.3, are two-sided. However, we will also encounter some exotic surfaces that are not two-sided, such as the Möbius strip and the Klein bottle.

We will also discuss how one measures the size—that is, the area—of a surface. For instance, the spheres S_2 and S_3 in Figure 1.3 have similar shapes but different sizes. In analogy with previous discussions for curves, this leads us to an integration theory along surfaces. Again, these *surface integrals* can used to model many phenoma; a classical example from physics is the electric flux through a surface.

1.5. **Putting It All Together.** In later parts of this module, we will gather all the theory we have developed—a hybrid of geometry, calculus, and linear algebra—and explore some topics and problems lying at the intersection of all these areas.

A classical problem in calculus that you have already encountered is that of finding the maximum and minimum values of a function, as well as where these values are attained. Here, we see how this analysis generalises to functions defined on curves and surfaces. For instance, can we use similar strategies to find, say, the locations on the earth's surface having the highest temperature?

Moreover, we will attack this same problem from a different viewpoint—we rebrand it as a problem of optimising a function (on the larger space) subject to contraints (i.e. we only care about the points lying on a curve or surface). This leads to the method of *Lagrange multipliers*, which has widespread applications in physics, engineering, and economics, as well as within mathematics itself.

We will conclude the module with some landmark theorems from vector calculus. A starting point of this discussion is the famous *fundamental theorem of calculus*,

$$\int_a^b f'(t)dt = f(b) - f(a), \tag{1.1}$$

which highlights a deep, and in many ways incredible, connection between differentiation and integration. We can then ask whether there are higher-dimensional, geometric generalisations of the fundamental theorem of calculus.

Here, we will study several such results: *Green's*, *Stokes'*, and *divergence theorems*. These theorems play important roles in describing many real-world phenomena. One well-known example in physics is *Gauss's law*, which connects the total electric charge enclosed inside a surface to the electric flux measured on this surface.

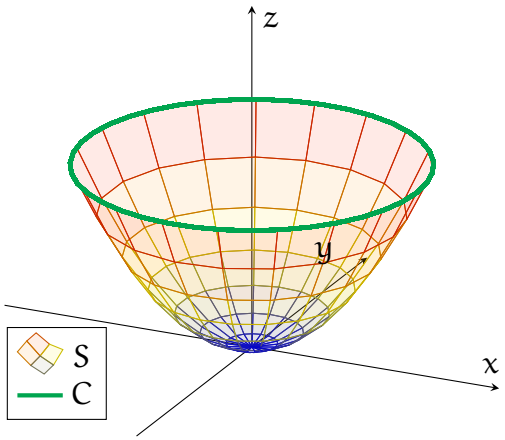


FIGURE 1.4. Stokes' theorem relates integrals over a 2-dimensional surface S with other integrals over its 1-dimensional boundary C.

Of course, what is covered in this module is but a mere introduction to a large and growing area of mathematics known as *differential geometry*. If you are especially interested in the material, then you are certainly encouraged to pursue more advanced courses in differential geometry, either in third year or beyond!

2. Vector Calculus

In this chapter, we revisit a number of concepts from vector algebra and vector calculus that you have seen before. This will include a sample of contents from:

- *MTH4*00: Calculus I*
- *MTH4*01: Calculus II*
- *MTH4*15: Vectors and Matrices*
- *MTH5*12: Linear Algebra I/Applied Linear Algebra*

One objective of this chapter is to refamiliarise yourself with content from these modules, as we will make ample use of them later in these notes.

In addition, we will approach these topics from a different, and more geometric, point of view. Thus, while all of this material should already be rather familiar to you, we will nonetheless aim to build a different understanding of it.

2.1. Tangent Vectors.

We begin this discussion with some very basic stuff: *vectors*. When you first learned of vectors, you were likely taught to visualise them as "arrows" lying in a plane or in space. More specifically, you probably viewed these arrows as "starting from a point" and "pointing in some direction".

On the other hand, for algebraic purposes, one had a different outlook:

Definition 2.1. Let $\mathbb{R}^n$ denote the n-dimensional Euclidean space,

$$\mathbb{R}^n = \{(x_1, x_2, \ldots, x_n) \mid x_1, x_2, \ldots, x_n \in \mathbb{R}\}, \tag{2.1}$$

consisting of all n-tuples of real numbers.

In linear algebra, you saw vectors as elements of the set $\mathbb{R}^n$ (or, more abstractly, as elements of a *vector space*). For example, a 2-dimensional vector would be a pair $\boldsymbol{v} = (v_1, v_2)$ of real numbers, while a 3-dimensional vector would likewise be a triple $\boldsymbol{w} = (w_1, w_2, w_3)$. You could then proceed to define several algebraic operations, such as vector addition and scalar multiplication, on $\mathbb{R}^n$.

Remark 2.2. In practice, the module will only involve 1, 2, and 3-dimensional settings. However, it is often not any more difficult to work in all dimensions.

While $\mathbb{R}^n$ is convenient for algebraic computations, the perspective of treating vectors as arrows contains a number of geometric intuitions that will prove useful. Therefore, we now develop this idea in a more formal manner.

To describe such an arrow, say in $\mathbb{R}^n$, we require two pieces of information:

- The *starting point* $\mathbf{p} \in \mathbb{R}^n$ of the arrow.
- The *vector component* $\boldsymbol{v} \in \mathbb{R}^n$ of the arrow, representing both the direction in which the arrow is pointing and the length of the arrow.

From this, we can craft a mathematical object "$\boldsymbol{v}_\mathbf{p}$" representing an arrow beginning at $\mathbf{p}$ and pointing along $\boldsymbol{v}$; see the first drawing in Figure 2.1. For reasons that will become apparent later, we refer to these arrows as *tangent vectors*:

Definition 2.3. A tangent vector in $\mathbb{R}^n$ is a pair $\boldsymbol{v}_\mathbf{p}$, with $\boldsymbol{v}, \mathbf{p} \in \mathbb{R}^n$, which formally represents the "arrow" in $\mathbb{R}^n$ starting at the point $\mathbf{p}$ and having pointing along $\boldsymbol{v}$.

The graphical interpretation of tangent vectors can be summarised as follows: $\boldsymbol{v}_\mathbf{p}$ *is drawn as an arrow starting at the point* $\mathbf{p}$ *and and terminating at* $\mathbf{p} + \boldsymbol{v}$.

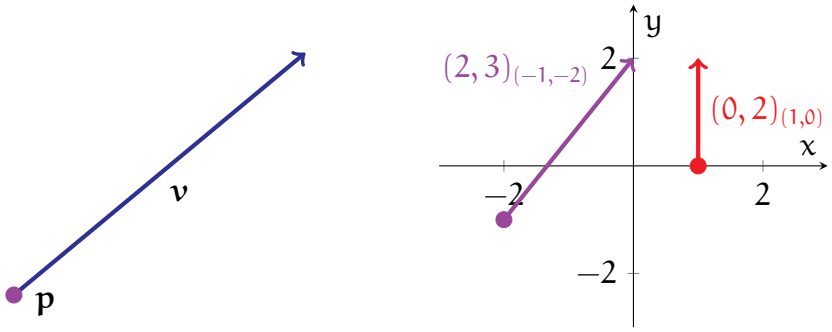


FIGURE 2.1. The left drawing depicts a tangent vector $\boldsymbol{v}_\mathbf{p}$ in $\mathbb{R}^n$. The right drawing shows the tangent vectors from Example 2.4.

Example 2.4. The tangent vectors $(0,2)_{(1,0)}$ and $(2,3)_{(-2,-1)}$ in $\mathbb{R}^2$ are depicted as arrows in red and purple, respectively, in the right drawing of Figure 2.1. In particular, $(2,3)_{(-2,-1)}$ is drawn as an arrow starting at $(-2,-1)$ with its tip at $(-2,-1) + (2,3) = (0,2)$.

Remark 2.5. Given a tangent vector $\boldsymbol{v}_\mathbf{p}$ in $\mathbb{R}^n$, we will generally refer to $\boldsymbol{v}$ as its *vector component*. Later on, we will discuss various interpretations of these objects.

It is often useful to distinguish tangent vectors by their starting points:

Definition 2.6. For any $\mathbf{p} \in \mathbb{R}^n$, the set of all tangent vectors in $\mathbb{R}^n$ starting from $\mathbf{p}$, which we denote by $\mathsf{T}_\mathbf{p}\mathbb{R}^n$, is called the tangent space of $\mathbb{R}^n$ at $\mathbf{p}$:

$$\mathsf{T}_\mathbf{p}\mathbb{R}^n = \{\boldsymbol{v}_\mathbf{p} \mid \boldsymbol{v} \in \mathbb{R}^n\}. \tag{2.2}$$

At each point $\mathbf{p} \in \mathbb{R}^n$, the tangent space $\mathsf{T}_\mathbf{p}\mathbb{R}^n$ *represents a "copy of* $\mathbb{R}^n$ *centred at* $\mathbf{p}$*"*, corresponding to all the arrows one can draw starting at $\mathbf{p}$.

Having gone through the trouble of formally defining tangent vectors, we should now ask what we can *do* with them that would be geometrically meaningful. Let us begin here with the most basic vector operations from linear algebra:

Definition 2.7. Let $\mathbf{p} \in \mathbb{R}^n$. We define the following operations on tangent vectors:

- Given tangent vectors $\mathbf{v}_\mathbf{p}, \mathbf{w}_\mathbf{p} \in \mathsf{T}_\mathbf{p}\mathbb{R}^n$, both starting from $\mathbf{p}$, we define

$$\mathbf{v}_\mathbf{p} + \mathbf{w}_\mathbf{p} = (\mathbf{v} + \mathbf{w})_\mathbf{p}. \tag{2.3}$$

- Given a tangent vector $\mathbf{v}_\mathbf{p} \in \mathsf{T}_\mathbf{p}\mathbb{R}^n$ and a scalar $c \in \mathbb{R}$, we define

$$c \cdot \mathbf{v}_\mathbf{p} = (c \cdot \mathbf{v})_\mathbf{p}. \tag{2.4}$$

Both operations in Definition 2.7 have natural geometric interpretations—in fact, the same ones you have seen since you first learned about vectors. These are demonstrated in the two pictures in Figure 2.2, both of which you should find quite familiar. In particular, the right graphic in Figure 2.2 is the standard parallelogram diagram that is commonly used to describe vector addition.

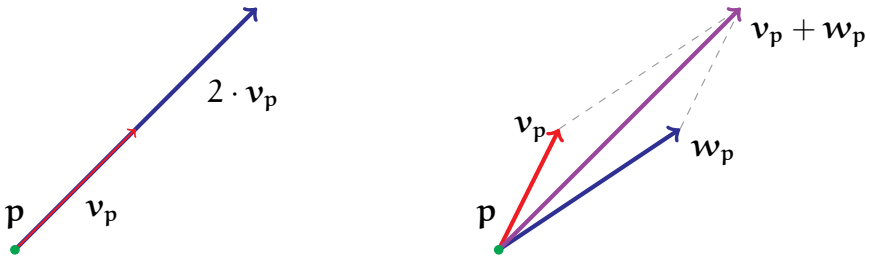


FIGURE 2.2. The left graphic demonstrates the geometric meaning of (2.4), while the right graphic shows the geometric meaning of (2.3).

Example 2.8. Addition and scalar multiplication of tangent vectors are very simple, since one just applies the usual operations to the vector components. The following example uses (2.3) to add two tangent vectors in $\mathbb{R}^2$, both starting from $(-1, 0)$:

$$\begin{aligned}(1,2)_{(-1,0)} + (3,-1)_{(-1,0)} &= [(1,2) + (3,-1)]_{(-1,0)} \\ &= (4,1)_{(-1,0)},\end{aligned}$$

Similarly, the following applies (2.4) to a tangent vector in $\mathbb{R}^3$:

$$\begin{aligned}-2 \cdot (1,2,3)_{(-1,-2,-3)} &= [-2 \cdot (1,2,3)]_{(-1,-2,-3)} \\ &= (-2,-4,-6)_{(-1,-2,-3)}.\end{aligned}$$

On the other hand, according to Definition 2.7, the sum

$$(0,1)_{(0,0)} + (1,0)_{(0,1)}$$

is not defined, since the two tangent vectors are based at different points!

Returning now to the language of linear algebra, one can verify (you should try it yourself!) that $T_p\mathbb{R}^n$ and the two operations in Definition 2.7 satisfy all the axioms of a *vector space*. Thus, the tangent spaces $T_p\mathbb{R}^n$ are concrete examples of the abstract vector spaces that you studied in linear algebra.

Definition 2.9. Given $v = (v_1, \ldots, v_n) \in \mathbb{R}^n$, we define its norm by

$$|v| = \sqrt{v_1^2 + v_2^2 + \cdots + v_n^2}.$$

Moreover, for a tangent vector $v_p \in T_p\mathbb{R}^n$, we define its norm to be

$$|v_p| = |v|. \tag{2.5}$$

In particular, the norm of a tangent vector is just the norm of its vector component. The geometric interpretation of this should be familiar to you:

- The norm $|v_p|$ captures the *length* of the arrow v_p.
- When $|v_p| \neq 0$, the unit tangent vector pointing in the same direction as v_p is

$$\frac{1}{|v_p|} \cdot v_p.$$

Example 2.10. The norm of the tangent vector $(1,2,-1,7)_{(0,0,0,0)}$ in $\mathbb{R}^4$ is

$$\begin{aligned} |(1,2,-1,7)_{(0,0,0,0)}| &= |(1,2,-1,7)| \\ &= \sqrt{55}. \end{aligned}$$

2.2. **Dot and Cross Products.** We now recall a very familiar vector operation—the *dot product*—and we extend it to tangent vectors.

Definition 2.11. Given two n-dimensional vectors,

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

we define their dot product by

$$v \cdot w = v_1 w_1 + v_2 w_2 + \cdots + v_n w_n \in \mathbb{R}. \tag{2.6}$$

Also, for tangent vectors $\boldsymbol{v}_{\mathbf{p}}$ and $\boldsymbol{w}_{\mathbf{p}}$ in $\mathbb{R}^n$, both starting at $\mathbf{p} \in \mathbb{R}^n$, we define

$$\boldsymbol{v}_{\mathbf{p}} \cdot \boldsymbol{w}_{\mathbf{p}} = \boldsymbol{v} \cdot \boldsymbol{w}. \tag{2.7}$$

In particular, (2.6) is the usual formula that you have seen for years. For tangent vectors, their dot product is simply the dot product of their vector components.

Remark 2.12. Note that while $\boldsymbol{v}$ and $\boldsymbol{w}$ are vectors, their dot product $\boldsymbol{v} \cdot \boldsymbol{w}$ is a scalar, not a vector. A similar statement holds for tangent vectors.

Example 2.13. An example of a dot product of tangent vectors in $\mathbb{R}^2$ is given below:

$$\begin{aligned}(1,2)_{(-1,-1)} \cdot (-1,2)_{(-1,-1)} &= (1,2) \cdot (-1,2) \\ &= 1 \cdot (-1) + 2 \cdot 2 \\ &= 3.\end{aligned}$$

On the other hand, according to Definition 2.11, the dot product

$$(1,2)_{(-1,-1)} \cdot (-1,2)_{(-1,0)}$$

is not defined, since the two tangent vectors are based at different points.

While (2.6) and (2.7) are most convenient for computations, another formula for dot products has more illuminating geometric interpretations:

Theorem 2.14. For $\boldsymbol{v}, \boldsymbol{w}, \mathbf{p}$ as in Definition 2.11, we have

$$\boldsymbol{v}_{\mathbf{p}} \cdot \boldsymbol{w}_{\mathbf{p}} = |\boldsymbol{v}_{\mathbf{p}}||\boldsymbol{w}_{\mathbf{p}}| \cos\theta, \tag{2.8}$$

where θ is the angle made between the arrows $\boldsymbol{v}_{\mathbf{p}}$ and $\boldsymbol{w}_{\mathbf{p}}$ at $\mathbf{p}$. In particular,

$$|\boldsymbol{v}_{\mathbf{p}}|^2 = \boldsymbol{v}_{\mathbf{p}} \cdot \boldsymbol{v}_{\mathbf{p}}. \tag{2.9}$$

The formula (2.8) indicates that dot products carry two pieces of information:

- The lengths of the tangent vectors involved.
- The angle between the tangent vectors involved.

In the case of (2.9), as there is only one vector involved (hence θ vanishes), the dot product $\boldsymbol{v}_{\mathbf{p}} \cdot \boldsymbol{v}_{\mathbf{p}}$ carries information only about the length of $\boldsymbol{v}_{\mathbf{p}}$.

Figure 2.3 gives the standard illustration of (2.8). Note that the use of tangent vectors formally realises this drawing, since the angle θ is now manifested as the angle made between two arrows at their common starting point $\mathbf{p}$.

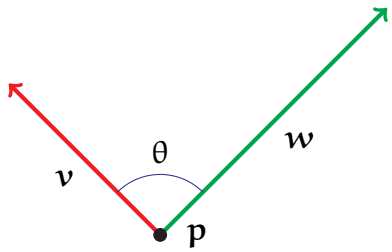


FIGURE 2.3. This drawing demonstrates the setting of (2.8).

Next, we embark on a similar discussion for *cross products*:

Definition 2.15. Given two 3-*dimensional* vectors

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

we define their cross product to be the 3-*dimensional* vector

$$\boldsymbol{v} \times \boldsymbol{w} = (v_2 w_3 - v_3 w_2,\ v_3 w_1 - v_1 w_3,\ v_1 w_2 - v_2 w_1) \in \mathbb{R}^3. \tag{2.10}$$

Moreover, for two tangent vectors $\boldsymbol{v}_p, \boldsymbol{w}_p \in T_p\mathbb{R}^3$, we define

$$\boldsymbol{v}_p \times \boldsymbol{w}_p = (\boldsymbol{v} \times \boldsymbol{w})_p. \tag{2.11}$$

Remark 2.16. An easy way to remember (2.10) is to express it informally as

$$\boldsymbol{v} \times \boldsymbol{w} = \det \begin{bmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ v_1 & v_2 & v_3 \\ w_1 & w_2 & v_3 \end{bmatrix}, \tag{2.12}$$

where $\mathbf{i}$, $\mathbf{j}$, $\mathbf{k}$ denote the canonical basis for $\mathbb{R}^3$:

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

Remark 2.17. Keep in mind that the cross product $\boldsymbol{v} \times \boldsymbol{w}$ of 3-dimensional vectors is itself a 3-dimensional vector, not a scalar. Similarly, the cross product of two tangent vectors in $\mathbb{R}^3$ is itself a tangent vector in $\mathbb{R}^3$, based at the same point.

Also, remember that cross products are specific to 3 dimensions. In particular, (please) *never try to take cross products of 2-dimensional vectors*!

To see why cross products are useful, we recall the following property:

Theorem 2.18. Let $\boldsymbol{v}, \boldsymbol{w}, \mathbf{p}$ be as in Definition 2.15.

- If $\boldsymbol{v}_p$ and $\boldsymbol{w}_p$ point in the same or the opposite directions, then $\boldsymbol{v}_p \times \boldsymbol{w}_p = 0$.

- Otherwise, $v_p \times w_p$ points in the direction that is perpendicular to both v_p and w_p and satisfies the right-hand rule. Furthermore,

$$|v_p \times w_p| = |v_p||w_p| \sin\theta, \tag{2.13}$$

where θ is the angle made between the tangent vectors v_p and w_p at p.

In particular, suppose you already have two vectors that span two of the three dimensions in $\mathbb{R}^3$. Then, their cross product provides a straightforward and computable way to generate the remaining third dimension.

The illustrations in Figure 2.4 demonstrate Theorem 2.18. Again, notice the use of tangent vectors formally realises Figure 2.4, as both the angle θ and the direction of the cross product can be manifested in terms of these arrows.

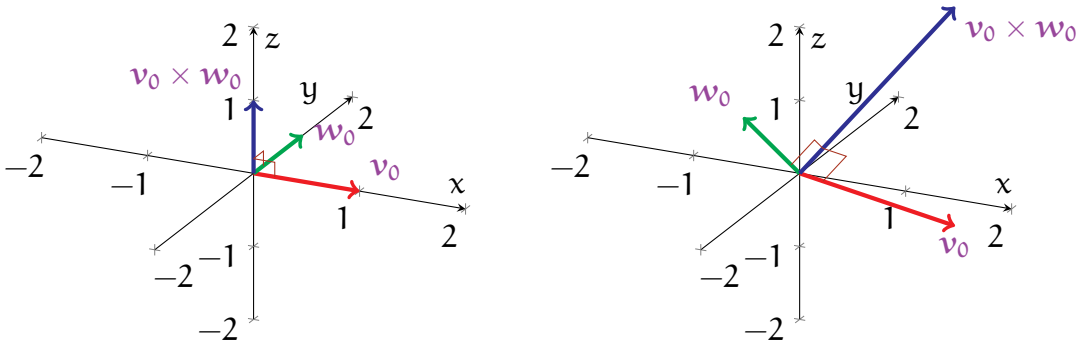


FIGURE 2.4. The diagrams demonstrate the cross products in Example 2.19. In both pictures, the tangent vectors are based at $0 = (0,0,0)$.

Example 2.19. Figure 2.4 contains graphical examples of cross products using tangent vectors based at the origin $0 = (0,0,0)$. The left plot depicts the tangent vectors

$$v_0 = (1,0,0)_0, \qquad w_0 = (0,1,0)_0, \qquad v_0 \times w_0 = (0,0,1)_0,$$

while the right plot depicts the tangent vectors

$$v_0 = (1,1,-1)_0, \qquad w_0 = (-1,1,0)_0, \qquad v_0 \times w_0 = (1,1,2)_0.$$

2.3. **Vector-Valued Functions.** Toying around with individual arrows is not that interesting. To study more complex mathematical objects and model real-world phenomena, we must broaden our horizons to vector-valued *functions*.

Definition 2.20. Consider a general function $f : A \to \mathbb{R}^n$, with A being a subset of $\mathbb{R}^m$.

- A is called the domain of $\mathbf{f}$.
- The image (or range) of $\mathbf{f}$ is the set of all possible values achieved by $\mathbf{f}$:

$$\mathbf{f}(A) = \{\mathbf{f}(\mathbf{x}) \mid \mathbf{x} \in A\}.$$

Moreover, we can view this map $\mathbf{f}$ in a variety of equivalent ways:

- We can think of $\mathbf{f}$ as mapping each vector $\mathbf{x} \in A$ to a vector $\mathbf{f}(\mathbf{x}) \in \mathbb{R}^n$.
- We could also view $\mathbf{f}$ as mapping each $\mathbf{x} \in A$ to n real values,

$$\mathbf{f}(\mathbf{x}) = (f_1(\mathbf{x}), f_2(\mathbf{x}), \dots, f_n(\mathbf{x})) \in \mathbb{R}^n.$$

  In other words, $\mathbf{f}$ splits into n real-valued functions,

$$\mathbf{f} = (f_1, f_2, \dots, f_n), \qquad f_k : A \to \mathbb{R}, \quad 1 \le k \le n.$$

- Similarly, we can split a vector $\mathbf{x} \in A$ into its components, $\mathbf{x} = (x_1, x_2, \dots, x_m)$, so that $\mathbf{f}$ can be viewed as a function of m real numbers:

$$\mathbf{f}(\mathbf{x}) = \mathbf{f}(x_1, \dots, x_m) = (f_1(x_1, \dots, x_m), \dots, f_n(x_1, \dots, x_m)).$$

You should be at least somewhat familiar with the conventions in Definition 2.20 from your previous experiences. In these notes, we will use both vector notations and expanded components interchangeably, depending on context.

In the remainder of this section, we discuss a few concrete examples of vector-valued functions. We begin with functions of one variable ($m = 1$):

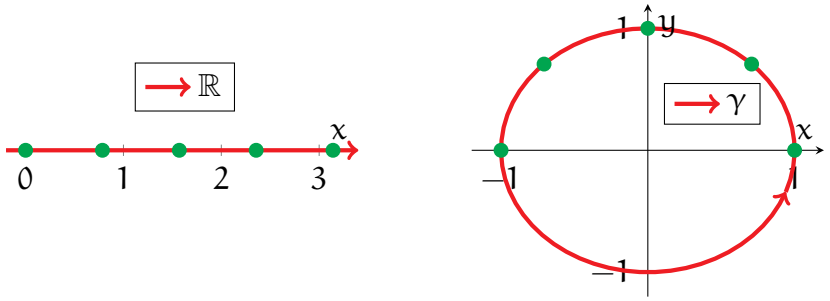


FIGURE 2.5. The red line in the left diagram shows the domain of γ from Example 2.21, while the red circle in the right plot shows the image of γ. The green dots correspond to the values from the table in Example 2.21—in particular, γ maps the green dots in the left plot to those in the right plot.

Example 2.21. Consider first the function

$$\gamma : \mathbb{R} \to \mathbb{R}^2, \qquad \gamma(t) = (\cos t, \sin t). \tag{2.14}$$

Note that γ maps any number t on the real line (i.e. the domain of γ, drawn in the left part of Figure 2.5) to the point $(\cos t, \sin t)$ on the xy-plane.

To better understand the behaviour of γ, let us now plot its image. By default, such a function can be plotted by hand using the following general procedure:

(1) The first step is to compute $\gamma(t)$ for a "large enough" sample of points t. Some specific values of γ are listed in the table below:

t	0	$\frac{\pi}{4}$	$\frac{\pi}{2}$	$\frac{3\pi}{4}$	π
$\gamma(t)$	$(1,0)$	$\left(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right)$	$(0,1)$	$\left(-\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right)$	$(-1,0)$

(2) Next, we plot the values $\gamma(t)$ from step (1) onto the xy-plane. The green dots in the left plot of Figure 2.5 show the values of t in the preceding table, while the green dots in the right plot represent the corresponding values for $\gamma(t)$.

(3) Once you have enough values of γ plotted, you can then try to guess the image of γ by "connecting the dots" in a reasonable manner.

By following the above steps, you should then be able to see that γ *maps out the unit circle about the origin*; this is the red path in the right plot of Figure 2.5.

(All this could also be deduced from (2.14) itself, since $\cos t$ and $\sin t$ are the x and y-coordinates of the point on the unit circle whose angle from the positive x-axis is t.)

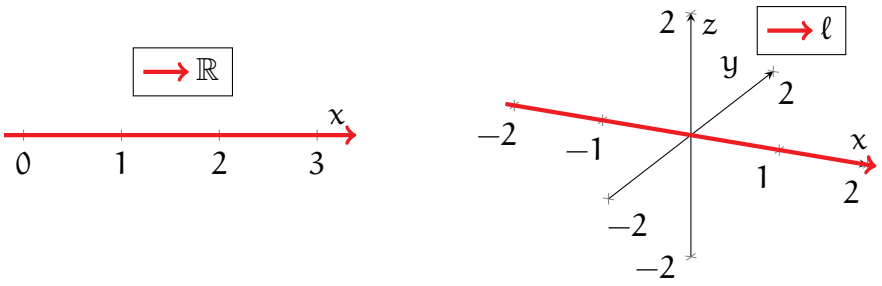


FIGURE 2.6. The left and right diagrams depict the domain and image, respectively, of the function ℓ from Example 2.22.

Example 2.22. For another simple example, this time in 3 dimensions, consider

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

We can plot the image of ℓ in 3-dimensional space in the same manner as in Example 2.21; see the right drawing of Figure 2.6. Note *the image of ℓ is the x-axis*.

In general, plotting single-variable functions onto a plane is rather straightforward, as one can follow the steps discussed in Example 2.21. Plotting functions in $\mathbb{R}^3$ is not any harder in principle, though drawing a 3-dimensional picture on paper can be tricky. Fortunately, you will not be asked to plot functions in 4 or more dimensions!

Finally, we turn our attention to vector-valued functions of multiple variables:

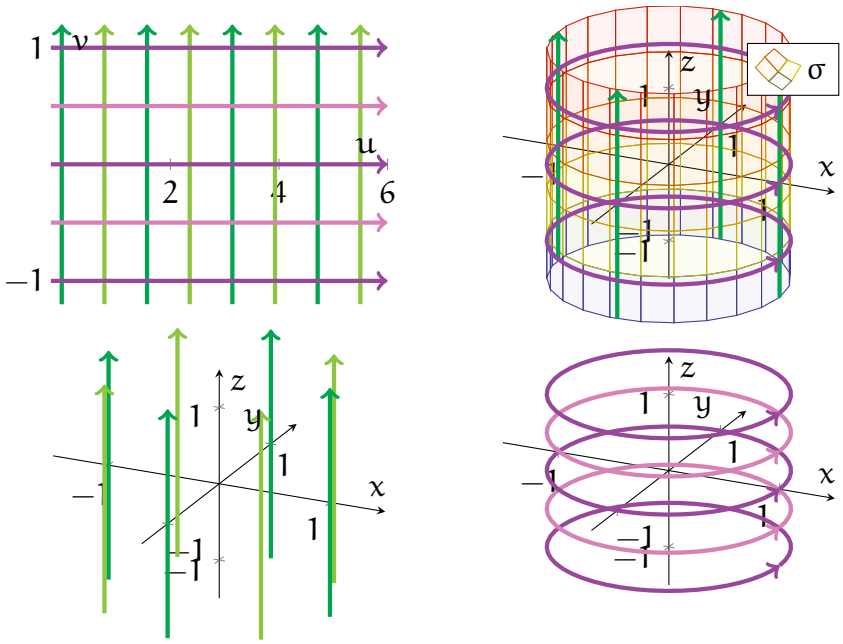


FIGURE 2.7. The top-left plot shows the domain $\mathbb{R}^2$ of σ from Example 2.23; the green lines are obtained by setting $u = 0, \frac{\pi}{4}, \frac{\pi}{2}, \ldots, 2\pi$ and varying v, while the purple lines are obtained by setting $v = -1, -\frac{1}{2}, 0, \frac{1}{2}, 1$ and varying u. The values of σ along these green and purple lines are shown in the bottom-left (constant u, varying v) and bottom-right (constant v, varying u) plots. Finally, the full image of σ—shown in the top-right—can be constructed from the green and purple paths in the bottom plots.

Example 2.23. Consider the function

$$\sigma : \mathbb{R}^2 \to \mathbb{R}^3, \qquad \sigma(u, v) = (\cos u, \sin u, v),$$

which maps two variables $(u, v) \in \mathbb{R}^2$ to a 3-dimensional vector $(\cos u, \sin u, v)$. The domain and image of σ are shown in the top-left and top-right drawings of Figure 2.7.

Again, there is a general process for plotting the image of σ:

(1) First, fix a particular value $u = u_0$ of the first variable in σ, and consider the function $v \mapsto \sigma(u_0, v)$ obtained by varying the second variable v. We can now

plot this function in the same manner as in Examples 2.21 and 2.22. If we repeat this for several values of u, then the resulting images—each of which is a path in 3-dimensional space—will give us a good idea of what the image of σ looks like. This process is shown in the bottom-left plot in Figure 2.7.

(2) A complementary set of paths can be obtained from fixing values of v and varying u instead. This is demonstrated in the bottom-right drawing in Figure 2.7.

(3) From the images obtained in steps (1) and (2), we can make a reasonable guess for what the full image of σ looks like; see the top-right plot in Figure 2.7.

From the above, we conclude that the image of σ is a *cylinder*, centred about the z-axis and having radius 1. Observe that if we fix u and vary v, then we obtain a vertical line of the cylinder, situated at a fixed polar angle. On the other hand, if we fix v and vary u, then we obtain a unit circle from the cylinder, situated at a fixed z-height.

Example 2.23 shows that plotting functions of two variables is also not so complicated in theory. However, the process can be more painstaking, since we may need to experiment by plotting several single-variable functions.

Remark 2.24. There is no need to panic if you are not comfortable with plotting functions yet. We will explore many more examples in upcoming sections!

2.4. **Limits and Continuity.** Let us continue our discussions of vector-valued functions, except we now shift our focus toward some topological concepts.

In mathematics, *topology* refers to the study of *properties preserved under continuous deformations*. To visualise this, imagine bending a piece of wire without breaking it, or stretching a ball of clay without poking a hole in the middle. In doing this, you alter the shapes of the objects, but there are other attributes that stay unchanged.

Before addressing continuity, we must first recall a familiar idea from calculus:

Definition 2.25. Let the vector-valued fuction $\mathbf{f} : A \to \mathbb{R}^n$ be as in Definition 2.20, and let $\mathbf{p} \in A$. We say that $\mathbf{L} \in \mathbb{R}^n$ is the limit of $\mathbf{f}$ at $\mathbf{p}$, denoted

$$\lim_{\mathbf{x}\to\mathbf{p}} \mathbf{f}(\mathbf{x}) = \mathbf{L}, \tag{2.15}$$

iff for every $\varepsilon > 0$, there exists some $\delta > 0$ such that the following statement holds: if $\mathbf{x} \in A$, $\mathbf{x} \neq \mathbf{p}$, and $|\mathbf{x} - \mathbf{p}| < \delta$, then $|\mathbf{f}(\mathbf{x}) - \mathbf{L}| < \varepsilon$.

Definition 2.25 is, in essence, a statement about the behaviour of $\mathbf{f}$ near the point $\mathbf{p}$. Roughly, it says that $\mathbf{f}(\mathbf{x})$ *gets as close as you want to* $\mathbf{L}$*, as long as* $\mathbf{x}$ *is close*

enough to $\mathbf{p}$. Moreover, the symbols "δ" and "ε" of calculus students' nightmares can be interpreted as follows: ε represents how close you want $\mathbf{f}(\mathbf{x})$ to be to $\mathbf{L}$, while δ represents how close $\mathbf{x}$ must be to $\mathbf{p}$ in order to achieve the desired ε-closeness.

Remark 2.26. One can still make sense of limits of $\mathbf{f}$ in Definition 2.25 at some points $\mathbf{p} \notin A$. However, we will not need this extended definition here.

The following shows how limits of vector-valued functions can be computed:

Theorem 2.27. Let $\mathbf{f} = (f_1, f_2, \ldots, f_n)$ be as in Definition 2.20, and let $\mathbf{p} \in A$. Then,

$$\lim_{\mathbf{x}\to\mathbf{p}} \mathbf{f}(\mathbf{x}) = \mathbf{L},$$

where $\mathbf{L} = (L_1, L_2, \ldots, L_n) \in \mathbb{R}^n$, if and only if for every $1 \leq k \leq n$,

$$\lim_{\mathbf{x}\to\mathbf{p}} f_k(\mathbf{x}) = L_k.$$

Proof. We avoid a formal proof here, as this would require some background from *MTH5104: Convergence and Continuity*. However, the key idea is the identity

$$|\mathbf{f}(\mathbf{x}) - \mathbf{L}|^2 = \sum_{k=1}^{n} |f_k(\mathbf{x}) - L_k|^2.$$

From the above, one sees that $\mathbf{f}(\mathbf{x})$ becomes arbitrarily close to $\mathbf{L}$ if and only if $f_k(\mathbf{x})$ becomes arbitrarily close to L_k for each $1 \leq k \leq n$. □

Theorem 2.27 implies that *limits of vector-valued functions can be taken componentwise*. In other words, to compute the limit of a vector-valued function $\mathbf{f}$, we need only compute the limits of the individual real-valued components of $\mathbf{f}$.

For completeness, we consider a simple example:

Example 2.28. Let us compute the limit, at $t_0 = -1$, of the function

$$\mathbf{h} : \mathbb{R} \to \mathbb{R}^3, \qquad h(t) = (t, t^2, t^3).$$

Applying Theorem 2.27, with $\mathbf{f} = \mathbf{h}$, we see that

$$\begin{aligned}\lim_{t\to-1} \mathbf{h}(t) &= \left(\lim_{t\to-1} t,\ \lim_{t\to-1} t^2,\ \lim_{t\to-1} t^3\right) \\ &= (-1, 1, -1),\end{aligned}$$

where in the final step, the individual real-valued limits can be evaluated directly.

Next, we apply limits to define continuous functions:

Definition 2.29. Let $\mathbf{f}$ be as in Definition 2.20. We say $\mathbf{f}$ is continuous at $\mathbf{p} \in A$ iff

$$\lim_{\mathbf{x}\to\mathbf{p}} \mathbf{f}(\mathbf{x}) = \mathbf{f}(\mathbf{p}). \tag{2.16}$$

In addition, we say $\mathbf{f}$ is continuous iff $\mathbf{f}$ is continuous at every $\mathbf{p} \in A$.

In particular, (2.16) says that *the values of* $\mathbf{f}$ *near the point* $\mathbf{p}$ *are close to* $\mathbf{f}(\mathbf{p})$ *itself*. In other words, the values of $\mathbf{f}$ do not "break", or "jump", while at $\mathbf{p}$.

Example 2.30. The function from Example 2.21,

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

is continuous; see Figure 2.5 for plots of its domain and image.

One can visualise this γ as taking an infinitely long wire (the left half of Figure 2.5) and bending it into a circular shape (the right half of Figure 2.5). That this γ is continuous can be seen from the fact that this bending does not break the wire at any point.

Example 2.31. Another example of a continuous function is that of Example 2.23,

$$\sigma : \mathbb{R}^2 \to \mathbb{R}^3, \qquad \sigma(u, v) = (\cos u, \sin u, v);$$

see Figure 2.7 for plots of the domain and image of σ.

One can visualise σ as taking a piece of paper (the top-left drawing in Figure 2.7) and rolling into up into a cylinder (the top-right drawing of Figure 2.7), though both the paper and the cylinder have infinite dimensions here. Again, σ is continuous because this "rolling up" bends the paper but does not tear it at any point.

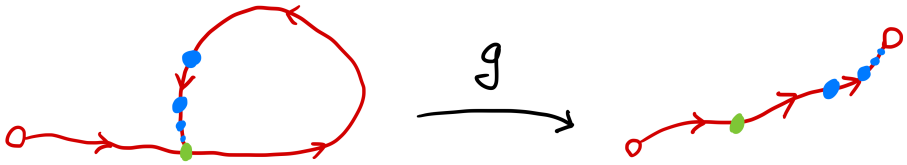


FIGURE 2.8. The above illustrates the function $\mathbf{g}$ from Example 2.32. The left and right drawings represent the domain and image of $\mathbf{g}$, respectively. Notice that $\mathbf{g}$ fails to be continuous at the green point.

Example 2.32. For an example of something discontinuous, we consider the function $\mathbf{g}$ roughly drawn in Figure 2.8. (The exact definition of $\mathbf{g}$ is not important here.)

Here, $\mathbf{g}$ maps the lasso-shaped red path in the left half of Figure 2.8 (its domain) to the path on the right half of this figure (its image). Moreover, the blue and green points in the domain are mapped to the blue and green points on the image, respectively.

To see why $\mathbf{g}$ fails to be continuous, note that in the domain of $\mathbf{g}$, the blue points converge toward the green point $\mathbf{p}$. On the other hand, the values of $\mathbf{g}$ at these blue points—that is, the blue points in the right drawing—do not tend to the green point $\mathbf{g}(\mathbf{p})$ in the right drawing. As a result, Definition 2.29 is violated at $\mathbf{p}$.

Less formally, one can view the function $\mathbf{g}$ as bending the left drawing in Figure 2.8 into the right drawing. However, to do this, one must also "cut" this left figure at the green point, hence making this deformation discontinuous.

Remark 2.33. We will not deal formally with limits and continuity in this module. However, these will be of some importance in some later conceptual discussions.

2.5. **Derivatives.** The next step of our revisions is to discuss aspects of differential calculus, which is primarily concerned with studying how a function is changing. In this section, we begin with a very familiar concept:

Definition 2.34. Let $I = (a, b)$ be an open interval, and let $\mathbf{f} : I \to \mathbb{R}^n$ be a vector-valued function (of one variable). Then, the derivative of $\mathbf{f}$ at $t_0 \in I$ is defined to be

$$\mathbf{f}'(t_0) = \lim_{t \to t_0} \frac{\mathbf{f}(t) - \mathbf{f}(t_0)}{t - t_0}, \tag{2.17}$$

whenever the limit on the right-hand side of (2.17) exists.

Moreover, whenever $\mathbf{f}'(t_0)$ exists for every $t_0 \in I$, we then define the derivative of $\mathbf{f}$ itself to be the function $\mathbf{f}' : I \to \mathbb{R}^n$ that maps each $t_0 \in I$ to $\mathbf{f}'(t_0)$.

One common interpretation for a function $\mathbf{f}$ as in Definition 2.34 comes from classical mechanics. For $t \in \mathbb{R}$, *the value* $\mathbf{f}(t) \in \mathbb{R}^n$ *can represent the position of a particle in* n*-dimensional space at time* t. (To be realistic, you can assume $n \leq 3$.) Then, $\mathbf{f}$ *itself would describe the total trajectory of this particle for all time.*

Now, if $t \in I$ as well, then the vector $\mathbf{f}(t) - \mathbf{f}(t_0)$ yields the *displacement*, or the total change in position, between the times t and t_0. Moreover, observe that $t - t_0$ is the total *time elapsed* during this displacement.

As a result, the quotient

$$\frac{\mathbf{f}(t) - \mathbf{f}(t_0)}{t - t_0} \tag{2.18}$$

describes the *average rate of change in position over the time interval* $[t_0, t]$. If we let t tend to t_0, so that the elapsed time $t - t_0$ tends to 0, then (2.18) becomes $\mathbf{f}'(t_0)$, describing the *instantaneous rate of change in position at the single time* t_0.

Remark 2.35. In physics, $f'(t_0)$ is known as the *velocity* of the particle represented by f at time t_0. If we take the norm $|f'(t_0)|$ of the velocity, thereby stripping away the directional information, we then obtain the *speed* of this particle.

In particular, at time t_0, the particle is located at $\mathbf{f}(t_0)$ and moving in the direction $\mathbf{f}'(t_0)$. A natural way to consolidate this information, both conceptually and visually, is as the arrow $\mathbf{f}'(t_0)_{\mathbf{f}(t_0)}$. This is illustrated in Figure 2.9 for a concrete $\mathbf{f}$; the point $\mathbf{f}(t_0)$ is marked in green, while the arrow $\mathbf{f}'(t_0)_{\mathbf{f}(t_0)}$ is drawn in orange.

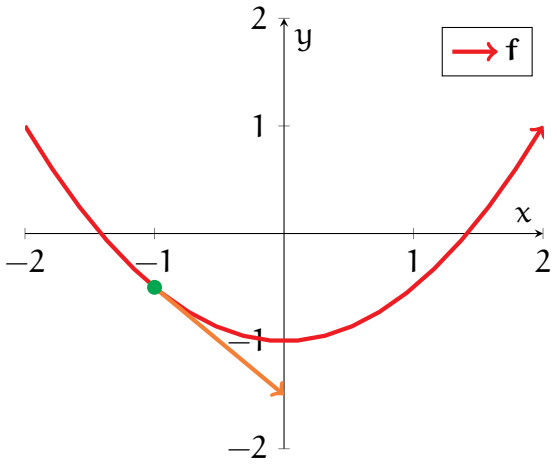


FIGURE 2.9. The vector-valued function $\mathbf{f}(t) = (t, \frac{1}{2}t^2 - 1)$ is drawn as a red path. The point $\mathbf{f}(-1) = (-1, -\frac{1}{2})$ is indicated in green, while the tangent vector $\mathbf{f}'(-1)_{\mathbf{f}(-1)} = (1, -1)_{(-1,-\frac{1}{2})}$ is drawn as an orange arrow.

Let us now recall how derivatives are computed in practice. First, when $n = 1$ (i.e. $\mathbf{f}$ is real-valued), (2.17) is the usual definition from first-year calculus.

Example 2.36. Consider the function

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

Then, at any point $t \in \mathbb{R}$, we can apply the usual calculus methods to compute

$$f'(t) = 6t^2 + \cos t + 2e^{2t}.$$

In particular, at $t = 0$ and $t = 1$, we have

$$f'(0) = 3, \qquad f'(1) = 6 + \cos 1 + 2e^2.$$

Next, for vector-valued functions of a single variable, one can compute its derivative by taking the usual derivative of each component:

Theorem 2.37. Let $\mathbf{f}$ be as in Definition 2.34, and write $\mathbf{f}$ in terms of its components:

$$\mathbf{f} = (f_1, f_2, \dots, f_n), \qquad f_k : I \to \mathbb{R}, \quad 1 \le k \le n.$$

Then, for any $t_0 \in I$, we have that

$$\mathbf{f}'(t_0) = (f_1'(t_0), f_2'(t_0), \dots, f_n'(t_0)), \tag{2.19}$$

as long as each of $f_1'(t_0), \dots, f_n'(t_0)$ exists.

Proof. By the definitions of vector addition and scalar multiplication, we have

$$\frac{\mathbf{f}(t) - \mathbf{f}(t_0)}{t - t_0} = \left(\frac{f_1(t) - f_1(t_0)}{t - t_0}, \dots, \frac{f_n(t) - f_n(t_0)}{t - t_0} \right).$$

Also, by Theorem 2.27, the limits of the above are also taken componentwise,

$$\begin{aligned} \lim_{t \to t_0} \frac{\mathbf{f}(t) - \mathbf{f}(t_0)}{t - t_0} &= \left(\lim_{t \to t_0} \frac{f_1(t) - f_1(t_0)}{t - t_0}, \dots, \lim_{t \to t_0} \frac{f_n(t) - f_n(t_0)}{t - t_0} \right) \\ &= (f_1'(t_0), \dots, f_n'(t_0)), \end{aligned}$$

which yields the desired formula (2.19). □

Thus, finding derivatives of vector-valued functions is not any more difficult than before; one just has to apply the usual steps from first-year calculus multiple times.

Example 2.38. Consider the function γ from Example 2.21,

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

and let us compute its derivative at any $t \in \mathbb{R}$. By Theorem 2.37, we need only compute the derivatives of the components $\cos t$ and $\sin t$ of γ:

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

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

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

The image of γ is drawn (in red) in the left part of Figure 2.10. The values of $\gamma(t)$ in the above table are drawn in green; the tangent vectors $\gamma'(t)_{\gamma(t)}$ are drawn in blue.

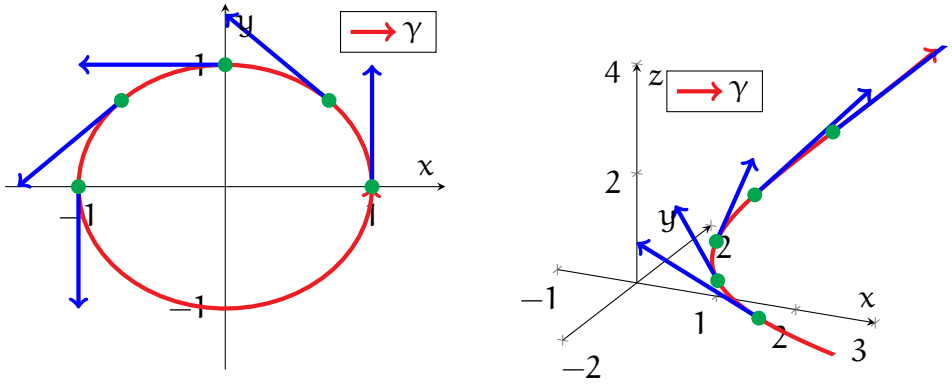


FIGURE 2.10. The left plot contains the image of γ from Example 2.38, while the right plot contains the image of γ from Example 2.39 (both in red). Both plots contain some sample values of $\gamma(t)$ (in green), as well as the corresponding tangent vectors $\gamma'(t)_{\gamma(t)}$ (in blue).

Example 2.39. Consider next the function

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

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

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

Some specific values of γ and γ' are given in the following table:

t	-1	$-\frac{1}{2}$	0	$\frac{1}{2}$	1
$\gamma(t)$	$(2, -1, e^{-1})$	$\left(\frac{5}{4}, -\frac{1}{2}, e^{-\frac{1}{2}}\right)$	$(1, 0, 1)$	$\left(\frac{5}{4}, \frac{1}{2}, e^{\frac{1}{2}}\right)$	$(2, 1, e)$
$\gamma'(t)$	$(-2, 1, e^{-1})$	$\left(-1, 1, e^{-\frac{1}{2}}\right)$	$(0, 1, 1)$	$\left(1, 1, e^{\frac{1}{2}}\right)$	$(2, 1, e)$

A plot of γ, along with the points $\gamma(t)$ and the arrows $\gamma'(t)_{\gamma(t)}$ from the above table, can be found in the right drawing of Figure 2.10.

Example 2.40. For a more abstract example, let us fix $\mathbf{p}, \mathbf{v} \in \mathbb{R}^n$ and define

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

Note first that $\ell(0) = \mathbf{p}$, so the image of ℓ passes through the point $\mathbf{p}$.

Next, writing out $\mathbf{p}$ and $\mathbf{v}$ in terms of their components, we have that

$$\ell(t) = (p_1 + tv_1, \ldots, p_n + tv_n), \qquad \ell'(t) = (v_1, \ldots, v_n) = \mathbf{v},$$

since the only non-constant quantity in the above is t. Thus, we see that ℓ is always moving in the (constant) direction $\mathbf{v}$. Combining all the above, we conclude that ℓ *maps out the line in* $\mathbb{R}^n$ *through* $\mathbf{p}$ *and along the direction* $\mathbf{v}$.

As the above examples 2.38 might suggest, vector-valued functions of a single variable will form the foundations for our study of curves later on in this module.

2.6. **Open and Connected Sets.** We now turn our attention to functions of multiple variables. Before discussing how such functions can be differentiated, let us first pose and answer the following more foundational question:

Question 2.41. Consider a function $\mathbf{f} : U \to \mathbb{R}^n$, where U is a subset of $\mathbb{R}^m$. What assumptions do we need on U in order to make sense of derivatives of $\mathbf{f}$?

Suppose we wish to differentiate $\mathbf{f}$ at a point $\mathbf{p} \in U$. As you know, finding derivatives corresponds to measuring how $\mathbf{f}$ is changing at $\mathbf{p}$. However, to understand this, we must measure how $\mathbf{f}$ behaves near $\mathbf{p}$. Therefore, *it makes sense to differentiate* $\mathbf{f}$ *at* $\mathbf{p}$ *only when* $\mathbf{f}$ *is also defined at points nearby* $\mathbf{p}$.

In other words, the crucial property needed for the domain U of $\mathbf{f}$ is the following: *whenever* $\mathbf{p} \in U$, *and whenever* $\mathbf{q}$ *is "sufficiently close" to* $\mathbf{p}$, *then* $\mathbf{q}$ *must also be in* U. This property is captured mathematically via the concept of *openness*:

Definition 2.42. A subset U of $\mathbb{R}^m$ is open iff for any $\mathbf{p} \in U$, there exists some $\delta > 0$ such that if $\mathbf{q} \in \mathbb{R}^n$ satisfies $|\mathbf{q} - \mathbf{p}| < \delta$, then $\mathbf{q} \in U$ as well.

The formal Definition 2.42, which could seem intimidating at first, can be directly connected to the prior discussion. In particular, δ corresponds to a "small enough distance" from $\mathbf{p}$, while the points $\mathbf{q}$ represent those that are "sufficiently close" to $\mathbf{p}$, measured in terms of our threshold δ. Putting this all together, Definition 2.42 states that *if you start from a point* $\mathbf{p} \in U$, *and you take a sufficiently small step away from* $\mathbf{p}$ *(of distance less than* δ*), then you will not leave the set* U.

One can also think of an open subset $U \subseteq \mathbb{R}^m$ *as one that does not contain any of its boundary, or edge, points.* Intuitively, if $\mathbf{p} \in U$ lies on the edge of U, then one can take as small a step as one wishes in some direction and end up outside of U.

Example 2.43. *Any open interval* $U = (a, b)$ *is an open subset of* $\mathbb{R}$. Indeed, given any $p \in (a, b)$, then points "near p" also lie within (a, b). This is shown in the left drawing of Figure 2.11; here, a point p is drawn in red, while points near p are shown as a dotted yellow region. (More formally, we can take $\delta = \min(p - a, b - p)$ in Definition 2.42.)

On the other hand, *a closed interval* $A = [a, b]$ *is not an open subset of* $\mathbb{R}$. To show this, we consider the boundary point $a \in A$; see the right drawing in Figure 2.11. Note any point to the left of a does not lie in A, no matter how close it is to a. In other words, the dotted green region in Figure 2.11, representing points "close to a", must contain points that do not lie in A. Thus, A violates the conditions of Definition 2.42.

FIGURE 2.11. The drawings show the open interval U (left) and the closed interval A (right) from Example 2.43. In the left picture, given a point $p \in U$ (in red), one can find a neighbourhood of p (e.g. the dotted yellow region) that is contained in U. On the other hand, in the right graphic, for the point a at the boundary of A, any neighbourhood of a (such as the dotted green region) must contain a point that is not in A.

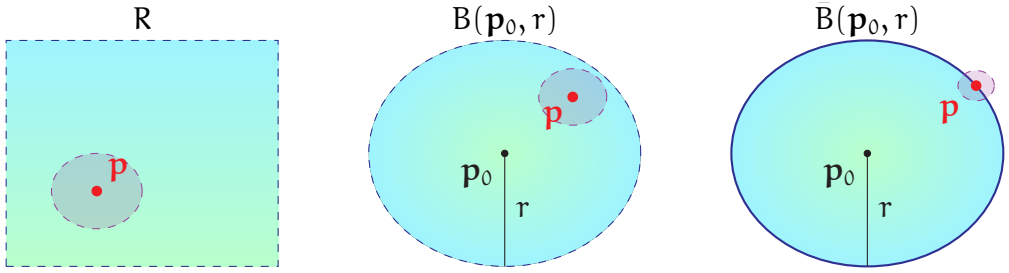


FIGURE 2.12. The left graphic depicts the open rectangle R (in blue and green) from Example 2.44(2); given some $\mathbf{p} \in R$ (drawn in red), one can find a disk around p (in grey) that is contained in U. Similarly, the middle graphic shows the open disk from Example 2.44(3). The drawing on the right shows a closed disk, which is not open; in particular, the point $\mathbf{p}$ (in red) on the boundary of $\bar{B}(\mathbf{p}_0, r)$ violates Definition 2.42.

Example 2.44. The following subsets of $\mathbb{R}^2$ are open:

(1) The *unit open disk* about the origin:

$$B(0, 1) = \{(x, y) \in \mathbb{R}^2 \mid x^2 + y^2 < 1\}.$$

(2) Any *open rectangle*:

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

(3) Any *open disk*, with centre $\mathbf{p}_0 \in \mathbb{R}^2$ and radius $r > 0$:

$$B(\mathbf{p}_0, r) = \{\mathbf{p} \in \mathbb{R}^2 \mid |\mathbf{p} - \mathbf{p}_0| < r\}.$$

On the other hand, every *closed disk*,

$$\bar{B}(\mathbf{p}_0, r) = \{\mathbf{p} \in \mathbb{R}^2 \mid |\mathbf{p} - \mathbf{p}_0| \leq r\},$$

is not an open subset of $\mathbb{R}^2$. See Figure 2.12 for illustrations.

Example 2.45. The ideas in Example 2.44 generalise to all dimensions. For instance, given any centre $\mathbf{p}_0 \in \mathbb{R}^m$ and radius $r > 0$, we can define the *open* and *closed balls*

$$B(\mathbf{p}_0, r) = \{\mathbf{p} \in \mathbb{R}^m \mid |\mathbf{p} - \mathbf{p}_0| < r\}, \qquad \bar{B}(\mathbf{p}_0, r) = \{\mathbf{p} \in \mathbb{R}^m \mid |\mathbf{p} - \mathbf{p}_0| \leq r\},$$

respectively. Again, $B(\mathbf{p}_0, r)$ is an open subset of $\mathbb{R}^m$, while $\bar{B}(\mathbf{p}_0, r)$ fails to be open.

There is another requirement, in addition to openness, that we can adopt to simplify matters. Notice that in the definition of the derivative (Definition 2.34), the domain I of the function was assumed to be an open *interval*. While Example 2.43 showed that I is an open subset of $\mathbb{R}$, one also has another observation:

- I *is connected*—more specifically, given any pair of points on I, the line segment connecting these two points also lies within I.

The reason for wanting connectedness is mainly a matter of convenience. Suppose instead that I is a union of disjoint open intervals, for example,

$$I = (-2, -1) \cup (1, 2);$$

see Figure 2.13 below. In particular, I is still open, but it is no longer connected. In this case, it is more sensible to treat a function $\mathbf{f} : I \to \mathbb{R}^n$ as two separate maps,

$$\mathbf{f}_1 : (-2, -1) \to \mathbb{R}^n, \qquad \mathbf{f}_2 : (1, 2) \to \mathbb{R}^n$$

one for each connected open interval within I. As a result of this, we can simply assume that our domain I is connected without losing any generality.

FIGURE 2.13. The set $I = (-2, -1) \cup (1, 2)$, shown in the above drawing, is an open subset of $\mathbb{R}$ but is not connected.

For the same reason, we can also restrict our attention to "connected" domains in higher dimensions. However, one point to keep in mind is that in higher dimensions, one can join two points in many more ways than just a line segment. As a result, we need the following more general definition of connectedness:

Definition 2.46. An open subset $U \subseteq \mathbb{R}^m$ is said to be connected iff for any two points $\mathbf{p}, \mathbf{q} \in U$, there is a path, lying entirely in U, going from $\mathbf{p}$ to $\mathbf{q}$.

Remark 2.47. A more formal description of a "path lying in U from $\mathbf{p}$ to $\mathbf{q}$" is as a continuous function $\alpha : [0, 1] \to U$ satisfying $\alpha(0) = \mathbf{p}$ and $\alpha(1) = \mathbf{q}$. However, for this module, we will not need to deal with connectness at such a formal level.

Example 2.48. Consider the three open subsets of $\mathbb{R}^2$ drawn in Figure 2.14.

(1) The disk U in the left drawing of Figure 2.14 is connected. To see this, we note that given two points $\mathbf{p}, \mathbf{q} \in U$, the line segment connecting $\mathbf{p}$ and $\mathbf{q}$ also lies within U; see, for instance, the red segment in the drawing of U.

(2) Next, consider the annulus A in the centre drawing of Figure 2.14. In contrast to the disk, the line segment between two points $\mathbf{p}, \mathbf{q} \in A$ needs not lie entirely in A. On the other hand, $\mathbf{p}$ and $\mathbf{q}$ can always be joined by a circular arc (drawn in red), hence A is connected by Definition 2.46.

(3) Finally, the set W in the right drawing of Figure 2.14 fails to be connected. Indeed, given a point $\mathbf{q}$ on the inner disk and a point $\mathbf{p}$ on the outer annulus, there is no way to join $\mathbf{p}$ to $\mathbf{q}$ using a path that lies entirely within W.

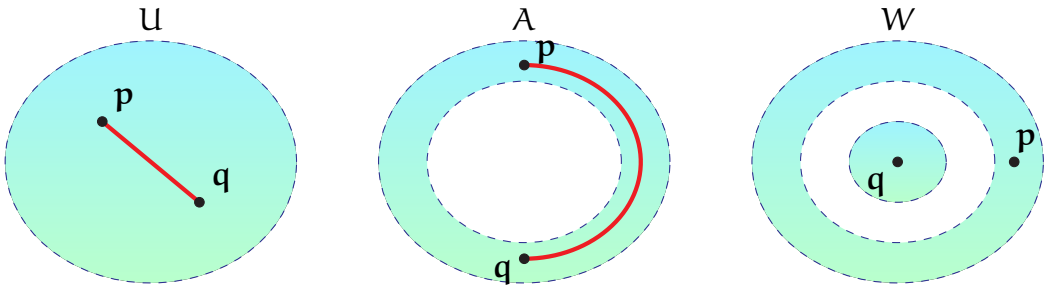


FIGURE 2.14. The figures depict a disk U (left), an annulus A (middle), and a region W with a disk enclosed by an annulus (right). As noted in Example 2.48, both U and A are connected, while W is not connected.

Remark 2.49. The notion of connected and open subsets can be viewed as a generalisation of open intervals to higher dimensions. While we do not give details here, one can prove, in fact, that *the only connected and open subsets of* $\mathbb{R}$ *are open intervals*.

2.7. **Partial Derivatives.** The preceding discussions yield a reasonable answer to Question 2.41: *to take derivatives of* $\mathbf{f} : U \to \mathbb{R}^n$, *we want* U *to be both open and connected*. With this in mind, we now recall how such derivatives are defined:

Definition 2.50. Let $U \subseteq \mathbb{R}^m$ be both open and connected, let $\mathbf{f} : U \to \mathbb{R}^n$, and let $\mathbf{p} = (p_1, \ldots, p_m) \in U$. Then, given any $1 \leq k \leq m$, we define the partial derivative with respect to the k-th variable of $\mathbf{f}$ at $\mathbf{p}$ to be

$$\partial_k \mathbf{f}(\mathbf{p}) = \lim_{q \to p_k} \frac{\mathbf{f}(p_1, \ldots, q, \ldots, p_m) - \mathbf{f}(p_1, \ldots, p_k, \ldots, p_m)}{q - p_k}, \tag{2.20}$$

whenever the limit on the right-hand side of (2.20) exists.

Moreover, whenever $\partial_k \mathbf{f}(\mathbf{p})$ exists for every $\mathbf{p} \in U$, we define the corresponding partial derivative of $\mathbf{f}$ to be the function $\partial_k \mathbf{f} : U \to \mathbb{R}^n$ mapping each $\mathbf{p} \in U$ to $\partial_k \mathbf{f}(\mathbf{p})$.

Remark 2.51. Note that when $m = 1$ in Definition 2.50, the partial derivative $\partial_1 \mathbf{f}(\mathbf{p})$ is the same as the ordinary derivative $\mathbf{f}'(\mathbf{p})$ from Definition 2.34.

First, when $n = 1$ in Definition 2.50 (that is, $\mathbf{f}$ is real-valued), the partial derivatives of $\mathbf{f}$ can be computed using the methods you previously learned in calculus:

Example 2.52. Consider the real-valued function

$$h : \mathbb{R}^2 \to \mathbb{R}, \qquad h(u, v) = u + v^2 + u^3 v.$$

To compute the partial derivative $\partial_1 h$, we treat the second variable v as a constant, and we differentiate as usual (see Example 2.36) with respect to the first variable u:

$$\partial_1 h(u, v) = 1 + 3u^2 v.$$

Similarly, to compute $\partial_2 h$, we hold u constant and differentiate with respect to v:

$$\partial_2 h(u, v) = 2v + u^3.$$

Furthermore, at $(u, v) = (1, 1)$, the partial derivatives of h evaluate to

$$\partial_1 h(1, 1) = 4, \qquad \partial_2 h(1, 1) = 3.$$

Next, we turn our attention to general vector-valued functions:

Theorem 2.53. Let $\mathbf{f}$ be as in Definition 2.50, and write $\mathbf{f}$ in terms of its components:

$$\mathbf{f} = (f_1, f_2, \ldots, f_n), \qquad f_l : U \to \mathbb{R}, \quad 1 \leq l \leq n.$$

Then, for any $1 \leq k \leq m$ and $\mathbf{p} \in U$, we have that

$$\partial_k \mathbf{f}(\mathbf{p}) = (\partial_k f_1(\mathbf{p}), \partial_k f_2(\mathbf{p}), \ldots, \partial_k f_n(\mathbf{p})), \tag{2.21}$$

as long as each of $\partial_k f_1(\mathbf{p}), \ldots, \partial_k f_n(\mathbf{p})$ exists.

Proof. This is analogous to the proof of Theorem 2.37. We again use that vector addition, scalar multiplication, and limits are all evaluated componentwise. □

Theorem 2.53 shows there is nothing novel in taking partial derivatives of a vector-valued function $\mathbf{f}$—you simply apply the usual tricks to each component of $\mathbf{f}$.

Example 2.54. Consider the function

$$\mathbf{f} : \mathbb{R}^2 \to \mathbb{R}^3, \qquad \mathbf{f}(u, v) = (uv, u + v, \sin u + \cos v).$$

To compute $\partial_1 \mathbf{f}$, we simply take this partial derivative of each component of $\mathbf{f}$:

$$\begin{aligned} \partial_1 \mathbf{f}(u, v) &= \left(\frac{\partial}{\partial u}(uv), \frac{\partial}{\partial u}(u + v), \frac{\partial}{\partial u}(\sin u + \cos v) \right) \\ &= (v, 1, \cos u). \end{aligned}$$

A similar computation yields $\partial_2 \mathbf{f}$, for each $(u, v) \in \mathbb{R}^2$:

$$\begin{aligned} \partial_2 \mathbf{f}(u, v) &= \left(\frac{\partial}{\partial v}(uv), \frac{\partial}{\partial v}(u + v), \frac{\partial}{\partial v}(\sin u + \cos v) \right) \\ &= (u, 1, -\sin v). \end{aligned}$$

Let us now discuss now partial derivatives can be interpreted. Consider a function $\mathbf{g} : V \to \mathbb{R}^n$, with V an open subset of $\mathbb{R}^2$ (i.e. $\mathbf{g}$ is a function of two variables).

Fix a point $(u_0, v_0) \in V$, and let γ denote the function given by

$$\gamma(u) = \mathbf{g}(u, v_0),$$

that is, we fix $v = v_0$ and vary only u in $\mathbf{g}$. Then, the derivative of γ at u_0 satisfies

$$\begin{aligned} \gamma'(u_0) &= \lim_{u \to u_0} \frac{\mathbf{g}(u, v_0) - \mathbf{g}(u_0, v_0)}{u - u_0} \\ &= \partial_1 \mathbf{g}(u_0, v_0), \end{aligned}$$

and it follows that

$$\partial_1 \mathbf{g}(u_0, v_0)_{\mathbf{g}(u_0, v_0)} = \gamma'(u_0)_{\gamma(u_0)}.$$

Recall that the tangent vector $\gamma'(u_0)_{\gamma(u_0)}$ is the arrow, starting from $\gamma(u_0)$, that indicates the direction and speed in which γ is changing while at parameter $u = u_0$. Thus, from the definition of γ, we conclude: *the tangent vector* $\partial_1 \mathbf{g}(u_0, v_0)_{\mathbf{g}(u_0, v_0)}$ *represents the direction and speed at which* $\mathbf{g}$ *is changing while at* $(u, v) = (u_0, v_0)$*, and while* v *is held constant and* u *is varied.*

Similarly, if we define

$$\lambda(v) = g(u_0, v),$$

then its derivative satisfies

$$\begin{aligned}\lambda'(v_0) &= \lim_{v \to v_0} \frac{g(u_0, v) - g(u_0, v_0)}{v - v_0} \\ &= \partial_2 g(u_0, v_0).\end{aligned}$$

Thus, similar to the previous case, *the arrow* $\partial_2 g(u_0, v_0)_{g(u_0,v_0)}$ *captures how* g *is changing while at* (u_0, v_0)*, and while* u *is held constant and* v *is varied.*

For a graphical demonstration of these curves γ and λ, see Figure 2.15 below.

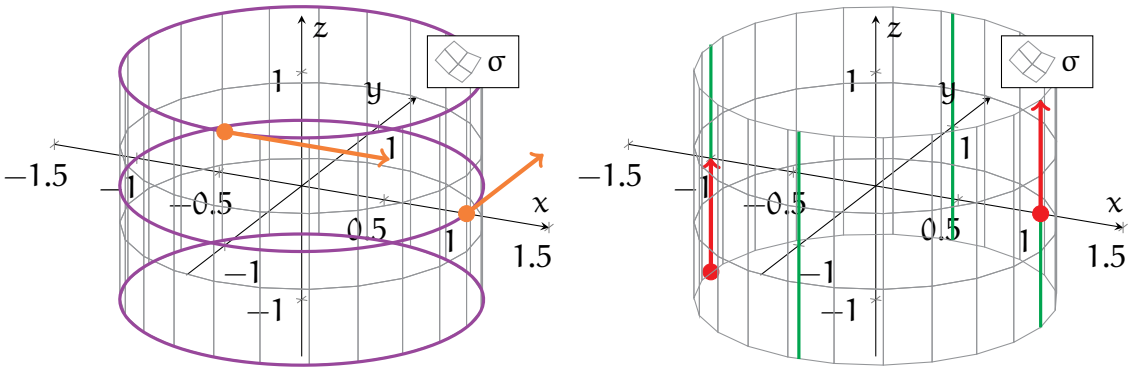


FIGURE 2.15. The plots show σ from Example 2.55. In the left diagram, the purple curves depict paths γ obtained by holding v constant, while the orange arrows depict tangent vectors of the form $\partial_1\sigma(u, v)_{\sigma(u,v)}$. In the right diagram, the green curves depict paths λ obtained by holding u constant, while the red arrows depict tangent vectors of the form $\partial_2\sigma(u, v)_{\sigma(u,v)}$.

Example 2.55. Let us return to the *cylinder* from Example 2.23:

$$\sigma : \mathbb{R}^2 \to \mathbb{R}^3, \qquad \sigma(u, v) = (\cos u, \sin u, v). \tag{2.22}$$

See Figure 2.15 (and Figure 2.7) for illustrations of the image of σ. The purple circles in the left drawing are obtained by holding v constant and varying u in (2.22); the green lines in the right drawing are obtained by holding u constant and varying v in (2.22).

We can compute the partial derivatives of σ using Theorem 2.53:

$$\partial_1\sigma(u, v) = (-\sin u, \cos u, 0), \qquad \partial_2\sigma(u, v) = (0, 0, 1).$$

Furthermore, the tangent vectors

$$\partial_1\sigma(0, 0)_{\sigma(0,0)} = (0, 1, 0)_{(1,0,0)}, \qquad \partial_1\sigma\left(-\frac{\pi}{2}, 1\right)_{\sigma\left(-\frac{\pi}{2},1\right)} = (1, 0, 0)_{(0,-1,1)}$$

are drawn as orange arrows in the left plot of Figure 2.15, while the tangent vectors

$$\partial_2\sigma(0,0)_{\sigma(0,0)} = (0,0,1)_{(1,0,0)}, \qquad \partial_2\sigma(\pi,-1)_{\sigma(\pi,-1)} = (0,0,1)_{(-1,0,-1)}$$

are drawn as red arrows in the right plot of Figure 2.15.

Note that, in accordance to our intuitions, the *orange arrows point along the purple paths*, while the *red arrows point along the green paths*.

Example 2.56. Consider now the function

$$\rho : \mathbb{R}^2 \to \mathbb{R}^3, \qquad \rho(u,v) = (u,v,u^2+v^2),$$

whose image is a *paraboloid* in 3-dimensional space; see Figure 2.16.

Differentiating ρ, we see that

$$\partial_1\rho(u,v) = (1,0,2u), \qquad \partial_2\rho(u,v) = (0,1,2v).$$

In particular, at $(u,v) = (0,0)$, we have

$$\partial_1\rho(0,0)_{\rho(0,0)} = (1,0,0)_{(0,0,0)}, \qquad \partial_2\rho(0,0)_{\rho(0,0)} = (0,1,0)_{(0,0,0)}.$$

These arrows are drawn in the middle and right plots of Figure 2.16, respectively.

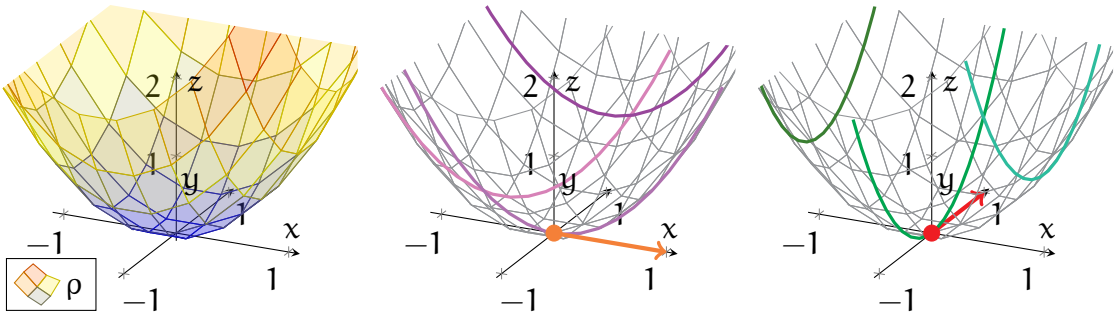


FIGURE 2.16. The left plot shows the image of ρ from Example 2.56. In the middle diagram, the purple curves depict paths obtained by holding v constant, while the orange arrow is the tangent vector $\partial_1\rho(0,0)_{\rho(0,0)}$. In the right diagram, the green curves depict paths obtained by holding u constant, while the red arrow is the tangent vector $\partial_2\rho(0,0)_{\rho(0,0)}$.

Finally, we remark that functions of two variables, such as those found in Examples 2.55 and 2.56, will form the basis for our study of surfaces later in this module.

2.8. **Vector Fields.** We now shake things up by bringing tangent vectors back into the discussion—we consider functions that have tangent vectors as values:

Definition 2.57. Let A be a subset of $\mathbb{R}^n$. A vector field on A is a function $\mathbf{F}$ that maps each point $\mathbf{p} \in A$ to a tangent vector $\mathbf{F}(\mathbf{p}) \in T_{\mathbf{p}}\mathbb{R}^n$ starting from $\mathbf{p}$.

In other words, *a vector field* $\mathbf{F}$ *maps each point* $\mathbf{p}$ *in its domain to an arrow* $\mathbf{F}(\mathbf{p})$ *beginning at the same point* $\mathbf{p}$. This, in particular, allows for a natural way to plot vector fields. Since $\mathbf{F}(\mathbf{p})$ is a tangent vector based at $\mathbf{p}$, we can depict $\mathbf{F}(\mathbf{p})$ by simply drawing the corresponding arrow in $\mathbb{R}^n$, with its starting point at $\mathbf{p}$. By drawn a large enough sample of arrows, we can gain an understanding of how $\mathbf{F}$ behaves.

Example 2.58. Consider the vector field $\mathbf{F}$ on $\mathbb{R}^2$ given by

$$\mathbf{F}(x, y) = (-y, x)_{(x,y)}.$$

Some values of $\mathbf{F}$ include the following:

$$\mathbf{F}(1, 0) = (0, 1)_{(1,0)}, \qquad \mathbf{F}(2, 1) = (-1, 2)_{(2,1)}, \qquad \mathbf{F}(0, -1) = (1, 0)_{(0,1)}. \tag{2.23}$$

Several values of $\mathbf{F}$ are drawn in the left plot of Figure 2.17 using green arrows. The three values computed in (2.23) are depicted within the same plot as red arrows.

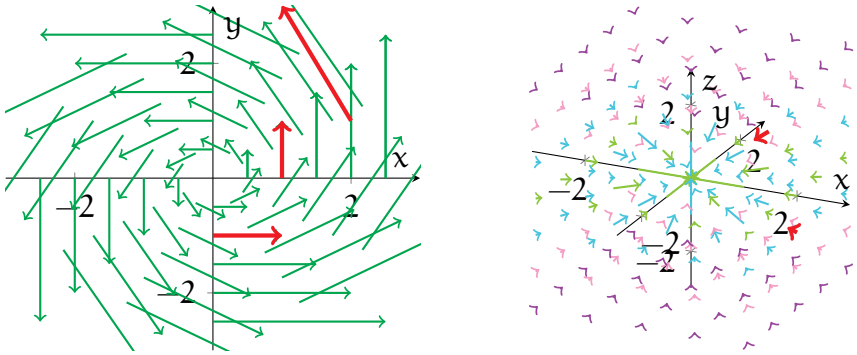


FIGURE 2.17. In the left plot, the green arrows represent several values of the vector field $\mathbf{F}$ from Example 2.58, while the red arrows indicate the values that were computed in (2.23). Similarly, in the right drawing, the arrows represent several values of the vector field $\mathbf{G}$ from Example 2.59, with the red arrows corresponding to the values computed in (2.24).

Example 2.59. Consider the vector field $\mathbf{G}$ on $\mathbb{R}^3 \setminus \{(0, 0, 0)\}$ given by

$$\mathbf{G}(\mathbf{x}) = -\frac{1}{|\mathbf{x}|^3} \cdot \mathbf{x}_{\mathbf{x}}.$$

Some values of $\mathbf{G}$ include (see the red arrows in the right plot of Figure 2.17):

$$\mathbf{G}(1, 1, 1) = -\frac{1}{3\sqrt{3}}(1, 1, 1)_{(1,1,1)}, \qquad \mathbf{G}(2, 0, -1) = -\frac{1}{5\sqrt{5}}(2, 0, -1)_{(2,0,-1)}. \tag{2.24}$$

Other values of $\mathbf{G}$ are also drawn (in various colours) in the right part of Figure 2.17.

In classical physics, $\mathbf{G}$ is often used to model the Newtonian gravitational force exerted by a massive particle sitting at the origin. At any point $\mathbf{x} \in \mathbb{R}^3$, with $\mathbf{x} \neq (0,0,0)$, the tangent vector $\mathbf{G}(\mathbf{x})$ points from $\mathbf{x}$ toward the origin. This is the gravitational force causing an object at $\mathbf{x}$ to accelerate toward the origin.

In addition, note that the length of $\mathbf{G}(\mathbf{x})$ satisfies

$$|\mathbf{G}(\mathbf{x})| = |\mathbf{x}|^{-2},$$

that is, the arrow becomes longer as $\mathbf{x}$ becomes closer to the origin. Thus, the gravitational force becomes stronger as one moves toward the particle at the origin.

Note that vector fields are hardly any different from vector-valued functions. The main difference is cosmetic: for a vector field $\mathbf{F}$, we also include $\mathbf{p}$ as part of the value of $\mathbf{F}(\mathbf{p})$—as the starting point of the tangent vector.

In other words, a vector field $\mathbf{F}$ can be made into a vector-valued function by taking only the vector component of each of its values $\mathbf{F}(\mathbf{p})$. Similarly, vector-valued fuctions are made into a vector fields by attaching a starting point to each of its values.

Remark 2.60. In fact, most calculus texts will define a vector field on $\mathbb{R}^n$ as just a vector-valued function $f : A \to \mathbb{R}^n$, without any reference to tangent vectors. However, for this module, we will maintain a conceptual distinction between vector-valued functions and vector fields, since they have rather different interpretations.

We now complete our discussion of differentiation with yet another familiar concept: *gradients*. However, here we adopt a slightly different definition:

Definition 2.61. Let $U \subseteq \mathbb{R}^m$ be open and connected, let $f : U \to \mathbb{R}$, and let $\mathbf{p} \in U$. We define the gradient of f at $\mathbf{p}$, denoted $\nabla f(\mathbf{p})$ or $\operatorname{grad} f(\mathbf{p})$, to be the tangent vector

$$\nabla f(\mathbf{p}) = (\partial_1 f(\mathbf{p}), \partial_2 f(\mathbf{p}), \ldots, \partial_m f(\mathbf{p}))_{\mathbf{p}} \in T_{\mathbf{p}}\mathbb{R}^m, \tag{2.25}$$

whenever the right-hand side is well-defined.

Furthermore, if $\nabla f(\mathbf{p})$ exists at every $\mathbf{p} \in U$, then we define the gradient of f itself to be the vector field mapping each $\mathbf{p} \in U$ to $\nabla f(\mathbf{p}) \in T_{\mathbf{p}}\mathbb{R}^m$.

Observe that the gradient of a function $\mathbf{f}$ aggregates all the partial derivatives of $\mathbf{f}$ into a single object. Moreover, the values $\nabla \mathbf{f}(\mathbf{p})$ are interpreted as arrows in $\mathbb{R}^m$ starting at $\mathbf{p}$. Later on, we will explore some geometric meanings of these arrows.

Remark 2.62. In contrast to derivatives (Definition 2.34) and partial derivatives (Definition 2.50), the gradient is only applied to real-valued functions.

Remark 2.63. Recall that in calculus, ∇f was (equivalently) defined to be the vector-valued function whose components are the partial derivatives of f. However, in either case, the values of the gradient are always interpreted as arrows.

Example 2.64. Consider the function

$$s : \mathbb{R}^2 \to \mathbb{R}, \qquad s(x, y) = \frac{1}{2}(x^2 + y^2).$$

The behaviour of s can be better visualised through the left drawing in Figure 2.18; there, the coloured circles indicate the sets on which the value of s is constant.

Next, differentiating s, we obtain, for every $(x, y) \in \mathbb{R}^2$,

$$\partial_1 s(x, y) = x, \qquad \partial_2 s(x, y) = y.$$

Thus, by Definition 2.61, the gradient of s satisfies

$$\nabla s(x, y) = (x, y)_{(x,y)}.$$

Some values of the gradient ∇s are plotted in the right side of Figure 2.18. Note that all of these arrows point radially, away from the origin.

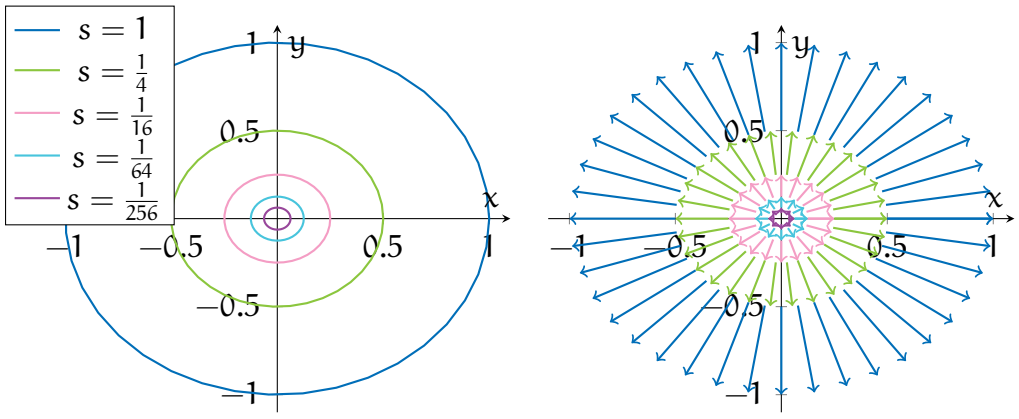


FIGURE 2.18. In the left drawing, the coloured circles are sets along which the function s in Example 2.64 is constant. In the right drawing, the arrows pointing away from the origin represent values of ∇s.

Example 2.65. For a 3-dimensional example, let us consider the function

$$g : \mathbb{R}^3 \setminus \{(0, 0, 0)\} \to \mathbb{R}, \qquad g(\mathbf{p}) = \frac{1}{|\mathbf{p}|}.$$

Alternatively, we can write g in terms of its component variables as

$$g(x,y,z) = \frac{1}{|(x,y,z)|} = \frac{1}{\sqrt{x^2+y^2+z^2}}.$$

Let us now take partial derivatives of g. First, by the chain and power rules,

$$\begin{aligned}\partial_1 g(x,y,z) &= \left(-\frac{1}{2}\right)\frac{1}{(x^2+y^2+z^2)^{\frac{3}{2}}}\cdot 2x \\ &= -\frac{x}{(x^2+y^2+z^2)^{\frac{3}{2}}}.\end{aligned}$$

By similar calculations, we also obtain

$$\partial_2 g(x,y,z) = -\frac{y}{(x^2+y^2+z^2)^{\frac{3}{2}}}, \qquad \partial_3 g(x,y,z) = -\frac{z}{(x^2+y^2+z^2)^{\frac{3}{2}}}.$$

Combining the above, we conclude that the gradient of g is

$$\nabla g(x,y,z) = -\frac{1}{(x^2+y^2+z^2)^{\frac{3}{2}}}\cdot (x,y,z)_{(x,y,z)}.$$

It is perhaps more informative to rewrite the preceding formula in vector form:

$$\nabla g(\mathbf{p}) = -\frac{1}{|\mathbf{p}|^3}\cdot \mathbf{p}_{\mathbf{p}}.$$

Notice that ∇g equals the vector field $\mathbf{G}$ from Example 2.59, modelling the Newtonian gravitational force from a particle at the origin. Thus, the gravitational force $\mathbf{G}$ has the form of a gradient of a scalar function. In physics terminology, the scalar function $-g$ is called the *gravitational potential* arising from this particle.

2.9. **Integrals.** We now turn our attention toward the other major aspect of calculus: *integration*. Here, we explore some interpretations of integrals, and we recall some basic methods you learned in calculus for evaluating them.

Let us begin with integrals of single-variable functions, which you learned to compute in *MTH4100/4200: Calculus I*. First, recall that if we integrate the constant function 1 along an interval $I = [a,b]$ of the real line, then we obtain

$$\int_I dx = \int_a^b 1\,dx = b - a.$$

In particular, *the integral of* 1 *over* I *yields the length of* I.

Now, one may want to alter how "length" is measured. For instance, one may be interested in a "weighted length", in which some points count for more than others.

This is where the power of integration comes in—*given* $g : I \to \mathbb{R}$, *we can interpret*

$$\int_I g\,dx = \int_a^b g(x)\,dx$$

as a "weighted length" of I, where *the "weight" applied to any* $x \in I$ *is given by* $g(x)$.

The term "weighted length" is left deliberately vague, as it can have many different meanings. We briefly mention a couple interpretations here:

Example 2.66. Suppose $g : [a, b] \to \mathbb{R}$ is everywhere positive, and consider the region

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

lying between the graph of g and the x-axis; see Figure 2.19. In calculus, you learned that the area of R is precisely the integral of g:

$$\mathcal{A}(R) = \int_a^b g(x)\,dx. \tag{2.26}$$

One way to interpret the area integral (2.26) is as a "weighted length". We can think of $\mathcal{A}(R)$ as a "length" of the interval (a, b), except that at each $x \in (a, b)$, we assign the height $g(x)$ as a weight. As a result, points at which g is larger count more toward this "weighted length" than points at which g is smaller.

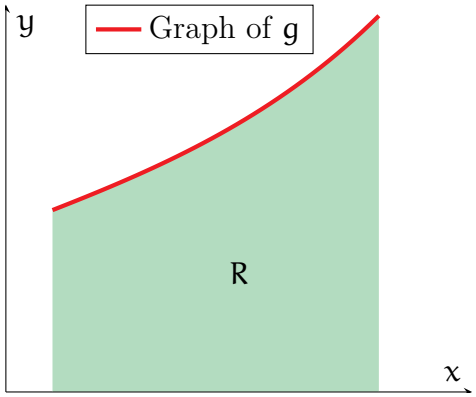


FIGURE 2.19. The red path is the graph of g from Example 2.66, while the green portion is the region R between g and the x-axis.

Example 2.67. Let $I = [a, b]$ represent a rod, and let $q : I \to \mathbb{R}$, such that $q(x)$ models the *electric charge density* (i.e. charge per unit length) of this rod at $x \in I$. Note that q can, at various points, be positive, negative, or zero—indicating positively charged, negatively charged, and uncharged sections of the rod.

The total charge Q of the rod can then be found by "summing up" its density along I:

$$Q = \int_a^b q(x)\,dx.$$

In particular, contributions to the total charge from parts of I where q is positive can be cancelled by other contributions from areas where q is negative.

Remark 2.68. While we discussed many interpretations of integrals, *we have not yet provided a precise definition of integrals*. Unfortunately, this is a more involved topic that is beyond this module. One rigorous definition is discussed in *MTH5105: Differential and Integral Analysis*. An even more powerful theory of integration (i.e. *Lebesgue integration*) is covered in the postgraduate module *MTH716U: Measure Theory and Probability*.

Next, we briefly revise how integrals of single-variable functions are calculated. The main tool, as you probably remember, is the *fundamental theorem of calculus*:

Theorem 2.69 (Fundamental theorem of calculus)**.** Let $f : [a, b] \to \mathbb{R}$ be continuous, and assume that its derivative $f'(x)$ is defined for all $x \in (a, b)$. Then,

$$\int_a^b f'(x)\,dx = f(b) - f(a). \tag{2.27}$$

The fundamental theorem of calculus is quite remarkable for multiple reasons. On the conceptual side, it establishes that differentiation and integration are inverses of each other—applying one operation undoes the other one.

Furthermore, the fundamental theorem of calculus provides practical ways to compute many integrals. The equation (2.27) relates integrals directly to derivatives, for which we already have many computational tricks. Therefore, if you know how to differentiate many functions, then you can also integrate many functions!

Example 2.70. Consider the function

$$f : [0, \pi] \to \mathbb{R}, \qquad f(x) = x^3 + \sin x.$$

Note f is the derivative of

$$g : [0, \pi] \to \mathbb{R}, \qquad g(x) = \frac{1}{4}x^4 - \cos x.$$

Thus, applying Theorem 2.69, we see that

$$\int_0^\pi (x^3 + \sin x)\,dx = \int_0^\pi g'(x)\,dx$$

$$= g(\pi) - g(0)$$
$$= \frac{1}{4}\pi^4 + 2.$$

In calculus, you have learned many methods for evaluating integrals. We will not cover most of these tricks in the lecture notes, as they are not directly relevant to this module. However, you can refer to your old notes and textbook [10] if needed.

2.10. **Multiple Integrals.** We conclude our revisions by discussing integrating functions of multiple variables. The first task is to extend our interpretations of single-variable integrals to *double* and *triple integrals*.

Consider a subset $B \subseteq \mathbb{R}^2$. Recall that, similar to the single integral case, *if we integrate the constant function* 1 *over* B*, then we obtain the area of* B:

$$\iint_B dA = \mathcal{A}(B). \tag{2.28}$$

Similarly, for a subset $C \subseteq \mathbb{R}^3$, *the integral of* 1 *over* C *yields the volume of* C:

$$\iiint_C dV = \mathcal{V}(C). \tag{2.29}$$

We can then think of general double and triple integrals as abstract measures of "weighted area" and "weighted volume". More specifically, given

$$f : B \to \mathbb{R}, \qquad g : C \to \mathbb{R},$$

the integrals

$$\iint_B f\, dA, \qquad \iiint_C g\, dV$$

give a weighted area and volume of B and C, with weights f and g, respectively.

Remark 2.71. In fact, one cannot define area or volume for all subsets of $\mathbb{R}^2$ or $\mathbb{R}^3$, respectively. (Look up the *Banach–Tarski paradox*!) One consequence of this is that we can only integrate over what are called *measurable subsets* of $\mathbb{R}^2$ or $\mathbb{R}^3$. However, in practice, any set that we encounter in this module (and in most other modules) will be measurable, so we do not stress about this point here.

Again, the meanings of "weighted area" and "weighted volume" are left vague in order to emphasise their applicability to a wide variety of situations.

Example 2.72. Let $B \subseteq \mathbb{R}^2$ represent a plate, and let $m : B \to \mathbb{R}$, with $m(x, y)$ being the *mass density* (i.e. mass per unit area) of the plate at the position $(x, y) \in B$. We can

then view the total mass M of the plate as a type of "weighted area":

$$M = \iint_B m \, dA.$$

Example 2.73. For an example in $\mathbb{R}^3$, we turn to quantum mechanics. Let $\Psi : \mathbb{R}^3 \to \mathbb{C}$ be the *wave function* of a particle, which describes its present state. In this context, the quantity $|\Psi(x, y, z)|^2 \in \mathbb{R}$ represents the *probability density* of the particle at (x, y, z).

Now, consider some region $C \subseteq \mathbb{R}^3$ of space. Then, the total probability that the particle is lying in C can be obtained by integrating the probability density over C:

$$\mathbb{P}(C) = \iiint_C |\Psi|^2 \, dV.$$

The main idea behind computing double and triple integrals is to transform them into quantities that we already know how to compute: single integrals. If you remember your calculus, then you know this is accomplished through *Fubini's theorem*, named after Italian mathematician Guido Fubini (1879–1943):

Theorem 2.74 (Fubini's theorem)**.** Let R denote the rectangle

$$R = [a, b] \times [c, d] = \{(x, y) \in \mathbb{R}^2 \mid a \le x \le b,\ c \le y \le d\}.$$

Then, for any bounded, piecewise continuous function $f : R \to \mathbb{R}$,

$$\iint_R f \, dA = \int_a^b \left[\int_c^d f(x, y) \, dy \right] dx = \int_c^d \left[\int_a^b f(x, y) \, dx \right] dy. \tag{2.30}$$

Furthermore, if Q denotes the rectangular prism

$$Q = [a, b] \times [c, d] \times [k, l] = \{(x, y, z) \in \mathbb{R}^3 \mid a \le x \le b,\ c \le y \le d,\ k \le z \le l\},$$

then for any bounded, piecewise continuous function $g : Q \to \mathbb{R}$,

$$\iiint_Q g \, dV = \int_k^l \int_c^d \int_a^b g(x, y, z) \, dx dy dz. \tag{2.31}$$

Similar formulas hold if the orders of the above integrals in x, y, and z are interchanged.

In particular, (2.30) reduces a double integral into two iterated single integrals, while (2.31) converts a triple integral into three iterated integrals.

Example 2.75. Let us compute the triple integral

$$\iiint_Q f \, dV, \tag{2.32}$$

where $Q = [0,1] \times [0,1] \times [0,2]$ is a rectangular prism in $\mathbb{R}^3$ (see the left picture in Figure 2.20), and where $f : \mathbb{R}^3 \to \mathbb{R}$ is the function given by

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

The first step is to apply (2.31) to convert (2.32) into iterated single integrals:

$$\iiint_Q f\,dV = \int_0^2 \int_0^1 \int_0^1 x(y+z)\,dx\,dy\,dz.$$

We can now evaluate each of the single integrals, starting from the innermost one. First, we integrate $x(y+z)$ with respect to x (treating y and z as constants):

$$\begin{aligned}\int_0^2 \int_0^1 \int_0^1 x(y+z)\,dx\,dy\,dz &= \int_0^2 \int_0^1 (y+z)\left[\int_0^1 x\,dx\right] dy\,dz \\ &= \frac{1}{2}\int_0^2 \int_0^1 (y+z)\,dy\,dz.\end{aligned}$$

We then continue by integrating with respect to y (while holding z constant):

$$\begin{aligned}\frac{1}{2}\int_0^2 \int_0^1 (y+z)\,dy\,dz &= \frac{1}{2}\int_0^2 \left[\frac{1}{2}y^2 + zy\right]\Big|_{y=0}^{y=1} dz \\ &= \int_0^2 \left(\frac{1}{4} + \frac{1}{2}z\right) dz.\end{aligned}$$

Combining all the above and integrating with respect to z yields the solution:

$$\int_0^2 \left(\frac{1}{4} + \frac{1}{2}z\right) dz = \frac{3}{2}.$$

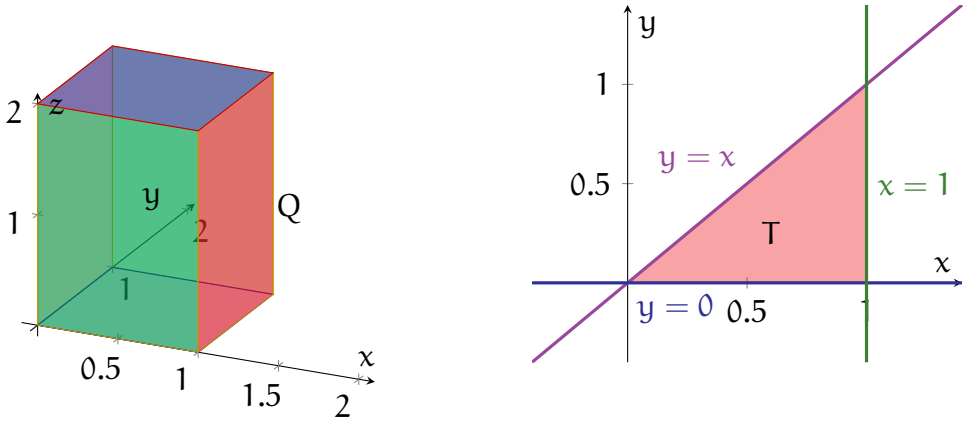


FIGURE 2.20. The solid figure on the left is the region Q from Example 2.75, while the shaded triangle on the right is T from Example 2.76.

Example 2.76. For a more involved example, let us compute the double integral

$$\iint_T g \, dA.$$

where $T \subseteq \mathbb{R}^2$ is the triangular region bounded by the lines $y = 0$, $x = 1$, and $y = x$ (see the right drawing in Figure 2.20), and where

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

The main idea is to apply (2.30) to the square $[0, 1] \times [0, 1]$ (which contains T) and to replace the integrand by another function that is equal to g on T and vanishes outside T:

$$g_T(x, y) = \begin{cases} g(x, y) = xy & (x, y) \in T, \\ 0 & \text{otherwise.} \end{cases}$$

From this, along with Fubini's theorem, we obtain

$$\begin{aligned} \iint_T g \, dA &= \iint_{[0,1]\times[0,1]} g_T \, dA \\ &= \int_0^1 \int_0^1 g_T(x, y) \, dy \, dx \\ &= \int_0^1 \left[\int_0^x xy \, dy \right] dx. \end{aligned}$$

Notice that in the last step, we can restrict the domain of the inner integral, since for any $x \in [0, 1]$, the value $g_T(x, y)$ is nonzero only when $0 < y \leq x$.

We can now evaluate the inner integral by treating x as a constant:

$$\begin{aligned} \int_0^1 \left[\int_0^x xy \, dy \right] dx &= \int_0^1 x \left[\frac{1}{2} y^2 \right] \Big|_{y=0}^{y=x} dx \\ &= \frac{1}{2} \int_0^1 x^3 \, dx. \end{aligned}$$

The answer now follows by evaluating this last integral in x:

$$\frac{1}{2} \int_0^1 x^3 \, dx = \frac{1}{8}.$$

Example 2.77. In Example 2.76, we integrated first in y and then in x. We can obtain the same answer with the reverse order of integration, provided we describe T correctly:

$$\iint_T g \, dA = \int_0^1 \int_0^1 g_T(x, y) \, dx \, dy$$

$$= \int_0^1 \int_y^1 xy \, dx dy.$$

Indeed, evaluating the above iterated integral, we obtain

$$\int_0^1 \int_y^1 xy \, dx dy = \frac{1}{2} \int_0^1 y(1 - y^2) \, dy$$
$$= \frac{1}{8}.$$

Example 2.78. For a more abstract example, we return to the region R from Example 2.66, containing the points lying between the x-axis and the graph of a positive function $g : [a, b] \to \mathbb{R}$ (see Figure 2.19 for an illustration).

First, the area of R is simply

$$\mathcal{A}(R) = \iint_R dA.$$

Like in Examples 2.76 and 2.77, this double integral can be expanded using (2.30):

$$\iint_R dA = \int_a^b \left[\int_0^{g(x)} 1 \, dy \right] dx$$
$$= \int_a^b g(x) \, dx.$$

In particular, this justifies the calculus formula (2.26) for the area of R.

Again, we will not cover all the various tips and tricks you learned from calculus to compute multiple integrals, as this is not the main focus of the module. An usual, you can consult your calculus notes or textbook [10] for additional details.

Finally, we remark that all of the preceding discussions for double and triple integrals extend directly to integrals involving functions of four or more variables. However, we will not encounter any such higher-dimensional integrals in this module.

3. The Geometry of Curves

With revisions behind us, we now advance to the main geometric topics. This chapter represents the first half of this effort, covering the *differential geometry of curves*. These are the simplest geometric objects that we will study, and they serve as warm-up for our future discussions involving surfaces.

You probably already have some solid intuitions for what a curve is. For example, you may visualise a curve as a string tracing out some path, or as a trajectory traced out by a particle over time. On the other hand, if you think more abstractly, then perhaps you view them as "1-dimensional objects" of some sort.

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 3.1. When we say "curve", what exactly do we mathematically mean?

3.1. Parametric Curves.

Let us begin with this vague idea of curves being particle trajectories or general "1-dimensional objects". Recall from the previous chapter that particle trajectories are modelled by vector-valued functions $\mathbf{f} : (a, b) \to \mathbb{R}^n$, with $\mathbf{f}(t)$ representing the position of the particle at time t.

Moreover, you may have noticed, in prior examples, that the images of such $\mathbf{f}$ trace out 1-dimensional paths; see, for instance, Examples 2.38, 2.39, and 2.40. Thus, we use such functions as the starting point of our formal discussions:

Definition 3.2. A parametric curve is a smooth vector-valued function $\gamma : I \to \mathbb{R}^n$, where n is a positive integer, and where I is an open (finite or infinite) interval.

A few remarks on Definition 3.2 are in order. First, by γ being *smooth*, we mean that *we can differentiate* γ *as many times as we like*. In particular, γ has no discontinuities or "jagged edges" that prevent its derivative from existing at any $t \in I$. The same is also assumed to hold for γ', as well as for higher derivatives of γ.

We also note the interval I in Definition 3.2 is allowed to be of any of the forms

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

The idea is that since I is 1-dimensional, we expect that the set of points $\gamma(t)$ mapped out by every $t \in I$ should also be 1-dimensional.

Lastly, if you prefer to be more concrete, then you can assume $n = 2$ or $n = 3$ in Definition 3.2. Almost all of our examples will be in these special cases.

Example 3.3. The function from Example 2.38,

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

is a parametric curve in $\mathbb{R}^2$. In particular, since $\cos$ and $\sin$ are infinitely differentiable, we can take as many derivatives of γ as we want, and therefore γ is smooth. Recall that γ maps out the *unit circle about the origin*; see the left plot in Figure 2.10.

By similar reasoning, given any $\mathbf{p}, \boldsymbol{v} \in \mathbb{R}^n$, the function

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

from Example 2.40, is a parametric curve in $\mathbb{R}^n$. Recall also that ℓ maps out the *line in $\mathbb{R}^n$ passing through the point $\mathbf{p}$ and aligned in the direction of $\boldsymbol{v}$*.

Now that we have a formal definition, we can evaluate it critically: *do parametric curves give an appropriate geometric description of curves?* One part of answering this question is to hunt for possible cases in which Definition 3.2 falls short.

Example 3.4. Fix a point $\mathbf{p} \in \mathbb{R}^n$, and consider the constant function

$$\lambda : \mathbb{R} \to \mathbb{R}^n, \qquad \lambda(t) = \mathbf{p}.$$

The not-so-interesting image of λ is shown in Figure 3.1.

Now, λ is a vector-valued function, and it is clearly smooth (all of its derivatives vanish). Thus, λ is indeed a parametric curve, by Definition 3.2. However, the image of λ, which is merely a single point $\mathbf{p}$, is apparently not 1-dimensional. (By convention, *we think of a single point, or a finite number of points, as 0-dimensional.*)

FIGURE 3.1. Well, that failed...

So, what went awry in Example 3.4? Why did this parametric curve fail to map out a 1-dimensional object? The deficiency here is rather apparent—*for λ to be "1-dimensional", its value $\lambda(t)$ needs to actually move as t changes.* When λ does not move, we end up with the rather sad situation in Example 3.4 and Figure 3.1.

Fortunately, we do know how to check if a parametric curve is moving—we simply check that its derivative is nonzero. This leads to the following definition:

Definition 3.5. A parametric curve $\gamma : I \to \mathbb{R}^n$ is regular iff $|\gamma'(t)| \neq 0$ for every $t \in I$.

In summary, our current progress on Question 3.1 is that we wish to define curves using parametric curves, but we must also assume that these are regular.

Example 3.6. Consider the parametric unit circle from Example 2.38,

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

Recall that at each $t \in \mathbb{R}$, we have

$$\gamma'(t) = (-\sin t, \cos t).$$

Taking the norm of the above (and recalling basic trigonometric rules), we see that

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

As a result, γ satisfies the condition in Definition 3.5 and hence is regular.

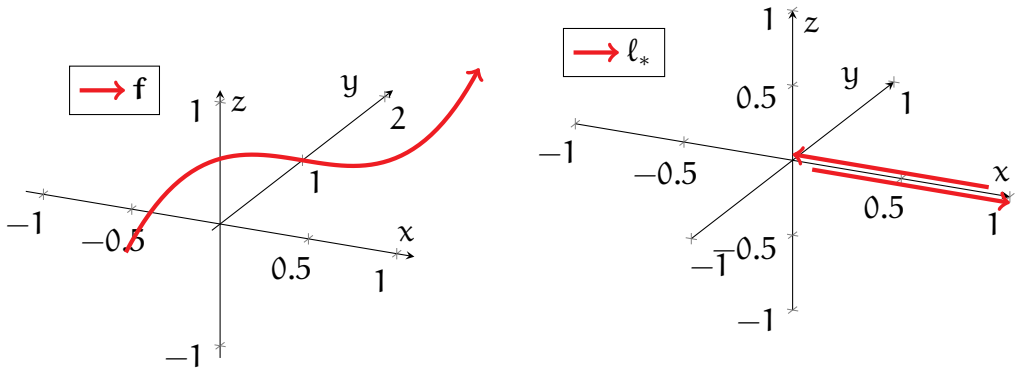


FIGURE 3.2. The left drawing shows the image of the regular parametric curve $\mathbf{f}$ from Example 3.7, while the right drawing shows the image of ℓ_* (which fails to be regular) from Example 3.8.

Example 3.7. Consider the following parametric curve in $\mathbb{R}^3$:

$$\mathbf{f} : \mathbb{R} \to \mathbb{R}^3, \qquad \mathbf{f}(t) = (t, 1, t^3).$$

(See the left plot in Figure 3.2.) Again, direct computations yield, at each $t \in \mathbb{R}$,

$$\mathbf{f}'(t) = (1, 0, 3t^2), \qquad |\mathbf{f}'(t)| = \sqrt{1 + 9t^4}.$$

Since $9t^4 \geq 0$ for any $t \in \mathbb{R}$, then

$$|\mathbf{f}'(t)| \geq \sqrt{1} = 1.$$

In particular, this shows $|\mathbf{f}'(t)|$ can never be zero, and thus $\mathbf{f}$ is regular.

Alternatively, one can simply note that $\mathbf{f}'(t)$ never vanishes for any $t \in \mathbb{R}$ (since its x-component is never 0), and hence $|\mathbf{f}'(t)|$ also cannot vanish.

Example 3.8. Consider now the parametric curve

$$\ell_* : \mathbb{R} \to \mathbb{R}^3, \qquad \ell_*(t) = (t^2, 0, 0).$$

The right plot of Figure 3.2 shows the behaviour of ℓ_*. For negative values of t, the value $\ell_*(t)$ moves leftward along the positive x-axis until it reaches the origin (when $t = 0$). When t is positive, the value $\ell_*(t)$ moves rightward along the same positive x-axis.

Taking a derivative of ℓ_*, we see that

$$\ell_*'(t) = (2t, 0, 0), \qquad |\ell_*'(t)| = 2|t|, \qquad t \in \mathbb{R}.$$

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

Although ℓ_* is not nearly as deficient as the constant function in Example 3.4, it still fails to be regular, since ℓ_* momentarily "stops" at $t = 0$ before changing directions. As we will see later, even this milder form of non-regularity will be undesirable, and we still want to eliminate such occurrences from our description of curves.

In addition, we wish to impose another restriction: *we exclude parametric curves that self-intersect or, in other words, "pass through themselves"*. Our main rationale for this is that the applications that we will study later on (for instance, *Lagrange multipliers* and *Green's theorem*) will require our curves to not self-intersect.

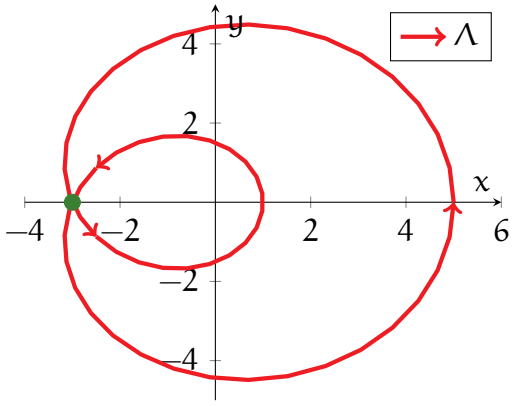


FIGURE 3.3. This is a plot of the limaçon Λ from Example 3.9.

Example 3.9. Consider the parametric curve given by

$$\Lambda : \mathbb{R} \to \mathbb{R}^2, \qquad \Lambda(t) = \left(2\cos t + 3\cos(2t),\, 2\sin t + 3\sin(2t)\right).$$

The image of Λ, illustrated in Figure 3.3, is known as a *limaçon*. Note this image passes through itself at the single point marked in green:

$$(-3,0) = \Lambda(t_\pm), \qquad t_\pm = \cos^{-1}\left(-\frac{1}{3}\right) \approx \pi \pm 1.23.$$

Even if a parametric curve is regular, this alone does not prevent self-intersections from occurring. Thus, to rule out objects like Λ from Example 3.9, we will have to impose further restrictions in our eventual definition of curves.

Remark 3.10. We emphasise that *our exclusion of self-intersections is a choice made for this module*, not a necessity. In fact, much of the literature does allow for self-intersecting curves (known as *immersed curves*), which have a rich theory of their own.

3.2. **Reparametrisations.** Thus far, we have argued that "good" parametric curves should be regular and should not pass through themselves. Now, we ask *what other shortcomings do parametric curves have, with regards to describing geometric curves?*

The following example highlights a more subtle issue:

Example 3.11. Consider the following two parametric curves:

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

The images of γ_1 and γ_2 are drawn in the left and right illustrations in Figure 3.4. Observe that γ_1 is regular by the computations in Example 3.6. Moreover, γ_2 is regular, since

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

is clearly nonzero for every $t \in (-1,1)$.

Now, γ_1 and γ_2 are apparently different parametric curves. On the other hand, Figure 3.4 shows that γ_1 and γ_2 trace out exactly the same points, the *upper half-circle*, and in the same order. Thus, in this sense, γ_1 and γ_2 are not so different after all!

If you think of γ_1 and γ_2 in Example 3.11 as travelling particles, then this simply means the two particles are traversing the same path at different times and speeds. From this perspective, it is sensible to think of γ_1 and γ_2 as distinct.

However, we are presently concerned with questions of *geometry*—of *shapes and sizes*. Here, γ_1 and γ_2 both trace out the same points and hence have the same shape

and size. Thus, although γ_1 and γ_2 are different parametric curves, one can argue convincingly that they should be considered "the same" geometrically.

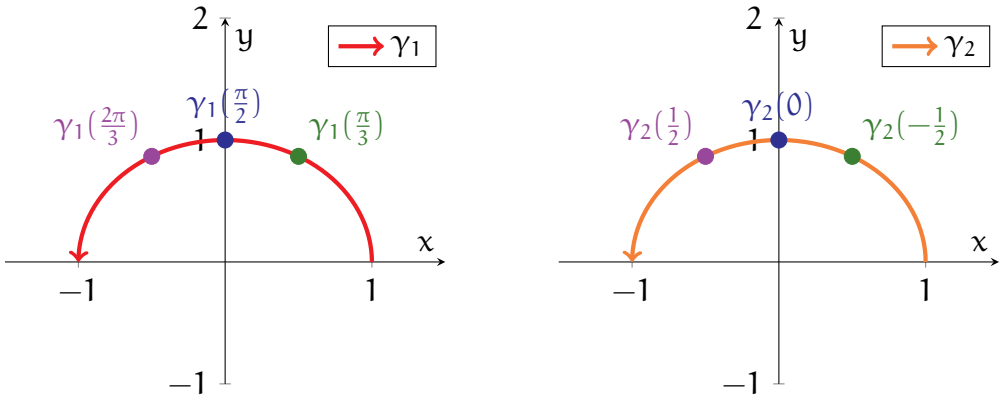


FIGURE 3.4. The plots show the common image of γ_1 (left) and γ_2 (right) from Example 3.11. The green, blue, and purple points show three specific points on this half-circle, as well as how they relate to γ_1 and γ_2.

To dig a bit deeper, let us now take a closer look at how γ_1 and γ_2 are related to each other. Suppose we are extra clever and magically know beforehand what has to be done—we adopt the following change of variables:

$$\tilde{t} = \phi(t) = -\cos(t). \tag{3.1}$$

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

There is a very good reason we chose (3.1) in particular. Applying γ_1 using the original parameter t and γ_2 with the new parameter $\tilde{t}$, we then obtain

$$\begin{aligned} \gamma_2(\tilde{t}) &= \left(\cos t, \sqrt{1 - (-\cos t)^2}\right) \\ &= (\cos t, \sin t) \\ &= \gamma_1(t). \end{aligned} \tag{3.2}$$

In other words, by switching t into $\tilde{t}$, we have transformed γ_1 into γ_2!

For any point $\mathbf{p}$ along this half-circle, the particle γ_1 reaches $\mathbf{p}$ at some time t, while γ_2 reaches $\mathbf{p}$ at some other time $\tilde{t}$. The change of variables (3.1) is the object that matches this time t for γ_1 to the corresponding time $\tilde{t}$ for γ_2. Figure 3.4 provides a graphical comparison of t and $\tilde{t}$ for three particular points. For instance, the point $(0, 1)$, the blue dot the top of the half-circle, corresponds to $t = \frac{\pi}{2}$ and $\tilde{t} = 0$.

Our intuition of γ_1 and γ_2 describing the same curve is closely tied to the observation that there is this change of variables (3.1) allowing us to convert from γ_1 to γ_2, and vice versa. We now formalise this idea for general parametric curves:

Definition 3.12. Let $\gamma : I \to \mathbb{R}^n$ and $\tilde{\gamma} : \tilde{I} \to \mathbb{R}^n$ be regular parametric curves. We say γ is a reparametrisation of $\tilde{\gamma}$ iff there exists a bijection $\phi : I \to \tilde{I}$ such that:

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

$$\tilde{\gamma}(\phi(t)) = \gamma(t). \tag{3.3}$$

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

Those unfamiliar with formal mathematics can find Definition 3.12 a bit scary, but there is nothing to fear! In fact, the idea is exactly the same as in Example 3.11:

- We can view $\phi(t)$ in (3.3) precisely as a "transformed time" $\tilde{t}$.
- Similarly, (3.3) generalises the conversion relation we had in (3.2).

Example 3.13. Let γ_1 and γ_2 be as in Example 3.11. Observe that (3.1) and (3.2) imply γ_1 is a reparametrisation of γ_2. In particular, in terms of Definition 3.12, we have

$$I = (0, \pi), \qquad \tilde{I} = (-1, 1), \qquad \phi(t) = -\cos t.$$

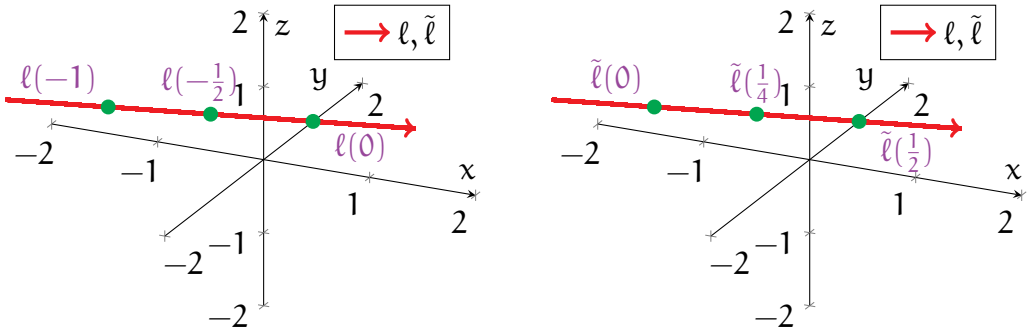


FIGURE 3.5. Both plots show the line mapped by ℓ and $\tilde{\ell}$ from Example 3.14, as well as three common points along this line (in green). The first plot is labelled in terms of ℓ, while the second is labelled in terms of $\tilde{\ell}$.

Example 3.14. Consider the two parametric lines (see Figure 3.5 for illustrations)

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

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

To show this, we find the change of variables ϕ identifying ℓ with $\tilde{\ell}$, that is, we solve $\ell(t) = \tilde{\ell}(\tilde{t})$ for a relation between t and $\tilde{t}$. The above expands to a system of 3 equations:

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

By doing a bit of algebra, we then see that the above has a simulaneous solution:

$$\tilde{t} = \frac{1}{2}(t + 1). \tag{3.4}$$

This then inspires us to define the map

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

which is clearly a smooth bijection between $\mathbb{R}$ and itself. Moreover, its inverse satisfies

$$\phi^{-1}(\tilde{t}) = 2\tilde{t} - 1,$$

which is also a smooth function. Finally, from (3.4) and the above, we have that

$$\tilde{\ell}(\phi(t)) = \tilde{\ell}(\tilde{t}) = \ell(t), \qquad t \in \mathbb{R},$$

and hence ℓ is indeed a reparametrisation of $\tilde{\ell}$.

The upshot, in the context of Definition 3.12, is that *whenever γ is a reparametrisation of $\tilde{\gamma}$, we view γ and $\tilde{\gamma}$ as describing the same curve.* This lies at the heart of the idea that *curves, and their properties, are independent of parametrisation.*

Remark 3.15. If γ is a reparametrisation of $\tilde{\gamma}$, then $\tilde{\gamma}$ is also a reparametrisation of γ. We often write "γ and $\tilde{\gamma}$ are reparametrisations of each other" due to this symmetry.

We previously mentioned that two parametric curves that are reparametrisations of each other can be viewed as two particles traversing the same path at different speeds. Using a bit of calculus, we can determine how these two speeds are related:

Theorem 3.16. Let $\gamma : I \to \mathbb{R}^n$ and $\tilde{\gamma} : \tilde{I} \to \mathbb{R}^n$ be parametric curves. Assume γ is a reparametrisation of $\tilde{\gamma}$, with change of variables $\phi : I \to \tilde{I}$ (i.e. (3.3) holds). Then,

$$\gamma'(t) = \phi'(t) \cdot \tilde{\gamma}'(\phi(t)), \qquad t \in I. \tag{3.5}$$

Proof. First, we express $\tilde{\gamma}$ in terms of its components,

$$\tilde{\gamma}(\tilde{t}) = (\tilde{\gamma}_1(\tilde{t}), \ldots, \tilde{\gamma}_n(\tilde{t})).$$

Using Theorem 2.37 and (3.3), we expand

$$\begin{aligned}\gamma'(t) &= \frac{d}{dt}[\tilde{\gamma}(\phi(t))] \\ &= \left(\frac{d}{dt}[\tilde{\gamma}_1(\phi(t))], \ldots, \frac{d}{dt}[\tilde{\gamma}_n(\phi(t))]\right)\end{aligned}$$

Applying the chain rule to the above then yields

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

which is precisely (3.5). □

Theorem 3.16 states that reparametrising a parametric curve will only transform its derivative by a scalar factor. This should not be surprising, since both functions are traversing the same paths and hence should be moving in the same directions.

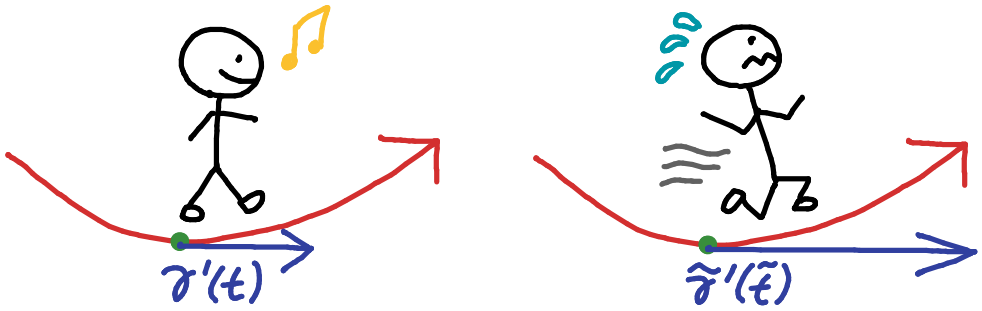


FIGURE 3.6. The pictures show one person γ walking and another person $\tilde{\gamma}$ running along the red path (the common image of γ and $\tilde{\gamma}$). At the green point, their velocities are indicated by blue arrows.

To help interpret this, let us consider the red path in Figure 3.6. The left drawing shows a person (modelled by the parametric curve γ) walking slowly along this path while humming a tune. The right drawing shows another person $\tilde{\gamma}$ running desperately along the same path because he is late to his *MTH5113* lecture. While both people are moving in the same direction, the runner $\tilde{\gamma}$ is moving at a higher speed.

3.3. **Geometric Curves.** So far, we have discussed the following principles:

(1) Curves are described using regular parametric curves.
(2) Curves, and their properties, are independent of parametrisation.
(3) We do not allow curves to pass through themselves, or "self-intersect".

With these ideas in mind, we can now formulate a precise definition of curves:

Definition 3.17. $C \subseteq \mathbb{R}^n$ is called a (smooth) curve iff for any $\mathbf{p} \in C$, there exist

- An open subset $V \subseteq \mathbb{R}^n$ such that $\mathbf{p} \in V$, and
- A regular and injective parametric curve $\gamma : I \to C$,

such that the following conditions hold:

- γ is a bijection between I and $C \cap V$.
- The inverse $\gamma^{-1} : C \cap V \to I$ of γ is also continuous.

Definition 3.17 draws upon some unfamiliar background from topology. For that reason, we will not focus on the precise definition in this module. However, we are in position to explore how Definition 3.17 relates to the above three principles.

Suppose $C \subseteq \mathbb{R}^n$ is a curve, as described in Definition 3.17. Then, given $\mathbf{p} \in C$, the parametric curve γ from Definition 3.17 provides a smooth one-to-one correspondence between points on the interval I and points of C near $\mathbf{p}$ (namely, in the subset $C \cap V$). This is illustrated in the left drawing of Figure 3.7; the open subset V is drawn in purple, while both I and $C \cap V$ are coloured red.

The interpretation is that *near* $\mathbf{p}$*, the curve* C *looks like a "deformed" version of* I (think of a bent piece of wire), while γ *describes how* I *was deformed.*

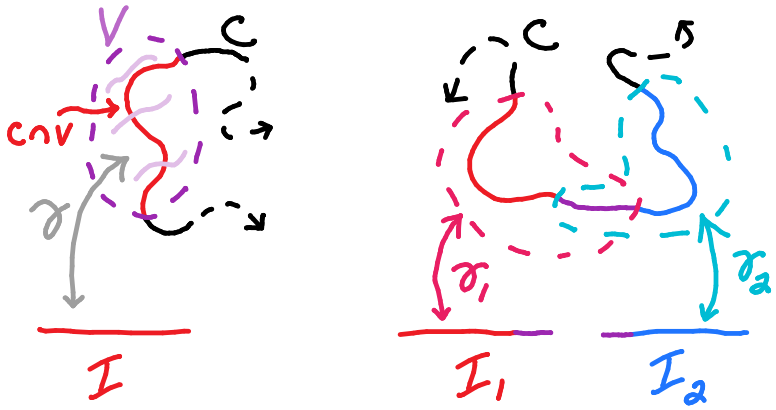


FIGURE 3.7. In the left drawing, a curve C is locally represented by a regular, injective parametric curve γ (indicated in red), as described in Definition 3.17. On the right, we have two parametric curves γ_1 (in red) and γ_2 (in blue) describing two parts of C. Moreover, at points where γ_1 and γ_2 overlap (drawn in purple), they are reparametrisations of each other.

Thus, C *can be viewed as the object obtained by "gluing together" one or more "deformed intervals".* In other words, using a collection of regular parametric curves,

each describing such a deformed interval, we can build the entire curve C. In this way, Definition 3.17 fulfills the principle (1) from the beginning of this discussion.

While Definition 3.17 states C can be described near any of its points using a regular parametric curve γ, it does not single out any particular γ. For any $\mathbf{p} \in C$, there are many possible γ satisfying the conditions of Definition 3.17—that is, there are many ways to describe a portion of C near $\mathbf{p}$ as a "deformed interval". This leads to the principle (2) of parametrisation independence: one views C as the underlying object that can be described equally well by many different parametric curves.

The above is also connected to the preceding discussion on reparametrisations. Suppose γ_1 and γ_2 are two parametric curves satisfying the hypotheses of Definition 3.17; see the right drawing in Figure 3.7 for an illustration. Then, *along the points where* γ_1 *and* γ_2 *overlap* (the purple segment in Figure 3.7), one can show that γ_1 *and* γ_2 *are reparametrisations of each other*. (We avoid proving a precise statement here, as it lies a bit beyond the scope of this module; the idea is that both $\gamma_2 \circ \gamma_1^{-1}$ and $\gamma_1 \circ \gamma_2^{-1}$ define changes of variables.) Thus, different parametric curves describing C are connected by the property of being reparametrisations of each other.

Example 3.18. Let $\mathcal{H}$ be the *helix*, defined as the subset

$$\mathcal{H} = \{(\cos t, \sin t, t) \mid t \in \mathbb{R}\} \subseteq \mathbb{R}^3.$$

See the left graphic in Figure 3.8 for a plot of $\mathcal{H}$.

To relate $\mathcal{H}$ to Definition 3.17, we consider the parametric curve

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

Note that $\mathbf{h}$ is indeed regular, since for any $t \in \mathbb{R}$,

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

Also, observe that $\mathbf{h}$ is injective, and that the image of $\mathbf{h}$ is precisely the set $\mathcal{H}$. Intuitively, $\mathbf{h}$ gives one description of how the real line can be deformed into the helix $\mathcal{H}$. One can also think of $\mathbf{h}$ as coiling a straight wire into a spring having the shape of $\mathcal{H}$.

For completeness, let us mention that given any $\mathbf{p} \in \mathcal{H}$, one can show that the open subset $V = \mathbb{R}^3$ and the regular parametric curve $\gamma = \mathbf{h}$ satisfy the hypotheses of Definition 3.17. (Due to a lack of background, we do not go into details of the proof here; most of this was already shown above, aside from the fact that the inverse of $\mathbf{h}$ is continuous.) As a result, we conclude that $\mathcal{H}$ *is indeed a curve*, in the sense of Definition 3.17.

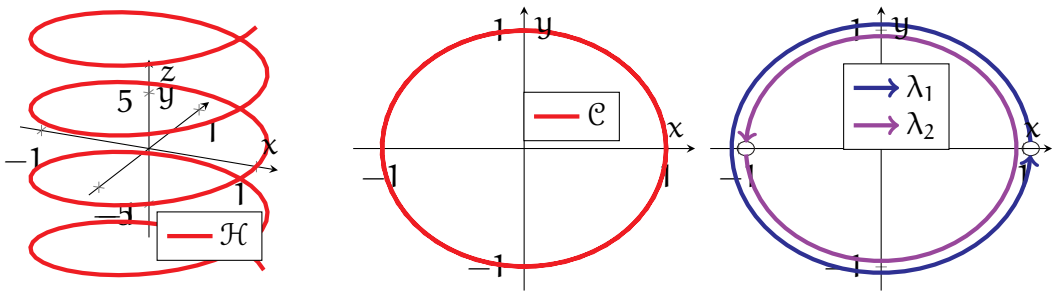


FIGURE 3.8. The left drawing is a plot of the helix $\mathcal{H}$ from Example 3.18, while the middle drawing is a plot of the circle $\mathcal{C}$ from Example 3.19. The right graphic shows how $\mathcal{C}$ can be constructed by patching together the two parametric curves λ_1 and λ_2 that were defined in Example 3.19.

Example 3.19. Next, we consider the *unit circle* $\mathcal{C}$ about the origin in $\mathbb{R}^2$,

$$\mathcal{C} = \{(x, y) \in \mathbb{R}^2 \mid x^2 + y^2 = 1\}.$$

See the middle drawing in Figure 3.8 for a drawing of $\mathcal{C}$.

To see why $\mathcal{C}$ is a curve, let us consider the following parametric curves:

$$\lambda_1 : (0, 2\pi) \to \mathbb{R}^2, \qquad \lambda_1(t) = (\cos t, \sin t),$$
$$\lambda_2 : (-\pi, \pi) \to \mathbb{R}^2, \qquad \lambda_2(t) = (\cos t, \sin t).$$

The same computations as in Example 3.6 yield that both λ_1 and λ_2 are regular. Moreover:

- λ_1 is injective, and its image is all of $\mathcal{C}$ except for the point $(1, 0)$.
- λ_2 is also injective, and its image is all of $\mathcal{C}$ except for the point $(-1, 0)$.

Both λ_1 and λ_2 describe an interval being bent into circular arcs; the right plot in Figure 3.8 shows the images of λ_1 and λ_2. Note that also although each arc misses one point of $\mathcal{C}$, the two arcs together describe $\mathcal{C}$ in its entirety. In other words, *the circle $\mathcal{C}$ is constructed by gluing together the images of λ_1 and λ_2*.

To connect the above with the formal definition, one can proceed as follows:

- If $\mathbf{p} \in \mathcal{C}$ and $\mathbf{p} \neq (1, 0)$, then the hypotheses of Definition 3.17 are satisfied, with
$$\gamma = \lambda_1, \qquad V = \mathbb{R}^2 \setminus \{(1, 0)\}.$$
- If $\mathbf{p} \in \mathcal{C}$ and $\mathbf{p} \neq (-1, 0)$, then one can instead take
$$\gamma = \lambda_2, \qquad V = \mathbb{R}^2 \setminus \{(-1, 0)\}.$$

(Again, we do not go into detailed proofs here.) As a result, $\mathcal{C}$ *is indeed a curve*.

The last point of discussion regarding Definition 3.17 is the exclusion of objects that self-intersect—principle (3) from before. This is achieved through the conditions satisfied by γ in Definition 3.17—in particular, that γ is a bijection, and that both γ and γ^{-1} are continuous. Roughly speaking, these conditions have the effect of ensuring that the image of γ has the structure of a "deformed line segment".

Consider the object C_1 depicted in the left drawing of Figure 3.9, and observe that C_1 contains a self-intersection point that is marked in green. If we let $\mathbf{p}$ be this green point, then for any open subset V specified in Definition 3.17, the intersection $C_1 \cap V$ must contain an "X-figure" around $\mathbf{p}$. The main idea is that this "X" has a different (topological) structure than a deformed interval, so that one cannot continuously map between an interval I and the "X" bijectively, as required in Definition 3.17.

To get a better sense of this, imagine bending a straight piece of wire into this X-shape. You should convince yourself that this is impossible without either breaking the wire (violating continuity of γ) or folding the wire over itself (violating injectivity).

As a result, we conclude that this C_1 is not a curve. In particular, this argument shows that the *limaçon* in Example 3.9 fails to be a curve.

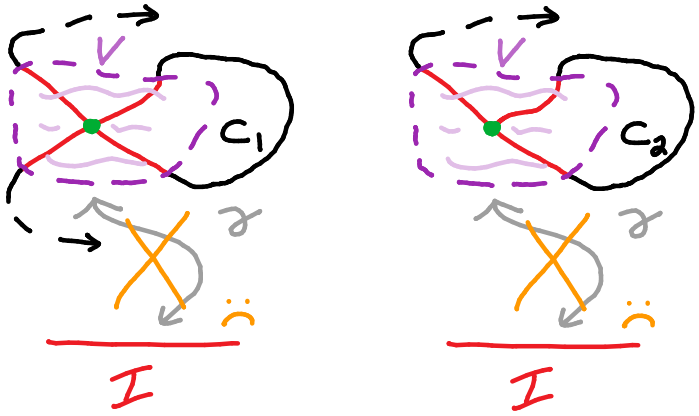


FIGURE 3.9. The object C_1 in the left drawing fails to be a curve, since any open subset V containing the self-intersecting green point includes an "X-figure". Moreover, C_2 in the right drawing is not a curve, as any open subset V containing the green point includes a "T-figure".

Similarly, the figure C_2 in the right drawing of Figure 3.9, resembling a piece of string with one of its ends touching the middle of the string, also fails to be a curve. The idea here is that for any open subset V containing the green point, the intersection $C_2 \cap V$ contains a "T-figure", which is also (topologically) different from an interval.

Remark 3.20. In more advanced terminology, the curves of Definition 3.17, which are not allowed to self-intersect, are called *embedded curves*. This is in contrast to the larger class of *immersed curves*, which are allowed to pass through themselves.

3.4. **Descriptions of Curves.** Now that we have a formal definition of curves, let us next explore some more practical aspects. Here, we discuss simpler ways to construct and describe curves, and we look at some additional examples.

First, recall that in Definition 3.17, the regular parametric curves γ that describe the curve in question are required to be injective. However, in many instances, it is more convenient to instead work with non-injective parametric curves.

Definition 3.21. Let $C \subseteq \mathbb{R}^n$ be a curve. Then, we refer to any regular (not necessarily injective) parametric curve $\gamma : I \to C$ as a parametrisation of C.

Example 3.22. Let us return to the *unit circle* from Example 3.19:

$$\mathcal{C} = \{(x, y) \in \mathbb{R}^2 \mid x^2 + y^2 = 1\}.$$

Recall from Example 3.19 that in order to formally describe $\mathcal{C}$ as a curve, in the sense of Definition 3.17, we need two injective regular parametric curves, e.g. λ_1 and λ_2.

On the other hand, consider the regular parametric curve

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

Recall that the image of γ (see the left side of Figure 2.10) is precisely all of $\mathcal{C}$. Therefore, γ *is*, by Definition 3.21, *a parametrisation of* $\mathcal{C}$.

However, note that γ, in contrast to λ_1 and λ_2, fails to be injective, as it hits each point of $\mathcal{C}$ infinitely many times. (For example, $\gamma(2\pi k) = (1, 0)$ for any $k \in \mathbb{Z}$.)

In summary, $\mathcal{C}$ can be described by a single parametrisation, as long as we do not require it to also be injective. In situations where injectivity is not essential, it is easier to work with γ, rather than with the pair λ_1 and λ_2 from Example 3.19.

In general, there are many different ways to parametrise a curve. One strategy is to set the parameter to represent either the x or the y-coordinate.

Example 3.23. Let us seek other parametrisations of the *unit circle* $\mathcal{C}$ in Example 3.22.

First, if we set the parameter t to be the x-coordinate (that is, $x = t$), then the defining equation $x^2 + y^2 = 1$ for $\mathcal{C}$ forces an equation for the y-coordinate:

$$y^2 = 1 - t^2, \qquad y = \pm\sqrt{1 - t^2}.$$

The above produces two parametrisations of $\mathcal{C}$,

$$\gamma_{x,+} : (-1,1) \to \mathcal{C}, \qquad \gamma_{x,+}(t) = \left(t, +\sqrt{1-t^2}\right),$$
$$\gamma_{x,-} : (-1,1) \to \mathcal{C}, \qquad \gamma_{x,-}(t) = \left(t, -\sqrt{1-t^2}\right),$$

which map out the upper and lower halves of $\mathcal{C}$, respectively. (The domain $(-1,1)$ is the largest open interval in which the defining formulas for $\gamma_{x,\pm}$ make sense.)

Moreover, by setting instead $y = t$, we obtain two more parametrisations of $\mathcal{C}$:

$$\gamma_{y,+} : (-1,1) \to \mathcal{C}, \qquad \gamma_{y,+}(t) = \left(+\sqrt{1-t^2}, t\right),$$
$$\gamma_{y,-} : (-1,1) \to \mathcal{C}, \qquad \gamma_{y,-}(t) = \left(-\sqrt{1-t^2}, t\right).$$

These map out the right and left halves of $\mathcal{C}$, respectively.

The images of $\gamma_{x,\pm}$ and $\gamma_{y,\pm}$ are illustrated in Figure 3.10. Note that $\mathcal{C}$ can be constructed by gluing together the four parametrisations $\gamma_{x,\pm}$ and $\gamma_{y,\pm}$.

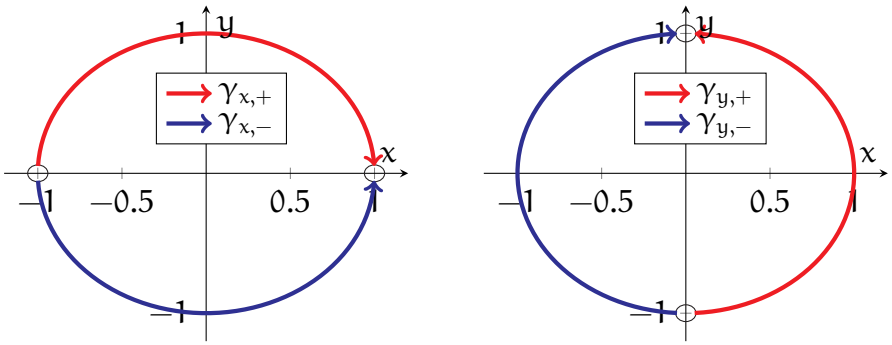


FIGURE 3.10. The two graphics illustrate the four parametrisations $\gamma_{x,\pm}$ and $\gamma_{y,\pm}$ of the unit circle $\mathcal{C}$ from Example 3.23.

Finally, if you wish to think more concretely, then you can view a curve C as a path, and a parametrisation of C as a particle that is moving along the path.

Now, you have probably noticed that, due to the technical nature of Definition 3.17, the process of showing that a set $C \subseteq \mathbb{R}^n$ is a curve is rather cumbersome. Thus, one may ask whether there are more convenient ways to generate examples of curves. The following theorem give one such method for curves in $\mathbb{R}^2$:

Theorem 3.24. Suppose $U \subseteq \mathbb{R}^2$ is open and connected, and let $f : U \to \mathbb{R}$ be a smooth function. In addition, let $c \in \mathbb{R}$, and let C denote the level set

$$C = \{(x,y) \in U \mid f(x,y) = c\}.$$

If $\nabla f(\mathbf{p})$ is nonzero for every $\mathbf{p} \in C$, then C is a curve.

We omit the proof of Theorem 3.24, as it relies on the *implicit function theorem*, a staple of advanced analysis and the theoretical basis for implicit differentiation in calculus. This is, unfortunately, a bit beyond the scope of this module.

Remark 3.25. One can prove a more specific version of Theorem 3.24. Given $\mathbf{p} \in C$:

- If $\partial_1 f(\mathbf{p}) \neq 0$, then one can parametrise C near $\mathbf{p}$ in the form $t \mapsto (t, g(t))$.
- If $\partial_2 f(\mathbf{p}) \neq 0$, then one can parametrise C near $\mathbf{p}$ in the form $t \mapsto (h(t), t)$.

Example 3.26. Consider the (clearly smooth) function

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

Observe that the unit circle $\mathcal{C}$ from Examples 3.19 and 3.22 is a level set of s:

$$\mathcal{C} = \{(x, y) \in U \mid s(x, y) = 1\}.$$

A direct computation shows that the gradient of s satisfies

$$\nabla s(x, y) = (2x, 2y)_{(x,y)}.$$

In particular, ∇s vanishes only when $(x, y) = (0, 0)$. Since $(0, 0) \notin \mathcal{C}$, we see that ∇s is non-vanishing on $\mathcal{C}$. Applying Theorem 3.24 (with $f = s$), we conclude that $\mathcal{C}$ is a curve.

Example 3.27. Consider next the (smooth) function

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

The standard *hyperbola* can be described as a level set of w:

$$H = \{(x, y) \in U \mid x^2 - y^2 = 1\}.$$

See the left plot in Figure 3.11 for an illustration of H.

By a similar computation as in Example 3.26, we see that

$$\nabla w(x, y) = (2x, -2y)_{(x,y)}, \qquad (x, y) \in \mathbb{R}^2,$$

which vanishes only when $(x, y) = (0, 0)$. Since $(0, 0) \notin H$, we can then apply Theorem 3.24 (with $f = w$) to conclude that H is a curve.

In addition, observe that the parametric curve

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

is a parametrisation of H, whose image is precisely the right half of H.

One consequence of Theorem 3.24 is that graphs of real-valued functions are curves:

Theorem 3.28. Let $h : I \to \mathbb{R}$ be smooth, with $I \subseteq \mathbb{R}$ an open interval. Then, both

$$G_h = \{(t, h(t)) \mid t \in I\}, \qquad G_h^* = \{(h(t), t) \mid t \in I\}$$

(i.e. the graph and "inverted graph" of h) are curves.

Proof. We only give the proof for G_h, as the argument for G_h^* is analogous. Let

$$F : I \times \mathbb{R} \to \mathbb{R}, \qquad F(x, y) = y - h(x),$$

and note that F is a smooth function defined on an open subset $I \times \mathbb{R}$ of $\mathbb{R}^2$.

Moreover, for any $(x, y) \in I \times \mathbb{R}$, we have that

$$\nabla F(x, y) = (-h'(x), 1)_{(x,y)} \neq (0, 0)_{(x,y)}.$$

The result now follows from Theorem 3.24, since

$$G_h = \{(x, y) \in I \times \mathbb{R} \mid F(x, y) = 0\}.$$ □

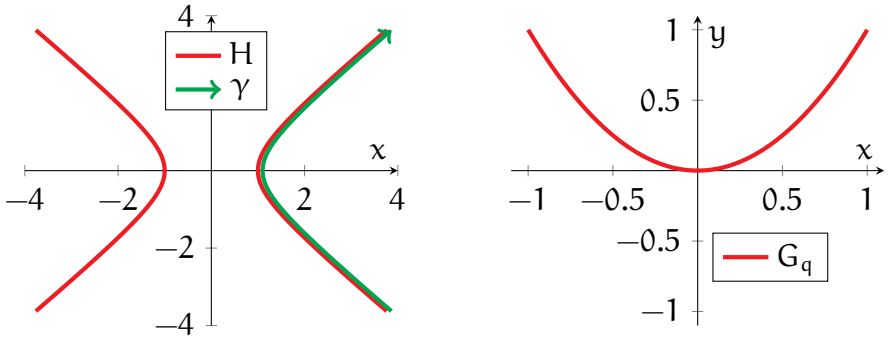


FIGURE 3.11. The left drawing depicts the hyperbola H (in red) and the parametrisation γ (in green) from Example 3.27. The right drawing depicts the parabola G_q (in red) from Example 3.29.

Example 3.29. Consider the quadratic function

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

By Theorem 3.28, the graph of q,

$$G_q = \{(t, t^2) \mid t \in \mathbb{R}\},$$

is a curve, known as a *parabola*. See the right drawing of Figure 3.11 for a plot of G_q.

3.5. **Tangent Lines.** The remainder of this chapter is dedicated to exploring *geometric properties* of curves. In practice, rather than study curves directly, we often measure properties of their parametrisations, which describe the points of the curve in a way that allows us to do calculus. But, this yields another problem: does a given parametric property actually reflect the geometry of the underlying curve?

For example, consider a curve $C \subseteq \mathbb{R}^n$, which one can think of as representing a path. Also, let $\gamma : I \to \mathbb{R}^n$ be a parametrisation of C, modelling a person moving along this path. Then, the derivative $\gamma'(t)$ at any $t \in I$ is certainly a property of γ, representing the person's velocity at a certain point of C.

Note, however, that this velocity carries information beyond the properties of the path itself. In other words, if you were given only the curve C, but not the particular trajectory γ, then you could not recover the information given by $\gamma'(t)$. (In particular, another person running along C could have a rather different velocity.)

Therefore, for something to be a geometric property of C:

- One should be able to extract this value using a parametrisation of C.
- One should obtain the *same* value from *any parametrisation of* C.

In other words, *a geometric property of* C *must be independent of parametrisation.*

Now, while the derivative of γ does not give a geometric property of C, we can, however, use this information to describe a property that is geometric in nature. In order to properly describe this, we first define the following:

Definition 3.30. Let $\gamma : I \to \mathbb{R}^n$ be a regular parametric curve, and let $t_0 \in I$. We define the tangent line to γ at t_0 to be the following subspace (of tangent vectors):

$$T_\gamma(t_0) = \left\{ s \cdot \gamma'(t_0)_{\gamma(t_0)} \middle| s \in \mathbb{R} \right\} \subseteq T_{\gamma(t_0)}\mathbb{R}^n. \tag{3.6}$$

Remark 3.31. Note $T_\gamma(t_0)$ is a vector subspace of $T_{\gamma(t_0)}\mathbb{R}^n$, in the sense that you learned in linear algebra, since $T_\gamma(t_0)$ is the span of the tangent vector $\gamma'(t_0)_{\gamma(t_0)}$.

Let us now make some sense of Definition 3.30. Recall that $\gamma'(t_0)_{\gamma(t_0)}$ represents the arrow starting at the point $\gamma(t_0)$ on γ, and pointing in the direction $\gamma'(t_0)$ along γ. Thus, *the arrows in* $T_\gamma(t_0)$, which are all the dilations of $\gamma'(t_0)_{\gamma(t_0)}$, *trace out a line that is tangent to* γ *at* $\gamma(t_0)$. This intuition justifies the name "tangent line".

Example 3.32. Consider the parametric *unit circle*,

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

At $t_0 = \frac{\pi}{4}$, we compute

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

By Definition 3.30, the tangent line to γ at $\frac{\pi}{4}$ is then given by

$$\begin{aligned} T_\gamma\left(\frac{\pi}{4}\right) &= \left\{ s \cdot \gamma'\left(\frac{\pi}{4}\right)_{\gamma(\frac{\pi}{4})} \middle| s \in \mathbb{R} \right\} \\ &= \left\{ s \cdot \left(-\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right)_{(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}})} \middle| s \in \mathbb{R} \right\}. \end{aligned}$$

Graphical representations of $T_\gamma(\frac{\pi}{4})$ are given in Figure 3.12. In the left illustration, the tangent vector $\gamma'(\frac{\pi}{4})_{\gamma(\frac{\pi}{4})}$ is drawn in blue. The arrows in the middle picture depict some other elements of $T_\gamma(\frac{\pi}{4})$. Lastly, in the right illustration, the tangent line $T_\gamma(\frac{\pi}{4})$ is represented as a purple line, which is traced out by all the arrows in $T_\gamma(\frac{\pi}{4})$.

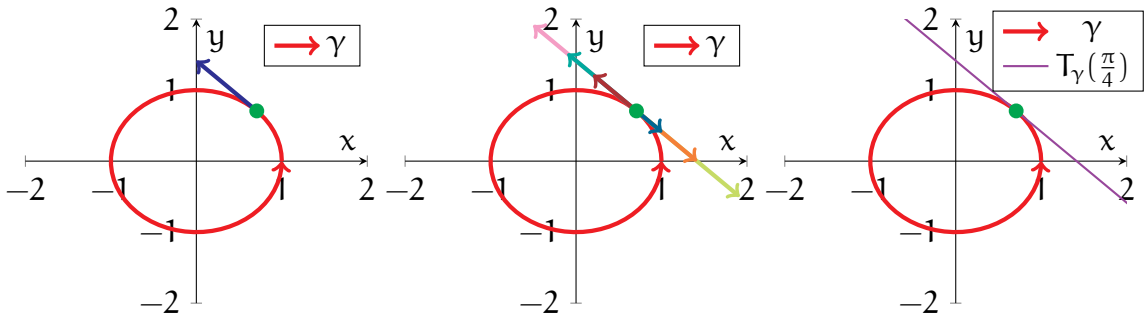


FIGURE 3.12. The pictures show the setup from Example 3.32. In the first plot, the blue arrow shows the tangent vector $\gamma'(\frac{\pi}{4})_{\gamma(\frac{\pi}{4})}$. The middle plot contains some additional elements of $T_\gamma(\frac{\pi}{4})$, drawn as coloured arrows. In the right plot, the tangent line $T_\gamma(\frac{\pi}{4})$ is represented as a purple line.

Example 3.33. Let $f : (-1, 1) \to \mathbb{R}$ be smooth, and consider the parametric curve

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

whose image is the *graph* of f. Then, for any $t \in (-1, 1)$, we have that

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

hence the tangent line of γ_f at t is

$$T_{\gamma_f}(t) = \left\{ s \cdot (1, f'(t))_{(t, f(t))} \middle| s \in \mathbb{R} \right\}.$$

Next, we show how these tangent lines define a geometric property of curves.

Theorem 3.34. Let $\gamma : I \to \mathbb{R}^n$ and $\tilde{\gamma} : \tilde{I} \to \mathbb{R}^n$ be regular parametric curves, and let γ be a reparametrisation of $\tilde{\gamma}$, with change of variables $\phi : I \to \tilde{I}$. Then, for any $t \in I$,

$$T_\gamma(t) = T_{\tilde{\gamma}}(\phi(t)). \tag{3.7}$$

Recall, in the setting of Theorem 3.34, that $\gamma(t)$ and $\tilde{\gamma}(\phi(t))$ refer to the same point. Thus, the formula (3.7) can be interpreted as follows: *at common points along the trajectories of γ and $\tilde{\gamma}$, one obtains identical tangent lines for both γ and $\tilde{\gamma}$.*

Proof of Theorem 3.34. Let us abbreviate $\tilde{t} = \phi(t)$. Then, Theorem 3.16 yields

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

Using Definition 3.30, we conclude that

$$\begin{aligned} T_\gamma(t) &= \left\{ s \cdot \gamma'(t)_{\gamma(t)} \middle| s \in \mathbb{R} \right\} \\ &= \left\{ s \cdot \tilde{\gamma}'(\tilde{t})_{\tilde{\gamma}(\tilde{t})} \middle| s \in \mathbb{R} \right\} \\ &= T_{\tilde{\gamma}}(\tilde{t}), \end{aligned}$$

since every scalar multiple of $\gamma'(t)_{\gamma(t)}$ is a multiple of $\tilde{\gamma}'(\tilde{t})_{\tilde{\gamma}(\tilde{t})}$, and vice versa. □

Though a parametric curve's speed can change after reparametrising, it will always point in one of two (opposite) directions. Thus, by taking all scalar multiples of the tangent vector, (3.6) precisely factors out all information about speed while preserving the directional information that is independent of parametrisation.

Using Theorem 3.34, we can now formally define a geometric property:

Definition 3.35. The tangent line to a curve $C \subseteq \mathbb{R}^n$ at a point $\mathbf{p} \in C$ is defined as

$$T_{\mathbf{p}}C = T_\gamma(t_0), \tag{3.8}$$

where $\gamma : I \to C$ is any parametrisation of C, and where $t_0 \in I$ satisfies $\gamma(t_0) = \mathbf{p}$.

In particular, a tangent line to C is obtained by computing the corresponding tangent line to *any* parametrisation of C. Indeed, Theorem 3.34 guarantees that we will always obtain the same answer regardless of our choice of parametrisation.

Example 3.36. Consider the *helix* $\mathcal{H}$ from Example 3.18 (see also Figure 3.8). Recall from Example 3.18 that the following function is a parametrisation of $\mathcal{H}$:

$$\mathbf{h} : \mathbb{R} \to \mathcal{H}, \qquad \mathbf{h}(t) = (\cos t, \sin t, t).$$

Let us compute the tangent lines to $\mathcal{H}$ at the points $(0, 1, \frac{\pi}{2})$ and $(-1, 0, \pi)$. Note that $(0, 1, \frac{\pi}{2})$ and $(-1, 0, \pi)$ correspond to $t = \frac{\pi}{2}$ and $t = \pi$, respectively:

$$\mathbf{h}\left(\frac{\pi}{2}\right) = \left(0, 1, \frac{\pi}{2}\right), \qquad \mathbf{h}(\pi) = (-1, 0, \pi).$$

Moreover, recall from Example 3.18 that

$$\mathbf{h}'(t) = (-\sin t, \cos t, 1), \qquad t \in \mathbb{R},$$

hence inserting the above values of t yields

$$\mathbf{h}'\left(\frac{\pi}{2}\right) = (-1, 0, 1), \qquad \mathbf{h}'(\pi) = (0, -1, 1).$$

As a result, by Definitions 3.30 and 3.35, the tangent line through $(0, 1, \frac{\pi}{2})$ is

$$\begin{aligned} T_{(0,1,\frac{\pi}{2})}\mathcal{H} &= T_{\mathbf{h}}\left(\frac{\pi}{2}\right) \\ &= \left\{ s \cdot (-1, 0, 1)_{(0,1,\frac{\pi}{2})} \,\middle|\, s \in \mathbb{R} \right\}, \end{aligned}$$

while the tangent line through $(-1, 0, \pi)$ is given by

$$\begin{aligned} T_{(-1,0,\pi)}\mathcal{H} &= T_{\mathbf{h}}(\pi) \\ &= \left\{ s \cdot (0, -1, 1)_{(-1,0,\pi)} \,\middle|\, s \in \mathbb{R} \right\}. \end{aligned}$$

See Figure 3.13 for illustrations of $\mathcal{H}$ and the above tangent lines.

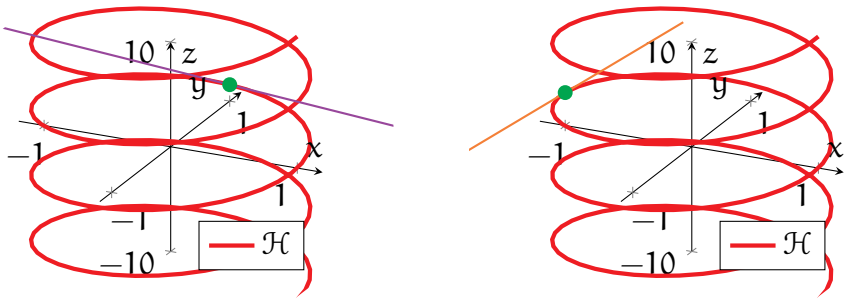


FIGURE 3.13. The two illustrations show the helix $\mathcal{H}$ (in red) from Example 3.36. In the left plot, the purple line is the tangent line to $\mathcal{H}$ at $(0, 1, \frac{\pi}{2})$, while in the right, the orange line is the tangent line to $\mathcal{H}$ at $(-1, 0, \pi)$.

Remark 3.37. In Definition 3.30, we viewed tangent lines as spaces of tangent vectors. An alternative, and likely more familiar, way to define tangent lines is as a set of points. For example, in the setting of Definition 3.30, we could have replaced (3.6) by

$$\mathfrak{T}_\gamma(t) = \{\gamma(t) + s \cdot \gamma'(t) \mid s \in \mathbb{R}\}. \tag{3.9}$$

Observe that the right-hand side of (3.9) is simply the image of the parametric line that passes through $\gamma(t)$ and that is aligned in the direction of $\gamma'(t)$.

The definition (3.9) has the advantage of being more straightforward; in fact, you may have already encountered this in your calculus modules. Moreover, the set $\mathfrak{T}_\gamma(t)$ coincides with the purple and orange lines drawn in Figures 3.12 and 3.13.

On the other hand, the more modern Definitions 3.30 and 3.35 *characterise tangent lines as a set of directions along a curve*, an intuition that will be quite useful in the future. Furthermore, (3.6) highlights the linear structure that is inherent to the tangent line. In particular, *the set* $T_\gamma(t_0)$ *is a* 1*-dimensional vector space* (see Remark 3.31), whereas the set $\mathfrak{T}_\gamma(t_0)$ in (3.9) has no such connection to linear algebra. For these reasons, we adopt Definitions 3.30 and 3.35 as our formal starting points for discussions.

3.6. **Tangents and Normals.** As before, let $C \subseteq \mathbb{R}^n$ be a curve, and consider a point $\mathbf{p} \in C$. We noted that the tangent line $T_\mathbf{p}C$ is a 1-dimensional vector space and hence has the same structure as the real line. Thus, *there are exactly two elements* $\mathbf{t}_\mathbf{p}^\pm \in T_\mathbf{p}C$ *having unit length*, which are *pointing in opposite directions*.

This above observation leads us to the following definition:

Definition 3.38. Let $C \subseteq \mathbb{R}^n$ be a curve, and let $\mathbf{p} \in C$. We define the unit tangents to C at $\mathbf{p}$ to be the two elements $\mathbf{t}_\mathbf{p}^\pm$ of $T_\mathbf{p}C$ such that $|\mathbf{t}_\mathbf{p}^\pm| = 1$.

Example 3.39. Consider the y*-axis* in $\mathbb{R}^3$,

$$L_y = \{(0, t, 0) \mid t \in \mathbb{R}\}.$$

One can show (though we omit this detail here) that L_y is a curve in $\mathbb{R}^3$. Let us now find the unit tangents to L_y at the point $\mathbf{p} = (0, -1, 0) \in L_y$.

The first step is compute the tangent line to L_y at $\mathbf{p}$. For this, note that

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

is a parametrisation of L_y, and that

$$\ell(-1) = (0, -1, 0), \qquad \ell'(-1) = (0, 1, 0).$$

As a result, we see from Definitions 3.30 and 3.35 that

$$\begin{aligned} T_\mathbf{p}L_y &= T_\ell(-1) \\ &= \{(0, s, 0)_{(0,-1,0)} \mid s \in \mathbb{R}\}. \end{aligned}$$

By Definition 3.38, the unit tangents to L_y at $(0,-1,0)$ are the two elements of $T_{\mathbf{p}}L_y$ with unit length. Clearly, these are given by the tangent vectors

$$\mathbf{t}^{\pm}_{(0,-1,0)} = (0,\pm 1,0)_{(0,-1,0)}.$$

See the left plot in Figure 3.14 for an illustration of the above—the two unit tangents to L_y at $\mathbf{p}$ are depicted as purple and blue arrows.

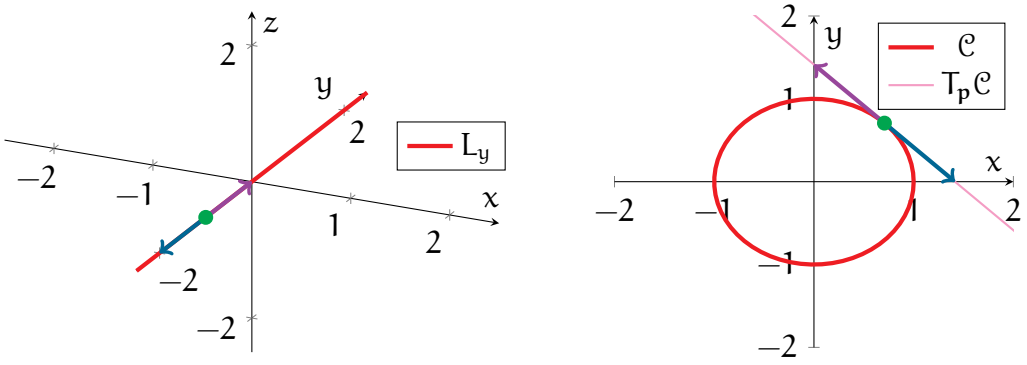


FIGURE 3.14. The graphics contain the settings of Examples 3.39 (left) and 3.42 (right). Furthermore, in both plots, the unit tangents at the given points $\mathbf{p}$ are drawn as purple and blue arrows.

Let us now return to the general curve $C \subseteq \mathbb{R}^n$ and point $\mathbf{p} \in C$. Recall (see Figures 3.12 and 3.13) that $T_{\mathbf{p}}C$ can be represented as a line in $\mathbb{R}^n$ that intersects C at $\mathbf{p}$, as well as points along the directions of C at $\mathbf{p}$. As a result, the *unit tangents to* C *at* $\mathbf{p}$, which point in opposite directions along this line, *correspond to the two directions that one can go along* C *from the point* $\mathbf{p}$.

To be more concrete, consider the right plot in Figure 3.14, illustrating the case where C is a circle in $\mathbb{R}^2$, and where $\mathbf{p}$ is the green point. Here, the purple and blue arrows, representing unit tangents $\mathbf{t}^{\pm}_{\mathbf{p}}$ to C at $\mathbf{p}$, point in opposite tangent directions along C. In particular, the purple arrow captures the anticlockwise direction along C, while the blue arrow corresponds to the clockwise direction.

Remark 3.40. The above also suggests another interpretation for $T_{\mathbf{p}}C$ itself. Since $T_{\mathbf{p}}C$ consists of all scalar multiples of the unit tangents $\mathbf{t}^{\pm}_{\mathbf{p}}$, it follows that $T_{\mathbf{p}}C$ is a combination of all the directions and speeds that one could move along C at $\mathbf{p}$. In other words, $T_{\mathbf{p}}C$ *represents all the possible velocities one could have at* $\mathbf{p}$ *while moving along* C.

Since we often work with curves through their parametrisations, it will be convenient to have a parametric formula for computing unit tangents:

Theorem 3.41. Let $C \subseteq \mathbb{R}^n$ be a curve, let $\gamma : I \to C$ be a parametrisation of C, and let $t_0 \in I$. Then, the unit tangents to C at the point $\mathbf{p} = \gamma(t_0)$ are given by

$$\pm |\gamma'(t_0)|^{-1} \cdot \gamma'(t_0)_{\gamma(t_0)}. \tag{3.10}$$

Proof. The tangent vectors $\mathbf{t}_\mathbf{p}^\pm = \pm|\gamma'(t_0)|^{-1} \cdot \gamma'(t_0)_{\gamma(t_p)}$ lie in $T_\mathbf{p}C$ by definition and clearly have unit length, thus they are unit tangents to C at $\mathbf{p}$. □

Example 3.42. Next, let $\mathcal{C}$ be the *unit circle* from Example 3.19. Recall, from Example 3.22, that the following function is a parametrisation of $\mathcal{C}$:

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

Let us find the unit tangents to $\mathcal{C}$ at the point

$$\mathbf{p} = \left(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right) = \gamma\left(\frac{\pi}{4}\right).$$

For this, we simply apply Theorem 3.41 using the above γ (with $t_0 = \frac{\pi}{4}$) to obtain

$$\begin{aligned} \pm\left|\gamma'\left(\frac{\pi}{4}\right)\right|^{-1} \cdot \gamma'\left(\frac{\pi}{4}\right)_{\gamma(\frac{\pi}{4})} &= \pm 1 \cdot \left(-\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right)_{\left(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right)} \\ &= \pm\left(-\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right)_{\left(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right)}. \end{aligned}$$

See the right plot of Figure 3.14; the unit tangents are drawn in purple and blue.

Next, *we restrict our attention only to the case* $n = 2$. Given a curve $C \subseteq \mathbb{R}^2$ and $\mathbf{p} \in C$, the tangent line $T_\mathbf{p}C$ can be viewed as a line sitting within a 2-dimensional plane. More formally, $T_\mathbf{p}C$ is a 1-dimensional subspace of the 2-dimensional vector space $T_\mathbf{p}\mathbb{R}^2$. Thus, there is one remaining dimension in $T_\mathbf{p}\mathbb{R}^2$ perpendicular to $T_\mathbf{p}C$, capturing the directions perpendicular, or *normal*, to C at $\mathbf{p}$.

As before, we focus on the *unit* vectors in these normal directions:

Definition 3.43. Let $C \subseteq \mathbb{R}^2$ be a curve, and let $\mathbf{p} \in C$. Then, $\mathbf{n}_\mathbf{p} \in T_\mathbf{p}\mathbb{R}^2$ is a unit normal to C at $\mathbf{p}$ iff $\mathbf{n}_\mathbf{p}$ is perpendicular to every element of $T_\mathbf{p}C$ and $|\mathbf{n}_\mathbf{p}| = 1$.

The unit normals to a curve can be obtained from the unit tangents:

Theorem 3.44. Let $C \subseteq \mathbb{R}^2$ be a curve, and let $\mathbf{p} \in C$. If $\pm(v_1, v_2)_\mathbf{p} \in T_\mathbf{p}C$ are the unit tangents to C at $\mathbf{p}$, then the unit normals to C at $\mathbf{p}$ are given by

$$\mathbf{n}_\mathbf{p}^\pm = \pm(-v_2, v_1)_\mathbf{p}. \tag{3.11}$$

Proof. Since (v_1, v_2) is a unit vector, then a direct computation shows that

$$\pm(-v_2, v_1)_{\mathbf{p}} \cdot \pm(v_1, v_2)_{\mathbf{p}} = 0, \qquad |\pm(-v_2, v_1)_{\mathbf{p}}| = 1,$$

and hence $\mathbf{n}_{\mathbf{p}}^{\pm}$ are indeed unit normals to C at $\mathbf{p}$. Since there is only one normal dimension, then $\mathbf{n}_{\mathbf{p}}^{\pm}$ must also be the only unit normals to C at $\mathbf{p}$. □

Remark 3.45. The intuition behind (3.11) is that the direction $(-v_2, v_1)$ is precisely a 90 degree rotation from (v_1, v_2). A slick way to see this is through complex numbers—an anticlockwise rotation of the complex number $v_1 + iv_2$ by 90 degrees yields

$$\begin{aligned} e^{i\cdot\frac{\pi}{2}}(v_1 + iv_2) &= i(v_1 + iv_2) \\ &= -v_2 + iv_1. \end{aligned}$$

Example 3.46. Consider again the *circle* $\mathcal{C}$ and the point $\mathbf{p} \in \mathcal{C}$ from Example 3.42:

$$\mathcal{C} = \{(x, y) \in \mathbb{R}^2 \mid x^2 + y^2 = 1\}, \qquad \mathbf{p} = \left(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right).$$

Recall, from Example 3.42, that unit tangents to $\mathcal{C}$ at $\mathbf{p}$ are

$$\pm\left(-\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right)_{\left(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right)}.$$

Then, by Theorem 3.44, the unit normals to $\mathcal{C}$ at $\mathbf{p}$ are given by

$$\pm\left(-\frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}}\right)_{\left(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right)} = \pm\left(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right)_{\left(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right)}.$$

These unit normals to $\mathcal{C}$ are illustrated (in orange and blue) in the left plot of Figure 3.15.

When the curve C is also a level set, in the sense of Theorem 3.24, there is another method with which one can compute the unit normals directly:

Theorem 3.47. Assume the setting of Theorem 3.24—let C be the curve

$$C = \{(x, y) \in U \mid f(x, y) = c\},$$

where $U \subseteq \mathbb{R}^2$, $c \in \mathbb{R}$, and $f : U \to \mathbb{R}$ is a smooth function such that $\nabla f(\mathbf{q})$ is nonzero for each $\mathbf{q} \in C$. Then, given any $\mathbf{p} \in C$, the unit normals to C at $\mathbf{p}$ are

$$\mathbf{n}_{\mathbf{p}}^{\pm} = \pm|\nabla f(\mathbf{p})|^{-1} \cdot \nabla f(\mathbf{p}). \tag{3.12}$$

Proof. Let $\gamma : I \to C$ be a parametrisation of C, whose values can be expanded as

$$\gamma(t) = (x(t), y(t)), \qquad t \in I,$$

and suppose $\gamma(t_0) = \mathbf{p}$ for some $t_0 \in I$. Since $f(\gamma(t)) = c$ for all $t \in I$, then taking a derivative (with respect to t) and applying the chain rule yields

$$\begin{aligned} 0 &= \frac{d}{dt}[f(\gamma(t))] \\ &= \partial_1 f(\gamma(t)) \cdot x'(t) + \partial_2 f(\gamma(t)) \cdot y'(t) \\ &= \nabla f(\gamma(t)) \cdot \gamma'(t)_{\gamma(t)}. \end{aligned}$$

Multiplying the preceding equation by any $s \in \mathbb{R}$ and taking $t = t_0$ yields

$$0 = \nabla f(\mathbf{p}) \cdot \left[s\gamma'(t_0)_{\gamma(t_0)} \right].$$

Recalling Definitions 3.30 and 3.35, we conclude that $\pm\nabla f(\mathbf{p})$ is perpendicular to every element of $T_\gamma(t_0) = T_{\mathbf{p}}C$. Finally, dividing $\pm\nabla f(\mathbf{p})$ by its norm yields the unit normals to C at $\mathbf{p}$, by Definition 3.43, and completes the proof. □

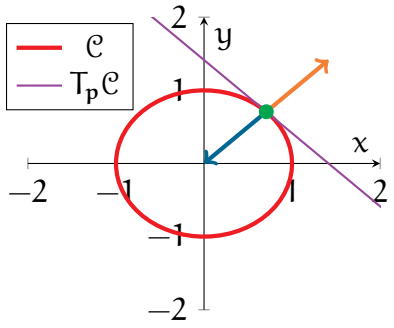


FIGURE 3.15. This graphic shows the setting of Examples 3.46 and 3.48. The unit normals to $\mathcal{C}$ are drawn, in orange and blue, at the green point.

Example 3.48. Let the circle $\mathcal{C}$ and the point $\mathbf{p}$ be as in Example 3.46. We now compute the unit normals to $\mathcal{C}$ at $\mathbf{p}$ once again, using Theorem 3.47.

Recall that $\mathcal{C}$ is a level set of the (smooth) function

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

Moreover, recalling the computations for s in Example 3.26, we see that

$$\nabla s(\mathbf{p}) = \left(\frac{2}{\sqrt{2}}, \frac{2}{\sqrt{2}} \right)_{\mathbf{p}}, \qquad |\nabla s(\mathbf{p})| = 2.$$

Thus, by Theorem 3.47 (with $f = s$), we see that the unit normals to $\mathcal{C}$ at $\mathbf{p}$ are

$$\pm|\nabla s(\mathbf{p})|^{-1} \cdot \nabla s(\mathbf{p}) = \pm\left(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}} \right)_{\left(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right)}.$$

3.7. **Orientation.** Our next topic, closely related to unit tangents, is the *orientation* of a curve, that is, the *choice of a direction that one goes along a curve.*

Example 3.49. Again, consider the y-*axis* L_y from Example 3.39:

$$L_y = \{(0, t, 0) \mid t \in \mathbb{R}\}.$$

We can traverse along L_y in two directions—with increasing or decreasing y-values. See Figure 3.16 for an illustration; the red path on the left shows the direction of increasing y-value, while the green path on the right shows the direction of decreasing y-value.

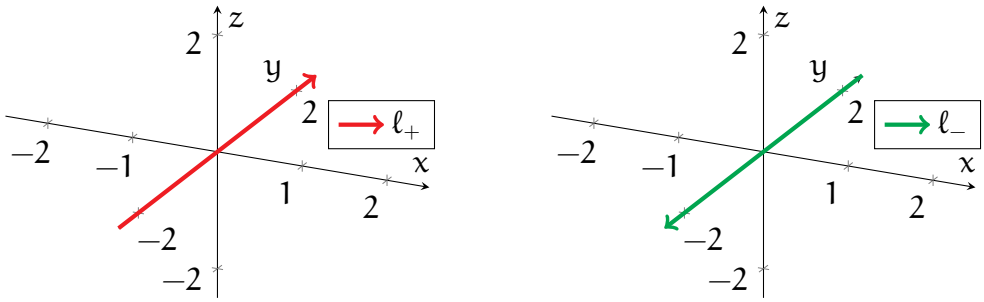


FIGURE 3.16. The two plots show two ways to traverse the y-axis L_y from Example 3.49. Moreover, the red and green paths are represented by the parametrisations $\ell_\pm$ of L_y, from Example 3.54.

Orientation plays an essential role in several geometric computations. In fact, it is connected to a number of concepts that are already quite familiar to you:

- A simple example comes from basic calculus, from which you know that
$$\int_0^1 dx = +1, \qquad \int_1^0 dx = -1.$$
The two different answers with opposite signs arise from the fact that the above integrals use two different orientations of the unit interval $[0, 1]$.
- In $\mathbb{R}^2$, one typically defines angles anticlockwise from the positive x-axis to be positive. However, it is equally valid to instead define clockwise angles as positive. The convention of anticlockwise angles being positive arises from choosing the anticlockwise-traversing orientation of the unit circle.

We now explore how orientation can be captured more rigorously. Consider again a curve $C \subseteq \mathbb{R}^n$. Recall that at any $\mathbf{p} \in C$, the unit tangents $\mathbf{t}_\mathbf{p}^\pm$ to C at $\mathbf{p}$ represent the two directions that you can travel along C from $\mathbf{p}$. Thus, by selecting either $\mathbf{t}_\mathbf{p}^+$ or $\mathbf{t}_\mathbf{p}^-$, you are essentially choosing a particular direction along C to move.

Now, we should keep in mind that the above selection was only for a single point. To settle on a direction along all of C, *we must choose a unit tangent* $\mathbf{t_q} \in T_qC$ *at each point* $\mathbf{q} \in C$. Furthermore, *our choices of directions must be consistent with each other*, in that one cannot select one direction at some point and then suddenly jump to the opposite direction as one slides along C.

All these considerations lead to the following definition:

Definition 3.50. An orientation of a curve $C \subseteq \mathbb{R}^n$ is a choice of a unit tangent $\mathbf{t_p} \in T_pC$ to C at each $\mathbf{p} \in C$, such that the $\mathbf{t_p}$'s vary continuously with respect to $\mathbf{p}$.

Example 3.51. We can now discuss more precisely the y-*axis* L_y from Example 3.49:

$$L_y = \{(0, t, 0) \mid t \in \mathbb{R}\}.$$

Recall (see Example 3.39) that the following is an injective parametrisation of (all of) L_y:

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

Moreover, at any point $(0, t_0, 0) \in L_y$, we have that

$$\ell(t_0) = (0, t_0, 0), \qquad \ell'(t_0) = (0, 1, 0), \qquad |\ell'(t_0)| = 1.$$

Thus, by Theorem 3.41, we conclude that the unit tangents to L_y at $(0, t_0, 0)$ are

$$\pm|\ell'(t_0)|^{-1} \cdot \ell'(t_0)_{\ell(t_0)} = (0, \pm 1, 0)_{(0,t_0,0)}.$$

From this, we see that one orientation of L_y can be constructed by selecting the unit tangent vector $(0, +1, 0)_{(0,t,0)}$ at each $(0, t, 0) \in L_y$; this corresponds to the direction of increasing y-value. The remaining orientation of L_y, representing the direction of decreasing y-value, is captured by the choice of $(0, -1, 0)_{(0,t,0)}$ at each $(0, t, 0) \in L_y$.

Note one cannot jump from $(0, +1, 0)$ to $(0, -1, 0)$ anywhere on L_y, or vice versa, as Definition 3.56 requires our choice of unit tangent vectors to vary continuously.

Orientations of curves can also be described using parametrisations. Roughly, a parametrisation $\gamma : I \to C$ of a curve $C \subseteq \mathbb{R}^n$ selects a direction of traversal along C. To be more precise, γ creates a natural choice of unit tangents

$$+|\gamma'(t)|^{-1} \cdot \gamma'(t)_{\gamma(t)}, \qquad t \in I, \tag{3.13}$$

to C (by Theorem 3.41), corresponding to the direction γ travels along C.

With the above motivation, we now make the following definition:

Definition 3.52. Let $C \subseteq \mathbb{R}^n$ be a curve, and let O be an orientation of C.

- A parametrisation $\gamma : I \to C$ of C generates the orientation O iff the quantities $|\gamma'(t)|^{-1} \cdot \gamma'(t)_{\gamma(t)}$ coincide with O for all $t \in I$.
- A parametrisation $\gamma : I \to C$ of C generates an orientation opposite to O iff the quantities $|\gamma'(t)|^{-1} \cdot \gamma'(t)_{\gamma(t)}$ do not coincide with O for any $t \in I$.

Remark 3.53. One important fact (which we will not prove here) is that for any curve $C \subseteq \mathbb{R}^n$, *there always exists an orientation of* C—that is, *every curve is orientable*. As we shall see later on, the analogous statement does not hold for surfaces.

Example 3.54. We return one more time to the y-*axis* L_y from Examples 3.49 and 3.51. Notice that the following two functions are parametrisations of L_y:

$$\ell_\pm : \mathbb{R} \to L_y, \qquad \ell_\pm(t) = (0, \pm t, 0).$$

See Figure 3.16 for plots of ℓ_+ (in red) and ℓ_- (in green).

Starting with ℓ_+, we see that for any $t \in \mathbb{R}$,

$$|\ell_+'(t)|^{-1} \cdot \ell_+'(t)_{\ell_+(t)} = (0, +1, 0)_{(0,t,0)}.$$

Thus, by Definition 3.52, we have that ℓ_+ generates the orientation associated with the direction of increasing y-value. Similarly, for ℓ_-, we obtain that

$$|\ell_-'(t)|^{-1} \cdot \ell_-'(t)_{\ell_-(t)} = (0, -1, 0)_{(0,-t,0)},$$

so ℓ_2 generates the opposite orientation, in the direction of decreasing y-value.

Example 3.55. Consider again the *circle* from Example 3.19:

$$\mathcal{C} = \{(x, y) \in \mathbb{R}^2 \mid x^2 + y^2 = 1\}.$$

Observe that the following are parametrisations of $\mathcal{C}$ (see also Example 3.22):

$$\gamma_1 : \mathbb{R} \to \mathcal{C}, \qquad \gamma_1(t) = (\cos t, \sin t),$$
$$\gamma_2 : \mathbb{R} \to \mathcal{C}, \qquad \gamma_2(t) = (\cos t, -\sin t).$$

Plots of γ_1 and γ_2 are given in Figure 3.17. Note that γ_1 is traversing in the anticlockwise direction along $\mathcal{C}$, while γ_2 is traversing in the clockwise direction. As a result, γ_1 and γ_2 generate the anticlockwise and clockwise orientations of $\mathcal{C}$, respectively.

To be more specific, at the points

$$\gamma_1(t) = (\cos t, \sin t), \qquad t \in \mathbb{R},$$

the parametrisation γ_1 generates the unit tangents

$$|\gamma_1'(t)|^{-1} \cdot \gamma_1'(t)_{\gamma_1(t)} = (-\sin t, \cos t)_{(\cos t, \sin t)}.$$

Some arrows representing the orientation generated by γ_1 are drawn in the left half of Figure 3.17. An analogous picture for γ_2 is given in the right half of Figure 3.17.

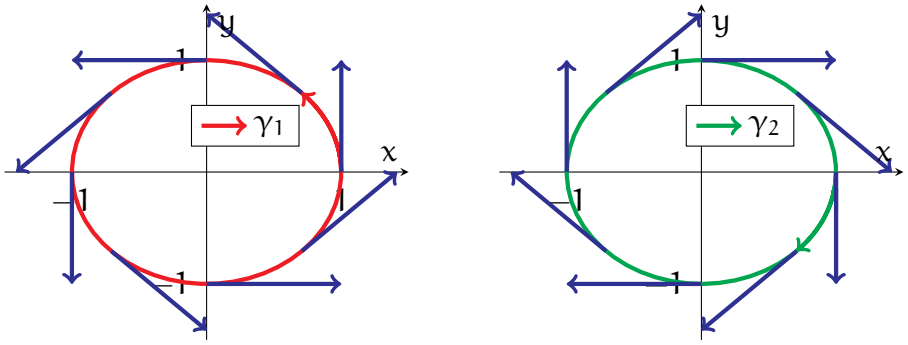


FIGURE 3.17. The left and right illustrations contain plots of the parametrisations γ_1 and γ_2, respectively, from Example 3.55. Each drawing also contains arrows (in blue) showing the associated orientation of the unit circle.

To conclude, we make formal the notion of a "curve with a direction of travel":

Definition 3.56. An oriented curve is a curve C, along with a chosen orientation of C.

Example 3.57. Returning to the setting from Example 3.55, the unit circle $\mathcal{C}$ along with the anticlockwise direction, which is generated by γ_1, defines an oriented curve.

The above gives one of two oriented curves made from $\mathcal{C}$. The other is constructed by pairing $\mathcal{C}$ with the clockwise orientation that is generated by γ_2.

3.8. **Parametric Integration.** Up to this point, the geometric properties of curves that we have discussed have been differential in nature. In the rest of this chapter, we turn our attention toward integral properties, representing various notions of "size".

A natural starting point is the following question:

Question 3.58. Given a curve, how do we define and compute its length?

Let us begin our discussion of Question 3.58 with something familiar. If we are given a line segment ℓ, say from a point $\mathbf{p}$ to another point $\mathbf{q}$, then what the length of ℓ should be is clear. Indeed, ℓ can be represented by the arrow from $\mathbf{p}$ to $\mathbf{q}$, the tangent vector $(\mathbf{q}-\mathbf{p})_{\mathbf{p}}$, and the length of ℓ is just the length of this arrow:

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

Now, for curved objects, we adopt a common strategy from integral calculus: we *approximate the curve as a finite number of line segments*. To implement this more precisely, let us first work at the parametric level. Let $\gamma : (a, b) \to \mathbb{R}^n$ be a parametric curve, and choose a finite number of points along γ:

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

See the plots in Figure 3.18 for an example of this situation; the image of γ is drawn in red, while the sample points $\gamma(t_i)$ 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 3.18, this is drawn as 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 is then

$$\begin{aligned} L(\Gamma) &= \sum_{i=1}^{N} |\gamma(t_i) - \gamma(t_{i-1})| \\ &= \sum_{i=1}^{N} \frac{|\gamma(t_i) - \gamma(t_{i-1})|}{|t_i - t_{i-1}|} \cdot \Delta t_i, \end{aligned} \tag{3.14}$$

where $\Delta t_i = t_i - t_{i-1}$ for each $1 \le i \le N$.

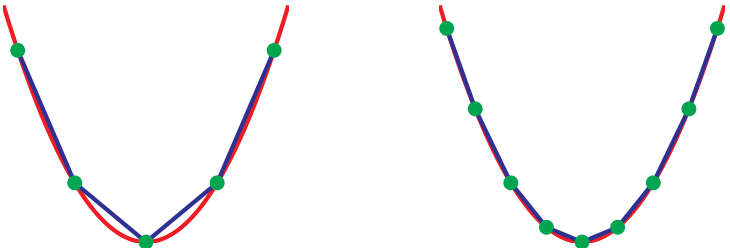


FIGURE 3.18. To approximate the length of the red parabolic curve, we pick a finite sample of points (in green) on the curve, and we compute the total length of the line segments (in blue) connecting the sampled points. In terms of (3.14), we took $N = 4$ on the left plot and $N = 8$ on the right.

Of course, $L(\Gamma)$ is only an approximation of the length of γ. If we are not satisfied with this, then can we refine our process. Choosing a larger number of points along γ (i.e. by increasing N) that are closer to each other, the resulting line segments yield a better approximation of γ than before—see the two plots in Figure 3.18. For this new Γ, we can calculate, as in (3.14), a better approximation of the length of γ.

Now, to obtain the *exact* length of γ, we resort to an "infinitely good" approximation. We let the number N of sample points tend to infinity, and we let the distance Δt_i between successive t_i's 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 $\gamma'(t_i)$. While we do not have the background for discussing this rigorously, the rough idea is that in taking the above limits, we obtain

$$\begin{aligned} \lim_{\substack{N\to\infty \\ \Delta t\to 0}} L(\Gamma) &= \lim_{\substack{N\to\infty \\ \Delta t\to 0}} \sum_{i=1}^{N} \frac{|\gamma(t_i) - \gamma(t_{i-1})|}{|t_i - t_{i-1}|} \cdot \Delta t_i \\ &= \int_a^b |\gamma'(t)| dt. \end{aligned} \tag{3.15}$$

Remark 3.59. Those interested in more rigorous developments of (3.15) should consult material from the module *MTH5105: Differential and Integral Analysis*.

In summary, the above argument motivates our next definition:

Definition 3.60. Given a parametric curve $\gamma : (a, b) \to \mathbb{R}^n$, we define its arc length by

$$L(\gamma) = \int_a^b |\gamma'(t)| dt. \tag{3.16}$$

We now apply Definition 3.60 to justify two common formulas:

Example 3.61. Fix $(x_0, y_0) \in \mathbb{R}^2$, and consider the parametric curve

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

Note that the image of ℓ is the *line segment* connecting the origin (when $t \searrow 0$) and the point (x_0, y_0) (when $t \nearrow 1$); this is shown in the left plot of Figure 3.19.

To calculate the length of ℓ, we first compute, for any $t \in (0, 1)$,

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

Since x_0, y_0 are constants, then by Definition 3.60, the arc length of ℓ satisfies

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

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

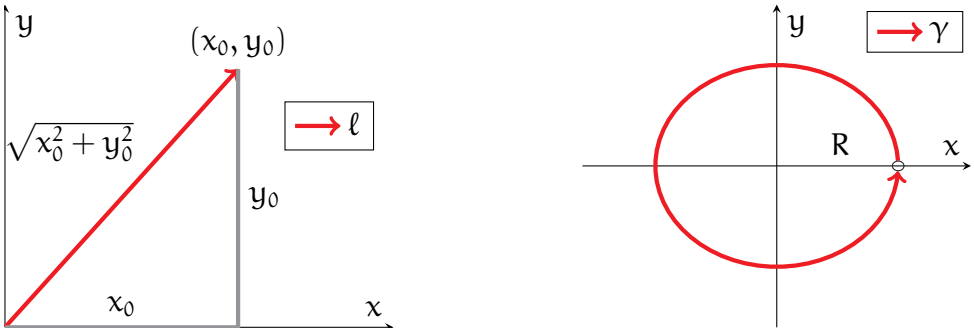


FIGURE 3.19. The left plot contains the setting from Example 3.61, while the right plot contains the setting from Example 3.62.

Example 3.62. Fix $R > 0$, and consider the parametric curve

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

which describes (one revolution of) a *circle of radius* R about the origin; see Figure 3.19.

A direct computation yields the speed of γ:

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

Consequently, using (3.16), we obtain the familiar circumference formula:

$$\begin{aligned} L(\gamma) &= R\int_0^{2\pi} dt \\ &= 2\pi R. \end{aligned}$$

Next, recall from our calculus revisions that the integrals

$$\int_a^b 1\, dx = b - a, \qquad \int_a^b f(x)\, dx$$

can be interpreted as the length of the interval (a, b) and a "weighted length" of (a, b), respectively, with the "weight" in the second expression given by the integrand f.

Having defined lengths of parametric curves, we ask whether the preceding ideas extend to curved settings. More specifically, we explore whether one can construct a similar notion of *integration on a parametric curve* $\gamma : (a, b) \to \mathbb{R}^n$ such that:

(1) Integrating the constant 1 over γ gives the arc length of γ:

$$\int_\gamma 1\, ds = \int_a^b |\gamma'(t)| dt. \tag{3.17}$$

(2) Integrating a function F over γ gives a weighted length of γ, with weight F.

For point (2), it makes sense, in light of the right-hand side of (3.17), to insert F into the integrand there, as this fits with the intuition of F being a weight. Moreover, since $|\gamma'(t)|$ represents the speed of γ *at* $\gamma(t)$, it would make sense that *the weight* F *is also being applied at the same point* $\gamma(t)$. This leads us to a formal definition:

Definition 3.63. Let $\gamma : (a, b) \to \mathbb{R}^n$ be a parametric curve, and let F be a real-valued function that is defined on the image of γ. We define the curve integral of F over γ by

$$\int_\gamma F\,ds = \int_a^b F(\gamma(t))|\gamma'(t)|dt. \tag{3.18}$$

Remark 3.64. Curve integrals are also commonly known as *path* or *line integrals*.

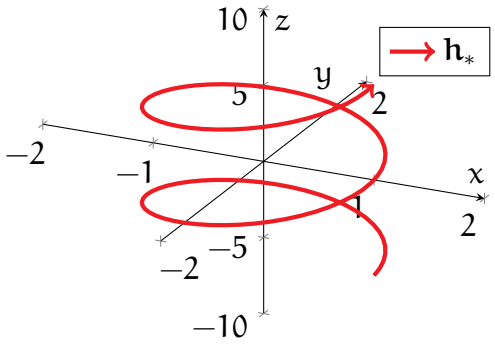


FIGURE 3.20. The above is a plot of $\mathbf{h}_*$ from Example 3.65.

Example 3.65. Consider a finite portion of the parametric *helix* from Example 3.18:

$$\mathbf{h}_* : (-2\pi, 2\pi) \to \mathbb{R}^3, \qquad \mathbf{h}_*(t) = (\cos t, \sin t, t),$$

In particular, the image of $\mathbf{h}_*$ is part a helix that revolves twice around the z-axis; see Figure 3.20. Let us now compute the integral over $\mathbf{h}_*$ of the function

$$G : \mathbb{R}^3 \to \mathbb{R}, \qquad G(x, y, z) = x^2 + y^2 + z^2.$$

First, observe that direct computations yield

$$|\mathbf{h}_*'(t)| = \sqrt{2}, \qquad G(\mathbf{h}_*(t)) = 1 + t^2.$$

As a result, applying Definition 3.63, we obtain

$$\begin{aligned}\int_{\mathbf{h}_*} G\,ds &= \sqrt{2}\int_{-2\pi}^{2\pi}(1 + t^2)dt \\ &= \sqrt{2}\left[4\pi + \frac{2}{3}(2\pi)^3\right].\end{aligned}$$

3.9. **Curve Integrals.** We had previously hinted that the arc length is a geometric property of curves. However, we have yet to justify this statement. To see why this might be true, let us first return to an earlier example:

Example 3.66. Recall the parametric curves from Example 3.11 (and Figure 3.4),

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

which both map out the *upper half of the unit circle*. In addition, since $|\gamma_1'(t)| = 1$ for all $t \in (0,\pi)$, then Definition 3.60 implies that the arc length of γ_1 is

$$\begin{aligned} L(\gamma_1) &= \int_0^\pi dt \\ &= \pi. \end{aligned}$$

The computation for the length of γ_2 is similar but a bit tricker. First, we compute

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

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

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

Thus, we obtain the same arc length π for γ_2 as we did for γ_1!

You may have noticed that the trigonometric substitution $t = -\cos u$ in Example 3.66 is the same as the change of variables (3.1) that linked γ_1 and γ_2. This is, of course, no accident, and it plays a key role in establishing general independence of parametrisation for arc length. In fact, we will be even more ambitious here—we establish this for not only the arc length, but for general curve integrals:

Theorem 3.67. Let $\gamma : (a, b) \to \mathbb{R}^n$ and $\tilde{\gamma} : (\tilde{a}, \tilde{b}) \to \mathbb{R}^n$ be regular parametric curves, and suppose that γ and $\tilde{\gamma}$ are reparametrisations of each other. Then, given any real-valued function F that is defined on the images of γ and $\tilde{\gamma}$, we have that

$$\int_\gamma F\,ds = \int_{\tilde{\gamma}} F\,ds. \tag{3.19}$$

In particular, γ and $\tilde{\gamma}$ have the same arc length: $L(\gamma) = L(\tilde{\gamma})$.

Proof. Let $\phi : (a, b) \to (\tilde{a}, \tilde{b})$ be the change of variables satisfying

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

Recall that by Theorem 3.16, we have

$$|\gamma'(t)| = |\phi'(t)||\tilde{\gamma}'(\phi(t))|, \qquad t \in (a, b).$$

As a result of the above, we see that

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

We now apply the substitution $\tilde{t} = \phi(t)$ and $d\tilde{t} = \phi'(t)\,dt$. From this, we obtain

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

which, by Definition 3.63, is precisely (3.19). (Note that some care is needed in the above, as the direction of integration is reversed when $\phi' < 0$.)

Finally, taking $F \equiv 1$ in (3.19) results in the remaining identity $L(\gamma) = L(\tilde{\gamma})$. □

Thanks to Theorem 3.67, it now makes sense to speak of *curve integrals over curves* and *arc lengths of curves*, rather than merely parametric curves:

Definition 3.68. Let $C \subseteq \mathbb{R}^n$ be a curve, and let $\gamma : (a, b) \to C$ be any injective parametrisation of C, whose image differs from C by only a finite number of points.

- For a real-valued function F defined on C, we define its curve integral over C as

$$\int_C F\,ds = \int_\gamma F\,ds. \tag{3.20}$$

- Moreover, we define the arc length of C by $L(C) = L(\gamma)$.

In particular, Definition 3.68 guarantees that *we can use any parametrisation to compute an integral over* C, as long as it is injective and its image matches C (with the possible exception of a finite number of points).

Example 3.69. Let $\mathcal{C}_+$ denote the *upper half* of the *unit circle* about the origin:

$$\mathcal{C}_+ = \{(x, y) \in \mathbb{R}^2 \mid x^2 + y^2 = 1,\ y > 0\}.$$

(One can apply Theorem 3.24 to show that $\mathcal{C}_+$ is indeed a curve, though we omit details here.) See the green curve in the left plot in Figure 3.21 for an illustration of $\mathcal{C}_+$.

Observe both γ_1 and γ_2 from Example 3.66 are injective parametrisations of $\mathcal{C}_+$. Moreover, $\mathcal{C}_+$ is precisely the image of both γ_1 and γ_2. As a result, Definition 3.68 states that one can use either γ_1 or γ_2 to compute the arc length of the upper semicircle $\mathcal{C}_+$:

$$L(\mathcal{C}_+) = L(\gamma_1) = \pi, \qquad L(\mathcal{C}_+) = L(\gamma_2) = \pi.$$

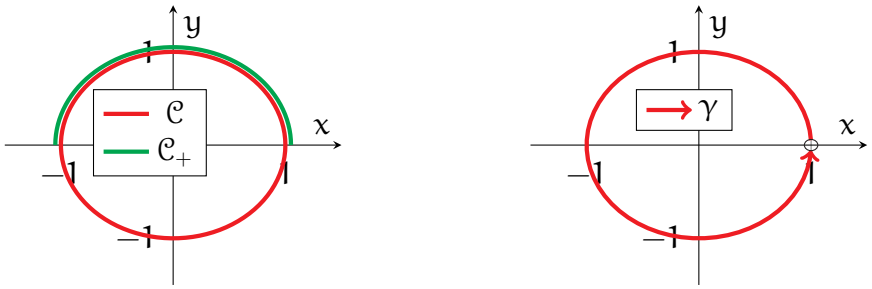


FIGURE 3.21. The left drawing shows both the unit circle $\mathcal{C}$ (in red) and the upper unit semicircle $\mathcal{C}_+$ (in green) from Examples 3.69 and 3.70. The right plot depicts the parametric curve γ from Examples 3.71 and 3.74.

Next, we show why the condition in Definition 3.68 that γ *is injective* is important:

Example 3.70. Let $\mathcal{C}$ be the *unit circle* from Example 3.19:

$$\mathcal{C} = \{(x, y) \in \mathbb{R}^2 \mid x^2 + y^2 = 1\}.$$

(See the left drawing in Figure 3.21.) Consider also the regular parametric curve

$$\lambda : (0, 4\pi) \to \mathcal{C}, \qquad \lambda(t) = (\cos t, \sin t),$$

which is a parametrisation of $\mathcal{C}$. We can then compute

$$\begin{aligned} L(\lambda) &= \int_0^{4\pi} |\lambda'(t)| dt \\ &= 4\pi. \end{aligned}$$

In particular, *the length of λ is twice what you would expect for the length of $\mathcal{C}$*!

To see what went wrong here, we observe that λ traverses two (anticlockwise) laps around $\mathcal{C}$. As a result, *in the above integral, we counted the contribution of each point of $\mathcal{C}$ twice*, and we hence obtained twice the expected length of $\mathcal{C}$.

Now, since λ traverses $\mathcal{C}$ twice, it fails to be injective. In particular, this shows that injectivity is absolutely necessary in order for Definition 3.68 to be sensible.

Example 3.71. In contrast, consider the parametric curve γ in Example 3.62, with $R = 1$:

$$\gamma : (0, 2\pi) \to \mathcal{C}, \qquad \gamma(t) = (\cos t, \sin t).$$

This γ is now injective; see the right half of Figure 3.21 for an illustration. Also, observe that the image of γ is all of $\mathcal{C}$ *except for the single point* $(1, 0)$, corresponding to $t = 0$ and $t = 2\pi$. Thus, we can apply Definition 3.63 (and recall Example 3.62) to obtain

$$L(\mathcal{C}) = L(\gamma) = 2\pi.$$

Remark 3.72. One expects that *a single point of a curve does not by itself contribute to its total arc length*. (This statement can be rigorously justified through formal integration theory, but we avoid doing this here.) This provides the rationale for the condition in Definition 3.63 that C can differ from the image of γ by a finite number of points.

Example 3.73. Let $\mathcal{H}_*$ be a finite segment of the *helix* from Example 3.18,

$$\mathcal{H}_* = \{(\cos t, \sin t, t) \mid t \in (-2\pi, 2\pi)\}.$$

Let us now integrate over $\mathcal{H}_*$ the function G from Example 3.65:

$$G : \mathbb{R}^3 \to \mathbb{R}, \qquad G(x, y, z) = x^2 + y^2 + z^2.$$

The main point is to note (see Examples 3.18 and 3.65) that the function

$$\mathbf{h}_* : \mathbb{R} \to \mathcal{H}_*, \qquad \mathbf{h}_*(t) = (\cos t, \sin t, t)$$

is an injective parametrisation of $\mathcal{H}_*$ whose image is precisely $\mathcal{H}_*$. Thus, applying Definition 3.68 and then recalling the computations from Example 3.65, we obtain

$$\begin{aligned} \int_{\mathcal{H}_*} G \, ds &= \int_{\mathbf{h}_*} G \, ds \\ &= \sqrt{2}\left[4\pi + \frac{2}{3}(2\pi)^3\right]. \end{aligned}$$

Example 3.74. Let $\mathcal{C}$ be the *unit circle* from Example 3.70, and let

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

To integrate F over $\mathcal{C}$, we recall that γ, from Example 3.71, is an injective parametrisation of $\mathcal{C}$, and that its image is $\mathcal{C} \setminus \{(1, 0)\}$. Thus, by Definitions 3.63 and 3.68,

$$\int_{\mathcal{C}} F \, ds = \int_0^{2\pi} F(\gamma(t))|\gamma'(t)|dt.$$

Now, from the definitions of F and γ, we have

$$|\gamma'(t)| = 1, \qquad F(\gamma(t)) = \sin t.$$

As a result of the above, we conclude that

$$\begin{aligned} \int_{\mathcal{C}} F \, ds &= \int_0^{2\pi} \sin t \, dt \\ &= 0. \end{aligned}$$

3.10. **Integration of Vector Fields.** Next, we introduce a different type of curve integration—*an integral along an oriented curve of a vector field.*

Of course, we should first ask *why we would want to make such a definition.* One reason is that such integrals often come up in physics. An example is the notion of *work* in classical mechanics, which can be viewed as "force applied over a distance".

Suppose there is a crate lying on the ground, and suppose you apply a force to the crate by pushing it in a certain direction, such as in the left drawing of Figure 3.22. This force can be modelled by a vector field $\mathbf{F}$ in $\mathbb{R}^3$; at a point $\mathbf{p} \in \mathbb{R}^3$, the arrow

$$\mathbf{F}(\mathbf{p}) = (F_1(\mathbf{p}), F_2(\mathbf{p}), F_3(\mathbf{p}))_{\mathbf{p}},$$

represents you standing at $\mathbf{p}$ and applying force in the direction $(F_1(\mathbf{p}), F_2(\mathbf{p}), F_3(\mathbf{p}))$.

Suppose that as you apply this force to the crate, it slides along a path given by a curve C in $\mathbb{R}^3$. Since this crate has travelled some distance as you applied force to it, you might claim that you have done work (in the physics sense).

But wait! There is still a question of how productive your force was. If the crate moved in the direction that you applied the force, as is the case in the middle drawing of Figure 3.22, then you can be happy with the fact that your force was effective.

On the other hand, suppose you push the crate rightwards, say, but it is moving downwards in a perpendicular direction, as is depicted in the right drawing in Figure

3.22. Even though you might be working (in a non-physics sense) hard, your force is not so productive; in the physics sense, you have done no work on the box. Moreover, if you push the crate one way, and the crate is instead moving the other way, then the outcome is opposite to your intention, and negative work has been done!

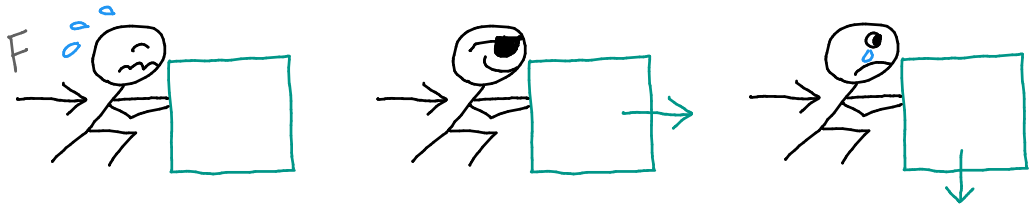


FIGURE 3.22. The first drawing shows a person applying a force to a box. In the second illustration, the box moves in the direction of the force, so positive work is being done by the force. In the third illustration, the box moves in a direction perpendicular to the force, so no work is being done.

Consequently, when measuring work, *the only part of the force* $\mathbf{F}$ *that counts is the component of* $\mathbf{F}$ *that points along the direction of the curve.* If you remember your basic vector geometry, then you will recall that this component can be captured by taking a *dot product* of $\mathbf{F}$ with a unit vector in the direction of the curve.

Putting all these ideas together, we end up at the following definition:

Definition 3.75. Let $C \subseteq \mathbb{R}^n$ be an oriented curve, and let $\mathbf{F}$ be a vector field on a subset of $\mathbb{R}^n$ containing C. Then, we define the curve integral of $\mathbf{F}$ over C by

$$\int_C \mathbf{F} \cdot d\mathbf{s} = \int_C (\mathbf{F} \cdot \mathbf{t}) ds, \tag{3.21}$$

where $\mathbf{t}$ denotes the unit tangents of C in the direction specified by the orientation of C, and where the integral on the right-hand side is defined as in (3.18).

First, note that since $\mathbf{t}$, in Definition 3.75, is the unit tangent in the direction of C, then $\mathbf{F} \cdot \mathbf{t}$ indeed captures the component of $\mathbf{F}$ along C. Also, as $\mathbf{F} \cdot \mathbf{t}$ is scalar-valued, the right-hand integral in (3.21) can indeed be interpreted as in Definition 3.68.

Next, we ask the following: *to what extent are curve integrals of vector fields geometric*, that is, *independent of parametrisation*? One observation, already noted above, is that the right-hand side of (3.21) is a curve integral of a scalar function, which we already know is parametrisation-independent by Theorem 3.67.

However, what slightly complicates this discussion is the unit tangent $\mathbf{t}$. If we reverse the orientation of C, then the corresponding unit tangent $\mathbf{t}$ is replaced by $-\mathbf{t}$.

Since everything else in the right-hand side of (3.21) is geometric and independent of parametrisation, we arrive at the following:

Theorem 3.76. Let C and $\mathbf{F}$ be as in Definition 3.75, and let C_* be the same curve as C, but given the opposite orientation. Then, we have that

$$\int_{C_*} \mathbf{F} \cdot d\mathbf{s} = -\int_C \mathbf{F} \cdot d\mathbf{s}. \tag{3.22}$$

In other words, Theorem 3.76 implies that *curve integrals of vector fields are not quite geometric properties of curves, but they are properties of oriented curves.*

To be more concrete, we return to our informal discussion of work that motivated Definition 3.75. If we apply a force to the crate, then whether the work done is positive or negative depends on whether the crate moves in or opposite to the direction of the force. Thus, the work done is sensitive to the direction that the crate travels.

Moving on to more practical matters, we now discuss how vector integrals can be computed. The next theorem shows that they, like scalar integrals, can be computed via parametrisations, except that we must be careful about orientation:

Theorem 3.77. Let C, $\mathbf{F}$ be as in Definition 3.75, and let $\gamma : (a, b) \to C$ be an injective parametrisation of C whose image differs from C by only a finite number of points.

- If γ generates the orientation of C, then

$$\int_C \mathbf{F} \cdot d\mathbf{s} = +\int_a^b \left[\mathbf{F}(\gamma(t)) \cdot \gamma'(t)_{\gamma(t)}\right] dt. \tag{3.23}$$

- If γ generates the orientation opposite to that of C, then

$$\int_C \mathbf{F} \cdot d\mathbf{s} = -\int_a^b \left[\mathbf{F}(\gamma(t)) \cdot \gamma'(t)_{\gamma(t)}\right] dt. \tag{3.24}$$

Proof. First, if γ generates the orientation of C, then Theorem 3.41 and Definition 3.52 imply that the unit tangent $\mathbf{t}$ chosen by the orientation at the point $\gamma(t)$ is given by $+|\gamma'(t)|^{-1} \cdot \gamma'(t)_{\gamma(t)}$. Thus, using Definitions 3.63, 3.68, and 3.75, we have

$$\begin{aligned}\int_C \mathbf{F} \cdot d\mathbf{s} &= \int_a^b \left\{\mathbf{F}(\gamma(t)) \cdot \left[+\frac{\gamma'(t)}{|\gamma'(t)|}\right]_{\gamma(t)}\right\} |\gamma'(t)|\, dt \\ &= \int_a^b \left[\mathbf{F}(\gamma(t)) \cdot \gamma'(t)_{\gamma(t)}\right] dt.\end{aligned}$$

The remaining case (3.24), in which γ generates the opposite orientation to C, is analogous, except that $\mathbf{t}$ now corresponds to $-|\gamma'(t)|^{-1} \cdot \gamma'(t)_{\gamma(t)}$. □

Remark 3.78. Note the base point $\gamma(t)$ of the tangent vectors $\mathbf{F}(\gamma(t))$ and $\gamma'(t)_{\gamma(t)}$ plays no role in the computation of the right-hand sides of (3.23) and (3.24). Although we ignore this $\gamma(t)$ in practice, we will still write them down here as a matter of principle, as reminders that the objects of interest here are arrows based at points along γ.

Example 3.79. Let $\mathcal{C}_*$ denote the *unit circle* about the origin,

$$\mathcal{C}_* = \{(x, y) \in \mathbb{R}^2 \mid x^2 + y^2 = 1\},$$

with the *anticlockwise orientation*. Also, let $\mathbf{F}$ be the vector field on $\mathbb{R}^2$ given by

$$\mathbf{F}(x, y) = (-y, x)_{(x,y)}.$$

See the left graphic in Figure 3.23 for an illustration of this setting. There, $\mathcal{C}_*$ is plotted in red, while some values of $\mathbf{F}$ along $\mathcal{C}_*$ are drawn as blue arrows.

Let us now compute the curve integral of $\mathbf{F}$ over $\mathcal{C}_*$ using Theorem 3.77. The first step is to find an appropriate parametrisation of $\mathcal{C}_*$; recall, from Example 3.74, that

$$\gamma : (0, 2\pi) \to \mathcal{C}_*, \qquad \gamma(t) = (\cos t, \sin t)$$

is an injective parametrisation of $\mathcal{C}_*$ whose image is all of $\mathcal{C}_*$ except for a single point. In Example 3.57, we noted that γ generates the same anticlockwise orientation.

As a result, by (3.23), we have that

$$\int_{\mathcal{C}_*} \mathbf{F} \cdot d\mathbf{s} = + \int_0^{2\pi} \left[\mathbf{F}(\gamma(t)) \cdot \gamma'(t)_{\gamma(t)}\right] dt.$$

The integrand can now be computed directly. First, note that

$$\gamma'(t)_{\gamma(t)} = (-\sin t, \cos t)_{(\cos t, \sin t)}, \qquad \mathbf{F}(\gamma(t)) = (-\sin t, \cos t)_{(\cos t, \sin t)},$$

for any $t \in (0, 2\pi)$. Taking a dot product of the above then yields

$$\mathbf{F}(\gamma(t)) \cdot \gamma'(t)_{\gamma(t)} = 1, \qquad t \in (0, 2\pi).$$

Putting all of the above together, we conclude that

$$\begin{aligned} \int_{\mathcal{C}_*} \mathbf{F} \cdot d\mathbf{s} &= \int_0^{2\pi} 1 \, dt \\ &= 2\pi. \end{aligned}$$

Finally, observe that $\mathbf{F}$ points along the same direction as $\mathcal{C}_*$. Thus, the integral of $\mathbf{F}$ over $\mathcal{C}_*$ should be positive, and the above calculation confirms this expectation.

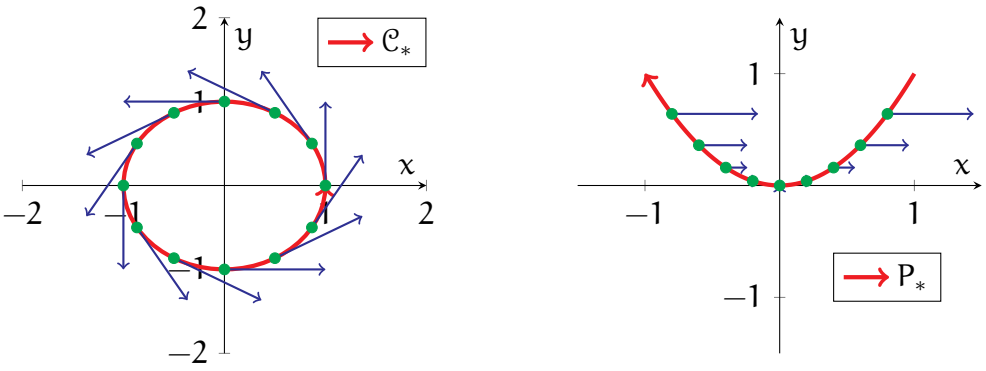


FIGURE 3.23. The left plot shows the setting of Example 3.79, with the oriented circle $\mathcal{C}_*$ in red and values of $\mathbf{F}$ in blue. The right plot is the setting of Example 3.80, with the parabola P_* in red and values of $\mathbf{H}$ in blue.

Example 3.80. Let P_* denote the *parabolic segment*

$$P_* = \{(t, t^2) \mid -1 < t < 1\},$$

oriented in the direction of decreasing x*-value.* (Note P_* is a curve by Theorem 3.28.)

Let us integrate over P_* the vector field $\mathbf{H}$ on $\mathbb{R}^2$ defined as

$$\mathbf{H}(x, y) = (y, 0)_{(x,y)}.$$

First, observe that P_* can be injectively parametrised in its entirety by

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

The usual computations then yield, for each $t \in (-1, 1)$,

$$\begin{aligned} \mathbf{F}(\lambda(t)) \cdot \lambda'(t)_{\lambda(t)} &= (t^2, 0) \cdot (1, 2t) \\ &= t^2. \end{aligned}$$

However, note that λ *generates the orientation in the direction of increasing* x*-value*, which is *opposite to our given orientation for* P_*. Thus, we apply (3.24) and obtain

$$\begin{aligned} \int_{P_*} \mathbf{F} \cdot d\mathbf{s} &= -\int_{-1}^{1} t^2 \, dt \\ &= -\frac{2}{3}. \end{aligned}$$

See the right graphic in Figure 3.23 for an illustration of P_* (drawn in red), as well as some values of $\mathbf{H}$ along P_* (drawn as blue arrows).

Remark 3.81. Lastly, assume the setting of Definition 3.75, and expand $\mathbf{F}$ as

$$\mathbf{F}(\mathbf{p}) = (F_1(\mathbf{p}), \ldots, F_n(\mathbf{p}))_{\mathbf{p}}, \qquad \mathbf{p} \in C.$$

In many calculus texts, the curve integral of $\mathbf{F}$ over C is often written as

$$\int_C \mathbf{F} \cdot d\mathbf{s} = \int_C (F_1\, dx_1 + F_2\, dx_2 + \cdots + F_n\, dx_n).$$

(When $n \leq 3$, one often writes dx, dy, dz in the place of dx_1, dx_2, dx_3.) Formally speaking, the integrand on the right-hand side is known as a *differential form*. Unfortunately, the theory behind these objects lies beyond the scope of this module.

4. The Geometry of Surfaces

In this portion of the module, we turn our attention toward the *differential geometry of surfaces*. You have probably heard the word "surface" used in many real-world contexts, such as "surface of a table" or "surface of the earth". More abstractly, one can view surfaces as "2-dimensional geometric objects".

However, for a systematic study, one need a more precise description of surfaces. Thus, like for curves, we spend the first part of the chapter addressing the following:

Question 4.1. When we say "surface", what exactly do we mathematically mean?

As we shall see, there will be some novel challenges in describing surfaces that we did not previously encounter when studying curves.

4.1. Parametric Surfaces.

Recall our discussions of curves began with the notion of *parametric curves*, which were used to map out the points of the underlying curve. Moreover, the expectation that parametric curves are "1-dimensional" is captured by their domains, which were open intervals in the (1-dimensional) real line.

We now employ a similar beginning to surface theory. The idea is once again to parametrise the points of a surface using vector-valued functions. Since we wish for surfaces to be "2-dimensional" in nature, then the above suggests that *we should now consider vector-valued functions whose domains are subsets of* $\mathbb{R}^2$.

This leads us to our 2-dimensional analogue of parametric curves:

Definition 4.2. A parametric surface is a smooth vector-valued function $\sigma : U \to \mathbb{R}^n$, where n is a positive integer, and where U is an open and connected subset of $\mathbb{R}^2$.

In Definition 4.2, the condition that σ is *smooth* means that *we can take as many partial derivatives of* σ *as we wish*, as all such derivatives are assumed to exist.

Remark 4.3. Although Definition 4.2 allows for parametric surfaces lying in any $\mathbb{R}^n$, almost all of our examples will be in the special case $n = 3$.

Example 4.4. First, recall the parametric *cylinder* from Examples 2.23 and 2.55:

$$\sigma : \mathbb{R}^2 \to \mathbb{R}^3, \qquad \sigma(u, v) = (\cos u, \sin u, v).$$

In particular, observe that the domain $\mathbb{R}^2$ of σ is both open and connected, thus σ is indeed a parametric surface. We also recall from the aforementioned examples that the image of σ is a cylinder of unit radius about the z-axis; see the left plot of Figure 4.1.

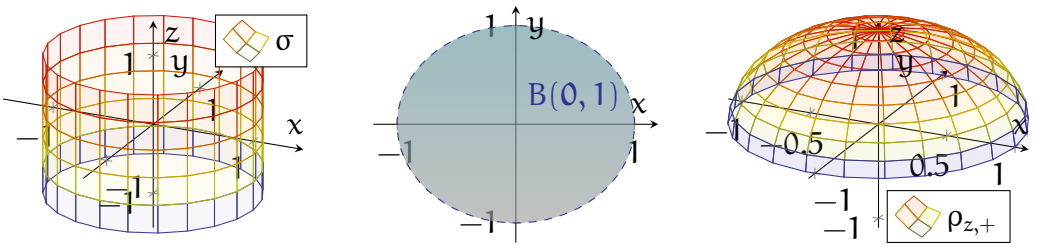


FIGURE 4.1. The left plot shows the image of the parametric surface σ from Example 4.4. The middle plot is the unit disk $B(0,1)$ about the origin, as described in Example 4.5. Finally, the right graphic shows the image of the parametric surface $\rho_{z,+}$ from Example 4.5.

Example 4.5. Let $B(0,1)$ denote the unit disk about the origin in $\mathbb{R}^2$,

$$B(0,1) = \{(x,y) \in \mathbb{R}^2 \mid x^2 + y^2 < 1\}$$

(see the middle part of Figure 4.1), and consider the vector-valued function

$$\rho_{z,+} : B(0,1) \to \mathbb{R}^3, \qquad \rho_{z,+}(u,v) = \left(u, v, \sqrt{1-u^2-v^2}\right).$$

Recall that $B(0,1)$ is both open (see Example 2.44) and connected (see Example 2.48). Also, $\rho_{z,+}$ is smooth, since the quantity $1-u^2-v^2$ under the square root is positive for all $(u,v) \in B(0,1)$. Thus, $\rho_{z,+}$ is indeed a parametric surface.

By plotting its image, we see that $\rho_{z,+}$ maps out the *upper unit hemisphere*, that is, the points of the unit sphere $x^2+y^2+z^2 = 1$ such that $z > 0$; see the right plot of Figure 4.1. We can also come to the same conclusion analytically:

- Any point $(x,y,z) = \rho_{z,+}(u,v)$ in the image of $\rho_{z,+}$ lies on the unit sphere, since

$$\begin{aligned} x^2 + y^2 + z^2 &= u^2 + v^2 + (1 - u^2 - v^2) \\ &= 1, \end{aligned}$$

  Furthermore, (x,y,z) lies on the upper half of the sphere, since

$$z = \sqrt{1-u^2-v^2} > 0.$$

- Conversely, if $x^2+y^2+z^2 = 1$ and $z > 0$, then it follows that

$$\begin{aligned} (x,y,z) &= \left(x, y, \sqrt{1-x^2-y^2}\right) \\ &= \rho_{z,+}(x,y). \end{aligned}$$

  In other words, (x,y,z) lies in the image of $\rho_{z,+}$.

Consider now a general parametric surface $\sigma : U \to \mathbb{R}^n$, and suppose you are standing at a point $\mathbf{p}_0 = \sigma(u_0, v_0)$. We now ask: *if you start walking from* $\mathbf{p}_0$ *along the image of* σ*, then what velocities could you initially have*?

Figure 4.2 provides a graphical depiction of our situation. Our objective, in terms of these plots, is then to "describe all the possible blue arrows".

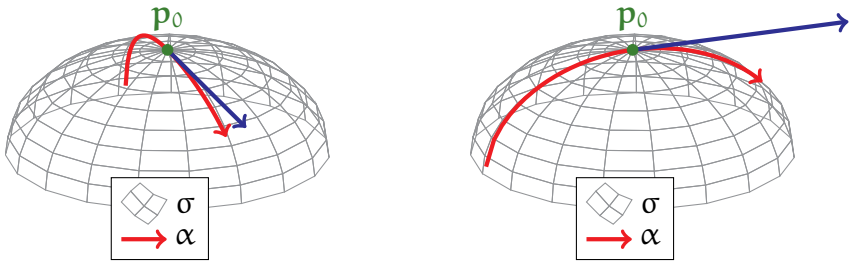


FIGURE 4.2. In the above, a person starts at a point $\mathbf{p}_0$ (in green) and can move along the parametric surface σ in various ways; a couple possibilities are drawn in red. The blue arrows then represent the corresponding velocity in which the person is initially moving while at $\mathbf{p}_0$.

To explore this, assume your position on σ is given by a parametric curve

$$\alpha(t) = \sigma(u(t), v(t)), \qquad (u(t), v(t)) \in U.$$

Furthermore, suppose that α passes through $\mathbf{p}_0$ at $t = 0$:

$$(u(0), v(0)) = (u_0, v_0), \qquad \alpha(0) = \sigma(u_0, v_0) = \mathbf{p}_0.$$

In terms of Figure 4.2, the image of α corresponds to the two red paths. The blue arrows then correspond to the tangent vectors $\alpha'(0)_{\mathbf{p}_0}$ at $\mathbf{p}_0$.

To find $\alpha'(0)$, we apply the (multivariable) chain rule:

$$\begin{aligned}\alpha'(0) &= \partial_1\sigma(u(t), v(t)) \cdot u'(t)|_{t=0} + \partial_2\sigma(u(t), v(t)) \cdot v'(t)|_{t=0} \\ &= u'(0) \cdot \partial_1\sigma(u_0, v_0) + v'(0) \cdot \partial_2\sigma(u_0, v_0).\end{aligned}$$

As a result, we conclude that

$$\alpha'(0)_{\mathbf{p}_0} = u'(0) \cdot \partial_1\sigma(u_0, v_0)_{\mathbf{p}_0} + v'(0) \cdot \partial_2\sigma(u_0, v_0)_{\mathbf{p}_0}. \tag{4.1}$$

Furthermore, by altering our trajectory α, we can arrange so that $(u'(0), v'(0))$ takes any value we want. Thus, *the possible velocities that you can have while at* $\mathbf{p}_0$ *is precisely given by all the linear combinations of* $\partial_1\sigma$ *and* $\partial_2\sigma$ *at* (u_0, v_0).

With all of this in mind, we now define the following:

Definition 4.6. Let $\sigma : U \to \mathbb{R}^n$ be a parametric surface, and let $(u_0, v_0) \in U$. Then, the tangent plane to σ at (u_0, v_0) is defined to be the following set:

$$T_\sigma(u_0, v_0) = \left\{ a \cdot \partial_1\sigma(u_0, v_0)_{\sigma(u_0,v_0)} + b \cdot \partial_2\sigma(u_0, v_0)_{\sigma(u_0,v_0)} \middle| a, b \in \mathbb{R} \right\}. \tag{4.2}$$

Remark 4.7. Observe that $T_\sigma(u_0, v_0)$, in Definition 4.6, is a vector space, since it is the linear span of the tangent vectors $\partial_1\sigma(u_0, v_0)_{\sigma(u_0,v_0)}$ and $\partial_2\sigma(u_0, v_0)_{\sigma(u_0,v_0)}$.

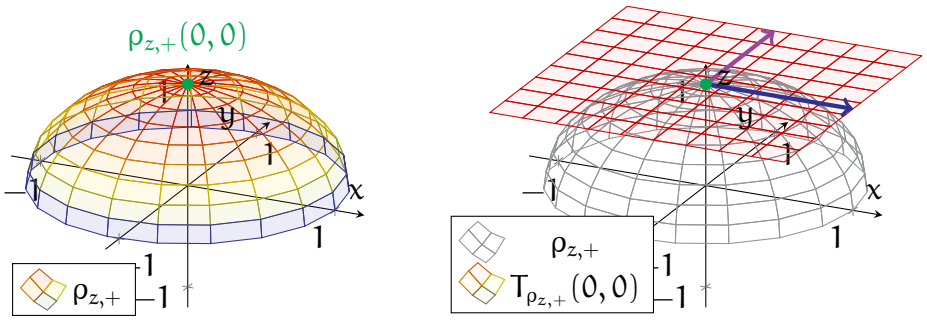


FIGURE 4.3. The left plot shows the image of $\rho_{z,+}$ from Example 4.8. In the right graphic, the arrows $\partial_1\rho_{z,+}(0,0)_{\rho_{z,+}(0,0)}$ (in blue) and $\partial_2\rho_{z,+}(0,0)_{\rho_{z,+}(0,0)}$ (in purple) are drawn on the image of $\rho_{z,+}$. Moreover, the tangent plane $T_{\rho_{z,+}}(0,0)$ is represented as a plane lying on top of $\rho_{z,+}$.

Example 4.8. Consider first the parametric *upper hemisphere* from Example 4.5:

$$\rho_{z,+} : B(0,1) \to \mathbb{R}^3, \qquad \rho_{z,+}(u,v) = \left(u, v, \sqrt{1-u^2-v^2}\right).$$

See the left half of Figure 4.3 for a plot of $\rho_{z,+}$. Let us find the tangent plane $T_{\rho_{z,+}}(0,0)$, i.e. the velocities along $\rho_{z,+}$ with which one can go from $\rho_{z,+}(0,0) = (0,0,1)$.

By Definition 4.6, the first step is to compute the partial derivatives of $\rho_{z,+}$. To find $\partial_1\rho_{z,+}$, we differentiate each component of $\rho_{z,+}$ with respect to the first variable u:

$$\begin{aligned}\partial_1\rho_{z,+}(u,v) &= \left(1, 0, \frac{1}{2} \cdot \frac{\partial_u(1-u^2-v^2)}{\sqrt{1-u^2-v^2}}\right) \\ &= \left(1, 0, -\frac{u}{\sqrt{1-u^2-v^2}}\right).\end{aligned}$$

A similar differentiation with respect to v yields that

$$\partial_2\rho_{z,+}(u,v) = \left(0, 1, -\frac{v}{\sqrt{1-u^2-v^2}}\right).$$

In particular, at $(u,v) = (0,0)$, we have

$$\partial_1\rho_{z,+}(0,0) = (1,0,0), \qquad \partial_2\rho_{z,+}(0,0) = (0,1,0).$$

Therefore, by (4.2), our desired tangent plane of $\rho_{z,+}$ is

$$\begin{aligned} T_{\rho_{z,+}}(0,0) &= \left\{ a \cdot \partial_1 \rho_{z,+}(0,0)_{\rho_{z,+}(0,0)} + b \cdot \partial_2 \rho_{z,+}(0,0)_{\rho_{z,+}(0,0)} \middle| a, b \in \mathbb{R} \right\} \\ &= \left\{ a \cdot (1,0,0)_{(0,0,1)} + b \cdot (0,1,0)_{(0,0,1)} \middle| a, b \in \mathbb{R} \right\}. \end{aligned}$$

The graphic on the right-hand side of Figure 4.3 demonstrates the above setting. The arrows $\partial_1 \rho_{z,+}(0,0)_{\rho_{z,+}(0,0)}$ and $\partial_2 \rho_{z,+}(0,0)_{\rho_{z,+}(0,0)}$ are drawn in blue and purple, respectively. All the tangent vectors within $T_{\rho_{z,+}}(0,0)$ (i.e. the linear combinations of the blue and purple arrows) then generate a 2-dimensional plane, which is drawn on top of $\rho_{z,+}$.

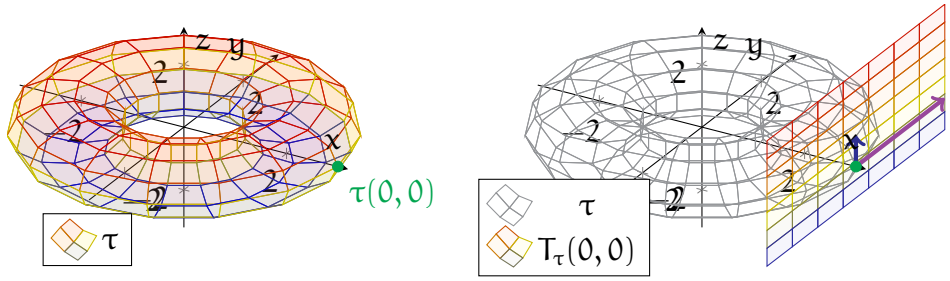


FIGURE 4.4. The left graphic shows the torus τ from Example 4.9. The right drawing also contains both $\partial_1 \tau(0,0)_{\tau(0,0)}$ (in blue) and $\partial_2 \tau(0,0)_{\tau(0,0)}$ (in purple), as well as the tangent plane $T_\tau(0,0)$.

Example 4.9. Next, consider the vector-valued function

$$\tau : \mathbb{R}^2 \to \mathbb{R}^3, \qquad \tau(u,v) = ((2+\cos u)\cos v, (2+\cos u)\sin v, \sin u),$$

which satisfies all the conditions of Definition 4.2 and hence is a parametric surface. The "doughnut-shaped" image of τ, called a *torus*, is drawn in the left part of Figure 4.4.

Let us now compute $T_\tau(0,0)$. First, some direct computations yield that

$$\begin{aligned} \partial_1 \tau(u,v) &= (-\sin u \cos v, -\sin u \sin v, \cos u), \\ \partial_2 \tau(u,v) &= (-(2+\cos u)\sin v, (2+\cos u)\cos v, 0). \end{aligned}$$

In particular, at $(u,v) = (0,0)$,

$$\partial_1 \tau(0,0)_{\tau(0,0)} = (0,0,1)_{(3,0,0)}, \qquad \partial_2 \tau(0,0)_{\tau(0,0)} = (0,3,0)_{(3,0,0)}.$$

Thus, using Definition 4.2, we conclude that

$$T_\tau(0,0) = \{ a \cdot (0,0,1)_{(3,0,0)} + b \cdot (0,3,0)_{(3,0,0)} \mid a, b \in \mathbb{R} \}.$$

See the right plot of Figure 4.4 for an illustration of $T_\tau(0,0)$.

Remark 4.10. Similar to Remark 3.37, we could also have defined tangent planes as sets of points. More specifically, in the setting of Definition 4.6, we could have set

$$\mathcal{T}_\sigma(u_0, v_0) = \{\sigma(u_0, v_0) + a \cdot \partial_1\sigma(u_0, v_0) + b \cdot \partial_2\sigma(u_0, v_0) \mid a, b \in \mathbb{R}\}. \tag{4.3}$$

Although (4.3) is more straightforward, it fails to capture the linear structure inherent in Definition 4.6—namely, that $T_\sigma(u_0, v_0)$ is a vector space.

4.2. **Regular Parametrisations.** Similar to parametric curves, there are ways that parametric surfaces can fall short of our goal of describing geometric surfaces. We explore some simple instances of this in the following examples:

Example 4.11. Consider the parametric surface

$$\zeta : \mathbb{R}^2 \to \mathbb{R}^3, \qquad \zeta(u, v) = (\cos u, \sin u, 1).$$

Here, the image of ζ is merely a (1-dimensional) circle; see Figure 4.5.

We can capture this shortcoming through the tangent planes of ζ. Since

$$\partial_1\zeta(u, v) = (-\sin u, \cos u, 0), \qquad \partial_2\zeta(u, v) = (0, 0, 0), \qquad (u, v) \in \mathbb{R},$$

then $T_\zeta(u, v)$ is only a 1-dimensional vector space (again, see Figure 4.5):

$$T_\zeta(u, v) = \{a \cdot (-\sin u, \cos u, 0)|_{(\cos u, \sin u, 1)} \mid a \in \mathbb{R}\}, \qquad (u, v) \in \mathbb{R}^2.$$

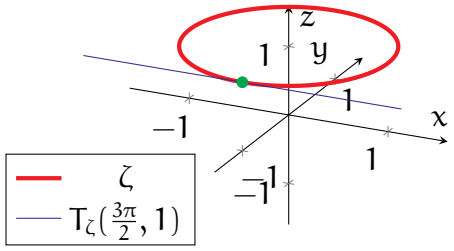


FIGURE 4.5. This plot contains the parametric surface ζ from Example 4.11, which fails miserably at being 2-dimensional. The point $\zeta(\frac{3\pi}{2}, 1)$ is indicated in green, while the tangent "plane" $T_\zeta(\frac{3\pi}{2}, 1)$ is drawn in blue.

The parametric surface in Example 4.11 was obviously deficient, as it depended on only one variable. However, excluding these worst-case scenarios, we can still find more benign situations where a parametric surface is undesirable:

Example 4.12. Consider the parametric surface

$$\rho_* : \mathbb{R} \times \left(-\frac{\pi}{2}, \frac{\pi}{2}\right) \to \mathbb{R}^3, \qquad \rho_*(u, v) = (\cos u \sin v, \sin u \sin v, \cos v).$$

By computing its values, one can see that ρ_* maps out the same *upper hemisphere* as $\rho_{z,+}$ in Examples 4.5 and 4.8. Moreover, note that the formula for ρ_* represents standard *spherical coordinates* (which you have probably seen in calculus). More specifically:

- u represents the *polar angle* in the xy-plane.
- v represents the angle from the positive z-axis (i.e. the *azimuthal angle*).

Unlike Example 4.11, this ρ_* represents an honest 2-dimensional object. See the left part of Figure 4.6 for a plot of ρ_*, along with some curves of constant u and v.

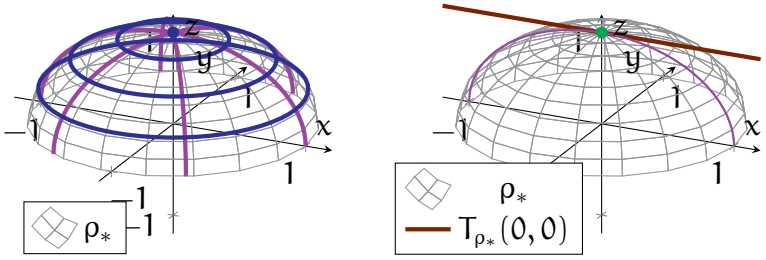


FIGURE 4.6. Both plots contain the upper hemisphere ρ_* from Example 4.12. In the left plot, the blue curves are generated by holding v constant and varying u, while the purple curves are generated by holding u constant and varying v. In the right plot, the point $\rho_*(0,0) = (0,0,1)$ is indicated in green, while $T_{\rho_*}(0,0)$ is drawn as a brown line.

Example 4.13. Let ρ_* be as in Example 4.12. Differentiating ρ_*, we see that

$$\partial_1\rho_*(u,v) = (-\sin u \sin v, \cos u \sin v, 0),$$
$$\partial_2\rho_*(u,v) = (\cos u \cos v, \sin u \cos v, -\sin v).$$

Evaluating at $(u,v) = (0,0)$, we then obtain

$$\rho_*(0,0) = (0,0,1), \qquad \partial_1\rho_*(0,0) = (0,0,0), \qquad \partial_2\rho_*(0,0) = (1,0,0).$$

Note that the parameter $(u,v) = (0,0)$ corresponds to the north pole.

In particular, the corresponding tangent plane of ρ_*, at $(0,0)$, is

$$T_{\rho_*}(0,0) = \{b \cdot (1,0,0)_{(0,0,1)} \mid b \in \mathbb{R}\},$$

which again gives a 1-dimensional line, not a 2-dimensional plane. See the right side of Figure 4.6, which shows $T_{\rho_*}(0,0)$ drawn on top of the image of ρ_*.

Although the image of ρ_* is 2-dimensional, it still fails to generate a reasonable tangent plane at the north pole. To see what went wrong, let us fix $v = 0$ and vary u:

$$\rho_*(u,0) = (0,0,1).$$

In particular, when $v = 0$, varying u does not change the value of ρ_*. This results in the loss of the "u-dimension" from the tangent plane $T_{\rho_*}(0,0)$.

One consequence of this is that ρ_* is an undesirable way to parametrise the upper hemisphere. Even though the upper hemisphere itself is genuinely 2-dimensional, the parametric surface ρ_* still fails to be 2-dimensional at the north pole.

What is needed, then, for a parametric surface to be "2-dimensional"? Recall that the tangent planes of a parametric surface σ capture all the directions that one can move along σ. Thus, *for σ to be truly 2-dimensional, there should be two dimensions of directions that one can move along* σ, i.e. *the $T_\sigma(u,v)$'s should be 2-dimensional.*

Definition 4.14. A parametric surface $\sigma : U \to \mathbb{R}^n$ is regular iff for any $(u,v) \in U$, the vectors $\partial_1\sigma(u,v)$ and $\partial_2\sigma(u,v)$ are linearly independent.

To better understand Definition 4.14, we recall a bit of linear algebra. Notice that Definition 4.14 implies σ is regular if and only if $\partial_1\sigma(u,v)$ and $\partial_2\sigma(u,v)$ always span a 2-dimensional vector space. By (4.2), this is true if and only if $T_\sigma(u,v)$ is also a 2-dimensional vector space. Thus, we can now conclude the following:

Theorem 4.15. A parametric surface $\sigma : U \to \mathbb{R}^n$ is regular if and only if $T_\sigma(u,v)$ is a 2-dimensional vector space for any $(u,v) \in U$. Moreover, when σ is regular, then $\partial_1\sigma(u,v)_{\sigma(u,v)}$ and $\partial_2\sigma(u,v)_{\sigma(u,v)}$ form a basis for $T_\sigma(u,v)$, for any $(u,v) \in U$.

Example 4.16. Recall our previous description of the upper hemisphere,

$$\rho_{z,+} : B(0,1) \to \mathbb{R}^3, \qquad \rho_{z,+}(u,v) = \left(u, v, \sqrt{1-u^2-v^2}\right).$$

from Example 4.8. There, we have already computed the partial derivatives of $\rho_{z,+}$:

$$\partial_1\rho_{z,+}(u,v) = \left(1, 0, -\frac{u}{\sqrt{1-u^2-v^2}}\right),$$
$$\partial_2\rho_{z,+}(u,v) = \left(0, 1, -\frac{v}{\sqrt{1-u^2-v^2}}\right).$$

For any $(u,v) \in B(0,1)$, the vector $\partial_1\rho_{z,+}(u,v)$ always has a nonzero x-component and a zero y-component, while $\partial_2\rho_{z,+}(u,v)$ always has a nonzero y-component and a zero x-component. Therefore, $\partial_1\rho_{z,+}(u,v)$ and $\partial_2\rho_{z,+}(u,v)$ always point in different directions and hence are linearly independent. By Definition 4.14, we conclude $\rho_{z,+}$ is regular.

In Example 4.16, it was clear by inspection that $\partial_1\rho_{z,+}$ and $\partial_2\rho_{z,+}$ are everywhere linearly independent. In general, however, this may not be so simple. Thus, we ask

whether there is a computational method to check whether a parametric surface is regular. A convenient method does exist for surfaces lying in 3-dimensional space:

Theorem 4.17. A parametric surface $\sigma : U \to \mathbb{R}^3$ is regular if and only if

$$|\partial_1\sigma(u,v) \times \partial_2\sigma(u,v)| \neq 0 \tag{4.4}$$

for every parameter $(u,v) \in U$.

Proof. Recall from Theorem 2.18 that

$$|\partial_1\sigma(u,v) \times \partial_2\sigma(u,v)| = |\partial_1\sigma(u,v)||\partial_2\sigma(u,v)| \cdot \sin\theta(u,v),$$

where $\theta(u,v)$ is the angle between $\partial_1\sigma(u,v)_{\sigma(u,v)}$ and $\partial_2\sigma(u,v)_{\sigma(u,v)}$. Thus, (4.4) holds if and only if both $\partial_u\sigma(u,v)$ and $\partial_v\sigma(u,v)$ are nonzero and $\sin\theta(u,v) \neq 0$. Furthermore, $\sin\theta(u,v)$ is nonzero if and only if $\partial_1\sigma(u,v)$ and $\partial_2\sigma(u,v)$ point in different directions (and hence are linearly independent). Combining all the above, we conclude that (4.4) is indeed equivalent to σ being regular. □

Example 4.18. Consider the *torus* from Example 4.9 (see also Figure 4.4),

$$\tau : \mathbb{R}^2 \to \mathbb{R}^3, \qquad \tau(u,v) = ((2+\cos u)\cos v, (2+\cos u)\sin v, \sin u).$$

Recall also from Example 4.9 that

$$\begin{aligned}\partial_1\tau(u,v) &= (-\sin u\cos v, -\sin u\sin v, \cos u),\\ \partial_2\tau(u,v) &= (-(2+\cos u)\sin v, (2+\cos u)\cos v, 0).\end{aligned}$$

Taking a cross product of the above yields

$$\partial_1\tau(u,v) \times \partial_2\tau(u,v) = -(2+\cos u)(\cos u\cos v,\ \cos u\sin v,\ \sin u),$$

for which the norm is

$$\begin{aligned}|\partial_1\tau(u,v) \times \partial_2\tau(u,v)| &= (2+\cos u)(\cos^2 u\cos^2 v + \cos^2 u\sin^2 v + \sin^2 u)\\ &= (2+\cos u)(\cos^2 u + \sin^2 u)\\ &= 2+\cos u.\end{aligned}$$

Since the above is everywhere nonzero, Theorem 4.17 implies that τ is regular.

Remark 4.19. Similarly, note that a parametric curve γ is regular if and only if all of its tangent lines $T_\gamma(t)$ are 1-dimensional vector spaces, i.e. they are actual lines.

4.3. **Geometric Surfaces.** We are now prepared to precisely define surfaces. The good news is that this will mostly mirror our previous discussions for curves.

There are three guiding principles that motivate how we wish to define surfaces:

(1) Surfaces are described using regular parametric surfaces.
(2) Surfaces, as well as their properties, should be independent of parametrisation.
(3) We do not allow surfaces to "self-intersect", i.e. to pass through themselves.

These match the principles for curves given prior to Definition 3.17. Also, we previously justified the first principle—we use parametric surfaces to map out points of a surface, and we need regularity to ensure these points form 2-dimensional sets.

These three principles are formally captured through the following definition:

Definition 4.20. A subset $S \subseteq \mathbb{R}^n$ is called a surface iff for any $\mathbf{p} \in S$, there exist

- An open subset $V \subseteq \mathbb{R}^n$ such that $\mathbf{p} \in V$, and
- A regular and injective parametric surface $\sigma : U \to S$,

such that the following conditions hold:

- σ is a bijection between U and $S \cap V$.
- The inverse $\sigma^{-1} : S \cap V \to U$ of σ is also continuous.

Observe that Definition 4.20 is a direct 2-dimensional analogue of Definition 3.17 for curves. Like for curves, here we will refrain from technical discussions of Definition 4.20, as it relies on background from topology that lies beyond this module. However, let us discuss informally how Definition 4.20 relates to the above three principles.

FIGURE 4.7. A surface S is locally mapped by a regular, injective parametric surface σ (in red and grey), as described in Definition 4.20.

Let $S \subseteq \mathbb{R}^n$ be a surface, and fix any point $\mathbf{p} \in S$. Then, the regular parametric surface σ described in Definition 4.20 yields a smooth one-to-one correspondence

between the region $U \subseteq \mathbb{R}^2$ and points of S near $\mathbf{p}$ (more specifically, the set $S \cap V$). See Figure 4.7 for an illustration; there, V is drawn in purple, while U and $S \cap V$ are drawn in red. In other words, *this part* $S \cap V$ *of* S *near* $\mathbf{p}$ *looks like a "deformation" of* U, with σ *describing how* U *is deformed into this part of* S.

One way to view this more concretely is as follows: suppose S represents the surface of the earth, and suppose you draw a map of a part of S (e.g. Europe) on a flat piece of paper. In this analogy, U would represent the part of the paper on which you draw the map, while u and v represent the horizontal and vertical coordinates of the paper. The parametric surface σ, which map points (u, v) of the paper to points $\sigma(u, v) \in S$, then describes which parts of the paper correspond to which parts of Europe.

Returning to the abstract setting, our surface S can then be viewed as being constructed by "gluing together" a bunch of "deformed 2-dimensional regions", each of which is represented by a regular parametric surface such as the above σ. In terms of our concrete example, we can map out the entire surface S of the earth by combining together several maps, each of which is represented by a parametric surface σ. This demonstrates how Definition 4.20 fulfills the first principle mentioned above.

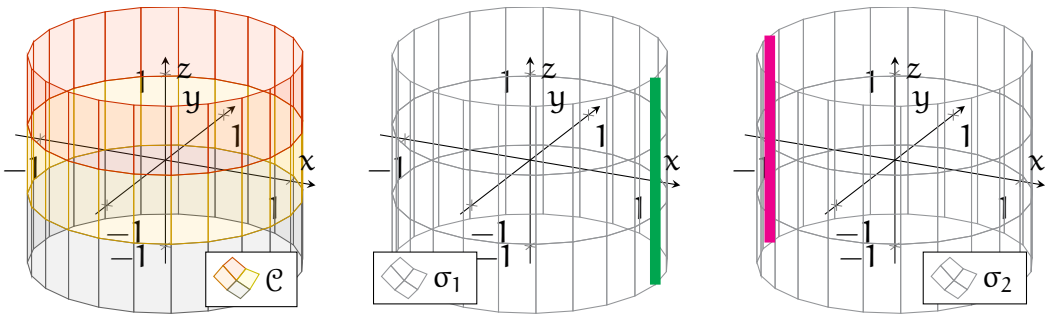


FIGURE 4.8. The left drawing shows the cylinder C from Example 4.21. The middle drawing shows the image of σ_1, while the right drawing shows the image of σ_2. (The green and pink lines are excluded from these images.)

Example 4.21. Consider the *cylinder*, defined as the following subset of $\mathbb{R}^3$:

$$C = \{(x, y, z) \in \mathbb{R}^3 \mid x^2 + y^2 = 1\}.$$

See the left part of Figure 4.8 for a drawing of C.

To argue that C is a surface, we consider the following parametric surfaces:

$$\sigma_1 : (0, 2\pi) \times \mathbb{R} \to \mathbb{R}^2, \qquad \sigma_1(u, v) = (\cos u, \sin u, v),$$
$$\sigma_2 : (-\pi, \pi) \times \mathbb{R} \to \mathbb{R}^2, \qquad \sigma_2(u, v) = (\cos u, \sin u, v).$$

We claim that both σ_1 and σ_2 are regular. For σ_1, we compute

$$\begin{aligned}|\partial_1\sigma_1(u,v)\times\partial_2\sigma_1(u,v)| &= |(-\sin u,\cos u,0)\times(0,0,1)|\\ &= 1,\end{aligned}$$

hence it follows from Theorem 4.17 that σ_1 is indeed regular. Similar computations, along with another application of Theorem 4.17, yield that σ_2 is also regular.

In addition, it is not too difficult to see that both σ_1 and σ_2 are injective, and:

- The image of σ_1 is all of $\mathcal{C}$ except for the vertical line $\mathcal{L}_1 = \{(1,0,v) \mid v\in\mathbb{R}\}$.
- The image of σ_2 is all of $\mathcal{C}$ except for the vertical line $\mathcal{L}_2 = \{(-1,0,v) \mid v\in\mathbb{R}\}$.

See the middle and right parts of Figure 4.8; the lines $\mathcal{L}_1$ and $\mathcal{L}_2$ are drawn in green and pink, respectively. In particular, both σ_1 and σ_2 map out only parts of $\mathcal{C}$. *To construct $\mathcal{C}$ in its entirety, we must then "glue together" the images of σ_1 and σ_2.*

For completeness, let us also connect σ_1 and σ_2 to Definition 4.20:

- If $\mathbf{p}\in\mathcal{C}$ and $\mathbf{p}\notin\mathcal{L}_1$, then the hypotheses of Definition 4.20 are satisfied, with
$$\sigma=\sigma_1,\qquad V=\{(x,y,z)\in\mathbb{R}^3 \mid x<1\}.$$
- If $\mathbf{p}\in\mathcal{C}$ and $\mathbf{p}\notin\mathcal{L}_2$, then the hypotheses of Definition 4.20 are satisfied, with
$$\sigma=\sigma_2,\qquad V=\{(x,y,z)\in\mathbb{R}^3 \mid x>-1\}.$$

(We omit detailed proofs here.) Therefore, *the cylinder $\mathcal{C}$ is indeed a surface.*

The notion of independence of parametrisation—principle (2) in the above—is manifested in the same way as for curves. There are many different parametric surfaces σ that one can use to describe S, and Definition 4.20 does not prefer any one over another. For instance, going back to our analogy, one might use another map of Europe with a different scale, or a bigger map that also includes Africa. Any of these would be a valid description of (a part of) the Earth's surface S.

Example 4.22. Next, consider the *unit sphere* (see the left plot in Figure 4.9),

$$\mathbb{S}^2 = \{(x,y,z)\in\mathbb{R}^3 \mid x^2+y^2+z^2=1\}.$$

Let us demonstrate one way $\mathbb{S}^2$ is described using regular, injective parametric surfaces.

First, we recall the parametric surface $\rho_{z,+}$ from Example 4.8:

$$\rho_{z,+}: B(0,1)\to\mathbb{R}^3,\qquad \rho_{z,+}(u,v)=\left(u,v,\sqrt{1-u^2-v^2}\right).$$

Observe that $\rho_{z,+}$ is injective, and the computations in Example 4.16 show that $\rho_{z,+}$ is regular. Moreover, Example 4.8 shows that the image of $\rho_{z,+}$ is the top half of the sphere:

$$\{(x, y, z) \in \mathbb{S}^2 \mid z > 0\}.$$

Next, to map the bottom half of the sphere, we simply negate the z-coordinate in $\rho_{z,+}$:

$$\rho_{z,-} : B(0, 1) \to \mathbb{R}^3, \qquad \rho_{z,-}(u, v) = \left(u, v, -\sqrt{1 - u^2 - v^2}\right).$$

Then, $\rho_{z,+}$ and $\rho_{z,-}$ together cover all of $\mathbb{S}^2$, except for the equatorial circle $z = 0$.

One (rather inefficient) way to cover this remaining equatorial circle is to define parametric surfaces that are like $\rho_{z,+}$ and $\rho_{z,-}$, except we switch around the roles of the three coordinates. More specifically, we define the following four parametric surfaces:

$$\rho_{y,\pm} : B(0, 1) \to \mathbb{R}^3, \qquad \rho_{y,\pm}(u, v) = \left(u, \pm\sqrt{1 - u^2 - v^2}, v\right),$$
$$\rho_{x,\pm} : B(0, 1) \to \mathbb{R}^3, \qquad \rho_{x,\pm}(u, v) = \left(\pm\sqrt{1 - u^2 - v^2}, u, v\right).$$

Notice that all six of the above parametric surfaces are regular and injective.

The right side of Figure 4.9 shows the images of all six parametric surfaces $\rho_{z,\pm}$, $\rho_{y,\pm}$, $\rho_{x,\pm}$. Note that these parametric surfaces together describe the full sphere—that is, $\mathbb{S}^2$ can be constructed by "gluing together" the images of these six parametric surfaces.

We also note that $\rho_{z,\pm}$, $\rho_{y,\pm}$, $\rho_{x,\pm}$ can be used to formally show that $\mathbb{S}^2$ satisfies the conditions of Definition 4.20 and hence is a surface. However, we omit the details here.

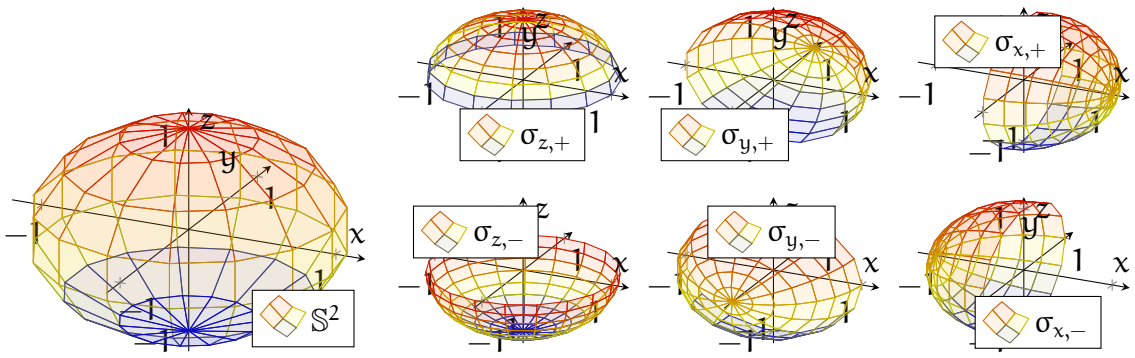


FIGURE 4.9. The left graphic is the sphere $\mathbb{S}^2$ from Example 4.22; the remaining plots show the images of the parametric surfaces $\rho_{z,\pm}$, $\rho_{y,\pm}$, $\rho_{x,\pm}$.

Finally, we give an informal discussion of the third principle—namely, that a surface should not be "self-intersecting". For this, we consider the object S in the left drawing of Figure 4.10, and we let $\mathbf{p}$ be the self-intersecting green point on S. Then, given any

open subset V as in Definition 4.20, the intersection $S \cap V$ must contain an "X-figure" around $\mathbf{p}$ made by two intersecting strips; see the right part of Figure 4.10.

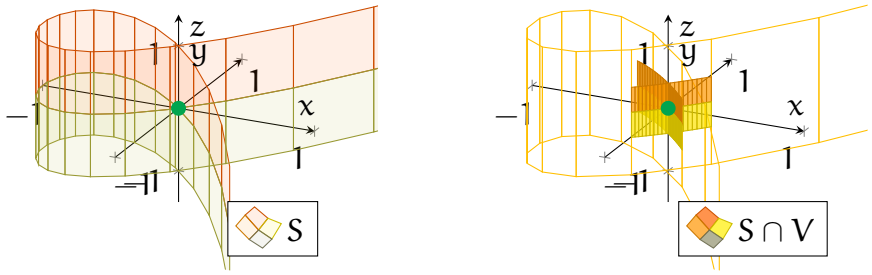


FIGURE 4.10. The object S in the above plots fails to be a surface. This is because at the green point $\mathbf{p}$, any open subset V that contains $\mathbf{p}$ must also contain an "X-figure" with two strips crossing each other. One particular example of this is illustrated in the drawing on the right.

The key point is that this X-figure has a different (topological) structure than an open set $U \subseteq \mathbb{R}^2$. The conditions on σ in Definition 4.20 prevent it from being able to map U onto this X-figure. To see this informally, you can convince yourself that it is not possible to bend a piece of paper (representing U) into this X-figure without either tearing it (violating continuity) or folding over itself (violating injectivity). Thus, we conclude that this S cannot satisfy Definition 4.20 and hence is not a surface.

Remark 4.23. The surfaces of Definition 4.20 are sometimes called *embedded surfaces*. One can, alternatively, study a more general class of 2-dimensional objects—*immersed surfaces*—that are allowed to pass through themselves.

4.4. **Descriptions of Surfaces.** We now discuss some more practical methods for constructing and describing surfaces. First, like for curves, it is often more convenient to describe surfaces using parametric surfaces that are not injective.

Definition 4.24. Let $S \subseteq \mathbb{R}^n$ be a surface. We refer to any regular (not necessarily injective) parametric surface $\sigma : U \to S$ as a parametrisation of S.

Example 4.25. Consider again the *cylinder* from Example 4.21:

$$\mathcal{C} = \{(x, y, z) \in \mathbb{R}^3 \mid x^2 + y^2 = 1\}.$$

In Example 4.21, we needed two injective regular parametric surfaces (σ_1 and σ_2) to describe all of $\mathcal{C}$. In fact, this is the best we can do, as one can show it is not possible to cover all of $\mathcal{C}$ using only a single injective and regular parametric surface.

However, we can do better if we *remove the requirement of injectivity*. Indeed, let us consider instead the parametric surface σ from Example 4.4:

$$\sigma : \mathbb{R}^2 \to \mathbb{R}^3, \qquad \sigma(u, v) = (\cos u, \sin u, v).$$

Observe (see also Example 2.23) that the image of σ is precisely all of $\mathcal{C}$. Furthermore, by the same computations as in Example 4.21, we obtain that σ is regular.

Thus, σ is a parametrisation of $\mathcal{C}$ whose image is all of $\mathcal{C}$. On the other hand, σ fails to be injective; for example, $\sigma(2\pi k, 0) = (1, 0, 0)$ for any $k \in \mathbb{Z}$.

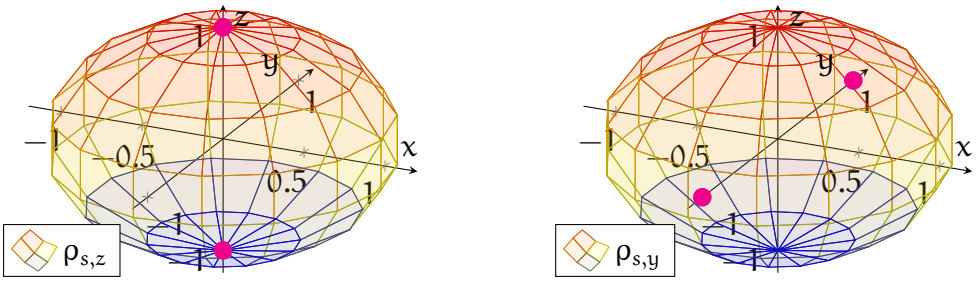


FIGURE 4.11. The two plots show the images of the parametrisations $\rho_{s,z}$ (left) and $\rho_{s,y}$ (right) of $\mathbb{S}^2$ from Example 4.26. In each drawing, the pair of pink-coloured points are excluded from the image.

Example 4.26. To contrast with the previous example, let us return to the *sphere*,

$$\mathbb{S}^2 = \{(x, y, z) \in \mathbb{R}^3 \mid x^2 + y^2 + z^2 = 1\}.$$

Example 4.25 demonstrated that the cylinder $\mathcal{C}$ could be described using only a single parametrisation. On the other hand, *it is impossible to cover all of* $\mathbb{S}^2$ *using only one parametrisation of* $\mathbb{S}^2$, injective or not. (Sadly, we cannot give a proof of this here.)

Let us instead describe $\mathbb{S}^2$ using two parametrisations. There are many ways to do this, but here make use of *spherical coordinates* (see also Example 4.12):

$$\rho_{s,z} : \mathbb{R} \times (0, \pi) \to \mathbb{S}^2, \qquad \rho_{s,z}(u, v) = (\cos u \sin v, \sin u \sin v, \cos v).$$

To see that $\rho_{s,z}$ is regular, we compute, for each $(u, v) \in \mathbb{R} \times (0, \pi)$:

$$\partial_1 \rho_{s,z}(u, v) = (-\sin u \sin v, \cos u \sin v, 0),$$
$$\partial_2 \rho_{s,z}(u, v) = (\cos u \cos v, \sin u \cos v, -\sin v).$$

Taking a cross product then yields

$$|\partial_1 \rho_{s,z}(u, v) \times \partial_2 \rho_{s,z}(u, v)| = |(-\cos u \sin^2 v, -\sin u \sin^2 v, -\sin v \cos v)|$$

$$= |\sin v|\,|(\cos u \sin v, \sin u \sin v, \cos v)|$$
$$= \sin v.$$

Since $v \in (0, \pi)$, then $\sin v \neq 0$, and Theorem 4.17 yields that $\rho_{s,z}$ is regular.

See the left part of Figure 4.11 for a plot of $\rho_{s,z}$. In particular, the image of $\rho_{s,z}$ is all of $\mathbb{S}^2$ except for the north pole $(0, 0, 1)$ (corresponding to $v = 0$) and the south pole $(0, 0, -1)$ (corresponding to $v = \pi$). We cannot include those two points into $\rho_{s,z}$, since $\sin 0 = \sin \pi = 0$, and hence the above computation implies that $\rho_{s,z}$ cannot be regular if we also allow $v = 0$ and $v = \pi$ into its domain.

To cover all of $\mathbb{S}^2$, we need a second parametrisation. An easy way to do this is to just permute the components of $\rho_{s,z}$, so that two different points are excluded:

$$\rho_{s,y} : \mathbb{R} \times (0, \pi) \to \mathbb{S}^2, \qquad \rho_{s,y}(u, v) = (\sin u \sin v, \cos v, \cos u \sin v),$$

By similar reasoning, $\rho_{s,y}$ is a parametrisation of $\mathbb{S}^2$, and its image is all of $\mathbb{S}^2$ except for the points $(0, \pm 1, 0)$. Thus, *the parametrisations $\rho_{s,z}$ and $\rho_{s,y}$ together cover all of $\mathbb{S}^2$.*

Next, we state an analogue of Theorem 3.24 for surfaces—that *level sets of sufficiently nice real-valued functions of three variables are surfaces*:

Theorem 4.27. Suppose $V \subseteq \mathbb{R}^3$ is open and connected, and let $f : V \to \mathbb{R}$ be a smooth function. Moreover, let $c \in \mathbb{R}$, and let S denote the level set

$$S = \{(x, y, z) \in V \mid f(x, y, z) = c\}.$$

If $\nabla f(\mathbf{p})$ is nonzero for every $\mathbf{p} \in S$, then S is a surface.

Due to lack of background, we omit the proof of Theorem 4.27. Like the proof of Theorem 3.24, it depends primarily on the *implicit function theorem*.

Next, we apply Theorem 4.27 to some simple objects we have already studied:

Example 4.28. Consider the function

$$s : \mathbb{R}^3 \to \mathbb{R}, \qquad s(x, y, z) = x^2 + y^2 + z^2.$$

Observe that the *unit sphere* $\mathbb{S}^2$ (see Example 4.22) is precisely the level set

$$\mathbb{S}^2 = \{(x, y, z) \in \mathbb{R}^3 \mid s(x, y, z) = 1\}.$$

Note also that s is smooth, and its gradient satisfies

$$\nabla s(x, y, z) = (2x, 2y, 2z)_{(x,y,z)}$$

which vanishes only when $(x, y, z) = (0, 0, 0)$. Now, since $(0, 0, 0) \notin \mathbb{S}^2$, it follows from Theorem 4.27 (with $f = s$ and $V = \mathbb{R}^3$) that $\mathbb{S}^2$ *is indeed a surface*.

Example 4.29. Next, notice that the cylinder $\mathcal{C}$ from Example 4.21 can be written as

$$\mathcal{C} = \{(x, y, z) \in \mathbb{R}^3 \mid b(x, y, z) = 1\},$$

where b is the real-valued function

$$b : \mathbb{R}^3 \to \mathbb{R}, \qquad b(x, y, z) = x^2 + y^2.$$

A direct computation shows that

$$\nabla b(x, y, z) = (2x, 2y, 0)_{(x,y,z)}$$

which vanishes only when $(x, y) = (0, 0)$. Since x and y cannot both vanish at any point $(x, y, z) \in \mathcal{C}$, then Theorem 4.27 (with $f = b$) implies that $\mathcal{C}$ is a surface.

Theorem 4.30. Let $h : U \to \mathbb{R}$ be smooth, with $U \subseteq \mathbb{R}^2$ open and connected. Then,

$$G_h^z = \{(u, v, h(u, v)) \mid (u, v) \in U\},$$
$$G_h^y = \{(u, h(u, v), v) \mid (u, v) \in U\},$$
$$G_h^x = \{(h(u, v), u, v) \mid (u, v) \in U\},$$

(i.e. the graphs of h) are surfaces.

Proof. The proof is analogous to that of Theorem 3.28. In particular, we note that

$$G_h^z = \{(x, y, z) \in U \times \mathbb{R} \mid z - h(x, y) = 0\},$$

and we apply Theorem 4.27 to the above, with the function

$$F : U \times \mathbb{R} \to \mathbb{R}, \qquad F(x, y, z) = z - h(x, y).$$

The remainings sets G_h^y and G_h^x can be treated similarly. □

Example 4.31. Consider the quadratic function

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

Then, by Theorem 4.30, the graph of q,

$$\mathcal{P} = \{(x, y, x^2 + y^2) \mid (x, y) \in \mathbb{R}^2\},$$

is a surface, known as a *paraboloid*; see Figure 4.12.

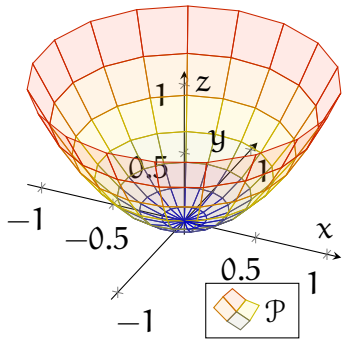


FIGURE 4.12. The above shows the paraboloid $\mathcal{P}$ from Example 4.31.

Thus far, our examples have only involved surfaces lying in 3-dimensional space. This will also be the case for the remainder of this module. However, let us take a bit of extra time here to consider the following question:

Question 4.32. Are there surfaces lying in higher dimensions that are "genuinely different" from all surfaces lying in $\mathbb{R}^3$. In other words, if we only consider surfaces in $\mathbb{R}^3$, then will we be missing any interesting surface geometries?

Unfortunately, the answer to Question 4.32 is *yes*. One 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 the left drawing in Figure 4.13. If we glue two opposite edges of this rectangle together, we then obtain a cylindrical strip. This is demonstrated in the middle drawing in Figure 4.13, where the left and right edges in the left drawing (marked in red) are now glued together. Finally, for the two remaining edges (marked in blue), we glue them together *in the opposite orientation*, as indicated by the blue arrows in Figure 4.13. Performing both gluings yields the right drawing in Figure 4.13.

Note that while this second gluing can be done abstractly, you could not physically do this with a piece of paper. You could not bring the blue edges of the paper together in this way without the paper passing through itself. (Try it yourself!)

However, such a gluing would be possible *if we had an extra dimension to move about*. If the paper is sitting in $\mathbb{R}^4$ instead of $\mathbb{R}^3$, 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 the geometric object one obtains as a result of this.

Finally, in a fantastically surprising contrast, one can prove that *every type of surface can be embedded in* $\mathbb{R}^4$. More specifically, given any surface S in $\mathbb{R}^n$, there is

an "equivalent" (we avoid making this term precise here) way to also describe S as a subset of $\mathbb{R}^4$. Thus, if we worked instead within $\mathbb{R}^4$ rather than in $\mathbb{R}^3$, then we would be able to essentially study every possible surface, including the Klein bottle.

The aforementioned result is a special case of the famous *Whitney embedding theorem*, named after *Hassler Whitney* (American mathematician, 1907–1989).

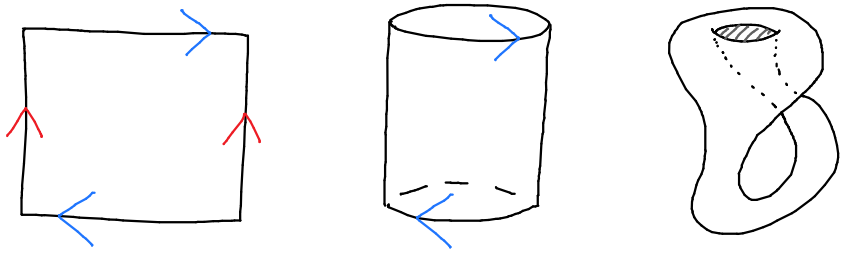


FIGURE 4.13. The left drawing shows a rectangle, with each opposite pair of edges marked with the same colour. Gluing together the red edges, with the orientations indicated by the red arrows, results in the cylindrical strip in the middle drawing. Next, gluing the blue edges, with the opposite orientations indicated by the blue arrows, yields the Klein bottle on the right.

4.5. **Tangent Planes.** Having formally defined surfaces, our next goal is to study their properties. Like for curves, we usually work with surfaces via their parametrisations, hence we tend to deal more directly with properties of parametric surfaces. It is then pertinent to ask whether such information is also a property of the underlying surface—in other words, *is this property independent of parametrisation?*

We now explore this question in the particular case of *tangent planes* of parametric surfaces; see Definition 4.6. Before making any formal statements, let us first motivate our discussion through a concrete example involving the sphere:

Example 4.33. Consider the *sphere* $\mathbb{S}^2$ from Example 4.22,

$$\mathbb{S}^2 = \{(x, y, z) \in \mathbb{R}^3 \mid x^2 + y^2 + z^2 = 1\},$$

along with the point $\mathbf{p} = (-1, 0, 0) \in \mathbb{S}^2$. As was noted in Example 4.22, one particular parametrisation of $\mathbb{S}^2$ that includes $\mathbf{p}$ in its image is given by

$$\rho_{x,-} : B(0, 1) \to \mathbb{S}^2, \qquad \rho_{x,-}(u, v) = \left(-\sqrt{1 - u^2 - v^2}, u, v\right).$$

Also, Example 4.26 provides another parametrisation of $\mathbb{S}^2$ containing $\mathbf{p}$:

$$\rho_{s,z} : \mathbb{R} \times (0, \pi) \to \mathbb{S}^2, \qquad \rho_{s,z}(u, v) = (\cos u \sin v, \sin u \sin v, \cos v).$$

For $\rho_{x,-}$, we first note that $\mathbf{p} = \rho_{x,-}(0,0)$. Taking partial derivatives yields

$$\partial_1\rho_{x,-}(u,v) = \left(\frac{u}{\sqrt{1-u^2-v^2}}, 1, 0\right), \qquad \partial_2\rho_{x,-}(u,v) = \left(\frac{v}{\sqrt{1-u^2-v^2}}, 0, 1\right),$$
$$\partial_1\rho_{x,-}(0,0) = (0,1,0), \qquad \partial_2\rho_{x,-}(0,0) = (0,0,1).$$

Recalling the equation (4.2), we then obtain

$$\mathsf{T}_{\rho_{x,-}}(0,0) = \left\{a\cdot(0,1,0)_{(-1,0,0)} + b\cdot(0,0,1)_{(-1,0,0)} \middle| a,b\in\mathbb{R}\right\}.$$

Similarly, for $\rho_{s,z}$, we begin by noting that $\mathbf{p} = \rho_{s,z}(\pi, \frac{\pi}{2})$. Recalling the computations in Example 4.26 for derivatives of $\rho_{s,z}$, we see that

$$\partial_1\rho_{s,z}\left(\pi,\frac{\pi}{2}\right) = (0,-1,0), \qquad \partial_2\rho_{s,z}\left(\pi,\frac{\pi}{2}\right) = (0,0,-1).$$

As a result, we obtain

$$\mathsf{T}_{\rho_{s,z}}\left(\pi,\frac{\pi}{2}\right) = \left\{a\cdot(0,-1,0)_{(-1,0,0)} + b\cdot(0,0,-1)_{(-1,0,0)} \middle| a,b\in\mathbb{R}\right\}.$$

Finally, comparing the above two computations, we conclude that

$$\mathsf{T}_{\rho_{x,-}}(0,0) = \mathsf{T}_{\rho_{s,z}}\left(\pi,\frac{\pi}{2}\right),$$

that is, *the tangent planes at* $\mathbf{p}$ *computed through the parametrisations* $\rho_{x,-}$ *and* $\rho_{s,z}$ *are the same*. See Figure 4.14 for illustrations of $\rho_{x,-}$, $\rho_{s,z}$, and their tangent planes.

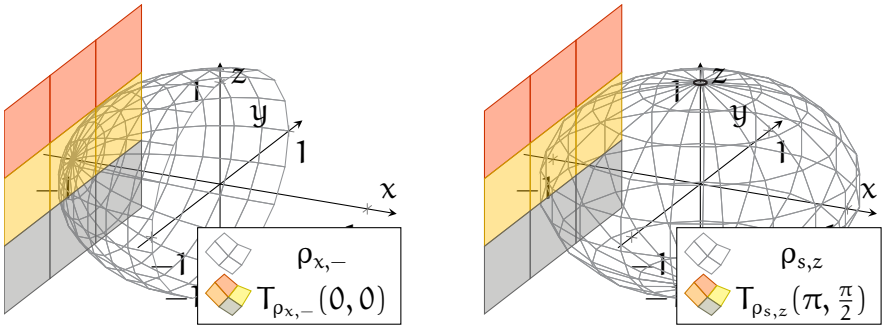


FIGURE 4.14. These figures show the two parametrisations of $\mathbb{S}^2$ discussed in Example 4.33. The left graphic depicts $\rho_{x,-}$ and the tangent plane $\mathsf{T}_{\rho_{x,-}}(0,0)$, while the right graphic depicts $\rho_{s,z}$ and $\mathsf{T}_{\rho_{s,z}}(\pi,\frac{\pi}{2})$.

Example 4.33 provides some initial evidence that tangent planes are independent of parametrisation and hence define a geometric property of surfaces. Our next objective is to show that this is indeed true in general. To discuss this more precisely, we adopt an approach that is similar to our previous discussion for curves.

Definition 4.34. Let $\sigma : U \to \mathbb{R}^n$ and $\tilde{\sigma} : \tilde{U} \to \mathbb{R}^n$ be regular parametric surfaces. We say σ is a reparametrisation of $\tilde{\sigma}$ iff there is a bijection $\Phi : U \to \tilde{U}$ such that:

- Both Φ and its inverse Φ^{-1} are smooth.
- The following holds for all $(u, v) \in U$:

$$\tilde{\sigma}(\Phi(u, v)) = \sigma(u, v). \tag{4.5}$$

In the above context, we refer to Φ as the corresponding change of variables.

Remark 4.35. Definition 4.34 is the analogue of Definition 3.12 for reparametrisations of parametric curves. In particular, the change of variables Φ matches the parameter (u, v) for σ to the parameter $(\tilde{u}, \tilde{v}) = \Phi(u, v)$ for $\tilde{\sigma}$ that maps to the same point.

Remark 4.36. One can again show that if σ is a reparametrisation of $\tilde{\sigma}$, then $\tilde{\sigma}$ is also a reparametrisation of σ. Thus, when convenient, we will write "σ and $\tilde{\sigma}$ are reparametrisations of each other" in the place of "σ is a reparametrisation of $\tilde{\sigma}$".

The next theorem shows that reparametrisations do not alter tangent planes:

Theorem 4.37. Let $\sigma : U \to \mathbb{R}^n$ and $\tilde{\sigma} : \tilde{U} \to \mathbb{R}^n$ be regular parametric surfaces. Moreover, assume σ is a reparametrisation of $\tilde{\sigma}$, with corresponding change of variables $\Phi : U \to \tilde{U}$. Then, given any $(u, v) \in U$, we have the following identity:

$$T_\sigma(u, v) = T_{\tilde{\sigma}}(\Phi(u, v)). \tag{4.6}$$

Proof. The proof is analogous to that of Theorem 3.34 for tangent lines. We let $\tilde{u}$ and $\tilde{v}$ denote the individual components of Φ—that is, we set

$$\Phi(u, v) = (\tilde{u}(u, v), \tilde{v}(u, v)).$$

Then, the relation (4.5) can be written as

$$\sigma(u, v) = \tilde{\sigma}(\tilde{u}(u, v), \tilde{v}(u, v)).$$

Using the above along with the (multivariable) chain rule, we can expand

$$\begin{aligned}
\partial_1\sigma(u, v) &= \partial_1\tilde{\sigma}(\tilde{u}(u, v), \tilde{v}(u, v)) \cdot \partial_1\tilde{u}(u, v) + \partial_2\tilde{\sigma}(\tilde{u}(u, v), \tilde{v}(u, v)) \cdot \partial_1\tilde{v}(u, v), \\
&= \partial_1\tilde{\sigma}(\Phi(u, v)) \cdot \partial_1\tilde{u}(u, v) + \partial_2\tilde{\sigma}(\Phi(u, v)) \cdot \partial_1\tilde{v}(u, v), \\
\partial_2\sigma(u, v) &= \partial_1\tilde{\sigma}(\tilde{u}(u, v), \tilde{v}(u, v)) \cdot \partial_2\tilde{u}(u, v) + \partial_2\tilde{\sigma}(\tilde{u}(u, v), \tilde{v}(u, v)) \cdot \partial_2\tilde{v}(u, v) \\
&= \partial_1\tilde{\sigma}(\Phi(u, v)) \cdot \partial_2\tilde{u}(u, v) + \partial_2\tilde{\sigma}(\Phi(u, v)) \cdot \partial_2\tilde{v}(u, v).
\end{aligned}$$

Combining (4.2) and the above, we see that

$$\begin{aligned}
T_\sigma(u,v) &= \left\{ a\cdot \partial_1\sigma(u,v)_{\sigma(u,v)} + b\cdot \partial_2\sigma(u,v)_{\sigma(u,v)} \middle| a,b\in\mathbb{R}\right\}\\
&= \Big\{ [a\cdot \partial_1\tilde{u}(u,v) + b\cdot\partial_2\tilde{u}(u,v)]\cdot \partial_1\tilde{\sigma}(\Phi(u,v))_{\tilde{\sigma}(\Phi(u,v))}\\
&\qquad + [a\cdot \partial_1\tilde{v}(u,v) + b\cdot\partial_2\tilde{v}(u,v)]\cdot \partial_2\tilde{\sigma}(\Phi(u,v))_{\tilde{\sigma}(\Phi(u,v))} \Big| a,b\in\mathbb{R}\Big\}\\
&\subseteq T_{\tilde{\sigma}}(\Phi(u,v)).
\end{aligned}$$

Finally, we can repeat all of the above work, but with the roles of σ and $\tilde{\sigma}$ interchanged; this yields the opposite relation $T_{\tilde{\sigma}}(\Phi(u,v)) \subseteq T_\sigma(u,v)$. □

In light of Theorem 4.37, we can now make sense of tangent planes to surfaces:

Definition 4.38. The tangent plane to a surface $S\subseteq\mathbb{R}^n$ at a point $\mathbf{p}\in S$ is defined as

$$T_{\mathbf{p}}S = T_\sigma(u_0,v_0), \tag{4.7}$$

where $\sigma : U\to S$ is any parametrisation of S, with $(u_0,v_0)\in U$ satisfying $\sigma(u_0,v_0)=\mathbf{p}$.

Example 4.39. Recall the *cylinder* from Example 4.21,

$$\mathcal{C} = \{(x,y,z)\in\mathbb{R}^3 \mid x^2+y^2=1\}.$$

Let us find the tangent plane $T_{\mathbf{p}}\mathcal{C}$, at the point $\mathbf{p}=(0,1,5)\in\mathcal{C}$.

The first step is to choose a parametrisation of $\mathcal{C}$ which covers $\mathbf{p}$. The simplest choice is to take the parametrisation σ from Example 4.25,

$$\sigma:\mathbb{R}^2\to\mathcal{C}, \qquad \sigma(u,v) = (\cos u, \sin u, v),$$

which satisfies $\mathbf{p}=\sigma(\frac{\pi}{2},5)$. Furthermore, we can compute that

$$\partial_1\sigma\left(\frac{\pi}{2},5\right) = (-1,0,0), \qquad \partial_2\sigma\left(\frac{\pi}{2},5\right) = (0,0,1).$$

As a result, by Definition 4.38, we have

$$\begin{aligned}
T_{\mathbf{p}}\mathcal{C} &= T_\sigma\left(\frac{\pi}{2},5\right)\\
&= \left\{ a\cdot(-1,0,0)_{(0,1,5)} + b\cdot(0,0,1)_{(0,1,5)} \middle| a,b\in\mathbb{R}\right\}.
\end{aligned}$$

A plot of $\mathcal{C}$ and $T_{\mathbf{p}}\mathcal{C}$ is found in the left drawing of Figure 4.15.

We conclude our discussion by highlighting a connection with linear algebra. Assume the setting of Definition 4.38. Recall $T_{\mathbf{p}}S = T_\sigma(u_0,v_0)$ is a 2-dimensional vector space, and the tangent vectors $\partial_1\sigma(u_0,v_0)_{\mathbf{p}}$ and $\partial_2\sigma(u_0,v_0)_{\mathbf{p}}$ form a basis for $T_{\mathbf{p}}S$.

Moreover, suppose $\tilde{\sigma} : \tilde{U} \to S$ is another parametrisation of S, with $\mathbf{p} = \tilde{\sigma}(\tilde{u}_0, \tilde{v}_0)$ for some $(\tilde{u}_0, \tilde{v}_0) \in \tilde{U}$. Since Definition 4.38 implies $T_\mathbf{p}S = T_{\tilde{\sigma}}(\tilde{u}_0, \tilde{v}_0)$ as well, the tangent vectors $\partial_1\tilde{\sigma}(\tilde{u}_0, \tilde{v}_0)_\mathbf{p}$ and $\partial_2\tilde{\sigma}(\tilde{u}_0, \tilde{v}_0)_\mathbf{p}$ yield a different basis for $T_\mathbf{p}S$.

In other words, *any parametrisation of* S *that contains* $\mathbf{p}$ *yields an associated basis of the* 2*-dimensional vector space* $T_\mathbf{p}S$. Furthermore, *when we change parametrisations, this is observed at the level of* $T_\mathbf{p}S$ *as a change of its basis.*

More intuitively, one can think of a parametrisation of S as a particular person's perspective (or, in physics terms, *frame of reference*) of S. Similarly, one can think of a basis of $T_\mathbf{p}S$ as a particular way of viewing this tangent plane. Consequently, when one switch parametrisations, the resulting change of basis of $T_\mathbf{p}S$ can hence be interpreted as a linear change of perspective for $T_\mathbf{p}S$.

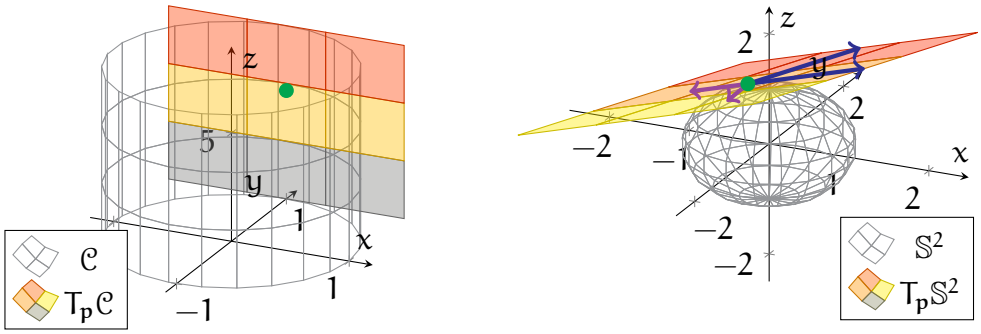


FIGURE 4.15. The left illustration contains the cylinder $\mathcal{C}$ and the tangent plane $T_\mathbf{p}\mathcal{C}$ from Example 4.39. The right plot shows $\mathbb{S}^2$ and $T_\mathbf{p}\mathbb{S}^2$ from Example 4.40; the basis vectors of $T_\mathbf{p}\mathbb{S}^2$ obtained from $\rho_{x,-}$ are drawn in blue, while the basis vectors of $T_\mathbf{p}\mathbb{S}^2$ from $\rho_{s,z}$ are drawn in purple.

Example 4.40. 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 $\rho_{x,-}$ and $\rho_{s,z}$ from Example 4.33, with

$$\mathbf{p} = \rho_{x,-}\left(\frac{1}{2}, \frac{1}{\sqrt{2}}\right) = \rho_{s,z}\left(\frac{3\pi}{4}, \frac{\pi}{4}\right).$$

We can then compute the associated partial derivatives:

$$\partial_1\rho_{x,-}\left(\frac{1}{2}, \frac{1}{\sqrt{2}}\right) = (1, 1, 0), \qquad \partial_2\rho_{x,-}\left(\frac{1}{2}, \frac{1}{\sqrt{2}}\right) = (\sqrt{2}, 0, 1),$$

$$\partial_1\rho_{s,z}\left(\frac{3\pi}{4}, \frac{\pi}{4}\right) = \left(-\frac{1}{2}, -\frac{1}{2}, 0\right), \qquad \partial_2\rho_{s,z}\left(\frac{3\pi}{4}, \frac{\pi}{4}\right) = \left(-\frac{1}{2}, \frac{1}{2}, -\frac{1}{\sqrt{2}}\right).$$

Then, the bases of $T_p\mathbb{S}^2$ associated with $\rho_{x,-}$ and $\rho_{s,z}$, respectively, are

$$\mathcal{B}_1 = \left\{(1,1,0)_p, (\sqrt{2},0,1)_p\right\}, \qquad \mathcal{B}_2 = \left\{\left(-\frac{1}{2},-\frac{1}{2},0\right)_p, \left(-\frac{1}{2},\frac{1}{2},-\frac{1}{\sqrt{2}}\right)_p\right\}.$$

These bases are shown in the right plot of Figure 4.15 in blue and purple, respectively. Note that while $\mathcal{B}_1$ and $\mathcal{B}_2$ are clearly different, they both span the same space T_pS.

4.6. **Normal Vectors.** In the remainder of this chapter, we will *restrict ourselves to surfaces lying only in* $\mathbb{R}^3$ (which we have already been doing in our examples).

Let $S \subseteq \mathbb{R}^3$ be a surface, and let $\mathbf{p} \in S$. Similar to our discussions for curves, the tangent plane T_pS is a 2-dimensional subspace of the 3-dimensional $T_p\mathbb{R}^3$. This leaves one remaining dimension in $T_p\mathbb{R}^3$ normal to T_pS. Since T_pS captures all the directions along S at $\mathbf{p}$, *this last dimension represents the directions normal to* S *at* $\mathbf{p}$.

Again, it will be useful to pick out *unit* vectors from this normal dimension:

Definition 4.41. Let $S \subseteq \mathbb{R}^3$ be a surface, and let $\mathbf{p} \in S$. Then, $\mathbf{n}_p \in T_p\mathbb{R}^3$ is a unit normal to S at $\mathbf{p}$ iff $\mathbf{n}_p$ is normal to every element of T_pS and $|\mathbf{n}_p| = 1$.

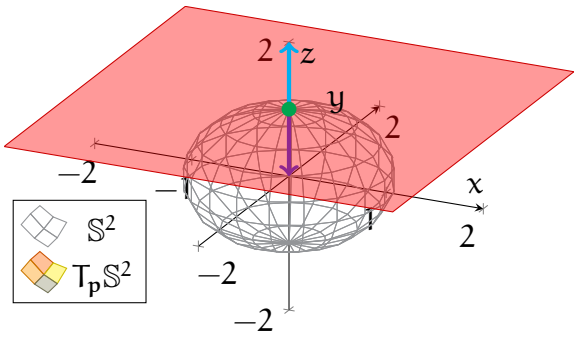


FIGURE 4.16. This is the setting of Example 4.42—the point $\mathbf{p}$ is drawn in green, while the unit normals $(0,0,\pm1)_p$ are in blue and purple.

Example 4.42. Consider the *sphere* $\mathbb{S}^2$ (see Example 4.22),

$$\mathbb{S}^2 = \{(x,y,z) \in \mathbb{R}^3 \mid x^2 + y^2 + z^2 = 1\},$$

along with the point $\mathbf{p} = (0,0,1) \in \mathbb{S}^2$. Since $\mathbf{p}$ is the north pole of $\mathbb{S}^2$, the tangent plane $T_p\mathbb{S}^2$ is simply a copy of the xy-plane on $\mathbf{p}$:

$$T_p\mathbb{S}^2 = \{a \cdot (1,0,0)_p + b \cdot (0,1,0)_p | a,b \in \mathbb{R}\}.$$

(This can also be computed using the parametrisation $\rho_{z,+}$ from Example 4.22.)

Thus, the z-direction is perpendicular to $T_p\mathbb{S}^2$. In particular, the unit length arrows in this direction are given by $(0,0,\pm 1)_p$. Definition 4.41 then implies $(0,0,\pm 1)_p$ are both unit normals to $\mathbb{S}^2$ at $\mathbf{p} = (0,0,1)$; see Figure 4.16 for an illustration of this setting.

A general observation, which you would likely find unsurprising, is that a surface in $\mathbb{R}^3$ always has two unit normals at each of its points:

Theorem 4.43. Let $S \subseteq \mathbb{R}^3$ be a surface, and let $\mathbf{p} \in S$. Then:

- There are exactly two unit normals to S at any $\mathbf{p}$.
- If $\mathbf{n}_p$ is a unit normal to S at $\mathbf{p}$, then so is $-\mathbf{n}_p$.

Proof. That there are two unit normals follows from the fact that the elements of $T_p\mathbb{R}^3$ that are perpendicular to T_pS form a 1-dimensional subspace, from which there are exactly two elements with norm 1. For the remaining statement, we simply note that if $\mathbf{n}_p$ satisfies the conditions of Definition 4.41, then so does $-\mathbf{n}_p$. □

The next part of our discussion concerns a more practical question: *how do we compute unit normals*? One general method is through parametrisations:

Theorem 4.44. Let $S \subseteq \mathbb{R}^3$ be a surface, let $\sigma : U \to S$ be a parametrisation of S, and fix $(u_0, v_0) \in U$. Then, the unit normals to S at $\mathbf{p} = \sigma(u_0, v_0)$ are given by

$$\mathbf{n}_p^{\pm} = \pm \left[\frac{\partial_1\sigma(u_0,v_0) \times \partial_2\sigma(u_0,v_0)}{|\partial_1\sigma(u_0,v_0) \times \partial_2\sigma(u_0,v_0)|} \right]_p. \tag{4.8}$$

Proof. Recall that since σ is regular, Theorem 4.15 and Definition 4.38 imply that $\partial_1\sigma(u_0,v_0)_p$ and $\partial_2\sigma(u_0,v_0)_p$ form a basis for the tangent plane $T_\sigma(u_0,v_0) = T_pS$. As a result, by Theorem 2.18, the cross product

$$\partial_1\sigma(u_0,v_0)_p \times \partial_2\sigma(u_0,v_0)_p = [\partial_1\sigma(u_0,v_0) \times \partial_2\sigma(u_0,v_0)]_p$$

is nonzero and is normal to both of its factors, and hence to T_pS. Finally, to obtain the unit normals, we divide the above by its norm and recall Theorem 4.43. □

Example 4.45. Consider next the *cylinder* from Example 4.21,

$$\mathcal{C} = \{(x,y,z) \in \mathbb{R}^3 \mid x^2 + y^2 = 1\}.$$

In addition, recall from Example 4.25 that

$$\sigma : \mathbb{R}^2 \to \mathcal{C}, \qquad \sigma(u,v) = (\cos u, \sin u, v)$$

is a parametrisation of $\mathcal{C}$ whose image is all of $\mathcal{C}$.

Direct computations show that for any $(u, v) \in \mathbb{R}^2$,

$$\partial_1\sigma(u,v) = (-\sin u, \cos u, 0), \qquad \partial_2\sigma(u,v) = (0,0,1),$$
$$\partial_1\sigma(u,v) \times \partial_2\sigma(u,v) = (\cos u, \sin u, 0), \qquad |\partial_1\sigma(u,v) \times \partial_2\sigma(u,v)| = 1.$$

Thus, by Theorem 4.44, we conclude that for any $\mathbf{p} = \sigma(u_0, v_0) \in \mathcal{C}$, the tangent vectors

$$\pm\left[\frac{\partial_1\sigma(u_0,v_0) \times \partial_2\sigma(u_0,v_0)}{|\partial_1\sigma(u_0,v_0) \times \partial_2\sigma(u_0,v_0)|}\right]_p = \pm(\cos u_0, \sin u_0, 0)_{(\cos u_0, \sin u_0, v_0)}.$$

are the unit normals to $\mathcal{C}$ at $\mathbf{p}$.

To state this more neatly, we set

$$(\cos u_0, \sin u_0, v_0) = \sigma(u_0, v_0) = (x, y, z).$$

From the above, we conclude that the unit normals to $\mathcal{C}$ at (x, y, z) are

$$\pm\mathbf{n}_p = \pm(\cos u_0, \sin u_0, 0)_{(\cos u_0, \sin u_0, v_0)} = \pm(x, y, 0)_{(x,y,z)}.$$

Some examples of unit normals to $\mathcal{C}$ are drawn in the left plot of Figure 4.17.

Furthermore, for surfaces that are level sets, in the sense of Theorem 4.27, there is a more direct method for computing their unit normals:

Theorem 4.46. Assume the setting of Theorem 4.27—let S be the surface

$$S = \{(x, y, z) \in V \mid f(x, y, z) = c\},$$

where $V \subseteq \mathbb{R}^3$ is open, $c \in \mathbb{R}$, and $f : V \to \mathbb{R}$ is a smooth function such that $\nabla f(\mathbf{q})$ is nonzero for each $\mathbf{q} \in S$. Then, for any $\mathbf{p} \in S$, the unit normals to S at $\mathbf{p}$ are given by

$$\mathbf{n}_p^{\pm} = \pm|\nabla f(\mathbf{p})|^{-1} \cdot \nabla f(\mathbf{p}). \tag{4.9}$$

Proof. Let $\sigma : U \to S$ be a parametrisation of S, and suppose $\sigma(u_0, v_0) = \mathbf{p}$, with $(u_0, v_0) \in U$. By the definition of S, we have, for any $(u, v) \in V$, that

$$f(\sigma(u,v)) = c, \qquad \partial_u[f(\sigma(u,v))] = 0, \qquad \partial_v[f(\sigma(u,v))] = 0.$$

The second and third equations can be expanded using the chain rule:

$$0 = \nabla f(\sigma(u,v)) \cdot \partial_1\sigma(u,v)_{\sigma(u,v)}, \qquad 0 = \nabla f(\sigma(u,v)) \cdot \partial_2\sigma(u,v)_{\sigma(u,v)}.$$

Taking linear combinations of the above and setting $(u, v) = (u_0, v_0)$ yields

$$0 = \nabla f(\mathbf{p}) \cdot \left[a \cdot \partial_1\sigma(u_0,v_0)_{\sigma(u_0,v_0)} + b \cdot \partial_2\sigma(u_0,v_0)_{\sigma(u_0,v_0)}\right],$$

for any $a, b \in \mathbb{R}$. Thus, $\pm\nabla f(\mathbf{p})$ is normal to every element of $T_\sigma(u_0, v_0)$, which by Definition 4.38 is $T_\mathbf{p}S$. The desired result now follows immediately. □

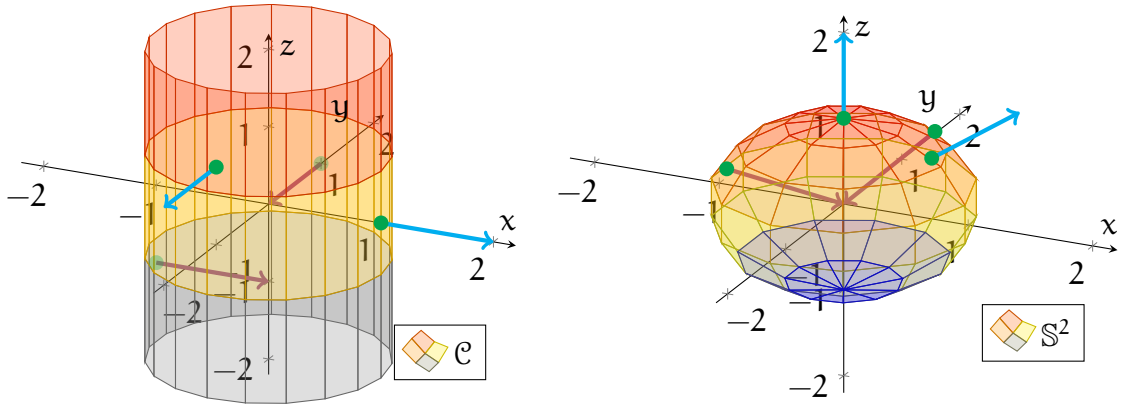


FIGURE 4.17. The above drawings contain plots of the cylinder $\mathcal{C}$ (left) from Examples 4.45 and 4.47 and the sphere $\mathbb{S}^2$ (right) from Example 4.48. Some unit normals for each surface are drawn in blue and purple.

Example 4.47. We return to the *cylinder* $\mathcal{C}$ from Example 4.45, and we now find its unit normals using Theorem 4.46. Recall, from Example 4.29, that $\mathcal{C}$ can be written as

$$\mathcal{C} = \{(x, y, z) \in \mathbb{R}^3 \mid b(x, y, z) = 1\},$$

where $b : \mathbb{R}^3 \to \mathbb{R}$ is the function given by $b(x, y, z) = x^2 + y^2$.

Furthermore, from computations in Example 4.29, we have

$$\nabla b(x, y, z) = (2x, 2y, 0)_{(x,y,z)}$$

for each $(x, y, z) \in \mathcal{C}$, and

$$\begin{aligned} |\nabla b(x, y, z)| &= 2\sqrt{x^2 + y^2} \\ &= 2, \end{aligned}$$

where we also used that $b(x, y, z) = 1$ on $\mathcal{C}$. Thus, by Theorem 4.46, we obtain

$$\pm\frac{\nabla b(x, y, z)}{|\nabla b(x, y, z)|} = \pm(x, y, 0)_{(x,y,z)}$$

for the unit normals to $\mathcal{C}$ at (x, y, z). Note this is the same answer as in Example 4.45.

Example 4.48. Recall from Example 4.28 that the *sphere* $\mathbb{S}^2$ can be expressed as

$$\mathbb{S}^2 = \{(x, y, z) \in \mathbb{R}^3 \mid s(x, y, z) = 1\},$$

where the function $s : \mathbb{R}^3 \to \mathbb{R}$ is given by $s(x, y, z) = x^2 + y^2 + z^2$. Furthermore, for any $(x, y, z) \in \mathbb{S}^2$, the gradient of s at (x, y, z) satisfies

$$\nabla s(x, y, z) = (2x, 2y, 2z)_{(x,y,z)}, \qquad |\nabla s(x, y, z)| = 2.$$

Thus, by Theorem 4.46, the unit normals to $\mathbb{S}^2$ at (x, y, z) are

$$\pm \frac{\nabla s(x, y, z)}{|\nabla s(x, y, z)|} = \pm (x, y, z)_{(x,y,z)}.$$

In other words, the unit normals of $\mathbb{S}^2$ at any of its points $\mathbf{p}$ is simply $\pm \mathbf{p}_{\mathbf{p}}$. See the right plot within Figure 4.17 for some illustrations of unit normals of $\mathbb{S}^2$.

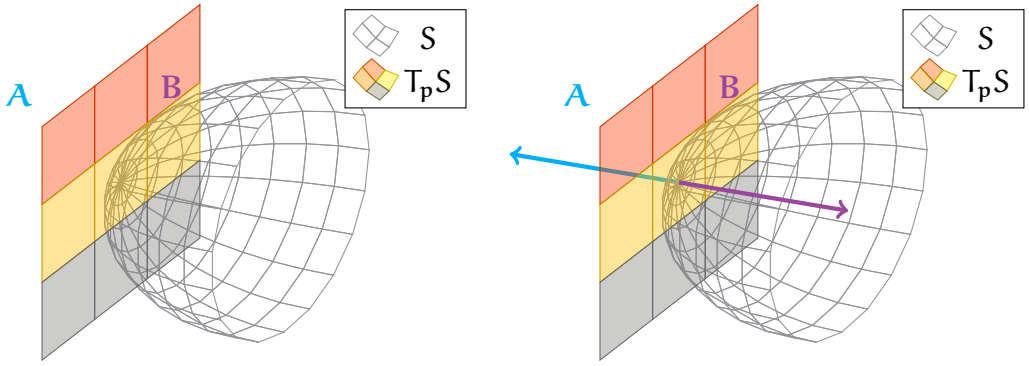


FIGURE 4.18. The left illustration shows the two sides (labelled "A" and "B") of a tangent plane T_pS. In the right graphic, two unit normals (in blue and purple) extending from T_pS are added; the blue normal represents side "A", while the purple normal represents side "B".

Finally, we relate unit normals to a basic property of planes—that they are "two-sided". Intuitively, we can think of a plane as a flat (infinite) piece of paper. Just as the paper has a front and a back side, an abstract plane also has two faces.

Now, for a surface $S \subseteq \mathbb{R}^3$, we can think of a tangent plane T_pS as splitting $T_p\mathbb{R}^3$ into two halves—with one side of T_pS facing one of the halves of $T_p\mathbb{R}^3$, and the other side of T_pS facing the other half; see the left part of Figure 4.18. One precise way to capture these "sides" of T_pS is through unit normals. In particular, *we simply associate a unit normal* $\mathbf{n}_p$ *with the side of* T_pS *facing the direction that* $\mathbf{n}_p$ *is pointing.* This is illustrated in the right part of Figure 4.18.

Remark 4.49. In fact, even for surfaces in higher dimensions ($S \subseteq \mathbb{R}^n$), one can still make sense of their tangent planes being "two-sided". However, this property is trickier to capture mathematically, hence we avoid discussing this here.

4.7. **Orientation.** Consider again a surface $S \subseteq \mathbb{R}^3$ and a point $\mathbf{p} \in S$. We already described how a unit normal to S at $\mathbf{p}$ can be viewed as selecting a side of $\mathsf{T}_\mathbf{p}S$. Now, since $\mathsf{T}_\mathbf{p}S$ can be seen as a plane extending from $\mathbf{p}$, *we can also think of a unit normal to* S *at* $\mathbf{p}$ *as choosing a side of* S *at* $\mathbf{p}$. In other words, a surface is two-sided at each point, and one side can be selected by choosing a unit normal at that point.

Example 4.50. Recall from our computations in Example 4.48 that the unit normals to the *sphere* $\mathbb{S}^2$ at any of its points $\mathbf{p} \in \mathbb{S}^2$ are given by $\pm\mathbf{p}_\mathbf{p}$.

- Observe that $+\mathbf{p}_\mathbf{p}$ points outward from the sphere; see the blue arrows in the left plot of Figure 4.19. Thus, $+\mathbf{p}_\mathbf{p}$ captures the "outward-facing" side of $\mathbb{S}^2$ at $\mathbf{p}$.
- Similarly, $-\mathbf{p}_\mathbf{p}$ points inward from the sphere; see the purple arrows in the right plot of Figure 4.19. Thus, $-\mathbf{p}_\mathbf{p}$ captures the "inward-facing" side of $\mathbb{S}^2$ at $\mathbf{p}$.

Now, the above only concerns the *local* geometry of S, in that it deals with a single point of S. Next, we pose a similar question involving the *global* geometry of S.

Question 4.51. Is S, as a whole, two-sided?

The first step is to make Question 4.51 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 choosing a unit normal at all points of S. However, this is not quite enough, as one must also ensure consistency—given any two points $\mathbf{p}, \mathbf{q} \in S$, the unit normals of S chosen at $\mathbf{p}$ and $\mathbf{q}$ represent the "same side" of S. These considerations now motivate the following:

Definition 4.52. An orientation of a surface $S \subseteq \mathbb{R}^3$ is a choice of a unit normal $\mathbf{n}_\mathbf{p}$ of S at every $\mathbf{p} \in S$ such that the $\mathbf{n}_\mathbf{p}$'s vary continuously with respect to $\mathbf{p}$.

Furthermore, we say that S is orientable iff such an orientation of S exists.

In particular, the assumption in Definition 4.52 that the normals $\mathbf{n}_\mathbf{p}$ vary continuously with respect to $\mathbf{p}$ prevents us from suddenly "jumping" from one side of S to the other as one moves to nearby points. Thus, when S is orientable, we are able to make a consistent choice of a side of S everywhere.

In other words, we can interpret S *being orientable as* S *being globally two-sided*. Then, *an orientation of* S *corresponds to a choice of one of the two sides of* S.

Example 4.53. Observe that the *sphere* $\mathbb{S}^2$ is orientable. To see this, we recall, from Example 4.48, that the unit normals of $\mathbb{S}^2$ at each $\mathbf{p} \in \mathbb{S}^2$ are given by $\pm\mathbf{p}_\mathbf{p}$. Then, one continuous choice of unit normals on $\mathbb{S}^2$ is to simply associate each $\mathbf{p} \in \mathbb{S}^2$ with $+\mathbf{p}_\mathbf{p}$.

- Since the $+\mathbf{p}_p$'s point outward from the sphere (see the left plot in Figure 4.19), the orientation given by the $+\mathbf{p}_p$'s represents the outward-facing side of $\mathbb{S}^2$.
- Had we chosen the $-\mathbf{p}_p$'s instead (see the right plot in Figure 4.19), then we would have captured the opposite, inward-facing orientation of $\mathbb{S}^2$.

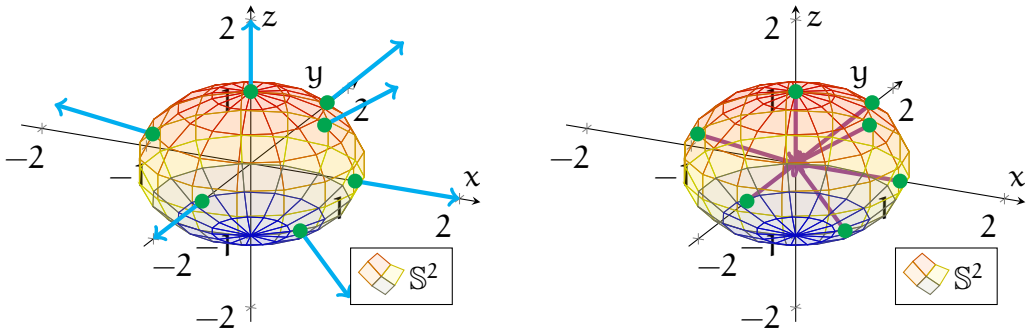


FIGURE 4.19. The left illustration shows $\mathbb{S}^2$ with the orientation representing its outward-facing side; some of the associated unit normals are drawn in blue. The right graphic shows the opposite inward-facing orientation of $\mathbb{S}^2$, with some associated unit normals drawn in purple.

Next, any parametrisation of a surface S naturally yields a choice of unit normals to S. More specifically, along the image of a parametrisation σ of S, one can just choose the candidate in equation (4.8) with the positive sign:

$$\mathbf{n}_\sigma(u,v) = +\left[\frac{\partial_1\sigma(u,v)\times\partial_2\sigma(u,v)}{|\partial_1\sigma(u,v)\times\partial_2\sigma(u,v)|}\right]_{\sigma(u,v)}. \tag{4.10}$$

Thus, when S is orientable, a parametrisation of S naturally selects an orientation of S. This is captured in the following analogue to Definition 3.52 for curves:

Definition 4.54. Let $S \subseteq \mathbb{R}^3$ be an orientable surface, and let O be an orientation of S.

- A parametrisation $\sigma : U \to S$ of S generates the orientation O iff the unit normals $\mathbf{n}_\sigma(u,v)$, defined as in (4.10), coincide with O for all $(u,v) \in U$.
- A parametrisation $\sigma : U \to S$ of S generates an orientation opposite to O iff $\mathbf{n}_\sigma(u,v)$, as defined in (4.10), does not coincide with O for any $(u,v) \in U$.

Example 4.55. Let us return again to the *sphere* $\mathbb{S}^2$. Consider the parametrisation

$$\rho_{s,z} : \mathbb{R}\times(0,\pi) \to \mathbb{S}^2, \qquad \rho_{s,z}(u,v) = (\cos u \sin v, \sin u \sin v, \cos v),$$

introduced in Example 4.33. Also, recall from Example 4.33 that

$$\partial_1\rho_{s,z}(u,v) = (-\sin u \sin v, \cos u \sin v, 0),$$

$$\partial_2 \rho_{s,z}(u,v) = (\cos u \cos v, \sin u \cos v, -\sin v).$$

Next, taking a cross product of the above vectors, we see that

$$\begin{aligned}\partial_1 \rho_{s,z}(u,v) \times \partial_2 \rho_{s,z}(u,v) &= (-\cos u \sin^2 v, -\sin u \sin^2 v, -\cos v \sin v)\\ &= -\sin v \cdot \rho_{s,z}(u,v),\\ |\partial_1 \rho_{s,z}(u,v) \times \partial_2 \rho_{s,z}(u,v)| &= \sin v.\end{aligned}$$

In particular, the formula (4.10), with $\rho_{s,z}$ in the place of σ, selects the unit normals

$$\begin{aligned}\mathbf{n}_{\rho_{s,z}}(u,v) &= +\left[\frac{\partial_1 \rho_{s,z}(u,v) \times \partial_2 \rho_{s,z}(u,v)}{|\partial_1 \rho_{s,z}(u,v) \times \partial_2 \rho_{s,z}(u,v)|}\right]_{\rho_{s,z}(u,v)}\\ &= -\rho_{s,z}(u,v)_{\rho_{s,z}(u,v)},\end{aligned}$$

for any $(u,v) \in \mathbb{R} \times (0,\pi)$. More explicitly, for any $\mathbf{p} = \rho_{s,z}(u,v)$, the parametrisation $\rho_{s,z}$ selects the inward-pointing unit normal $-\mathbf{p}_\mathbf{p}$. Therefore, according to Definition 4.54, the parametrisation $\rho_{s,z}$ *generates the inward-facing orientation of* $\mathbb{S}^2$.

In contrast, consider instead the parametrisation

$$\tilde{\rho}_{s,z} : (0,\pi) \times \mathbb{R} \to \mathbb{S}^2, \qquad \tilde{\rho}_{s,z}(u,v) = (\cos v \sin u, \sin v \sin u, \cos u),$$

that is, $\rho_{s,z}$ with the roles of u and v interchanged. Similar computations then yield

$$\begin{aligned}\mathbf{n}_{\tilde{\rho}_{s,z}}(u,v) &= +\left[\frac{\partial_1 \tilde{\rho}_{s,z}(u,v) \times \partial_2 \tilde{\rho}_{s,z}(u,v)}{|\partial_1 \tilde{\rho}_{s,z}(u,v) \times \partial_2 \tilde{\rho}_{s,z}(u,v)|}\right]_{\tilde{\rho}_{s,z}(u,v)}\\ &= +\tilde{\rho}_{s,z}(u,v)_{\tilde{\rho}_{s,z}(u,v)},\end{aligned}$$

hence $\tilde{\rho}_{s,z}$ *generates the outward-facing orientation of* $\mathbb{S}^2$.

The following definition formalises the idea of a "surface with a chosen side":

Definition 4.56. An oriented surface is a surface S along with a chosen orientation of S.

Finally, we look at an example of a *non-orientable* surface: the *Möbius strip*. One simple way to construct this Möbius strip is to take a rectangle and glue two of its opposite edges together "in reverse order"; see Figure 4.20. To try this yourself, take a thin strip of paper, twist it halfway, and then tape the two ends together.

Intuitively, the Möbius strip is not orientable since it has only one side, rather than two. 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.

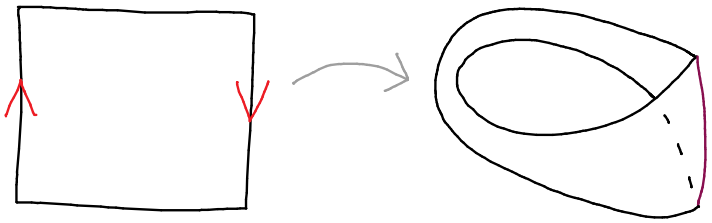


FIGURE 4.20. These drawings demonstrate how the Möbius strip can be constructed by gluing together two opposite edges of a rectangle.

Example 4.57. Consider the parametric surface $\sigma : (-1, 1) \times \mathbb{R} \to \mathbb{R}^3$ defined by

$$\sigma(u, v) = \left(\left(1 - \frac{u}{2}\sin\frac{v}{2}\right)\cos v, \left(1 - \frac{u}{2}\sin\frac{v}{2}\right)\sin v, \frac{u}{2}\cos\frac{v}{2}\right),$$

whose image is plotted in the left side of Figure 4.21. The *Möbius strip* can then be more concretely described as the image M of σ. (With some effort, one can in fact show that M satisfies the hypotheses of Definition 4.20.)

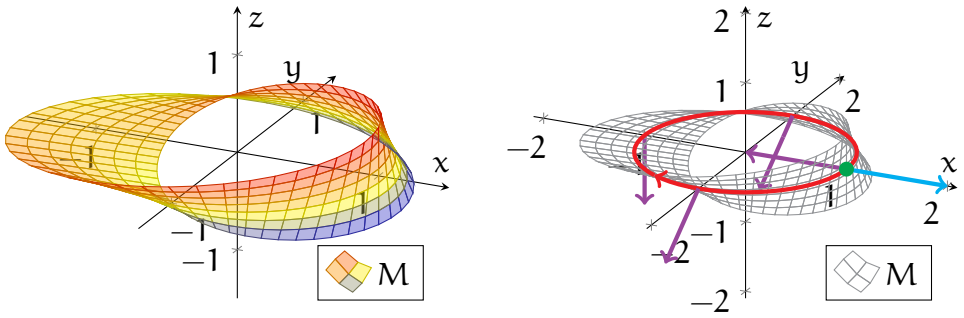
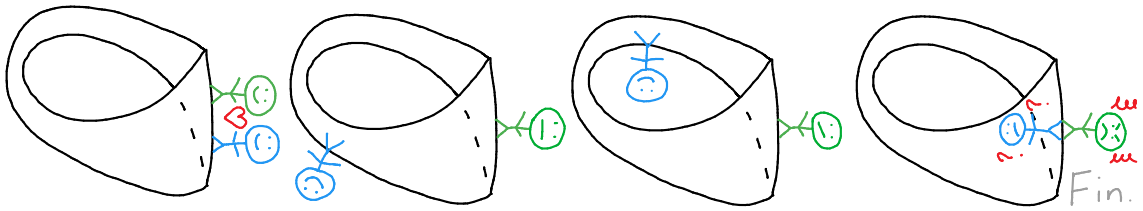


FIGURE 4.21. The left plot contains the image M of the parametric surface σ from Example 4.57, representing the Möbius strip. The right plot demonstrates why M fails to be orientable—any choice of unit normals along the strip must fail to vary continuously at some point.

We can also give an informal argument for why the Möbius strip fails to be orientable. Consider the strip M in the right plot of Figure 4.21, and consider the blue unit normal $\mathbf{n}_p$ at the point $\mathbf{p} \in M$ (in green). Let us now try to construct from $\mathbf{n}_p$ a continuous choice of unit normals along all of M.

As we traverse (anticlockwise) along the red path on M, we see that our hand is forced. The unit normals are determined by the requirement that they vary continuously along this red path; some of these unit normals are drawn in purple in the right side of Figure 4.21. The problem arises once we travel a full lap around M along the red path. 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. This shows that our attempt to construct a continuously varying global choice of unit normals must inevitably fail.



Remark 4.58. The above argument can be made into a formal proof using the parametrisation σ from Example 4.57 and the unit normals constructed using Theorem 4.44.

Remark 4.59. Note that Definition 4.52 is quite similar to Definition 3.50 for the orientation of curves. However, one key difference is that *one can always choose an orientation for a curve, while not every surface is orientable*.

Remark 4.60. One can also make sense of orientability and orientation for general surfaces $S \subseteq \mathbb{R}^n$. However, this is trickier to capture mathematically, so we do not go into this point here. For example, the Klein bottle (see Figure 4.13) fails to be orientable.

4.8. **Parametric Integration.** We now turn our attention toward studying the "sizes" of surfaces, in effect moving from differential to integral properties. Our discussion begins with a basic and familiar question:

Question 4.61. How do we define and compute the area of a surface?

Recall that to find the arc length of a curve, we approximated it as line segments joining chosen sample points along the curve; measuring the lengths of these line segments yielded an approximation of the arc length. The exact length was then obtained by taking a "limit" as the number of sample points tended to infinity. A graphical demonstration of this was given in Figure 3.18.

The idea is similar in the context of surfaces. We can approximate the area of a surface S by "tiling" it with parallelograms and by calculating the total area of these parallelograms. Finally, taking a "limit" as the number of such parallelograms tends toward infinity results in a formula for area of S.

To be more specific, let us consider a regular parametric surface $\sigma : U \to \mathbb{R}^3$, sitting in 3-dimensional space. To obtain a tiling of σ, we partition its domain U into a rectangular grid, with each rectangle having length Δu and height Δv. The function σ maps each of these rectangles to a "curved rectangle"; see Figure 4.22.

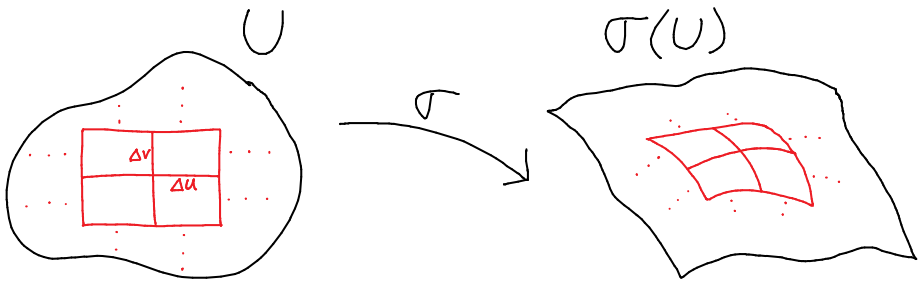


FIGURE 4.22. The drawing shows how the image of a parametric surface σ can be tiled. The domain U of σ is partitioned into rectangles, which are then mapped onto curved rectangles in the image of σ.

To measure the total area of σ, we must find and then add up the areas of all the "curved rectangles" $\mathcal{R}$ comprising the image of σ:

$$\mathcal{A}(\sigma) = \sum_{\text{Curved rectangles } \mathcal{R}} \mathcal{A}(\mathcal{R}).$$

Although we cannot directly compute the area of such a "curved rectangle" $\mathcal{R}$, we can approximate it if we replace $\mathcal{R}$ with a parallelogram with straight edges.

In particular, the four corners of $\mathcal{R}$ are given by points

$$\sigma(u, v), \qquad \sigma(u + \Delta u, v), \qquad \sigma(u, v + \Delta v), \qquad \sigma(u + \Delta u, v + \Delta v).$$

Thus, we can approximate $\mathcal{R}$ by the parallelogram $\mathcal{P}_{\mathcal{R}}$ with two sides given by

$$\begin{aligned} \mathbf{a}_{\sigma(u,v)} &= [\sigma(u + \Delta u, v) - \sigma(u, v)]_{\sigma(u,v)} \in \mathsf{T}_{\sigma(u,v)}\mathbb{R}^3, \\ \mathbf{b}_{\sigma(u,v)} &= [\sigma(u, v + \Delta v) - \sigma(u, v)]_{\sigma(u,v)} \in \mathsf{T}_{\sigma(u,v)}\mathbb{R}^3. \end{aligned} \tag{4.11}$$

This process is shown in the left drawing of Figure 4.23.

Claim 4.62. The area of the parallelogram $\mathcal{P}_{\mathcal{R}}$ defined above is

$$\mathcal{A}(\mathcal{P}_{\mathcal{R}}) = |\mathbf{a} \times \mathbf{b}|.$$

Proof. Let θ be the angle between $\mathbf{a}_{\sigma(u,v)}$ and $\mathbf{b}_{\sigma(u,v)}$. If we take $\mathbf{a}_{\sigma(u,v_0)}$ to represent the base of $\mathcal{P}_{\mathcal{R}}$, then the height h of $\mathcal{P}_{\mathcal{R}}$ is the length of the purple line segment in the right drawing of Figure 4.23; a bit of trigonometry shows that $h = |\mathbf{b}| \sin\theta$.

Thus, by Definition 2.15, Theorem 2.18, and the above, we obtain

$$\begin{aligned} \mathcal{A}(\mathcal{P}_{\mathcal{R}}) &= |\mathbf{a}| \cdot |\mathbf{b}| \sin\theta \\ &= |\mathbf{a} \times \mathbf{b}|. \end{aligned}$$

□

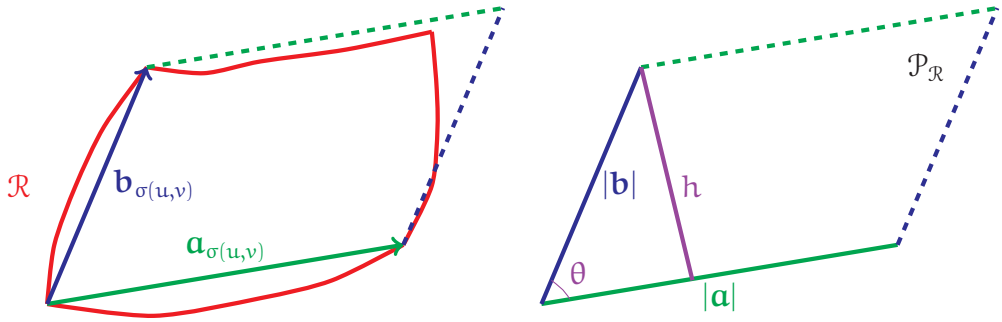


FIGURE 4.23. The left graphic is a crude rendering of how a curved rectangle $\mathcal{R}$ is approximated by a parallelogram $\mathcal{P}_{\mathcal{R}}$. The boundary of $\mathcal{R}$ is drawn in red, while the tangent vectors $\mathbf{a}_{\sigma(u,v)}$ and $\mathbf{b}_{\sigma(u,v)}$ defining $\mathcal{P}_{\mathcal{R}}$ are drawn in blue and green, respectively. The right graphic describes the setup for computing the area of $\mathcal{P}_{\mathcal{R}}$ in the proof of Claim 4.62.

Summing the areas of all the $\mathcal{P}_{\mathcal{R}}$'s and applying Claim 4.62 then yields

$$\begin{aligned} \mathcal{A}(\sigma) &\approx \sum_{\text{Curved rectangles } \mathcal{R}} \mathcal{A}(\mathcal{P}_{\mathcal{R}}) \\ &= \sum_{\mathcal{R}} \Delta u \Delta v \left| \frac{\mathbf{a}}{\Delta u} \times \frac{\mathbf{b}}{\Delta v} \right|. \end{aligned} \tag{4.12}$$

The above approximation can be refined by taking a smaller rectangular grid in U, i.e. by decreasing Δu and Δv. However, for the actual area of σ, we need an "infinitely good" approximation, that is, we take $\Delta u, \Delta v \searrow 0$. Observe that by (4.11),

$$\lim_{\Delta u \searrow 0} \frac{\mathbf{a}}{\Delta u} = \partial_1 \sigma(u_0, v_0), \qquad \lim_{\Delta v \searrow 0} \frac{\mathbf{b}}{\Delta v} = \partial_2 \sigma(u_0, v_0).$$

Also, in this limit, the summation in (4.12) becomes an integral over U.

Combining all the above, we (at least informally) see that

$$\begin{aligned} \mathcal{A}(\mathcal{R}) &= \text{“} \lim_{\Delta u, \Delta v \searrow 0} \text{”} \sum_{\mathcal{R}} \left| \frac{\mathbf{a}}{\Delta u} \times \frac{\mathbf{b}}{\Delta v} \right| \Delta u \Delta v \\ &= \iint_U |\partial_1 \sigma(u, v) \times \partial_2 \sigma(u, v)| \, du dv. \end{aligned}$$

(Again, we avoid discussing why the above limit holds formally, as this lies beyond the scope of this module.) Thus, our derivation leads us to the following:

Definition 4.63. For a parametric surface $\sigma : U \to \mathbb{R}^3$, we define its surface area as

$$\mathcal{A}(\sigma) = \iint_U |\partial_1 \sigma(u, v) \times \partial_2 \sigma(u, v)| \, du dv. \tag{4.13}$$

Example 4.64. Fix $h > 0$, and consider the parametric surface

$$\sigma_h : (0,\pi) \times (0,h) \to \mathbb{R}^3, \qquad \sigma_h(u,v) = (\cos u, \sin u, v).$$

Observe σ_h maps out a *half-cylinder* of height h; see the left part of Figure 4.24.

Moreover, the same computations as in Example 4.21 yield the following:

$$\begin{aligned} \partial_1\sigma_h(u,v) &= (-\sin u, \cos u, 0), \\ \partial_2\sigma_h(u,v) &= (0,0,1), \\ |\partial_1\sigma_h(u,v) \times \partial_2\sigma_h(u,v)| &= 1. \end{aligned}$$

Applying Definition 4.63 and Fubini's theorem results in the area of σ_h:

$$\begin{aligned} \mathcal{A}(\sigma_h) &= \iint_{(0,\pi)\times(0,h)} |\partial_1\sigma_h(u,v) \times \partial_2\sigma_h(u,v)| \, du dv \\ &= \iint_{(0,\pi)\times(0,h)} 1 \, du dv \\ &= \int_0^\pi du \int_0^h dv \\ &= \pi h. \end{aligned}$$

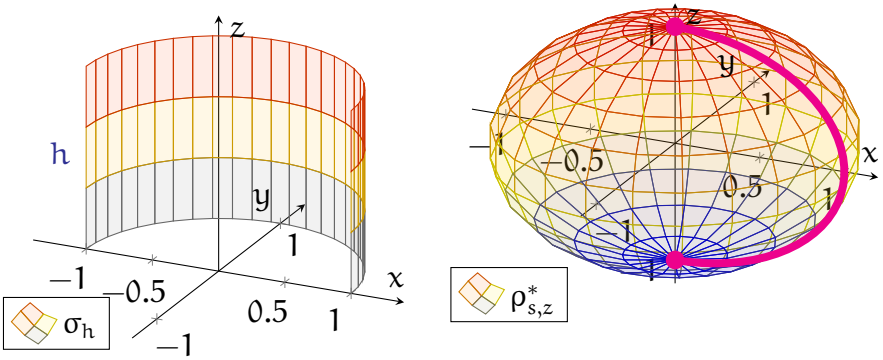


FIGURE 4.24. The left plot shows the image of σ_h from Example 4.64, while the right plot shows the image of $\rho^*_{s,z}$ from Example 4.65.

Example 4.65. Next, consider the following parametric surface:

$$\rho^*_{s,z} : (0,2\pi) \times (0,\pi) \to \mathbb{R}^3, \qquad \rho^*_{s,z}(u,v) = (\cos u \sin v, \sin u \sin v, \cos v).$$

(Note this is identical to $\rho_{s,z}$ from Example 4.26, except we now restrict the u-value to $(0,2\pi)$.) The image of $\rho^*_{s,z}$ is shown in the right half of Figure 4.24. In particular, $\rho^*_{s,z}$ maps out all of the unit sphere $\mathbb{S}^2$ except for a closed arc, drawn in pink.

Now, from computations similar to those in Example 4.26, we see that

$$|\partial_1\rho^*_{s,z}(u,v)\times\partial_2\rho^*_{s,z}(u,v)| = \sin v, \qquad (u,v)\in(0,2\pi)\times(0,\pi).$$

Thus, by Definition 4.63 and Fubini's theorem, the area of $\rho^*_{s,z}$ is

$$\begin{aligned}\mathcal{A}(\rho^*_{s,z}) &= \iint_{(0,2\pi)\times(0,\pi)} |\partial_1\rho^*_{s,z}(u,v)\times\partial_2\rho^*_{s,z}(u,v)|\,du dv\\ &= \int_0^{2\pi} du\int_0^{\pi}\sin v\,dv\\ &= 4\pi.\end{aligned}$$

Since the image of $\rho^*_{s,z}$ differs from the sphere $\mathbb{S}^2$ only by a 1-dimensional arc (which one thinks should have negligible area), this suggests (as expected) that the *surface area of $\mathbb{S}^2$ itself should also be 4π*. We will revisit this point more formally later on.

Remark 4.66. With a bit of algebra, one can show the following holds for any $\mathbf{a},\mathbf{b}\in\mathbb{R}^3$:

$$|\mathbf{a}\times\mathbf{b}|^2 = \left|\det\begin{bmatrix}\mathbf{a}\cdot\mathbf{a} & \mathbf{a}\cdot\mathbf{b}\\ \mathbf{b}\cdot\mathbf{a} & \mathbf{b}\cdot\mathbf{b}\end{bmatrix}\right|$$

Thus, assuming the setting of Definition 4.63, and setting $\mathbf{a}$ and $\mathbf{b}$ to be the values of $\partial_1\sigma$ and $\partial_2\sigma$, respectively, we obtain an alternative formula for surface area:

$$\mathcal{A}(\sigma) = \iint_U \sqrt{\mathcal{F}(u,v)}\,du dv, \qquad \mathcal{F} = \left|\det\begin{bmatrix}\partial_1\sigma\cdot\partial_1\sigma & \partial_1\sigma\cdot\partial_2\sigma\\ \partial_2\sigma\cdot\partial_1\sigma & \partial_2\sigma\cdot\partial_2\sigma\end{bmatrix}\right|. \tag{4.14}$$

While (4.14) is no simpler than (4.13), it does have the advantage that *the quantity $\mathcal{F}(u,v)$ makes sense when σ takes values in any $\mathbb{R}^n$*, and not only $\mathbb{R}^3$. Thus, (4.14) *gives a sensible definition of surface area for parametric surfaces in all dimensions*.

4.9. **Surface Integrals.** Having defined the area of parametric surfaces, the next aim is to extend our understanding to "weighted" areas. Similar to our discussions for curves, we will achieve this by constructing a theory of surface integration.

Consider the area formula (4.13) for a parametric surface σ. There, the integrand

$$|\partial_1\sigma(u,v)\times\partial_2\sigma(u,v)| \tag{4.15}$$

represents the area of an "infinitesimal" parallelogram *at the point* $\sigma(u,v)$. Thus, if we wish to add a weight F to the area (4.13), then the factor (4.15) should be paired with F, *applied to the same point* $\sigma(u,v)$. These considerations lead to the following:

Definition 4.67. Let $\sigma : U \to \mathbb{R}^3$ be a parametric surface, and let F be a real-valued function defined on the image of σ. We define the surface integral of F over σ by

$$\iint_\sigma F \, dA = \iint_U F(\sigma(u,v))|\partial_1\sigma(u,v) \times \partial_2\sigma(u,v)| \, du dv. \tag{4.16}$$

Example 4.68. Consider the *half-cylinder* from Example 4.64, but with $h = 1$:

$$\sigma : (0,\pi) \times (0,1) \to \mathbb{R}^3, \qquad \sigma(u,v) = (\cos u, \sin u, v).$$

(See the left part of Figure 4.25.) Let us integrate over σ the function

$$G : \mathbb{R}^3 \to \mathbb{R}, \qquad G(x,y,z) = x + y + z.$$

First, recall from the computations in Example 4.64 that

$$|\partial_1\sigma(u,v) \times \partial_2\sigma(u,v)| = 1, \qquad (u,v) \in (0,\pi) \times (0,1).$$

In addition, observe that for the same values of (u,v),

$$\begin{aligned} G(\sigma(u,v)) &= G(\cos u, \sin u, v) \\ &= \cos u + \sin u + v. \end{aligned}$$

Therefore, applying Definition 4.67 and Fubini's theorem, we conclude that

$$\begin{aligned} \iint_\sigma G \, dA &= \int_{(0,\pi)\times(0,1)} G(\sigma(u,v))|\partial_1\sigma(u,v) \times \partial_2\sigma(u,v)| \, du dv \\ &= \int_0^1 \left[\int_0^\pi (\cos u + \sin u + v) \cdot 1 \cdot du \right] dv \\ &= \int_0^1 (0 + 2 + \pi v) \, dv \\ &= 2 + \frac{\pi}{2}. \end{aligned}$$

Example 4.69. Consider the following portion of a parametric *paraboloid*:

$$\mathcal{P} : (0,1) \times (0,1) \to \mathbb{R}^3, \qquad \mathcal{P}(u,v) = (u, v, u^2 + v^2).$$

(See the right side of Figure 4.25.) Observe that a direct computation yields

$$\begin{aligned} |\partial_1\mathcal{P}(u,v) \times \partial_2\mathcal{P}(u,v)| &= |(1,0,2u) \times (0,1,2v)| \\ &= \sqrt{1 + 4u^2 + 4v^2}, \end{aligned}$$

for any parameters $(u,v) \in (0,1) \times (0,1)$.

Let us now integrate the function

$$H : \mathbb{R}^3 \to \mathbb{R}, \qquad H(x, y, z) = \sqrt{1 + 2x^2 + 2y^2 + 2z}$$

over $\mathcal{P}$. To accomplish this, we apply Definition 4.67 and expand:

$$\begin{aligned}\iint_{\mathcal{P}} H \, dA &= \iint_{(0,1)\times(0,1)} H(\mathcal{P}(u,v)) |\partial_1 \mathcal{P}(u,v) \times \partial_2 \mathcal{P}(u,v)| \, du dv \\ &= \iint_{(0,1)\times(0,1)} \sqrt{1 + 2u^2 + 2v^2 + 2(u^2+v^2)} \sqrt{1 + 4u^2 + 4v^2} \, du dv \\ &= \iint_{(0,1)\times(0,1)} (1 + 4u^2 + 4v^2) \, dv du.\end{aligned}$$

Using Fubini's theorem, the above can be evaluated directly:

$$\begin{aligned}\iint_{\mathcal{P}} H \, dA &= \int_0^1 \int_0^1 1 \, du dv + 4 \int_0^1 \int_0^1 u^2 \, du dv + 4 \int_0^1 \int_0^1 v^2 \, du dv \\ &= 1 + 4 \int_0^1 u^2 du + 4 \int_0^1 v^2 dv \\ &= \frac{11}{3}.\end{aligned}$$

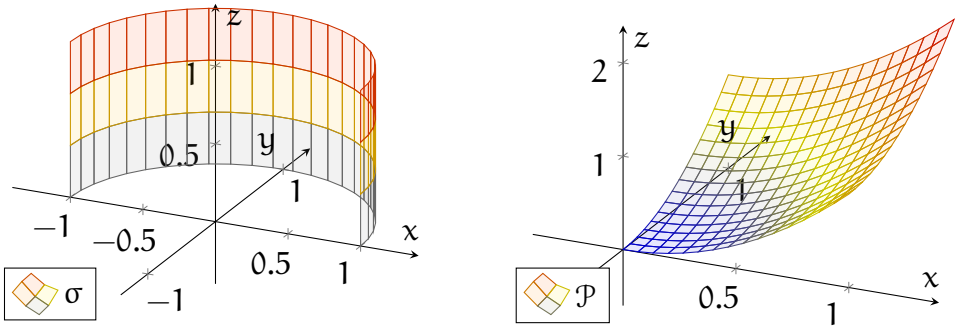


FIGURE 4.25. The left plot is the image of σ from Example 4.68, while the right plot is the image of $\mathcal{P}$ from Example 4.69.

The following theorem, a direct analogue of Theorem 3.67 for curve integrals, affirms that the integrals of Definition 4.67 are *independent of parametrisation*:

Theorem 4.70. Let $\sigma : U \to \mathbb{R}^n$ and $\tilde{\sigma} : \tilde{U} \to \mathbb{R}^n$ be regular parametric surfaces, and suppose σ and $\tilde{\sigma}$ are reparametrisations of each other (see Definition 4.34). Then, given any real-valued function F defined on the images of σ and $\tilde{\sigma}$, we have that

$$\iint_{\sigma} F \, dA = \iint_{\tilde{\sigma}} F \, dA. \tag{4.17}$$

In particular, σ and $\tilde{\sigma}$ have the same surface area: $\mathcal{A}(\sigma) = \mathcal{A}(\tilde{\sigma})$.

Proof. We give a brief sketch of the proof here, though we omit many computational details. Let Φ be the change of variables for σ and $\tilde{\sigma}$, for which the following holds:

$$\sigma(u, v) = \tilde{\sigma}(\Phi(u, v)), \qquad (u, v) \in U.$$

The key step is to substitute $(\tilde{u}, \tilde{v}) = \Phi(u, v)$ into the right-hand side of (4.16):

$$\begin{aligned}\iint_\sigma F \, dA &= \iint_U F(\tilde{\sigma}(\Phi(u, v)))|\partial_1\sigma(u, v) \times \partial_2\sigma(u, v)| \, du dv \\ &= \iint_{\tilde{U}} F(\tilde{\sigma}(\tilde{u}, \tilde{v})) \, |\partial_1\sigma(\Phi^{-1}(\tilde{u}, \tilde{v})) \times \partial_2\sigma(\Phi^{-1}(\tilde{u}, \tilde{v}))| \, |\mathcal{J}_\Phi(\tilde{u}, \tilde{v})| \, d\tilde{u} d\tilde{v}.\end{aligned}$$

Here, $\mathcal{J}_\Phi$ denotes the Jacobian determinant associated with this change of variables.

The main observation—also the main computation of the proof—is the identity

$$|\partial_1\sigma(\Phi^{-1}(\tilde{u}, \tilde{v})) \times \partial_2\sigma(\Phi^{-1}(\tilde{u}, \tilde{v}))| \, |\mathcal{J}_\Phi(\tilde{u}, \tilde{v})| = |\partial_1\tilde{\sigma}(\tilde{u}, \tilde{v}) \times \partial_2\tilde{\sigma}(\tilde{u}, \tilde{v})|. \tag{4.18}$$

(In particular, slightly annoying calculuations using the chain rule yield that

$$|\partial_1\sigma(\Phi^{-1}(\tilde{u}, \tilde{v})) \times \partial_2\sigma(\Phi^{-1}(\tilde{u}, \tilde{v}))|, \qquad |\partial_1\tilde{\sigma}(\tilde{u}, \tilde{v}) \times \partial_2\tilde{\sigma}(\tilde{u}, \tilde{v})|$$

differ by a factor that is the determinant of a matrix. In addition, this is precisely the matrix found in the Jacobian $\mathcal{J}_\Phi$, and hence the identity (4.18) follows.)

Combining all the above and recalling again (4.16), we conclude that

$$\begin{aligned}\iint_\sigma F \, dA &= \iint_{\tilde{U}} F(\tilde{\sigma}(\tilde{u}, \tilde{v}))|\partial_1\tilde{\sigma}(\tilde{u}, \tilde{v}) \times \partial_2\tilde{\sigma}(\tilde{u}, \tilde{v})| \, d\tilde{u} d\tilde{v} \\ &= \iint_{\tilde{\sigma}} F \, dA,\end{aligned}$$

which is (4.17). Finally, taking $F \equiv 1$ in (4.17) yields $\mathcal{A}(\sigma) = \mathcal{A}(\tilde{\sigma})$. □

In particular, we can use Definitions 4.63 and 4.67 to make sense of *areas of surfaces* and *integrals over surfaces*, i.e. both are geometric properties of surfaces.

Definition 4.71. Let $S \subseteq \mathbb{R}^3$ be a surface, and let $\sigma : U \to S$ be an injective parametrisation of S, whose image differs from S by only a finite number of points and curves.

- For any real-valued function F on S, we define its surface integral over S by

$$\iint_S F \, dA = \iint_\sigma F \, dA. \tag{4.19}$$

- Moreover, we define the surface area of S by $\mathcal{A}(S) = \mathcal{A}(\sigma)$.

The assumptions found in Definition 4.71 mirror those found in Definition 3.68 for curve integrals. In particular, we assume σ is injective in order to avoid counting any points of the underlying surface S more than once in our integrals.

Moreover, one does not expect isolated (0-dimensional) points and (1-dimensional) curves within a surface to contribute to the total surface area. (This can be proved rigorously, but the details lie beyond this module.) This motivates the condition in Definition 4.71 that S can differ from the image of σ by these "negligible" sets.

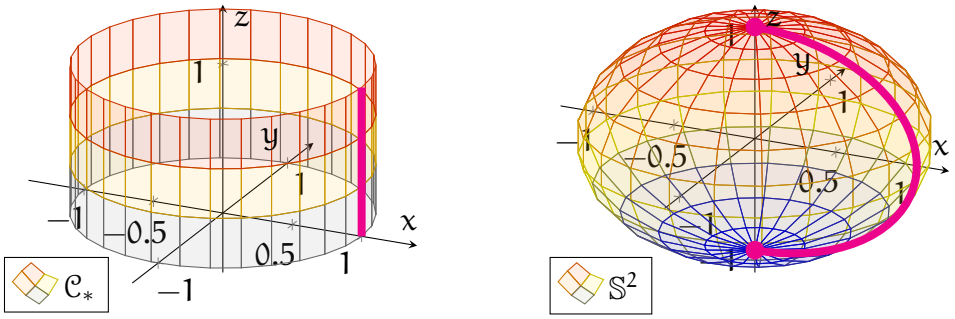


FIGURE 4.26. The left plot shows the cylindrical segment $\mathcal{C}_*$ from Example 4.72; the image of the parametrisation σ_* is all of $\mathcal{C}_*$ except for the pink line segment. The right plot shows the sphere $\mathbb{S}^2$ from Example 4.73; the image of $\rho^*_{s,z}$ is all of $\mathbb{S}^2$ except for the points marked in pink.

Example 4.72. Let $\mathcal{C}_*$ denote the following finite segment of the cylinder from Example 4.21 (for an illustration, see the left plot of Figure 4.26):

$$\mathcal{C}_* = \{(x, y, z) \in \mathbb{R}^3 \mid x^2 + y^2 = 1,\ 0 < z < 1\}.$$

Let us now integrate the following function over $\mathcal{C}_*$:

$$G : \mathbb{R}^3 \to \mathbb{R}, \qquad G(x, y, z) = x + y + z.$$

The first step is to parametrise $\mathcal{C}_*$ in a manner that is consistent with Definition 4.71. As inspiration, we can look at the injective parametrisations of the full cylinder used in Example 4.21. In fact, we need only restrict one of these to $0 < z < 1$:

$$\sigma_* : (0, 2\pi) \times (0, 1) \to \mathcal{C}_*, \qquad \sigma_*(u, v) = (\cos u, \sin u, v).$$

Note its image is all of $\mathcal{C}_*$ except for one line segment; see the left half of Figure 4.26.

Thus, σ_* satisfies the conditions in Definition 4.71, and hence

$$\iint_{\mathcal{C}_*} G \, dA = \iint_{\sigma_*} G \, dA$$

$$= \iint_{(0,2\pi)\times(0,1)} G(\sigma_*(u,v))|\partial_1\sigma_*(u,v) \times \partial_2\sigma_*(u,v)|\,du\,dv.$$

The right-hand side can now be computed in the same manner as in Example 4.68:

$$\begin{aligned}\iint_{e_*} G\,dA &= \int_0^1 \left[\int_0^{2\pi} (\cos u + \sin u + v)1\,du\right] dv \\ &= 2\pi \int_0^1 v\,dv \\ &= \pi.\end{aligned}$$

Example 4.73. Next, consider the *sphere* $\mathbb{S}^2$, and recall the parametrisation

$$\rho^*_{s,z} : (0,2\pi) \times (0,\pi) \to \mathbb{S}^2, \qquad \rho^*_{s,z}(u,v) = (\cos u \sin v, \sin u \sin v, \cos v)$$

of $\mathbb{S}^2$ from Example 4.65. Note the domain restrictions $u \in (0,2\pi)$ and $v \in (0,\pi)$ imply that $\rho^*_{s,z}$ is injective. Moreover, the image of $\rho^*_{s,z}$ is all of $\mathbb{S}^2$, except for the poles $(0,0,\pm1)$ and a semicircle—see the pink arc in the right plot of Figure 4.26.

As a result, the parametrisation $\rho^*_{s,z}$ satisfies the assumptions of Definition 4.71. Furthermore, recall from Example 4.65 that the area of $\rho^*_{s,z}$ is 4π. Thus, by Definition 4.71, we obtain the standard formula for the surface area of a unit sphere:

$$\mathcal{A}(\mathbb{S}^2) = \mathcal{A}(\rho^*_{s,z}) = 4\pi.$$

4.10. **Integration of Vector Fields.** In this final section, we discuss a different notion of surface integration, in which *the integrand is a vector field*.

The main motivation for this comes from the notion of *flux* in physics, which roughly quantifies a "rate of flow through a surface". The term "flow" could refer to many different phenomena, including the motion of a fluid or an electromagnetic force. Mathematically, such a flow is expressed using a vector field $\mathbf{F}$. In the case of fluids, the arrow $\mathbf{F}(\mathbf{p})$ represents the velocity of the fluid at the position $\mathbf{p}$.

The flux of $\mathbf{F}$ *through a surface* S *should measure "the total amount that the arrows of* $\mathbf{F}$ *pass through* S*"*. More specifically, we must keep in mind the following:

- If the values of $\mathbf{F}$ lie tangent to S (that is, the fluid flows along but not through S), then one should measure zero flux. This is shown in the left drawing of Figure 4.27. Thus, *only the component of* $\mathbf{F}$ *that is normal to* S *matters*.
- Moreover, when $\mathbf{F}$ has a component normal to S, we also *want to measure which way* $\mathbf{F}$ *passes through* S. (See the middle and right drawings of Figure

4.27, which show two vector fields flowing through S in opposite directions.) To achieve this, *we will associate one direction through S with positive flux, and the other direction with negative flux.*

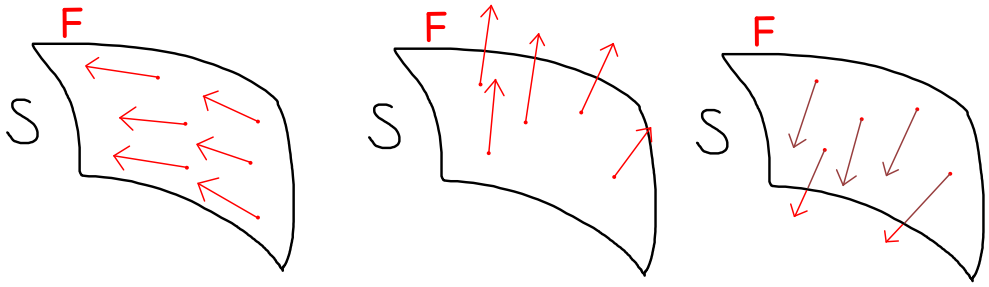


FIGURE 4.27. In the left drawing, the vector field $\mathbf{F}$ is tangent to the surface S; in this case, there is zero flux. On the other hand, in the middle and right drawings, $\mathbf{F}$ is normal to S, but pointing in opposite directions; the flux is nonzero in both cases, however they will have opposite signs.

However, the above discussion leaves open a key question: *which of the two directions through S should correspond to positive flux?* In general, both directions are equally plausible, thus we will have to make an extra choice.

Now, this choice of a direction through S is connected to the choice of *unit normals* to S. In particular, if $\mathbf{F}(\mathbf{p})$ points in the same direction as the chosen unit normal $\mathbf{n}_\mathbf{p}$ at $\mathbf{p}$, then the flow can be considered "positive". On the other hand, if $\mathbf{F}(\mathbf{p})$ points in the opposite direction as $\mathbf{n}_\mathbf{p}$, then the flow is deemed "negative".

Notice the above can be very conveniently captured by the dot product $\mathbf{F}(\mathbf{p}) \cdot \mathbf{n}_\mathbf{p}$. Indeed, this is precisely the component of $\mathbf{F}$ in the $\mathbf{n}_\mathbf{p}$-direction, and it is positive if and only if $\mathbf{F}$ is pointing in the same normal direction as $\mathbf{n}_\mathbf{p}$.

Moreover, by Definition 4.52, such a choice of unit normals $\mathbf{n}$ is determined by an *orientation* of S. Thus, *we can make sense of the flux through S only after we have chosen an orientation of S, which gives the direction corresponding to positive flux.* For example, in Figure 4.27, if we assign the "upward-facing" orientation of S, then the middle and right drawings demonstrate positive and negative flux, respectively.

Combining all the above, we arrive at our formal notion of "flux integral":

Definition 4.74. Let $S \subseteq \mathbb{R}^3$ be an oriented surface, and let $\mathbf{F}$ be a vector field on a subset of $\mathbb{R}^3$ containing S. Then, we define the surface integral of $\mathbf{F}$ over S by

$$\iint_S \mathbf{F} \cdot d\mathbf{A} = \iint_S (\mathbf{F} \cdot \mathbf{n}) dA, \tag{4.20}$$

where $\mathbf{n}$ denotes the unit normals of S in the direction specified by the orientation of S, and where the integral on the right-hand side is defined as in (4.19).

Remark 4.75. Note in particular that if one reverses the orientation of S, then the only change to the right-hand side of (4.20) is that $\mathbf{n}$ is replaced by $-\mathbf{n}$. Therefore, if S^* denotes the same surface S, but with the opposite orientation, then

$$\iint_{S^*} \mathbf{F} \cdot d\mathbf{A} = -\iint_S \mathbf{F} \cdot d\mathbf{A}.$$

Remark 4.76. Definition 4.74 is similar to Definition 3.75 for curve integrals of vector fields. However, one key difference is that Definition 4.74 measures our vector field with respect to the unit normal, whereas Definition 3.75 uses the unit tangent.

The next task is to find a practical method for computing surface integrals of vector fields. The subsequent theorem, which is an analogue of Theorem 3.77 for curve integrals, provides a convenient formula using parametrisations:

Theorem 4.77. Let S, $\mathbf{F}$ be as in Definition 4.74, and let $\sigma : U \to S$ be an injective parametrisation of S, whose image differs from S by a finite collection of points and curves.

- If σ generates the orientation of S, then

$$\iint_S \mathbf{F} \cdot d\mathbf{A} = +\iint_U \left\{ \mathbf{F}(\sigma(u,v)) \cdot [\partial_1\sigma(u,v) \times \partial_2\sigma(u,v)]_{\sigma(u,v)} \right\} du dv. \tag{4.21}$$

- If σ generates the orientation opposite to that of S, then

$$\iint_S \mathbf{F} \cdot d\mathbf{A} = -\iint_U \left\{ \mathbf{F}(\sigma(u,v)) \cdot [\partial_1\sigma(u,v) \times \partial_2\sigma(u,v)]_{\sigma(u,v)} \right\} du dv. \tag{4.22}$$

Proof. First, if σ generates the orientation of S, then Definition 4.54 implies that the chosen unit normal to S at $\sigma(u,v)$, for any $(u,v) \in U$, is

$$\mathbf{n}_{\sigma(u,v)} = +\left[\frac{\boldsymbol{w}(u,v)}{|\boldsymbol{w}(u,v)|} \right]_{\sigma(u,v)}, \qquad \boldsymbol{w}(u,v) = \partial_1\sigma(u,v) \times \partial_2\sigma(u,v).$$

Thus, using Definitions 4.67, 4.71, and 4.74, we conclude that

$$\begin{aligned} \iint_S \mathbf{F} \cdot d\mathbf{A} &= \iint_U \left\{ \mathbf{F}(\sigma(u,v)) \cdot \left[\frac{\boldsymbol{w}(u,v)}{|\boldsymbol{w}(u,v)|} \right]_{\sigma(u,v)} \right\} |\boldsymbol{w}(u,v)| \, du dv \\ &= \iint_U \left[\mathbf{F}(\sigma(u,v)) \cdot \boldsymbol{w}(u,v)_{\sigma(u,v)} \right] du dv, \end{aligned}$$

which is the formula (4.21). The remaining equation (4.22) is proved similarly. □

Example 4.78. Let $\mathcal{C}_*$ be the *cylindrical segment* from Example 4.72,

$$\mathcal{C}_* = \{(x, y, z) \in \mathbb{R}^3 \mid x^2 + y^2 = 1,\ 0 < z < 1\},$$

along with the "outward-facing" orientation. We now integrate, over $\mathcal{C}_*$, the vector field

$$\mathbf{F}(x, y, z) = (x, y, z)_{(x,y,z)}, \qquad (x, y, z) \in \mathbb{R}^3.$$

The first step is find an appropriate parametrisation of $\mathcal{C}_*$. For this, we can use

$$\sigma_* : (0, 2\pi) \times (0, 1) \to \mathbb{R}^3, \qquad \sigma_*(u, v) = (\cos u, \sin u, v),$$

from Example 4.72. Recall σ_* is an injective parametrisation of $\mathcal{C}_*$, and its image is all of $\mathcal{C}_*$ except for a vertical line segment; see the left half of Figure 4.28.

We also observe that for any $(u, v) \in (0, 2\pi) \times (0, 1)$, the quantity

$$+[\partial_1\sigma_*(u, v) \times \partial_2\sigma_*(u, v)]_{\sigma_*(u,v)} = (\cos u, \sin u, 0)_{(\cos u, \sin u, v)},$$

which is in the same direction as the unit normal generated by σ_* (defined in (4.10)), points outward from $\mathcal{C}_*$. Thus, σ_* generates our "outward-facing" orientation of $\mathcal{C}_*$.

Thus, we can apply Theorem 4.77—in particular (4.21)—in order to obtain

$$\iint_{\mathcal{C}_*} \mathbf{F} \cdot d\mathbf{A} = \iint_{(0,2\pi)\times(0,1)} \left\{\mathbf{F}(\sigma_*(u, v)) \cdot [\partial_1\sigma_*(u, v) \times \partial_2\sigma_*(u, v)]_{\sigma_*(u,v)}\right\} du dv.$$

Finally, direct computations and Fubini's theorem lead to our desired solution:

$$\begin{aligned}
\iint_{\mathcal{C}_*} \mathbf{F} \cdot d\mathbf{A} &= \iint_{(0,2\pi)\times(0,1)} [(\cos u, \sin u, v) \cdot (\cos u, \sin u, 0)]\, du dv \\
&= \iint_{(0,2\pi)\times(0,1)} 1\, du dv \\
&= \int_0^1 \int_0^{2\pi} 1\, du dv \\
&= 2\pi.
\end{aligned}$$

Example 4.79. Next, consider the *unit sphere*,

$$\mathbb{S}^2 = \{(x, y, z) \in \mathbb{R}^3 \mid x^2 + y^2 + z^2 = 1\},$$

with the "outward-facing" orientation. Let us integrate $\mathbf{F}$ from Example 4.78 over $\mathbb{S}^2$.

First, recall from Example 4.65 that

$$\rho^*_{s,z} : (0, 2\pi) \times (0, \pi) \to \mathbb{R}^3, \qquad \rho^*_{s,z}(u, v) = (\cos u \sin v, \sin u \sin v, \cos v)$$

is an injective parametrisation of $\mathbb{S}^2$, whose image is all of $\mathbb{S}^2$ except for a closed arc; see the right half of Figure 4.28. Moreover, direct computations (see Example 4.26) yield

$$+[\partial_1\rho^*_{s,z}(u,v)\times\partial_2\rho^*_{s,z}(u,v)]_{\rho^*_{s,z}(u,v)} = -\sin v\cdot\rho^*_{s,z}(u,v)_{\rho^*_{s,z}(u,v)},$$

which, by inspection, can be shown to point inward from the sphere at each point. Thus, it follows that $\rho^*_{s,z}$ *generates the "inward-facing" orientation of* $\mathbb{S}^2$.

Since this is opposite to our given orientation, we apply (4.22) instead of (4.21):

$$\begin{aligned}\iint_{\mathbb{S}^2}\mathbf{F}\cdot d\mathbf{A} &= -\int_0^{2\pi}\int_0^{\pi}\left\{\mathbf{F}(\rho^*_{s,z}(u,v))\cdot[\partial_1\rho^*_{s,z}(u,v)\times\partial_2\rho^*_{s,z}(u,v)]_{\rho^*_{s,z}(u,v)}\right\}dvdu\\ &= -\int_0^{2\pi}\int_0^{\pi}\left\{\rho^*_{s,z}(u,v)_{\rho^*_{s,z}(u,v)}\cdot[-\sin v\,\rho^*_{s,z}(u,v)]_{\rho^*_{s,z}(u,v)}\right\}dvdu\\ &= \int_0^{2\pi}\int_0^{\pi}\sin v\,[\rho^*_{s,z}(u,v)\cdot\rho^*_{s,z}(u,v)]\,dvdu.\end{aligned}$$

Finally, since $\rho^*_{s,z}(u,v)\cdot\rho^*_{s,z}(u,v)=1$ everywhere (either from direct computation, or by noting the values of $\rho^*_{s,z}$ always lie on the unit sphere), we conclude that

$$\begin{aligned}\iint_{\mathbb{S}^2}\mathbf{F}\cdot d\mathbf{A} &= \int_0^{2\pi}\int_0^{\pi}\sin v\,dvdu\\ &= 2\pi\int_0^{\pi}\sin v\,dv\\ &= 4\pi.\end{aligned}$$

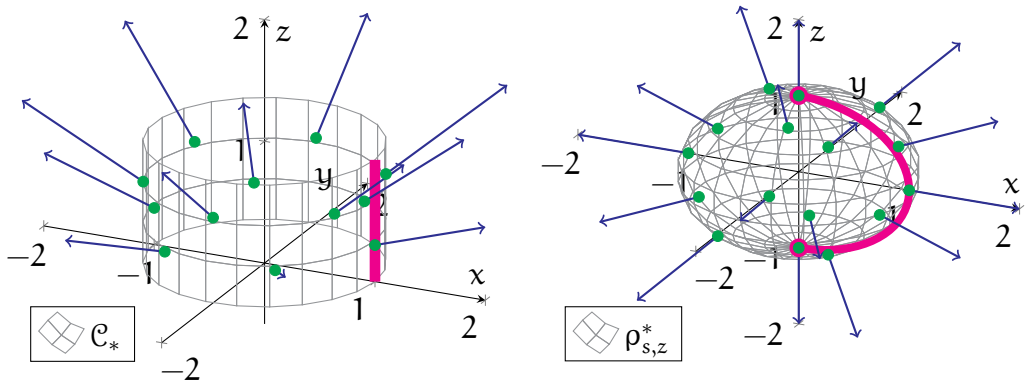


FIGURE 4.28. The left illustration shows the cylindrical segment $\mathcal{C}_*$ from Example 4.78, along with some values of $\mathbf{F}$ along $\mathcal{C}_*$; the image of the parametrisation σ_* is all of $\mathcal{C}_*$ except for the pink line segment. The right illustration shows the sphere $\mathbb{S}^2$ from Example 4.79, again with some values of $\mathbf{F}$; the image of $\rho^*_{s,z}$ covers all of $\mathbb{S}^2$ except for the pink arc.

5. Applications

In this final chapter, we consolidate the knowledge we have gained on the geometry of curves and surfaces, along with our background in calculus and linear algebra. We apply this knowledge to explore a couple important topics in vector calculus:

- *Constrained optimisation problems*, and the method of *Lagrange multipliers*.
- The *integral theorems of vector calculus*, which extend the fundamental theorem of calculus to higher dimensional objects.

In fact, both topics are based on geometric foundations. Thus, unlike more traditional calculus modules, here we approach them from a more geometric point of view.

5.1. Constrained Optimisation.

Let us begin with a hypothetical real world problem. Suppose you are given material to build a fence, say with a total length of 160 metres. In addition, suppose that you are allowed to fence in a rectangular area, and you can then keep the land that you enclosed with your fence.

Question 5.1. How should you build your fence so that you enclose, and hence take, the maximum amount—that is, the maximum area—of land?

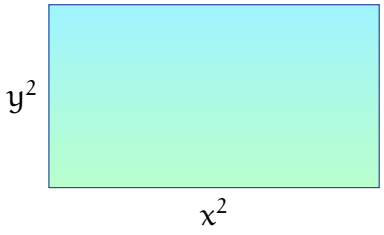


Figure 5.1. This illustration shows the setting of Question 5.1: a rectangular region, with dimensions x^2 and y^2, enclosed by a 160-metre fence.

To convert this into a mathematical problem, we consider a possible rectangular area enclosed by this fence. Let x^2 and y^2 denote the horizontal and vertical dimensions of this rectangle; see Figure 5.1. According to Question 5.1, our objective is to *optimise*—more specifically, to *maximise*—the area $\mathcal{A} = x^2y^2$ of this rectangle.

However, we are not allowed to adjust x^2 and y^2 freely. The condition that our fence is 160 metres long imposes an additional constraint on our problem:

$$2x^2 + 2y^2 = 160.$$

Our objective is to solve problems such as the above, in which we optimise a quantity, but with the stipulation that our parameters are somehow constrained. In

order to discuss the mathematics behind these problems, let us now widen our view to more general classes of *constrained optimisation problems*:

Problem 5.2. Consider the following constrained optimisation problems:

(1) Maximise or minimise $f(x, y)$, subject to the constraint $g(x, y) = c$.
(2) Maximise or minimise $F(x, y, z)$, subject to the constraint $G(x, y, z) = c$.

The only difference between parts (1) and (2) in Problem 5.2 is the number of variables involved in both the constraint and the quantity to be optimised.

Let us first focus on Problem 5.2(1)—the two-variable case. Our first step toward understanding this problem, and toward developing a solution, is to connect it to the geometric theory of curves that we have developed in earlier chapters.

Consider the set C of all points satisfying our constraint:

$$C = \{(x, y) \in \mathbb{R}^2 \mid g(x, y) = c\}. \tag{5.1}$$

If g is "nice", in that its gradient does not vanish anywhere on C, then Theorem 3.24 implies that C is a curve. As a result, our constrained optimisation problem can be equivalently reformulated as a geometric optimisation problem:

Problem 5.3. Maximise or minimise $f(\mathbf{p})$, for all points $\mathbf{p}$ on the curve C from (5.1).

Note this is similar to the problem of *finding extrema* (i.e. maxima or minima) *of functions* that you studied in first-year calculus, except the functions are now defined on *curves* rather than on intervals. To attack this, we will combine our background from first-year calculus with our knowledge on curve geometry.

The key idea behind Problem 5.3 is captured in the following theorem:

Theorem 5.4. Consider the curve (see Theorem 3.24)

$$C = \{(x, y) \in U \mid g(x, y) = c\},$$

where $c \in \mathbb{R}$, where $U \subseteq \mathbb{R}^2$ is open and connected, and where $g : U \to \mathbb{R}$ is a smooth function such that $\nabla g(\mathbf{q})$ is nonzero for every $\mathbf{q} \in C$.

In addition, let $f : U \to \mathbb{R}$ be another smooth function, and assume f achieves either its maximum or minimum value on C at a point $\mathbf{p} \in C$. Then:

- $\nabla f(\mathbf{p})$ is normal to every element of $T_p C$.
- There exists $\lambda \in \mathbb{R}$ such that

$$\nabla f(\mathbf{p}) = \lambda \cdot \nabla g(\mathbf{p}). \tag{5.2}$$

Proof. Consider a parametrisation $\gamma : I \to C$ of C, with $\gamma(t_0) = \mathbf{p}$. Since f achieves an extremal value at $\mathbf{p}$, then $f(\gamma(t))$ is extremised at $t = t_0$, and

$$\frac{d}{dt}[f(\gamma(t))]\Big|_{t=t_0} = 0.$$

Repeating the computations in the proof of Theorem 3.47, we then obtain

$$0 = \nabla f(\mathbf{p}) \cdot \left[s\,\gamma'(t_0)_{\gamma(t_0)}\right], \qquad s \in \mathbb{R}.$$

By Definitions 3.30 and 3.35, we conclude that $\nabla f(\mathbf{p})$ is normal to every element of $T_\gamma(t_0) = T_\mathbf{p}C$. Finally, as Theorem 3.47 implies $\nabla g(\mathbf{p})$ is also normal to every element of $T_\mathbf{p}C$, then $\nabla f(\mathbf{p})$ must be a scalar multiple of $\nabla g(\mathbf{p})$, and (5.2) follows. □

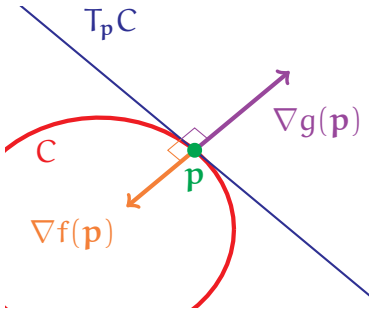


FIGURE 5.2. The illustration shows the setting of Theorem 5.4. If f achieves an extremum at $\mathbf{p}$, then both $\nabla f(\mathbf{p})$ and $\nabla g(\mathbf{p})$ are normal to $T_\mathbf{p}C$.

The setting of Theorem 5.4 is illustrated in Figure 5.2. In particular, f *could only achieve a maximum or minimum value at a point* $\mathbf{p} \in C$ *when* $\nabla f(\mathbf{p})$ *and* $\nabla g(\mathbf{p})$ *are both aligned along a common direction* as in (5.2).

The equation (5.2) is the key observation for handling Problem 5.3—and hence Problem 5.2(1)—in practice. Indeed, (5.2) tells us *we need not search over all of* C *in order to optimise* f. Rather, *we need only check the points of* C *at which* (5.2) *holds*; in general, this represents a much smaller set of points to check.

We now turn our attention to Problem 5.2(2). This is nearly identical to before, except that the quantities F and G of interest now depend on three variables. The main novelty here is that our *constraint is now described by a surface*,

$$S = \{(x, y, z) \in \mathbb{R}^2 \mid G(x, y, z) = c\}. \tag{5.3}$$

Consequently, Problem 5.2(2) can be equivalently reformulated as the following:

Problem 5.5. Maximise or minimise $F(\mathbf{p})$, for all points $\mathbf{p}$ on the surface S from (5.3).

To deal with Problem 5.5, we rely on the following geometric observations:

Theorem 5.6. Consider the surface (see Theorem 4.27)

$$S = \{(x, y, z) \in U \mid G(x, y, z) = c\},$$

where $c \in \mathbb{R}$, where $U \subseteq \mathbb{R}^3$ is open and connected, and where $G : U \to \mathbb{R}$ is a smooth function such that $\nabla G(\mathbf{q})$ is nonzero for every $\mathbf{q} \in S$.

In addition, let $F : U \to \mathbb{R}$ be another smooth function, and assume F achieves either its maximum or minimum value on S at a point $\mathbf{p} \in S$. Then:

- $\nabla F(\mathbf{p})$ is normal to every element of $T_p S$.
- There exists $\lambda \in \mathbb{R}$ such that

$$\nabla F(\mathbf{p}) = \lambda \cdot \nabla G(\mathbf{p}). \tag{5.4}$$

Proof. Let $\sigma : V \to S$ be a parametrisation of S, with $\sigma(u_0, v_0) = \mathbf{p}$. Then, we have

$$\partial_u [F(\sigma(u, v))]|_{(u,v)=(u_0,v_0)} = 0, \qquad \partial_v [F(\sigma(u, v))]|_{(u,v)=(u_0,v_0)} = 0,$$

since $F(\sigma(u, v))$ is extremised when $(u, v) = (u_0, v_0)$. Applying computations that are analogous to the proof of Theorem 4.46 (in particular, the chain rule) yields

$$0 = \nabla F(\mathbf{p}) \cdot \left[a \cdot \partial_1 \sigma(u_0, v_0)_{\sigma(u_0,v_0)} + b \cdot \partial_2 \sigma(u_0, v_0)_{\sigma(u_0,v_0)} \right], \qquad a, b \in \mathbb{R}.$$

By Definitions 4.6 and 4.38, we see that $\nabla F(\mathbf{p})$ is normal to $T_\sigma(u_0, v_0) = T_p S$. Finally, since both $\nabla F(\mathbf{p})$ and $\nabla G(\mathbf{p})$ are normal to $T_p S$ (see Theorem 4.46), it follows that $\nabla F(\mathbf{p})$ must be a scalar multiple of $\nabla G(\mathbf{p})$, which is (5.4). □

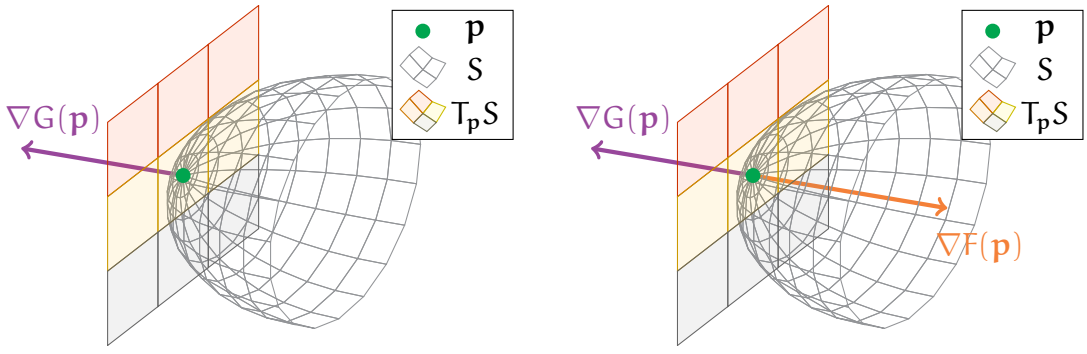


FIGURE 5.3. The illustrations show the setting of Theorem 5.6. In the left plot, the gradient $\nabla G(\mathbf{p})$ (in purple) is normal to the tangent plane $T_p S$. The right graphic shows the case in which F achieves an extremal value at $\mathbf{p}$; here, both $\nabla F(\mathbf{p})$ and $\nabla G(\mathbf{p})$ are normal to $T_p S$.

The conclusions from Theorem 5.6 are analogous to those from Theorem 5.4. In the context of Problem 5.5, we see that *we need not consider all of* S *when seeking to optimise* F. Indeed, *we need only check the points of* S *at which* (5.4) *holds*.

5.2. **Lagrange Multipliers.** We have discussed how Theorems 5.4 and 5.6 can contribute toward solving Problem 5.2, in particular by restricting the possible points where extrema can occur. From this observation, we can now craft a process, known as the *method of Lagrange multipliers*, for dealing with Problem 5.2.

Let us first look at constrained optimisation problems with two variables—Problem 5.2(1). Here, the method of Lagrange multipliers can be summarised as follows:

- **Step 1.** Consider the following system of equations:

$$\partial_1 f(x,y) = \lambda \cdot \partial_1 g(x,y), \qquad \partial_2 f(x,y) = \lambda \cdot \partial_2 g(x,y), \qquad g(x,y) = c. \tag{5.5}$$

- **Step 2.** Solve the system (5.5) for the unknowns $x, y, \lambda \in \mathbb{R}$.
- **Step 3.** For each solution (x,y,λ) from Step 2, check directly whether (x,y) is a maximum or a minimum of f, subject to the constraint $g(x,y) = c$.

Notice that the third equation in (5.5) is simply our given constraint. Furthermore, first two equations in (5.5) capture the gradient relation (5.2).

Remark 5.7. The term "Lagrange multiplier" refers to the real constant λ in (5.5), which is an extra unknown that needs to be found. The method is named after Joseph-Louis Lagrange (French and Italian mathematician, 1736–1813).

Example 5.8. To demonstrate how the method of Lagrange multipliers works, let us apply it to the original fence-building problem that was posed in Question 5.1: *maximise the quantity* x^2y^2*, subject to the constraint* $x^2 + y^2 = 80$.

To set up the system (5.5), we define the functions

$$f(x,y) = x^2y^2, \qquad g(x,y) = x^2 + y^2, \qquad (x,y) \in \mathbb{R}^2,$$

representing the quantity to be optimised and the constraint, respectively. Note that

$$\partial_1 f(x,y) = 2xy^2, \qquad \partial_2 f(x,y) = 2x^2y,$$
$$\partial_1 g(x,y) = 2x, \qquad \partial_2 g(x,y) = 2y.$$

As a result, the system that we must solve is

$$2xy^2 = \lambda \cdot 2x, \qquad 2x^2y = \lambda \cdot 2y, \qquad x^2 + y^2 = 80. \tag{5.6}$$

The first two equations in (5.6) must be treated differently depending on whether x or y vanishes (since we will divide by x and y). Thus, we split our analysis into cases:

- *Case 1:* Suppose $x = 0$. Then, the first equation in (5.6) becomes $0 = 0$ and gives no information. However, the third equation implies $y^2 = 80$, and the second equation then yields $\lambda = 0$. As a result, we obtain two solutions to (5.6):
$$(x, y, \lambda) = \left(0, \pm\sqrt{80}, 0\right).$$
- *Case 2:* Suppose instead $y = 0$. Similar to the preceding case, the second equation in (5.6) yields no information, while the third and first equations imply $x^2 = 80$ and $\lambda = 0$, respectively. Thus, we obtain another pair of solutions to (5.6):
$$(x, y, \lambda) = \left(\pm\sqrt{80}, 0, 0\right).$$
- *Case 3:* Finally, suppose both $x \neq 0$ and $y \neq 0$. Then, we can divide the first two equations of (5.6) by $2x$ and $2y$, respectively. This yields a pair of equations,
$$y^2 = \lambda = x^2.$$
Combining this with the third equation of (5.6) yields
$$x^2 + x^2 = 80, \qquad x = \pm\sqrt{40}.$$
Since $y^2 = x^2 = \lambda$, we obtain four more solutions to (5.6):
$$(x, y, \lambda) = \left(\pm\sqrt{40}, \pm\sqrt{40}, 40\right).$$

Note that the above three cases exhaust all the possibilites for (x, y). Therefore, we have produced all the possible solutions (x, y, λ) for (5.6).

Finally, we compare the values $f(x, y)$ for each solution (x, y, λ) of (5.6):

$$f\left(0, \pm\sqrt{80}\right) = 0, \qquad f\left(\pm\sqrt{80}, 0\right) = 0, \tag{5.7}$$
$$f\left(\pm\sqrt{40}, \pm\sqrt{40}\right) = 1600.$$

In particular, the highest value of f in (5.7) is 1600, which is achieved when

$$x^2 = y^2 = 40.$$

As a result, we conclude that 1600 *may be* the maximum value of x^2y^2 subject to the constraint $x^2 + y^2 = 80$. (In particular, we only know that *if* f achieves a maximum, then it must correspond to a solution of (5.6); however, such a maximum may not exist.)

In fact, 1600 is indeed the maximum value here. (Let us just accept this for now; we will justify it later on.) Thus, in terms of Question 5.1, we conclude that the area of the fenced region is maximised at 1600 square metres when the region is a square, with side lengths of 40 metres; see the right part of Figure 5.4 for an illustration.

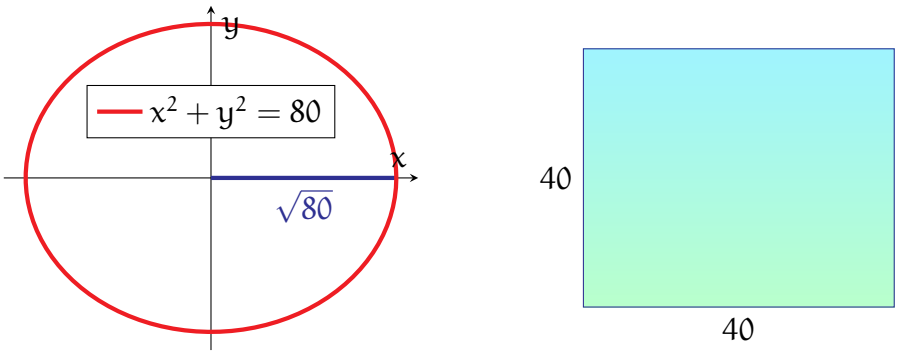


FIGURE 5.4. The two illustrations demonstrate the setting of Question 5.1 and Example 5.8. The left part shows the circle representing all the points satisfying the constraint $x^2+y^2 = 80$, while the right part shows the optimal solution, which is a square region with side lengths of 40 metres.

Remark 5.9. Although λ is one of the unknowns of the system (5.5), its value does not affect our final answer. In particular, we are only interested in values of $f(x, y)$.

Next, constrained optimisation problems involving 3 variables—that is, Problem 5.2(2)—can be treated using an analogous *method of Lagrange multipliers*:

- **Step 1.** Consider the following system of equations:

$$\begin{aligned} &\partial_1 F(x,y,z) = \lambda \cdot \partial_1 G(x,y,z), \qquad \partial_2 F(x,y,z) = \lambda \cdot \partial_2 G(x,y,z), \\ &\partial_3 F(x,y,z) = \lambda \cdot \partial_3 G(x,y,z), \qquad G(x,y,z) = c. \end{aligned} \tag{5.8}$$

- **Step 2.** Solve the system (5.8) for the unknowns $x, y, z, \lambda \in \mathbb{R}$.
- **Step 3.** For each solution (x, y, z, λ) from Step 2, check whether (x, y, z) is a maximum or a minimum of F, subject to the constraint $G(x, y, z) = c$.

The last equation in (5.8) represents our constraint on (x, y, z), while the remaining equations correspond to the gradient relation (5.4).

Example 5.10. Let us solve the following: *find the extremal values of* $2xy + z$ *subject to the constraint* $x^2 + y^2 + z^2 = 1$, *and find the points where these values are achieved.*

In order to apply the method of Lagrange multipliers, we first define the functions

$$F(x, y, z) = 2xy + z, \qquad G(x, y, z) = x^2 + y^2 + z^2, \qquad (x, y, z) \in \mathbb{R}^3.$$

Observe that the partial derivatives of F and G satisfy

$$\partial_1 F(x,y,z) = 2y, \qquad \partial_2 F(x,y,z) = 2x, \qquad \partial_3 F(x,y,z) = 1,$$
$$\partial_1 G(x,y,z) = 2x, \qquad \partial_2 G(x,y,z) = 2y, \qquad \partial_3 G(x,y,z) = 2z.$$

Thus, the system (corresponding to (5.8)) that we must solve is

$$2y = 2\lambda x, \qquad 2x = 2\lambda y, \qquad 1 = 2\lambda z, \qquad x^2 + y^2 + z^2 = 1. \tag{5.9}$$

To solve (5.9) (which again involves dividing by x and y), we split into cases:

- *Case 1:* Suppose $x = 0$. Then, the first equation in (5.9) implies $y = 0$, and the fourth equation yields $z^2 = 1$. Thus, in this case, we obtain the solutions
$$(x,y,z,\lambda) = \left(0,0,+1,+\frac{1}{2}\right), \qquad (x,y,z,\lambda) = \left(0,0,-1,-\frac{1}{2}\right)$$
to (5.9). (Note the values of λ follow from the third equation in (5.9).) Furthermore, the values of F corresponding to the above solutions are
$$F(0,0,\pm 1) = \pm 1. \tag{5.10}$$
- *Case 2:* Suppose $y = 0$. The second equation of (5.9) then yields $x = 0$. Thus, we obtain the same solutions of (5.9) as in the preceding case.
- *Case 3:* Finally, suppose both $x \neq 0$ and $y \neq 0$. Then, dividing the first and second equations of (5.9) by $2x$ and $2y$, respectively, we see that
$$\frac{y}{x} = \lambda = \frac{x}{y}, \qquad x^2 = y^2.$$
From the above and the third equation of (5.9), we obtain two possibilities:
$$x = +y, \qquad \lambda = +1, \qquad z = +\frac{1}{2},$$
$$x = -y, \qquad \lambda = -1, \qquad z = -\frac{1}{2}.$$
Combining the above with the constraint equation yields
$$x^2 + x^2 + \left(\pm\frac{1}{2}\right)^2 = 1, \qquad x = \pm\sqrt{\frac{3}{8}}.$$
Thus, in this case, we end up with four solutions of (5.9):
$$(x_1,y_1,z_1,\lambda_1) = \left(+\sqrt{\frac{3}{8}}, +\sqrt{\frac{3}{8}}, +\frac{1}{2}, +1\right), \qquad F(x_1,y_1,z_1) = +\frac{5}{4}, \tag{5.11}$$

$$(x_2, y_2, z_2, \lambda_2) = \left(+\sqrt{\frac{3}{8}}, -\sqrt{\frac{3}{8}}, -\frac{1}{2}, -1\right), \qquad F(x_1, y_1, z_1) = -\frac{5}{4},$$
$$(x_3, y_3, z_3, \lambda_3) = \left(-\sqrt{\frac{3}{8}}, +\sqrt{\frac{3}{8}}, -\frac{1}{2}, -1\right), \qquad F(x_1, y_1, z_1) = -\frac{5}{4},$$
$$(x_4, y_4, z_4, \lambda_4) = \left(-\sqrt{\frac{3}{8}}, -\sqrt{\frac{3}{8}}, +\frac{1}{2}, +1\right), \qquad F(x_1, y_1, z_1) = +\frac{5}{4}.$$

In particular, (5.10) and (5.11) contain the values of F for *all* the solutions of (5.9).

Comparing all these values, we conclude that:

- The maximum of F *may be* $+\frac{5}{4}$, which is achieved at (x_1, y_1, z_1) and (x_4, y_4, z_4).
- The minimum of F *may be* $-\frac{5}{4}$, which is achieved at (x_2, y_2, z_2) and (x_3, y_3, z_3).

(In fact, $+\frac{5}{4}$ and $-\frac{5}{4}$ are indeed the extreme values of F. At this point, however, we cannot be certain of this, as we have not yet excluded the possibility that the maximum or the minimum does not exist. We will address this issue in the subsequent section.)

5.3. **Existence of Extrema.** While the method of Lagrange multipliers yields information about where the extrema in a constrained optimisation problem *may be* attained, it does not tell us whether these values are actually achieved. In fact, there are many simple examples for which the maximum or the minimum does not exist:

Example 5.11. Let $f(x, y)$ denote the squared distance between (x, y) and the origin:

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

Let us try to *optimise* $f(x, y)$, *subject to the constraint* $x = 1$.

The setting is illustrated in the left half of Figure 5.5. The origin is marked in green, while the constraint $x = 1$ is described by the red vertical line. Our goal is to optimise the (squared) distance between any point on the red line and the green point.

The solutions to this question should be clear just by inspection. The minimum value of f is achieved at the point $(x, y) = (1, 0)$, directly to the right of the origin; the smallest distance is $f(1, 0) = 1$. On the other hand, there is no maximum value of f, since $f(x, y)$ becomes arbitrarily large as we slide further up or down the red line. Thus, $f(x, y)$ *has a minimum but no maximum when subject to the constraint* $x = 1$.

Example 5.12. Consider the upper half-circle given by

$$C = \{(x, y) \in \mathbb{R}^2 \mid x^2 + y^2 = 1,\ y > 0\}.$$

See the right part of Figure 5.5, in which C is drawn as a red curve. Let us optimise

$$f(x, y) = x^2 + (y + 1)^2,$$

over all the points $(x, y) \in C$. Note $f(x, y)$ is precisely the squared distance from (x, y) to the point $(0, -1)$, that is, the green point in the right graphic of Figure 5.5.

From Figure 5.5, one can see that the maximum of f is achieved at the uppermost point $(0, 1)$ of C. On the other hand, f has no minimum on C. In particular, while f becomes smaller as one gets closer toward the left and right boundary points of C, it does not actually achieve a minimum value on C itself.

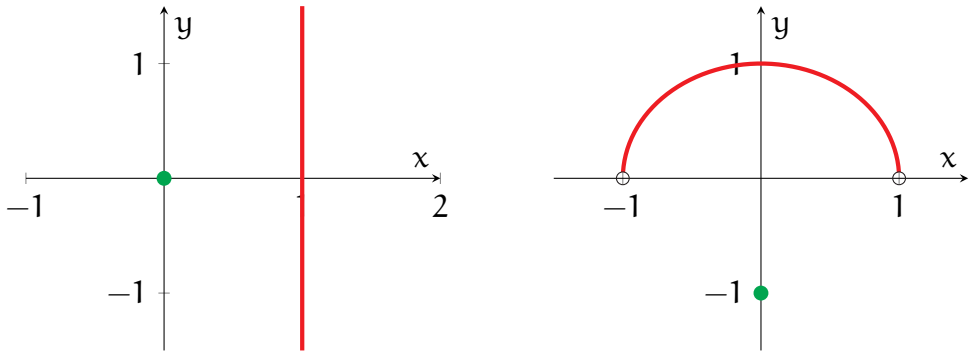


FIGURE 5.5. The left illustration shows the setting from Example 5.11; here, the goal is to optimise the squared distance between the green point (at the origin) and any point on the red line. The right graphic illustrates the setting of Example 5.12, in which the aim is to optimise the squared distance between the green point and red half-circle.

As you may imagine, this possible non-existence of extrema could add a considerable amount of trouble to our analysis. Therefore, we ask whether there are some especially nice situations where optimal values are guaranteed to exist.

In fact, we do indeed know of such favourable settings—this can be established using arguments from abstract analysis and topology. While these proofs do (unfortunately) lie beyond the scope of this module, we will give an informal description of the results below. For this, we will need some additional terminology:

Definition 5.13. Let $M \subseteq \mathbb{R}^n$ be a curve or a surface.

- M is bounded iff M "does not go off toward infinity".
- M is closed iff M "does not contain any boundary, or edge, points".

Rather than discussing precise definitions (which requires additional background), we instead demonstrate these concepts informally through visual examples:

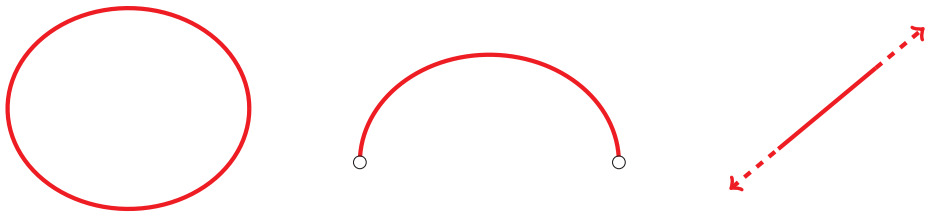


FIGURE 5.6. These three curves are discussed in Example 5.14: the circle (left) is both closed and bounded, the half-circle (middle) is bounded but not closed, while the line (right) is closed but not bounded.

Example 5.14. Consider the three curves in Figure 5.6:

- The circle in the left illustration is bounded, since no part of it goes off to infinity. Indeed, all the points of the circle are contained in a finite region of the plane. Moreover, the circle is closed, since it has no boundary points. More precisely, one can go indefinitely in either direction along the circle without it "terminating".
- The half-circle in the middle illustration is also bounded—in particular, since the half-circle is a subset of the full circle, it is also contained within a finite region. On the other hand, the half-circle fails to be closed, since it terminates (in both directions) at two boundary points marked in black.
- The infinite line in the right illustration fails to be bounded, as it goes off to infinity in both directions. However, this line is closed, since it does not terminate no matter how far one travels along it in either direction.

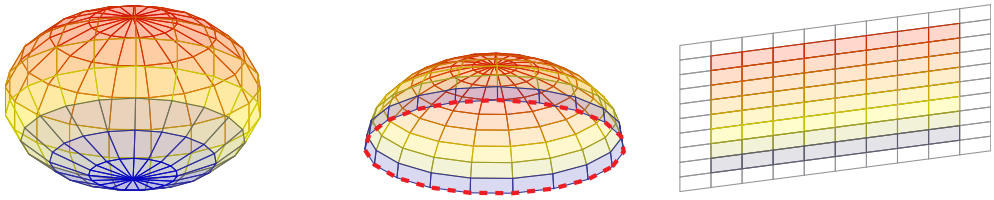


FIGURE 5.7. These three surfaces are discussed in Example 5.15: the sphere (left) is both closed and bounded, the half-sphere (middle) is bounded but not closed, while the plane (right) is closed but not bounded.

Example 5.15. Let us next consider the surfaces in Figure 5.7:

- The sphere in the left is both closed and bounded. In particular, it is contained within a finite region, and it does not have any boundary points.
- The half-sphere in the middle is bounded but not closed. Like the sphere, it is contained within a finite region. On the other hand, the half-sphere does have boundary points (namely, the red dashed equatorial circle in the figure).

- The infinite plane on the right is closed but not bounded. In particular, the plane has no boundary points at which it terminates, but it does go off to infinity.

Remark 5.16. Of course, there do exist precise, formal definitions of closed and bounded sets. These topics are further explored in *MTH6127: Metric Spaces and Topology*.

The main result regarding the existence of extrema is stated below:

Theorem 5.17. Let $M \subseteq \mathbb{R}^n$ be a curve or a surface, and let f be a smooth real-valued function whose domain includes M. If M is both closed and bounded, then f achieves both a maximum value and a minimum value on M.

Remark 5.18. Theorem 5.17 is a consequence of the *Heine–Borel theorem* from real analysis. This is also further explored in *MTH6127: Metric Spaces and Topology*.

In particular, Examples 5.14 and 5.15 show that both circles and spheres are closed and bounded. Thus, it is convenient to consider problems with constraints given by circles or spheres, since Theorem 5.17 guarantees that we must achieve both a maximum and a minimum. Moreover, by Theorems 5.4 and 5.6, we know that these extrema can always be found through the method of Lagrange multipliers.

Example 5.19. We return to the problem, from Example 5.8, of finding the maximum of $f(x, y) = x^2y^2$ subject to the constraint $x^2 + y^2 = 80$. Note the constraint curve,

$$C = \{(x, y) \in \mathbb{R}^2 \mid x^2 + y^2 = 80\},$$

which is a circle of radius $\sqrt{80}$, is both closed and bounded; see the left graphic in Figure 5.4. Then, from Theorems 5.4 and 5.17, we deduce that the maximum of f exists, and that it must be achieved at a point corresponding to a solution of the system (5.6).

To find this maximum of f, we need only compare all the values of f at the solution points of (5.6). This was already done in (5.7), where we found that the largest value of f is 1600. This proves that the maximum value of $f(x, y)$, subject to the constraint $x^2 + y^2 = 80$, is 1600, and this maximum is achieved whenever $x^2 = y^2 = 40$.

Example 5.20. Returning to Example 5.10, we again notice that the constraint surface,

$$S = \{(x, y, z) \in \mathbb{R}^3 \mid x^2 + y^2 + z^2 = 1\},$$

a unit sphere, is both closed and bounded. Theorem 5.17 then implies $F(x, y, z) = 2xy+z$ achieves both a maximum and a minimum value on S. To find these extrema, we again compare the values of F at the solution points of (5.9).

This was already done in (5.10) and (5.11); from there, we conclude that the maximum value of $2xy + z$, subject to the constraint $x^2 + y^2 + z^2 = 1$, is $+\frac{5}{4}$, and this value is achieved at the points (x_1, y_1, z_1) and (x_4, y_4, z_4). Similarly, the minimum value is $-\frac{5}{4}$, and this value is achieved at (x_2, y_2, z_2) and (x_3, y_3, z_3).

Finally, when a curve or surface M fails to be closed or bounded, the analysis becomes more complicated. Here, one has several possibilities to consider:

(1) A maximum or minimum value is achieved at a point of M. The value and the point can then be found via the method of Lagrange multipliers, as before.
(2) One could approach a maximum or minimum at a boundary point of M.
(3) One could also approach a maximum or minimum "toward infinity".

If either case (2) or case (3) occurs, then a maximum or a minimum value does not exist, since neither the boundary of M nor "infinity" are points of M.

Example 5.21. As a simple example, let us optimise $f(x, y) = y^2$ on the *half-line*

$$L = \{(x, y) \in \mathbb{R}^2 \mid x = 0,\ y > -1\}.$$

(The constraint is $g(x, y) = x = 0$, with the additional restriction $y > -1$.)

To find potential extrema on L, we first apply the method of Lagrange multipliers. Here, the system corresponding to (5.5), with the above f and g, is given by

$$0 = \lambda \cdot 1, \qquad 2y = \lambda \cdot 0, \qquad x = 0,$$

which only has one solution corresponding to a point of L:

$$(x, y, \lambda) = (0, 0, 0), \qquad f(0, 0) = 0. \tag{5.12}$$

Since L has a boundary point $(0, -1)$, we must also check what happens there. Here, we see that as one approaches $(0, -1)$, the value of f approaches

$$f(0, -1) = 1. \tag{5.13}$$

Also, since L is unbounded, we must also check what happens as one goes "toward infinity" along L. Here, as one traverses infinitely far up L, the value of f approaches

$$\lim_{y \nearrow +\infty} f(0, y) = +\infty. \tag{5.14}$$

Finally, comparing the values of f in (5.12)–(5.14), we conclude that:

- f attains a genuine minimum value, 0, at the point $(0, 0) \in L$.

- f does not attain a maximum value on L. Instead, the largest value ($+\infty$) of f is reached as one goes upward toward infinity along L.

On the other hand, the value of f at the boundary point $(0, -1)$ of L (which, by definition, is not counted as a point of L) is not an extremum.

5.4. **Conservative Fields.** The final topics for this module involve *integral theorems from vector calculus*, which, though usually associated with calculus, are geometric in nature. Let us first recall the familiar *fundamental theorem of calculus*:

$$\int_a^b f'(x)\,dx = f(b) - f(a). \tag{5.15}$$

While you should already know (5.15) intimately by now, here we note the following:

- The left-hand side is an *integral*, over the 1-*dimensional* interval (a, b). Moreover, the integrand here is the *derivative* of a function f.
- The right-hand side is the difference between the values of f at b and a. This can be viewed as an *integral* of f over the 0-*dimensional boundary* of (a, b).

The general question we want to pose is the following:

Question 5.22. Are there geometric generalisations of the fundamental theorem of calculus? Also, do these generalisations extend to higher dimensional settings?

First, let us see what happens when the interval (a, b) in (5.15) is replaced by an oriented curve. The general result is give by the following theorem:

Theorem 5.23. Let $C \subseteq \mathbb{R}^n$ be a bounded oriented curve, and suppose $\mathbf{p}, \mathbf{q} \in \mathbb{R}^n$ are the initial and final points of C, respectively. Then, for any smooth map $f : U \to \mathbb{R}$, for which its domain $U \subseteq \mathbb{R}^n$ contains C, $\mathbf{p}$, and $\mathbf{q}$, we have the identity

$$\int_C \nabla f \cdot d\mathbf{s} = f(\mathbf{q}) - f(\mathbf{p}). \tag{5.16}$$

Let us not be too precise on the definitions of initial and final points in Theorem 5.23, though the intuition should be clear. Here, we simply informally view $\mathbf{p}$ and $\mathbf{q}$ as the endpoints of C, arranged so that C *goes from* $\mathbf{p}$ *to* $\mathbf{q}$; see Figure 5.8.

Proof of Theorem 5.23. The assumptions imposed on C—as well as on $\mathbf{p}$ and $\mathbf{q}$—imply there exists an injective parametrisation $\gamma : (a, b) \to C$ of C satisfying

$$\lim_{t \searrow a} \gamma(t) = \mathbf{p}, \qquad \lim_{t \nearrow b} \gamma(t) = \mathbf{q}.$$

By Theorem 3.77, we then have

$$\int_C \nabla f \cdot d\mathbf{s} = \int_a^b \left[\nabla f(\gamma(t)) \cdot \gamma'(t)_{\gamma(t)}\right] dt.$$

Moreover, computations analogous to those in the proof of Theorem 3.47 yield

$$\nabla f(\gamma(t)) \cdot \gamma'(t)_{\gamma(t)} = \frac{d}{dt}[f(\gamma(t))].$$

Combining the above with the fundamental theorem of calculus, we obtain

$$\begin{aligned}\int_C \nabla f \cdot d\mathbf{s} &= \int_a^b \frac{d}{dt}[f(\gamma(t))]\, dt \\ &= f(\mathbf{q}) - f(\mathbf{p}).\end{aligned}$$

□

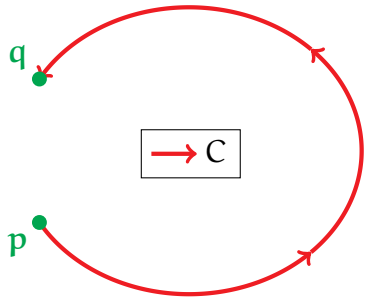


FIGURE 5.8. The above shows the objects from Theorem 5.23: the oriented curve C, its initial point $\mathbf{p}$, and its final point $\mathbf{q}$.

Observe that (5.16) is analogous in spirit to the original fundamental theorem of calculus (5.15). Again, the left-hand side of (5.16) is an integral, over a 1-*dimensional* curve C, of a specially chosen *derivative* (the gradient) of a function. Similarly, the right-hand side is an integral of f itself over the 0-*dimensional boundary* of C.

Remark 5.24. If we take $n = 1$ and C to be an open interval in the setting of Theorem 5.23, then we recover precisely the fundamental theorem of calculus (5.15). Thus, Theorem 5.23 can be viewed as a direct generalisation of (5.15).

Example 5.25. Let $\mathbf{G}$ be the vector field on $\mathbb{R}^3 \setminus \{(0,0,0)\}$ from Example 2.59:

$$\mathbf{G}(\mathbf{x}) = -\frac{1}{|\mathbf{x}|^3} \cdot \mathbf{x}_{\mathbf{x}}.$$

Moreover, let C denote *any* bounded oriented curve in $\mathbb{R}^3$ such that:

- C does not pass through the origin.
- $(-1,0,0)$ and $(0,1,1)$ are the initial and final points of C, respectively.

Let us now integrate $\mathbf{G}$ over this curve C.

The key observation comes from Example 2.65, namely, that $\mathbf{G}$ can be expressed as a gradient, $\mathbf{G} = \nabla g$, where g is the real-valued function

$$g : \mathbb{R}^3 \setminus \{(0,0,0)\} \to \mathbb{R}, \qquad g(\mathbf{p}) = \frac{1}{|\mathbf{p}|}.$$

By our assumptions for C, we see that g is well-defined on both C and its endpoints. As a result, applying Theorem 5.23 to C and g, we conclude that

$$\begin{aligned} \int_C \mathbf{G} \cdot d\mathbf{s} &= \int_C \nabla g \cdot d\mathbf{s} \\ &= g(0,1,1) - g(-1,0,0) \\ &= \frac{1}{\sqrt{2}} - 1. \end{aligned}$$

Recall also, from Example 2.59, that $\mathbf{G}$ represents the Newtonian gravitational force exerted by a particle at the origin. As a result, the integral of $\mathbf{G}$ over C represents the total work done by this gravitational force on another object travelling along C, from $(-1,0,0)$ to $(0,1,1)$. In particular, observe that *the work done depends only on the values of the gravitational potential g at the endpoints of C.*

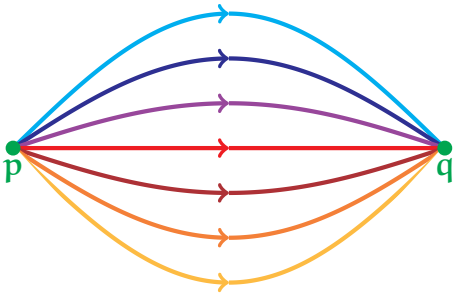


FIGURE 5.9. The above illustration shows oriented curves starting and ending at common points $\mathbf{p}$ and $\mathbf{q}$, respectively. Integrating a gradient ∇f along any of these oriented curves will yield the same value.

Example 5.25 illuminates a rather interesting consequence of Theorem 5.23. Returning to the abstract setting of the theorem, suppose wish to compute

$$\int_C \nabla f \cdot d\mathbf{s}.$$

Then, (5.16) implies that, *although the above integral involves the values of f on all of C, it ultimately depends only on the values of f at the endpoints $\mathbf{p}$ and $\mathbf{q}$.*

In addition, consider a different bounded oriented curve, C', which has the same initial and final points as C. Then, applying Theorem 5.23 twice, we see that

$$\begin{aligned}\int_{C'} \nabla f \cdot d\mathbf{s} &= f(\mathbf{q}) - f(\mathbf{p}) \\ &= \int_C \nabla f \cdot d\mathbf{s}.\end{aligned}$$

In particular, *even though* C *and* C' *may be radically different, Theorem 5.23 ensures that the integrals of* ∇f *over* C *and* C' *must be the same.* Moreover, this holds for *any* pair of curves, provided they have the same starting and end points; see Figure 5.9.

In summary, we conclude that curve integrals of ∇f are *independent of path*, in that the values of these integrals do not depend on what the designated curve is doing, but only on where the curve begins and ends. In other words, we can think of gradients as belonging to a very special class of vector fields.

Definition 5.26. Let F be a vector field on an open and connected subset $U \subseteq \mathbb{R}^n$. We say that F is *conservative* iff for any $\mathbf{p}, \mathbf{q} \in U$, and for any bounded oriented curves $C, C' \subseteq U$ with both the same initial point $\mathbf{p}$ and the same final point $\mathbf{q}$, we have

$$\int_C F \cdot d\mathbf{s} = \int_{C'} F \cdot d\mathbf{s}. \tag{5.17}$$

While Theorem 5.23 implies that gradients are conservative fields, the converse is also true: *every conservative field is the gradient of some real-valued function.*

Theorem 5.27. Let F be a smooth vector field on an open and connected subset $U \subseteq \mathbb{R}^n$. Then, F is conservative if and only if there is a smooth $f : U \to \mathbb{R}$ such that $F = \nabla f$.

Proof. See the *MTH5102: Calculus III* lecture notes [2] or a calculus textbook (e.g. [10]) for details. In fact, if F is conservative, then for a fixed $\mathbf{p}_0 \in U$, one can define

$$f(\mathbf{p}) = \int_{C(\mathbf{p})} F \cdot d\mathbf{s}, \qquad \mathbf{p} \in U,$$

where $C(\mathbf{p}) \subseteq U$ is *any* oriented curve with starting point $\mathbf{p}_0$ and final point $\mathbf{p}$. □

Remark 5.28. The term "conservative fields" comes from physics, as such vector fields represent forces within systems that have a conserved energy.

5.5. **Green's Theorem.** The next step is to take the viewpoint we have developed with the fundamental theorem of calculus (5.15) and its generalisation, Theorem 5.23, and to then explore their analogues in higher dimensions.

Let us start here with the case when all objects are just one dimension higher. More specifically, let us consider integral identities of the form:

$$\iint_{\text{2-dimensional object } A} (\text{“derivative” of } F) = \int_{\text{1-dimensional boundary of } A} F. \tag{5.18}$$

Note the connection between (5.18) and our interpretation of (5.15):

- The left-hand side is a 2-dimensional integral of a derivative.
- The right-hand side is an integral over the 1-dimensional boundary.

Suppose, moreover, that the “2-dimensional object A” in (5.18) is a subset of $\mathbb{R}^2$. In this setting, the precise statement of (5.18) is known as *Green's theorem*:

Theorem 5.29 (Green's theorem)**.** Let $D \subseteq \mathbb{R}^2$ be open and bounded, and assume the boundary of D consists of a union of curves $C_1, C_2, \ldots, C_k$. Also, let $\mathbf{F}$ be a smooth vector field defined on D and its boundary, and expand $\mathbf{F}$ in terms of its components as

$$\mathbf{F}(\mathbf{p}) = (F_1(\mathbf{p}), F_2(\mathbf{p}))_{\mathbf{p}}. \tag{5.19}$$

Then, we have the integral identity,

$$\iint_D [\partial_1 F_2(x, y) - \partial_2 F_1(x, y)]\, dx dy = \int_{C_1} \mathbf{F} \cdot d\mathbf{s} + \cdots + \int_{C_k} \mathbf{F} \cdot d\mathbf{s}, \tag{5.20}$$

where the curves $C_1, \ldots, C_k$ are positively oriented with respect to D:

- For each $1 \leq i \leq k$, we choose the orientation of C_i that is *leftward* (i.e. anti-clockwise) from the *outward-pointing normal direction* from D.

Remark 5.30. Theorem 5.29 is named after *George Green* (British mathematician and physicist, 1793–1841), who had a rather unusual background. Up until the point when Green published his most renowned work in 1828, which contained a less modern version of Theorem 5.29, he was almost entirely self-taught. During this time, he worked as a baker and miller, and he had only one year of formal schooling.

Notice that (5.20) fits into the framework from (5.18). Here, the 2-dimensional region A is the domain D, while the object F being integrated is the vector field $\mathbf{F}$. In addition, the special “derivative” of $\mathbf{F}$ is the integrand in the left-hand side of (5.20).

The notion of *positive orientation* in the statement of Theorem 5.29 also deserves further explanation. However, rather than giving an extended treatment with precise definitions, it is perhaps more instructive (and less time-consuming) to explain this informally through some illustrated examples:

Example 5.31. Let D be the blue shaded disk in the left part of Figure 5.10. Observe that in this case, the boundary of D is the single circular curve C (in red). As a result, for an appropriate vector field $\mathbf{F}$, expanded as in (5.19), Green's theorem yields

$$\iint_D [\partial_1 F_2(x,y) - \partial_2 F_1(x,y)]\,dxdy = \int_C \mathbf{F}\cdot d\mathbf{s}. \tag{5.21}$$

It remains to determine the orientation of C, which is needed for the integral on the right-hand side of (5.21) to be well-defined. For this, we note the purple arrows in the left illustration of Figure 5.10, which depict outward normals from D. If we turn leftward from each of these purple arrows, then we face the *anticlockwise* direction along C; this is precisely the positive orientation of C described in Theorem 5.29.

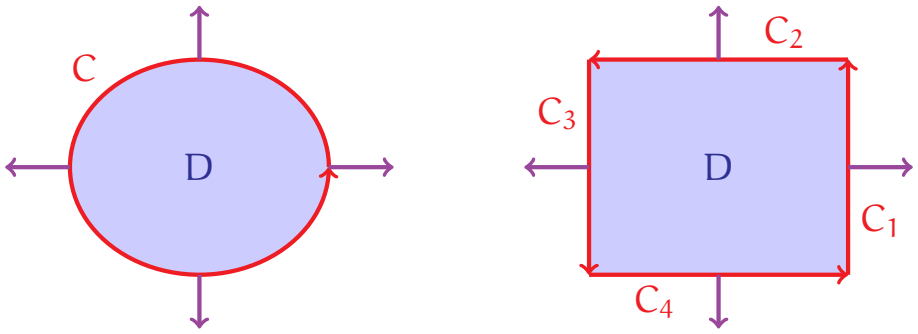


FIGURE 5.10. The left and right graphics illustrate the settings of Examples 5.31 and 5.32, respectively. In both cases, D is the blue shaded region, while the oriented boundary curves are shown in red. In addition, some instances of the outward normal from D are drawn in purple.

Example 5.32. Next, let D be the rectangular region in the right side of Figure 5.10. This is similar to Example 5.31, except the boundary of D is now given by four separate (linear) curves C_1, C_2, C_3, C_4. In this case, Green's theorem implies

$$\iint_D [\partial_1 F_2(x,y) - \partial_2 F_1(x,y)]\,dxdy = \int_{C_1} \mathbf{F}\cdot d\mathbf{s} + \cdots + \int_{C_4} \mathbf{F}\cdot d\mathbf{s}, \tag{5.22}$$

where the vector field $\mathbf{F}$ satisfies the assumptions in Theorem 5.29. In particular, the right-hand side of (5.22) is now a sum of four curve integrals, rather than just one.

Again, to make full sense of (5.22), we must find the correct orientations of each of the line segments $C_1, \ldots, C_4$. Once more, the purple arrows in the right part of Figure 5.10 depict outward normals from D. Turning leftward from each of these purple arrows results in the directions along $C_1, \ldots, C_4$ indicated in the figure. Note the positive orientations of $C_1, \ldots, C_4$ are again those given by traversing anticlockwise around D.

Examples 5.31 and 5.32 give a basic idea about how Green's theorem works. We now explore some situations where this theorem would be especially useful.

Example 5.33. Let us return to the setting of Example 5.31. More specifically, we let

$$D = \{(x, y) \in \mathbb{R}^2 \mid x^2 + y^2 < 1\}$$

be a circular disk, and we let C be its boundary,

$$C = \{(x, y) \in \mathbb{R}^2 \mid x^2 + y^2 = 1\},$$

with the positive (i.e. anticlockwise) orientation in relation to D.

Consider now the vector field $\mathbf{F}$ on $\mathbb{R}^2$ given by

$$\mathbf{F}(x, y) = (F_1(x, y), F_2(x, y))_{(x,y)} = (x^2, y^4)_{(x,y)},$$

and suppose you want to compute the curve integral

$$\int_C \mathbf{F} \cdot d\mathbf{s}.$$

Of course, one way to do this is to calculate it directly using Theorem 3.77. As you know, this involves parametrising C appropriately and then computing a single integral.

Green's theorem provides an alternative—and in this case, much easier—method for calculating this integral. Indeed, from the equation (5.20), we see that

$$\begin{aligned}\int_C \mathbf{F} \cdot d\mathbf{s} &= \iint_D [\partial_1 F_2(x, y) - \partial_2 F_1(x, y)]\, dx dy \\ &= \iint_D [\partial_x(y^4) - \partial_y(x^2)]\, dx dy.\end{aligned}$$

In particular, the integrand $\partial_x(y^4) - \partial_y(x^2)$ vanishes, and we immediately obtain

$$\int_C \mathbf{F} \cdot d\mathbf{s} = 0.$$

Remark 5.34. We note *it is not intuitively obvious that the integral in Example 5.33 should be equal to zero*. If you plot the values of $\mathbf{F}$ on C, then you will see it is generally not pointing in directions normal to C. Nonetheless, there is a deeper cancellation occurring within this integral, and this is implicitly captured through Green's theorem.

To better understand this, let us consider some point of C at which one is moving rightward. Here, we are integrating x^2—the x-component of $\mathbf{F}(x, y)$—in this rightward direction. The idea is that as C is a loop, there must be another point of C with the same

x-coordinate at which one is going leftward. Here, this leftward integral of x^2 will cancel with the previous rightward integral of x^2. More generally, *each point where we integrate x^2 rightward is cancelled by some corresponding point in which we integrate x^2 leftward.*

A similar argument can be made for vertical motion along C, with the y-component y^4 of $\mathbf{F}(x,y)$. As a result, the total integral of $\mathbf{F}$ over C must be 0.

For our next application, let us consider the region $D \subseteq \mathbb{R}^2$ drawn in Figure 5.11. Imagine that D represents the surface of a lake, and suppose you wish to find its area. Normally, to do this, you would probably take a boat out into the lake and make various measurements as you sail about.

Now, suppose D is also full of ravenous giant sharks that have eaten your boat, so you can no longer venture out onto the lake! Would you still be able to somehow find the area of the lake, even though you are landbound?

In more abstract terms, the above question can be formulated as follows:

Question 5.35. Can we measure the area of a region $D \subseteq \mathbb{R}^2$ (say, the one from Figure 5.11) using only information gathered at the boundary of D?

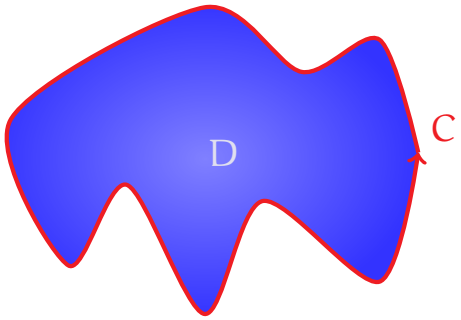


FIGURE 5.11. Corollary 5.36 gives a way to compute the area of the (blue) region D using only the information on its (red) boundary curve C.

The key idea here is to look at the vector field $\mathbf{A}$ on $\mathbb{R}^2$, given by

$$\mathbf{A}(x,y) = \frac{1}{2} \cdot (-y,x)_{(x,y)}. \tag{5.23}$$

In particular, if we were to apply Green's theorem, with this vector field $\mathbf{F} = \mathbf{A}$ and with the region D in Figure 5.11, then a direct computation yields

$$\begin{aligned} \int_C \mathbf{A} \cdot d\mathbf{s} &= \iint_D \left[\partial_x \left(\frac{1}{2} x \right) - \partial_y \left(-\frac{1}{2} y \right) \right] dx dy \\ &= \iint_D 1 \, dx dy. \end{aligned}$$

The above tells us that the area of D can be expressed in terms of an integral along its boundary curve C. In other words, we need not venture into the lake D in order to measure its area! We need only go one lap around the edge C of the lake and sum up the information read from $\mathbf{A}$ while doing so. (Observe that $\mathbf{A}$ contains information about one's horizontal and vertical positions as one travels around the lake.)

Repeating the above for more general regions, we obtain the following result:

Corollary 5.36. Let D and $C_1, C_2, \ldots, C_k$ be as in the statement of Theorem 5.29. Moreover, let $\mathbf{A}$ be the vector field on $\mathbb{R}^2$ given by (5.23). Then,

$$\mathcal{A}(D) = \int_{C_1} \mathbf{A} \cdot d\mathbf{s} + \cdots + \int_{C_k} \mathbf{A} \cdot d\mathbf{s}. \tag{5.24}$$

Remark 5.37. We mention that $\mathbf{A}$ is not the only vector field we can use here. For example, if we define another pair of vector fields $\mathbf{A}_x$ and $\mathbf{A}_y$ on $\mathbb{R}^2$ via the formulas

$$\mathbf{A}_x(x, y) = (0, x)_{(x,y)}, \qquad \mathbf{A}_y(x, y) = (-y, 0)_{(x,y)},$$

then one can see that *Corollary 5.36 still holds when $\mathbf{A}$ is replaced by either $\mathbf{A}_x$ or $\mathbf{A}_y$.*

Remark 5.38. Unfortunately, we lack the time in this module to discuss the proof of Green's theorem. However, let us mention that Green's theorem can be derived as a special case of the divergence theorem, which we will discuss later on.

5.6. **Divergence and Curl.** We now take a brief detour from our study of integral theorems to define two additional differential operations on vector fields: the *divergence* and *curl*. As we will soon see, these operations represent very special derivatives and will play prominent roles in upcoming integral theorems.

First, we look at the divergence, which maps vector fields to scalar functions:

Definition 5.39. Let $U \subseteq \mathbb{R}^n$ be open and connected. In addition, let $\mathbf{F}$ be a smooth vector field on U, and write $\mathbf{F}$ in terms of its components as

$$\mathbf{F}(\mathbf{p}) = (F_1(\mathbf{p}), F_2(\mathbf{p}), \ldots, F_n(\mathbf{p}))_{\mathbf{p}}, \qquad \mathbf{p} \in U. \tag{5.25}$$

We define the divergence of $\mathbf{F}$, denoted $\nabla \cdot \mathbf{F}$ or $\operatorname{div} \mathbf{F}$, to be the (real-valued) function

$$(\nabla \cdot \mathbf{F}) : U \to \mathbb{R}, \qquad (\nabla \cdot \mathbf{F})(\mathbf{p}) = \partial_1 F_1(\mathbf{p}) + \partial_2 F_2(\mathbf{p}) + \cdots + \partial_n F_n(\mathbf{p}). \tag{5.26}$$

Remark 5.40. A helpful way to remember (5.26) is to write it as a "dot product",

$$\nabla \cdot \mathbf{F} = (\partial_1, \partial_2, \ldots, \partial_n) \cdot (F_1, F_2, \ldots, F_n).$$

The above is also the inspiration for the notation "$\nabla \cdot \mathbf{F}$".

Observe that although $\mathbf{F}(\mathbf{p})$ is an arrow starting at $\mathbf{p}$, the corresponding divergence $(\nabla \cdot \mathbf{F})(\mathbf{p})$ *is a scalar*. Thus, do make sure that you are always dealing with the correct type of object in your computations, so that you avoid getting confused!

Example 5.41. Consider the vector field $\mathbf{F}$ on $\mathbb{R}^2$ given by

$$\mathbf{F}(x, y) = (F_1(x, y),\, F_2(x, y))_{(x,y)} = (x, y)_{(x,y)}.$$

See the left illustration in Figure 5.12 for a partial plot of $\mathbf{F}$.

To find the divergence of $\mathbf{F}$, we simply apply Definition 5.39 to each $(x, y) \in \mathbb{R}^2$:

$$\begin{aligned}(\nabla \cdot \mathbf{F})(x, y) &= \partial_1 F_1(x, y) + \partial_2 F_2(x, y)\\ &= \partial_x(x) + \partial_y(y)\\ &= 2.\end{aligned}$$

Thus, $\nabla \cdot \mathbf{F}$ is a constant function, and the divergence of $\mathbf{F}$ at every point is 2.

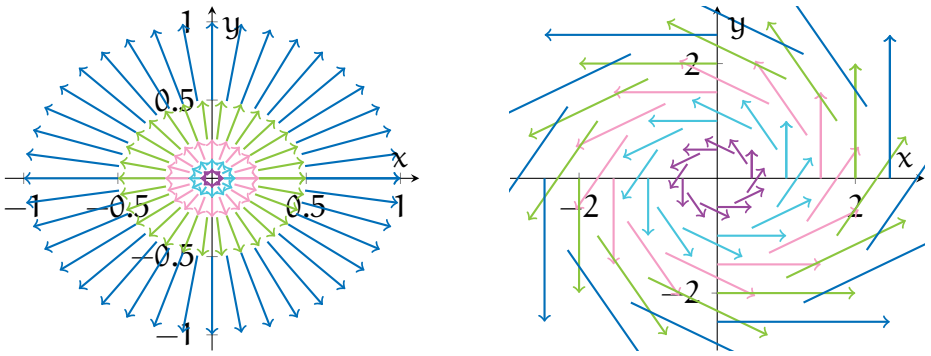


FIGURE 5.12. The arrows in the left plot represent several values of the vector field $\mathbf{F}$ from Example 5.41; similarly, the arrows in the right plot indicate some values of $\mathbf{H}$ from Example 5.42.

Example 5.42. Next, let $\mathbf{H}$ be the vector field on $\mathbb{R}^2$ given by

$$\mathbf{H}(x, y) = (-y, x)_{(x,y)}.$$

(See the right plot in Figure 5.12.) Again, we can compute $\nabla \cdot \mathbf{H}$ using Definition 5.39:

$$\begin{aligned}(\nabla \cdot \mathbf{H})(x, y) &= \partial_x(-y) + \partial_y(x)\\ &= 0.\end{aligned}$$

In particular, $\nabla \cdot \mathbf{H}$ vanishes everywhere, i.e. $\mathbf{H}$ is *divergence-free*.

Example 5.43. For a 3-dimensional example, we define the vector field $\mathbf{Y}$ on $\mathbb{R}^3$ by

$$\mathbf{Y}(x,y,z) = (xy,\, x^2+yz,\, z^4+xyz)_{(x,y,z)}.$$

Once again, applying Definition 5.39, we obtain, for any $(x,y,z)\in\mathbb{R}^3$,

$$\begin{aligned}(\nabla\cdot\mathbf{Y})(x,y,z) &= \partial_x(xy)+\partial_y(x^2+yz)+\partial_z(z^4+xyz)\\ &= y+z+4z^3+xy.\end{aligned}$$

Examples 5.41–5.43 show that the divergence is rather simple to compute using (5.26). On the other hand, (5.26) fails to shed much light on how the divergence should be interpretated. Thus, let us also pose the following question here:

Question 5.44. What is the intuitive meaning of the divergence of a vector field?

Unfortunately, we cannot address this question until a bit later, after discussing the *divergence theorem*. As a result, we defer our answer to Question 5.44 until then.

Next, we give an analogous exposition for the curl of a 3-dimensional vector field:

Definition 5.45. Let $U\subseteq\mathbb{R}^3$ be open and connected. In addition, let $\mathbf{F}$ be a smooth vector field on U, and write $\mathbf{F}$ in terms of its components as

$$\mathbf{F}(\mathbf{p}) = (F_1(\mathbf{p}),F_2(\mathbf{p}),F_3(\mathbf{p}))_{\mathbf{p}},\qquad \mathbf{p}\in U. \tag{5.27}$$

We define the curl of $\mathbf{F}$, denoted $\nabla\times\mathbf{F}$ or $\operatorname{curl}\mathbf{F}$, to be the vector field on U given by

$$(\nabla\times\mathbf{F})(\mathbf{p}) = (\partial_2F_3(\mathbf{p})-\partial_3F_2(\mathbf{p}),\,\partial_3F_1(\mathbf{p})-\partial_1F_3(\mathbf{p}),\,\partial_1F_2(\mathbf{p})-\partial_2F_1(\mathbf{p}))_{\mathbf{p}}. \tag{5.28}$$

Remark 5.46. The formula (5.28) can be a serious pain to remember! One trick to recall (5.28) correctly (which also inspires the notation "$\nabla\times\mathbf{F}$") is to informally write

$$(\nabla\times\mathbf{F})(\mathbf{p}) = \det\begin{bmatrix}\mathbf{i} & \mathbf{j} & \mathbf{k}\\ \partial_1 & \partial_2 & \partial_3\\ F_1 & F_2 & F_3\end{bmatrix}.$$

Expanding the right-hand side as one would a 3×3 determinant results in (5.28).

Equation (5.28) tells us that, like for the divergence, the curl can be computed by taking a bunch of partial derivatives. However, unlike the divergence, the value $(\nabla\times\mathbf{F})(\mathbf{p})$ is *an arrow starting from* $\mathbf{p}$, not a scalar value.

Another point to keep in mind is that *the curl is only defined for vector fields in 3-dimensional space*. (In particular, please do refrain from trying to compute curls

of 2-dimensional vector fields.) This again differs from the divergence, which is well-defined through (5.26) for vector fields in any dimension.

Example 5.47. Let $\mathbf{X}$ be the vector field on $\mathbb{R}^3$ defined by

$$\begin{aligned}\mathbf{X}(x,y,z) &= (X_1(x,y,z),\, X_2(x,y,z),\, X_3(x,y,z))_{(x,y,z)}\\ &= (yz, xz, xy)_{(x,y,z)}.\end{aligned}$$

Let us now find the curl of $\mathbf{X}$.

According to Definition 5.45, the components of $\nabla \times \mathbf{X}$ are given by

$$\begin{aligned}\partial_2 X_3(x,y,z) - \partial_3 X_2(x,y,z) &= \partial_y(xy) - \partial_z(xz)\\ &= x - x\\ &= 0,\\ \partial_3 X_1(x,y,z) - \partial_1 X_3(x,y,z) &= \partial_z(yz) - \partial_x(xy)\\ &= y - y\\ &= 0,\\ \partial_1 X_2(x,y,z) - \partial_2 X_1(x,y,z) &= \partial_x(xz) - \partial_y(yz)\\ &= z - z\\ &= 0.\end{aligned}$$

As a result, by (5.28), we conclude that $\mathbf{X}$ is *curl-free*:

$$(\nabla \times \mathbf{X})(x,y,z) = (0,0,0)_{(x,y,z)}, \qquad (x,y,z) \in \mathbb{R}^3.$$

Example 5.48. Let us return to the vector field $\mathbf{Y}$ from Example 5.43:

$$\begin{aligned}\mathbf{Y}(x,y,z) &= (Y_1(x,y,z),\, Y_2(x,y,z),\, Y_3(x,y,z))_{(x,y,z)}\\ &= (xy, x^2 + yz, z^4 + xyz)_{(x,y,z)}.\end{aligned}$$

By Definition 5.45, the components of $\nabla \times \mathbf{Y}$ satisfy

$$\begin{aligned}\partial_2 Y_3(x,y,z) - \partial_3 Y_2(x,y,z) &= \partial_y(z^4 + xyz) - \partial_z(x^2 + yz)\\ &= xz - y,\\ \partial_3 Y_1(x,y,z) - \partial_1 Y_3(x,y,z) &= \partial_z(xy) - \partial_x(z^4 + xyz)\\ &= -yz,\end{aligned}$$

$$\begin{aligned}\partial_1 Y_2(x,y,z) - \partial_2 Y_1(x,y,z) &= \partial_x(x^2+yz) - \partial_y(xy)\\ &= x.\end{aligned}$$

Therefore, we obtain, for the curl of $\mathbf{Y}$,

$$(\nabla \times \mathbf{Y})(x,y,z) = (xz-y,-yz,x)_{(x,y,z)}, \qquad (x,y,z)\in\mathbb{R}^3.$$

On the other hand, Definition 5.45 reveals little about what the curl means:

Question 5.49. What is the intuitive interpretation of the curl of a vector field?

Once again, we cannot answer this question until later, after we have understood *Stokes' theorem*. Thus, we will return to address Question 5.49 at that point.

5.7. **Stokes' Theorem.** Recall that Green's theorem (Theorem 5.29) related an integral over a 2-dimensional region $D\subseteq\mathbb{R}^2$ with an integral over its 1-dimensional boundary—see Figures 5.10 and 5.11. Our next task is to consider the following:

Question 5.50. Could Theorem 5.29 be further generalised, so that the region $D\subseteq\mathbb{R}^2$ is replaced by a surface $S\subseteq\mathbb{R}^3$? If so, then what is the corresponding formula?

The answer to Question 5.50 is given by an integral identity known as *Stokes' theorem*, primarily attributed to George Stokes (British mathematician and physicist, 1819–1903) and Lord Kelvin (Scottish mathematician and physicist, 1824–1907). (The latter is perhaps better known for the unit of temperature bearing his name.)

Theorem 5.51 (Stokes' theorem)**.** Let $S\subseteq\mathbb{R}^3$ be a bounded and oriented surface, and assume the boundary of S is comprised of a union of curves $C_1, C_2, \dots, C_k$. In addition, let $\mathbf{F}$ be a smooth vector field that is defined on S and its boundary. Then,

$$\iint_S (\nabla\times\mathbf{F})\cdot d\mathbf{A} = \int_{C_1}\mathbf{F}\cdot d\mathbf{s} + \cdots + \int_{C_k}\mathbf{F}\cdot d\mathbf{s}, \tag{5.29}$$

where $C_1,\dots,C_k$ are positively oriented with respect to the chosen orientation of S:

- For each $1\le i\le k$, we choose the orientation of C_i obtained by turning *leftward from the outward-pointing unit normal direction* from S, whenever S is *viewed from the side that corresponds to its chosen orientation*.

Note that Theorem 5.51 also fits into the framework described by (5.18). Here, the 2-dimensional object A is the surface S, while the "derivative" is given by the curl.

Similar to our discussions for Theorem 5.29, we omit formal definitions of boundaries and positive orientations. Instead, we demonstrate these through examples.

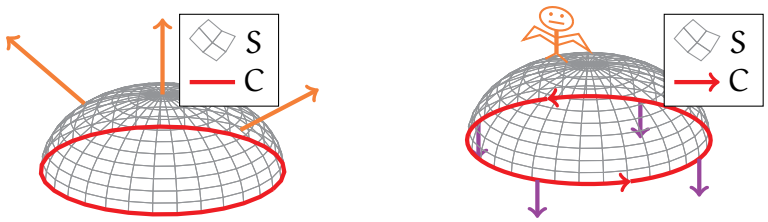


FIGURE 5.13. The left drawing shows the half-sphere S from Example 5.52, as well as its boundary C (in red); moreover, the orientation of S is indicated by some orange arrows. In the right drawing, we add some instances of outward normals to S (drawn as purple arrows), and we indicate the orientation of C that is generated from applying Stokes' theorem to S.

Example 5.52. Let S denote the upper half-sphere,

$$S = \{(x, y, z) \in \mathbb{R}^3 \mid x^2 + y^2 + z^2 = 1,\ z > 0\}.$$

See the left illustration in Figure 5.13 for a plot of S. In addition, let us assign to S its *upward-facing orientation*; this is indicated in Figure 5.13 by orange arrows.

Observe that the boundary of S, where its points "terminate", is the equatorial circle

$$C = \{(x, y, 0) \in \mathbb{R}^3 \mid x^2 + y^2 = 1\},$$

drawn in red in Figure 5.13. Applying Theorem 5.51 to this S and an appropriate vector field $\mathbf{F}$ (satisfying the assumptions in Theorem 5.51), we obtain

$$\iint_S (\nabla \times \mathbf{F}) \cdot d\boldsymbol{A} = \int_C \mathbf{F} \cdot d\mathbf{s}. \tag{5.30}$$

To make full sense of (5.30), we still must describe the correct orientation for C. For this, you can imagine yourself floating "over" S, *on the side given by its orientation*. Since we have chosen the upward-facing orientation of S, we should be looking down on this upward-facing side of S, like the orange bat-creature in the right side of Figure 5.13.

Now, some outward-facing normals to S are drawn in purple in the right part of Figure 5.13. If we look down at these purple arrows from the position of the orange bat-creature in the figure, and if we turn left from these purple arrows, then we would be facing the direction of C that is indicated in this figure (i.e. anticlockwise from the creature's point of view). This is the *positive orientation* of C described in Theorem 5.51.

As a cautionary remark, we mention that the orientation of S plays an essential role. Had we chosen the opposite orientation of S (so the bat-creature would float below S in Figure 5.13), the above exercise would have yielded the opposite orientation of C!

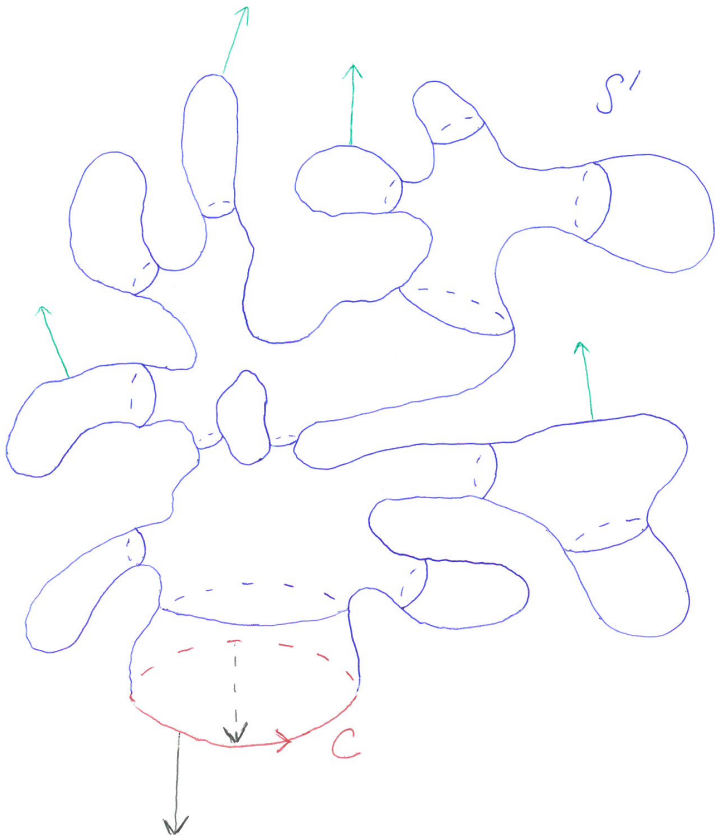


FIGURE 5.14. This drawing shows the surface S' (in blue) from Examples 5.53 and 5.54, along with its boundary C (in red). In addition, the green arrows indicate the outward-facing orientation of S', while the black arrows represent the outward-pointing normal direction from S'.

Example 5.53. Let C be the same circle as in Example 5.52. However, we attach to C a different, and more complicated, surface S', which is crudely illustrated in Figure 5.14. Like in Example 5.52, we select the *outward-facing orientation* of S' (see the green arrows in Figure 5.14). Suppose, for some perverse reason, we wish to compute

$$\iint_{S'} \mathbf{G} \cdot d\mathbf{A},$$

where $\mathbf{G}$ is the vector field on $\mathbb{R}^3$ given by the formula

$$\mathbf{G}(x, y, z) = (xz - y, -yz, x)_{(x,y,z)}.$$

At first glance, this seems utterly hopeless—see how terribly complicated S' looks! However, we are saved from certain despair thanks to two crucial observations:

- The boundary S' is simply C.

- $\mathbf{G}$ is the curl of the vector field $\mathbf{Y}$ from Example 5.48:

$$\mathbf{Y}(x, y, z) = (xy, x^2 + yz, z^4 + xyz)_{(x,y,z)}, \qquad (x, y, z) \in \mathbb{R}^3.$$

As a result, applying Theorem 5.51 to S' and $\mathbf{G}$, we see that

$$\begin{aligned} \iint_{S'} \mathbf{G} \cdot d\mathbf{A} &= \iint_{S'} (\nabla \times \mathbf{Y}) \cdot d\mathbf{A} \\ &= \int_C \mathbf{Y} \cdot d\mathbf{s}. \end{aligned}$$

Thus, the integral of $\mathbf{G}$ over S' is equal to a much simpler integral over C! Because $\mathbf{G}$ has this special curl structure, all the complexities arising from S' will cancel.

Example 5.54. Let S, S', C, $\mathbf{G}$, and $\mathbf{Y}$ be as in Examples 5.52 and 5.53. Recalling again that $\mathbf{G} = \nabla \times \mathbf{Y}$ and applying Theorem 5.51 twice (to S' and then to S), we obtain

$$\begin{aligned} \iint_{S'} \mathbf{G} \cdot d\mathbf{A} &= \int_C \mathbf{Y} \cdot d\mathbf{s} \\ &= \iint_S \mathbf{G} \cdot d\mathbf{A}. \end{aligned}$$

Again, although the integral over S' seems exceedingly complicated, Stokes' theorem tells us it is actually equal to an integral, of the same vector field, over the far simpler surface S. This is yet another use of Stokes' theorem to simplify computations.

For completeness, let us compute all three of the above integrals by computing the integral of $\mathbf{Y}$ over C. Consider the following parametrisation of C:

$$\gamma : (0, 2\pi) \to C, \qquad \gamma(t) = (\cos t, \sin t, 0).$$

Observe that γ is an injective parametrisation of all of C except for a single point $(1, 0, 0)$, and that γ generates the positive orientation of C obtained from Stokes' theorem. Thus,

$$\begin{aligned} \int_C \mathbf{Y} \cdot d\mathbf{s} &= \int_0^{2\pi} [\mathbf{Y}(\gamma(t)) \cdot \gamma'(t)_{\gamma(t)}] \, dt \\ &= \int_0^{2\pi} [(\cos t \sin t, \cos^2 t, 0) \cdot (-\sin t, \cos t, 0)] \, dt \\ &= \int_0^{2\pi} (-\sin^2 t \cos t + \cos^3 t) \, dt \\ &= \int_0^{2\pi} (\cos t - 2 \sin^2 t \cos t) \, dt \\ &= 0. \end{aligned}$$

As a result, combining all the above, we conclude that

$$\iint_{S'} \mathbf{G} \cdot d\boldsymbol{A} = 0, \qquad \iint_{S} \mathbf{G} \cdot d\boldsymbol{A} = 0, \qquad \int_{C} \mathbf{Y} \cdot d\boldsymbol{s} = 0.$$

Example 5.55. Let T denote the *torus* illustrated in the left half of Figure 5.15, and let $\mathbf{F}$ be a smooth vector field on $\mathbb{R}^3$. Observe that, unlike our previous examples, T *has no boundary*! Therefore, in this case, Stokes' theorem tells us that

$$\iint_{T} (\nabla \times \mathbf{F}) \cdot d\boldsymbol{A} = 0.$$

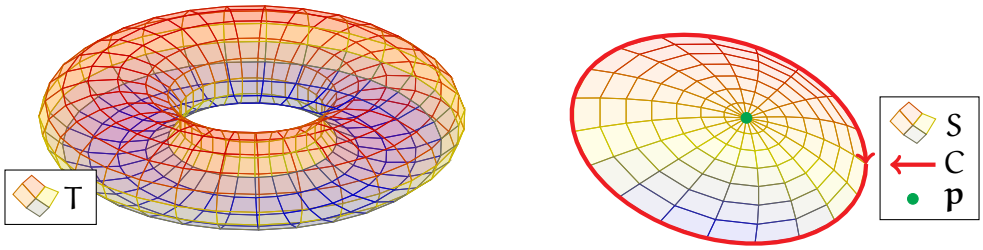


FIGURE 5.15. The left plot shows the boundaryless torus T from Example 5.55. The right drawing illustrates the meaning of the curl of a vector field $\mathbf{F}$; in particular, the total amount of $\nabla \times \mathbf{F}$ passing through S is equal to the amount that $\mathbf{F}$ circles around S along C.

Another application of Stokes' theorem is to address Question 5.49—the meaning of the curl of a vector field. Suppose $\mathbf{F}$ is a smooth vector field on $U \subseteq \mathbb{R}^3$, and fix $\mathbf{p} \in U$. Our goal, then, is to give an interpretation of the value $(\nabla \times \mathbf{F})(\mathbf{p})$.

Suppose C is any curve that circles around $\mathbf{p}$; see, for instance, the right graphic in Figure 5.15. Moreover, let S be a bounded surface that passes through $\mathbf{p}$ and has C as its boundary (again, see Figure 5.15). Applying Theorem 5.51, we have

$$\iint_{S} (\nabla \times \mathbf{F}) \cdot d\boldsymbol{A} = \int_{C} \mathbf{F} \cdot d\boldsymbol{s}, \tag{5.31}$$

as long as S and C are given appropriate orientations.

Let us now look more carefully at what each side of (5.31) tells us:

- The surface integral on the left-hand side of (5.31) represents the flux of the vector field $\nabla \times \mathbf{F}$ through S. Note that since S is situated near $\mathbf{p}$, the value of this integral measures a certain direction of $\nabla \times \mathbf{F}$ near $\mathbf{p}$.
- The curve integral on the right-hand side of (5.31) measures how much the vector field $\mathbf{F}$ points along C. Since C wraps around $\mathbf{p}$, then the value of this integral can be interpreted as the tendency of $\mathbf{F}$ to "circle around $\mathbf{p}$".

Also, keep in mind that these observations hold for *any* such curve C and surface S.

Thus, from the above, we arrive at the following rough interpretation: $(\nabla \times \mathbf{F})(\mathbf{p})$ *quantifies how, and how much, the vector field* $\mathbf{F}$ *"circles around* $\mathbf{p}$*" nearby* $\mathbf{p}$.

Remark 5.56. In fact, a more careful look at (5.31) yields a more precise conclusion: *the component of* $(\nabla \times \mathbf{F})(\mathbf{p})$ *in any direction* $\mathbf{X} \in T_p\mathbb{R}^3$ *tells us how much* $\mathbf{F}$ *"circles around* $\mathbf{p}$*" nearby* $\mathbf{p}$ *along the plane that is perpendicular to* $\mathbf{X}$.

5.8. **The Divergence Theorem.** Green's and Stokes' theorems related integrals over 2-dimensional objects to integrals over their 1-dimensional boundaries. In the next (and last) integral identity, we again raise the dimension—we connect integrals over 3-*dimensional regions* to integrals over their 2-*dimensional boundaries*.

The classical result for this setting is known as the *divergence theorem*. Its first versions were primarily attributed to *Carl Friedrich Gauss* (German mathematician and physicist, 1777–1855) and *Joseph-Louis Lagrange* (French-Italian mathematician and physicist, 1736–1817). (In case you were wondering, "Italian" is not a mistake; Lagrange's birth name was Giuseppe Luigi Lagrangia.) Others, including George Green, have also made contributions toward the theorem in later years.

We now give a precise statement of the divergence theorem:

Theorem 5.57 (Divergence theorem)**.** Let $W \subseteq \mathbb{R}^3$ be open and bounded, and assume the boundary of W consists of a union of surfaces $S_1, S_2, \ldots, S_k$. In addition, let $\mathbf{F}$ be a smooth vector field that is defined on W and its boundary. Then,

$$\iiint_W (\nabla \cdot \mathbf{F})\, dV = \iint_{S_1} \mathbf{F} \cdot d\mathbf{A} + \cdots + \iint_{S_k} \mathbf{F} \cdot d\mathbf{A}, \tag{5.32}$$

where the surfaces $S_1, \ldots, S_k$ are given the *outward-facing* (from W) *orientation*.

Again, rather than delving into the formal definitions of boundaries and outward orientations, we instead demonstrate Theorem 5.57 through examples:

Example 5.58. As usual, let $\mathbb{S}^2$ denote the *unit sphere* centred at the origin,

$$\mathbb{S}^2 = \{(x, y, z) \in \mathbb{R}^3 \mid x^2 + y^2 + z^2 = 1\},$$

here with the *outward-facing orientation*, and let B denote the interior of $\mathbb{S}^2$,

$$B = \{(x, y, z) \in \mathbb{R}^2 \mid x^2 + y^2 + z^2 < 1\}.$$

Note that the boundary of B is precisely $\mathbb{S}^2$; see the left illustration in Figure 5.16.

Suppose now that $\mathbf{F}$ is the vector field from Example 4.78:

$$\mathbf{F}(x,y,z) = (x,y,z)_{(x,y,z)}, \qquad (x,y,z) \in \mathbb{R}^3.$$

Let us now apply the divergence theorem to compute the integral of $\mathbf{F}$ over $\mathbb{S}^2$.

First, using (5.26), we compute

$$\begin{aligned}(\nabla \cdot \mathbf{F})(x,y,z) &= \partial_x x + \partial_y y + \partial_z z \\ &= 3.\end{aligned}$$

Applying Theorem 5.57 to the above B and $\mathbf{F}$ then yields

$$\begin{aligned}\iint_{\mathbb{S}^2} \mathbf{F} \cdot d\mathbf{A} &= \iiint_B (\nabla \cdot \mathbf{F})\, dV \\ &= 3 \iiint_B dV.\end{aligned}$$

In other words, the integral of $\mathbf{F}$ over $\mathbb{S}^2$ is simply three times the volume of B.

Recall that the volume of B, which is a ball of radius 1, is simply $\frac{4\pi}{3} \cdot 1^3 = \frac{4\pi}{3}$. (This result could be derived with a bit of multivariable calculus—in particular, by changing to spherical coordinates.) As a result, we obtain

$$\iint_{\mathbb{S}^2} \mathbf{F} \cdot d\mathbf{A} = 3 \cdot \frac{4\pi}{3} = 4\pi,$$

which is the same answer as was obtained in Example 4.79.

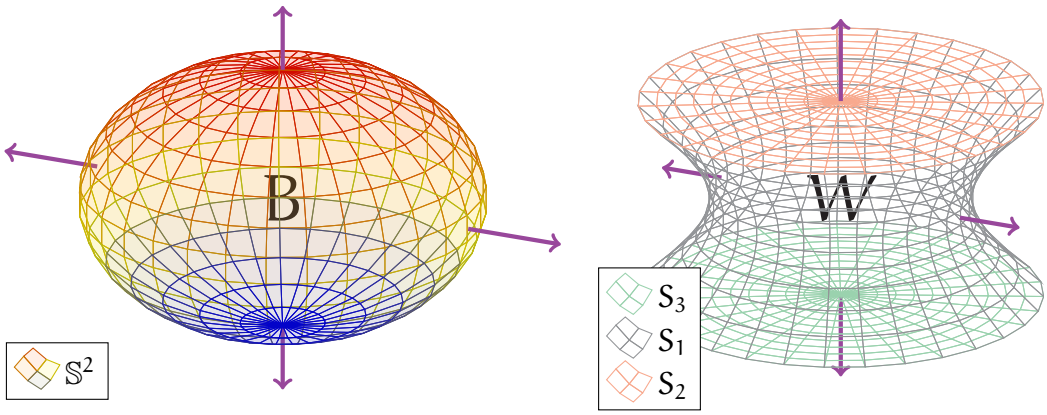


FIGURE 5.16. The left illustration shows the setting of Example 5.58, where the region B is the inside of the drawn sphere $\mathbb{S}^2$. Similarly, the right illustration shows the setting of Example 5.59; the region W is the interior of the surfaces S_1, S_2, S_3 drawn there. In both graphics, the orientations of the boundary surfaces are represented by purple arrows.

Example 5.59. Next, consider the graphic in the right half of Figure 5.16:

- S_1 is the "wormhole-like" surface (in grey) that looks like a squished cylinder.
- S_2 is the top "lid" (in orange) attached to the upper end of S_1.
- S_3 is the bottom "lid" (in green) attached to the lower end of S_1.

Let $W \subseteq \mathbb{R}^3$ be the region that is bounded by (i.e. in the interior of) S_1, S_2, and S_3. In particular, the boundary of W is exactly given by the surfaces S_1, S_2, S_3.

Then, applying Theorem 5.57 to this V, we obtain

$$\iiint_W (\nabla \cdot \mathbf{F})\, dx\,dy\,dz = \iint_{S_1} \mathbf{F} \cdot d\mathbf{A} + \iint_{S_2} \mathbf{F} \cdot d\mathbf{A} + \iint_{S_3} \mathbf{F} \cdot d\mathbf{A},$$

for any smooth vector field $\mathbf{F}$ that is defined on W and its boundary. Here, S_1, S_2, and S_3 are all given the *outward-facing orientation* (see the purple arrows in Figure 5.16).

Another application of the divergence theorem is that it gives an interpretation of the divergence of a vector field, i.e. it can be used to answer Question 5.44.

For this purpose, let $\mathbf{F}$ be a vector field on some $U \subseteq \mathbb{R}^3$, and fix $\mathbf{p} \in U$. Moreover, let $S \subseteq U$ be a surface that encloses $\mathbf{p}$, and let $W \subseteq U$ denote the interior of S; this is shown in the left part of Figure 5.17. In this setting, Theorem 5.57 implies

$$\iiint_W (\nabla \cdot \mathbf{F})\, dV = \iint_S \mathbf{F} \cdot d\mathbf{A}, \tag{5.33}$$

where S is given the outward-facing orientation.

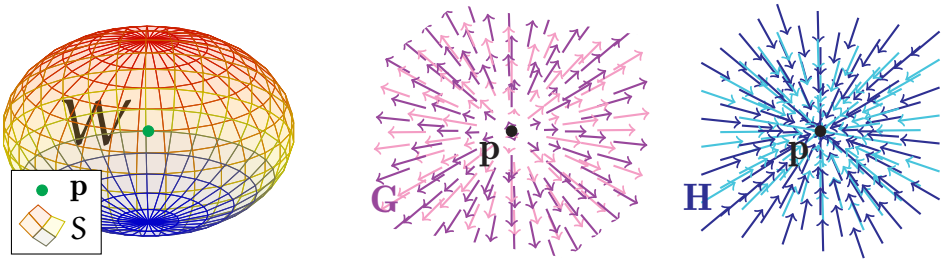


FIGURE 5.17. The left graphic demonstrates the meaning of the divergence of a vector field $\mathbf{F}$; the total measure of $\nabla \cdot \mathbf{F}$ on W is equal to the amount that $\mathbf{F}$ passes outward through the surface S. The middle and right illustrations show some values of two vector fields $\mathbf{G}$ and $\mathbf{H}$ (see Example 5.60). In particular, observe that $(\nabla \cdot \mathbf{G})(\mathbf{p}) > 0$ and $(\nabla \cdot \mathbf{H})(\mathbf{p}) < 0$.

We can now make sense of each side of (5.33):

- The left-hand side of (5.33) measures the total divergence of $\mathbf{F}$ on W. Since W contains $\mathbf{p}$, this integral can be viewed as measuring $\nabla \cdot \mathbf{F}$ near $\mathbf{p}$.

- The right-hand side of (5.33) measures the outward flux of $\mathbf{F}$ through S—the degree to which the arrows of $\mathbf{F}$ are leaving S. Since S encloses $\mathbf{p}$, this can be viewed as a measure of how much $\mathbf{F}$ is pointing away from $\mathbf{p}$.

In addition, keep in mind that (5.33) holds for *any* such surface S and interior W.

Thus, the divergence can be roughly interpreted as follows: $(\nabla \cdot \mathbf{F})(\mathbf{p})$ *quantifies how much* $\mathbf{F}$ *is "pointing away from* $\mathbf{p}$*" nearby* $\mathbf{p}$. If the values of $\mathbf{F}$ near $\mathbf{p}$ are pointing away from $\mathbf{p}$, then $(\nabla \cdot \mathbf{F})(\mathbf{p})$ will be positive. On the other hand, if the values of $\mathbf{F}$ near $\mathbf{p}$ are pointing toward $\mathbf{p}$, then $(\nabla \cdot \mathbf{F})(\mathbf{p})$ will be negative.

Example 5.60. Consider the middle and right drawings of Figure 5.17, which illustrate some values of two vector fields $\mathbf{G}$ (in purple) and $\mathbf{H}$ (in blue), respectively.

Now, since the purple arrows in the middle drawing are pointing outward and away from $\mathbf{p}$, we deduce that $(\nabla \cdot \mathbf{G})(\mathbf{p}) > 0$. Similarly, since the blue arrows in the right drawing are all pointing inward toward $\mathbf{p}$, we thus conclude that $(\nabla \cdot \mathbf{H})(\mathbf{p}) < 0$.

We also note that Theorem 5.57 can be generalised to all dimensions: roughly, *for an open and bounded* $D \subseteq \mathbb{R}^n$*, the integral over* D *of* $\nabla \cdot \mathbf{F}$ *is equal to the integral of the outward normal component of* $\mathbf{F}$ *over the boundary of* D. We cannot give a precise statement of this here, since we have not discussed higher-dimensional geometric objects. However, we can give a formal statement in the case of $n = 2$:

Theorem 5.61 (Divergence theorem)**.** Let $D \subseteq \mathbb{R}^2$ be open and bounded, and assume the boundary of D consists of a union of curves $C_1, C_2, \ldots, C_k$. In addition, let $\mathbf{F}$ be a smooth vector field that is defined on D and its boundary. Then,

$$\iint_D (\nabla \cdot \mathbf{F})\, dA = \int_{C_1} (\mathbf{F} \cdot \mathbf{n})\, ds + \cdots + \int_{C_k} (\mathbf{F} \cdot \mathbf{n})\, ds, \tag{5.34}$$

where $\mathbf{n}$ denotes the *outward-pointing* (from D) unit normal to $C_1, \ldots, C_k$.

Remark 5.62. In fact, one can use Theorem 5.61 to prove Theorem 5.29, and vice versa.

Remark 5.63. Moreover, when $n = 1$, the divergence theorem reduces to (5.15).

Finally, let us mention that there is a *general* integral theorem that includes all the identities we have studied—the fundamental theorem of calculus, Green's theorem, Stokes' theorem, and the divergence theorem—as special cases. This all-encompassing result is called the *generalised Stokes' theorem* and is most succinctly expressed as

$$\int_M d\alpha = \int_{\partial M} \alpha. \tag{5.35}$$

Although we certainly do not have the time nor the background to discuss (5.35) in detail, let us at least summarise what every object in (5.35) means:

- M is the n-dimensional geometric object (a *manifold*) over which we integrate.
- ∂M denotes the $(n-1)$-dimensional boundary of M.
- The integrand α is a special object known as a *differential form*.
- $d\alpha$ denotes a special "derivative" (called an *exterior derivative*) of α.

The generalised Stokes' theorem is a topic that one encounters in a more advanced differential geometry module. If you continue your studies in geometry and analysis, then you will most likely come across these topics in the future!

5.9. **Some Equations of Physics.** Historically, Stokes' theorem and the divergence theorem have played essential roles in the development of physics.

One well-known example of this comes from classical electromagnetism. For this discussion, we consider two vector fields on $\mathbb{R}^3$:

- The *electric field* $\mathbf{E}$, with $\mathbf{E}(\mathbf{p})$ representing the electric force at $\mathbf{p}$.
- The *magnetic field* $\mathbf{B}$, with $\mathbf{B}(\mathbf{p})$ representing the magnetic force at $\mathbf{p}$.

Two of the fundamental laws of electromagnetism, from the early 1800s, are as follows:

(1) *Faraday's law:* The change in time of the magnetic flux through a surface S is equal, up to sign, to the electric force around its boundary curve C (see the left graphic in Figure 5.18 for an illustration of S and C):

$$-\frac{\partial}{\partial t}\iint_S \mathbf{B}\cdot d\mathbf{A} = \int_C \mathbf{E}\cdot d\mathbf{s}. \tag{5.36}$$

(2) *Gauss's law:* The outward electric flux through a closed and bounded surface S is equal to the total charge enclosed in the region W inside S:

$$\begin{aligned}\iint_S \mathbf{E}\cdot d\mathbf{A} &= Q(W)\\ &= \iiint_W \rho\, dV.\end{aligned} \tag{5.37}$$

Here, $Q(W)$ denotes the total charge inside W, and ρ represents the *charge density*. (See the right graphic in Figure 5.18 for an illustration.)

Remark 5.64. There are two additional electromagnetic laws—*Ampère's law* and *Gauss's law for magnetism*—which have forms similar to (5.36) and (5.37). However, to keep our discussions brief, we restrict our attention only to Faraday's and Gauss's laws.

Remark 5.65. More precisely, the vector fields $\mathbf{E}$ and $\mathbf{B}$ should also depend on a fourth variable, t, representing the time. For simplicity, we omit this from the notation.

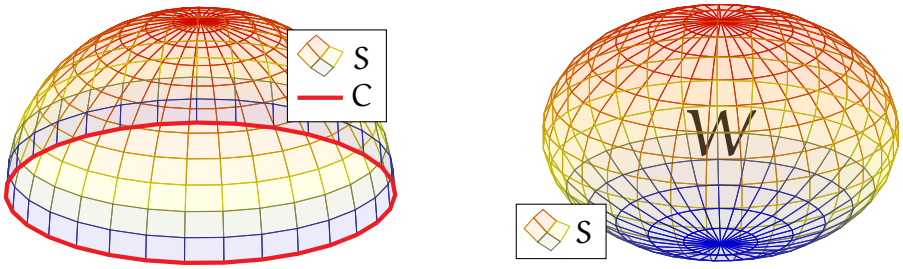


FIGURE 5.18. The left illustration contains the setting of Faraday's law, which involves a bounded surface S and its boundary curve C. The right drawing shows the setting of Gauss's law, which involves a closed and bounded surface S along with the region W enclosed by S.

Applying Stokes' theorem to the right-hand side of Faraday's law, (5.36), we obtain

$$-\frac{\partial}{\partial t}\iint_S \mathbf{B}\cdot d\mathbf{A} = \iint_S (\nabla\times\mathbf{E})\cdot d\mathbf{A}, \tag{5.38}$$

for *any* such bounded surface $S\subseteq\mathbb{R}^3$. Similarly, applying the divergence theorem to the left-hand side of Gauss's law, (5.37), yields the integral identity

$$\iiint_W (\nabla\cdot\mathbf{E})\,dV = \iiint_W \rho\,dV, \tag{5.39}$$

again for *any* such region $W\subseteq\mathbb{R}^3$.

Since (5.38) and (5.39) hold for *every* possible surface S and region W, one can, in fact, conclude that the integrands in these identities are equal to each other. (We refrain from giving a proof of this here.) This results in the *differential* identities,

$$-\frac{\partial}{\partial t}\mathbf{B} = \nabla\times\mathbf{E}, \qquad \nabla\cdot\mathbf{E} = \rho, \tag{5.40}$$

which are two of the four *Maxwell's equations*. (The other two equations are similarly obtained using Ampère's law and Gauss's law for magnetism.)

The equalities in (5.40) are examples of *partial differential equations*. One advantage of writing the laws of electromagnetism in this differential form is that one can proceed to *solve* these equations for $\mathbf{E}$ and $\mathbf{B}$, for all positions and times. This allows one to *predict* the future behaviours of our electric and magnetic fields.

Remark 5.66. If you are interested in partial differential equations, then you should certainly consider the third year module *MTH6151: Partial Differential Equations*.

As an additional example, we play a similar game with fluid mechanics. Suppose we have a fluid, which we mathematically model using the following:

- A scalar function $\rho : \mathbb{R}^3 \to \mathbb{R}$, with each value $\rho(x, y, z)$ representing the *mass density* of the fluid particle at position (x, y, z).
- A vector field $\boldsymbol{v}$ on $\mathbb{R}^3$, such that each tangent vector $\boldsymbol{v}(x, y, z)$ represents the *velocity* of the fluid particle at position (x, y, z).

In particular, the total mass of the fluid within a bounded region $W \subseteq \mathbb{R}^3$ is given by

$$M(W) = \iiint_W \rho \, dV.$$

Now, the guiding physical principle in our discussion is the following:

- *Continuity equation:* The rate of change of the mass of the fluid in a region $W \subseteq \mathbb{R}^3$ is equal to the total flux of the fluid into W through its boundary S:

$$\frac{\partial}{\partial t} \iiint_W \rho \, dV = -\iint_S \rho \boldsymbol{v} \cdot d\boldsymbol{A}. \tag{5.41}$$

Note the left-hand side of (5.41) indeed captures the change in the total fluid mass in W. In the right-hand side, S is given the *outward-facing* (from W) *orientation*, hence the total fluid flow *into* W is given by the *negative* of the flux integral through S. See Figure 5.19 for an illustration of the setting of (5.41).

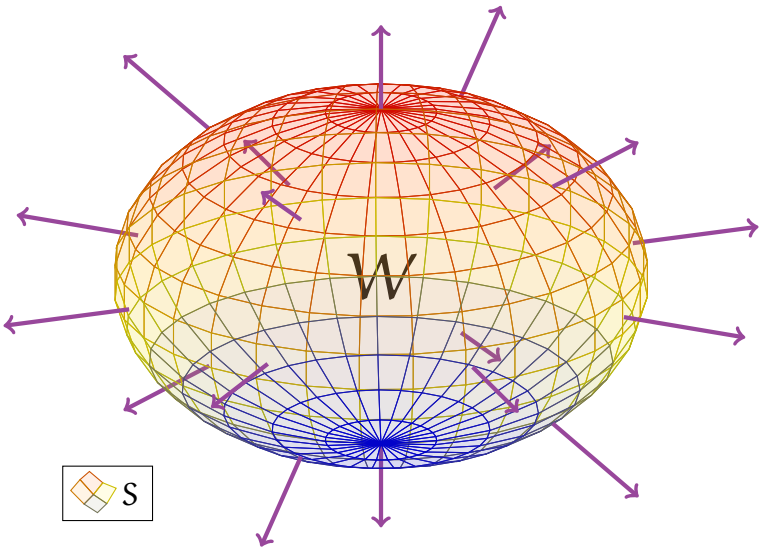


FIGURE 5.19. The illustration shows a fluid flowing out of a spherical region W through its boundary surface S; this is indicated by purple arrows. As a result of this, the total mass of the fluid in W is decreasing.

One interpretion of (5.41) is as a statement of *mass conservation*: fluid particles cannot instantaneously "teleport" from one point to another. Instead, the only way that such a particle can enter or leave W is to pass through its boundary S.

Remark 5.67. Again, to be more precise, both ρ and $\boldsymbol{v}$ should depend on an additional time variable, which we omit from our notations for simplicity.

Applying the divergence theorem to the right-hand side of (5.41) yields

$$\frac{\partial}{\partial t}\iiint_W \rho \, dV = -\iiint_W [\nabla \cdot (\rho \boldsymbol{v})] \, dV.$$

(Recall that S is given the outward-facing orientation.) Since the above holds for *every* region W, we conclude from this the following *differential* identity:

$$-\frac{\partial}{\partial t}\rho = \nabla \cdot (\rho \boldsymbol{v}). \tag{5.42}$$

The equation (5.42) is yet another example of a partial differential equation that plays an important role in research within mathematics, physics, and engineering. In particular, this equation can be solved to predict future behaviours of fluids.

Notes and Acknowledgments

MTH5113: Introduction to Differential Geometry was designed from the topics of two previously existing modules, *MTH5102: Calculus III* and *MTH5109: Geometry II: Knots and Surfaces*. One goal of *MTH5113* was to better streamline the contents of *MTH5102* and *MTH5109*, which had significant overlaps.

The present notes are a lightly updated version of the *MTH5113* lecture notes from previous years. The most recent changes include the following:

- A new section (*2.4: Limits and Continuity*) was added to provide an informal discussion on continuous vector-valued functions. This was done to make the formal definitions of curves and surfaces more understandable.
- The discussions in a few sections were made more concise.

I thank the students from the previous iterations of *MTH5113* for their feedback on the module contents and on the lecture notes.

The contents of these lecture notes are largely based on the following sources:

- Previous lecture notes [3, 5, 6] for *MTH5109*.
- Previous lecture notes [2] for *MTH5102*.

However, I do regret that several interesting topics from these notes (in particular, the curvature of curves and surfaces) could not be covered in *MTH5113*. Hopefully, these will be found in third-year modules in the future.

While these notes are intended to be entirely self-contained, the interested reader may also wish to consult the following texts from the official module reading list:

(1) Thomas, Weir, Hass. *Thomas' Calculus*. [10].
(2) Spiegel, Lipschutz, Spellman. *Vector Analysis and an Introduction to Tensor Analysis*. [9].
(3) Simons. *Vector Analysis for Mathematicians, Scientists and Engineers*. [8].
(4) Bär. *Elementary Differential Geometry*. [1].
(5) Pressley. *Elementary Differential Geometry*. [4].

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. M. MacCallum, C. D. Murray, P. Saha, W. Sutherland, and M. J. Thompson, MTH5102 (Calculus III) Lecture Notes, 2010, Queen Mary University of London.
3. S. Majid, MTH5109 (Geometry II) Lecture Notes, 2015–2016, Queen Mary University of London.
4. A. Pressley, Elementary Differential Geometry, Springer, 2010.
5. A. Shao, MTH5109 (Geometry II) Lecture Notes, 2016–2017, Queen Mary University of London.
6. ———, MTH5109 (Geometry II) Lecture Notes, 2017–2018, Queen Mary University of London.
7. ———, MTH5109 (Introduction to Differential Geometry) Lecture Notes, 2018–2019, Queen Mary University of London.
8. S. Simons, Vector analysis and an introduction to tensor analysis, Pergamon, 1970.
9. M. R. Spiegel, S. Lipschutz, and D. Spellman, Vector analysis and an introduction to tensor analysis, McGraw-Hill, 2009.
10. G. B. Thomas, M. D. Weir, and J. Hass, Thomas' calculus, Addison-Wesley, 2010.
11. Wikipedia: Geometry, `http://en.wikipedia.org/wiki/Geometry`.