Multivariable Calculus
Semester 2: Multivariable Calculus

Anderson Trimm
Gwinnett School of Mathematics, Science and Technology

These are the notes for the Spring Semester 2020 course in Multivariable Calculus at GSMST. They will continually be updated throughout the course. The latest PDF can always be accessed at https://github.com/atrimm/mvc/blob/master/Course%20Notes/multivariable_calculus_2020.pdf.

Contents

Quick reference

1 Curves

In this section we study functions with one input and multiple outputs.

1.1 Vector-valued functions

1.1.1 Definitions

Suppose a particle moves in the plane along the following curve C: Since the curve fails the

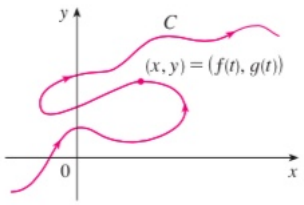


Figure 1: Trajectory of a particle moving in the plane.

vertical line test, C cannot be described as the graph of a function $y = f(x)$. Note however that the x- and y-coords of the particle are functions of time

$$x = f(t), \quad y = g(t)$$

so the curve C can be described as the image of function $\mathbf{r} : I \to \mathbb{R}^2$ defined by

$$\mathbf{r}(t) = (f(t), g(t)),$$

where $I = [a, b]$ is an interval in $\mathbb{R}$.

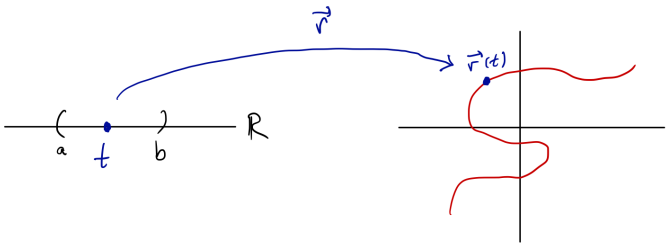


Figure 2: A mapping diagram showing the function $\mathbf{r} : I \to \mathbb{R}^2$.

Definition 1.1 (Vector-valued function). Let $I \subseteq \mathbb{R}$. A mapping $\mathbf{r} : I \to \mathbb{R}^n$ called a *vector-valued function*. The value of $\mathbf{r}$ at $t \in I$ can be written as

$$\mathbf{r}(t) = (r_1(t), r_2(t), \ldots, r_n(t))$$

where the n functions $r_i : I \to \mathbb{R}$, $i = 1, \ldots, n$ are called the *component functions* of $\mathbf{r}$.

Unless specified otherwise, we will take the domain I of a vector-valued function to be the largest domain on which all of the component functions are defined.

Example 1.2. Consider the vector-valued function $\mathbf{r} : I \to \mathbb{R}^3$ defined by

$$\mathbf{r}(t) = (t^3, \ln(3-t), \sqrt{t}).$$

The component functions of $\mathbf{r}(t)$ are

$$r_1(t) = t^3, \quad r_2(t) = \ln(3-t), \quad r_3(t) = \sqrt{t}.$$

The domains of each of these functions, respectively, are

$$I_1 = \mathbb{R}, \quad I_2 = (-\infty, 3), \quad I_3 = [0, \infty),$$

so the domain I of $\mathbf{r}(t)$ is

$$I = I_1 \cap I_2 \cap I_3 = [0, 3).$$

This function can be visualized in *Mathematica* using"ParametricPlot3D":

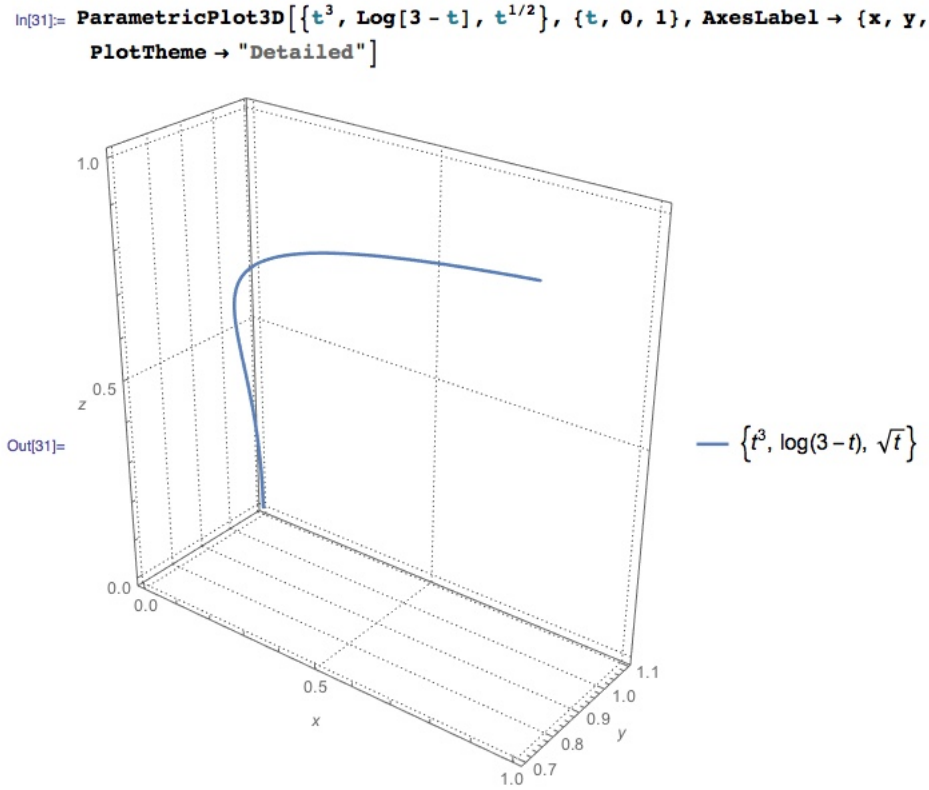


Figure 3: Plot of the vector-valued function $\mathbf{r}(t) = (t^3, \ln(3-t), \sqrt{t})$ from $t = 0$ to $t = 1$.

Exercise 1.1. Consider the vector-valued function $\mathbf{r} : I \to \mathbb{R}^3$ defined by

$$\mathbf{r}(t) = \left(\frac{t-2}{t+2}, \sin t, \ln(9-t^2)\right).$$

What is the domain I of the function?

Solution. The component functions of $\mathbf{r}(t)$ are

$$r_1 = \frac{t-2}{t+2}, \qquad r_2 = \sin t, \qquad r_3 = \ln(9-t^2).$$

The domains of $r_1(t)$ and $r_2(t)$ are given, respectively, by

$$I_1 = (-\infty, -2) \cup (-2, \infty), \qquad I_2 = \mathbb{R}.$$

To find the domain of $r_3(t)$, we need to solve the inequality

$$9 - t^2 > 0.$$

The graph of the function $y = 9 - x^2$ is a concave-down parabola with y-intercept 9 and x-intercepts ± 3.

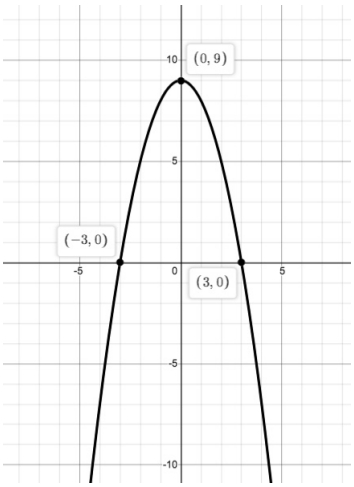


Figure 4: Graph of $y = 9 - x^2$.

We have $y > 0$ where the graph is above the x-axis, so $y > 0$ when $-3 < x < 3$. The domain of $r_3(t)$ is therefore

$$I_3 = (-3, 3).$$

The domain of $\mathbf{r}(t)$ is then

$$I = I_1 \cap I_2 \cap I_3 = (-3, -2) \cup (-2, 3).$$

□

1.1.2 Review: limits of single-variable functions

Before considering limits of vector-valued functions, let's review the definition for a real-valued function $y = f(x)$ of a single real variable x.

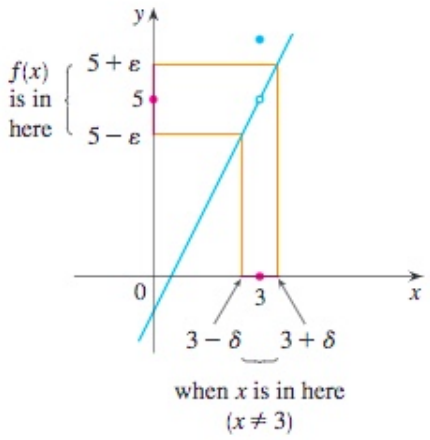


Figure 5: Graph of the function $y = f(x)$ in the example above.

To motivate the definition, consider the function

$$f(x) = \begin{cases} 2x - 1, & \text{if } x \neq 3 \\ 6, & \text{if } x = 3 \end{cases}$$

whose graph is shown in the figure below.

From the graph, we see that when x is close to 3 but not equal to 3, then $f(x)$ is close to 5, and so $\lim_{x \to 3} f(x) = 5$.

However, it is important to be able to state this precisely. For instance, we may ask: *How close to 3 does x have to be so that $f(x)$ differs from 5 by less than* 0.1*?*

The distance from x to 3 is $|x - 3|$ and the distance from $f(x)$ to 5 is $|f(x) - 5|$, so our problem is to find a number δ such that

$$|f(x) - 5| < 0.1 \quad \text{if} \quad 0 < |x - 3| < \delta.$$

If $x \neq 3$, then

$$|f(x) - 5| = |(2x - 1) - 5| = |2x - 6| = 2|x - 3|$$

so we see that by taking $\delta = \frac{1}{2}(0.1) = 0.05$, we have $|f(x) - 5| < 2(0.05) = 0.1$. Thus, an answer to the problem is given by $\delta = 0.05$; that is, if x is within a distance of 0.05 from 3, then $f(x)$ will be within a distance of 0.1 from 5.

If we change the number 0.1 in our problem to the smaller number 0.01, then by using the same method we find that $f(x)$ will differ from 5 by less than 0.01 provided that x differs from 3 by less than $\frac{1}{2}(0.01) = 0.005$; that is,

$$|f(x) - 5| < 0.01 \quad \text{if} \quad 0 < |x - 3| < 0.005.$$

Similarly,

$$|f(x) - 5| < 0.001 \quad \text{if} \quad 0 < |x - 3| < 0.0005.$$

Think of the numbers $0.1, 0.01, 0.001$ above as *error tolerances* that we might allow. That is, when challenged with an error tolerance, it is our task to find a corresponding δ so that whenever x is within a distance of δ from 3, $f(x) \approx 5$, within the given error tolerance.

Now for 5 to be the precise limit of $f(x)$ as x approaches 3, we must not only be able to bring the difference between $f(x)$ and 5 below each of these numbers; we must be able to bring it below *any* positive number. And, by exactly the same reasoning, we can. That is, if ϵ is any positive number, then by choosing $\delta = \frac{\epsilon}{2}$, we find

$$|f(x) - 5| < \epsilon \quad \text{if} \quad 0 < |x - 3| < \delta = \frac{\epsilon}{2}. \tag{1.1}$$

This is a precise way of saying that $f(x)$ is close to 5 when x is close to 3, because Equation (1.1) says that we can make the values of $f(x)$ within an arbitrary distance ϵ from 5 by taking the values of x within a distance $\frac{\epsilon}{2}$ from 3 (but $x \neq 3$).

Note that Equation (1.1) can be rewritten as follows:

$$\text{if} \quad 3 - \delta < x < 3 + \delta \ \ (x \neq 3) \quad \text{then} \quad 5 - \epsilon < f(x) < 5 + \epsilon$$

as illustrated in the figure above. This says that by taking the values of x ($x \neq 3$) to lie in the interval $(3 - \delta, 3 + \delta)$ we can make the values of $f(x)$ lie in the interval $(5 - \epsilon, 5 + \epsilon)$.

Following the reasoning in this example, the precise definition of a limit is the following.

Definition 1.3 (Limit of a single-variable function). Let (a, b) be an open interval containing the point x_0 and let $f(x)$ be a real-valued function defined on this interval, except possibly at x_0 itself. A number L is called the *limit of $f(x)$ as x approaches x_0* if for every $\epsilon > 0$ there exists a $\delta > 0$ such that $|f(x) - L| < \epsilon$ whenever $0 < |x - x_0| < \delta$. If such an L exists, we write

$$\lim_{x \to x_0} f(x) = L.$$

In computing limits using Definition 1.3, we use the following properties of the absolute value function:

Theorem 1.4 (Properties of the absolute value function). The absolute value function $f(x) = |x|$ has the following properties

(1) $|x| \geq 0$ for all $x \in \mathbb{R}$, and $|x| = 0$ if and only if $x = 0$.

(2) $|xy| = |x|\,|y|$ for all $x, y \in \mathbb{R}$.

(3) For all $x, y \in \mathbb{R}$,

$$|x + y| \leq |x| + |y|.$$

The third property is called the *triangle inequality*. This is because, geometrically, given two line segments of lengths $|x|$ and $|y|$, the third line segment must have length $< |x + y|$ to be able to form a triangle (the case $|x + y| = |x| + |y|$ corresponds to a straight line).

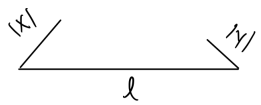


Figure 6: We see that if $\ell > |x| + |y|$, then the three line segments cannot form a triangle.

Proof. (1) We can define $|x|$ as

$$|x| = \begin{cases} x, & \text{if } x \geq 0, \\ -x, & \text{if } x < 0, \end{cases}$$

so this property is immediate from the definition.

(2) If either $x = 0$ or $y = 0$, then $xy = 0$ so by property (1) $|xy| = 0 = |x||y|$. Suppose now that neither x nor y are zero. If $x, y > 0$ then $xy > 0$, so $|xy| = xy = |x|\,|y|$. If $x, y < 0$, then $xy > 0$, so $|xy| = xy = (-x)(-y) = |x|\,|y|$. If $x > 0$ and $y < 0$, then $xy < 0$, so $|xy| = -xy = x(-y) = |x|\,|y|$. If $x < 0$ and $y > 0$ then $xy < 0$, so $|xy| = -xy = (-x)y = |x|\,|y|$.

(3) For any $x, y \in \mathbb{R}$ we have

$$\begin{aligned} |x+y|^2 = (x+y)^2 &= x^2 + y^2 + 2xy \\ &= |x|^2 + |y|^2 + 2xy \\ &\leq |x|^2 + |y|^2 + 2|x||y| \\ &= (|x| + |y|)^2. \end{aligned}$$

Since both $|x+y|$ and $|x| + |y|$ are nonnegative, this implies that [1]

$$|x+y| \leq |x| + |y|.$$

□

Theorem 1.5 (Uniqueness of limits). If $f(x)$ has a limit L at x_0, then the limit is unique.

Proof. Suppose that $\lim_{x \to x_0} f(x) = L$ and $\lim_{x \to x_0} f(x) = L'$. Then, given any $\epsilon > 0$ there exist positive numbers δ_1 and δ_2 such that

$$|f(x) - L| < \frac{\epsilon}{2} \quad \text{if} \quad |x - x_0| < \delta_1$$

and

$$|f(x) - L'| < \frac{\epsilon}{2} \quad \text{if} \quad |x - x_0| < \delta_2.$$

Let $\delta \equiv \min\{\delta_1, \delta_2\}$ be the *minimum* of δ_1 and δ_2. [2] Then if we take $|x - x_0| < \delta$, *both* of these inequalities hold at the same time, so we have

$$|L - L'| = |L - L' + f(x) - f(x)| = |L - f(x) + f(x) - L| \leq |f(x) - L| + |f(x) - L'| < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon.$$

[1] Think about the graph of $f(x) = x^2$: it is strictly decreasing on $(-\infty, 0)$ and increasing on $(0, \infty)$. Thus, it is only true in general that $x_1^2 < x_2^2 \implies x_1 < x_2$ if both of these numbers are nonnegative. Otherwise, this might not be true. For instance, $2^2 < (-4)^2$, but $-4 < 2$.

[2] Note that $\delta > 0$, since the minimum of two positive numbers is positive.

For this to be true for all $\epsilon > 0$ (that is, $|L - L'|$ is smaller than *any* positive number), we must have $L - L' = 0$, or $L = L'$. □

Exercise 1.2. Use Definition 1.3 to prove that $\lim_{x\to 3}(4x - 5) = 7$.

Solution. Let $\epsilon > 0$. For all $x \neq 3$,

$$|f(x) - 7| = |(4x - 5) - 7| = |4x - 12| = 4|x - 3|.$$

By taking $\delta = \frac{\epsilon}{4}$, we have $0 < |x - 3| < \frac{\epsilon}{4}$ and therefore

$$|f(x) - 7| = 4|x - 3| < 4 \cdot \frac{\epsilon}{4} = \epsilon,$$

which proves that $\lim_{x\to 3}(4x - 5) = 7$. □

Example 1.6. We now use Definition 1.3 to prove that $\lim_{x\to 3} x^2 = 9$.

Let $\epsilon > 0$. For all $x \neq 3$, we have

$$|f(x) - 9| = |x^2 - 9| = |(x + 3)(x - 3)| = |x + 3||x - 3|.$$

Notice that if we can find a positive number C such that $|x + 3| < C$, then

$$|x + 3||x - 3| < C|x - 3|$$

and we can make $C|x - 3| < \epsilon$ by taking $|x - 3| < \frac{\epsilon}{C} = \delta$. We can find such a number C if we restrict x to lie in some interval centered at 3. Since we are only interested in values of x that are close to 3, this is exactly what we want. Let's assume that $|x - 3| < \alpha$ for some positive number α, say $\alpha = 1$ (it does not matter what positive number we take here). Then we have

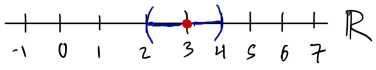


Figure 7: The interval $(2, 4)$ of radius 1 centered at 3.

$$x - 3 < 1 \quad \text{or} \quad -x + 3 < 1.$$

The first inequality says $x < 4$ and the second says $2 < x$, so $|x - 3| < 1$ implies that

$$2 < x < 4.$$

Adding 3 to both sides of this inequality gives

$$5 < x + 3 < 7,$$

and therefore $|x + 3| < |7| = 7 = C$. But now there are two restrictions on $|x - 3|$, namely

$$|x - 3| < 1 \quad \text{and} \quad |x - 3| < \frac{\epsilon}{C} = \frac{\epsilon}{7}.$$

To make sure that both of these inequalities are satisfied, we take $\delta = \min\{1, \frac{\epsilon}{7}\}$. Since $0 < |x - 3| < \delta$ implies $|x^2 - 9| < \epsilon$, this proves that $\lim_{x\to 3} x^2 = 9$.

The previous example shows that it is not always easy to prove that a function has a particular limit using Definition 1.3. In fact, if we had considered a more complicated function such as

$$f(x) = \frac{6x^2 - 8x + 9}{2x^2 - 1}$$

then proving that $\lim_{x\to 1} f(x) = 7$ using Definition 1.3 would require a great deal of ingenuity. Instead, we prove the following theorems, which makes evaluating limits much easier. In the proofs, note the crucial role played by the properties of the absolute value function from Theorem 1.4

Theorem 1.7 (Limit laws for single-variable functions). Suppose $f(x)$ and $g(x)$ are defined on the same open set containing x_0, and that

$$\lim_{x\to x_0} f(x) = L \qquad \text{and} \qquad \lim_{x\to x_0} g(x) = M.$$

Then

(i) $\lim_{x\to x_0} c = c$ for any constant $c \in \mathbb{R}$.

(ii) $\lim_{x\to x_0} x = x_0$.

(iii) $\lim_{x\to x_0} cf(x) = cL$ for any $c \in \mathbb{R}$;

(iv) $\lim_{x\to x_0}(f(x) + g(x)) = L + M$;

(v) $\lim_{x\to x_0}(f(x)g(x)) = LM$;

(vi) $\lim_{x\to x_0} = \frac{f(x)}{g(x)} = \frac{L}{M}$ whenever $M \neq 0$.

Proof. (i) Let $\epsilon > 0$. Since $|c - c| = 0$, $|c - c| < \epsilon$ whenever $|x - x_0| < \delta$ for any positive number δ.

(ii) Given $\epsilon > 0$, by taking $\delta = \epsilon$ we have $|x - x_0| < \epsilon$ whenever $|x - x_0| < \delta = \epsilon$.

(iii) Since $\lim_{x\to x_0} f(x) = L$, given $\epsilon > 0$ there exists a corresponding $\delta > 0$ such that $|f(x) - L| < \epsilon$ whenever $0 < |x - x_0| < \delta$. Then $|cf(x) - cL| = |c||f(x) - L| < \epsilon$ whenever $0 < |x - x_0| < \frac{\epsilon}{|c|}$.

(iv) We have

$$|f(x) + g(x) - (L + M)| = |(f(x) - L) + (g(x) - M)| \leq |f(x) - L| + |g(x) - M|$$

by the Triangle Inequality. Since $\lim_{x\to x_0} f(x) = L$ and $\lim_{x\to x_0} g(x) = M$, given $\epsilon > 0$ there exist positive numbers δ_1 and δ_2 such that

$$|f(x) - L| < \frac{\epsilon}{2} \qquad \text{if} \qquad |x - x_0| < \delta_1$$

and

$$|g(x) - M| < \frac{\epsilon}{2} \qquad \text{if} \qquad |x - x_0| < \delta_2.$$

By taking $|x - x_0| < \delta = \min\{\delta_1, \delta_2\}$, we have

$$|f(x) + g(x) - (L + M)| \leq |f(x) - L| + |g(x) - M| < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon,$$

which proves that $\lim_{x\to x_0}(f(x) + g(x)) = L + M$.

(v) First, note that

$$f(x)g(x) - LM = (f(x) - L)(g(x) - M) + L(g(x) - M) + M(f(x) - L).$$

Let $\epsilon > 0$. Since $\lim_{x\to x_0} f(x) = L$ there exists $\delta_1 > 0$ such that $|f(x) - L| < \sqrt{\epsilon}$ whenever $|x - x_0| < \delta_1$. Since $\lim_{x\to x_0} g(x) = M$ there exists $\delta_2 > 0$ such that $|g(x) - M| < \sqrt{\epsilon}$ whenever $|x - x_0| < \delta_2$. Then, whenever $|x - x_0| < \delta = \min\{\delta_1, \delta_2\}$, we have

$$|(f(x) - L)(g(x) - M)| = |f(x) - L||g(x) - M| < (\sqrt{\epsilon})^2 = \epsilon$$

which shows that $\lim_{x-x_0}(f(x) - L)(g(x) - M) = 0$. By (iii),

$$\lim_{x\to x_0} L(g(x) - M) = L \lim_{x\to x_0} (g(x) - M) = L \cdot 0 = 0,$$

and

$$\lim_{x\to x_0} M(f(x) - L) = M \lim_{x\to x_0} (f(x) - L) = M \cdot 0 = 0.$$

Applying (iv),

$$\begin{aligned}\lim_{x\to x_0} (f(x)g(x) - LM) &= \lim_{x\to x_0} (f(x) - L)(g(x) - M) + \lim_{x\to x_0} L(g(x) - M) + \lim_{x\to x_0} M(f(x) - L) \\ &= 0 + 0 + 0 \\ &= 0,\end{aligned}$$

and therefore

$$\lim_{x\to x_0} f(x)g(x) = LM.$$

(vi) First, note that since $|M| > 0$ and $\lim_{x\to x_0} g(x) = M$, there exists $\delta_1 > 0$ such that $|g(x)| > \frac{1}{2}|M|$ whenever $|x - x_0| < \delta_1$. Let $\epsilon > 0$. Choose $\delta_2 > 0$ such that $|x - x_0| < \delta_2$ implies that $|g(x) - M| < \frac{1}{2}|M|^2\epsilon$. Then, for $|x - x_0| < \delta = \min\{\delta_1, \delta_2\}$, we have

$$\begin{aligned}\left|\frac{1}{g(x)} - \frac{1}{M}\right| &= \left|\frac{M - g(x)}{Mg(x)}\right| \\ &= \frac{|g(x) - M|}{|Mg(x)|} \\ &< \frac{\frac{1}{2}|M|^2\epsilon}{\frac{1}{2}|M|^2} \\ &= \epsilon,\end{aligned}$$

and therefore $\lim_{x\to x_0} \frac{1}{g(x)} = \frac{1}{M}$.

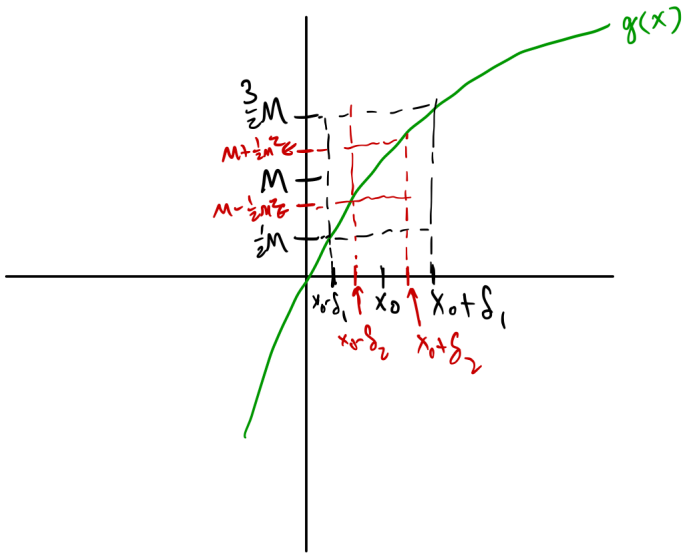


Figure 8: Illustration of bounds on $g(x)$ in proof of Limit Law (vi).

It then follows from (v) that

$$\begin{aligned}\lim_{x\to x_0}\frac{f(x)}{g(x)} &= \lim_{x\to x_0} f(x)\lim_{x\to x_0}\frac{1}{g(x)}\\ &= \frac{L}{M}.\end{aligned}$$

□

Using Theorem 1.7, it is much easier to prove the limits in the examples above. For instance

$$\begin{aligned}\lim_{x\to 3}(4x-5) &= (\lim_{x\to 3}4)(\lim_{x\to 3}x)+(-\lim_{x\to 3}(5))\\ &= 4(3)+(-5)\\ &= 12-5\\ &= 7,\end{aligned}$$

and

$$\lim_{x\to 3}x^2 = (\lim_{x\to 3}x)(\lim_{x\to 3}x) = (3)(3) = 9.$$

Note that, in both of these examples, the function $f(x)$ is actually defined at x_0 and $\lim_{x\to x_0} f(x) = f(x_0)$; that is, the limit of $f(x)$ as x approaches x_0 is equal to the value of $f(x)$ at x_0.

Definition 1.8 (Continuity). Let $f(x)$ be defined on an open interval (a,b) containing a point x_0. We say that $f(x)$ is *continuous at* x_0 if $\lim_{x\to x_0} f(x) = f(x_0)$. We then say that $f(x)$ is *continuous on* (a,b) if $f(x)$ is continuous at every point in (a,b).

We emphasize that Definition 1.8 says that for f to be continuous at x_0, the following three things must be true:

1. f is defined at x_0 (i.e., $f(x_0)$ exists),
2. $\lim_{x\to x_0}$ exists,
3. $\lim_{x\to x_0} = f(x_0)$.

The limit laws in Theorem 1.7 immediately imply the following

Theorem 1.9 (New continuous functions from old). Let $f(x)$ and $g(x)$ be defined on the same open interval containing x_0. If $f(x)$ and $g(x)$ are continuous at x_0, then so are

(i) $cf(x)$

(ii) $f(x) + g(x)$

(iii) $f(x)g(x)$

(iv) $f(x)/g(x)$, whenever $g(x_0) \neq 0$.

Example 1.10 (Examples of Continuous Functions).

- Polynomials are continuous on $\mathbb{R}$;
- Rational functions are continuous wherever they are defined;
- The absolute value function $f(x) = |x|$ is continuous on $\mathbb{R}$;

Trig functions, and exponential and logarithmic functions are all also continuous wherever they are defined (though the proofs don't depend on Theorem 1.9).

Exercise 1.3. Prove that $f(x) = |x|$ is continuous on $\mathbb{R}$.

Solution. If $x > 0$, then $f(x) = x$ which is continuous since it is a polynomial. The same is true for $x < 0$ since then $f(x) = -x$. By taking $\delta = \epsilon$, $|f(x) - 0| = ||x|| = |x| < \epsilon$ whenever $|x| < \delta = \epsilon$, so $\lim_{x\to 0} f(x) = 0 = f(0)$, which shows that $f(x)$ is also continuous at $x = 0$. Thus, $f(x)$ is continuous on $\mathbb{R}$. □

Theorem 1.11 (A composition of continuous functions is continuous). Suppose $f(x)$ is defined on an open interval containing x_0 and $g(x)$ is defined on an open interval containing $f(x_0)$. If f is continuous at x_0 and $g(x)$ is continuous at $f(x_0)$, then $(g \circ f)(x)$ is continuous at x_0.

Proof. Let $\epsilon > 0$. Since g is continuous at $f(x_0)$, corresponding to ϵ there exists $\eta > 0$ such that $|g(f(x)) - g(f(x_0))| < \epsilon$ whenever $|f(x) - f(x_0)| < \eta$. Since f is continuous at x_0, corresponding to η there exists $\delta > 0$ such that $|f(x) - f(x_0)| < \eta$ whenever $|x - x_0| < \delta$. This shows that $|g(f(x)) - g(f(x_0))| < \epsilon$ whenever $|x - x_0| < \delta$, proving that $(g \circ f)(x)$ is continuous at x_0. □

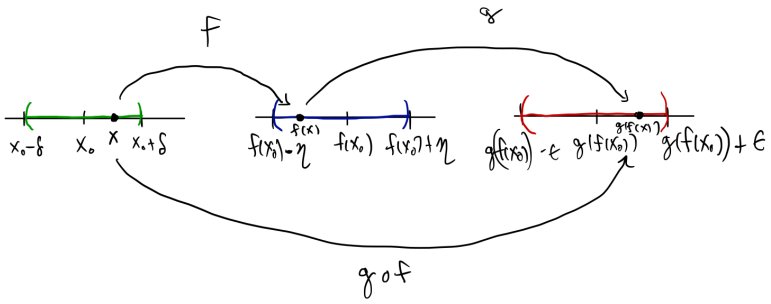


Figure 9: A composition of continuous functions is continuous.

Example 1.12. Consider the function $f(x) = e^{x^2}$. We can view $f(x)$ as the composition $(h \circ g)(x)$, where $h(x) = e^x$ and $g(x) = x^2$. Since $h(x)$ and $g(x)$ are continuous on $\mathbb{R}$, by Theorem 1.11 so is $f(x)$.

1.1.3 Limits of vector-valued functions

Throughout this section, let $I = (a, b)$ denote an open interval in $\mathbb{R}$ containing a point t_0, and let $\mathbf{r}(t)$ be a vector-valued function defined on I, except perhaps at t_0 itself. For a vector $\mathbf{x} = (x_1, \ldots, x_n)$ in $\mathbb{R}^n$, we write

$$||\mathbf{x}|| = \sqrt{\sum_{i=1}^{n} x_i^2}$$

for the length of $\mathbf{x}$. Recall that the distance between two points $\mathbf{x}$ and $\mathbf{y}$ in $\mathbb{R}^n$ is then given by the length of the vector $\mathbf{x} - \mathbf{y}$:

$$||\mathbf{x} - \mathbf{y}|| = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}.$$

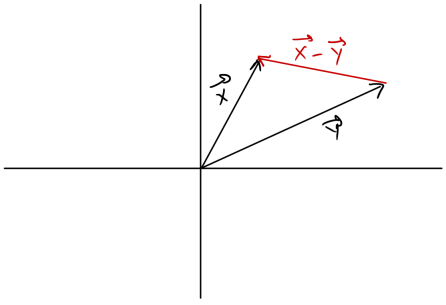


Figure 10: The distance between points $\mathbf{x}$ and $\mathbf{y}$ in $\mathbb{R}^n$ is the length of the vector $\mathbf{x} - \mathbf{y}$.

The definition of the limit of a vector-valued function is the obvious generalization of the limit of a real-valued function.

Definition 1.13 (Limit of a vector-valued function). A fixed vector $\mathbf{L} \in \mathbb{R}^n$ is said to be the *limit of* $\mathbf{r}(t)$ *as* $\mathbf{r}(t)$ *approaches* t_0 if for every $\epsilon > 0$ there exists a corresponding $\delta > 0$ such that

$$0 < |t - t_0| < \delta \implies ||\mathbf{r}(t) - \mathbf{L}|| < \epsilon.$$

If $\mathbf{L}$ exists, we write $\lim_{t \to t_0} \mathbf{r}(t) = \mathbf{L}$.

We will now show that the limit of a vector-valued function can be computed in terms of the limits of its component functions. We will need the following lemma.

Lemma 1.14. Let $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ and $\mathbf{y} = (y_1, y_2, \ldots, y_n)$ be vectors in $\mathbb{R}^n$. Then

$$|x_i - y_i| \leq ||\mathbf{x} - \mathbf{y}|| \leq \sum_{i=1}^{3} |x_i - y_i|$$

for all $i = 1, 2, \ldots, n$.

Proof. For any fixed index i, we have

$$\begin{aligned} ||\mathbf{x} - \mathbf{y}||^2 &= \sum_{j=1}^{n} (x_j - y_j)^2 \\ &= (x_i - y_i)^2 + \underbrace{\sum_{j \neq i} (x_j - y_j)^2}_{\geq 0} \\ &\geq (x_i - y_i)^2 \\ &= |x_i - y_i|^2. \end{aligned}$$

Since both $||\mathbf{x} - \mathbf{y}||$ and $|x_i - y_i|$ nonnegative, this implies that

$$||\mathbf{x} - \mathbf{y}|| \geq |x_i - y_i|,$$

so the first inequality holds.

To see that the second inequality holds, note that

$$\begin{aligned} \left(\sum_{i=1}^{n} |x_i - y_i| \right)^2 &= \sum_{i=1}^{n} |x_i - y_i|^2 + \underbrace{2 \sum_{1 \leq i < j \leq n} |x_i - y_i||x_j - y_j|}_{\geq 0} \\ &\geq \sum_{i=1}^{n} |x_i - y_i|^2 \\ &= \sum_{i=1}^{n} (x_i - y_i)^2 \\ &= ||\mathbf{x} - \mathbf{y}||^2. \end{aligned}$$

Since both $\sum_{i=1}^{n} |x_i - y_i|$ and $||\mathbf{x} - \mathbf{y}||$ are nonnegative, this implies that

$$\sum_{i=1}^{n} |x_i - y_i| \geq ||\mathbf{x} - \mathbf{y}||,$$

so the second inequality holds. □

Exercise 1.4. Verify that

$$\left(\sum_{i=1}^{n}|x_i-y_i|\right)^2=\sum_{i=1}^{n}|x_i-y_i|^2+2\sum_{1\le i<j\le n}|x_i-y_i||x_j-y_j|$$

for $n=3$ by explicitly writing out both sides.

Theorem 1.15 (Limit of a vector-valued function). Let $\mathbf{r}:I\to\mathbb{R}^n$ be a vector-valued function. Then

$$\lim_{t\to t_0}\mathbf{r}(t)=(\lim_{t\to t_0}r_1(t),\lim_{t\to t_0}r_2(t),\lim_{t\to t_0}r_3(t)). \tag{1.2}$$

Proof. Let $\mathbf{L}=(L_1,L_2,\dots,L_n)$ be a fixed vector in $\mathbb{R}^n$. We will prove that $\lim_{t\to t_0}\mathbf{r}(t)=\mathbf{L}$ if and only if $\lim_{t\to t_0}r_i(t)=L_i$ for all $i=1,\dots,n$; that is, both sides of Equation (1.2) are either undefined, or they are both equal to $\mathbf{L}$ and hence to each other.

($\Longrightarrow$) First, suppose that $\lim_{t\to t_0}\mathbf{r}(t)=\mathbf{L}$. Then, given $\epsilon>0$, there exists $\delta>0$ such that $||\mathbf{r}(t)-\mathbf{L}||<\epsilon$ whenever $0<|t-t_0|<\delta$. By Lemma 1.14, for each $i=1,\dots,n$

$$|r_i(t)-L_i|<||\mathbf{r}(t)-\mathbf{L}||$$

so we have $|r_i(t)-L_i|<\epsilon$ for each $i=1,\dots,n$ whenever $0<|t-t_0|<\delta$. Thus, $\lim_{t\to t_0}\mathbf{r}(t)=\mathbf{L}$ implies that $\lim_{t\to t_0}r_i(t)=L_i$ for all $i=1,\dots,n$.

($\Longleftarrow$) Now suppose that $\lim_{t\to t_0}r_i(t)=L_i$ for all $i=1,\dots,n$. Given $\epsilon>0$, there exist positive numbers $\delta_1,\delta_2,\dots,\delta_n$ such that $|r_i(t)-L_i|<\frac{\epsilon}{n}$ whenever $0<|t-t_0|<\delta_i$. By Lemma 1.14,

$$||\mathbf{r}(t)-\mathbf{L}||<\sum_{i=1}^{n}|r_i(t)-L_i|,$$

so by taking $\delta=\min\{\delta_1,\delta_2,\dots,\delta_n\}$, we have

$$||\mathbf{r}(t)-\mathbf{L}||<\sum_{i=1}^{n}|r_i(t)-L_i|<\frac{\epsilon}{n}+\cdots+\frac{\epsilon}{n}=n\frac{\epsilon}{n}=\epsilon$$

whenever $|t-t_0|<\delta$. Thus, $\lim_{t\to t_0}r_i(t)=L_i$ for all $i=1,\dots,n$ implies that $\lim_{t\to t_0}\mathbf{r}(t)=\mathbf{L}$. □

Corollary 1.16 (Uniqueness of the limit of a vector-valued function). If $\lim_{t\to t_0}\mathbf{r}(t)=\mathbf{L}$, then the limit is unique.

Proof. Since the limits $\lim_{t\to t_0}r_i(t)=L_i$ are unique (if they exist) by Theorem 1.5, it follows immediately from Theorem 1.15 that $\lim_{t\to t_0}\mathbf{r}(t)=\mathbf{L}$ is unique if it exists. □

Example 1.17. Let $\mathbf{r}(t)=(1+t^3,te^{-t},\frac{\sin t}{t})$. Since

$$\begin{aligned}\lim_{t\to0}(1+t^3)&=1,\\ \lim_{t\to0}te^{-t}&=\lim_{t\to0}t\lim_{t\to0}e^{-t}=0\cdot1=0,\\ \lim_{t\to0}\frac{\sin t}{t}&=\lim_{t\to0}\cos t=1\quad\text{(by L'Hospital's rule)}\end{aligned}$$

by Theorem 1.15

$$\begin{aligned}\lim_{t\to0}\mathbf{r}(t)&=\left(\lim_{t\to0}(1+t^3),\lim_{t\to0}te^{-t},\lim_{t\to0}\frac{\sin t}{t}\right)\\ &=(1,0,1).\end{aligned}$$

Exercise 1.5. Find $\lim_{t\to 1}\mathbf{r}(t)$, where $\mathbf{r}(t)=\left(\frac{t^2-t}{t-1},\sqrt{t+8},\frac{\sin(\pi t)}{\ln(t)}\right)$, if it exists.

Solution. Since

$$\begin{aligned}\lim_{t\to 1}\frac{t^2-t}{t-1}&=\lim_{t\to 1}\frac{t(t-1)}{t-1}=\lim_{t\to 1}t=1,\\ \lim_{t\to 1}\sqrt{t+8}&=\sqrt{1+8}=\sqrt{9}=3,\\ \lim_{t\to 1}\frac{\sin(\pi t)}{\ln(t)}&=\lim_{t\to 1}\frac{\pi\cos(\pi t)}{\frac{1}{t}}=\lim_{t\to 1}\pi t\cos(\pi t)=\pi(1)\cos(\pi)=-\pi,\end{aligned}$$

by Theorem 1.15

$$\begin{aligned}\lim_{t\to 1}\mathbf{r}(t)&=\left(\lim_{t\to 1}\frac{t^2-t}{t-1},\lim_{t\to 1}\sqrt{t+8},\lim_{t\to 1}\frac{\sin(\pi t)}{\ln(t)}\right)\\ &=(1,3,-\pi).\end{aligned}$$

□

Theorem 1.18 (Limit laws for vector-valued functions). Let $\mathbf{u},\mathbf{v}$ be vector valued functions into $\mathbb{R}^n$ defined on the same open interval containing t_0 and let $c\in\mathbb{R}$ be a constant. Then

(i) $\lim_{t\to t_0}(c_1,c_2,\ldots,c_n)=(c_1,c_2,\ldots,c_n)$ if $(c_1,c_2,\ldots,c_n)$ is a constant vector in $\mathbb{R}^n$.

(ii) $\lim_{t\to t_0}c\mathbf{u}(t)=c\lim_{t\to t_0}\mathbf{u}(t)$

(iii) $\lim_{t\to t_0}[\mathbf{u}(t)+\mathbf{v}(t)]=\lim_{t\to t_0}\mathbf{u}(t)+\lim_{t\to t_0}\mathbf{v}(t)$

(iv) $\lim_{t\to t_0}[\mathbf{u}(t)\cdot\mathbf{v}(t)]=\lim_{t\to t_0}\mathbf{u}(t)\cdot\lim_{t\to t_0}\mathbf{v}(t)$

(v) $\lim_{t\to t_0}[\mathbf{u}(t)\times\mathbf{v}(t)]=\lim_{t\to t_0}\mathbf{u}(t)\times\lim_{t\to t_0}\mathbf{v}(t)$ (for $n=3$)

Proof. The proof of each of these follows by applying Theorems 1.7 and 1.15.

(i) If $(c_1,c_2,\ldots,c_n)$ is a constant vector in $\mathbb{R}^n$, then

$$\lim_{t\to t_0}(c_1,c_2,\ldots,c_n)=(\lim_{t\to t_0}c_1,\lim_{t\to t_0}c_2,\ldots,\lim_{t\to t_0}c_n)=(c_1,c_2,\ldots,c_n).$$

(ii) If $c\in\mathbb{R}$ is a constant, then

$$\begin{aligned}\lim_{t\to t_0}c\mathbf{u}(t)&=\lim_{t\to t_0}c(u_1(t),u_2(t),\ldots,u_n(t))\\ &=\lim_{t\to t_0}(cu_1(t),cu_2(t),\ldots,cu_n(t))\\ &=(\lim_{t\to t_0}cu_1(t),\lim_{t\to t_0}cu_2(t),\ldots,\lim_{t\to t_0}cu_n(t))\\ &=(c\lim_{t\to t_0}u_1(t),c\lim_{t\to t_0}u_2(t),\ldots,c\lim_{t\to t_0}u_n(t))\\ &=c(\lim_{t\to t_0}u_1(t),\lim_{t\to t_0}u_2(t),\ldots,\lim_{t\to t_0}u_n(t))\\ &=c\lim_{t\to t_0}(u_1(t),u_2(t),\ldots,u_n(t))\\ &=c\lim_{t\to t_0}\mathbf{u}(t).\end{aligned}$$

(iii) If $\mathbf{u}(t), \mathbf{v}(t)$ are vector-valued functions into $\mathbb{R}^n$, then

$$\begin{aligned}
\lim_{t\to t_0}[\mathbf{u}(t)+\mathbf{v}(t)] &= \lim_{t\to t_0}[(u_1(t),\dots,u_n(t))+(v_1(t),\dots,v_n(t))] \\
&= \lim_{t\to t_0}(u_1(t)+v_1(t),\dots,u_n(t)+v_n(t)) \\
&= (\lim_{t\to t_0}(u_1(t)+v_1(t)),\dots,\lim_{t\to t_0}(u_n(t)+v_n(t))) \\
&= (\lim_{t\to t_0}u_1(t)+\lim_{t\to t_0}v_1(t),\dots,\lim_{t\to t_0}u_n(t)+\lim_{t\to t_0}v_n(t)) \\
&= (\lim_{t\to t_0}u_1(t),\lim_{t\to t_0}u_2(t),\lim_{t\to t_0}u_3(t))+(\lim_{t\to t_0}v_1(t),\lim_{t\to t_0}v_2(t),\lim_{t\to t_0}v_3(t)) \\
&= \lim_{t\to t_0}(u_1(t),u_2(t),u_3(t))+\lim_{t\to t_0}(v_1(t),v_2(t),v_3(t)) \\
&= \lim_{t\to t_0}\mathbf{u}(t)+\lim_{t\to t_0}\mathbf{v}(t).
\end{aligned}$$

(iv) If $\mathbf{u}(t), \mathbf{v}(t)$ are vector-valued functions into $\mathbb{R}^n$, then

$$\begin{aligned}
\lim_{t\to t_0}[\mathbf{u}(t)\cdot\mathbf{v}(t)] &= \lim_{t\to t_0}\sum_{i=1}^{n}u_i(t)v_i(t) \\
&= \sum_{i=1}^{n}\lim_{t\to t_0}u_i(t)v_i(t) \\
&= \sum_{i=1}^{n}\lim_{t\to t_0}u_i(t)\lim_{t\to t_0}v_i(t) \\
&= \lim_{t\to t_0}\mathbf{u}(t)\cdot\lim_{t\to t_0}\mathbf{v}(t).
\end{aligned}$$

(v) If $\mathbf{u}, \mathbf{v}$ are vector-valued functions into $\mathbb{R}^3$, then

$$\begin{aligned}
&\lim_{t\to t_0}[\mathbf{u}(t)\times\mathbf{v}(t)] = \lim_{t\to t_0}(u_2(t)v_3(t)-u_3(t)v_2(t),-u_1(t)v_3(t)+u_3(t)v_1(t),u_1(t)v_2(t)-u_2(t)v_1(t)) \\
&= (\lim_{t\to t_0}(u_2(t)v_3(t)-u_3(t)v_2(t)),\lim_{t\to t_0}(-u_1(t)v_3(t)+u_3(t)v_1(t)),\lim_{t\to t_0}(u_1(t)v_2(t)-u_2(t)v_1(t))) \\
&= (\lim_{t\to t_0}u_2(t)\lim_{t\to t_0}v_3(t)-\lim_{t\to t_0}u_3(t)\lim_{t\to t_0}v_2(t),-\lim_{t\to t_0}u_1(t)\lim_{t\to t_0}v_3(t)+\lim_{t\to t_0}u_3(t)\lim_{t\to t_0}v_1(t), \\
&\qquad\qquad \lim_{t\to t_0}u_1(t)\lim_{t\to t_0}v_2(t)-\lim_{t\to t_0}u_2(t)\lim_{t\to t_0}v_1(t)) \\
&= \lim_{t\to t_0}\mathbf{u}(t)\times\lim_{t\to t_0}\mathbf{v}(t).
\end{aligned}$$

□

[Add examples.]

Definition 1.19 (Continuity). Let $\mathbf{r}(t)$ be defined on an open interval (a,b) containing a point t_0. We say that $\mathbf{r}(t)$ is *continuous at* t_0 if $\lim_{t\to t_0}\mathbf{r}(t) = \mathbf{r}(t_0)$. We then say that $\mathbf{r}(t)$ is *continuous on* (a,b) if $\mathbf{r}(t)$ is continuous at every point in (a,b).

Theorem 1.20 (Continuity of vector-valued functions). A vector-valued function $\mathbf{r}(t) = (r_1(t),\dots,r_n(t))$ is continuous at t_0 if and only if its component functions are all continuous at t_0.

Proof. This follows immediately from Theorem 1.15. □

Example 1.21. The vector-valued function $\mathbf{r}(t) = (\cos t, \sin t, t)$ is continuous on $\mathbb{R}$ since its component functions are each continuous on $\mathbb{R}$.

Theorem 1.22 (New continuous vector-valued functions from old). Let $\mathbf{u}$ and $\mathbf{v}$ be two vector-valued functions on $\mathbb{R}^n$ which are continuous at t_0 and let c be a constant. Then the following functions are also continuous at t_0:

(i) $c\mathbf{u}(t)$

(ii) $\mathbf{u}(t) + \mathbf{v}(t)$

(iii) $\mathbf{u}(t) \cdot \mathbf{v}(t)$

(iv) $\mathbf{u}(t) \times \mathbf{v}(t)$ (for $n = 3$)

Proof. The proof follows immediately from Theorem 1.18. □

The following theorem will be used extensively in the next section.

Theorem 1.23 (Continuity of composite function). Let $\mathbf{r} : \mathbb{R} \to \mathbb{R}^n$ be a continuous vector-valued function and $\varphi : \mathbb{R} \to \mathbb{R}$ a continuous real-valued function. Then the composite function $\tilde{\mathbf{r}} = \mathbf{r} \circ \varphi : \mathbb{R} \to \mathbb{R}^n$ is continuous.

Proof. Since $\mathbf{r}(t)$ is continuous, given $\epsilon > 0$ there exists $\eta > 0$ such that $||\mathbf{r}(\varphi(t)) - \mathbf{r}(\varphi(t_0))|| < \epsilon$ whenever $|\varphi(t) - \varphi(t_0)| < \eta$. Since $\varphi(t)$ is continuous, corresponding to η there exists $\delta > 0$ such that $|\varphi(t) - \varphi(t_0)| < \eta$ whenever $|t - t_0| < \delta$. Thus, given $\epsilon > 0$, there exists $\delta > 0$ such that $||\mathbf{r}(\varphi(t)) - \mathbf{r}(\varphi(t_0))|| < \epsilon$ whenever $|t - t_0| < \delta$. □

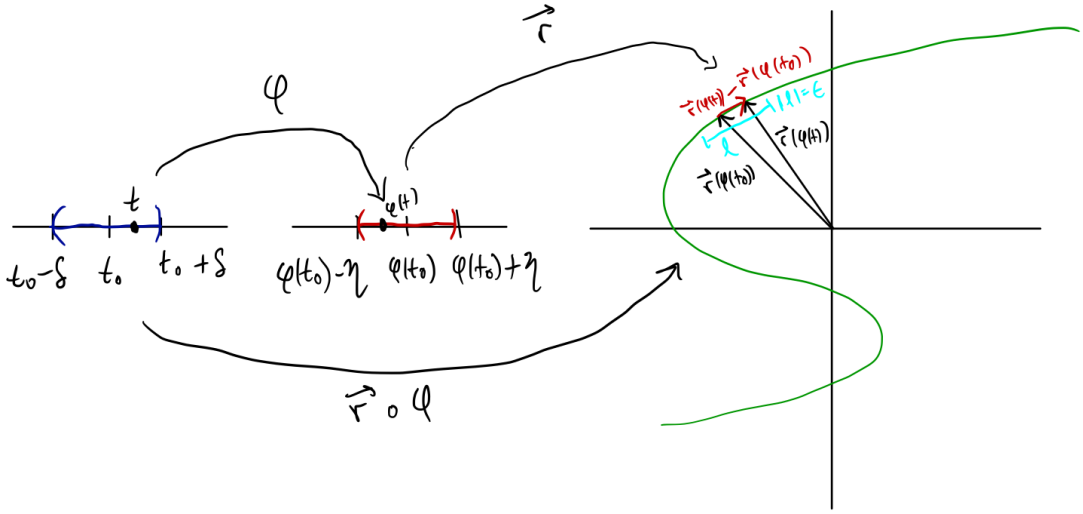


Figure 11: The composition of a continuous vector-valued function on the left of a continuous real-valued function is continuous.

Example 1.24. Let $\mathbf{r}(t) = (\sin t, \cos t, e^t)$ and let $\varphi(t) = t^2 - 1$. Since $\mathbf{r}, \varphi$ are both continuous, so is

$$\mathbf{r}(\varphi(t)) = (\sin(t^2-1), \cos(t^2-1), e^{t^2-1}).$$

1.1.4 Derivatives of vector-valued functions

Definition 1.25 (Derivative of a vector-valued function). The *derivative* of a vector-valued function $\mathbf{r}(t)$ is the limit

$$\mathbf{r}'(t) = \lim_{h\to 0} \frac{\mathbf{r}(t+h) - \mathbf{r}(t)}{h}.$$

The function $\mathbf{r}$ is said to be *differentiable at* t_0 if $\mathbf{r}'(t_0)$ exists, and $\mathbf{r}$ is said to be *differentiable on* (a, b) if $\mathbf{r}'(t)$ exists for all $t \in (a, b)$.

Geometrically, $\mathbf{r}'(t_0)$ is the *tangent vector* to the curve $\mathbf{r}$ at $\mathbf{r}(t)$.

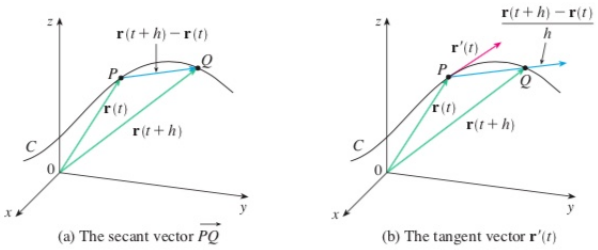


Figure 12: The derivative $\mathbf{r}'(t)$ is the tangent vector to the curve $\mathbf{r}$ at $\mathbf{r}(t)$.

The next theorem shows that we can compute the derivative of a vector-valued function in terms of the derivatives of its component functions.

Theorem 1.26 (Derivative of a vector-valued function). The derivative of a vector-valued function $\mathbf{r}(t) = (r_1(t), r_2(t), r_3(t))$ is given by

$$\mathbf{r}'(t) = (r_1'(t), r_2'(t), r_3'(t)).$$

Proof. By Theorem 1.15

$$\begin{aligned}
\mathbf{r}'(t) &= \lim_{h\to 0} \frac{\mathbf{r}(t+h) - \mathbf{r}(t)}{h} \\
&= \lim_{h\to 0} \frac{(r_1(t+h), r_2(t+h), r_3(t+h)) - (r_1(t), r_2(t), r_3(t))}{h} \\
&= \lim_{h\to 0} \frac{(r_1(t+h) - r_1(t), r_2(t+h) - r_2(t), r_3(t+h) - r_3(t))}{h} \\
&= \lim_{h\to 0} \left(\frac{r_1(t+h) - r_1(t)}{h}, \frac{r_2(t+h) - r_2(t)}{h}, \frac{r_3(t+h) - r_3(t)}{h} \right) \\
&= \left(\lim_{h\to 0} \frac{r_1(t+h) - r_1(t)}{h}, \lim_{h\to 0} \frac{r_2(t+h) - r_2(t)}{h}, \lim_{h\to 0} \frac{r_3(t+h) - r_3(t)}{h} \right) \\
&= (r_1'(t), r_2'(t), r_3'(t)).
\end{aligned}$$

□

Example 1.27. Let $\mathbf{r}(t) = (\sin t, \cos t, e^t)$. Then

$$\begin{aligned}\mathbf{r}'(\varphi(t)) &= ((\sin t)', (\cos t)', (e^t)') \\ &= (\cos t, -\sin t, e^t).\end{aligned}$$

Theorem 1.28 (Differentiation formulas). Let $\mathbf{u}, \mathbf{v}$ be differentiable vector-valued functions into $\mathbb{R}^n$, c a constant, and $\varphi : \mathbb{R} \to \mathbb{R}$ a real-valued function. Then

(1) $[\mathbf{u}(t) + \mathbf{v}(t)]' = \mathbf{u}'(t) + \mathbf{v}'(t)$

(2) $[c\mathbf{u}(t)]' = c\mathbf{u}'(t)$

(3) $[\varphi(t)\mathbf{u}(t)]' = \varphi'(t)\mathbf{u}(t) + \varphi(t)\mathbf{u}'(t)$

(4) $[\mathbf{u}(t) \cdot \mathbf{v}(t)]' = \mathbf{u}'(t) \cdot \mathbf{v}(t) + \mathbf{u}(t) \cdot \mathbf{v}'(t)$

(5) $[\mathbf{u}(t) \times \mathbf{v}(t)]' = \mathbf{u}'(t) \times \mathbf{v}(t) + \mathbf{u}(t) \times \mathbf{v}'(t)$ (for $n = 3$)

(6) $[\mathbf{u}(\varphi(t))]' = \varphi'(t)\mathbf{u}'(\varphi(t))$

Proof. The proof follows immediately from Theorem 1.26 and the corresponding properties for real-valued functions. □

Definition 1.29 (Higher derivatives). (a) The *second derivative*, $\mathbf{r}''$, of a vector-valued function $\mathbf{r}$ is the derivative of $\mathbf{r}'$: $\mathbf{r}'' = (\mathbf{r}')'$. Thus,

$$\mathbf{r}''(t) = (f''(t), g''(t), h''(t))$$

Similarly, for k a positive integer, the kth derivative of $\mathbf{r}$ is given by the formula

$$\mathbf{r}^{(k)}(t) = (f^{(k)}(t), g^{(k)}(t), h^{(k)}(t)).$$

(b) A vector-valued function is said to be *of class* $\mathscr{C}^k$ if its first k derivatives exist and are continuous and *of class* $\mathscr{C}^\infty$ if all of its derivatives exist. Functions in the class $\mathscr{C}^\infty$ are also called *smooth*.

1.1.5 Integrals of Vector valued functions

Similar to derivatives, we define the integral of a vector-valued function in terms of the integrals of its component functions.

Definition 1.30 (Integral of a vector-valued function). If $\mathbf{r}(t) = (r_1(t), \ldots, r_n(t))$ is a vector-valued function, then

$$\int_a^b \mathbf{r}(t)dt = (\int_a^b r_1(t)dt, \cdots, \int_a^b r_n(t)dt).$$

Indefinite integrals are defined similarly.

Example 1.31. (a) Let $\mathbf{r}(t) = (\cos t, 1, -2t)$. Then

$$\begin{aligned}\int (\cos t, 1, -2t)dt &= \left(\int \cos t dt, \int dt, -\int 2t dt\right) \\ &= (\sin t + c_1, t + c_2, -(t^2 + c_3)) \\ &= (\sin t, t, -t^2) + (c_1, c_2, -c_3).\end{aligned}$$

(b) Now a definite integral:

$$\begin{aligned}\int_0^{\pi}(\cos t, 1, -2t)dt &= \left(\int_0^{\pi}\cos t dt, \int_0^{\pi} dt, -\int_0^{\pi} 2t dt\right)\\ &= (\sin t|_0^{\pi}, t|_0^{\pi}, -t^2|_0^{\pi})\\ &= (0-0, \pi-0, -(\pi^2-0^2))\\ &= (0, \pi, -\pi^2).\end{aligned}$$

Exercise 1.6. Solve the first order differential equation

$$\mathbf{r}'(t) = (\cos t, -\sin t, 1)$$

subject to the initial condition $\mathbf{r}(0) = (2, 0, 1)$.

Solution. This differential equation can be solved by integrating with respect to t:

$$\begin{aligned}\mathbf{r}(t) &= \int \mathbf{r}'(t)dt\\ &= (\int \cos t dt, -\int \sin t dt, \int dt)\\ &= (\sin t, \cos t, t) + (c_1, c_2, c_3).\end{aligned}$$

The initial condition says that

$$\begin{aligned}\mathbf{r}(0) &= (\sin 0, \cos 0, 0) + (c_1, c_2, c_3)\\ &= (0, 1, 0) + (c_1, c_2, c_3)\\ &= (c_1, c_2 + 1, c_3)\\ &= (2, 0, 1),\end{aligned}$$

and therefore

$$\begin{aligned}c_1 &= 2\\ c_2 + 1 &= 0\\ c_3 &= 1.\end{aligned}$$

Thus, the solution to the differential equation is

$$\begin{aligned}\mathbf{r}(t) &= (\sin t, \cos t, t) + (2, -1, 1)\\ &= (\sin t + 2, \cos t - 1, t + 1).\end{aligned}$$

□

1.2 Parametrized Curves

We will now focus on a special class of vector-valued functions, which model the motion of a particle through space. Our interest will primarily be in the geometry of the particle's trajectory.

Definition 1.32 (Parametrized curve). Let I be an interval. A *parametrized curve* is a smooth mapping $\mathbf{r} : I \to \mathbb{R}^n$. For $n = 2$, a curve is also called a *plane curve* while for $n = 3$ it is also called a *space curve*. The variable t is called the *parameter*. The image $\mathbf{r}(I) \subseteq \mathbb{R}^n$ is called the *trace* of the curve $\mathbf{r}$. [3]

Remark. Every interval in $\mathbb{R}$ is of one of the following forms:

$$(-\infty, b), (-\infty, b], (a, b), [a, b), (a, b], [a, b], [a, \infty), (a, \infty).$$

If I is an interval containing boundary points, such as $[a, b]$, then we define $f'(a)$ as the right-hand limit

$$f'(a) = \lim_{h \to 0^+} \frac{f(a+h) - f(a)}{h}$$

and, $f'(b)$ as the left-hand limit

$$f'(b) = \lim_{h \to 0^-} \frac{f(b+h) - f(b)}{h}.$$

□

To avoid the proliferation of primes in what follows, we will use the following physical terminology.

Definition 1.33 (Velocity, speed, accleration). (a) We will call the tangent vector $\mathbf{r}'(t)$ the *velocity* at t. We call its length $||\mathbf{r}'(t)||$ the *speed* at t.

(b) We call the vector $\mathbf{r}''(t)$ the *acceleration* at t.
From now on we will write $\mathbf{v}(t) \equiv \mathbf{r}'(t)$, $||\mathbf{v}(t)|| \equiv ||\mathbf{r}'(t)||$, and $\mathbf{a}(t) \equiv \mathbf{r}''(t)$.

[3] The *image of I under* $\mathbf{r}$ is the set $\mathbf{r}(I) = \{\mathbf{r}(t) : t \in I\}$.

Example 1.34. The graph of any smooth function $y = f(x)$ can be written as a parametrized curve by defining

$$\mathbf{r} : \mathbb{R} \to \mathbb{R}^2$$
$$\mathbf{r}(t) = (t, f(t)).$$

For example, the function $y = x^2$ can be written the parametrized curve $\mathbf{r}(t) = (t, t^2)$ for all $-\infty < t < \infty$. The trace of a plane curve can be plotted in *Mathematica* as shown below:

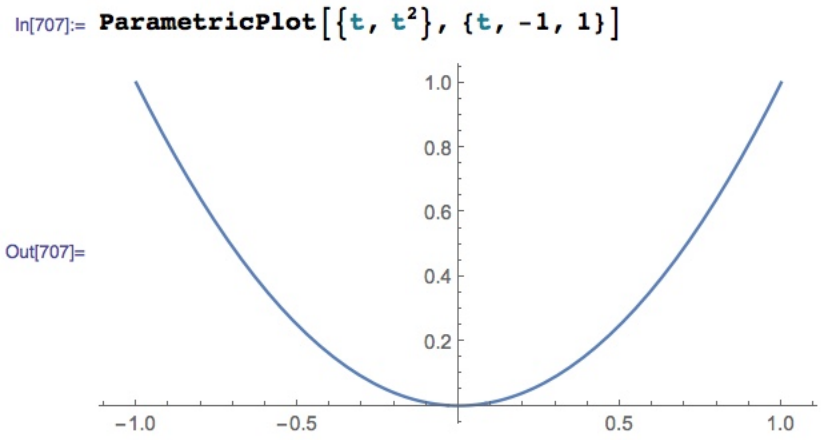


Figure 13: The trace of the plane curve $\mathbf{r}(t) = (t, t^2)$.

The particle moves down the parabola from the upper left starting at $t = -\infty$, reaches the origin at $t = 0$, and then continues up the parabola to the upper right for all $t > 0$. [4]

Exercise 1.7. Sketch the trace of the plane curve $\mathbf{r}(t) = (t^2 - 2t, t + 1)$, $-\infty < t < \infty$ in by making a table of the coordinates $(x(t), y(t))$ for integer values of t from $-2 \leq t \leq 4$. Compare your sketch with the curve produced in *Mathematica* using the "ParametricPlot" command.

Solution. The curve is a parabola, which we can confirm by eliminating t as follows. First, solve for t in terms of y to obtain

$$y = t + 1 \implies y - 1 = t.$$

Substituting into $x(t)$ then gives x in terms of y:

$$x = t^2 - 2t = (y-1)^2 - 2(y-1) = y^2 - 4y + 3 = (y-2)^2 - 1.$$

which is the equation of a parabola in vertex form, with vertex at $(x, y) = (-1, 2)$.

The particle travels upwards along the parabola from $t = -\infty$, reaches the vertex at $t = 1$, and then continues upward to the right for $t > 0$. □

[4] Some nice animations of parametrized curves can be found at https://demonstrations.wolfram.com/ParametricTrace/.

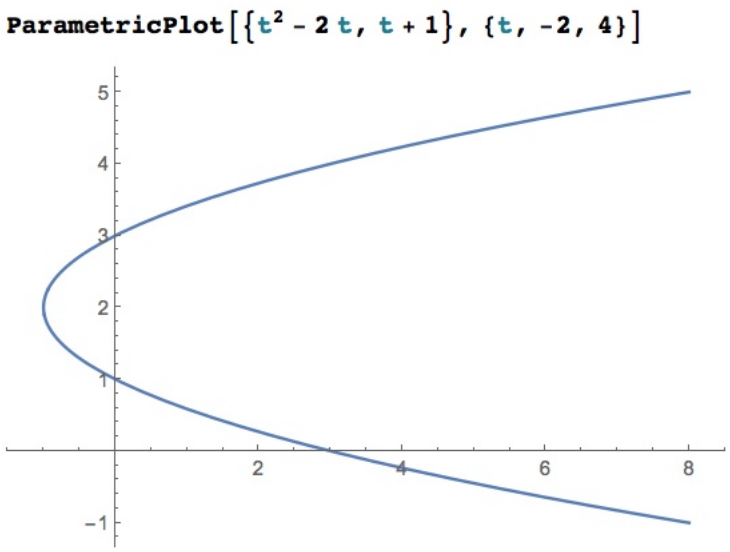


Figure 14: The trace of the plane curve $\mathbf{r}(t) = (t^2 - 2t, t + 1)$ from $-2 \leq t \leq 4$.

Example 1.35. Consider the following three plane curves

(1) $\mathbf{r}_1(t) = (\cos t, \sin t), 0 \leq t \leq 2\pi$,

(2) $\mathbf{r}_2(t) = (-\sin 2t, \cos 2t), 0 \leq t \leq 2\pi$,

(3) $\mathbf{r}_3(t) = (\cos(-t), \sin(-t)), 0 \leq t \leq 2\pi$.

For all three curves, $x(t)^2 + y(t)^2 = 1$, so the trace of each curve is the unit circle. However, as trajectories of a particle, they are different. The reader can easily verify the following by making a table for each curve and sketching the trace:

(1) In $\mathbf{r}_1$, the particle begins at $(x, y) = (1, 0)$ at $t = 0$ and completes one full circle, moving in the counterclockwise direction.

(2) In $\mathbf{r}_2$, the particle begins at $(x, y) = (0, 1)$ at $t = 0$ and completes *two* full circles, moving in the counterclockwise direction.

(3) In $\mathbf{r}_3$, the particle begins at $(x, y) = (1, 0)$ at $t = 0$ and completes one full circle, moving in the *clockwise* direction.

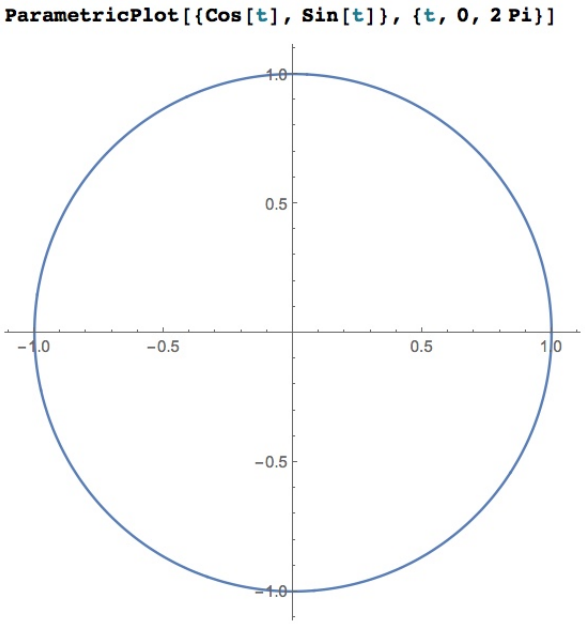
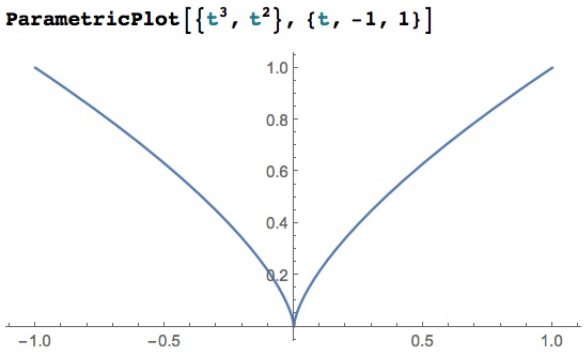


Figure 15: The trace of the curves $\mathbf{r}_1, \mathbf{r}_2, \mathbf{r}_3$.

The previous example shows that a parametrized curve is more than just the trace: it is the trace together with a choice of how to traverse the trace of curve.

Example 1.36. Consider now the trace of the plane curve $\mathbf{r}(t) = (t^3, t^2)$, $-\infty < t < \infty$.



Since the component functions are smooth, the reader may be surprised by the cusp at the origin. Solving for y in terms of x by eliminating t, we find that $y = x^{2/3}$, which is indeed not differentiable at the origin. Note that $\mathbf{v}(t) = (3t^2, 2t) \neq (0, 0)$ everywhere except the origin, where the tangent vector vanishes.

Definition 1.37 (Regular curve). Let $\mathbf{r} : I \to \mathbb{R}^n$ be a parametrized curve.

(a) A point where $\mathbf{v}(t) = \mathbf{0}$ is called a *singular point*, while a point where $\mathbf{v}(t) \neq \mathbf{0}$ is called a *regular point*.

(b) A *regular curve* is a curve with no singular points; that is, it is a curve for whice $\mathbf{v}(t) \neq \mathbf{0}$ for all $t \in I$.

In our study of the differential geometry of curves, we will assume our curve has a tangent vector at every point. Thus, from now on we will assume all curves are regular.

1.3 Reparametrizations

Recall the three regular curves from Example 1.35:

(1) $\mathbf{r}_1(t) = (\cos t, \sin t), 0 \leq t \leq 2\pi$,

(2) $\mathbf{r}_2(t) = (-\sin 2t, \cos 2t), 0 \leq t \leq 2\pi$,

(3) $\mathbf{r}_3(t) = (\cos(-t), \sin(-t)), 0 \leq t \leq 2\pi$.

We have noticed that all three curves have exactly the same trace; namely, the unit circle. If we are only interested in the geometry of the trace of the curve, then we should view all parametrized curves with the same trace as equivalent. We now formalize this idea.

First, notice that if we define the function

$$\varphi : [0, 2\pi] \to [0, 2\pi]$$

by

$$\varphi(t) = 2t + \frac{\pi}{2}$$

then we see that $\mathbf{r}_2$ is equal to the composition $\mathbf{r}_2 = \mathbf{r}_1 \circ \varphi$, since for all $t \in [0, 2\pi]$

$$\begin{aligned}
\mathbf{r}_1(\varphi(t)) &= (\cos(2t + \frac{\pi}{2}), \sin(2t + \frac{\pi}{2})) \\
&= (\cos(2t) \underbrace{\cos(\frac{\pi}{2})}_{=0} - \sin(2t) \underbrace{\sin(\frac{\pi}{2})}_{=1}, \sin(2t) \underbrace{\cos(\frac{\pi}{2})}_{=0} + \cos(2t) \underbrace{\sin(\frac{\pi}{2})}_{=1}) \\
&= (-\sin(2t), \cos(2t)) \\
&= \mathbf{r}_2(t).
\end{aligned}$$

To understand this relationship, recall that a parametrized curve $\mathbf{r}(t)$ is a function which gives the position of the particle at time t, with respect to a given clock. If φ denotes the time with respect to a different clock, then $\mathbf{r}(\varphi)$ gives the position of the particle at time φ. The particles follow the same trajectory, but are at different positions at different times if the two clocks are not synchronized. In this example, the clocks are related by $\varphi = 2t + \frac{\pi}{2}$; that is, the φ-clock is $\frac{\pi}{2}$ seconds ahead of the t-clock (since $\varphi = \frac{\pi}{2}$ when $t = 0$), and runs half as fast (since if the time interval between two successive ticks of the t-clock is $\Delta t = 1$ s, then the time interval between two successive ticks of the φ-clock is $\Delta\varphi = 2\Delta t = 2$ s; that is, if the t-clock ticks once every second, then the φ-clock ticks once every two seconds). Thus if $\mathbf{r}_1(t)$ describes a single trip around the unit circle in the counterclockwise sense at unit speed, then, with respect to the t-clock, $\mathbf{r}_1(\varphi) = \mathbf{r}_2(t)$ describes a particle that is already at $(0, 1)$ when $t = 0$ and completes two trips around the circle at double speed in 2π seconds.

Exercise 1.8. Verify that by defining $\varphi : [0, 2\pi] \to [0, 2\pi]$ by $\varphi(t) = 2\pi - t$, then $\mathbf{r}_3 = \mathbf{r}_1 \circ \varphi$. Interpret this as above.

Solution. We see that $\mathbf{r}_1(\varphi(t)) = (\cos(2\pi - t), \sin(2\pi - t)) = (\cos(-t), \sin(-t)) = \mathbf{r}_3(t)$ for all t. The φ-clock is related to the t-clock by "time-reversal"; that is, the clocks run at the same speed, but the φ clock runs backwards with respect to the t-clock, since $\varphi(t)$ is a decreasing function of t with $\varphi(0) = 2\pi$ and $\varphi(2\pi) = 0$. □

Definition 1.38 (Reparametrization). Let $\mathbf{r} : I \to \mathbb{R}^n$ be a regular curve. A *reparametrization* of $\mathbf{r}$ is a function of the form $\widetilde{\mathbf{r}} = \mathbf{r} \circ \varphi : \tilde{I} \to \mathbb{R}^n$, where $\tilde{I}$ is an interval and $\varphi : \tilde{I} \to I$ is a smooth bijection with nowhere vanishing derivative ($\varphi'(t) \neq 0$ for all $t \in \tilde{I}$). [5]

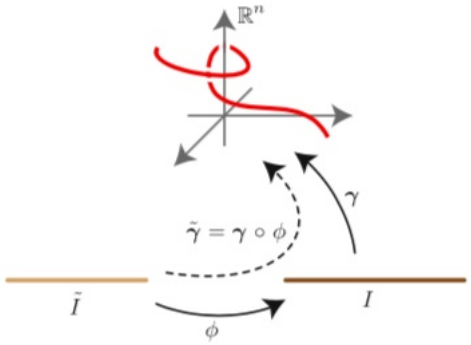


Figure 16: A reparametrization $\widetilde{\mathbf{r}} = \mathbf{r} \circ \varphi$ of a curve $\mathbf{r}$.

Remark. Note that the hypotheses on φ are all required to ensure that $\widetilde{\mathbf{r}} = \mathbf{r} \circ \varphi$ has the same trace as $\mathbf{r}$:

- The requirement that φ is smooth ensures that $\widetilde{\mathbf{r}}$ is smooth, since the composition of smooth functions is smooth.
- The requirement that φ is a bijection is also essential. If φ is not surjective, then the trace of $\widetilde{\mathbf{r}}$ will only be a proper subset of the trace of $\mathbf{r}$. If φ is not injective, then $\widetilde{\mathbf{r}}$ will have self-intersections at points where $\mathbf{r}$ does not (since if $\varphi(t_1) = \varphi(t_2)$ for $t_1 \neq t_2$, then even if $\mathbf{r}(t_1) \neq \mathbf{r}(t_2)$, we will have $\widetilde{\mathbf{r}}(\varphi(t_1)) = \widetilde{\mathbf{r}}(\varphi(t_2))$). Thus, if φ is not a bijection, then the two curves won't have the same trace.
- The requirement that $\varphi'(t) \neq 0$ for all t ensures that $\widetilde{\mathbf{r}}(t)$ is regular, since if $\mathbf{r}(t)$ is regular and $\varphi'(t) \neq 0$, then by the chain rule

$$\widetilde{\mathbf{v}}(t) = \varphi'(t)\mathbf{r}'(t) \neq 0.$$

□

Proposition 1.39 (Reparametrized curves are equivalent). The relation $\widetilde{\mathbf{r}} \sim \mathbf{r}$ if $\widetilde{\mathbf{r}}$ is a reparametrization of $\mathbf{r}$ is an equivalence relation.

[5] While in each of the examples above we had $\varphi : \tilde{I} \to I$ with $\tilde{I} = I$, these intervals need not be the same. For instance, in Example 1.8, we could have instead taken $\varphi : [0, 2\pi] \to [-2\pi, 0]$ where $\varphi(t) = -t$.

Proof. (1) (Reflexivity) Let $\mathbf{r} : I \to \mathbb{R}^n$ be a parametrized curve. Take $\varphi = id_I : T \to I$ to be the identity map on I. This is a smooth bijection, and since $id_I(t) = t$, $id_I'(t) = 1 \neq 0$ for all $t \in I$. Since $\mathbf{r} = \mathbf{r} \circ id_I$, $\mathbf{r} \sim \mathbf{r}$. Thus, $\sim$ is reflexive.

(2) (Symmetry) Suppose $\mathbf{r} : I \to \mathbb{R}^n$ and $\widetilde{\mathbf{r}} : \tilde{I} \to \mathbb{R}^n$ are parametrized curves with $\widetilde{\mathbf{r}} \sim \mathbf{r}$. Then there exists a smooth bijection $\varphi : \tilde{I} \to I$ with $\varphi'(t) \neq 0$ for all $t \in \tilde{I}$ such that $\widetilde{\mathbf{r}} = \mathbf{r} \circ \varphi$. It follows from the Inverse Function Theorem that φ has a smooth inverse, $\varphi^{-1} : I \to \tilde{I}$, where $(\varphi^{-1})'(t) = 1/\varphi'(t)$ Since $\varphi'(t)$ is never zero, neither is $(\varphi^{-1})'(t)$. [6] Then

$$\begin{aligned}\widetilde{\mathbf{r}} \circ \varphi^{-1} &= \mathbf{r} \circ \varphi^{-1} \circ \varphi \\ &= \mathbf{r} \circ id_I \\ &= \mathbf{r}.\end{aligned}$$

This proves that $\widetilde{\mathbf{r}} \sim \mathbf{r}$ implies that $\mathbf{r} \sim \widetilde{\mathbf{r}}$. Thus, $\sim$ is symmetric.

(3) (Transitivity) Let

$$\begin{aligned}\mathbf{r}_1 &: I_1 \to \mathbb{R}^n, \\ \mathbf{r}_2 &: I_2 \to \mathbb{R}^n, \\ \mathbf{r}_3 &: I_3 \to \mathbb{R}^n,\end{aligned}$$

be parametrized curves and suppose $\mathbf{r}_1 \sim \mathbf{r}_2$ and $\mathbf{r}_2 \sim \mathbf{r}_3$. Then there exist smooth bijections

$$\begin{aligned}\varphi &: I_2 \to I_1, \\ \psi &: I_3 \to I_2\end{aligned}$$

with nowhere vanishing first derivatives such that

$$\begin{aligned}\mathbf{r}_2 &= \mathbf{r}_1 \circ \varphi, \\ \mathbf{r}_3 &= \mathbf{r}_2 \circ \psi\end{aligned}$$

Then $\varphi \circ \psi : I_3 \to I_1$ is a smooth bijection with nowhere vanishing derivative such that

$$\mathbf{r}_3 = \mathbf{r}_1 \circ \varphi \circ \psi.$$

Thus, $\mathbf{r}_1 \sim \mathbf{r}_3$.

This shows that $\sim$ is reflexive, symmetric, and transitive, and is therefore an equivalence relation. □

Definition 1.40 (Curve). The we call the equivalence class

$$[\mathbf{r}] \equiv \{\widetilde{\mathbf{r}} : \widetilde{\mathbf{r}} = \mathbf{r} \circ \varphi \text{ for some } \varphi\}$$

a *curve* in $\mathbb{R}^n$. Any parametrized curve $\widetilde{\mathbf{r}}(t)$ in $[\mathbf{r}]$ is said to be a *representative* of the equivalence class $[\mathbf{r}]$.

[6] The *Inverse Function Theorem* states that if $\varphi : I \to \mathbb{R}$ is continuously differentiable with $\varphi'(t_0) \neq 0$ then φ is invertible in a neighborhood of t_0, the inverse is continuously differentiable, and the derivative of the inverse function at $b = \varphi(t_0)$ is the reciprocal of the derivative of φ at t_0. Applying this to the k-th derivative, it follows as a corollary that we can replace "continuously differentiable" everywhere in the theorem above by "smooth". For a regular curve $\varphi'(t)$ is never zero on I, so φ must be either increasing or decreasing on I. Either way φ is 1-1 on all of I, and therefore the "neighborhood" of t_0 appearing in the theorem is all of I.

Let $\mathbf{r}(t)$ be a parametrized curve and let $\widetilde{\mathbf{r}} = \mathbf{r} \circ \varphi$ be a reparametrization of $\mathbf{r}(t)$. Since $\varphi'(t)$ is nowhere 0, φ must be either monotonically increasing ($\varphi' > 0$) or monotonically decreasing ($\varphi' < 0$) on all of $\tilde{I}$.

Definition 1.41 (Oriented curves). A reparametrization $\widetilde{\mathbf{r}} = \mathbf{r} \circ \varphi$ is said to be *orientation-preserving* if $\varphi' > 0$ and *orientation-reversing* if $\varphi' < 0$. Thus, each curve $[\mathbf{r}]$ is partitioned into two subsets

$$[\mathbf{r}]_+ = \{\widetilde{\mathbf{r}} : \widetilde{\mathbf{r}} = \mathbf{r} \circ \varphi \text{ with } \varphi' > 0\},$$
$$[\mathbf{r}]_- = \{\widetilde{\mathbf{r}} : \widetilde{\mathbf{r}} = \mathbf{r} \circ \varphi \text{ with } \varphi' < 0\}.$$

Each subset is called an *oriented curve*. A choice of orientation of a curve $[\mathbf{r}]$ is a choice of one of these two subsets.

Exercise 1.9. Consider again three parametrized curves from Example 1.35 and :

(1) $\mathbf{r}_1(t) = (\cos t, \sin t), 0 \leq t \leq 2\pi$,

(2) $\mathbf{r}_2(t) = (-\sin 2t, \cos 2t), 0 \leq t \leq 2\pi$,

(3) $\mathbf{r}_3(t) = (\cos(-t), \sin(-t)), 0 \leq t \leq 2\pi$.

Which of these curves have the same orientation?

Solution. We have seen that $\mathbf{r}_2 = \mathbf{r}_1 \circ \varphi$ where $\varphi(t) = 2t + \frac{\pi}{2}$. Since $\varphi'(t) = 2 > 0$, these two curves have the same orientation (i.e., they are representatives of the same oriented curve). We also saw that $\mathbf{r}_3 = \mathbf{r}_1 \circ \psi$ where $\psi(t) = 2\pi - t$. Since $\psi'(t) = -1 < 0$, these two curves have opposite orientation. □

1.4 Arc Length

We have just seen that a given curve can be parametrized in many ways. Out of these many options, there is a natural choice for the parameter t, namely, the *arc length* measured from any point $\mathbf{r}_0$ on the curve. This is because the arc length is an *invariant* of the curve, which is independent of the parametrization. For a curve given by the graph of a function $y = f(x)$, the arc length from a to b is given by

$$s = \int_a^b \sqrt{1 + (f'(x))^2}\, dx. \tag{1.3}$$

Writing the function as a parametrized curve $\mathbf{r}(t) = (t, f(t))$, we have $\mathbf{v}(t) = (1, f'(t))$ and $||\mathbf{v}(t)|| = \sqrt{1 + (f'(t))^2}$, so we can write Equation (1.3) as

$$s = \int_a^b ||\mathbf{v}(t)||\, dt. \tag{1.4}$$

How do we compute the arc length if $\mathbf{r}(t)$ is not the graph of a function? If $\mathbf{r}(t)$ is not the graph of a function, it turns out that Equation (1.4) is still valid. In this case, we obtain the formula as the limit of a sequence of polygonal approximations of the curve. For those interested in the details, see Appendix A.

Definition 1.42. Arc length Let $\mathbf{r} : [a, b] \to \mathbb{R}^n$ be a regular curve. We define the *arc length* between the points $\mathbf{r}(a)$ and $\mathbf{r}(b)$ on the curve by

$$s = \int_a^b ||\mathbf{v}(t)||dt.$$

Example 1.43. We now compute the arc length along the portion of the unit circle from $(1, 0)$ to $(0, 1)$. If we use the parametrization

$$\mathbf{r}_1(t) = (\cos t, \sin t), \quad 0 \leq t \leq \frac{\pi}{2},$$

then we have

$$\begin{aligned} \mathbf{v}_1(t) &= (-\sin t, \cos t), \\ ||\mathbf{v}_1(t)|| &= \sqrt{(-\sin t)^2 + (\cos t)^2} = \sqrt{1} = 1. \end{aligned}$$

Equation (1.4) then gives

$$s = \int_0^{\frac{\pi}{2}} ||\mathbf{v}_1(t)||dt = \int_0^{\frac{\pi}{2}} dt = \frac{\pi}{2}.$$

Exercise 1.10. Use the parametrization $\mathbf{r}_2(t) = (\cos(2t), \sin(2t))$, $0 \leq t \leq \frac{\pi}{4}$ to compute the arc length of the same portion of the unit circle as in Example 1.43. Verify that you get the same answer.

Solution. Using the parametrization $\mathbf{r}_2(t) = (\cos(2t), \sin(2t))$, $0 \leq t \leq \frac{\pi}{4}$, we have

$$\begin{aligned} \mathbf{v}_2(t) &= (-2\sin(2t), 2\cos(2t)), \\ ||\mathbf{v}_2(t)|| &= \sqrt{4\sin^2(2t) + 4\cos^2(2t)} = \sqrt{4(\sin^2(2t) + \cos^2(2t))} = \sqrt{4} = 2 \end{aligned}$$

and therefore

$$s = \int_0^{\frac{\pi}{4}} ||\mathbf{v}_2(t)||dt = 2\int_0^{\frac{\pi}{4}} dt = 2 \cdot \frac{\pi}{4} = \frac{\pi}{2}.$$

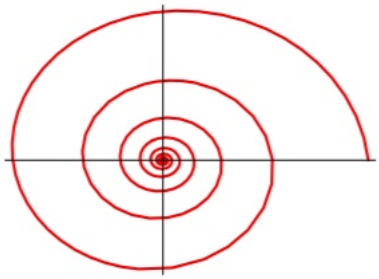
□

Exercise 1.11. A *logarithmic spiral* is a plane curve of the form $\mathbf{r}(t) = c(e^{\lambda t}\cos t, e^{\lambda t}\sin t)$, $t \in \mathbb{R}$ where $c, \lambda \in \mathbb{R}$ and $c \neq 0$. Below is the restriction of $\mathbf{r}(t)$ to $[0, \infty)$ with $\lambda < 0$.

Use an improper integral to prove that such a restriction has finite arc length even though it makes infinitely many loops around the origin.

Solution. We have

$$\begin{aligned} \mathbf{v}(t) &= c(\lambda e^{\lambda t}\cos t - e^{\lambda t}\sin t, \lambda e^{\lambda t}\sin t + e^{\lambda t}\cos t) \\ &= ce^{\lambda t}(\lambda\cos t - \sin t, \lambda\sin t + \cos t) \end{aligned}$$



and therefore

$$\begin{aligned}||\mathbf{v}(t)|| &= |c|e^{\lambda t}\sqrt{(\lambda\cos t - \sin t)^2 + (\lambda\sin t + \cos t)^2} \\ &= |c|e^{\lambda t}\sqrt{\lambda^2\cos^2 t + \sin^2 t - 2\lambda\sin t\cos t + \lambda^2\sin^2 t + \cos^2 t + 2\lambda\sin t\cos t} \\ &= |c|e^{\lambda t}\sqrt{\lambda^2+1}.\end{aligned}$$

The arc length is then given by the integral

$$s = \int_0^\infty ||\mathbf{v}(t)||dt = |c|\sqrt{\lambda^2+1}\int_0^\infty e^{-|\lambda|t}dt.$$

Substituting

$$u = -|\lambda|t,\ du = -|\lambda|dt,$$

this becomes

$$\begin{aligned}s &= -\frac{|c|\sqrt{\lambda^2+1}}{|\lambda|}\int_0^{-\infty} e^u du \\ &= -\frac{|c|\sqrt{\lambda^2+1}}{|\lambda|}\lim_{k\to-\infty}(e^k - 1) \\ &= -\frac{|c|\sqrt{\lambda^2+1}}{|\lambda|}(0-1) \\ &= \frac{|c|\sqrt{\lambda^2+1}}{|\lambda|}.\end{aligned}$$

□

Proposition 1.44 (Invariance of arc length). The arc length is independent of parametrization.

Proof. Let $\widetilde{\mathbf{r}} = \mathbf{r}\circ\varphi$ be a reparametrization of $\mathbf{r}$. Then since

$$\widetilde{\mathbf{v}}(t) = [\mathbf{r}(\varphi(t))]' = \varphi'(t)\mathbf{v}(\varphi(t)),$$

we have

$$\begin{aligned} s &= \int_{t_0}^{t_1} ||\widetilde{\mathbf{v}}(t)||dt \\ &= \int_{t_0}^{t_1} ||\varphi'(t)\mathbf{v}(\varphi(t))||dt \\ &= \int_{t_0}^{t_1} ||\mathbf{v}(\varphi(t))||\,|\varphi'(t)|dt \end{aligned}$$

If $\varphi'(t) > 0$, then $|\varphi'(t)| = \varphi'(t)$ and by substituting

$$u = \varphi(t),\ du = \varphi'(t)dt$$

we obtain

$$s = \int_{\varphi(t_0)}^{\varphi(t_1)} ||\mathbf{v}(u)||du.$$

If $\varphi'(t) < 0$, then $|\varphi'(t)| = -\varphi'(t)$ and the same substitution gives

$$s = -\int_{\varphi(t_0)}^{\varphi(t_1)} ||\mathbf{v}(u)||du = \int_{\varphi(t_0)}^{\varphi(t_1)} ||\mathbf{v}(u)||du.$$

□

Proposition 1.45 (Any regular curve can be parametrized by arc length). Any regular curve $\mathbf{r} : I \to \mathbb{R}^n$ can be parametrized by arc length.

Proof. Choose $t_0 \in I$ and consider the arc length function $s : I \to \mathbb{R}$ defined by

$$s(t) = \int_{t_0}^{t} ||\mathbf{v}(u)||du.$$

Let $\tilde{I} = s(I)$ denote the image of I under s. By the Fundamental Theorem of Calculus, $s'(t) = ||\mathbf{v}(t)|| \neq 0$ (since the curve is regular), so by the Inverse Function Theorem s has an inverse $\varphi : \tilde{I} \to I$, which is also a smooth bijection with a nowhere-vanishing derivative. Then $\widetilde{\mathbf{r}} = \mathbf{r} \circ \varphi$ is parametrized by arc length, since $\mathbf{r}(\varphi(s))$ is the position of a point along the curve at the time when it achieves arc length s measured from t_0. □

Example 1.46. Consider the helix $\mathbf{r}(t) = (\cos t, \sin t, t)$, $t \in \mathbb{R}$.

Let us reparametrize the helix with respect to the arc length measured from $(1, 0, 0)$ in the direction of increasing t. Since $\mathbf{r}(0) = (1, 0, 0)$, following the proof above we define

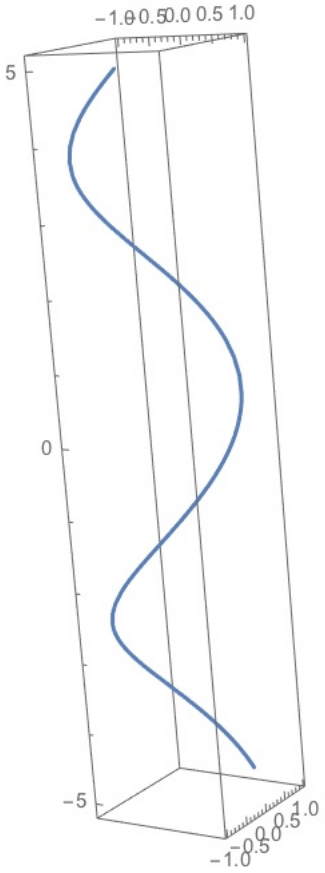
$$s(t) = \int_{0}^{t} ||\mathbf{v}(u)||du.$$

Since $\mathbf{v}(u) = (-\sin u, \cos u, 1)$, we have $||\mathbf{v}(u)|| = \sqrt{2}$ and therefore

$$\begin{aligned} s(t) &= \int_0^t ||\mathbf{v}(u)||du \\ &= \sqrt{2}\int_0^t du \\ &= \sqrt{2}t. \end{aligned}$$

```
In[728]:= ParametricPlot3D[{Cos[t], Sin[t], t}, {t, -5, 5}]
```

Out[728]=



The inverse function is then easily obtained:

$$\varphi(s) = \frac{s}{\sqrt{2}}$$

and

$$\mathbf{r}(\varphi(s)) = \left(\cos\left(\frac{s}{\sqrt{2}}\right), \sin\left(\frac{s}{\sqrt{2}}\right), \frac{s}{\sqrt{2}}\right).$$

Example 1.47. Let's reparametrzie the plane curve $\mathbf{r}(t) = (\cos(3t), \sin(2t))$ with respect to arc length measured from $(1,0)$ in the direction of increasing t. Since

$$\mathbf{v}(t) = (-3\sin(3t), 2\cos(2t))$$
$$||\mathbf{v}(t)|| = \sqrt{9\sin^2(3t) + 4^2\cos(2t)},$$

the arc length function is given by

$$s(t) = \int_0^t \sqrt{9\sin^2(3t) + 4\cos^2(2t)}dt$$

This integral cannot be evaluated in closed form. This example illustrates that, while our proof gave an explicit method to parametrize any curve by arc length, in practice it is usually not computationally reasonable to implement. However, the fact that every curve can be parametrized by arc length is very important for theoretical purposes, as we will see in the following sections.

Proposition 1.48. A curve parametrized by arc length is a *unit speed curve*; that is, $||\mathbf{v}(t)|| = 1$ for all $t \in I$.

Proof. If $\mathbf{r}$ is parametrized by arc length, then by the Fundamental Theorem of Calculus

$$s'(t) = \frac{d}{dt}\int_{t_0}^{t} ||\mathbf{v}(u)||du = ||\mathbf{v}(t)|| = 1.$$

Conversely, suppose $||\mathbf{v}(t)|| = 1$ for all t. Then

$$s(t) = \int_{t_0}^{t} ||\mathbf{v}(u)||du = \int_{t_0}^{t} du = t - t_0$$

so t is equal to the arc length accumulated from the point t_0. □

Exercise 1.12. Verify that the helix parametrized by arc length in Example 1.46 is a unit speed curve.

Solution. We have $\mathbf{v}(s) = \frac{1}{\sqrt{2}}(-\sin(\frac{s}{\sqrt{2}}), \cos(\frac{s}{\sqrt{2}}), 1)$ and therefore

$$\begin{aligned} ||\mathbf{v}(s)|| &= \frac{1}{\sqrt{2}}\sqrt{\sin^2(\frac{s}{\sqrt{2}}) + \cos^2(\frac{s}{\sqrt{2}}) + 1} \\ &= \frac{1}{\sqrt{2}}\sqrt{2} \\ &= 1. \end{aligned}$$

□

Exercise 1.13. Parametrize each of the following curves by arc length (measured from $t = 0$) and verify that $||\mathbf{v}(s)|| = 1$ for all s.

(1) Helix: $\mathbf{r}(t) = (a\cos t, a\sin t, bt)$, $-\infty < t < \infty$, a, b fixed positive real numbers.

(2) Logarithmic Spiral: $\mathbf{r}(t) = (e^t \cos t, e^t \sin t)$, $-\infty < t < \infty$.

Solution. (1) $\mathbf{r}(s) = (a\cos(\frac{s}{\sqrt{a^2+b^2}}), a\sin(\frac{s}{\sqrt{a^2+b^2}}), \frac{bs}{\sqrt{a^2+b^2}})$

(2) $\mathbf{r}(s) = ((\frac{s}{\sqrt{2}}+1)\cos(\ln(\frac{s}{\sqrt{2}}+1)), (\frac{s}{\sqrt{2}}+1)\sin(\ln(\frac{s}{\sqrt{2}}+1)))$

□

1.5 Curvature

In this section, we seek a quantity which measures how sharply a curve "bends" at each point. We call this number the *curvature* at that point. Since the curvature can vary from point to point along the curve, we seek a *curvature function* $\kappa : I \to \mathbb{R}^{\geq 0}$ which gives the curvature at each point along the curve. By definition, the curvature function should be identically zero for a straight line. Our curvature function should also be constant for a circle (due to the rotational symmetry), and inversely proportional to the radius (since a circle of smaller radius is "more curved" than a circle of a larger radius). We choose units such that $\kappa(s) = \frac{1}{R}$ for a circle of radius R.

1.5.1 Curvature of a unit speed curve

We begin by considering curves parametrized by arc length. The following lemma will be extremely useful for unit-speed curves.

Lemma 1.49. Let $\mathbf{r}_1, \mathbf{r}_2 : I \to \mathbb{R}^n$ be a pair of curves.

(a) If $\mathbf{r}_1$ has constant nonzero length (that is, if $||\mathbf{r}_1(t)|| = c > 0$ for all $t \in I$), then $\mathbf{r}_1'(t)$ is orthogonal to $\mathbf{r}(t)$ for all $t \in I$.

(b) If $\mathbf{r}_1(t)$ is orthogonal to $\mathbf{r}_2(t)$ for all $t \in I$, then

$$\mathbf{r}_1'(t) \cdot \mathbf{r}_2(t) = -\mathbf{r}_1(t) \cdot \mathbf{r}_2'(t) = 0$$

for all $t \in I$.

Proof. (1) Suppose $||\mathbf{r}_1(t)|| = c$ for all t. Differentiating both sides, gives $||\mathbf{r}_1(t)||' = 0$. That is,

$$\begin{aligned} 0 &= ||\mathbf{r}_1(t)||' \\ &= \sqrt{\mathbf{r}_1(t) \cdot \mathbf{r}_1(t)}' \\ &= \frac{2\mathbf{r}_1'(t) \cdot \mathbf{r}_1(t)}{2\sqrt{\mathbf{r}_1(t) \cdot \mathbf{r}_1(t)}} \\ &= \frac{\mathbf{r}_1'(t) \cdot \mathbf{r}_1(t)}{||\mathbf{r}_1(t)||} \\ &= \frac{\mathbf{r}_1'(t) \cdot \mathbf{r}_1(t)}{c} \end{aligned}$$

and therefore $\mathbf{r}_1'(t) \cdot \mathbf{r}_1(t) = 0$ for all t.

(2) Suppose $\mathbf{r}_1(t) \cdot \mathbf{r}_2(t) = 0$ for all t. Differentiating both sides, we obtain

$$0 = \mathbf{r}_1'(t) \cdot \mathbf{r}_2(t) + \mathbf{r}_1(t) + \mathbf{r}_2'(t)$$

and therefore

$$\mathbf{r}_1'(t) \cdot \mathbf{r}_2(t) = -\mathbf{r}_1(t) + \mathbf{r}_2'(t)$$

for all t.

□

Note that both of these hypotheses of Lemma 1.49 are true for an orthonormal set of vectors $\{\mathbf{r}_1, \mathbf{r}_2\}$. [7]

Let $\mathbf{r} : I \to \mathbb{R}^n$ be a curve parametrized by arc length. Since $\mathbf{v}(s)$ has unit length, by part (a) of Lemma 1.49, $\mathbf{a}(s)$ is orthogonal to $\mathbf{v}(s)$. Thus, $||\mathbf{a}(s)||$ measures the rate at which the curve is pulling away from the tangent line at $\mathbf{r}(s)$. The function $||\mathbf{a}(s)||$ is therefore a good candidate for our curvature function.

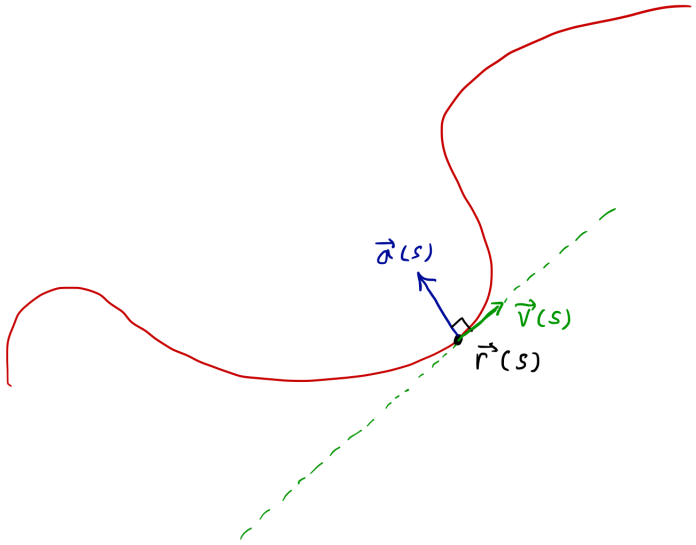


Figure 17: For a unit-speed curve, the magnitude of the acceleration vector measures the rate at which the curve is pulling away from the tangent line at $\mathbf{r}(s)$.

Note also that

(a) If $\mathbf{r}$ is a straight line, then $\mathbf{r}(s) = \hat{\mathbf{v}}(s_0)s + \mathbf{r}(s_0)$. We have $\mathbf{a}(s) \equiv 0$. Conversely, if $\kappa = ||\mathbf{a}(s)|| \equiv 0$, then $\mathbf{a}(s) \equiv 0$, and integrating twice gives $\mathbf{r}(s) = \hat{\mathbf{v}}(s_0)s + \mathbf{r}(s_0)$, so the curve is a straight line. Thus, a curve is a straight line if and only if $||\mathbf{a}(s)|| \equiv 0$.

(b) If $\mathbf{r}(s) = (R\cos(\frac{s}{R}), R\sin(\frac{s}{R}))$ is a circle of radius R, then

$$\mathbf{a}(s) = (-\frac{1}{R}\cos(\frac{s}{R}), -\frac{1}{R}\sin(\frac{s}{R}))$$

and therefore

$$\kappa(s) = ||\mathbf{a}(s)|| = \frac{1}{R}.$$

Thus, $||\mathbf{a}(s)||$ satisfies the requirements we set out in the beginning for our curvature function.

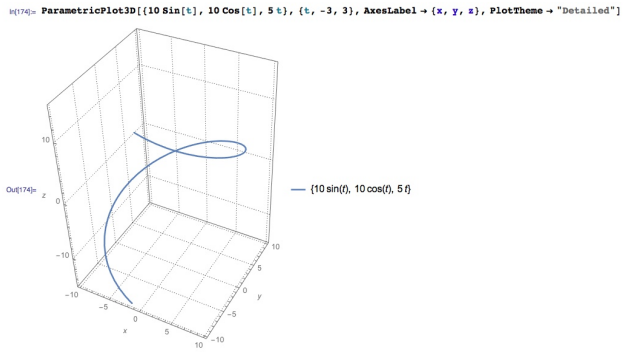
Definition 1.50 (Curvature). Let $\mathbf{r} : I \to \mathbb{R}^n$ be a curve parametrized by arc length $s \in I$. Define the *curvature function* $\kappa : I \to [0, \infty)$ by $\kappa(s) = ||\mathbf{a}(s)||$. The number $\kappa(s)$ is called the *curvature* of the curve $\mathbf{r}$ at s.

[7] See Section 5.4 of my Linear Algebra Notes for the definition of an orthonormal set of vectors.

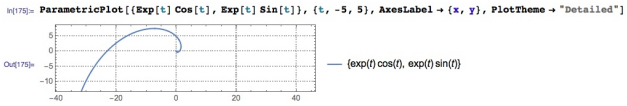
Exercise 1.14. Compute the curvature function of each of the following unit-speed curves:

(a) $\mathbf{r}(s) = (a\cos(\frac{s}{\sqrt{a^2+b^2}}), a\sin(\frac{s}{\sqrt{a^2+b^2}}), \frac{bs}{\sqrt{a^2+b^2}}),\ s > 0.$

(b) $\mathbf{r}(s) = ((\frac{s}{\sqrt{2}}+1)\cos(\ln(\frac{s}{\sqrt{2}}+1)), (\frac{s}{\sqrt{2}}+1)\sin(\ln(\frac{s}{\sqrt{2}}+1))),\ s > 0.$



(a) A portion of a helix.



(b) A portion of a logarithmic spiral.

Figure 18: The curves in Exercise 1.14.

Solution.

(a) $\kappa(s) = ||\mathbf{a}(s)|| = \frac{a}{a^2+b^2}$. Note that if $b = 0$, then the curve is a circle of radius a and $\kappa(s) = \frac{a}{a^2+0^2} = \frac{1}{a}$ as it should be. If $a = 0$, then the curve is a straight line, and $\kappa(s) = \frac{0}{0^2+b^2} \equiv 0$, again as expected.

(b) $\kappa(s) = ||\mathbf{a}(s)|| = \frac{\sqrt{2}}{\sqrt{2}s+2}$. Note that the curvature is highest at $s = 0$, and monotonically decreases as s increases.

□

Definition 1.51 (Unit normal vector). At points where $\kappa(s) \neq 0$, the vector

$$\hat{\mathbf{n}}(s) = \frac{\mathbf{a}(s)}{||\mathbf{a}(s)||} = \frac{\mathbf{a}(s)}{\kappa(s)}$$

is a unit vector orthogonal to $\mathbf{v}(s)$. This is called the *unit normal vector* to $\mathbf{r}$ at s.

The vectors $\{\hat{\mathbf{t}}, \hat{\mathbf{n}}\}$ play an important role in the geometry of the curve in the neighborhood of a point s_0 where $\kappa(s_0) \neq 0$, which we will see next.

Exercise 1.15. Compute the unit tangent and unit normal vectors for the helix

$$\mathbf{r}(s) = (a\cos(\frac{s}{\sqrt{a^2+b^2}}), a\sin(\frac{s}{\sqrt{a^2+b^2}}), \frac{bs}{\sqrt{a^2+b^2}}),\ s > 0.$$

Solution.

$$\hat{\mathbf{t}}(s) = \mathbf{v}(s) = \frac{1}{\sqrt{a^2+b^2}}(-a\sin(\frac{s}{\sqrt{a^2+b^2}}), a\cos(\frac{s}{\sqrt{a^2+b^2}}), \frac{b}{\sqrt{a^2+b^2}})$$

$$\hat{\mathbf{n}}(s) = \frac{\mathbf{a}(s)}{||\mathbf{a}(s)||} = (-\cos(\frac{s}{\sqrt{a^2+b^2}}), -\sin(\frac{s}{\sqrt{a^2+b^2}}), 0)$$

□

1.5.2 Osculating Plane

Let $\mathbf{r}(s)$ be a unit-speed curve. Consider a point s_0 where $\kappa(s_0) \neq 0$. Then, for sufficiently small $h = s - s_0$, we can approximate $\mathbf{r}(s)$ by its second-order Taylor polynomial:

$$\mathbf{r}(s_0 + h) = \mathbf{r}(s_0) + h\mathbf{v}(s_0) + \frac{h^2}{2}\mathbf{a}(s_0) + \mathbf{E}(h),$$
$$\mathbf{r}(s_0 + h) = \mathbf{r}(s_0) + h\hat{\mathbf{t}} + \frac{h^2}{2}\kappa(s_0)\hat{\mathbf{n}}(s_0) + \mathbf{E}(h),$$

where $\lim_{h\to 0} \frac{||\mathbf{E}(h)||}{h^2} = 0$. Thus, sufficiently near $\mathbf{r}(s_0)$ (where we can ignore the error term $\mathbf{E}(h)$), the trace of a curve lies in the plane spanned by $\{\hat{\mathbf{t}}(s_0), \hat{\mathbf{n}}(s_0)\}$, called the *osculating plane at* $\mathbf{r}(s_0)$. With respect to the coordinate $\mathbf{D}(h) \equiv \mathbf{r}(s_0 + h) - \mathbf{r}(s_0)$ centered at $\mathbf{r}(s_0)$, the trace of the curve in the osculating plane is given by

$$\mathbf{D}(I) = \{(h, \frac{\kappa(s_0)}{2}h^2) : -\epsilon < h < \epsilon\} \subseteq \text{Span}\{\hat{\mathbf{t}}(s_0), \hat{\mathbf{n}}(s_0)\} \cong \mathbb{R}^2,$$

which is the parabola $y = \frac{\kappa(s_0)}{2}x^2$ whose vertex is at the origin and whose concavity is $y'' = \kappa(s_0)$.

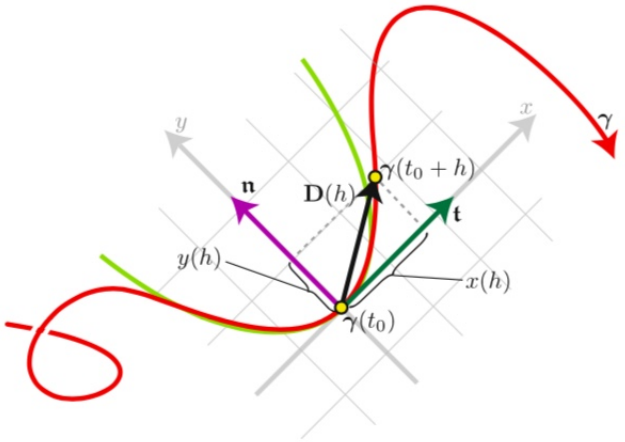


Figure 19: The trace of $\mathbf{r}$ is well-approximated near $\mathbf{r}(s_0)$ by the parabola with concavity $\kappa(s_0)$ in the osculating plane.

The trace of the curve in the neighborhood of a point s_0 where $\kappa(s_0) \neq 0$ is also well approximated by a certain circle in the osculating plane.

Definition 1.52 (Osculating circle). The *osculating circle* to is the circle of radius $\kappa(s_0)$ in the osculating plane centered at $\mathbf{r}(s_0) + \frac{1}{\kappa(s_0)}\hat{\mathbf{n}}$.

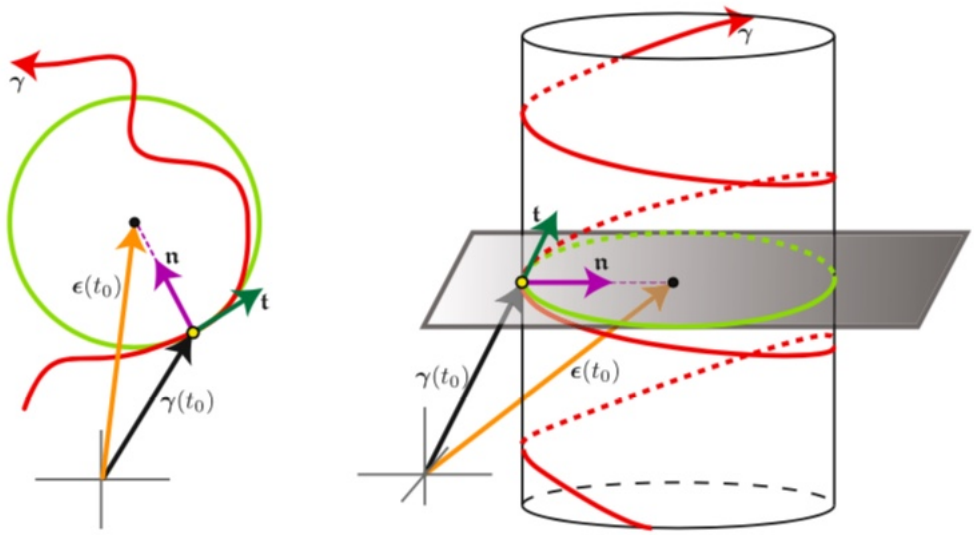


Figure 20: The osculating circle for a plane curve (left) and a space curve (right).

Thus, we may also interpret the curvature as the radius of the osculating circle.

Note that $s \mapsto \epsilon(s)$ is itself a parametrized curve (not necessarily regular) on any neighborhood of s_0 along which $\kappa(s_0) \neq 0$. This curve is called the *evolute* of $\mathbf{r}$. [8]

At points where $\kappa(s) = 0$, the normal vector (and therefore the osculating plane) is not defined. To proceed with our analysis, we will need the osculating plane, so from now on we will only consider curves for which $\kappa(s) \neq 0$ for all $s \in I$.

1.5.3 Arbitrary parametrizations

As noted previously, in practice it is often too difficult to parametrize a curve by arc length, and we are forced to work with other parametrizations.

Let $\mathbf{r} : I \to \mathbb{R}^n$ be a curve with an arbitrary parametrization $t \in I$. Since the speed $||\mathbf{v}(t)||$ is no longer constant, the acceleration vector $\mathbf{a}(t)$ will have a nonzero component in the direction of the velocity vector $\mathbf{v}(t)$. We might therefore be tempted to define the curvature function $\kappa(t)$ as the magnitude of the component of the acceleration perpendicular to $\mathbf{v}(t)$; that is, to define $\kappa(t) := ||\mathbf{a}_\perp(t)||$, where

$$\begin{aligned}\mathbf{a}_\perp(t) &= \mathbf{a}(t) - \text{proj}_{\mathbf{v}(t)}\mathbf{a}(t) \\ &= \mathbf{a}(t) - \frac{\mathbf{a}(t)\cdot\mathbf{v}(t)}{\mathbf{v}(t)\cdot\mathbf{v}(t)}\mathbf{v}(t).\end{aligned}$$

[8] A nice demonstration of osculating circles and evolutes can be found at https://demonstrations.wolfram.com/EvolutesOfSomeBasicCurves/.

However, this definition is problematic, as the next example shows.

Example 1.53. Consider the parabola which is the graph of $y = x^2$. Describe the parabola as a parametrized curve in the following two parametrizations,

$$\begin{aligned}\mathbf{r}_1(t) &= (t, t^2), & t \in \mathbb{R},\\ \mathbf{r}_2(t) &= (2t-2, (2t-2)^2), & t \in \mathbb{R}.\end{aligned}$$

In the first parametrization, the origin is the point $\mathbf{r}(0) = (0,0)$. Since

$$\begin{aligned}\mathbf{v}_1(t) &= (1, 2t),\\ \mathbf{a}_1(t) &= (0, 2),\end{aligned}$$

$$\begin{aligned}\mathbf{a}_{1\,\perp}(0) &= (0,2) - \frac{(0,2)\cdot(1,0)}{(1,0)\cdot(1,0)}(1,0)\\ &= (0,2),\end{aligned}$$

and therefore $\kappa(0) = ||\mathbf{a}_{1\,\perp}(0)|| = 2$.

In the second parametrization, the origin is $\mathbf{r}_2(1) = (0,0)$. Since

$$\begin{aligned}\mathbf{v}_2(t) &= (2, 4(2t-2)) = (2, 8t-8)\\ \mathbf{a}_2(t) &= (0, 8)\end{aligned}$$

we have

$$\begin{aligned}\mathbf{a}_{2\,\perp}(1) &= (0,8) - \frac{(0,8)\cdot(2,0)}{(2,0)\cdot(2,0)}(2,0)\\ &= (0,8)\end{aligned}$$

and therefore

$$\kappa(1) = ||\mathbf{a}_{2\,\perp}(1)|| = 8.$$

Using our definition of $\kappa(t)$, we have just obtained *different* values of the curvature at the origin of the parabola by using different parametrizations. This makes no sense, as curvature is a geometric property of the curve that does not depend on parametrization, so this definition is bad. The problem, as we will see in a moment, is that $||\mathbf{a}_\perp(t)||$ depends not only on the shape of the curve, but also on the speed at time t.

To find a satisfactory definition of curvature, we need to study more carefully the dependence of the derivatives of $\mathbf{r}(t)$ on parametrization. Our function $\kappa(t)$ should be independent of parametrization, and should agree with our previous definition when $t = s$ is the arc length.

Let $\mathbf{r} : I \to \mathbb{R}^n$ be a regular curve with an arbitrary paramtrization, and let $\tilde{\mathbf{r}} = \mathbf{r} \circ \varphi$ be a reparametrization of $\mathbf{r}$. Then, by the chain rule,

$$\begin{aligned}\widetilde{\mathbf{v}}(t) &= \mathbf{v}(\varphi(t))\varphi'(t),\\ \widetilde{\mathbf{a}}(t) &= \mathbf{a}(\varphi(t))(\varphi'(t))^2 + \mathbf{v}(\varphi(t))\varphi''(t),\\ \widetilde{\mathbf{a}}_\perp(t) &= \mathbf{a}_\perp(\varphi(t))(\varphi'(t))^2.\end{aligned}$$

From these expressions, we see that the quantity

$$\frac{||\mathbf{a}_\perp(t)||}{||\mathbf{v}(t)||^2} \tag{1.5}$$

is invariant under reparametrizations. Moreover, when $t = s$ is the arc length, $||\mathbf{v}(s)|| = 1$ and $\mathbf{a}(s) = \mathbf{a}_\perp(s)$, so the expression in (1.5) agrees with our definition of the curvature function of a unit speed curve in Definition 1.50. Thus, we have arrived at our desired definition of the curvature function in an arbitrary parametrization.

Definition 1.54 (Curvature in an arbitrary parametrization). Let $\mathbf{r} : I \to \mathbb{R}^n$ be a regular curve with an arbitrary parametrization $t \in I$. We define the *curvature function* $\kappa : I \to [0, \infty)$ for all $t \in I$ by

$$\kappa(t) = \frac{||\mathbf{a}_\perp(t)||}{||\mathbf{v}(t)||^2}. \tag{1.6}$$

Rewriting (1.6) as

$$||\mathbf{a}_\perp(t)|| = \kappa(t)||\mathbf{v}(t)||^2,$$

this expression says, in physical terms, that the magnitude of the centripetal acceleration at t depends not only on the curvature, but also on the speed at t. This explains the results of Example 1.53.

Definition 1.55 (Unit tangent and Unit normal in an arbitrary parametrization). Let $\mathbf{r} : I \to \mathbb{R}^n$ be a regular curve with an arbitrary parametrization and let $t \in I$ be a point where $\kappa(t) \neq 0$. We define the *unit tangent* and *unit normal vectors* at t to be

$$\hat{\mathbf{t}}(t) = \frac{\mathbf{v}(t)}{||\mathbf{v}(t)||}, \quad \hat{\mathbf{n}}(t) = \frac{\mathbf{a}_\perp(t)}{||\mathbf{a}_\perp(t)||}.$$

The next proposition shows that we can compute the unit normal vector in terms of the unit tangent vector.

Proposition 1.56. Let $\mathbf{r} : I \to \mathbb{R}^n$ be a regular curve. For any point where $\kappa(t) \neq 0$, we have

$$\hat{\mathbf{n}}(t) = \frac{\hat{\mathbf{t}}'(t)}{||\hat{\mathbf{t}}'(t)||}.$$

Proof. By the quotient rule,

$$\hat{\mathbf{t}}'(t) = \frac{||\mathbf{v}(t)||\mathbf{a}(t) - \mathbf{v}(t)||\mathbf{v}(t)||'}{||\mathbf{v}(t)||^2}.$$

Now

$$\begin{aligned} ||\mathbf{v}(t)||' &= \frac{d}{dt}(\mathbf{v}(t) \cdot \mathbf{v}(t))^{1/2} \\ &= \frac{1}{2}(\mathbf{v}(t) \cdot \mathbf{v}(t))^{-1/2}(2\mathbf{a}(t) \cdot \mathbf{v}(t)) \\ &= \frac{\mathbf{a}(t) \cdot \mathbf{v}(t)}{||\mathbf{v}(t)||} \end{aligned}$$

so we have

$$\begin{aligned}\hat{\mathbf{t}}'(t) &= \frac{||\mathbf{v}(t)||\mathbf{a}(t) - \mathbf{v}(t)||\mathbf{v}(t)||'}{||\mathbf{v}(t)||^2} \\ &= \frac{\mathbf{a}(t)}{||\mathbf{v}(t)||} - \frac{\mathbf{a}(t)\cdot\mathbf{v}(t)}{||\mathbf{v}(t)||^3}\mathbf{v}(t) \\ &= \frac{1}{||\mathbf{v}(t)||}\left(\mathbf{a}(t) - \frac{\mathbf{a}(t)\cdot\mathbf{v}(t)}{||\mathbf{v}(t)||^2}\mathbf{v}(t)\right) \\ &= \frac{\mathbf{a}_\perp(t)}{||\mathbf{v}(t)||}.\end{aligned}$$

and therefore

$$\begin{aligned}\frac{\hat{\mathbf{t}}'(t)}{||\hat{\mathbf{t}}'(t)||} &= \frac{\frac{\mathbf{a}_\perp(t)}{||\mathbf{v}(t)||}}{\frac{||\mathbf{a}_\perp(t)||}{||\mathbf{v}(t)||}} \\ &= \frac{\mathbf{a}_\perp(t)}{||\mathbf{a}_\perp(t)||} \\ &\equiv \hat{\mathbf{n}}(t).\end{aligned}$$

□

[Assign exercises from the book.]

1.6 Space Curves

In this section, we will consider regular curves $\mathbf{r} : I \to \mathbb{R}^3$ where $\kappa(t) \neq 0$ for all $t \in I$. In the previous section, we found that the unit tangent vector, unit normal vector, and curvature function are given by

$$\hat{\mathbf{t}}(t) = \frac{\mathbf{v}(t)}{||\mathbf{v}(t)||}, \quad \hat{\mathbf{n}}(t) = \frac{\hat{\mathbf{t}}'(t)}{||\hat{\mathbf{t}}'(t)||},$$
$$\kappa(t) = \frac{||\mathbf{a}_\perp(t)||}{||\mathbf{v}(t)||^2}.$$

In $\mathbb{R}^3$, we can use the cross-product to simplify the computation of the curvature function.

Proposition 1.57. Let $\mathbf{r} : I \to \mathbb{R}^3$ be a regular space curve. Then for all $t \in I$,

$$\kappa(t) = \frac{||\mathbf{v}(t) \times \mathbf{a}(t)||}{||\mathbf{v}(t)||^3}.$$

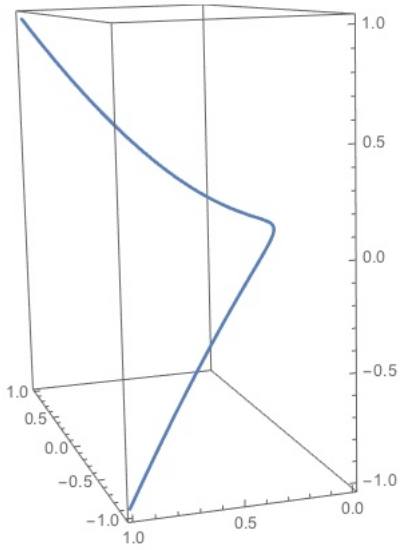
Proof. Since $||\mathbf{a}_\perp(t)|| = ||\mathbf{a}(t)|| \sin\theta$ (where θ is the angle between $\mathbf{v}(t)$ and $\mathbf{a}(t)$), we have

$$\kappa(t) = \frac{||\mathbf{a}_\perp(t)||}{||\mathbf{v}(t)||^2} = \frac{||\mathbf{a}(t)|| \sin\theta}{||\mathbf{v}(t)||^2} = \frac{||\mathbf{v}(t)||\,||\mathbf{a}(t)||}{||\mathbf{v}(t)||^3} = \frac{||\mathbf{v}(t) \times \mathbf{a}(t)||}{||\mathbf{v}(t)||^3}.$$

□

In[731]:= `ParametricPlot3D[{t, t^2, t^3}, {t, -1, 1}]`

Out[731]=



Example 1.58. The space curve $\mathbf{r}(t) = (t, t^2, t^3)$, $t \in \mathbb{R}$ is called the *twisted cubic*. The unit tangent and unit normal vectors are rather tedious to compute by hand. These can easily be computed in *Mathematica* as follows:

One can easily verify that these vectors form an orthonormal set: The curvature function is then given by

$$\begin{aligned}
\kappa(t) &= \frac{1}{(1+4t^2+9t^4)^{3/2}} \left|\left| \begin{vmatrix} \hat{\mathbf{i}} & \hat{\mathbf{j}} & \hat{\mathbf{k}} \\ 1 & 2t & 3t^2 \\ 0 & 2 & 6t \end{vmatrix} \right|\right| \\
&= \frac{2||(1,-3t,3t^2)||}{(1+4t^2+9t^4)^{3/2}} \\
&= \frac{2\sqrt{1+9t^2+9t^4}}{(1+4t^2+9t^4)^{3/2}} \\
&= 2\sqrt{\frac{1+9t^2+9t^4}{(1+4t^2+9t^4)^3}},
\end{aligned}$$

or using *Mathematica*:

We can also use the cross product to add one more vector to the set $\{\hat{\mathbf{t}}(t), \hat{\mathbf{n}}(t)\}$.

Definition 1.59 (Unit binormal vector). The vector

$$\hat{\mathbf{b}}(t) = \hat{\mathbf{t}}(t) \times \hat{\mathbf{n}}(t)$$

is called the *unit binormal vector*.

Proposition 1.60. The set $\{\hat{\mathbf{t}}(t), \hat{\mathbf{n}}(t), \hat{\mathbf{b}}(t)\}$ is an orthonormal basis for $\mathbb{R}^3$. [9]

[9] Recall that an *orthonormal basis* for a vector space V is a basis B for V such that all vectors in B have unit length and all pairs of distinct vectors in B are orthogonal. For more details, see Section 5.4 of my linear algebra notes.

```
In[752]:= r[t_] := {t, t^2, t^3};

In[753]:= T[t_] := Simplify[r'[t] / Norm[r'[t]], Assumptions → t ∈ Reals]

In[754]:= n[t_] := Simplify[T'[t] / Norm[T'[t]], Assumptions → t ∈ Reals]

In[762]:= T[t]

Out[762]= {1/Sqrt[1 + 4 t^2 + 9 t^4], 2 t/Sqrt[1 + 4 t^2 + 9 t^4], 3 t^2/Sqrt[1 + 4 t^2 + 9 t^4]}

In[764]:= n[t]

Out[764]= {-(t (2 + 9 t^2))/Sqrt[(1 + 4 t^2 + 9 t^4) (1 + 9 t^2 + 9 t^4)],
  (1 - 9 t^4)/Sqrt[1 + 13 t^2 + 54 t^4 + 117 t^6 + 81 t^8], (3 t (1 + 2 t^2))/Sqrt[(1 + 4 t^2 + 9 t^4) (1 + 9 t^2 + 9 t^4)]}

In[767]:= Simplify[Norm[T[t]], Assumptions → t ∈ Reals]
Out[767]= 1

In[768]:= Simplify[Norm[n[t]], Assumptions → t ∈ Reals]
Out[768]= 1

In[769]:= Simplify[Dot[T[t], n[t]], Assumptions → t ∈ Reals]
Out[769]= 0
```

Proof. We have already seen that $\{\hat{\mathbf{t}}(t), \hat{\mathbf{n}}(t)\}$ is an orthonormal set. It follows immediately from a property of the cross product that $\hat{\mathbf{b}}(t)$ is orthogonal to both $\hat{\mathbf{t}}(t)$ and $\hat{\mathbf{n}}(t)$. [10] Also

$$||\hat{\mathbf{b}}(t)|| = ||\hat{\mathbf{t}}(t)||\,||\hat{\mathbf{n}}(t)||\,|\sin\left(\frac{\pi}{2}\right)| = 1 \cdot 1 \cdot 1 = 1.$$

Thus, $\{\hat{\mathbf{t}}(t), \hat{\mathbf{n}}(t), \hat{\mathbf{b}}(t)\}$ is an orthonormal basis for $\mathbb{R}^3$. □

Definition 1.61 (Frenet frame). A basis for $\mathbb{R}^n$ together with a choice of origin is called a *frame*. The orthonormal frame $\{\hat{\mathbf{t}}(t), \hat{\mathbf{n}}(t), \hat{\mathbf{b}}(t)\}$ for $\mathbb{R}^3$ (whose origin is understood to be $\mathbf{r}(t)$) is called the *Frenet frame*.

The curvature function of the curve $\mathbf{r}$ gave us a quantitative measure of to what extent the trace of the curve fails to be a straight line in a neighborhood of the point $\mathbf{r}(t)$. We now want a quantitative measure of the extent to which the curve fails to lie in the osculating plane in a neighborhood of $\mathbf{r}(t)$. Similar to the definition of curvature, a good candidate for such a function is $||\mathbf{b}'(t)||$, which gives the rate at which the osculating plane is tilting at $\mathbf{r}(t)$. However, we will instead find it more useful to define a *signed* measurement of the rate at which the osculating plane is tilting.

[10] See Proposition 3.24(a) in my linear algebra notes.

```
In[760]:= κ = Simplify[Norm[Cross[r'[t], r''[t]]] / (Norm[r'[t]])^3, Assumptions → t ∈ Reals]
```

$$\text{Out[760]= } 2\sqrt{\frac{1+9t^2+9t^4}{\left(1+4t^2+9t^4\right)^3}}$$

First, note that $\hat{\mathbf{b}}'(t)$ is parallel to $\hat{\mathbf{n}}(t)$. To see this, note that note that since $\{\hat{\mathbf{t}}(t), \hat{\mathbf{n}}(t), \hat{\mathbf{b}}(t)\}$ is an orthonormal basis, we can expand $\hat{\mathbf{b}}'$ in this basis as [11]

$$\hat{\mathbf{b}}'(t) = (\hat{\mathbf{b}}'(t)\cdot\hat{\mathbf{t}}(t))\hat{\mathbf{t}}(t) + (\hat{\mathbf{b}}'(t)\cdot\hat{\mathbf{n}}(t))\hat{\mathbf{n}}(t) + (\hat{\mathbf{b}}'(t)\cdot\hat{\mathbf{b}}(t))\hat{\mathbf{b}}(t).$$

Since $||\hat{\mathbf{b}}(t)|| = 1$ for all $t \in I$, by part (a) of Lemma 1.49 we have $\hat{\mathbf{b}}'(t)\cdot\hat{\mathbf{b}}(t) = 0$. By part (b) of the same lemma we also have

$$\hat{\mathbf{b}}'(t)\cdot\hat{\mathbf{t}}(t) = -\hat{\mathbf{b}}(t)\cdot\hat{\mathbf{t}}'(t) = -\hat{\mathbf{b}}(t)\cdot(||\mathbf{t}'(t)||\hat{\mathbf{n}}(t)) = -||\mathbf{t}'(t)||(\hat{\mathbf{b}}(t)\cdot\hat{\mathbf{n}}(t)) = 0,$$

since $\hat{\mathbf{b}}(t)$ and $\hat{\mathbf{n}}(t)$ are orthogonal. Thus, $\hat{\mathbf{b}}'(t) = (\hat{\mathbf{b}}'(t)\cdot\hat{\mathbf{n}}(t))\hat{\mathbf{n}}(t)$.

Definition 1.62 (Torsion for unit speed curve). Let $\mathbf{r} : I \to \mathbb{R}^3$ be a regular curve parametrized by arc length s, such that $\kappa(s) \neq 0$ for all $s \in I$. The *torsion function* of $\mathbf{r}$ is defined by

$$\tau(s) = -\hat{\mathbf{b}}'(s)\cdot\hat{\mathbf{n}}(s).$$

The minus sign in the definition above is conventional, and many textbooks define torsion without this sign. Before we move on to interpreting this formula, we need a formula valid for a regular curve with arbitrary parametrization.

Proposition 1.63. Let $\mathbf{r} : I \to \mathbb{R}^3$ be a regular space curve with an arbitrary parametrization. Then for every $t \in I$ with $\kappa(t) \neq 0$, the expression

$$\tau(t) = \frac{-\hat{\mathbf{b}}'(t)\cdot\hat{\mathbf{n}}(t)}{||\mathbf{r}'(t)||} \tag{1.7}$$

is invariant under reparametrizations and agrees with the torsion function $\tau(s)$ of Definition 1.62.

Proof. If the curve is parametrized by arc length, then $||\mathbf{r}'(t)|| = 1$ for all t and thus the formula becomes that of Definition 1.62.

To see that this formula is independent of parametrization, let us first consider the transformation of the vectors in the Frenet frame change under a reparametrization:

$$\begin{aligned}
\widetilde{\hat{\mathbf{t}}}(t) &= \frac{\widetilde{\mathbf{v}}(t)}{||\widetilde{\mathbf{v}}(t)||} = \frac{\varphi'(t)\mathbf{v}(\varphi(t))}{||\varphi'(t)\mathbf{v}(\varphi(t))||} = \frac{\varphi'(t)\mathbf{v}(\varphi(t))}{|\varphi'(t)|||\mathbf{v}(\varphi(t))||} = \operatorname{sgn}(\varphi'(t))\hat{\mathbf{t}}(\varphi(t)) \\
\widetilde{\hat{\mathbf{n}}}(t) &= \frac{\widetilde{\mathbf{a}}_\perp(t)}{||\widetilde{\mathbf{a}}_\perp(t)||} = \frac{(\varphi'(t))^2\mathbf{a}_\perp(\varphi(t))}{||(\varphi'(t))^2\mathbf{a}_\perp(\varphi(t))||} = \frac{(\varphi'(t))^2\mathbf{a}_\perp(\varphi(t))}{(\varphi'(t))^2||\mathbf{a}_\perp(\varphi(t))||} = \hat{\mathbf{n}}(\varphi(t)) \\
\widetilde{\hat{\mathbf{b}}}(t) &= \widetilde{\hat{\mathbf{t}}}(t)\times\widetilde{\hat{\mathbf{n}}}(t) = \operatorname{sgn}(\varphi'(t))\hat{\mathbf{b}}(\varphi(t))
\end{aligned}$$

[11] See Equation 5.9 in my linear algebra notes.

If φ is an orientation-preserving reparametrization ($\varphi'(t) > 0$), then $\widetilde{\hat{\mathbf{t}}} = \hat{\mathbf{t}} \circ \varphi$, $\widetilde{\hat{\mathbf{n}}} = \hat{\mathbf{n}} \circ \varphi$, $\widetilde{\hat{\mathbf{b}}} = \hat{\mathbf{b}} \circ \varphi$, and

$$\widetilde{\tau}(t) = -\frac{\widetilde{\hat{\mathbf{b}}}'(t) \cdot \widetilde{\hat{\mathbf{n}}}(t)}{||\widetilde{\mathbf{r}}'(t)||} = \frac{-\varphi'(t)\hat{\mathbf{b}}(\varphi(t)) \cdot \hat{\mathbf{n}}(\varphi(t))}{||\varphi'(t)\mathbf{r}'(\varphi(t)||} = \frac{-\hat{\mathbf{b}}(\varphi(t)) \cdot \hat{\mathbf{n}}(\varphi(t))}{||\mathbf{r}'(\varphi(t)||} = \tau(\varphi(t)).$$

If φ is an orientation-reversion reparametrization ($\varphi'(t) < 0$), then $\widetilde{\hat{\mathbf{t}}} = -\hat{\mathbf{t}} \circ \varphi$, $\widetilde{\hat{\mathbf{n}}} = \hat{\mathbf{n}} \circ \varphi$, $\widetilde{\hat{\mathbf{b}}} = -\hat{\mathbf{b}} \circ \varphi$, and

$$\widetilde{\tau}(t) = -\frac{\widetilde{\hat{\mathbf{b}}}'(t) \cdot \widetilde{\hat{\mathbf{n}}}(t)}{||\widetilde{\mathbf{r}}'(t)||} = \frac{\varphi'(t)\hat{\mathbf{b}}(\varphi(t)) \cdot \hat{\mathbf{n}}(\varphi(t))}{-\varphi'(t)||\mathbf{r}'(\varphi(t)||} = \frac{-\hat{\mathbf{b}}(\varphi(t)) \cdot \hat{\mathbf{n}}(\varphi(t))}{||\mathbf{r}'(\varphi(t)||} = \tau(\varphi(t)).$$

so the sign changes in the formula cancel, yielding the same conclusion: $\widetilde{\tau} = \tau \circ \varphi$. Thus, the expression in (1.8) is invariant under reprametrizations. □

Definition 1.64. Let $\mathbf{r} : I \to \mathbb{R}^3$ be a regular space curve with an arbitrary parametrization. Then for every $t \in I$ with $\kappa(t) \neq 0$, we define the *torsion function* of the curve by the expression

$$\tau(t) = \frac{-\hat{\mathbf{b}}'(t) \cdot \hat{\mathbf{n}}(t)}{||\mathbf{r}'(t)||}. \tag{1.8}$$

Proposition 1.65. Let $\mathbf{r} : I \to \mathbb{R}^3$ be a regular space curve with $\kappa(t) \neq 0$ for all $t \in I$. Then the trace of $\mathbf{r}$ is constrained to a plane if and only if $\tau(t) = 0$ for all $t \in I$.

Proof. Let $\mathbf{w} = (a, b, c)$ be a fixed vector in $\mathbb{R}^3$ to serve as the normal vector of our plane. First, suppose that the trace of $\mathbf{r}$ is constrained to the plane

$$P = \{\mathbf{x} \in \mathbb{R}^3 : \mathbf{x} \cdot \mathbf{w} = d\}.$$

Since the trace of $\mathbf{r}$ lies in the plane P, $\mathbf{r}(t) \cdot \mathbf{w} = d$ is a constant function of t, so its derivatives vanish:

$$0 = (\mathbf{r}(t) \cdot \mathbf{w})' = \mathbf{v}(t) \cdot \mathbf{w}$$
$$0 = (\mathbf{r}(t) \cdot \mathbf{w})'' = \mathbf{a}(t) \cdot \mathbf{w}.$$

From this it follows that $\hat{\mathbf{t}}(t)$ and $\hat{\mathbf{n}}(t)$ are both orthogonal to $\mathbf{w}$, so their cross product must be parallel to $\mathbf{w}$: $\hat{\mathbf{b}}(t) = \pm\frac{\mathbf{w}}{||\mathbf{w}||}$. Since $\hat{\mathbf{b}}(t)$ is a continuous function of t, this sign cannot change abruptly, so it must be constant on I. Thus, $\hat{\mathbf{b}}$ is constant, so $\tau(t) = 0$ for all $t \in I$.

Conversely, suppose that $\tau(t) = 0$ for all $t \in I$. This implies that $\hat{\mathbf{b}}'(t) = 0$ for all $t \in I$, so $\hat{\mathbf{b}}(t) = \mathbf{w}$ (a constant vector) for all $t \in I$. Notice that

$$(\mathbf{w} \cdot \mathbf{r}(t))' = \mathbf{w} \cdot \mathbf{v}(t) = ||\mathbf{v}(t)||(\hat{\mathbf{b}}(t) \cdot \hat{\mathbf{t}}(t)) = 0,$$

since $\hat{\mathbf{b}}(t)$ and $\hat{\mathbf{t}}(t)$ are orthogonal. This $\mathbf{w} \cdot \mathbf{r}(t) = d$ is a constant function. In other words, the trace of $\mathbf{r}$ lies in the plane

$$P = \{\mathbf{x} \in \mathbb{R}^3 : \mathbf{x} \cdot \mathbf{w} = d\}.$$

□

Thus, roughly, torsion measures the failure of the trace of the curve to remain in a single plane. To formulate this idea more precisely, we must first compute the derivatives of the vectors in the Frenet frame. In the following analysis, we will assume that $\mathbf{r}$ is parametrized by arc length.

Proposition 1.66. Let $\mathbf{r} : I \to \mathbb{R}^3$ be a regular curve parametrized by arc length. At every time $s \in I$ with $\kappa(s) \neq 0$, the derivatives of the unit tangent, normal, and binormal vectors are given by the *Frenet equations*:

$$\begin{aligned}\hat{\mathbf{t}}'(s) &= \kappa(s)\hat{\mathbf{n}}(s)\\ \hat{\mathbf{n}}'(s) &= -\kappa(s)\hat{\mathbf{t}}(s) + \tau(s)\hat{\mathbf{b}}(s)\\ \hat{\mathbf{b}}'(s) &= -\tau(s)\hat{\mathbf{n}}(s).\end{aligned}$$

Proof. The first and third of these follow immediately from the definitions of κ and τ, since

$$\begin{aligned}\kappa(s) &= \hat{\mathbf{t}}'(s)\cdot\hat{\mathbf{n}}(s)\\ \tau(s) &= -\hat{\mathbf{b}}'(s)\cdot\hat{\mathbf{n}}(s).\end{aligned}$$

To show the second, expand $\hat{\mathbf{n}}'(s)$ in the Frenet frame as

$$\begin{aligned}\hat{\mathbf{n}}'(s) &= \underbrace{(\hat{\mathbf{n}}'(s)\cdot\hat{\mathbf{t}}(s))}_{=-\hat{\mathbf{n}}(s)\cdot\hat{\mathbf{t}}'(s)=-\kappa(s)}\hat{\mathbf{t}}(s) + \underbrace{(\hat{\mathbf{n}}'(s)\cdot\hat{\mathbf{n}}(s))}_{=0}\hat{\mathbf{n}}(s) + \underbrace{(\hat{\mathbf{n}}'(s)\cdot\hat{\mathbf{b}}(s))}_{-\hat{\mathbf{n}}(s)\cdot\hat{\mathbf{b}}'(s)=\tau(s)}\hat{\mathbf{b}}(s)\\ &= -\kappa(s)\hat{\mathbf{t}}(s) + \tau(s)\hat{\mathbf{b}}(s).\end{aligned}$$

□

We can now describe more precisely how torsion measures the failure of the trace of the curve to remain in a single plane. Assume now that $\mathbf{r} : I \to \mathbb{R}^3$ is a unit-speed curve, and that $s \in I$ with $\kappa(s) \neq 0$. We saw above that the trace of a second-order Taylor polynomial for $\mathbf{r}$ at s is a parabola in the osculating plane. We will now show that if $\tau(s) \neq 0$, the third-order Taylor polynomial for $\mathbf{r}$ at s leaves the osculating plane.

The third-order Taylor polynomial for the displacement vector $\mathbf{D}(h) \equiv \mathbf{r}(s+h) - \mathbf{r}(s)$ at $s \in I$ is given by

$$\mathbf{D}(h) \equiv \mathbf{r}(s+h) - \mathbf{r}(s) = h\mathbf{v}(s) + \frac{h^2}{2}\mathbf{a}(s) + \frac{h^3}{6}\mathbf{j}(s),$$

where, in physics, $\mathbf{j}(s) \equiv \mathbf{r}'''(s)$ is called the *jerk*, since a curve with $\mathbf{r}'''(s) \neq 0$ experinces jerky motion.

Using Proposition 1.66,

$$\begin{aligned}\mathbf{v}(s) &= \hat{\mathbf{t}}(s)\\ \mathbf{a}(s) &= \kappa(s)\hat{\mathbf{n}}(s)\\ \mathbf{j}(s) &= (\kappa(s)\hat{\mathbf{n}}(s))' = \kappa'(s)\hat{\mathbf{n}}(s) + \kappa(s)\hat{\mathbf{n}}'(s)\\ &= \kappa'(s)\hat{\mathbf{n}}(s) + \kappa(s)(-\kappa(s)\hat{\mathbf{t}}(s) + \tau(s)\hat{\mathbf{b}}(s))\\ &= \kappa'(s)\hat{\mathbf{n}}(s) - (\kappa(s))^2\hat{\mathbf{t}}(s) + \kappa(s)\tau(s)\hat{\mathbf{b}}(s)\end{aligned}$$

so the third-order Taylor polynomial of $\mathbf{D}(h)$ is

$$\begin{aligned}\mathbf{D}(h) &= h\hat{\mathbf{t}}(s) + \kappa(s)\frac{h^2}{2}\hat{\mathbf{n}}(s) + \frac{h^3}{6}(\kappa'(s)\hat{\mathbf{n}}(s) - (\kappa(s))^2\hat{\mathbf{t}}(s) + \kappa(s)\tau(s)\hat{\mathbf{b}}(s))\\ &= (h - (\kappa(s))^2\frac{h^3}{6})\hat{\mathbf{t}}(s) + (\kappa(s)\frac{h^2}{2} + \kappa'(s)\frac{h^3}{6})\hat{\mathbf{n}}(s) + \kappa(s)\tau(s)\frac{h^3}{6}\hat{\mathbf{b}}(s)\end{aligned}$$

If we choose a coordinate system centered at $\mathbf{r}(s)$ where $\hat{\mathbf{t}}(s) = (1,0,0), \hat{\mathbf{n}}(s) = (0,1,0)$, and $\hat{\mathbf{b}}(s) = (0,0,1)$, then we have

$$[\mathbf{D}(h)]_B = (x(h), y(h), z(h)) = \left(h - (\kappa(s))^2 \frac{h^3}{6}, \kappa(s)\frac{h^2}{2} + \kappa'(s)\frac{h^3}{6}, \kappa(s)\tau(s)\frac{h^3}{6} \right).$$

This representation is called the *local canonical form of* $\mathbf{r}$ *in a neighborhood of* s. Below, we plot the projections of the trace of $\mathbf{r}$, for small h, in the tn, tb, and nb planes. The tn plane is the osculating plane, while the tb and nb planes are called the *rectifying plane* and *normal plane*, respectively.

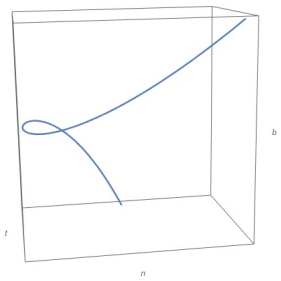
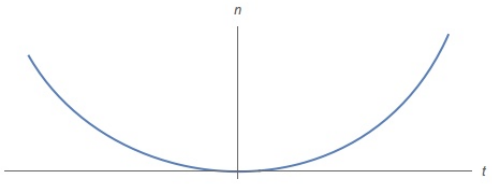
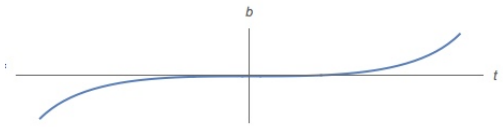


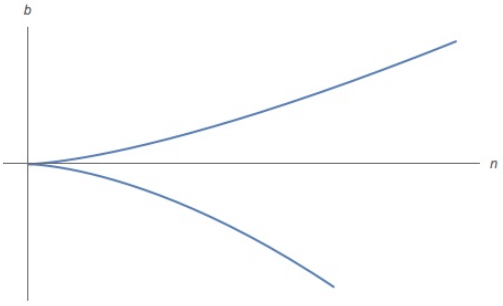
Figure 21: Trace of a regular space curve in the neighborhood of a point with non-zero curvature and torsion.



(a) Projection onto the osculating plane.



(b) Projection onto the rectifying plane.



(c) Projection onto the normal plane.

Figure 22: Projections of a regular space curve in the neighborhood of a point with non-zero curvature and torsion onto the osculating, rectifying, and normal planes.

Since $\kappa(s) > 0$, the Taylor polynomial for $z(h)$ implies that if $\tau(s) > 0$, then $z(h) > 0$ for sufficiently small positive h. On the other hand, if $\tau(s) < 0$, then $z(h) < 0$ for sufficiently small positive h. Thus *positive torsion* at s implies that the the curve $\mathbf{r}$ is passing through the osculating

plane at s from below. Negative torsion implies from above. Here "above" means the direction of $\hat{\mathbf{b}}(s)$.

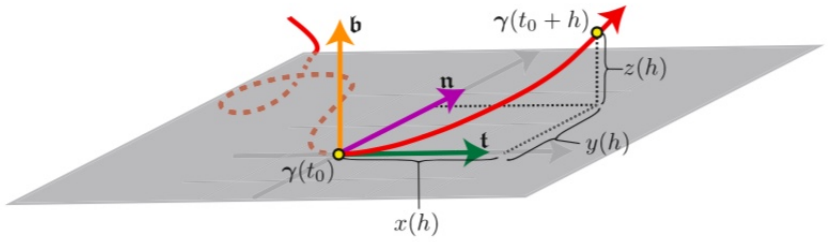


Figure 23: Positive torsion implies that the curve passes through the osculating plane from below.

To avoid taking the derivative of the cross product, here is another formula for the torsion which is often more computationally useful.

Proposition 1.67. Let $\mathbf{r} : I \to \mathbb{R}^3$ be a regular space curve with an arbitrary parametrization. Then for every $t \in I$ with $\kappa(t) \neq 0$, the torsion is given by

$$\tau(t) = \frac{\mathbf{v}(t) \times \mathbf{a}(t) \cdot \mathbf{j}(t)}{||\mathbf{v}(t) \times \mathbf{a}(t)||^2}. \tag{1.9}$$

Proof. We will prove that this formula holds for a curve parametrized by arc length, and that it is also independent of parametrization, which then implies it holds for a curve with any parametrization.

For a curve parametrized by arc length, $\mathbf{r}'(s) = \hat{\mathbf{t}}(s)$ and $\mathbf{r}''(s) = \kappa(s)\hat{\mathbf{n}}(s)$, and therefore $\mathbf{r}'(s) \times \mathbf{r}''(s) = \kappa(s)(\hat{\mathbf{t}}(s) \times \hat{\mathbf{n}}(s)) = \kappa(s)\hat{\mathbf{b}}(s)$ and $||\mathbf{r}'(s) \times \mathbf{r}''(s)|| = ||\kappa(s)\hat{\mathbf{b}}(s)|| = \kappa(s)$. Plugging into (1.9), we have

$$\begin{aligned}
\tau(s) &= \frac{\hat{\mathbf{b}}(s) \cdot \mathbf{r}'''(s)}{\kappa(s)} \\
&= \frac{\hat{\mathbf{b}}(s) \cdot (\kappa(s)\hat{\mathbf{n}}(s))'}{\kappa(s)} \\
&= \frac{\kappa'(s)\hat{\mathbf{b}}(s) \cdot \hat{\mathbf{n}}(s) + \kappa(s)\hat{\mathbf{b}}(s) \cdot \hat{\mathbf{n}}'(s)}{\kappa(s)} \\
&= \frac{\kappa'(s) \cdot 0 + \kappa(s)\hat{\mathbf{b}}(s) \cdot \hat{\mathbf{n}}'(s)}{\kappa(s)} \\
&= \hat{\mathbf{b}}(s) \cdot \hat{\mathbf{n}}'(s) \\
&= -\hat{\mathbf{b}}'(s) \cdot \hat{\mathbf{n}}(s)
\end{aligned}$$

in agreement with Definition 1.62.

To check that this formula is independent of parametrization, as we computed previously, if

$\widetilde{\mathbf{r}} = \mathbf{r} \circ \varphi$, then

$$\begin{aligned}
\widetilde{\mathbf{v}}(t) &= \varphi'(t)\mathbf{v}(\varphi(t)) \\
\widetilde{\mathbf{a}}(t) &= \varphi''(t)\mathbf{v}(\varphi(t)) + (\varphi'(t))^2\mathbf{a}(\varphi(t)) \\
\widetilde{\mathbf{j}}(t) &= \varphi'''(t)\mathbf{v}(\varphi(t)) + 3\varphi'(t)\varphi''(t)\mathbf{a}(\varphi(t)) + (\varphi'(t))^3\mathbf{j}(\varphi(t)).
\end{aligned}$$

Then

$$\begin{aligned}
\widetilde{\mathbf{v}}(t) \times \widetilde{\mathbf{a}}(t) &= \varphi'(t)\mathbf{v}(\varphi(t)) \times [\varphi''(t)\mathbf{v}(\varphi(t)) + (\varphi'(t))^2\mathbf{a}(\varphi(t))] \\
&= \varphi'(t)\varphi''(t)\underbrace{\mathbf{v}(\varphi(t)) \times \mathbf{v}(\varphi(t))}_{=0} + (\varphi'(t))^3(\mathbf{v}(\varphi(t)) \times \mathbf{a}(\varphi(t))) \\
&= (\varphi'(t))^3(\mathbf{v}(\varphi(t)) \times \mathbf{a}(\varphi(t)))
\end{aligned}$$

and therefore

$$\widetilde{\tau}(t) = \frac{(\varphi'(t))^3(\mathbf{v}(\varphi(t)) \times \mathbf{a}(\varphi(t))) \cdot (\varphi'''(t)\mathbf{v}(\varphi(t)) + 3\varphi'(t)\varphi''(t)\mathbf{a}(\varphi(t)) + (\varphi'(t))^3\mathbf{j}(\varphi(t)))}{(\varphi'(t))^6||\mathbf{r}'(\varphi(t)) \times \mathbf{a}(\varphi(t))||^2}$$

Using the cyclic property of the triple product [12], we have

$$\mathbf{v}(t) \times \mathbf{a}(t) \cdot \mathbf{v}(t) = \underbrace{\mathbf{v}(t) \times \mathbf{v}(t)}_{=0} \cdot \mathbf{a}(t) = 0$$

and

$$\mathbf{v}(t) \times \mathbf{a}(t) \cdot \mathbf{a}(t) = \underbrace{\mathbf{a}(t) \times \mathbf{a}(t)}_{=0} \cdot \mathbf{v}(t) = 0$$

so the formula becomes

$$\begin{aligned}
\widetilde{\tau}(t) &= \frac{(\varphi'(t))^6(\mathbf{v}(\varphi(t)) \times \mathbf{a}(\varphi(t))) \cdot \mathbf{j}(\varphi(t)))}{(\varphi'(t))^6||\mathbf{v}(\varphi(t)) \times \mathbf{a}(\varphi(t))||^2} \\
&= \frac{(\mathbf{v}(\varphi(t)) \times \mathbf{a}(\varphi(t))) \cdot \mathbf{j}(\varphi(t)))}{||\mathbf{v}(\varphi(t)) \times \mathbf{a}(\varphi(t))||^2} \\
&= \tau(\varphi(t))
\end{aligned}$$

which shows τ is invariant under reparametrizations. □

Example 1.68. Consider the twisted cubic $\mathbf{r}(t) = (t, t^2, t^3)$ of Example 1.58. The unit binormal vector is given by

The torsion function is

Notice that the torsion takes its minimal value of 3 at $t = 0$, and then strictly decreases as $|t| \to \infty$.

Physically, we can think of a curve in $\mathbb{R}^3$ a being obtained from a straight line by bending (curvature) and twisting (torsion).

[12] See Proposition 3.33 in my linear algebra notes

```
In[863]:= Simplify[Cross[T[t], n[t]], Assumptions → t ∈ Reals]
```

$$\text{Out[863]=} \left\{\frac{3\,t^2}{\sqrt{1+9\,t^2+9\,t^4}}, -\frac{3\,t}{\sqrt{1+9\,t^2+9\,t^4}}, \frac{1}{\sqrt{1+9\,t^2+9\,t^4}}\right\}$$

```
In[865]:= τ = Simplify[Dot[Cross[r'[t], r''[t]], r'''[t]] / (Norm[Cross[r'[t], r''[t]]])^2, Assumptions → t ∈ Reals]
```

$$\text{Out[865]=} \frac{3}{1+9\,t^2+9\,t^4}$$

Exercise 1.16. Show that the helix $\mathbf{r}(t) = (a\cos t, a\sin t, bt)$ has constant torsion $\tau = \frac{b}{a^2+b^2}$.

Theorem 1.69 (Fundamental Theorem of the Local Theory of Space Curves). Given smooth functions $\kappa(s) > 0$ and $\tau(s)$, $s \in I$, there exists a regular parametrized curve $\mathbf{r} : I \to \mathbb{R}^3$ such that s is the arc length, $\kappa(s)$ is the curvature, and $\tau(s)$ is the torsion of $\mathbf{r}$. This curve is unique up to rotations and translations.

The proof of this theorem requires differential equations prerequisites, but the theorem says that a space curve is uniquely determined by its curvature and torsion functions. Intuitively, given a straight line, these functions tell you how to obtain the space curve by "bending" (curvature) and "twisting" (torsion).

1.7 Summary of formulas for space curves

The following formulas hold for arbitrary parametrizations. Here, given a curve $\mathbf{r}$,

$$\begin{aligned}\mathbf{v} &\equiv \mathbf{r}',\\ \mathbf{a} &\equiv \mathbf{r}'',\\ \mathbf{j} &\equiv \mathbf{r}'''.\end{aligned}$$

1. Frenet frame

$$\hat{\mathbf{t}} = \frac{\mathbf{v}}{||\mathbf{v}||}, \quad \hat{\mathbf{n}} = \frac{\hat{\mathbf{t}}'}{||\hat{\mathbf{t}}'||}, \quad \hat{\mathbf{b}} = \hat{\mathbf{t}} \times \hat{\mathbf{n}}.$$

2. Curvature function

$$\kappa = \frac{||\mathbf{v} \times \mathbf{a}||}{||\mathbf{v}||^3}.$$

3. Torsion function

$$\tau = \frac{\mathbf{v} \times \mathbf{a} \cdot \mathbf{j}}{||\mathbf{v} \times \mathbf{a}||^2}.$$

2 Multivariable Functions

2.1 Basic Definitions

In the last section we studied functions with one input and multiple outputs. In this section, we study functions with multiple input and one output. That is, we study functions of the form

$$f : U \subseteq \mathbb{R}^n \to \mathbb{R}.$$

Example 2.1. (a) The volume of a right circular cylinder of radius r and height h is a function $V : U \subseteq \mathbb{R}^2 \to \mathbb{R}$ defined by $V(r,h) = \pi r^2 h$.

(b) The distance function on $\mathbb{R}^n$ is a function $\rho : \mathbb{R}^n \to \mathbb{R}$ defined by

$$\rho(x_1, x_2, \ldots, x_n) = \sqrt{x_1^2 + x_2^2 + \cdots + x_n^2}.$$

(c) The ideal gas law is a function giving the pressure in terms of the temperature and volume:

$$p : \mathbb{R}^2 \to \mathbb{R}$$
$$p(T,V) = \frac{nRT}{V}.$$

In calculus, one generally defines a function by writing $y = f(x)$, where it is understood that the domain of the function is the largest set for which it is defined. This is common practice for multivariable functions as well, and one writes $z = f(x,y)$ meaning that the value of the function f at $(x,y) \in \mathbb{R}^2$ is z. Similarly, one writes $u = f(x,y,z)$, $w = f(x,y,z,u)$, etc., if f depends on more variables.

Example 2.2. Consider the function $f : D \subseteq \mathbb{R}^2 \to \mathbb{R}$ defined by $f(x,y) = \sqrt{y - x^2}$. For the square root to be defined, we must have $y - x^2 \geq 0$. Taking the domain D to be the largest set where the function is defined, we have then $D = \{(x,y) \in \mathbb{R}^2 : y \geq x^2\}$, which is the shaded region above (and including) the parabola $y = x^2$ shown below. Note that D is closed (since it contains its boundary $y = x^2$) and unbounded. (See Appendix B for definitions of these terms.)

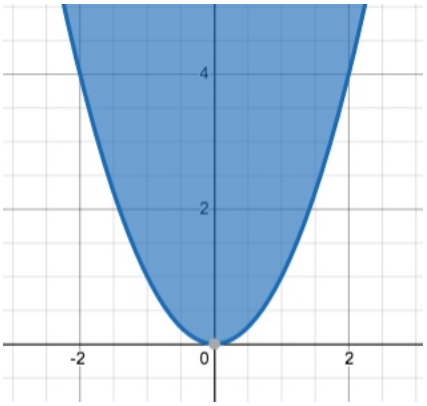


Figure 24: The domain of the function $z = \sqrt{y - x^2}$.

Exercise 2.1. Describe the domain D of the function $f : D \subseteq \mathbb{R}^2 \to \mathbb{R}$ defined by

$$f(x,y) = \frac{1}{\sqrt{16 - x^2 - y^2}}.$$

Solution. For the square root to be defined, we must have $16 \geq x^2 + y^2$. However, we cannot take $16 = x^2 + y^2$ since then we divide by zero. Therefore

$$D = \{(x,y) : x^2 + y^2 < 16\} = B_4(0)$$

is the open ball of radius 4 centered at $(0,0)$. Note that D is open and bounded. □

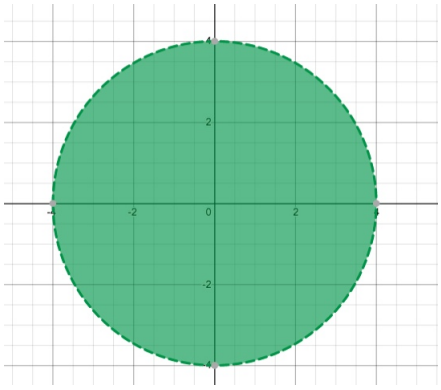


Figure 25: The domain of the function $z = \frac{1}{\sqrt{16-x^2-y^2}}$

.

Exercise 2.2. For each of the following functions, evaluate $f(3,2)$ and find and sketch the domain.

1. $f(x,y) = \frac{\sqrt{x+y+1}}{x-1}$
2. $f(x,y) = x\ln(y^2 - x)$

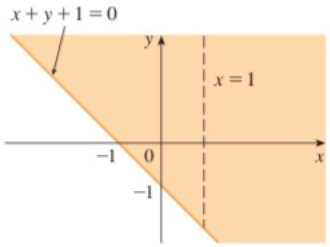
SOLUTION

(a) $$f(3,2) = \frac{\sqrt{3+2+1}}{3-1} = \frac{\sqrt{6}}{2}$$

The expression for f makes sense if the denominator is not 0 and the quantity under the square root sign is nonnegative. So the domain of f is

$$D = \{(x,y) \mid x + y + 1 \geqslant 0,\ x \neq 1\}$$

The inequality $x + y + 1 \geqslant 0$, or $y \geqslant -x - 1$, describes the points that lie on or above



the line $y = -x - 1$, while $x \neq 1$ means that the points on the line $x = 1$ must be excluded from the domain. (See Figure 2.)

(b) $$f(3, 2) = 3 \ln(2^2 - 3) = 3 \ln 1 = 0$$

Since $\ln(y^2 - x)$ is defined only when $y^2 - x > 0$, that is, $x < y^2$, the domain of f is $D = \{(x, y) \mid x < y^2\}$. This is the set of points to the left of the parabola $x = y^2$. (See Figure 3.)

2.2 Visualizing Multivariable Functions

2.2.1 Shapes and Functions

Consider the following three expressions:

$$(\cos(t), \sin(t))$$
$$x^2 + y^2 = 1$$
$$\sqrt{1 - x^2}$$

Each evokes a shape: a circle. (The last may evoke only an arc of a circle.) Let us describe each of these evocations in the language of sets and functions: for each, we define sets X, Y, a function $f : X \to Y$, and tell how the shape appears as a subset of either Y, X, or $X \times Y$.

- *Image*. As in the previous section, take $X = [0, 2\pi)$, $Y = \mathbb{R}^2$, and define

$$f : [0, 2\pi) \to \mathbb{R}^2$$
$$t \mapsto (\cos(t), \sin(t)).$$

  Then the circle is the *image* $f(\mathbb{R}) \subseteq Y$ of the function f, a subset of the codomain. The image construction parametrizes a shape.

- *Preimage*. Now take $X = \mathbb{R}^2$, $Y = \mathbb{R}$ and define

$$f : \mathbb{R}^2 \to \mathbb{R},$$
$$(x, y) \mapsto x^2 + y^2.$$

  Then the circle is the *preimage* $f^{-1}(1) \subseteq X$, a subset of the domain.

- *Graph*. Take $X = (-1, 1)$ and $Y = \mathbb{R}$. Define

$$f : [-1, 1] \to \mathbb{R}$$
$$x \mapsto \sqrt{1 - x^2}.$$

  Then a half-circle is the graph $\Gamma(f) \subseteq X \times Y$, the subset of the Cartesian product defined by

$$\Gamma(f) = \{(x, f(x)) : x \in X\}.$$

2.2.2 Visualizing Multivariable Functions

Definition 2.3. Let $f : U \subset \mathbb{R}^2 \to \mathbb{R}$ be a real-valued function of two variables.

(1) The *graph* of f is the set

$$\Gamma(f) = \{(x, y, z) \subset \mathbb{R}^3 : z = f(x, y)\}$$

.

(2) A *level curve* of f is the set

$$L(f) = \{(x, y) \in U \subset \mathbb{R}^2 : f(x, y) = \text{constant}\}$$

(3) To each level curve in the domain we associate a *contour line*, which is the intersection of the plane $z = c$ with the surface $z = f(x, y)$.

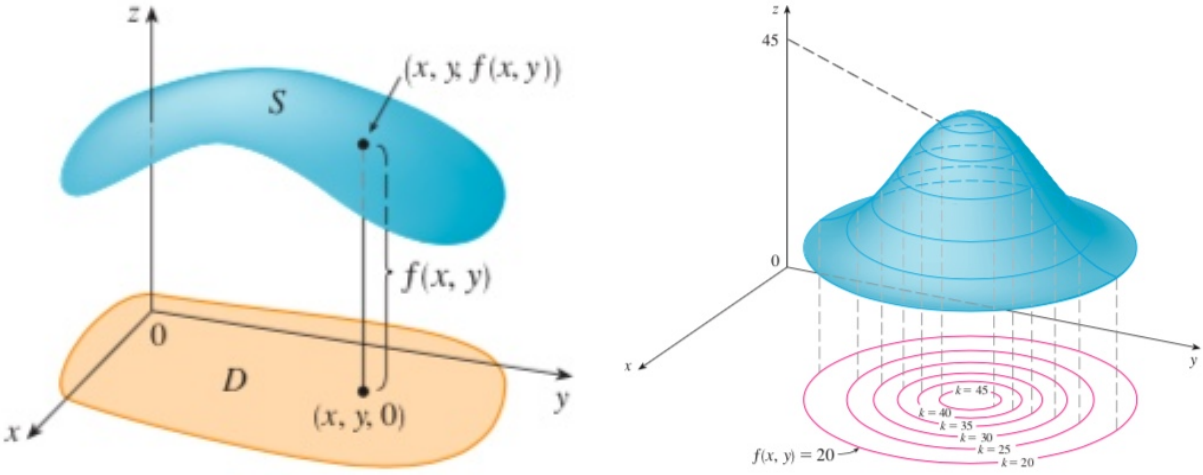


Figure 26: The graph of a function $z = f(x, y)$ is shown on the left. On the right, contour lines are shown in the domain which are the projections onto the xy-plane of the contour lines on the graph.

Example 2.4. The graph of the function $z = 100 - x^2 - y^2$ is shown below, along with the contour line at $z = 75$, and the corresponding level curve.

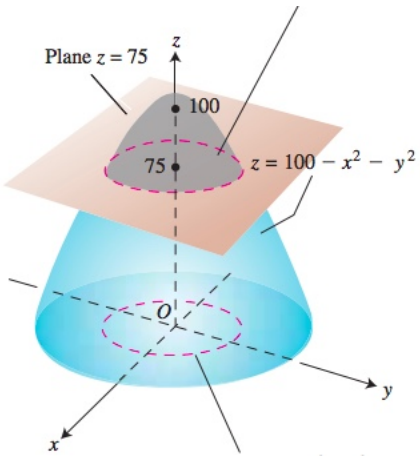
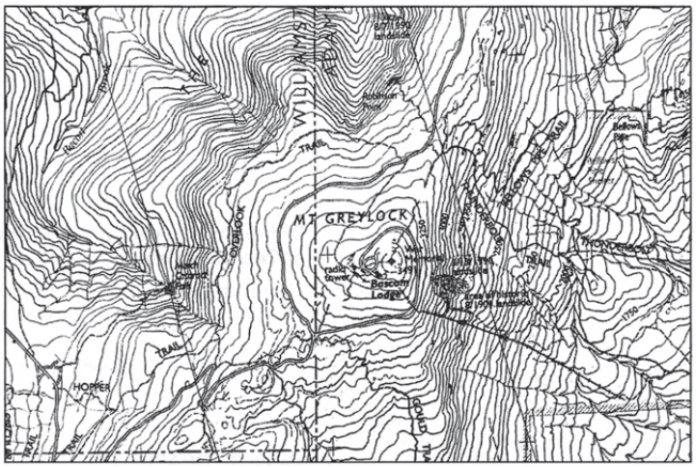


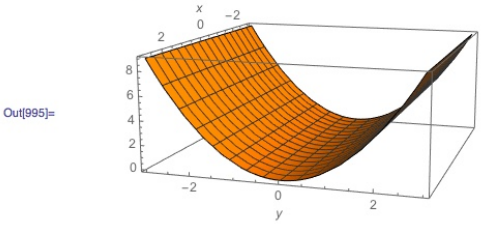
Figure 27: Graph of $z = 100 - x^2 - y^2$. The contour line at $z = 75$ and corresponding level curve are shown.

Example 2.5 (Elevation maps). On maps, contour lines (really level curves; many people use these terms interchangibly) represent constant elevation.



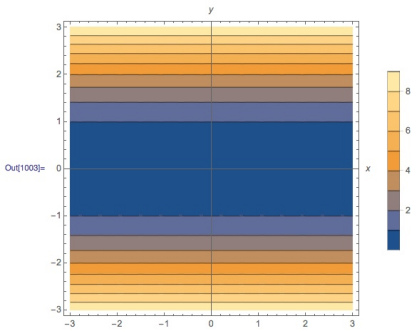
Example 2.6. Consider the function $f : \mathbb{R}^2 \to \mathbb{R}$ defined by $f(x,y) = y^2$. For any fixed value x_0 of x, a cross section of the graph is the parabola $z = y^2$ in the zy plane at x_0. Since $f(x,y)$ is independent of x, the graph has a translational symmetry in the x direction. The graph and level curves can be plotted in *Mathematica* using the commands shown below.

```
In[995]:= Plot3D[y^2, {x, -3, 3}, {y, -3, 3}, AxesLabel → Automatic]
```



(a) The graph of the function $z = y^2$.

```
In[1003]:= ContourPlot[y^2, {x, -3, 3}, {y, -3, 3}, AxesLabel → Automatic, PlotTheme → "Detailed"]
```

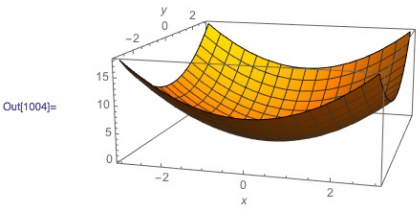


(b) Level curves of the function $z = y^2$.

Figure 28: The graph and level curves of the function $z = y^2$.

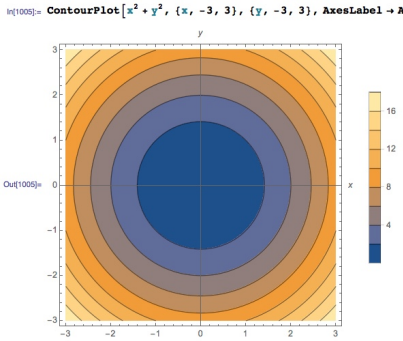
Exercise 2.3. Sketch the graph and level curves of the function $f : \mathbb{R}^2 \to \mathbb{R}$ defined by $f(x,y) = x^2 + y^2$.

```
In[1004]:= Plot3D[x^2 + y^2, {x, -3, 3}, {y, -3, 3}, AxesLabel → Automatic]
```



(a) The graph of the function $z = x^2 + y^2$.

```
In[1005]:= ContourPlot[x^2 + y^2, {x, -3, 3}, {y, -3, 3}, AxesLabel → Automatic, PlotTheme → "Detailed"]
```



(b) Level curves of the function $z = x^2 + y^2$.

Figure 29: The graph and level curves of the function $z = x^2 + y^2$.

Solution. □

If $f : U \subset \mathbb{R}^3 \to \mathbb{R}$ is a function of three variables, its graph $\{(x,y,z,f(x,y,z)\}$ is a subset of $\mathbb{R}^4$, and therefore is impossible to visualize. Instead, we visualize the function by its three-dimensional *level surfaces*.

Definition 2.7 (Level surfaces). A *level surface* of f is the set

$$L(f) = \{(x,y,z) \in U \subset \mathbb{R}^3 : f(x,y,z) = \text{constant}\}$$

Example 2.8. The level surfaces of the function $f : \mathbb{R}^3 \to \mathbb{R}$ defined by $f(x, y, z) = x^2 + y^2 + z^2$ are concentric spheres. These can be plotted in *Mathematica* using the command shown below:

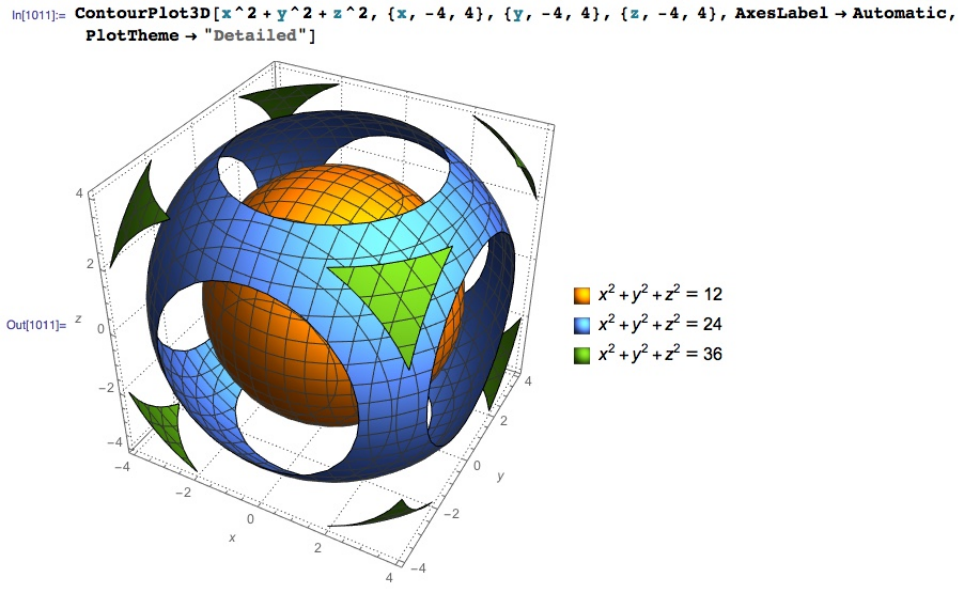


Figure 30: Level surfaces of $f(x, y, z) = x^2 + y^2 + z^2$.

Exercise 2.4. Plot the level surfaces of the function $f : \mathbb{R}^3 \to \mathbb{R}$ defined by $f(x, y, z) = z$.

Solution.

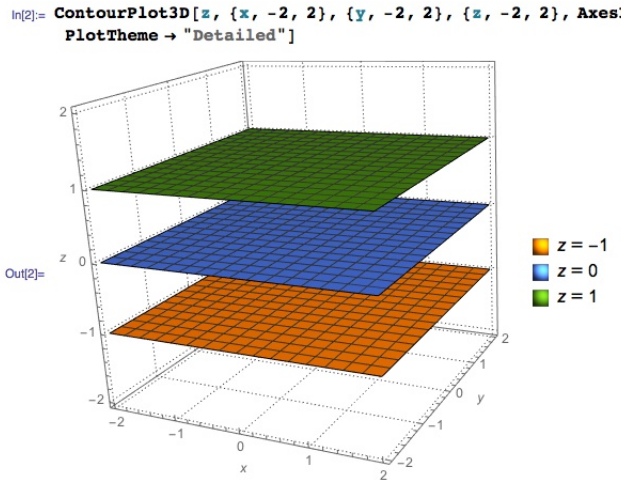


Figure 31: Level surfaces of $f(x, y, z) = z$.

□

Exercise 2.5. Plot the level surfaces of the function $f : \mathbb{R}^3 \to \mathbb{R}$ defined by $f(x, y, z) = x + z$.

Solution.

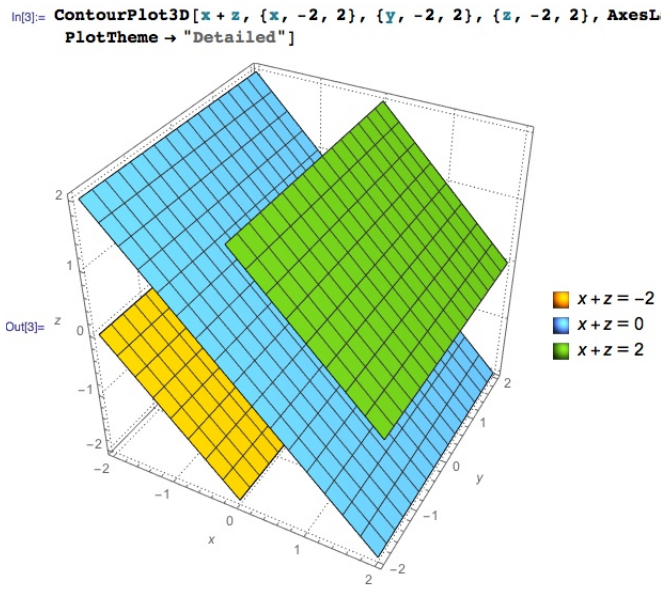


Figure 32: Level surfaces of $f(x, y, z) = x + z$.

□

2.3 Quadric Surfaces

Definition 2.9 (Quadric Surface). Let $A, B, C, D, E, F, G, H, I, J, K$ be fixed real numbers and define a function

$$\begin{aligned} f : \mathbb{R}^3 &\to \mathbb{R} \\ (x, y, z) &\mapsto Ax^2 + By^2 + Cz^2 + Dxy + Eyz + Fxz + Gx + Hy + Jz + K. \end{aligned}$$

A *quadric surface* is the preimage $f^{-1}(0)$; that is, it is the set of all points $(x, y, z) \in \mathbb{R}^3$ such that

$$Ax^2 + By^2 + Cz^2 + Dxy + Eyz + Fxz + Gx + Hy + Jz + K = 0.$$

Here, we will consider some famous quadric surfaces and plot them in *Mathematica*. Many examples of functions of two variables in this unit will have these quadric surfaces as graphs.

2.3.1 Ellipsoids

Definition 2.10 (Ellipsoid). An *ellipsoid* is the quadric surface whose equation is of the form

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} + \frac{z^2}{c^2} = 1.$$

Note that the intersection of this surface with each coordinate plane is an ellipse. For example,

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1 \qquad \text{when } z = 0.$$

If any two of the parameters a, b, c are equal, then the surface is an *ellipsoid of revolution*. If all three are equal, the surface is a sphere.

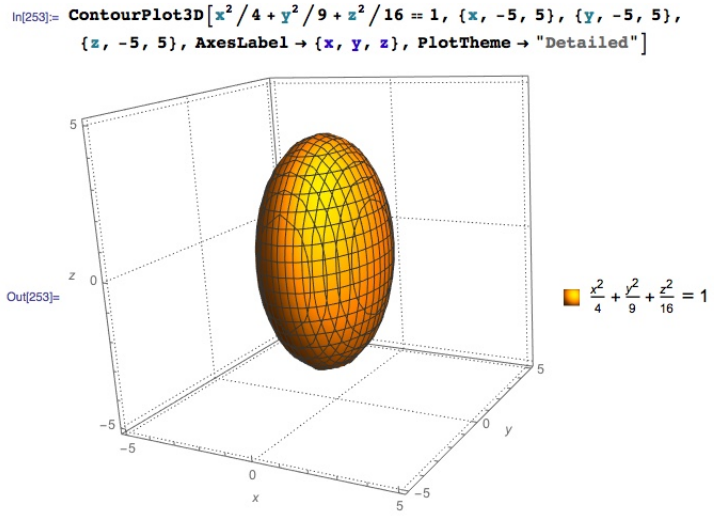


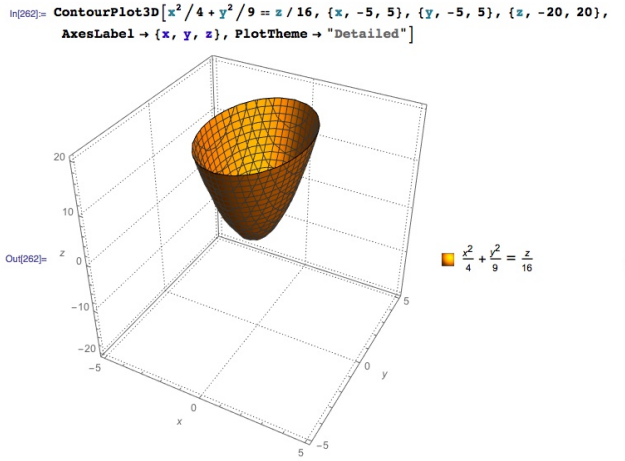
Figure 33: The ellipsoid $\frac{x^2}{4} + \frac{y^2}{9} + \frac{z^2}{16} = 1$.

2.3.2 Paraboloids

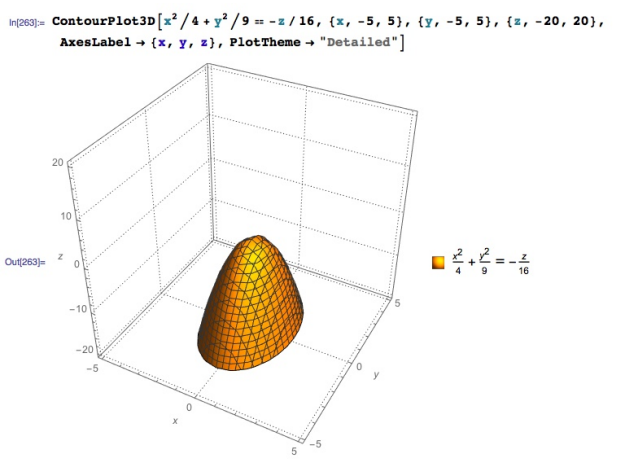
Definition 2.11 (Elliptic Paraboloid). An *elliptic paraboloid* is the quadric described by the equation

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = \frac{z}{c}.$$

Except for the point $(0,0,0)$, the surface lies entirely above or entirely below the xy-plane, depending on the sign of c.



(a) The elliptic paraboloid $\frac{x^2}{4} + \frac{y^2}{9} = \frac{z}{16}$.



(b) The elliptic paraboloid $\frac{x^2}{4} + \frac{y^2}{9} = -\frac{z}{16}$.

The intersections of the surface with the coordinate planes are

$$\begin{aligned} x = 0 &: \text{ the parabola } z = \frac{c}{b^2}y^2 \\ y = 0 &: \text{ the parabola } z = \frac{c}{a^2}x^2 \\ z = 0 &: \text{ the point } (0,0,0). \end{aligned}$$

Each plane $z = z_0$ above the xy-plane intersects the surface in the ellipse

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = \frac{z_0}{c}.$$

If $a = b$, the surface is called a *circular paraboloid* or a *paraboloid of revolution*.

2.3.3 Cones

Definition 2.12 (Cone). An *elliptic cone* is the quadric surface described by the equation

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = \frac{z^2}{c^2}.$$

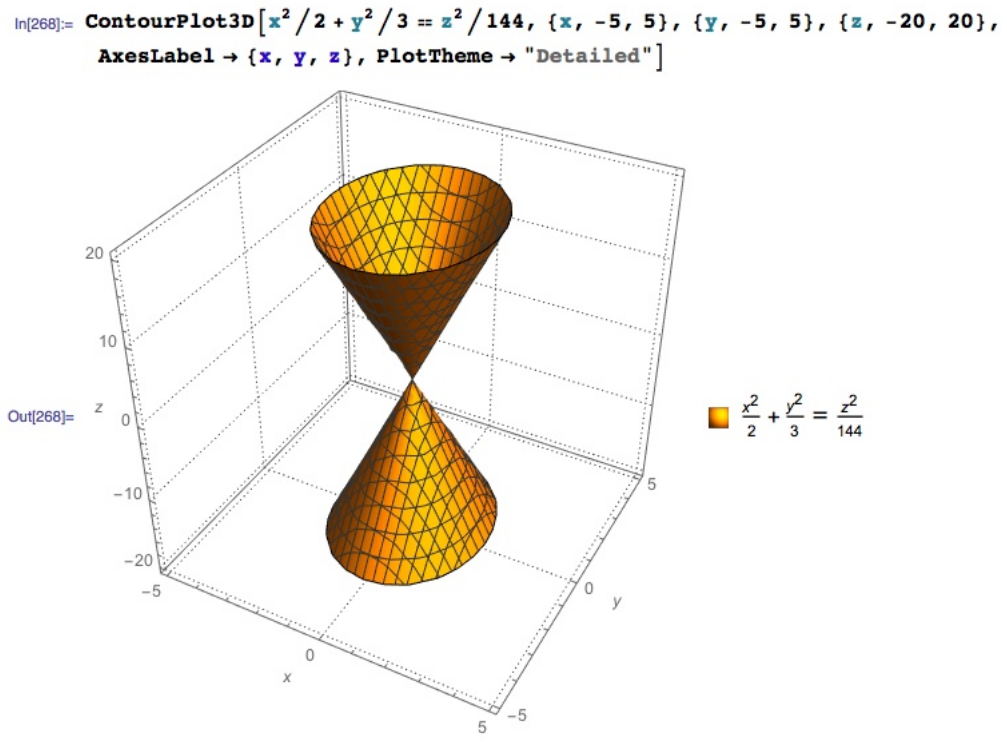


Figure 35: A cone.

The intersections of the surface with the coordinate planes are

$$x = 0: \text{ the lines } z = \pm\frac{c}{b}y$$
$$y = 0: \text{ the parabola } z = \pm\frac{c}{a}x$$
$$z = 0: \text{ the point } (0,0,0).$$

The intersections with the planes $z = z_0$ above and below the xy-plane are ellipses whose centers lie on the z-axis and whose vertices lie on the above lines.

If $a = b$, the cone is a *right circular cone*.

2.3.4 Hyperboloids

Definition 2.13 (Hyperboloid of one sheet). A *hyperboloid of one sheet* is the quadric surface described by the equation

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} - \frac{z^2}{c^2} = 1.$$

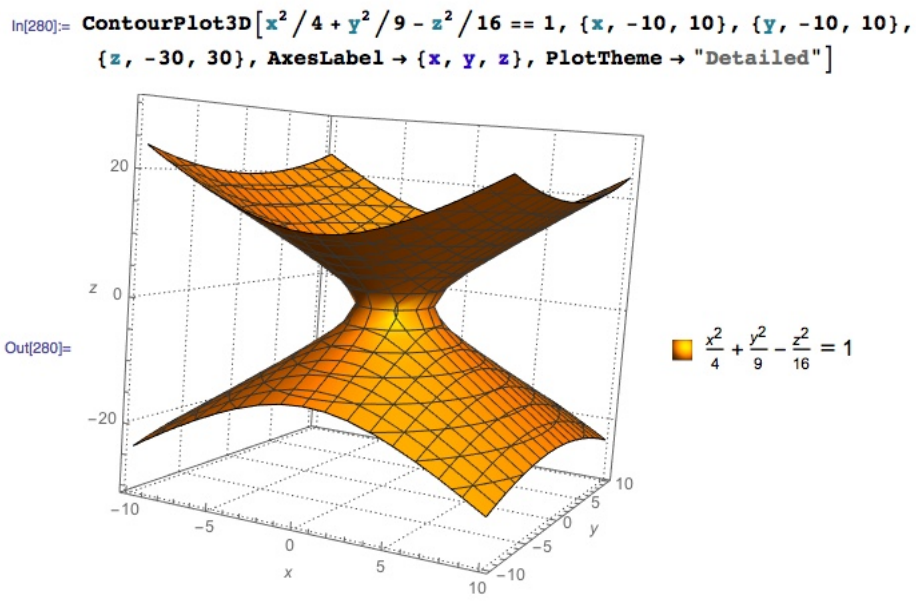


Figure 36: A hyperboloid of one sheet.

The intersections with the coordinate planes are

$$x = 0: \text{ the hyperbola } \frac{y^2}{b^2} - \frac{z^2}{c^2} = 1$$
$$y = 0: \text{ the hyperbola } \frac{x^2}{a^2} - \frac{z^2}{c^2} = 1$$
$$z = 0: \text{ the ellipse } \frac{x^2}{a^2} + \frac{y^2}{b^2} = 1.$$

The plane $z = z_0$ intersects the surface in an ellipse with center on the z-axis and vertices on one of the above hyperbolas.

The surface is connected, meaning that it is possible to travel from one point on it to any other without leaving the surface. For this reason, it is said to have *one sheet*, in contrast to the hyperboloid in the next example, which has two sheets.

If $a = b$, the hyperboloid is a surface of revolution.

Definition 2.14 (Hyperboloid of two sheets). A *hyperboloid of two sheets* is the quadric surface described by the equation

$$-\frac{x^2}{a^2} - \frac{y^2}{b^2} + \frac{z^2}{c^2} = 1.$$

The plane $z = 0$ does not intersect the surface; in fact, for a horizontal plane to intersect the

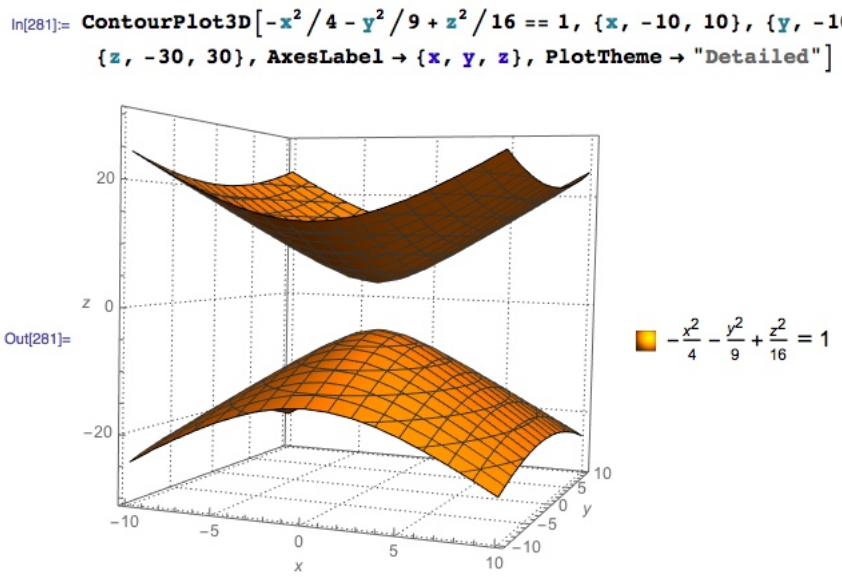


Figure 37: A hyperboloid of two sheets.

surface, we must have $|z| \geq |c|$. The hyperbolic cross-sections

$$x = 0 : \frac{z^2}{c^2} - \frac{y^2}{b^2} = 1$$
$$y = 0 : \frac{z^2}{c^2} - \frac{x^2}{a^2} = 1$$

have their vertices and foci on the z-axis. The surface has two disconnected components, one above the plane $z = c$ and the other below the plane $z = -c$.

2.3.5 Hyperbolic Paraboloids

Definition 2.15 (Hyperbolic paraboloid). A *hyperbolic paraboloid* is the quadric surface described by the equation

$$\frac{y^2}{b^2} - \frac{x^2}{a^2} = \frac{z}{c}, \qquad c > 0.$$

The intersections with the coordinate planes are

$$x = 0 : \text{ the parabola } z = \frac{c}{b^2} y^2$$
$$y = 0 : \text{ the parabola } z = -\frac{c}{a^2} x^2.$$

In the plane $x = 0$, the parabola opens upward from the origin. The parabola in the plane $y = 0$ opens downward.

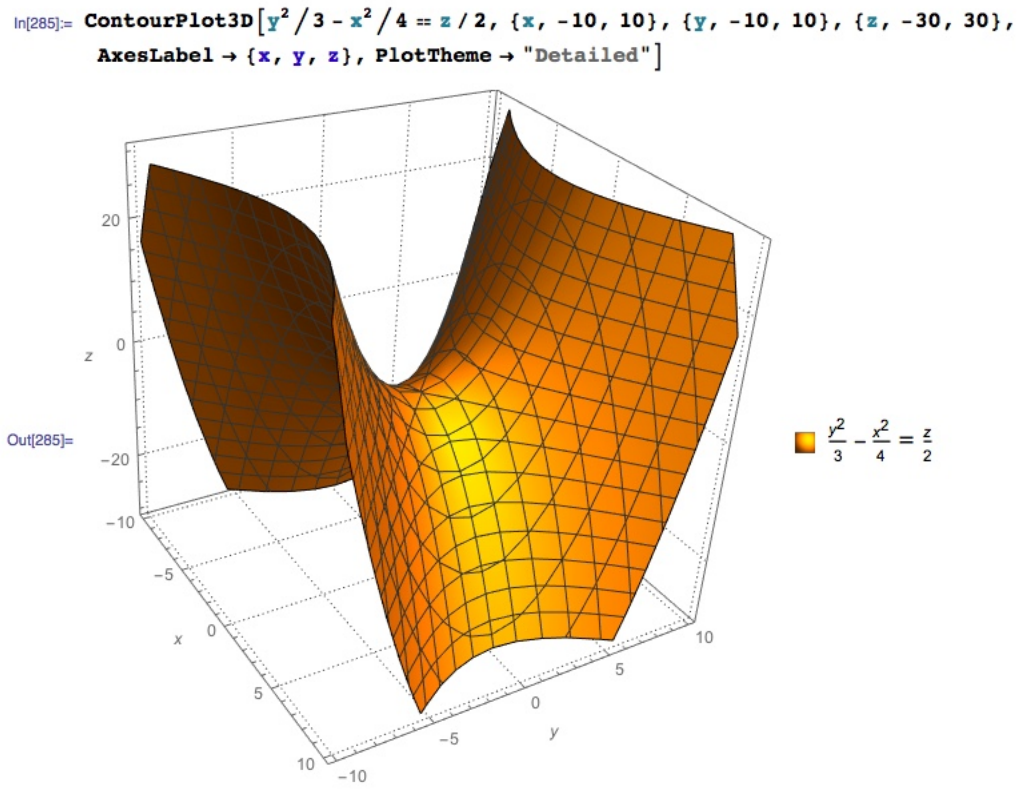


Figure 38: A hyperbolic paraboloid.

The intersection of the surface with a plane $z = z_0 > 0$ is the hyperbola

$$\frac{y^2}{b^2} - \frac{x^2}{a^2} = \frac{z_0}{c}.$$

Near the origin, the surface is shaped like a saddle. To a person traveling along the surface in the yz-plane, the origin looks like a minimum. To a person traveling in the xz-plane, the origin looks like a maximum. Such a point is called a *saddle point* of a surface.

Exercise 2.6. Work out the various cross-sections of the quadric surfaces described above and verify the claims in this section.

2.4 Limits

Let's compare the behavior of the functions

$$f(x,y) = \frac{\sin(x^2+y^2)}{x^2+y^2} \quad \text{and} \quad g(x,y) = \frac{x^2-y^2}{x^2+y^2}$$

as $(x,y) \to (0,0)$. Note that neither function is defined at $(0,0)$:

It looks like $\lim_{(x,y)\to(0,0)} f(x,y) = 1$ (since the values of the function approach 1 from all directions) while $\lim_{(x,y)\to(0,0)} g(x,y)$ does not exist (since the values of the function near $(0,0)$ do not approach any fixed value).

TABLE 1 Values of $f(x, y)$

x \ y	−1.0	−0.5	−0.2	0	0.2	0.5	1.0
−1.0	0.455	0.759	0.829	0.841	0.829	0.759	0.455
−0.5	0.759	0.959	0.986	0.990	0.986	0.959	0.759
−0.2	0.829	0.986	0.999	1.000	0.999	0.986	0.829
0	0.841	0.990	1.000		1.000	0.990	0.841
0.2	0.829	0.986	0.999	1.000	0.999	0.986	0.829
0.5	0.759	0.959	0.986	0.990	0.986	0.959	0.759
1.0	0.455	0.759	0.829	0.841	0.829	0.759	0.455

TABLE 2 Values of $g(x, y)$

x \ y	−1.0	−0.5	−0.2	0	0.2	0.5	1.0
−1.0	0.000	0.600	0.923	1.000	0.923	0.600	0.000
−0.5	−0.600	0.000	0.724	1.000	0.724	0.000	−0.600
−0.2	−0.923	−0.724	0.000	1.000	0.000	−0.724	−0.923
0	−1.000	−1.000	−1.000		−1.000	−1.000	−1.000
0.2	−0.923	−0.724	0.000	1.000	0.000	−0.724	−0.923
0.5	−0.600	0.000	0.724	1.000	0.724	0.000	−0.600
1.0	0.000	0.600	0.923	1.000	0.923	0.600	0.000

Definition 2.16 (Limit). Let $f : U - \{\mathbf{x}_0\} \subset \mathbb{R}^n \to \mathbb{R}$, and $L \in \mathbb{R}$. We say that $f(\mathbf{x}) \to L$ as $\mathbf{x} \to \mathbf{x}_0$ if for every $\epsilon > 0$ there exists a corresponding $\delta > 0$ such that

$$0 < ||\mathbf{x} - \mathbf{x}_0|| < \delta \implies |f(\mathbf{x}) - L| < \epsilon.$$

In this case L is called the *limit* of f as $\mathbf{x}$ approaches $\mathbf{x}_0$, and we write $\lim_{\mathbf{x}\to\mathbf{x}_0} f(\mathbf{x}) = L$.

This definition says that the distance between $f(\mathbf{x})$ and L can be made arbitrarily small by making the distance between $\mathbf{x}$ and $\mathbf{x}_0$ sufficiently small. [13]

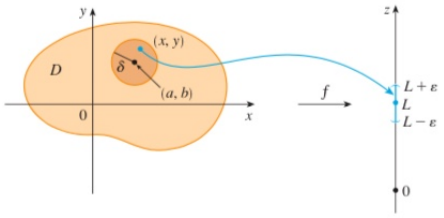


Figure 39: A function $f(\mathbf{x})$ has limit L as $\mathbf{x} \to \mathbf{x}_0$ if $f(\mathbf{x})$ can be made arbitrarily close to L by taking $\mathbf{x}$ sufficiently close to $\mathbf{x}_0$.

[13] Two points x_1, x_2 are said to be *arbitrarily close* if, for *every* $\epsilon > 0$, $d(x_1, x_2) < \epsilon$. A point x_1 is said to be *sufficiently close* to x_2 if there exists *some* (fixed) $\delta > 0$ such that $d(x_1, x_2) < \delta$.

As in single-variable calculus, another illustration of the definition of a limit can be given in terms of its graph: For any $\epsilon > 0$, we can find $\delta > 0$ such that if $\mathbf{x} \in B_\delta(\mathbf{x}_0)$ and $\mathbf{x} \neq \mathbf{x}_0$, then $f(x)$ lies between the horizontal planes $z = L - \epsilon$ and $z = L + \epsilon$.

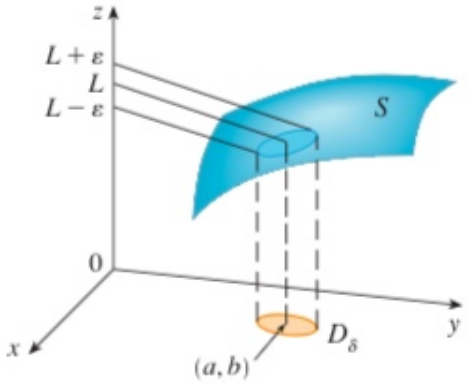


Figure 40: A function f has limit L as $\mathbf{x} \to \mathbf{x}_0$ if the graph can be bounded between two arbitrarily close planes whenever for all $\mathbf{x}$ sufficiently close to $\mathbf{x}_0$.

For functions of a single variable, when we let $x \to a$, there are only two directions of approach: from the left or from the right. Recall that if $\lim_{x\to a^-} f(x) \neq \lim_{x\to a^+} f(x)$, then $\lim_{x\to a} f(x)$ does not exist. For functions of two variables, we can let (x, y) approach (a, b) from an infinite number of directions, as long as (x, y) stays within the domain of f.

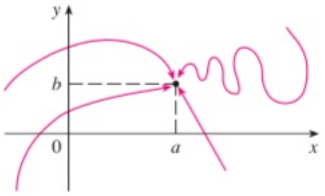


Figure 41: Various paths in the domain of a function $f : \mathbb{R}^2 \to \mathbb{R}$ along which we can approach a point (a, b).

Note that the definition of a limit only depends on the *distance* from (x, y) and (a, b). It does *not* refer to the direction of the approach. Therefore, if the limit exists, then $f(x, y)$ must approach the same limit no matter how (x, y) approaches (a, b). Thus, if we can find two different paths of approach along which the function $f(x, y)$ has different limits, then it follows that $\lim_{(x,y)\to(a,b)} f(x, y)$ does not exist.

Example 2.17. Consider $\lim_{(x,y)\to(0,0)} \frac{x^2-y^2}{x^2+y^2}$. Consider approaching $(0, 0)$ along the x-axis, that is, along the curve $\mathbf{r}(t) = (t, 0)$. Then

$$\lim_{(x(t),y(t))\to(0,0)} \frac{x(t)^2 - y(t)^2}{x(t)^2 + y(t)^2} = \lim_{t\to 0} \frac{t^2 - 0}{t^2 + 0} = \lim_{t\to 0} \frac{t^2}{t^2} = \lim_{t\to 0} 1 = 1.$$

Approaching $(0,0)$ along the y-axis, i.e., along the path $\mathbf{r}(t) = (0,t)$, we find

$$\lim_{(x(t),y(t))\to(0,0)} \frac{x(t)^2 - y(t)^2}{x(t)^2 + y(t)^2} = \lim_{t\to 0} \frac{0 - t^2}{0 + t^2} = \lim_{t\to 0} \frac{-t^2}{t^2} = \lim_{t\to 0}(-1) = -1.$$

We have found two paths along which $\lim_{(x,y)\to(0,0)} \frac{x^2-y^2}{x^2+y^2}$ has different values, so $\lim_{(x,y)\to(0,0)} \frac{x^2-y^2}{x^2+y^2}$ does not exist.

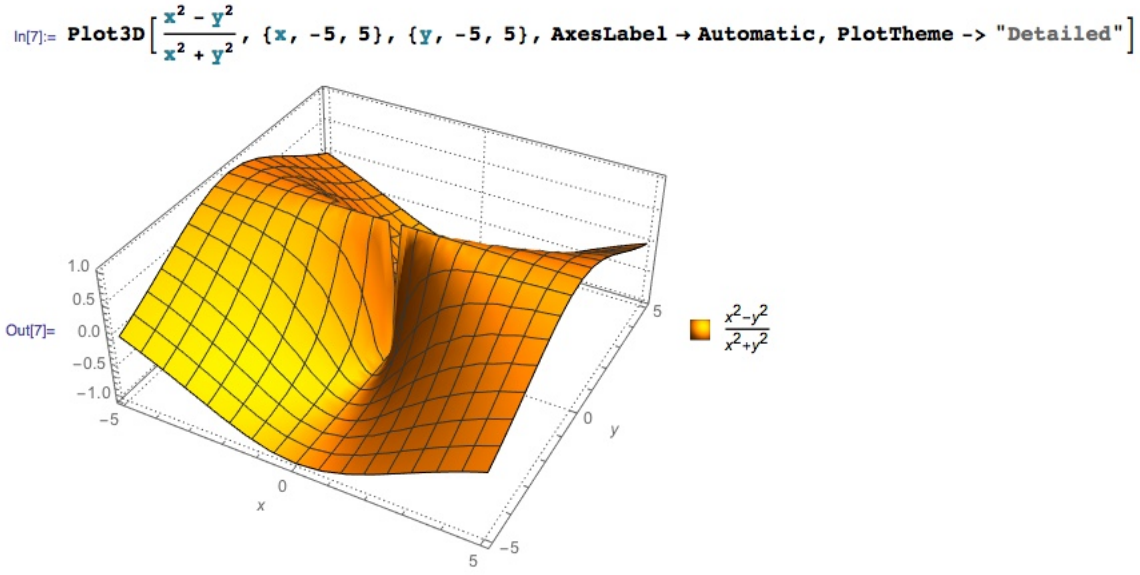


Figure 42: The function $f(x,y) = \frac{x^2-y^2}{x^2+y^2}$ has no limit as $(x,y) \to (0,0)$.

Example 2.18. Consider $\lim_{(x,y)\to(0,0)} \frac{xy}{x^2+y^2}$. Approaching $(0,0)$ along the x-axis gives

$$\lim_{x\to 0} \frac{x \cdot 0}{x^2 + 0} = \lim_{x\to 0} \frac{0}{x^2} = 0.$$

Approaching $(0,0)$ along the y-axis gives

$$\lim_{y\to 0} \frac{0 \cdot y}{0 + y^2} = \lim_{y\to 0} \frac{0}{y^2} = 0.$$

While these limits agree, this is *not* sufficient to conclude that $\lim_{(x,y)\to(0,0)} \frac{xy}{x^2+y^2} = 0$, since we must show that we obtain the same limit along *every* path approaching $(0,0)$. Indeed, if we approach $(0,0)$ along the line $y = x$, we obtain

$$\lim_{x\to 0} \frac{x^2}{x^2 + x^2} = \lim_{x\to 0} \frac{x^2}{2x^2} = \lim_{x\to 0} \frac{1}{2} = \frac{1}{2}.$$

Thus, $\lim_{(x,y)\to(0,0)} \frac{xy}{x^2+y^2}$ does not exist.

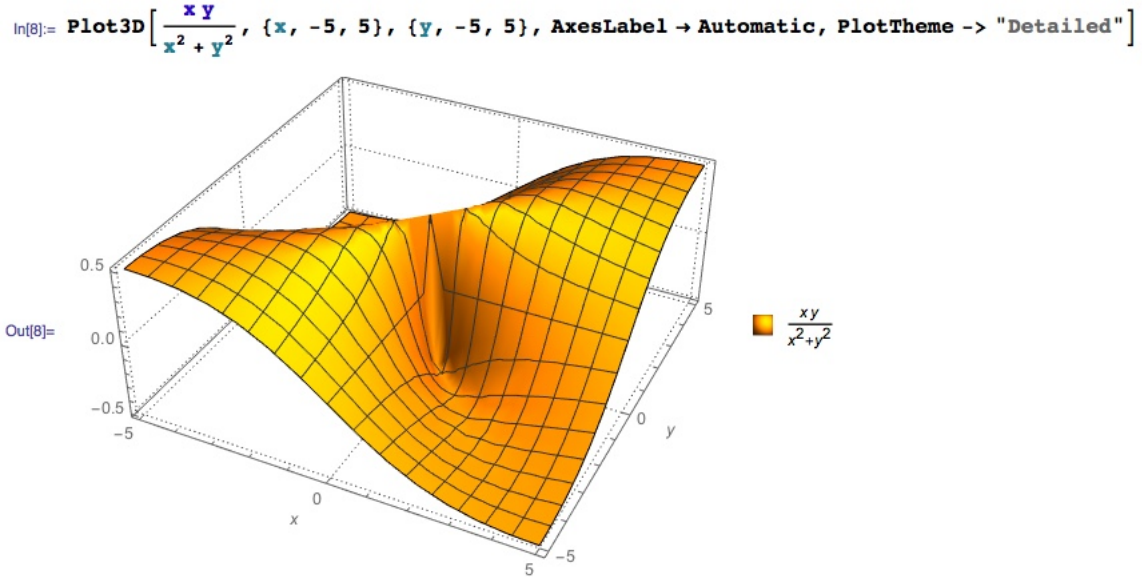


Figure 43: The function $f(x,y) = \frac{xy}{x^2+y^2}$ has no limit as $(x,y) \to (0,0)$.

The previous example showed that it is not sufficient to compute the limit of a multivariable function by restricting the function to paths in the domain. We now show how to use Definition 2.16 to show that a limit *does* exist.

Theorem 2.19. Let a, b, c be fixed real numbers. Then

(a) $\lim_{(x,y)\to(a,b)} c = c$.

(b) $\lim_{(x,y)\to(a,b)} x = a$.

(c) $\lim_{(x,y)\to(a,b)} y = b$.

Proof.

(a) Let $\epsilon > 0$. Since $|c - c| = 0$, by taking δ to be any positive real number, then $0 < \sqrt{(x-a)^2 + (y-b)^2} < \delta$ implies $|f(x,y) - f(a,b)| = |c - c| = 0 < \epsilon$. This proves that $\lim_{(x,y)\to(a,b)} c = c$.

(b) Let $\epsilon > 0$. We have $|f(x,y) - f(a,b)| = |x - a|$. Since

$$|x-a| = \sqrt{(x-a)^2} \leq \sqrt{(x-a)^2 + (y-b)^2},$$

by taking $\delta = \epsilon$, $0 < \sqrt{(x-a)^2 + (y-b)^2} < \delta$ implies that $|f(x,y) - f(a,b)| = |x - a| < \epsilon$. This proves that $\lim_{(x,y)\to(a,b)} x = a$.

(c) The proof is essentially the same as that of part (b) and is left as an exercise.

□

Exercise 2.7. Prove part (c) of Proposition **??**.

Solution. Let $\epsilon > 0$. We have $|f(x,y) - f(a,b)| = |y - b|$. Since

$$|y-b| = \sqrt{(y-b)^2} \leq \sqrt{(x-a)^2 + (y-b)^2},$$

by taking $\delta = \epsilon$, $0 < \sqrt{(x-a)^2 + (y-b)^2} < \delta$ implies that $|f(x,y) - f(a,b)| = |y-b| < \epsilon$. This proves that $\lim_{(x,y)\to(a,b)} y = b$. □

The next theorem shows that all of the limit properties of single-variable functions carry over to multivariable functions.

Theorem 2.20 (Limit laws for multivariable functions). Suppose $f(x,y)$ and $g(x,y)$ are defined on the same open set containing (x_0, y_0), and that

$$\lim_{(x,y)\to(x_0,y_0)} f(x,y) = L \quad \text{and} \quad \lim_{(x,y)\to(x_0,y_0)} g(x,y) = M.$$

Then

(i) $\lim_{(x,y)\to(x_0,y_0)} cf(x,y) = cL$ for any $c \in \mathbb{R}$;

(ii) $\lim_{(x,y)\to(x_0,y_0)} (f(x,y) + g(x,y)) = L + M$;

(iii) $\lim_{(x,y)\to(x_0,y_0)} (f(x,y)g(x,y)) = LM$;

(iv) $\lim_{(x,y)\to(x_0,y_0)} = \frac{f(x,y)}{g(x,y)} = \frac{L}{M}$ whenever $M \neq 0$.

Proof. The proofs of these are identical to the proofs of the corresponding properties in Theorem 1.7, with the metric $d_{\mathbb{R}}(x,y) = |x-y|$ on $\mathbb{R}$ replaced by the metric $d_{\mathbb{R}^2}(\mathbf{x},\mathbf{y}) = ||\mathbf{x}-\mathbf{y}||$ on $\mathbb{R}^2$. [14] □

Example 2.21. To evaluate $\lim_{(x,y)\to(1,2)} (x^2y^3 - x^3y^2 + 3x + 2y)$ using Theorems 2.20 and 2.19, we have

$$\begin{aligned}
&\lim_{(x,y)\to(1,2)} (x^2y^3 - x^3y^2 + 3x + 2y) = \lim_{(x,y)\to(1,2)} (x^2y^3) + \lim_{(x,y)\to(1,2)} (-x^3y^2) + \lim_{(x,y)\to(1,2)} (3x+2y) \\
&= \lim_{(x,y)\to(1,2)} x^2 \cdot \lim_{(x,y)\to(1,2)} y^3 - \lim_{(x,y)\to(1,2)} x^3 \cdot \lim_{(x,y)\to(1,2)} y^2 + 3 \lim_{(x,y)\to(1,2)} x + 2 \lim_{(x,y)\to(1,2)} y \\
&= (\lim_{(x,y)\to(1,2)} x)^2 \cdot (\lim_{(x,y)\to(1,2)} y)^3 - (\lim_{(x,y)\to(1,2)} x)^3 \cdot (\lim_{(x,y)\to(1,2)} y)^2 + 3 \lim_{(x,y)\to(1,2)} x + 2 \lim_{(x,y)\to(1,2)} y \\
&= (1)^2(2)^3 - (1)^3(2)^2 + 3(1) + 2(2) \\
&= 8 - 4 + 3 + 4 \\
&= 11.
\end{aligned}$$

Definition 2.22 (Continuous function). A function f of two variables is *continuous at* (a,b) if

$$\lim_{(x,y)\to(a,b)} f(x,y) = f(a,b).$$

The function f is *continuous on* D if f is continuous at every point (a,b) in D.

[14] In fact, these properties hold for maps between any metric spaces. For a function $f : X \to Y$ between metric spaces, the definition of $\lim_{x\to x_0} f(x) = L$ becomes: for every $\epsilon > 0$ there exists $\delta > 0$ such that $d_Y(f(x), L) < \epsilon$ whenever $0 < d_X(x, x_0) < \delta$. Since the proofs of the limit properties only depend on properties shared by every metric (e.g., the triangle inequality), the proofs are exactly the same, with the Euclidean metrics on $\mathbb{R}^2, \mathbb{R}$ replaced by the metrics on X and Y.

As in single-variable calculus, continuous functions map nearby points in D to nearby points in $\mathbb{R}$, so a surface that is the graph of a continuous function has no holes or breaks.

Using the limit properties, we find that sums, products, and quotients of continuous functions are continuous on their domains. Similarly, compositions of continuous functions are continuous.

It follows from Theorem 2.19 that $f(x,y) = x, g(x,y) = y$, and $h(x,y) = c$ are continuous. Since any polynomial, e.g.,

$$f(x,y) = x^4 + 5x^3y^2 + 6xy^4 - 7y + 6$$

can be built out of these by multiplication and addition, all polynomials are continuous on $\mathbb{R}^2$. Likewise, any rational function is continuous on its domain because it is a quotient of continuous functions. E.g.,

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

Example 2.23. The function $f(x,y) = \frac{x^2-y^2}{x^2+y^2}$ is continuous everywhere except at $(0,0)$.

Exercise 2.8. Where is the function

$$g(x,y) = \begin{cases} \frac{x^4-x^2y^2}{x+y}, & \text{if } (x,y) \neq (0,0), \\ 0, & \text{if } (x,y) = (0,0). \end{cases}$$

continuous?

Solution. Since a rational function is continuous everywhere it is defined, $g(x,y)$ is continuous for all $(x,y) \neq (0,0)$. Since for $(x,y) \neq (0,0)$

$$\frac{x^4-x^2y^2}{x+y} = \frac{x^2(x^2-y^2)}{x+y} = \frac{x^2(x+y)(x-y)}{x+y} = x^2(x-y),$$

we have

$$\lim_{(x,y)\to(0,0)} \frac{x^4-x^2y^2}{x+y} = \lim_{(x,y)\to(0,0)} x^2(x-y) = 0.$$

Thus, $g(x,y)$ is continuous everywhere. □

2.5 Partial Derivatives

2.5.1 Motivation

Example 2.24 (Heat index).

On a hot day, extreme humidity makes us think the temperature is higher than it really is, whereas in very dry air we perceive the temperature to be lower than the thermometer indicates. The National Weather Service has devised the *heat index* (also called the temperature-humidity index, or humidex, in some countries) to describe the combined effects of temperature and humidity. The heat index I is the perceived air temperature when the actual temperature is T and the relative humidity is H. So I is a function of T and H and we can write $I = f(T, H)$. The following table of values of I is an excerpt from a table compiled by the National Weather Service.

Relative humidity (%)

Actual temperature (°F) T \ H	50	55	60	65	70	75	80	85	90
90	96	98	100	103	106	109	112	115	119
92	100	103	105	108	112	115	119	123	128
94	104	107	111	114	118	122	127	132	137
96	109	113	116	121	125	130	135	141	146
98	114	118	123	127	133	138	144	150	157
100	119	124	129	135	141	147	154	161	168

If we concentrate on the highlighted column of the table, which corresponds to a relative humidity of $H = 70\%$, we are considering the heat index as a function of the single variable T for a fixed value of H. Let's write $g(T) = f(T, 70)$. Then $g(T)$ describes how the heat index I increases as the actual temperature T increases when the relative humidity is 70%. The derivative of g when $T = 96°\text{F}$ is the rate of change of I with respect to T when $T = 96°\text{F}$:

$$g'(96) = \lim_{h \to 0} \frac{g(96 + h) - g(96)}{h} = \lim_{h \to 0} \frac{f(96 + h, 70) - f(96, 70)}{h}$$

We can approximate $g'(96)$ using the values in Table 1 by taking $h = 2$ and -2:

$$g'(96) \approx \frac{g(98) - g(96)}{2} = \frac{f(98,70) - f(96,70)}{2} = \frac{133 - 125}{2} = 4$$

$$g'(96) \approx \frac{g(94) - g(96)}{-2} = \frac{f(94,70) - f(96,70)}{-2} = \frac{118 - 125}{-2} = 3.5$$

Averaging these values, we can say that the derivative $g'(96)$ is approximately 3.75. This means that, when the actual temperature is 96°F and the relative humidity is 70%, the apparent temperature (heat index) rises by about 3.75°F for every degree that the actual temperature rises!

Now let's look at the highlighted row in Table 1, which corresponds to a fixed temperature of $T = 96$°F. The numbers in this row are values of the function $G(H) = f(96, H)$, which describes how the heat index increases as the relative humidity H increases when the actual temperature is $T = 96$°F. The derivative of this function when $H = 70\%$ is the rate of change of I with respect to H when $H = 70\%$:

$$G'(70) = \lim_{h \to 0} \frac{G(70 + h) - G(70)}{h} = \lim_{h \to 0} \frac{f(96, 70 + h) - f(96, 70)}{h}$$

By taking $h = 5$ and -5, we approximate $G'(70)$ using the tabular values:

$$G'(70) \approx \frac{G(75) - G(70)}{5} = \frac{f(96,75) - f(96,70)}{5} = \frac{130 - 125}{5} = 1$$

$$G'(70) \approx \frac{G(65) - G(70)}{-5} = \frac{f(96,65) - f(96,70)}{-5} = \frac{121 - 125}{-5} = 0.8$$

By averaging these values we get the estimate $G'(70) \approx 0.9$. This says that, when the temperature is 96° F and the relative humidity is 70%, the heat index rises about 0.9° F for every percent the relative humidity rises.

In general, if $f(x, y)$ is a function of two variables x and y, and we consider the values of the function obtained by varying x (say) while keeping y fixed to a constant value $y = y_0$, then we are really considering a function of a *single* variable $g(x) = f(x, y_0)$. Then if $g(x)$ is differentiable at $x = x_0$, we call $g'(x)$ a *partial derivative* of $f(x, y)$. We now define this formally.

2.5.2 Definition of Partial Derivatives

Definition 2.25 (Partial derivative). Let $f : U \subset \mathbb{R}^2 \to \mathbb{R}$ be defined on an open subset of $\mathbb{R}^2$ containing the point (x_0, y_0). [15] Then the real number

$$\frac{\partial f}{\partial x}(x_0, y_0) = \lim_{h \to 0} \frac{f(x+h, y) - f(x, y)}{h}$$

is called the *partial derivative of f with respect to x at the point (x_0, y_0)*, provided this limit exists. Similarly, the real number

$$\frac{\partial f}{\partial y}(x_0, y_0) = \lim_{h \to 0} \frac{f(x, y+h) - f(x, y)}{h}$$

is called the *partial derivative of f with respect to y at the point (x_0, y_0)*, provided this limit exists. If $\frac{\partial f}{\partial x}$ and $\frac{\partial f}{\partial y}$ are defined at all points of U, then they are functions on U.

Remark. Note that $\frac{\partial f}{\partial x}$ is not a quotient, but a *limit* of a quotient. Thus, the symbols "∂f" or "∂x" do not have independent meaning. Similar remarks apply to $\frac{\partial f}{\partial y}$. While this is the most common notation for the partial derivatives (and the one we will use almost exclusively), other notations are also frequently used. □

Notations for Partial Derivatives If $z = f(x, y)$, we write

$$f_x(x, y) = f_x = \frac{\partial f}{\partial x} = \frac{\partial}{\partial x} f(x, y) = \frac{\partial z}{\partial x} = f_1 = D_1 f = D_x f$$

$$f_y(x, y) = f_y = \frac{\partial f}{\partial y} = \frac{\partial}{\partial y} f(x, y) = \frac{\partial z}{\partial y} = f_2 = D_2 f = D_y f$$

Figure 44: Notations for partial derivatives.

To compute the partial derivative of a function $f(x, y)$ with respect to x, one simply treats y as a constant and differentiates $f(x, y)$ with respect to x. To compute the partial derivative with respect to y, treat x as a constant and differentiate $f(x, y)$ with respect to y.

Example 2.26. If $f(x, y) = x^3 + x^2y^3 - 2y^2$, then

$$\frac{\partial f}{\partial x} = 3x^2 + 2xy^3$$
$$\frac{\partial f}{\partial y} = 3x^2y^2 - 4y.$$

[15] Since U is open, there exists a neighborhood of (x_0, y_0) contained in U. Thus, if h is sufficiently close to zero, then all points of the form $(x_0 + h, y_0)$ will be in U, so this definition makes sense.

2.5.3 Interpretation of Partial Derivatives

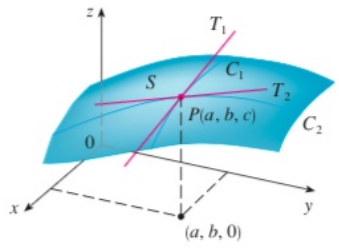


FIGURE 1
The partial derivatives of f at (a, b) are the slopes of the tangents to C_1 and C_2.

To give a geometric interpretation of partial derivatives, we recall that the equation $z = f(x, y)$ represents a surface S (the graph of f). If $f(a, b) = c$, then the point $P(a, b, c)$ lies on S. By fixing $y = b$, we are restricting our attention to the curve C_1 in which the vertical plane $y = b$ intersects S. (In other words, C_1 is the trace of S in the plane $y = b$.) Likewise, the vertical plane $x = a$ intersects S in a curve C_2. Both of the curves C_1 and C_2 pass through the point P. (See Figure 1.)

Notice that the curve C_1 is the graph of the function $g(x) = f(x, b)$, so the slope of its tangent T_1 at P is $g'(a) = f_x(a, b)$. The curve C_2 is the graph of the function $G(y) = f(a, y)$, so the slope of its tangent T_2 at P is $G'(b) = f_y(a, b)$.

Thus the partial derivatives $f_x(a, b)$ and $f_y(a, b)$ can be interpreted geometrically as the slopes of the tangent lines at $P(a, b, c)$ to the traces C_1 and C_2 of S in the planes $y = b$ and $x = a$.

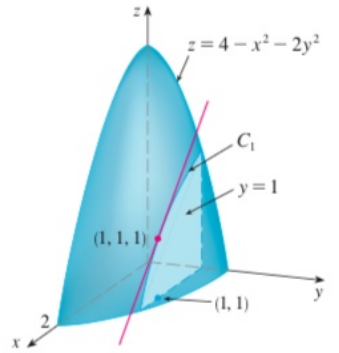


FIGURE 2

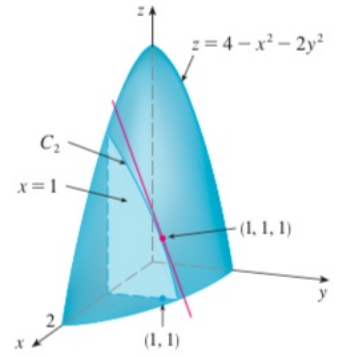
As we have seen in the case of the heat index function, partial derivatives can also be interpreted as *rates of change*. If $z = f(x, y)$, then $\partial z/\partial x$ represents the rate of change of z with respect to x when y is fixed. Similarly, $\partial z/\partial y$ represents the rate of change of z with respect to y when x is fixed.

EXAMPLE 2 If $f(x, y) = 4 - x^2 - 2y^2$, find $f_x(1, 1)$ and $f_y(1, 1)$ and interpret these numbers as slopes.

SOLUTION We have

$$f_x(x, y) = -2x \qquad f_y(x, y) = -4y$$

$$f_x(1, 1) = -2 \qquad f_y(1, 1) = -4$$



The graph of f is the paraboloid $z = 4 - x^2 - 2y^2$ and the vertical plane $y = 1$ intersects it in the parabola $z = 2 - x^2$, $y = 1$. (As in the preceding discussion, we label it C_1 in Figure 2.) The slope of the tangent line to this parabola at the point $(1, 1, 1)$ is $f_x(1, 1) = -2$. Similarly, the curve C_2 in which the plane $x = 1$ intersects the paraboloid is the parabola $z = 3 - 2y^2$, $x = 1$, and the slope of the tangent line at $(1, 1, 1)$ is $f_y(1, 1) = -4$. (See Figure 3.)

Figure 4 is a computer-drawn counterpart to Figure 2. Part (a) shows the plane $y = 1$ intersecting the surface to form the curve C_1 and part (b) shows C_1 and T_1. [We have used the vector equations $\mathbf{r}(t) = \langle t, 1, 2 - t^2 \rangle$ for C_1 and $\mathbf{r}(t) = \langle 1 + t, 1, 1 - 2t \rangle$ for T_1.] Similarly, Figure 5 corresponds to Figure 3.

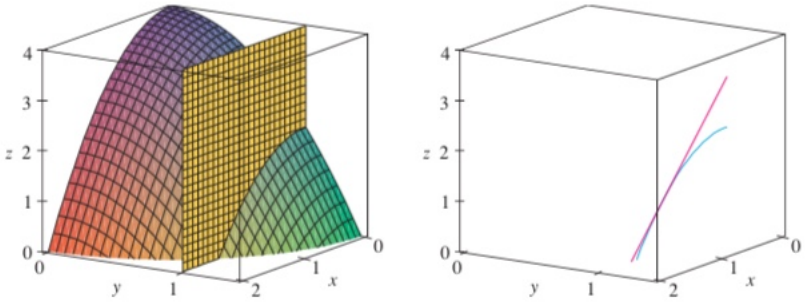


FIGURE 4

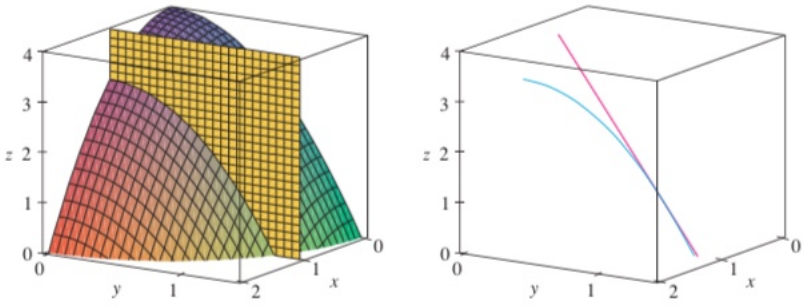


FIGURE 5

Exercise 2.9. If $f(x,y) = \sin\left(\frac{x}{1+y}\right)$, calculate $\frac{\partial f}{\partial x}$ and $\frac{\partial f}{\partial y}$.

SOLUTION Using the Chain Rule for functions of one variable, we have

$$\frac{\partial f}{\partial x} = \cos\left(\frac{x}{1+y}\right) \cdot \frac{\partial}{\partial x}\left(\frac{x}{1+y}\right) = \cos\left(\frac{x}{1+y}\right) \cdot \frac{1}{1+y}$$

$$\frac{\partial f}{\partial y} = \cos\left(\frac{x}{1+y}\right) \cdot \frac{\partial}{\partial y}\left(\frac{x}{1+y}\right) = -\cos\left(\frac{x}{1+y}\right) \cdot \frac{x}{(1+y)^2}$$

Exercise 2.10. If $f(x,y) = y\sin(xy)$, calculate $\frac{\partial f}{\partial x}$ and $\frac{\partial f}{\partial y}$.

Solution.

$$\frac{\partial f}{\partial x} = y^2\cos(xy), \quad \frac{\partial f}{\partial y} = \sin(xy) + xy\cos(xy).$$

□

Exercise 2.11. If $f(x,y) = \frac{2y}{y+\cos(x)}$, calculate $\frac{\partial f}{\partial x}$ and $\frac{\partial f}{\partial y}$.

Solution.

$$\frac{\partial f}{\partial x} = \frac{2y\sin(x)}{(y+\cos(x))^2}, \quad \frac{\partial f}{\partial y} = \frac{2\cos(x)}{(y+\cos(x))^2}.$$

□

2.5.4 Functions of More than Two Variables

Partial derivatives can also be defined for functions of three or more variables. For example, if f is a function of three variables x, y, and z, then its partial derivative with respect to x is defined as

$$f_x(x,y,z) = \lim_{h\to 0} \frac{f(x+h,y,z) - f(x,y,z)}{h}$$

and it is found by regarding y and z as constants and differentiating $f(x,y,z)$ with respect to x. If $w = f(x,y,z)$, then $f_x = \partial w/\partial x$ can be interpreted as the rate of change of w with respect to x when y and z are held fixed. But we can't interpret it geometrically because the graph of f lies in four-dimensional space.

In general, if u is a function of n variables, $u = f(x_1, x_2, \ldots, x_n)$, its partial derivative with respect to the ith variable x_i is

$$\frac{\partial u}{\partial x_i} = \lim_{h\to 0} \frac{f(x_1,\ldots,x_{i-1},x_i+h,x_{i+1},\ldots,x_n) - f(x_1,\ldots,x_i,\ldots,x_n)}{h}$$

That is, we hold all the variables other than x_i constant and differentiate f with respect to x_i.

Exercise 2.12. Compute $\frac{\partial f}{\partial x}$, $\frac{\partial f}{\partial y}$, and $\frac{\partial f}{\partial z}$ if $f(x,y,z) = e^{xy}\ln z$.

SOLUTION Holding y and z constant and differentiating with respect to x, we have

$$f_x = ye^{xy}\ln z$$

Similarly, $f_y = xe^{xy}\ln z$ and $f_z = \dfrac{e^{xy}}{z}$

2.5.5 Implicit Partial Differentiation

Implicit partial differentiation works exactly as in the single variable case.

Example 2.27. If the equation

$$yz - \ln z = x + y$$

defines z implicitly as a function of x and y, then we compute $\frac{\partial z}{\partial x}$ by differentiating each side with respect to x:

$$y\frac{\partial z}{\partial x} - \frac{1}{z}\frac{\partial z}{\partial x} = 1$$
$$\Longrightarrow \frac{\partial z}{\partial x} = \frac{1}{y - \frac{1}{z}}.$$

Exercise 2.13. Compute $\frac{\partial z}{\partial y}$ if the equation

$$yz - \ln z = x + y$$

defines z implicitly as a function of x and y.

Solution.

$$\frac{\partial z}{\partial y} = \frac{1 - z}{y}.$$

□

2.5.6 Higher Partial Derivatives

If f is a function of two variables, then its partial derivatives f_x and f_y are also functions of two variables, so we can consider their partial derivatives $(f_x)_x$, $(f_x)_y$, $(f_y)_x$, and $(f_y)_y$, which are called the **second partial derivatives** of f. If $z = f(x, y)$, we use the following notation:

$$(f_x)_x = f_{xx} = f_{11} = \frac{\partial}{\partial x}\left(\frac{\partial f}{\partial x}\right) = \frac{\partial^2 f}{\partial x^2} = \frac{\partial^2 z}{\partial x^2}$$

$$(f_x)_y = f_{xy} = f_{12} = \frac{\partial}{\partial y}\left(\frac{\partial f}{\partial x}\right) = \frac{\partial^2 f}{\partial y\,\partial x} = \frac{\partial^2 z}{\partial y\,\partial x}$$

$$(f_y)_x = f_{yx} = f_{21} = \frac{\partial}{\partial x}\left(\frac{\partial f}{\partial y}\right) = \frac{\partial^2 f}{\partial x\,\partial y} = \frac{\partial^2 z}{\partial x\,\partial y}$$

$$(f_y)_y = f_{yy} = f_{22} = \frac{\partial}{\partial y}\left(\frac{\partial f}{\partial y}\right) = \frac{\partial^2 f}{\partial y^2} = \frac{\partial^2 z}{\partial y^2}$$

Thus the notation f_{xy} (or $\partial^2 f/\partial y\,\partial x$) means that we first differentiate with respect to x and then with respect to y, whereas in computing f_{yx} the order is reversed.

Exercise 2.14. Compute the second partial derivatives of

$$f(x,y) = x^3 + x^2y^3 - 2y^2.$$

SOLUTION In Example 1 we found that

$$f_x(x, y) = 3x^2 + 2xy^3 \qquad f_y(x, y) = 3x^2y^2 - 4y$$

Therefore

$$f_{xx} = \frac{\partial}{\partial x}(3x^2 + 2xy^3) = 6x + 2y^3 \qquad f_{xy} = \frac{\partial}{\partial y}(3x^2 + 2xy^3) = 6xy^2$$

$$f_{yx} = \frac{\partial}{\partial x}(3x^2y^2 - 4y) = 6xy^2 \qquad f_{yy} = \frac{\partial}{\partial y}(3x^2y^2 - 4y) = 6x^2y - 4$$

Notice for the function in Exercise 2.14 that $\frac{\partial f}{\partial x \partial y} = \frac{\partial f}{\partial y \partial x}$. This is no coincidence.

Theorem 2.28 (Equality of mixed partial derivatives). Let $f : U \subseteq \mathbb{R}^2 \to \mathbb{R}$. If $\frac{\partial f}{\partial x \partial y}$ and $\frac{\partial f}{\partial y \partial x}$ are both continuous on U, then $\frac{\partial f}{\partial x \partial y} = \frac{\partial f}{\partial y \partial x}$ throughout U.

Proof. Let $h \neq 0$ and consider the quantity

$$\Delta(h) \equiv [f(a+h, b+h) - f(a+h, b)] - [f(a, b+h) - f(a, b)]. \tag{2.1}$$

If we let $g(x) \equiv f(x, b+h) - f(x, b)$, then

$$\Delta(h) = g(a+h) - g(a).$$

If g is differentiable on U, by the Mean Value Theorem there exists $c \in (a, a+h)$ such that

$$\begin{aligned} g(a+h) - g(a) &= hg'(c) \\ &= h\left[\frac{\partial f}{\partial x}(c, b+h) - \frac{\partial f}{\partial x}(c, b)\right]. \end{aligned}$$

Applying the Mean Value Theorem again to $\frac{\partial f}{\partial x}$, there exists $d \in (b, b+h)$ such that

$$\frac{\partial f}{\partial x}(c, b+h) - \frac{\partial f}{\partial x}(c, b) = h\frac{\partial^2 f}{\partial y \partial x}(c, d).$$

Thus,

$$\Delta(h) = h^2 \frac{\partial^2 f}{\partial y \partial x}(c, d).$$

As $h \to 0$, $(c, d) \to (a, b)$, so by continuity of $\frac{\partial^2 f}{\partial y \partial x}$,

$$\lim_{h \to 0} \frac{\Delta(h)}{h^2} = \lim_{(c,d) \to (a,b)} \frac{\partial^2 f}{\partial y \partial x}(c, d) = \frac{\partial^2 f}{\partial y \partial x}(a, b).$$

Now regroup the terms in (2.1) as

$$\Delta(h) = [f(a+h, b+h) - f(a, b+h)] - [f(a+h, b) - f(a, b)]$$

and let $u(y) \equiv f(a+h, y) - f(a, y)$ so that $\Delta(h) = u(b+h) - u(b)$. By the Mean Value Theorem there exists $d \in (b, b+h)$ such that

$$\begin{aligned} u(b+h) - u(b) &= hg'(d) \\ &= h\left[\frac{\partial f}{\partial y}(a+h, d) - \frac{\partial f}{\partial y}(a, d)\right]. \end{aligned}$$

Applying the Mean Value Theorem again, there exists $c \in (a, a+h)$ such that

$$\frac{\partial f}{\partial y}(a+h, d) - \frac{\partial f}{\partial y}(a, d) = h\frac{\partial^2 f}{\partial x \partial y}(c, d),$$

so

$$\Delta(h) = h^2 \frac{\partial^2 f}{\partial x \partial y}(c, d).$$

Since $(c, d) \to (a, b)$ as $h \to 0$, by continuity of $\frac{\partial^2 f}{\partial x \partial y}$ we have

$$\lim_{h \to 0} \frac{\Delta(h)}{h^2} = \lim_{(c,d) \to (a,b)} \frac{\partial^2 f}{\partial x \partial y}(c, d) = \frac{\partial^2 f}{\partial x \partial y}(a, b).$$

Thus,

$$\frac{\partial^2 f}{\partial y \partial y}(a, b) = \frac{\partial^2 f}{\partial x \partial y}(a, b).$$

□

Exercise 2.15. Compute the mixed partial derivatives of the function

$$f(x, y) = xy + e^y(y^2 + 1)$$

and verify that these are equal. In which order do we get the answer more quickly?

Exercise 2.16. For each function below, determine which order of computing the mixed partial derivatives will give you the answer the fastest and compute them.

(a) $f(x, y) = x \sin y + e^y$

(b) $f(x, y) = 1/x$

(c) $f(x, y) = y + x^2 y + 4y^3 - \ln(y^2 + 1)$

(d) $f(x, y) = x \ln(xy)$

We can partially-differentiate a function as many times as we like as long as the derivatives exist:

$$\frac{\partial^3 f}{\partial y \partial y \partial x}(x,y), \frac{\partial^4 f}{\partial y \partial y \partial x \partial x}(x,y), \dots.$$

By repeated application of Theorem 2.28, the order does not matter as long as the partial derivatives are all continuous.

Exercise 2.17. Compute

$$\frac{\partial^5 f}{\partial x \partial x \partial y \partial y \partial y}(x,y)$$

if $f(x,y) = y^5 x^4 e^x + 2\sin(3x)\cos(2y)$.

Example 2.29 (A function whose mixed partials are unequal). Consider the following function

$$f(x,y) = \begin{cases} \frac{xy^3 - x^3 y}{x^2+y^2}, & (x,y) \neq (0,0), \\ 0, & (x,y) = (0,0). \end{cases}$$

For $(x,y) \neq (0,0)$, the quotient rule gives

$$\begin{aligned} \frac{\partial f}{\partial x} &= \frac{y^3 - 3x^2 y}{x^2+y^2} - \frac{2x(xy^3 - x^3 y)}{(x^2+y^2)^2} \\ &= \frac{y^5 - 4x^2 y^3 - x^4 y}{(x^2+y^2)^2}. \end{aligned}$$

The partial derivative with respect to x at $(0,0)$ is

$$\frac{\partial f}{\partial x}(0,0) = \lim_{h \to 0} \frac{f(h,0) - f(0,0)}{h} = 0,$$

so

$$\frac{\partial f}{\partial x} = \begin{cases} \frac{y^5 - 4x^2 y^3 - x^4 y}{(x^2+y^2)^2}, & (x,y) \neq (0,0) \\ 0, & (x,y) = (0,0). \end{cases}$$

A similar calculation gives

$$\frac{\partial f}{\partial y} = \begin{cases} \frac{-x^5 + 4x^3 y^2 + x y^4}{(x^2+y^2)^2}, & (x,y) \neq (0,0) \\ 0, & (x,y) = (0,0). \end{cases}$$

The second partial derivative with respect to y is given by, when $(x,y) \neq (0,0)$

$$\frac{\partial f}{\partial y \partial x} = \frac{-x^6 - 9x^4 y^2 + 9x^2 y^4 + y^6}{(x^2+y^2)^3}$$

while at $(0,0)$ it is given by

$$\frac{\partial f}{\partial y \partial x}(0,0) = \lim_{h \to 0} \frac{\frac{\partial f}{\partial x}(0,h) - \frac{\partial f}{\partial x}(0,0)}{h} = \lim_{h \to 0} \frac{1}{h} \frac{h^5 - 0 - 0}{h^4} = \lim_{h \to 0} \frac{h}{h} = 1.$$

Thus,

$$\frac{\partial f}{\partial y \partial x} = \begin{cases} \frac{-x^6 - 9x^4y^2 + 9x^2y^4 + y^6}{(x^2+y^2)^3}, & (x,y) \neq (0,0), \\ 1, & (x,y) = (0,0). \end{cases}$$

Now for $(x,y) \neq (0,0)$,

$$\frac{\partial f}{\partial x \partial y} = \frac{-x^6 - 9x^4y^2 + 9x^2y^4 + y^6}{(x^2+y^2)^3}$$

and for $(x,y) = (0,0)$

$$\frac{\partial f}{\partial x \partial y}(0,0) = \lim_{h \to 0} \frac{1}{h} \frac{-h^5 + 0 + 0}{h^4} = \lim_{h \to 0} \frac{-h}{h} = -1.$$

Thus,

$$\frac{\partial f}{\partial x \partial y} = \begin{cases} \frac{-x^6 - 9x^4y^2 + 9x^2y^4 + y^6}{(x^2+y^2)^3}, & (x,y) \neq (0,0) \\ -1, & (x,y) = (0,0). \end{cases}$$

We see that the mixed partial derivatives are not equal at the origin. The reason for this is that the mixed partial derivatives are not continuous at the origin. To see this, consider approaching the origin along the path along the line $y = x$. Then

$$\lim_{(h,h) \to (0,0)} \frac{\partial f}{\partial x \partial y} = \lim_{(h,h) \to (0,0)} \frac{\partial f}{\partial y \partial x} \frac{0}{8h^6} = 0.$$

Since the mixed partials are discontinuous at the origin, they do not satisfy the hypothesis of Theorem 2.28 at the origin, and they are unequal.

2.5.7 Application: Partial Differential Equations

A *partial differential equation* (PDE) is an equation relating a function to one or more of its partial derivatives. PDEs arise frequently in physics and engineering. A famous one is *Laplace's equation*

$$\frac{\partial^2}{\partial x^2} u(x,y) + \frac{\partial^2}{\partial y^2} u(x,y) = 0.$$

Solutions of Laplace's equation which are continuous are called *harmonic functions*. These play an important role in heat conduction, fluid flow, and electric potential.

Exercise 2.18. Show that the function $u(x,y) = e^x \sin y$ is harmonic.

Another important PDE in physics and engineering is the *wave equation*

$$\frac{\partial^2}{\partial t^2} u(x,t) - v^2 \frac{\partial^2}{\partial x^2} u(x,t) = 0,$$

which describes the motion of a wave propagating in the x-direction, such as a wave traveling on a vibrating string. The function $u(x,t)$ describes the displacement of the string at a distance x from one end of the string at time t. The constant v is the speed, which depends on the density of the string and the string tension.

Exercise 2.19. Verify that the function $u(x,t) = \sin(x - vt)$ is a solution of the wave equation.

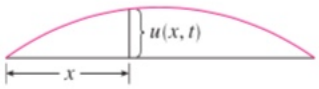


Figure 45: A solution $u(x,t)$ of the wave equation describes the displacement of a wave from equilibrium at position x and time t.

2.6 Differentiability

If $f : U \subseteq \mathbb{R}$ is differentiable at $x_0 \in U$, then the limit

$$f'(x_0) = \lim_{h \to 0} \frac{f(x_0 + h) - f(x_0)}{h} \tag{2.2}$$

exists. The existence of this limit implies that $f(x)$ is well-approximated by its tangent line

$$f(x) = f(x_0) + f'(x_0)(x - x_0)$$

for x in a neighborhood of x_0. Recall also that a differentiable function is automatically continuous, since if f is differentiable at x_0, then

$$\begin{aligned}
\lim_{h \to 0}(f(x_0 + h) - f(x_0)) &= \lim_{h \to 0} \frac{h}{h}(f(x_0 + h) - f(x_0)) \\
&= \lim_{h \to 0} \left(h \cdot \frac{f(x_0 + h) - f(x_0)}{h} \right) \\
&= \lim_{h \to 0} h \cdot \lim_{h \to 0} \frac{f(x_0 + h) - f(x_0)}{h} \\
&= \lim_{h \to 0} h f'(x_0) \\
&= 0 \cdot f'(x_0) \\
&= 0.
\end{aligned}$$

However, the next example shows that a multivariable function can be discontinuous at a point where the partial derivatives exist.

Example 2.30. Let

$$f(x,y) = \begin{cases} 0 \text{ if } xy \neq 0, \\ 1 \text{ if } xy = 0 \end{cases}$$

Since $\lim_{(x,y) \to (0,0)} f(x,y)$ is 1 along the x-axis and 0 along the line $y = x$, the limit does not exist and therefore $f(x,y)$ is discontinuous at $(0,0)$. However,

$$\begin{aligned}
\frac{\partial f}{\partial x}(0,0) &= \lim_{h \to 0} \frac{f(h,0) - f(0,0)}{h} = \lim_{h \to 0} \frac{1-1}{h} = 0 \\
\frac{\partial f}{\partial y}(0,0) &= \lim_{h \to 0} \frac{f(0,h) - f(0,0)}{h} = \lim_{h \to 0} \frac{1-1}{h} = 0
\end{aligned}$$

so the partial derivatives both exist at $(0,0)$.

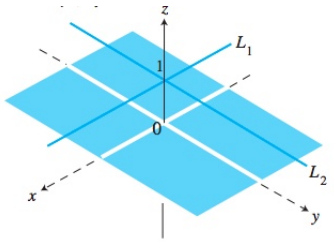


Figure 46: Graph of the function $f(x,y)$ in Example 2.30

The reason for this is that continuity at (a,b) requires that $\lim_{(x,y)\to(a,b)} f(x,y) = f(a,b)$ when the limit is taken along *any* path passing through (a,b). The partial derivatives, on the other hand, only depend on what is happening along lines through (a,b) parallel to the x and y axes, and therefore they failed to detect the discontinuity in the previous example. This suggests we need a stronger condition in our definition of differentiability for multivariable functions than the existence of the partial derivatives if we want differentiable multivariable functions to behave similarly to differentiable single variable functions.

To see what definition to take, it is useful to view the definition of differentiability of a single-variable function from another perspective. We do this as follows. First, let $\Delta f(h) \equiv f(x+h) - f(x)$, where we view $\Delta f : \mathbb{R} \to \mathbb{R}$ as a function of h with x held fixed, which gives the change in f when we change x to $x+h$. We then rearrange (2.2) as

$$\begin{aligned}
\lim_{h\to 0}\left[\frac{\Delta f(h)}{h}\right] - f'(x) &= 0 \\
\lim_{h\to 0}\left[\frac{\Delta f(h)}{h} - f'(x)\right] &= 0 \\
\lim_{h\to 0}\left[\frac{\Delta f(h)}{h} - \frac{f'(x)h}{h}\right] &= 0 \\
\lim_{h\to 0}\left[\frac{\Delta f(h) - f'(x)h}{h}\right] &= 0
\end{aligned}$$

If we let $r(h) \equiv \Delta f(h) - f'(x)h$, Then existence of the limit in equation (2.2) implies that

$$\Delta f(h) = f'(x)h + r(h) \tag{2.3}$$

where $\lim_{h\to 0} \frac{r(h)}{h} = 0$. That is, for all points in a sufficiently small neighborhood of x,

$$\Delta f(h) \approx f'(x)h \tag{2.4}$$

which is a *linear* function of h.

Definition 2.31. If f is differentiable at x, we call the linear function

$$df_x : \mathbb{R} \to \mathbb{R}$$

defined by $df_x(h) = f'(x)h$ the *differential of f at x*.

Thus existence of the limit in (2.2) is equivalent to the statement that there exists a *linear* function (the differential of f at x) that approximates the change of f near x, up to an error term which can be made arbitrarily small by taking points sufficiently close to x.

Definition 2.32. Let $f : U \subseteq \mathbb{R}^n \to \mathbb{R}$ be a multivariable function defined on an open subset U of $\mathbb{R}^n$ and let $\Delta f : \mathbb{R}^n \to \mathbb{R}$ be defined by $\Delta f(\mathbf{h}) = f(\mathbf{x}+\mathbf{h}) - f(\mathbf{x})$. We say that f is *differentiable* at $\mathbf{x} \in U$ if there exists a linear function $T : \mathbb{R}^n \to \mathbb{R}$ such that

$$\Delta f(\mathbf{h}) = T(\mathbf{h}) + r(\mathbf{h})$$

where $r(\mathbf{h})$ is a "small" error term, in the sense that $\lim_{\mathbf{h}\to\mathbf{0}} \frac{|r(\mathbf{h})|}{||\mathbf{h}||} = 0$. If T exists, we denote it by $df_{\mathbf{x}}$ and call it the *differential of f at* $\mathbf{x}$.

Writing this out, this says that f is differentiable at $\mathbf{x}$ if for every $\epsilon > 0$, there exists a linear function $df_{\mathbf{x}} : \mathbb{R}^n \to \mathbb{R}$ and a number $\delta > 0$ such that

$$|f(\mathbf{x}+\mathbf{h}) - f(\mathbf{x}) - df_{\mathbf{x}}(\mathbf{h})| < \epsilon||\mathbf{h}||$$

whenever $||\mathbf{h}|| < \delta$.

We see immediately that if the differential of f at $\mathbf{x}$ exists, it is unique.

Proposition 2.33 (Uniqueness of the differential). If $T, T' : \mathbb{R}^n \to \mathbb{R}$ are two linear maps which satisfy Definition 2.32, then $T' = T$.

Proof. Suppose $T, T' : \mathbb{R}^n \to \mathbb{R}$ are linear maps such that

$$\begin{aligned} f(\mathbf{x}+\mathbf{h}) &= f(\mathbf{x}) + T(\mathbf{h}) + r(\mathbf{h}), \\ f(\mathbf{x}+\mathbf{h}) &= f(\mathbf{x}) + T'(\mathbf{h}) + r'(\mathbf{h}) \end{aligned}$$

with

$$\lim_{\mathbf{h}\to\mathbf{0}} \frac{|r(\mathbf{h})|}{||\mathbf{h}||} = \lim_{\mathbf{h}\to\mathbf{0}} \frac{|r'(\mathbf{h})|}{||\mathbf{h}||} = 0.$$

Then, since T, T' are linear,

$$\begin{aligned} |(T-T')(\mathbf{h})| &= |T(\mathbf{h}) - T'(\mathbf{h})| \\ &= |T(\mathbf{h}) - T'(\mathbf{h}) + f(\mathbf{x}+\mathbf{h}) - f(\mathbf{x}) - (f(\mathbf{x}+\mathbf{h}) - f(\mathbf{x}))| \\ &= |T(\mathbf{h}) - (f(\mathbf{x}+\mathbf{h}) - f(\mathbf{x})) + f(\mathbf{x}+\mathbf{h}) - f(\mathbf{x}) - T'(\mathbf{h})| \\ &\leq |T(\mathbf{h}) - (f(\mathbf{x}+\mathbf{h}) - f(\mathbf{x}))| + |f(\mathbf{x}+\mathbf{h}) - f(\mathbf{x}) - T'(\mathbf{h})| \\ &= |f(\mathbf{x}+\mathbf{h}) - f(\mathbf{x}) - T(\mathbf{h})| + |f(\mathbf{x}+\mathbf{h}) - f(\mathbf{x}) - T'(\mathbf{h})| \\ &= |r(\mathbf{h})| + |r'(\mathbf{h})| \end{aligned}$$

and therefore

$$\frac{|(T-T')(\mathbf{h})|}{||\mathbf{h}||} \leq \frac{|r(\mathbf{h})|}{||\mathbf{h}||} + \frac{|r'(\mathbf{h})|}{||\mathbf{h}||}.$$

Since

$$\lim_{\mathbf{h}\to\mathbf{0}} \frac{|r(\mathbf{h})|}{||\mathbf{h}||} = \lim_{\mathbf{h}\to\mathbf{0}} \frac{|r'(\mathbf{h})|}{||\mathbf{h}||} = 0,$$

given $\epsilon > 0$ there exist positive real numbers δ_1, δ_2 such that

$$\frac{|r(\mathbf{h})|}{||\mathbf{h}||} < \frac{\epsilon}{2}$$

whenever $||\mathbf{h}|| < \delta_1$ and

$$\frac{|r'(\mathbf{h})|}{||\mathbf{h}||} < \frac{\epsilon}{2}$$

whenever $||\mathbf{h}|| < \delta_2$. By choosing $\delta = \min\{\delta_1, \delta_2\}$, we have

$$\frac{|(T - T')(\mathbf{h})|}{||\mathbf{h}||} \leq \frac{|r(\mathbf{h})|}{||\mathbf{h}||} + \frac{|r'(\mathbf{h})|}{||\mathbf{h}||} < \frac{\epsilon}{2} + \frac{\epsilon}{2}$$

whenever $||\mathbf{h}|| < \delta$, and therefore

$$\lim_{\mathbf{h} \to \mathbf{0}} \frac{|(T - T')(\mathbf{h})|}{||\mathbf{h}||} = 0.$$

Since this limit holds as $\mathbf{h} \to \mathbf{0}$ along any path, for any fixed $\mathbf{h} \neq \mathbf{0}$ we must have

$$\lim_{\mathbf{h} \to \mathbf{0}} \frac{|(T - T')(t\mathbf{h})|}{||t\mathbf{h}||} = 0.$$

Since $T - T'$ is linear, this becomes

$$\begin{aligned}
0 &= \lim_{t \to 0} \frac{|(T - T')(t\mathbf{h})|}{||t\mathbf{h}||} \\
&= \lim_{t \to 0} \frac{|t(T - T')(\mathbf{h})|}{||t\mathbf{h}||} \\
&= \lim_{t \to 0} \frac{|t| \, |(T - T')(\mathbf{h})|}{|t| \, ||\mathbf{h}||} \\
&= \lim_{t \to 0} \frac{|(T - T')(\mathbf{h})|}{||\mathbf{h}||} \\
&= \frac{|(T - T')(\mathbf{h})|}{||\mathbf{h}||},
\end{aligned}$$

which implies that $(T - T')(\mathbf{h}) = 0$ and therefore $\mathbf{h} \in \ker(T - T')$. Since $\mathbf{h}$ was an arbitrary nonzero vector, this means that $\ker(T - T') = \mathbb{R}^2$, so $T - T'$ is the zero map, and therefore $T = T'$. □

From Definition 2.32, we see that if $f : U \subseteq \mathbb{R}^n \to \mathbb{R}$ is a linear function, then f is differentiable at each point $\mathbf{x} \in U$ and $df_{\mathbf{x}} = f$.

Proposition 2.34. If $f : U \subseteq \mathbb{R}^n \to \mathbb{R}$ is linear, then for each $\mathbf{x} \in U$, $df_{\mathbf{x}} = f$.

Proof. Since f is linear,

$$\begin{aligned}
\lim_{\mathbf{h} \to \mathbf{0}} \frac{|f(\mathbf{x} + \mathbf{h}) - f(\mathbf{x}) - f(\mathbf{h})|}{||\mathbf{h}||} &= \lim_{\mathbf{h} \to \mathbf{0}} \frac{|f(\mathbf{x}) + f(\mathbf{h}) - f(\mathbf{x}) - f(\mathbf{h})|}{||\mathbf{h}||} \\
&= \lim_{\mathbf{h} \to \mathbf{0}} \frac{0}{||\mathbf{h}||} \\
&= 0.
\end{aligned}$$

Thus, $df_{\mathbf{x}} = f(\mathbf{x})$. □

Theorem 2.35 (A differentiable function is continuous). If $f : U \subset \mathbb{R}^n \to \mathbb{R}$ is differentiable at $\mathbf{x} \in U$, then it is continuous at $\mathbf{x} \in U$.

Proof. Let $\epsilon > 0$. Then

$$\begin{aligned} |f(\mathbf{x}+\mathbf{h}) - f(\mathbf{x})| &= |df_{\mathbf{x}}(\mathbf{h}) + r(\mathbf{h})| \\ &\leq |df_{\mathbf{x}}(\mathbf{h})| + |r(\mathbf{h})|. \end{aligned}$$

Choose $\delta_0 > 0$ such that $|r(\mathbf{h})| < \epsilon||\mathbf{h}||$ whenever $||\mathbf{h}|| < \delta_0$. Then

$$|f(\mathbf{x}+\mathbf{h}) - f(\mathbf{x})| < |df_{\mathbf{x}}(\mathbf{h})| + \epsilon||\mathbf{h}||.$$

By ???, $|df_{\mathbf{x}}(\mathbf{h})| \leq ||df_{\mathbf{x}}||\,||\mathbf{h}||$, so

$$\begin{aligned} |f(\mathbf{x}+\mathbf{h}) - f(\mathbf{x})| &< ||df_{\mathbf{x}}||\,||\mathbf{h}|| + \epsilon||\mathbf{h}|| \\ &= (||df_{\mathbf{x}}|| + \epsilon)||\mathbf{h}||. \end{aligned}$$

Let $\delta = \min\{\delta_0, \frac{\epsilon}{||df_{\mathbf{x}}||+\epsilon}\}$. Then by taking $||\mathbf{h}|| < \delta$, we have

$$\begin{aligned} |f(\mathbf{x}+\mathbf{h}) - f(\mathbf{x})| &< (||df_{\mathbf{x}}|| + \epsilon)||\mathbf{h}|| \\ &< (||df_{\mathbf{x}}|| + \epsilon)\frac{\epsilon}{(||df_{\mathbf{x}}|| + \epsilon)} \\ &= \epsilon, \end{aligned}$$

which shows that $\lim_{\mathbf{h}\to\mathbf{0}}(f(\mathbf{x}+\mathbf{h}) - f(\mathbf{x})) = 0$. Hence, f is continuous. □

In the next section, we show how to compute the differential of a differentiable function.

2.6.1 Directional Derivatives

Our next task is to show how to compute $df_{\mathbf{x}}(\mathbf{v})$ for any $\mathbf{v} \in \mathbb{R}^n$. We now show that this computation can be reduced to the derivative of a single-variable function.

Theorem 2.36 (Computation of the differential). Let $f : U \subseteq \mathbb{R}^n \to \mathbb{R}$ be differentiable at $\mathbf{x} \in U$ and let $\mathbf{r} : I \subseteq \mathbb{R} \to U$ be the straight-line path $\mathbf{r}(t) = \mathbf{x} + t\mathbf{v}$ passing through $\mathbf{x}$ in the direction of $\mathbf{v}$ at $t = 0$. Then the single-variable function $f \circ \mathbf{r} : I \subseteq \mathbb{R} \to \mathbb{R}$ is the single-variable function given by $f(\mathbf{r}(t))$ is differentiable and

$$df_{\mathbf{x}}(\mathbf{v}) = \frac{d}{dt}f(\mathbf{x}+t\mathbf{v})|_{t=0}.$$

Proof. Since f is differentiable at $\mathbf{x}$, for every $\epsilon > 0$ there exists $\delta > 0$ such that $||\mathbf{h}|| < \delta$ implies

$$|f(\mathbf{x}+\mathbf{h}) - f(\mathbf{x}) - df_{\mathbf{x}}(\mathbf{h})| < \epsilon||\mathbf{h}||.$$

Let $\mathbf{v}$ be a fixed nonzero vector and choose $t \in \mathbb{R}$ sufficiently small so that $||t\mathbf{v}|| < \delta$. Then

$$|f(\mathbf{x}+t\mathbf{v}) - f(\mathbf{x}) - df_{\mathbf{x}}(t\mathbf{v})| < \epsilon||t\mathbf{v}||.$$

Since $df_{\mathbf{x}}$ is linear, this implies

$$|f(\mathbf{x}+t\mathbf{v}) - f(\mathbf{x}) - tdf_{\mathbf{x}}(\mathbf{v})| < \epsilon|t|\,||\mathbf{v}||.$$

Dividing through by $|t|$, we have

$$\left| \frac{f(\mathbf{x}+t\mathbf{v}) - f(\mathbf{x})}{t} - df_{\mathbf{x}}(\mathbf{v}) \right| < \epsilon ||\mathbf{v}||.$$

Since the quantity on the left can be made smaller than any positive number by taking t sufficiently small, it must be that

$$\lim_{t\to 0} \frac{f(\mathbf{x}+t\mathbf{v}) - f(\mathbf{x})}{t} \equiv \frac{d}{dt} f(\mathbf{x}+t\mathbf{v})|_{t=0} = df_{\mathbf{x}}(\mathbf{v}).$$

□

Definition 2.37 (Directional derivative). The real number $\frac{d}{dt} f(\mathbf{x}+t\mathbf{v})|_{t=0}$ is called a *directional derivative* of f, and is often denoted $D_{\mathbf{v}} f(\mathbf{x})$.

Remark. Strictly speaking, in Definition 2.37 we are misusing the word "direction", since if $\mathbf{w} = c\mathbf{v}$ with $c > 0$, then $\mathbf{w}$ and $\mathbf{v}$ point in the same direction but Theorem 2.36 shows that $D_{\mathbf{w}} f(\mathbf{x}) = df_{\mathbf{x}}(\mathbf{w}) = df_{\mathbf{x}}(c\mathbf{v}) = cD_{\mathbf{v}}(\mathbf{x})$. Thus, when we speak of directional derivatives we will always take $\mathbf{v}$ in Definition 2.36 to be a unit vector. □

Corollary 2.38. If $f : U \subseteq \mathbb{R}^n \to \mathbb{R}$ is differentiable at $\mathbf{x}$, then all of its directional derivatives exist at $\mathbf{x}$. In particular, all the partial derivatives of f exist at $\mathbf{x}$.

Proof. By taking $\mathbf{v} = e_i$ to be the ith standard basis vector for $\mathbb{R}^n$, by Theorem 2.36,

$$\begin{aligned} df_{\mathbf{x}}(e_i) &= \lim_{t\to 0} \frac{f(\mathbf{x}+te_i) - f(\mathbf{x})}{t} \\ &= \lim_{t\to 0} \frac{f(x_1, \ldots, x_i + t, \ldots, x_n) - f(x_1, \ldots, x_i, \ldots, x_n)}{t} \\ &\equiv \frac{\partial f}{\partial x_i}(x_1, \ldots, x_n). \end{aligned}$$

□

If $f : \mathbb{R}^2 \to \mathbb{R}$ is differentiable at $\mathbf{x}$, the equation of the *tangent plane* to the graph of f at $(\mathbf{x}, f(\mathbf{x}))$ is

$$\begin{aligned} f(x,y) &= f(x_0,y_0) + T(x-x_0, y-y_0) \\ &= f(x_0,y_0) + \begin{bmatrix} \frac{\partial f(x_0,y_0)}{\partial x} & \frac{\partial f(x_0,y_0)}{\partial y} \end{bmatrix} \begin{bmatrix} x-x_0 \\ y-y_0 \end{bmatrix} \\ &= f(x_0,y_0) + \frac{\partial f(x_0,y_0)}{\partial x}(x-x_0) + \frac{\partial f(x_0,y_0)}{\partial y}(y-y_0). \end{aligned}$$

The converse to Corollary 2.38 is *false*, as the next example shows.

Example 2.39. A function $f : \mathbb{R}^n \to \mathbb{R}$ is said to be *homogeneous* if $f(c\mathbf{x}) = cf(\mathbf{x})$ for all $c \in \mathbb{R}$ and $\mathbf{x} \in \mathbb{R}^n$. If f is homogeneous, then for any nonzero $\mathbf{v} \in \mathbb{R}^n$ we have

$$\begin{aligned} D_{\mathbf{v}}(0) &= \lim_{t\to 0} \frac{f(0+t\mathbf{v}) - f(0)}{t} \\ &= \lim_{t\to 0} \frac{f(t\mathbf{v})}{t} \\ &= \lim_{t\to 0} \frac{tf(\mathbf{v})}{t} \\ &= \lim_{t\to 0} f(\mathbf{v}) \\ &= f(\mathbf{v}), \end{aligned}$$

which shows that all the directional derivatives of a homogeneous function exist at 0 at $D_{\mathbf{v}}f(0) = f(\mathbf{v})$. If f is also differentiable at 0, then

$$df_0(\mathbf{v}) = D_{\mathbf{v}}(0) = f(\mathbf{v}),$$

thus $f = df_0$, so f is linear by Proposition 2.34. Therefore, if f is any nonlinear homogeneous function, then all of its directional derivatives will exist, but f will not be differentiable. For instance, define $f : \mathbb{R}^2 \to \mathbb{R}$ by

$$f(x,y) = \begin{cases} \frac{x^3}{x^2+y^2}, & \text{if } (x,y) \neq (0,0) \\ 0, & \text{if } (x,y) = (0,0). \end{cases}$$

Then $f(tx,ty) = tf(x,y)$, so f is homogeneous, but not linear.

The converse to Corollary 2.38 *does* hold if we add the condition that the directional derivatives are *continuous*. It suffices to show this for the directional derivatives along a basis for $\mathbb{R}^n$, i.e., for the partial derivatives.

Theorem 2.40. Let U be an open subset of $\mathbb{R}^n$ and let $f : U \to \mathbb{R}$. Then, if each of the partial derivatives $\frac{\partial f}{\partial x_i}$ exists and is continuous on U, then f is differentiable on U.

Proof. To cut down on writing, we will prove this for the case of $n = 2$. The proof for $n > 2$ is exactly the same. We need to show that, for any $(x,y) \in U$ and $\epsilon > 0$, there exists a $\delta > 0$ such that

$$|f(x+h_1, y+h_2) - f(x,y) - df_{(x,y)}(\mathbf{h})| < \epsilon||\mathbf{h}||$$

whenever $||\mathbf{h}|| \equiv \sqrt{h_1^2 + h_2^2} < \delta$.

$$\begin{aligned} &|f(x+h_1, y+h_2) - f(x,y) - df_{(x,y)}(\mathbf{h})| \\ &= |f(x+h_1, y+h_2) - f(x,y) - \frac{\partial f}{\partial x}(x,y)h_1 - \frac{\partial f}{\partial y}(x,y)h_2| \\ &= |f(x+h_1, y+h_2) - f(x, y+h_2) + f(x, y+h_2) - f(x,y) - \frac{\partial f}{\partial x}(x,y)h_1 - \frac{\partial f}{\partial y}(x,y)h_2| \end{aligned}$$

By the Mean Value Theorem, there exists $u_1 \in (x, x+h_1)$ and $u_2 \in (y, y+h_2)$ such that

$$f(x+h_1, y+h_2) - f(x, y+h_2) = \frac{\partial f}{\partial x}(u_1, y+h_2)h_1$$
$$f(x, y+h_2) - f(x, y) = \frac{\partial f}{\partial y}(x, u_2)h_2$$

so the above is equal to

$$\begin{aligned}
&|(\frac{\partial f}{\partial x}(u_1, y+h_2) - \frac{\partial f}{\partial x}(x,y))h_1 + (\frac{\partial f}{\partial y}(x, u_2) - \frac{\partial f}{\partial y}(x,y))h_2| \\
\leq &|(\frac{\partial f}{\partial x}(u_1, y+h_2) - \frac{\partial f}{\partial x}(x,y), \frac{\partial f}{\partial y}(x, u_2) - \frac{\partial f}{\partial y}(x,y))|\ ||\mathbf{h}|| \\
\leq &(|\frac{\partial f}{\partial x}(u_1, y+h_2) - \frac{\partial f}{\partial x}(x,y)| + |\frac{\partial f}{\partial y}(x, u_2) - \frac{\partial f}{\partial y}(x,y)|)||\mathbf{h}||
\end{aligned}$$

where in the second line we have used the Cauchy-Schwarz inequality [16] and in the third line the fact that $\sqrt{x^2+y^2} \leq |x| + |y|$. Since the partial derivatives are continuous, and since u_1 and u_2 lie between x and $x+h_1$ and y and $y+h_2$, respectively, there exists a δ_1 such that the first absolute value in the parenthesis is less than $\epsilon/2$ whenever $||\mathbf{h}|| < \delta_1$ and a δ_2 such that the second absolute value in the parenthesis is less than $\epsilon/2$ whenever $||\mathbf{h}|| < \delta_2$. Thus, by taking $||\mathbf{h}|| < \delta = \min\{\delta_1, \delta_2\}$, the entire parenthesis is less than ϵ. □

2.6.2 Explicit computation of the differential

To compute the differential explicitly, we observe that it obeys the usual rules of differentiation.

Theorem 2.41 (Algebraic properties of the differential). Let f_1, f_2 be differentiable at $\mathbf{x}$. Then

(1) d is linear: $d(cf_1 + f_2) = cdf_1 + df_2$.

(2) d obeys the Leibnitz rule: $d(f_1 \cdot f_2) = df_1 \cdot f_2 + f_1 \cdot df_2$.

Proof. (1) For any $\mathbf{v} \in \mathbb{R}^n$, we have

$$d(cf_1 + f_2)(\mathbf{v}) = D_{\mathbf{v}}(cf_1 + f_2)(\mathbf{x}) = cD_{\mathbf{v}}f_1(\mathbf{x}) + D_{\mathbf{v}}f_2(\mathbf{x}) = cdf_1(\mathbf{v}) + df_2(\mathbf{v})$$

and therefore $d(cf_1 + f_2) = cdf_1 + df_2$.

(2) For any $\mathbf{v} \in \mathbb{R}^n$, we have

$$d(f_1 \cdot f_2)(\mathbf{v}) = D_{\mathbf{v}}(f_1(\mathbf{v}) \cdot f_2(\mathbf{v})) = (D_{\mathbf{v}}f_1(\mathbf{x}))f_2 + f_1(D_{\mathbf{v}}f_2(\mathbf{x})) = df_1(\mathbf{v}) \cdot f_2(\mathbf{v}) + f_1(\mathbf{v})df_2(\mathbf{v})$$

and therefore $d(f_1 \cdot f_2) = df_1 \cdot f_2 + f_1 \cdot df_2$.

□

Theorem 2.42 (Chain rule). Let $\mathbf{r} : I \to \mathbb{R}^2$ and $f : U \subseteq \mathbb{R}^2 \to \mathbb{R}$, where I is an interval in $\mathbb{R}$ and U is an open subset of $\mathbb{R}^2$ containing the trance of the curve $\mathbf{r}$. If $\mathbf{r}$ is differentiable at t and f is differentiable at $\mathbf{r}(t)$, then $f \circ \mathbf{r} : I \to \mathbb{R}$ is differentiable at t and

$$d(f \circ \mathbf{r})_t = df_{\mathbf{r}(t)} \circ d\mathbf{r}_t.$$

[16] The Cauchy-Schwarz inequality states that $|df_{(x,h)} \cdot \mathbf{h}| \leq |df_{(x,h)}|\ ||\mathbf{h}||$.

[Draw a picture.]

Proof. Since $\mathbf{r}$ is differentiable at t and f is differentiable at $\mathbf{r}(t)$, given $\epsilon_1, \epsilon_2 > 0$ there exist $\delta_1, \delta_2 > 0$ such that if $|k| < \delta_1$ and $||\mathbf{h}|| < \delta_2$, then $t + k \in I$, $\mathbf{r}(t) + \mathbf{h} \in U$ and

$$||\mathbf{r}(t+k) - \mathbf{r}(t) - d\mathbf{r}_t(k)|| \leq \epsilon_1 |k|, \tag{2.5}$$
$$|f(\mathbf{r}(t) + \mathbf{h}) - f(\mathbf{r}(t)) - df_{\mathbf{r}(t)}(\mathbf{h})| \leq \epsilon_2 ||\mathbf{h}||. \tag{2.6}$$

Set $\delta = \min\{\delta_1, \frac{\delta_2}{\epsilon_1 + ||d\mathbf{r}_t||}\}$. Let $k \in \mathbb{R}$ with $|k| < \delta$ and set $\mathbf{h} = \mathbf{r}(t+k) - \mathbf{r}(t)$. Then (2.5) implies

$$||\mathbf{h} - d\mathbf{r}_t(k)|| \leq \epsilon_1 |k|.$$

By the triangle inequality, we then have

$$\begin{aligned}
||\mathbf{h}|| &= ||\mathbf{h} - d\mathbf{r}_t(k) + d\mathbf{r}_t(k)|| \\
&\leq ||\mathbf{h} - d\mathbf{r}_t(k)|| + ||d\mathbf{r}_t(k)|| \\
&\leq \epsilon_1 |k| + ||d\mathbf{r}_t|| |k| \\
&= (\epsilon_1 + ||d\mathbf{r}_t||)|k| \\
&< (\epsilon_1 + ||d\mathbf{r}_t||) \frac{\delta_2}{(\epsilon_1 + ||d\mathbf{r}_t||)} \\
&= \delta_2
\end{aligned}$$

and therefore by (2.6)

$$\begin{aligned}
&|f(\mathbf{r}(t+k)) - f(\mathbf{r}(t)) - df_{\mathbf{r}(t)}(d\mathbf{r}_t(k))| = |f(\mathbf{r}(t+k) - \mathbf{r}(t) + \mathbf{r}(t)) - f(\mathbf{r}(t)) - df_{\mathbf{r}(t)}(d\mathbf{r}_t(k))| \\
=&|f(\mathbf{r}(t) + \mathbf{h}) - f(\mathbf{r}(t)) - df_{\mathbf{r}(t)}(d\mathbf{r}_t(k))| \\
=&|f(\mathbf{r}(t) + \mathbf{h}) - f(\mathbf{r}(t)) - df_{\mathbf{r}(t)}(\mathbf{h}) + df_{\mathbf{r}(t)}(\mathbf{h}) - df_{\mathbf{r}(t)}(d\mathbf{r}_t(k))| \\
\leq&|f(\mathbf{r}(t) + \mathbf{h}) - f(\mathbf{r}(t)) - df_{\mathbf{r}(t)}(\mathbf{h})| + |df_{\mathbf{r}(t)}(\mathbf{h}) - df_{\mathbf{r}(t)}(d\mathbf{r}_t(k))| \\
=&|f(\mathbf{r}(t) + \mathbf{h}) - f(\mathbf{r}(t)) - df_{\mathbf{r}(t)}(\mathbf{h})| + |df_{\mathbf{r}(t)}(\mathbf{h} - d\mathbf{r}_t(k))| \\
\leq&\epsilon_2 ||\mathbf{h}|| + ||df_{\mathbf{r}(t)}|| \epsilon_1 |k| \\
\leq&\epsilon_2(\epsilon_1 + ||d\mathbf{r}_t||)|k| + \epsilon_1 ||df_{\mathbf{r}(t)}||\ |k| \\
&= (\epsilon_2(\epsilon_1 + ||d\mathbf{r}_t||) + \epsilon_1 ||df_{\mathbf{r}(t)}||)|k|
\end{aligned}$$

Thus, given any $\epsilon > 0$, we can find $\delta_1, \delta_2 > 0$ such that (2.5) and (2.6) hold for

$$\epsilon_1 = \frac{1}{2} \frac{\epsilon}{||df_{\mathbf{r}(t)}||}$$
$$\epsilon_2 = \frac{1}{2} \frac{\epsilon}{\epsilon_1 + ||d\mathbf{r}_t||}.$$

Defining δ as before, by taking $|k| < \delta$ we therefore have

$$|f(\mathbf{r}(t+k)) - f(\mathbf{r}(t)) - df_{\mathbf{r}(t)}(d\mathbf{r}_t(k))| \leq \epsilon |k|.$$

This proves that $f \circ \mathbf{r}$ is differentiable at t with differential $df_{\mathbf{r}(t)} \circ d\mathbf{r}_t$. □

Since the matrix representing a composition of linear maps is the product of the matrices representing each linear map [Refer to my Linear Algebra notes], the standard matrix of $d(f \circ \mathbf{r})_t$ is

$$\begin{aligned}[d(f \circ \mathbf{r})_t] &= \begin{bmatrix} \frac{\partial f}{\partial x} & \frac{\partial f}{\partial y} \end{bmatrix} \begin{bmatrix} \frac{dx}{dt} \\ \frac{dx}{dt} \end{bmatrix} \\ &= \frac{\partial f}{\partial x}\frac{dx}{dt} + \frac{\partial f}{\partial y}\frac{dy}{dt},\end{aligned}$$

evaluated at $(x(t), y(t))$.

Example 2.43. Let $f(x,y) = x^2y + 3xy^4$ and $\mathbf{r}(t) = (x(t), y(t)) = (\sin 2t, \cos t)$. If $F(t) = f(x(t), y(t))$, then

$$\begin{aligned}\frac{dF}{dt} &= \frac{\partial f}{\partial x}\frac{dx}{dt} + \frac{\partial f}{\partial y}\frac{dy}{dt} \\ &= (2xy + 3y^4)2\cos(2t) + (x^2 + 12xy^3)(-\sin t)|_{(x,y)=(\sin 2t, \cos t)} \\ &= (2\sin(2t)\cos(t) + 3\cos^4(t))2\cos(2t) - (\sin^2(2t) + 12\sin(2t)\cos^3(t))\sin(t) \\ &= 4\sin(2t)\cos(2t)\cos(t) + 6\cos^4(t)\cos(2t) - \sin^2(2t)\sin(t) - 12\sin(2t)\cos^3(t)\sin(t).\end{aligned}$$

Exercise 2.20. Find $(f \circ \mathbf{r})'(t)$ for each function.

(1) $f(x,y) = x^2 + y^2 + xy$, $\mathbf{r}(t) = (\sin t, e^t)$.

(2) $f(x,y) = \cos(x + 4y)$, $\mathbf{r}(t) = (5t^4, t^{-1})$.

Theorem 2.44. If U is connected, f differentiable, and $df_p = 0$ for all p, then f is constant.

[Does this fit in somewhere? If so, discuss and prove this.]

2.6.3 Implicit Differentiation

Consider two real variables x and y related by an equation $F(x,y) = 0$. We would like to say that this implicitly defines a function $y = f(x)$, and we would like to compute $\frac{dy}{dx}$. One generally cannot solve for y explicitly, so it is important to know when such a function does indeed exist without having to solve for it.

For example, consider the function $F(x,y) = x^2 + y^2 - 1$. We are interested in those x and y related by the equation equation $F(x,y) = 0$, which is the equation of the unit circle centered at the origin. A function $f(x)$ is a "solution" if and only if $F(x, f(x)) = 0$ for all x in the domain of f. In this example, f must be given by $f(x) = \pm\sqrt{1 - x^2}$, and either of these is a solution. This shows that f need not be unique. Given (x_0, y_0) such that $F(x_0, y_0) = 0$, we would like to know if we can find $f(x)$ such that $F(x, f(x)) = 0$ and f is differentiable and *unique* near (x_0, y_0). In our example, if $x_0 \neq \pm 1$, then this is true if f is taken to be the appropriate square root. The (sign of the) given y_0 determines which square root must be selected. The points $x_0 = \pm 1$ are exceptional for several reasons. First, f is not differentiable there and second, near $x_0 = \pm 1$, f could be either square root, so f is not uniquely determined. The exceptional points are exactly the places where $\partial F/\partial y = 0$. Thus, in general, we want some condition like $\partial F/\partial y \neq 0$ to guarantee that, locally at least, we can find a unique differentiable f such that $F(x, f(x)) = 0$. The exact condition is given by the following theorem

Theorem 2.45 (Implicit Function Theorem). Let $F : U \subseteq \mathbb{R}^2 \to \mathbb{R}$ be a function of class C^p defined on an open subset U of $\mathbb{R}^2$. Suppose $(x_0, y_0) \subseteq U$ and $F(x_0, y_0) = 0$. If $\partial F/\partial y \neq 0$, then there is a neighborhood $U \subseteq \mathbb{R}$ of x_0 and a neighborhood V of y_0 and a unique function $f : U \to V$ of class C^p such that

$$F(x, f(x)) = 0$$

for all $x \in U$.

Proof. Omitted. □

If $F(x, y) = 0$ defines y implicitly as a function of x, say $y = f(x)$, then by the chain rule

$$\begin{aligned} 0 &= \frac{dF}{dx} \\ &= \frac{\partial F}{\partial x}\frac{dx}{dx} + \frac{\partial F}{\partial y}\frac{dy}{dx} \\ &= \frac{\partial F}{\partial x} + \frac{\partial F}{\partial y}\frac{dy}{dx} \end{aligned}$$

and therefore

$$\frac{dy}{dx} = -\frac{\frac{\partial F}{\partial x}}{\frac{\partial F}{\partial y}}.$$

Example 2.46. Suppose $F(x, y) = x^2 - xy + y^2 - 7$. Then $\partial F/\partial y = -x + 2y$. By the Implicit Function Theorem, for all $y \neq 2x$, F defines y implicitly as a function of x and

$$\frac{dy}{dx} = -\frac{\frac{\partial F}{\partial x}}{\frac{\partial F}{\partial y}} = -\frac{2x - y}{-x + 2y} = \frac{y - 2x}{2y - x}.$$

2.6.4 The Gradient

Since $df_{\mathbf{x}} : \mathbb{R}^n \to \mathbb{R}$ is a linear map, it can be represented by an $1 \times n$ matrix i.e., a row vector. [Refer to my linear algebra notes.] Note that $df_{\mathbf{x}}(e_i) = \frac{\partial f}{\partial x_i}(\mathbf{x})$ is the ith column of the matrix representing $df_{\mathbf{x}}$ with respect to the standard bases. Thus, the standard matrix of $df_{\mathbf{x}}$ is

$$[df_{\mathbf{x}}] = \begin{bmatrix} \frac{\partial f}{\partial x_1}(\mathbf{x}) & \frac{\partial f}{\partial x_2}(\mathbf{x}) & \cdots & \frac{\partial f}{\partial x_n}(\mathbf{x}) \end{bmatrix}.$$

Definition 2.47 (Gradient). The vector $\begin{bmatrix} \frac{\partial f}{\partial x_1}(\mathbf{x}) & \frac{\partial f}{\partial x_2}(\mathbf{x}) & \cdots & \frac{\partial f}{\partial x_n}(\mathbf{x}) \end{bmatrix}$ is called the *gradient of f at* $\mathbf{x}$, which we denote by $\nabla f(\mathbf{x})$. [17]

[17] The symbol "∇" is called the *nabla* symbol. The name, by reason of the symbol's shape, from the Hellenistic Greek work *νάβλα* for a Phoenician harp.

Given a vector $\mathbf{v} \in \mathbb{R}^n$ with coordinate vector $[\mathbf{v}] = \begin{bmatrix} v_1 & v_2 & \cdots & v_n \end{bmatrix}$, we therefore have

$$\begin{aligned}
[df_{\mathbf{x}}(\mathbf{v})] &= [df_{\mathbf{x}}][\mathbf{v}] \\
&= \begin{bmatrix} \frac{\partial f}{\partial x_1}(\mathbf{x}) & \frac{\partial f}{\partial x_2}(\mathbf{x}) & \cdots & \frac{\partial f}{\partial x_n}(\mathbf{x}) \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{bmatrix} \\
&= v_1 \frac{\partial f}{\partial x_1}(\mathbf{x}) + v_2 \frac{\partial f}{\partial x_2}(\mathbf{x}) + \cdots + v_n \frac{\partial f}{\partial x_n}(\mathbf{x}) \\
&= \nabla f(\mathbf{x}) \cdot \mathbf{v}.
\end{aligned}$$

Since $df_{\mathbf{x}}(\mathbf{v}) = D_{\mathbf{v}}(\mathbf{x})$, we therefore have

$$D_{\mathbf{v}}(\mathbf{x}) = \nabla f(\mathbf{x}) \cdot \mathbf{v}. \tag{2.7}$$

Exercise 2.21. Rederive Equation (2.7) by computing

$$D_{\mathbf{v}}(\mathbf{x}) = \frac{d}{dt} f(\mathbf{x} + t\mathbf{v})|_{t=0}$$

using the chain rule.

Example 2.48. We now show how to compute $D_{\mathbf{v}} f(1,2)$ if $f(x,y) = x^3 - 3xy + 4y^2$ and $\hat{\mathbf{v}} = (\frac{\sqrt{3}}{2}, \frac{1}{2})$. The gradient of f is

$$\nabla f(x,y) = \begin{bmatrix} 3x^2 - 3y & -3x + 8y \end{bmatrix}$$

so $\nabla f(1,2) = (-3, 13)$ and therefore

$$\begin{aligned}
D_{\mathbf{v}} f(1,2) &= \nabla f(1,2) \cdot \mathbf{v} \\
&= (-3,13) \cdot (\frac{\sqrt{3}}{2}, \frac{1}{2}) \\
&= -\frac{3\sqrt{3}}{2} + \frac{13}{2} \\
&= \frac{13 - 3\sqrt{3}}{2}.
\end{aligned}$$

Exercise 2.22. Let $f(x,y) = x^2y^3 - 4y$. Find the directional derivative at $(2,-1)$ in the direction of $\mathbf{v} = (2,5)$. Make sure to normalize $\mathbf{v}$ first.

Exercise 2.23. Let $f(x,y,z) = x\sin(yz)$. Find the directional derivative at $(1,3,0)$ in the direction of $\mathbf{v} = (1,2,-1)$

Theorem 2.49 (Algebraic Properties of the Gradient). The gradient of a differentiable function has the following properties.

1. $\nabla(cf) = c\nabla f$, where $c \in \mathbb{R}$
2. $\nabla(f+g) = \nabla f + \nabla g$

3. $\nabla(f \cdot g) = (\nabla f)g + g\nabla f$
4. $\nabla\left(\frac{f}{g}\right) = \frac{g\nabla f - f\nabla g}{g^2}$

Proof. The proof follows immediately from Theorem 2.41. □

Example 2.50. If $f(x,y,z) = x - y$ and $g(x,y,z) = z$, then

$$\begin{aligned}\nabla\left(\frac{f}{g}\right) &= \frac{g\nabla f - f\nabla g}{g^2} \\ &= \frac{z(1,-1,0) - (x-y)(0,0,1)}{z^2} \\ &= \frac{1}{z^2}(z,-z,-x+y) \\ &(\frac{1}{z},-\frac{1}{z},\frac{-x+y}{z^2}).\end{aligned}$$

3 Finding Maxima and Minima

[Infinitesimal to local to global. Example: if df_p zero at every point of U and U connected then f is constant on U]

3.1 Review: Single-variable Functions

Let $f : I \subset \mathbb{R} \to \mathbb{R}$, $x_0 \in U$, and (a,b) an *open* interval containing x_0. Then $f(x_0)$ is called a

- *local maximum* of f on (a,b) if $f(x_0) \geq f(x)$ for all $x \in (a,b)$
- *local minimum* of f on (a,b) if $f(x_0) \leq f(x)$ for all $x \in (a,b)$

If $(a,b) = U$, then f is the *absolute maximum* or *absolute minimum*, respectively. Note that every absolute max or min is, in particular, a local max or min, respectively.

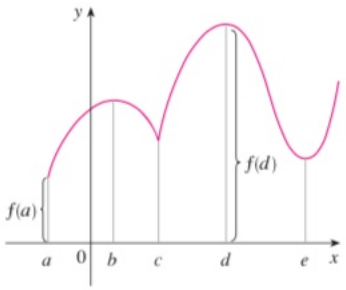


Figure 47: Function with various local maxima and minima.

Example 3.1. The function $f(x) = \cos(x)$ has a local maximum when $x = 2\pi n$, $n \in \mathbb{Z}$ and a local minimum when $x = 2\pi(n + \frac{1}{2})$, $n \in \mathbb{Z}$. Each of these points is also an absolute max and an absolute min, respectively.

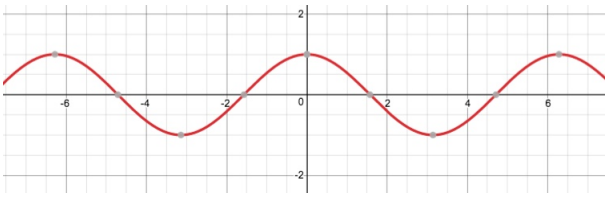


Figure 48: The graph of $f(x) = \cos x$.

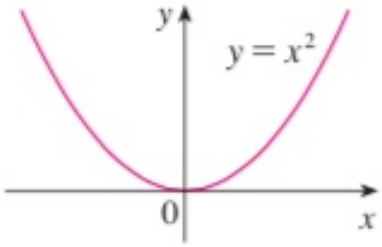


Figure 49: The graph of $f(x) = x^2$.

Example 3.2. The function $f(x) = x^2$ has a local minimum at $x = 0$, which is a global minimum. It has no local maxima.

Example 3.3. The function $f(x) = x^3$ has no local maxima or minima.

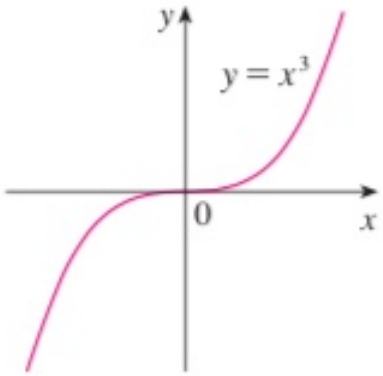


Figure 50: The graph of $f(x) = x^3$.

We have just seen that some functions have extreme values while others do not. When will a function have extreme values?

Theorem 3.4 (Extreme Value Theorem). If f is continuous on a *closed* interval, then f attains an absolute maximum $f(x_1)$ and an absolute minimum $f(x_2)$ for some $x_1, x_2 \in [a, b]$.

Proof. The proof rests on several results beyond the scope of the course, so we omit it. □

We illustrate the Extreme Value Theorem with the following examples

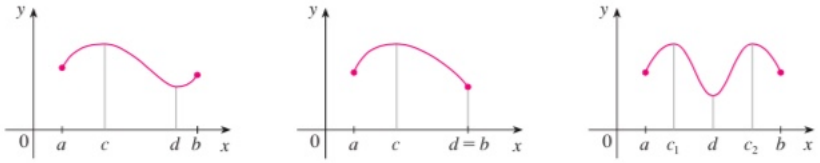


Figure 51: Examples of the Extreme Value Theorem

and non-examples

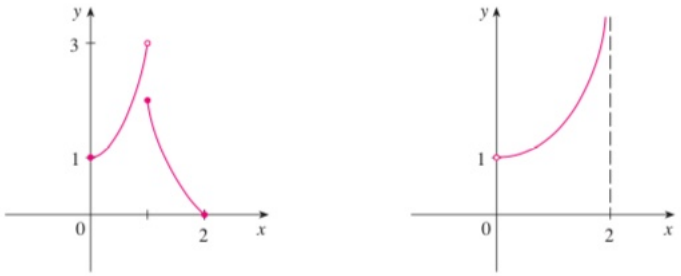


Figure 52: The graph on the left is not continuous at every point on $[0,2]$. The graph on the right is defined on $[0,2)$, which is not closed.

The Extreme Value Theorem tells us that the extreme values of f exist, but not how to find them.

Let's start by looking for local maxima and minima. The following theorem gives a necessary condition for a function to have a local max or min at $x_0 \in (a,b)$.

Theorem 3.5 (Fermat's Theorem). Suppose f is differentiable on an interval (a,b). If f has a local max or min at a point $x_0 \in (a,b)$, then $f'(x_0) = 0$.

Proof. Suppose f has a local max at x_0. Then $f(x_0) \geq f(x)$ for all $x \in (a,b)$. Then for any number h sufficiently close to zero, we have

$$f(x_0) \geq f(x_0 + h)$$

and therefore

$$f(x_0 + h) - f(x_0) \leq 0. \tag{3.1}$$

If $h > 0$, dividing both sides by h gives

$$\frac{f(x_0 + h) - f(x_0)}{h} \leq 0.$$

Taking the right-hand limit of both sides of this inequality [Justify this.], we get

$$\lim_{h\to 0^+} \frac{f(x_0+h)-f(x_0)}{h} \leq 0 \leq \lim_{h\to 0^+} 0 = 0.$$

Since $f'(x_0)$ exists, we have

$$f'(x_0) = \lim_{h\to 0} \frac{f(x_0+h)-f(x_0)}{h} = \lim_{h\to 0^+} \frac{f(x_0+h)-f(x_0)}{h},$$

so $f'(x_0) \leq 0$.

If $h < 0$, the the direction of the inequality in (3.1) is reversed when we divide by h:

$$\frac{f(x_0+h)-f(x_0)}{h} \geq 0 \quad (h<0).$$

Taking the left-hand limit [Justify this.], we have

$$f'(x_0) = \lim_{h\to 0} \frac{f(x_0+h)-f(x_0)}{h} = \lim_{h\to 0^-} \frac{f(x_0+h)-f(x_0)}{h} \geq 0.$$

Since $f'(x_0) \leq 0$ and $f'(x_0) \geq 0$, we must have $f'(x_0) = 0$. The proof in the case of a local minimum is similar. □

While this condition is necessary, it is not sufficient.

Example 3.6. If $f(x) = x^3$ then $f'(x) = 3x^2$ and therefore $f'(0) = 0$, but f has no local max or min at 0 (since $x^3 > 0$ for $x > 0$ and $x^3 < 0$ for $x < 0$).

Theorem 3.5 also assumed that f was differentiable on (a, b). Points on (a, b) where f is not differentiable may also be a local max or min.

Example 3.7. The function $f(x) = |x|$ has a local min at $x = 0$. Since f is not differentiable at $x = 0$, we cannot use Theorem 3.5 to find this point.

Definition 3.8 (Critical point). A *critical point* of a function f is a number x_0 in the domain of f such that either $f'(x_0) = 0$ or $f'(x_0)$ does not exist.

Example 3.9. Let $f(x) = x^{3/5}(4-x)$. Since

$$\begin{aligned} f'(x) &= \frac{3}{5}x^{-2/5}(4-x) + x^{3/5}(-1) \\ &= \frac{12-8x}{5x^{2/5}} \end{aligned}$$

we see that $x = 0$ is a critical point (since $f'(0)$ does not exist) and $x = \frac{3}{2}$ is a critical point (since $f'(3/2) = 0$). These are the only critical points of $f(x)$.

A Geometric Justification of the Arc Length Formula

Proposition A.1. The arc length formula

$$s = \int_a^b ||\mathbf{r}'(t)||dt. \tag{A.1}$$

can be obtained as the limit of a polygonal approximation of the curve.

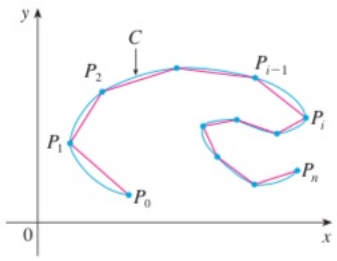


Figure 53: A polygonal approximation of a regular curve.

Proof. [This proof references an idea here and there which are most likely unfamiliar. Attempts to give a rough explanation are given, but may not be entirely satisfactory.] Let $\mathbf{r} : I \to \mathbb{R}^n$ be a parametrized curve and let $[a, b] \subseteq I$. Let $a \equiv t_0 < t_1 < t_2 < \cdots < t_n \equiv b$ be a partition of $[a, b]$ where, for simplicity, we take each interval $t_i - t_{i-1}$ to have equal width $\Delta t \equiv \frac{b-a}{n}$. Approximate the length of curve between $\mathbf{r}(t_{i-1})$ and $\mathbf{r}(t_i)$ by a straight line. We then define the arc length s between $\mathbf{r}(a)$ and $\mathbf{r}(b)$ by

$$\begin{aligned} s &\equiv \lim_{n\to\infty} \sum_{i=1}^{n} ||\mathbf{r}(t_i) - \mathbf{r}(t_{i-1})|| \\ &=\equiv \lim_{n\to\infty} \sum_{i=1}^{n} \sqrt{(x(t_i) - x(t_{i-1}))^2 + (y(t_i) - y(t_{i-1}))^2}. \end{aligned}$$

Since the coordinate functions $x(t)$ and $y(t)$ are continuous on $[a, b]$ and differentiable on (a, b), by the Mean Value Theorem there exist t_i^* and t_i^{**} in (a, b) such that

$$\begin{aligned} x(t_i) - x(t_{i-1}) &= x'(t_i^*)(t_i - t_{i-1}) \\ &= x'(t_i^*)\Delta t \\ y(t_i) - y(t_{i-1}) &= y'(t_i^{**})(t_i - t_{i-1}) \\ &= y'(t_i^{**})\Delta t \end{aligned}$$

and therefore

$$s = \lim_{n\to\infty} \sum_{i=1}^{n} \sqrt{(x'(t_i^*))^2 + (y'(t_i^{**}))^2}\Delta t.$$

If t_i^* were equal to t_i^{**} for all $i = 1, \cdots, n$, then this would be a Riemann sum for $\int_a^b ||\mathbf{r}'(t)||dt$. However, in general they are not equal, so we will instead show that the difference between this expression and the Riemann sum

$$\int_a^b ||\mathbf{r}'(t)||dt \equiv \lim_{n\to\infty} \sum_{i=1}^{n} \sqrt{(x'(t_i^*))^2 + (y'(t_i^*))^2}\Delta t$$

goes to zero as $n \to \infty$. [18] First, note that since $\mathbf{r}$ is, in particular, continuously differentiable on each closed interval $[t_{i-1}, t_i]$, the functions x' and y' are uniformly continuous on $[t_{i-1}, t_i]$. [19] Thus,

[18] Intuitively, this should make sense, since t_i^* and t_i^{**} both lie within $[t_{i-1}, t_i]$ and by shrinking Δt, each of these subintervals also shrinks and therefore we should be able to make the points t_i^* and t_i^{**} become arbitrarily close to each other by making Δt arbitrarily small.

[19] The proof of this involves the notion of "compactness", the explanation of which would take us beyond the scope of this course. A function is said to be *uniformly continuous* if for every $\epsilon > 0$ there exists $\delta > 0$

given any $\epsilon > 0$ there exists a $\delta > 0$ such that $|t_i^* - t_i^{**}| < \delta$ implies $|y'(t_i^*) - y'(t_i^{**})| < \epsilon$. Since $|t_i^* - t_i^{**}| < t_i - t_{i-1} = \Delta t$, if we take $\Delta t < \delta$, then we will have $|y'(t_i^*) - y'(t_i^{**})| < \epsilon$. Note that this is possible since, given any $\delta > 0$, by the Archimedean property of the real numbers we can find a positive integer n such that $\Delta t = \frac{b-a}{n} < \delta$. Assuming now that we have chosen $\Delta t < \delta$, we then have

$$\begin{aligned}
&\left|\sqrt{(x'(t_i^*))^2 + (y'(t_i^*))^2} - \sqrt{(x'(t_i^*))^2 + (y'(t_i^{**}))^2}\right| \\
&= \left|\sqrt{(x'(t_i^*))^2 + (y'(t_i^*))^2} - \sqrt{(x'(t_i^*))^2 + (y'(t_i^{**}))^2 + (y'(t_i^*))^2 - (y'(t_i^*))^2}\right| \\
&= \left|\sqrt{(x'(t_i^*))^2 + (y'(t_i^*))^2} - \sqrt{(x'(t_i^*))^2 + (y'(t_i^*))^2 + (y'(t_i^{**}))^2 - (y'(t_i^*))^2}\right|.
\end{aligned}$$

Now $(y'(t_i^{**}))^2 - (y'(t_i^*))^2 \leq |(y'(t_i^{**}))^2 - (y'(t_i^*))^2|$, so, by the triangle inequality,

$$\begin{aligned}
\sqrt{(x'(t_i^*))^2 + (y'(t_i^*))^2 + (y'(t_i^{**}))^2 - (y'(t_i^*))^2} &\leq \sqrt{(x'(t_i^*))^2 + (y'(t_i^*))^2 + |(y'(t_i^{**}))^2 - (y'(t_i^*))^2|} \\
&\leq \sqrt{(x'(t_i^*))^2 + (y'(t_i^*))^2} + \sqrt{|(y'(t_i^{**}))^2 - (y'(t_i^*))^2|}
\end{aligned}$$

and therefore

$$\begin{aligned}
&\left|\sqrt{(x'(t_i^*))^2 + (y'(t_i^*))^2} - \sqrt{(x'(t_i^*))^2 + (y'(t_i^{**}))^2}\right| \\
&= \left|\sqrt{(x'(t_i^*))^2 + (y'(t_i^*))^2} - \sqrt{(x'(t_i^*))^2 + (y'(t_i^*))^2 + (y'(t_i^{**}))^2 - (y'(t_i^*))^2}\right| \\
&\leq \left|\sqrt{(x'(t_i^*))^2 + (y'(t_i^*))^2} - \left(\sqrt{(x'(t_i^*))^2 + (y'(t_i^*))^2} + \sqrt{|(y'(t_i^{**}))^2 - (y'(t_i^*))^2|}\right)\right| \\
&= \left|-\sqrt{|(y'(t_i^{**}))^2 - (y'(t_i^*))^2|}\right| \\
&= \sqrt{|(y'(t_i^{**}))^2 - (y'(t_i^*))^2|} \\
&= \sqrt{|(y'(t_i^{**}) + (y'(t_i^*))(y'(t_i^{**}) - (y'(t_i^*))|} \\
&= \sqrt{|(y'(t_i^{**}) + (y'(t_i^*))|}\sqrt{|(y'(t_i^{**}) - (y'(t_i^*))|}
\end{aligned}$$

Since y' is continuous and $[a, b]$ a closed interval, $y'([a, b])$ is bounded by some $M > 0$, so [20]

$$\begin{aligned}
&= \sqrt{|(y'(t_i^{**}) + (y'(t_i^*))|}\sqrt{|(y'(t_i^{**}) - (y'(t_i^*))|} \\
&\leq \sqrt{2M}\sqrt{|(y'(t_i^{**}) - (y'(t_i^*))|},
\end{aligned}$$

such that for any point $t \in [t_{i-1}, t_i]$, $|t' - t| < \delta$ implies $||f(t) - f(t')|| < \epsilon$; that is, the *same* δ works for every point $t \in [t_{i-1}, t_i]$. This is not true when a function is continuous but not uniformly continuous, since in that case δ depends on both ϵ *and* t.

[20] Here we have used the facts that the image of a compact set under a continuous mapping is compact, and that a compact set is bounded.

so by choosing $\Delta t < \delta$ such that $|(y'(t_i^{**}) - (y'(t_i^*))| < \frac{\epsilon^2}{4M^2(b-a)^2}$, then

$$\begin{aligned}
&\left|\sqrt{(x'(t_i^*))^2 + (y'(t_i^*))^2} - \sqrt{(x'(t_i^*))^2 + (y'(t_i^{**}))^2}\right| \\
&\leq \sqrt{2M}\sqrt{|(y'(t_i^{**}) - (y'(t_i^*))|} \\
&< \sqrt{2M}\sqrt{\frac{\epsilon^2}{4M^2(b-a)^2}} \\
&= \frac{\epsilon}{b-a}
\end{aligned}$$

and therefore

$$\begin{aligned}
&\left|\sum_{i=1}^n \sqrt{(x'(t_i^*))^2 + (y'(t_i^*))^2}\Delta t - \sum_{i=1}^n \sqrt{(x'(t_i^*))^2 + (y'(t_i^{**}))^2}\Delta t\right| \\
&= \left|\sum_{i=1}^n \left(\sqrt{(x'(t_i^*))^2 + (y'(t_i^*))^2} - \sqrt{(x'(t_i^*))^2 + (y'(t_i^{**}))^2}\right)\right| \Delta t \\
&\leq \sum_{i=1}^n \left|\sqrt{(x'(t_i^*))^2 + (y'(t_i^*))^2} - \sqrt{(x'(t_i^*))^2 + (y'(t_i^{**}))^2}\right| \Delta t \\
&< n\frac{\epsilon}{b-a}\Delta t \\
&= n\frac{\epsilon}{b-a}\frac{(b-a)}{n} \\
&= \epsilon.
\end{aligned}$$

□

B Topology of $\mathbb{R}^n$

In this section we begin our study of those basic properties of $\mathbb{R}^n$ which allow us to define the notion of a continuous function. We will study open sets, which generalize open intervals on $\mathbb{R}$, and closed sets, which generalize closed intervals.

Most of what follows depends only on the basic properties of the distance function

$$||\mathbf{x} - \mathbf{y}|| = \sqrt{\sum_{i=1}^n (x_i - y_i)^2}, \quad \mathbf{x}, \mathbf{y} \in \mathbb{R}^n.$$

Theorem B.1 (Properties of the distance function).

(i) $||\mathbf{x} - \mathbf{y}|| \geq 0$, with equality if and only if $\mathbf{x} = \mathbf{y}$;

(ii) $||\mathbf{x} - \mathbf{y}|| = ||\mathbf{y} - \mathbf{x}||$;

(iii) $||\mathbf{x} - \mathbf{y}|| \leq ||\mathbf{x} - \mathbf{z}|| + ||\mathbf{z} - \mathbf{y}||$ (triangle inequality).

Proof. Properties (i) and (ii) are obvious from the definition. We prove (iii):
Let $\mathbf{u}, \mathbf{v} \in \mathbb{R}^n$. Then

$$\begin{aligned}
||\mathbf{u}+\mathbf{v}||^2 &= (\mathbf{u}+\mathbf{v}) \cdot (\mathbf{u}+\mathbf{v}) \\
&= ||\mathbf{u}||^2 + ||\mathbf{v}||^2 + 2\mathbf{u} \cdot \mathbf{v} \\
&\leq ||\mathbf{u}||^2 + ||\mathbf{v}||^2 + 2|\mathbf{u} \cdot \mathbf{v}| \\
&\leq ||\mathbf{u}||^2 + ||\mathbf{v}||^2 + 2||\mathbf{u}|| \cdot ||\mathbf{v}|| \text{ (by the Cauchy-Schwarz inequality)} \\
&= (||\mathbf{u}|| + ||\mathbf{v}||)^2.
\end{aligned}$$

Since $||\mathbf{u}+\mathbf{v}||$ and $||\mathbf{u}|| + ||\mathbf{v}||$ are nonnegative, it follows that

$$||\mathbf{u}+\mathbf{v}|| \leq ||\mathbf{u}|| + ||\mathbf{v}||.$$

The proof of the theorem then follows by taking $\mathbf{u} = \mathbf{x} - \mathbf{z}$ and $\mathbf{v} = \mathbf{z} - \mathbf{y}$. □

B.1 Open Sets

In order to define open sets, we first introduce the notion of an open ball.

Definition B.2. Let $\mathbf{x}_0 \in \mathbb{R}^n$ and let $r > 0$. The *open ball of radius r centered at $\mathbf{x}_0$ is the set*

$$B_r(\mathbf{x}_0) = \{\mathbf{x} \in \mathbb{R}^n : ||\mathbf{x} - \mathbf{x}_0|| < r\}.$$

Example B.3.

(a) In $\mathbb{R}$, the open ball $B_r(x_0)$ is the open interval $(x_0 - r, x_0 + r)$.

(b) In $\mathbb{R}^2$, the open ball $B_r(\mathbf{x}_0)$ is the open disc

$$B_r(\mathbf{x}_0) = \{(x, y) \in \mathbb{R}^2 : \sqrt{(x - x_0)^2 + (y - y_0)^2} < r\}.$$

Definition B.4 (Open Set). A subset $A \subseteq \mathbb{R}^n$ is open if, given any point $\mathbf{x} \in A$, there exists an open ball centered at $\mathbf{x}$ contained in A.

That is, an open set A is one for which if you are originally located at any point $\mathbf{x} \in A$, then you can move a sufficiently small distance and still remain in A.

Definition B.5 (Neighborhood). If A is an open set containing a point $\mathbf{x}$, we say that A is a *neighborhood* of $\mathbf{x}$.

It is immediate from Definition B.4 that $\mathbb{R}^n$ itself and $\emptyset$ are open sets. Here is a nontrivial example:

Proposition B.6. An open ball is an open set.

Proof. Let $\mathbf{x} \in B_r(\mathbf{x}_0)$. Then $h = r - d(\mathbf{x}, \mathbf{x}_0) > 0$. Let $\mathbf{y} \in B_h(\mathbf{x})$. Then $d(\mathbf{y}, \mathbf{x}) < h$. It then follows from the triangle inequality that

$$\begin{aligned}
d(\mathbf{x}_0, \mathbf{y}) &\leq d(\mathbf{x}_0, \mathbf{x}) + d(\mathbf{x}, \mathbf{y}) \\
&< r - h + h = r,
\end{aligned}$$

which shows that $\mathbf{y} \in B_r(\mathbf{x}_0)$, and therefore $B_h(\mathbf{x}) \subseteq B_r(\mathbf{x}_0)$. Thus, $B_r(\mathbf{x}_0)$ is open. □

Note that whether or not a set is open depends not only on the set itself, but on which $\mathbb{R}^n$ we are considering it to be a subset of. For instance, the open interval $(0,1)$ is open in $\mathbb{R}$, but it is not open in $\mathbb{R}^2$, since every open ball $B_r(\mathbf{x})$ with $\mathbf{x} \in \{(x,0) : 0 < x < 1\}$ contains points $(x,y) \in \mathbb{R}^2$ with $y \neq 0$.

Theorem B.7 (Properties of Open Sets).

(i) The intersection of a finite collection of open sets is open.

(ii) The union of an arbitrary collection of open sets is open.

Proof. (i) Let $\{U_i\}_{i=1}^n$ be a finite collection of open sets and let $\mathbf{x} \in U \equiv \cap_{i=1}^n U_i$. Then $\mathbf{x} \in U_i$ for all $i = 1, \ldots, n$. Since each U_i is open, there exists $r_i > 0$ such that $\mathbf{x} \in B_{r_i}(\mathbf{x}) \subseteq U_i$. By taking $r = \min\{r_1, \ldots, r_n\}$, $\mathbf{x} \in B_r(\mathbf{x}) \subseteq U_i$ for all $i = 1, \ldots, n$, and therefore $\mathbf{x} \in B_r(\mathbf{x}) \subseteq U$.

(ii) Let $\{U_\alpha\}_{\alpha \in \mathscr{A}}$ be an arbitrary collection of open sets. Let $\mathbf{x} \in \cup_{\alpha \in \mathscr{A}} U_\alpha$. Then $\mathbf{x} \in U_\alpha$ for some $\alpha \in \mathscr{A}$. Since U_α is open, there exists $r > 0$ such that $\mathbf{x} \in B_r(\mathbf{x}) \subseteq U_\alpha$ and therefore $\mathbf{x} \in B_r(\mathbf{x}) \subseteq U$.

□

Note that it is *not* true that the intersection of an arbitrary collection of open sets is necessarily open. For example, let $x_0 \in \mathbb{R}$ and consider the collection of open balls $B_r(x_0) = (x_0 - r, x_0 + r)$. Since $x_0 \in B_r(x_0)$ for all $r > 0$, $\cap_{r>0} B_r(x_0) \neq \emptyset$. Suppose there is a point y different from x in $\cap_{r>0} B_r(x_0)$. Let $d = |y - x|$. Then for $0 < r < d$, y is not in $B_r(x_0)$ and therefore not in $\cap_{r>0} B_r(x_0)$, which is a contradiction. Hence, $\cap_{r>0} B_r(x_0) = \{x_0\}$, which is not open, since for every $r > 0$, $B_r(x_0)$ contains points of $\mathbb{R}$ other than x_0.

Definition B.8 (Topological Space). Let X be a set and let $\mathscr{T}$ be a collection of subsets of X (called *open sets*) with the following properties:

(i) X is open.

(ii) $\emptyset$ is open.

(iii) The intersection of finitely many open sets is open.

(iv) The union of an arbitrary collection of open sets is open.

Then $\mathscr{T}$ is called a *topology* on X and the ordered pair $(X, \mathscr{T})$ is called a *topological space*.

Thus, $\mathbb{R}^n$ together with the collection of open subsets of $\mathbb{R}^n$ as defined in Definition B.4 is a topological space. We will not consider other topological spaces in these notes, but many of the results we discuss will hold in any topological space.

Definition B.9 (Interior of a Set). Let A be any subset of $\mathbb{R}^n$. A point $\mathbf{x} \in A$ is an *interior point* of A if there is an open ball centered $\mathbf{x}$ contained in A; that is, if there exists $r > 0$ such that $\mathbf{x} \in B_r(\mathbf{x}) \subseteq A$. The *interior* of A is the set of all interior points of A, denoted $\text{Int}(A)$.

Note that $\text{Int}(A)$ might be empty. For example, if $A = \{x_0\}$ contains a single point, then $\text{Int}(A)$ is empty.

Proposition B.10. The interior of A is the largest open set contained in A, in the sense of that if U is any other open set contained in A, then $U \subseteq \text{Int}(A)$. Since the intersection of all open sets contained in A has this property, $\text{Int}(A)$ is the union of all open subsets contained in A.

Proof. Let $A \subseteq \mathbb{R}^n$, let $\{U_\alpha\}_{\alpha\in\mathscr{A}}$ be the collection of open subsets of $\mathbb{R}^n$ contained in A, and let $U \equiv \cup_{\alpha\in\mathscr{A}} U_\alpha$. If $\mathbf{x} \in U$, then $\mathbf{x} \in U_\alpha$ for some α. Since U_α is open, there exists $r > 0$ such that $\mathbf{x} \in B_r(\mathbf{x}) \subseteq U_\alpha \subseteq A$, which shows that $\mathbf{x} \in \text{Int}(A)$. This shows that $U \subseteq \text{Int}(A)$. Conversely, suppose that $\mathbf{x} \in \text{Int}(A)$. Then there exists $r > 0$ such that $\mathbf{x} \in B_r(\mathbf{x}) \subseteq A$. By Proposition B.6, $B_r(\mathbf{x})$ is open subset of $\mathbb{R}^n$. Since $B_r(\mathbf{x})$ is an open subset of $\mathbb{R}^n$ contained in A, it is contained in the union of all such subsets. Hence, $\mathbf{x} \in U$. This shows that $\text{Int}(A) \subseteq U$ and therefore $U = \text{Int}(A)$. □

Proposition B.11. A subset $A \subseteq \mathbb{R}^n$ is open if and only if $\text{Int}(A) = A$.

Proof. Definition B.4 says precisely that A is open if and only if every point is an interior point, that is, if and only if $\text{Int}(A) = A$. □

B.2 Closed Sets

Definition B.12 (Closed set). A set B in $\mathbb{R}^n$ is *closed* if its complement $(\mathbb{R}^n - B)$ is open.

Example B.13. A set $\{\mathbf{x}_0\} \subseteq \mathbb{R}^n$ containing a single point is closed. To see that its complement is open, let $\mathbf{x}$ be any point in $\mathbb{R}^n - \{\mathbf{x}_0\}$, and take $r = ||\mathbf{x} - \mathbf{x}_0||$. Then the open ball $B_r(\mathbf{x})$ is contained in $\mathbb{R}^n - \{\mathbf{x}_0\}$. Thus, $\mathbb{R}^n - \{\mathbf{x}_0\}$, so $\{\mathbf{x}_0\}$ is closed.

It is important to note from Definitions B.4 and B.12 that a set can be *neither* open nor closed or *both* open and closed.

Example B.14. (a) The half open interval $(0, 1] \subseteq \mathbb{R}$ is neither open nor closed in $\mathbb{R}$. It is not open because every open ball centered at 1 contains points > 1. Its complement is $\mathbb{R} - (0, 1] = (-\infty, 0] \cup (1, \infty)$ is not open, since every open ball centered at 0 contains points < 0, so $(0, 1]$ is not closed.

(b) Since $\mathbb{R}^n$ is open $\emptyset = \mathbb{R}^n - \mathbb{R}^n$ is closed. Similarly, since $\emptyset$ is open, $\mathbb{R}^n - \emptyset = \mathbb{R}^n$ is closed. Hence, $\mathbb{R}^n$ and $\emptyset$ are both open and closed. [21]

Recall the following set theoretic identities:

If $\{U_\alpha\}_{\alpha\in\mathscr{A}}$ is a collection of subsets of a set X, then

$$X - \cup_{\alpha\in\mathscr{A}} U_\alpha = \cap_{\alpha\in\mathscr{A}} (X - U_\alpha)$$
$$X - \cap_{\alpha\in\mathscr{A}} U_\alpha = \cup_{\alpha\in\mathscr{A}} (X - U_\alpha)$$

The following properties of closed sets follow from these identities and from the corresponding properties of open sets.

Theorem B.15 (Properties of closed sets).

(i) The union of finitely many closed sets is closed.

(ii) The intersection of an arbitrary collection of closed sets is closed.

[21] Such sets are sometimes referred to as *clopen*.

Proof. (i) Let $\{U_i\}_{i=1}^n$ be a finite collection of closed subsets of $\mathbb{R}^n$. Then $\mathbb{R}^n - U_i$ is open for each $i = 1, \ldots, n$ and therefore by part (ii) of Theorem B.7

$$\cup_{i=1}^n (\mathbb{R}^n - U_i) = \mathbb{R}^n - \cap_{i=1}^n U_i$$

is open, and therefore $\cap_{i=1}^n U_i$ is closed.

(ii) Let $\{U_\alpha\}_{\alpha \in \mathscr{A}}$ be an arbitrary collection of closed sets. Then $\mathbb{R}^n - U_\alpha$ is open for each $\alpha \in \mathscr{A}$, and by part (i) of Theorem B.7,

$$\cup_{\alpha \in \mathscr{A}} (\mathbb{R}^n - U_\alpha) = \mathbb{R}^n - \cap_{\alpha \in \mathscr{A}} U_\alpha$$

is open, and therefore $\cap_{\alpha \in \mathscr{A}} U_\alpha$ is closed.

□

Example B.16. Let $\mathbf{x}_0 \in \mathbb{R}^n$ and $r > 0$. Define the *closed ball of radius r centered at* $\mathbf{x}_0$ to be the set

$$\overline{B}_r(\mathbf{x}_0) \equiv \{\mathbf{x} \in \mathbb{R} : ||\mathbf{x} - \mathbf{x}_0|| < r\}.$$

To see that this set is closed, let $\mathbf{x} \in \mathbb{R}^n - \overline{B}_r(\mathbf{x}_0)$. Letting $\epsilon = ||\mathbf{x} - \mathbf{x}_0|| - r$ (note that $\epsilon > 0$), the open ball $B_\epsilon(\mathbf{x})$ is contained in $\mathbb{R}^n - \overline{B}_r(\mathbf{x}_0)$. This shows that $\mathbb{R}^n - \overline{B}_r(\mathbf{x}_0)$ is open and therefore $\overline{B}_r(\mathbf{x}_0)$ is closed.

Example B.17. We have seen that one point sets are closed. It follows from part (i) of Theorem B.15 that any finite subset of $\mathbb{R}^n$ is closed.

B.3 Limit points

There is another very useful way to determine whether or not a set is closed. This depends on the important concept of a limit point.

Definition B.18 (Limit point). Let $A \subseteq \mathbb{R}^n$. A point $\mathbf{x} \in \mathbb{R}^n$ is called a *limit point* of A if every open ball centered at $\mathbf{x}$ contains a point of A other than $\mathbf{x}$.

That is, a limit point $\mathbf{x}$ of a set A is a point such that there are other points of A arbitrarily close to $\mathbf{x}$. [22]

Example B.19. (a) A one point set $\{\mathbf{x}_0\} \subseteq \mathbb{R}^n$ has no limit points. Since $\mathbb{R}^n - \{\mathbf{x}_0\}$ is open, a one-point subset of $\mathbb{R}^n$ has no limit points, because for every point $\mathbf{x} \in \mathbb{R}^n - \{\mathbf{x}_0\}$, there exists some open ball which is contained in $\mathbb{R}^n - \{\mathbf{x}_0\}$, and therefore does not contain $\mathbf{x}_0$. For an open ball centered at $\mathbf{x}_0$ itself, it is not possible for it to contain a point of $\{\mathbf{x}_0\}$ different from $\mathbf{x}_0$ since that is the only point in the set.

(b) In $\mathbb{R}$, the set of limit points of the open interval $(0, 1)$ is given by the closed interval $[0, 1]$, since for any $x \in [0, 1]$ and any $r > 0$, the open ball $B_r(x)$ contains a point of $(0, 1)$ different from x. This shows that a limit point $\mathbf{x}$ of a set A need not lie in A.

Definition B.20 (Isolated point). If $\mathbf{x} \in A$ and $\mathbf{x}$ is not a limit point of A, then $\mathbf{x}$ is called an *isolated point* of A.

[22] Note that the concept of a limit point of a set is distinct from that of a limit of a function.

Example B.21. Consider the set $\{\frac{1}{n}\}_{n\in\mathbb{N}}$. Each point in this set is an isolated point, since by choosing $r = |\frac{1}{n+1} - \frac{1}{n}|$, the open ball $B_r(\frac{1}{n})$ contains only the point $\frac{1}{n}$. A set in which every point is isolated is called a *discrete* set.

The definitions of limit points and closed sets are closely related, as shown by the next theorem.

Theorem B.22 (A closed sets is one that contains all its limit points). Let $A \subseteq \mathbb{R}^n$. Then A is closed if and only if A contains all of its limit points.

Proof. Suppose A is closed and let $\mathbf{x}$ be a limit point of A. Suppose $\mathbf{x} \in \mathbb{R}^n - A$. Since $\mathbb{R}^n - A$ is open, there exists an open ball centered at $\mathbf{x}$ contained in $\mathbb{R}^n - A$. This contradicts the fact that $\mathbf{x}$ is a limit point of A. Thus, $\mathbf{x}$ must be in A.

Now suppose A contains all its limit points. Let $\mathbf{x} \in \mathbb{R}^n - A$. Then $\mathbf{x}$ is not a limit point of A, so there exists some open ball centered at $\mathbf{x}$ which is contained in $\mathbb{R}^n - A$. This shows that $\mathbb{R}^n - A$ is open, and therefore A is closed. □

Note that a set need not have any limit points (e.g., any finite subset of $\mathbb{R}^n$). It is then vacuously true that such a set contains all its limit points, so Theorem B.22 still applies and we conclude that the set is closed.

B.4 Closure of a Set

The interior of a set A is the largest open subset contained in A. Similarly, we can form the smallest closed set containing a set A. What it means for a set U to be the smallest closed set containing A is that if W is any other closed set containing A, then $U \subseteq W$. If we let $\{U_\alpha\}_{\alpha\in\mathscr{A}}$ be the collection of all closed sets containing A, then we see that $\cap_{\alpha\in\mathscr{A}} U_\alpha$ has this property, and therefore is the smallest closed set containing A.

Definition B.23 (Closure). Let $A \subseteq \mathbb{R}^n$. The *closure* of A is defined to be the intersection of all closed sets containing A. We denote the closure of A by $\bar{A}$.

By part (ii) of Theorem B.15, $\bar{A}$ is closed.

Example B.24. Let $A = (0,1) \subseteq \mathbb{R}$. Then $\bar{A} = [0,1]$.

Proposition B.25. A is closed if and only if $A = \bar{A}$.

Proof. Suppose A is closed. Then, since A is a closed subset containing A, the intersection of all such subsets is clearly A. Thus, $A = \bar{A}$.

Conversely, suppose $A = \bar{A}$. By definition, if $\{U_\alpha\}_{\alpha\in\mathscr{A}}$ is the collection of closed subsets containing A, then $\bar{A} = \cap_{\alpha\in\mathscr{A}} U_\alpha$. We have then

$$\begin{aligned} X - A &= X - \cap_{\alpha\in\mathscr{A}} U_\alpha \\ &= \cup_{\alpha\in\mathscr{A}} (X - U_\alpha) \end{aligned}$$

is open is each $X - U_\alpha$ is open and arbitrary unions of open sets are open. Thus A is closed. □

The connected between closure and limit points is given by the following theorem.

Theorem B.26. Let $A \subseteq \mathbb{R}^n$. Let A' denote the set of limit points of A. Then $\bar{A} = A \cup A'$.

Proof. Since a closed set contains all of its limit points, any closed set containing A contains $A \cup A'$. Thus, it suffices to show that $A \cup A'$ is closed, for then $A \cup A'$ will be the smallest closed set containing A. Let $\mathbf{x}$ be a limit point of $A \cup A'$. If $\mathbf{x} \in A$, then $\mathbf{x} \in A \cup A'$, so assume $\mathbf{x} \notin A$. Let $B_r(\mathbf{x})$ be an open ball containing $\mathbf{x}$. Since $\mathbf{x}$ is a limit point of $A \cup A'$, $B_r(\mathbf{x}) \cap (A \cup A')$ contains a point $\mathbf{y}$ of $A \cup A'$ distinct from $\mathbf{x}$. Now either $\mathbf{y} \in A$ or $\mathbf{y} \in A'$. If $\mathbf{y} \in A$, this shows that $\mathbf{x}$ is a limit point of A and we are done. If $\mathbf{y} \in A'$, then, since $B_r(\mathbf{x})$ is a neighborhood of $\mathbf{y}$, $B_r(\mathbf{x})$ contains a point $\mathbf{z}$ of A distinct from $\mathbf{y}$. Since $\mathbf{x} \notin A$, $\mathbf{z} \neq \mathbf{x}$. Thus, $\mathbf{x} \in A'$ and therefore $\mathbf{x} \in A \cup A'$. This shows that $A \cup A'$ contains all of its limit points, and is therefore closed by Theorem B.22. □

We saw previously in Example B.19 (b) that if $A = (0,1)$, then $A' = [0,1]$, in agreement with Example B.24.

B.5 Boundary of a Set

If we consider the unit disc in $\mathbb{R}^2$, we know what we would like to call the boundary - the obvious choice is the unit circle. But, for more complicated sets, such as the rationals, it is not as intuitively clear what the boundary should be. Therefore a precise definition is needed.

Definition B.27 (Boundary of a set). Given $A \subseteq \mathbb{R}^n$, the *boundary* of A, denoted ∂A, is defined to be the set

$$\partial A = \bar{A} \cap \overline{\mathbb{R}^n - A}.$$

By part (ii) of Theorem B.15, ∂A is closed. Also note that $\partial A = \partial(\mathbb{R}^n - A)$. It follows from Theorem B.26, that we can also describe the boundary is the following intuitive way.

Theorem B.28. Let $A \subseteq \mathbb{R}^n$. Then $\mathbf{x} \in \partial A$ if and only if for every $\epsilon > 0$, $D_\epsilon(\mathbf{x})$ contains points of A and $\mathbb{R}^n - A$ (these points might included $\mathbf{x}$ it self).

[Draw a picture.]

Proof. Let $\mathbf{x} \in \partial A = \bar{A} \cap \overline{\mathbb{R}^n - A}$. Now, either $\mathbf{x} \in A$ or $\mathbf{x} \in \mathbb{R}^n - A$. If $\mathbf{x} \in A$, then by Theorem B.26 $\mathbf{x}$ is a limit point of $\mathbb{R}^n - A$, so for every $\epsilon > 0$, the open ball $B_\epsilon(\mathbf{x})$ contains a point of $\mathbb{R}^n - A$. The case $x \in \mathbb{R}^n - A$ and the converse as similar. [Finish.] □

Example B.29. Let $A = \mathbb{Q} \cap [0,1] \subseteq \mathbb{R}$. Then $\partial A = [0,1]$, since for any $\epsilon > 0$ and $x \in A$, $D_\epsilon(x)$ contains both rational and irrational points. [Need to prove rationals dense in $\mathbb{R}$.]

Note that if $\mathbf{x} \in \partial A$, then $\mathbf{x}$ need not be a limit point of A. For example, if $A = \{0\} \subseteq \mathbb{R}$, then A has no limit points, but $\partial A = \{0\}$.

C Every linear map on a finite-dimensional vector space is continuous

C.1 Least Upper Bounds

At some point in your study of calculus, you may have wondered why we use $\mathbb{R}$ to do calculus. Is there something stopping us from, say, doing calculus over $\mathbb{Q}$? There certainly nothing stopping us from defining limits, continuity, derivatives, etc. for functions $f : \mathbb{Q} \to \mathbb{Q}$ exactly as we have done for real functions. [Add example.] However, the reason we are more interested in calculus

over $\mathbb{R}$ rather than $\mathbb{Q}$ is that various important theorems become *false* over $\mathbb{Q}$. For instance, the intermediate value theorem and the extreme value theorem are false over $\mathbb{Q}$. [Add examples.] Since these theorems are used to prove even more important theorems - such as the Mean Value Theorem and the Fundamental Theorem of Calculus, calculus is pretty boring over $\mathbb{Q}$, and is not capable of applications in physics, etc.

The reason why these theorems are false over $\mathbb{Q}$ is that $\mathbb{R}$ has an important property which is not shared by $\mathbb{Q}$, even though $\mathbb{R}$ and $\mathbb{Q}$ have the same *algebraic* properties. Already, $\mathbb{Q}$ is insufficient to solve basic equations; for instance, the equation $x^2 = 2$ has no solution in $\mathbb{Q}$ (since $\sqrt{2}$ is not rational). It turns out there are infinitely many *irrational* numbers, so one may visualize the set of rational numbers as lying along a number line which has certain "gaps". The real number system is constructed to *extend* $\mathbb{Q}$; that is, $\mathbb{R}$ contains (an isomorphic copy of) $\mathbb{Q}$ and has exactly the same algebraic properties of $\mathbb{Q}$, but in such a way the *real* number line has no gaps.

We need to formalize this property, which $\mathbb{Q}$ fails to have, that ensures a set has no gaps when the real numbers are ordered and placed along a number line. Since this property should only depend on the set in question, this seems immediately puzzling: if $\mathbb{Q}$ is all we have, then how are we to state that there is a "gap" at $\sqrt{2}$ without making reference to $\sqrt{2}$?! The following example sheds some light on the situation:

Example C.1. Let $A = \{p \in \mathbb{Q} : p^2 < 2\}$ and $B = \{p \in \mathbb{Q} : p^2 > 2\}$. We will show that A contains no largest element and B contains no smallest element. Explicitly, for every $p \in A$ we can find a $q \in A$ such that $p < q$, and for every $p \in B$ we can find $q \in B$ such that $q < p$.

To do this, associate with each rational number $p > 0$ the number

$$q = p - \frac{p^2 - p}{p + 2} = \frac{2p + 2}{p + 2}. \tag{C.1}$$

Then

$$q^2 - 2 = \frac{2(p^2 - 2)}{(p + 2)^2}. \tag{C.2}$$

If $p \in A$ then $p^2 - 2 < 0$, (C.1) shows that $q > p$, and (C.2) shows that $q^2 < 2$. Thus, $q \in A$.

If $p \in B$ then $p^2 - 2 > 0$, (C.1) shows that $0 < q < p$, and (C.2) shows that $q^2 > 2$. Thus, $q \in B$.

Thus, elements in A get closer and closer to $\sqrt{2}$ without ever reaching it, and elements of B get closer and closer to $\sqrt{2}$ without ever reaching it.

To discuss the issue, we make the following definitions.

Definition C.2 (Upper and lower bounds).

1. A set A of numbers is *bounded above* if there is a real number x (which need not be in A) such that $a \leq x$ for every $a \in A$. The number x is called an *upper bound* for A.

2. A set A of numbers is *bounded below* if there is a real number x (which need not be in A) such that $x \geq a$ for every $a \in A$. The number x is called a *lower bound* for A.

Example C.3. (a) Looking back at Example C.1, we see that every element of B is an upper bound of A, and every element of A is a lower bound of B.

(b) The entire collections $\mathbb{R}$, $\mathbb{Q}$, and $\mathbb{N}$ are examples of sets which are *not* bounded above.

(c) The set $S = \{x \in \mathbb{R} : 0 \leq x < 1\}$ is bounded above.

The set S in part (c) of Example C.3 has many upper bounds; any real $x \geq 1$ will do. However, the upper bound 1 is the *least* upper bound, since if ϵ is any positive number, the interval $(1-\epsilon, 1)$ is nonempty, and thus contains a number smaller than 1, which is therefore not an upper bound of S. We now define this formally.

Definition C.4 (Least upper bound). Let A be a set of numbers which is bounded above. A number x is a *least upper bound* of A if

(1) x is an upper bound of A, and

(2) if y is an upper bound of A, then $x \leq y$.

Thus, unlike the set S of Example C.3(c), the set A of Example C.1 is bounded above, but has no least upper bound, since B has no smallest element.

Proposition C.5. Let A be a set of numbers. If A has a least upper bound, then it is unique.

Proof. Suppose x and y are both least upper bounds of A. Then

$$\begin{aligned} x \leq y, &\quad \text{since } y \text{ is an upper bound, and } x \text{ is a least upper bound, and} \\ y \leq x, &\quad \text{since } x \text{ is an upper bound, and } y \text{ is a least upper bound.} \end{aligned}$$

Thus, $x = y$. □

Therefore, we speak of *the* least upper bound of A, when it exists. The least upper bound is also called the *supremum* of A, and is denoted $\sup A$ (pronounced "soup A").

There is an analogous definition for sets bounded below.

Definition C.6 (Greatest lower bound). Let A be a set of numbers. A number x is the *greatest lower bound* of A if

(1) x is a lower bound of A, and

(2) if y is a lower bound of A, then $x \geq y$.

The greatest lower bound of A is also called the *infimum* of A, denoted $\inf A$.

We have seen that some sets have a least upper bound while others do not (similarly for greatest lower bounds). We may then ask, "Which sets have a least upper bound?".

Definition C.7 (The least upper bound property). A set S of numbers is said to have the *least upper bound property* if every nonempty subset of S which is bounded above has a least upper bound. [23]

Example C.1 shows that $\mathbb{Q}$ does not have the least upper bound property.

Theorem C.8 ($\mathbb{R}$ has the least upper bound property.). $\mathbb{R}$ has the least upper bound property.

Theorem C.1 is proved by constructing $\mathbb{R}$ from $\mathbb{Q}$. This is straightfoward, but rather long, so we do not go through this here. It is a fact that $\mathbb{R}$ is the *unique* extension of $\mathbb{Q}$ which has the least upper bound property (up to isomorphism). Going forward, we will simply assume this to be true and we will continue to use real numbers the way you are accustomed to using them. In the

[23] Consider $\emptyset \subseteq S$. Then by Definition C.4, every element of S is an upper bound for $\emptyset$, since $\emptyset$ has no elements (i.e., it is *vacuously true* that, given any $x \in S$, $y \leq x$ for all $y \in \emptyset$). Since any number $x \in S$ is an upper bound for $\emptyset$, if S is infinite then there is no least upper bound for $\emptyset$.

following, we will illustrate the least upper bound property and some consequences which are important for the study of calculus.

It follows from Theorem that $\mathbb{R}$ also has the *greatest lower bound property* (defined analogously to Definition C.7).

Theorem C.9. The fact that $\mathbb{R}$ has the least upper bound property implies that $\mathbb{R}$ also has the greatest lower bound property.

Proof. Let B be a nonempty subset of $\mathbb{R}$ which is bounded below. Let L be the set of all lower bounds of B, which is nonempty by assumption. Then, every $y \in L$ has the property that $y \leq x$ for all $x \in B$. This says that every $x \in B$ is an *upper bound* of L. Since L is a nonempty subset of $\mathbb{R}$ which is bounded above, it has a least upper bound by Theorem C.1. Let $\alpha = \sup L$. We claim that $\alpha = \inf B$.

Suppose $y < \alpha$. Then y is not an upper bound of L, hence $y \notin B$. This shows that $\alpha \leq x$ for every $x \in B$. Thus, $\alpha \in L$.

If $\alpha < y$, then $y \notin L$, since α is an upper bound of L.

We have shown that $\alpha \in L$ but $y \notin L$ if $y > \alpha$. In other words, α is a lower bound of B, but y is not if $y > \alpha$. Thus, $\alpha = \inf B$.

We have shown that every nonempty subset of $\mathbb{R}$ which is bounded below has a greatest lower bound. Thus, $\mathbb{R}$ satisfies the greatest lower bound property.

□

Theorem C.10 ($\mathbb{N} \subseteq \mathbb{R}$ is not bounded above.). $\mathbb{N} \subseteq \mathbb{R}$ is not bounded above.

Proof. (By contradiction.) Suppose $\mathbb{N} \subseteq \mathbb{R}$ is bounded above. Since $\mathbb{N} \neq \emptyset$, there exists a least upper bound for $\mathbb{N}$. Let $\alpha \equiv \sup \mathbb{N}$. Then

$$\alpha \geq n \qquad \text{for all } n \in \mathbb{N}.$$

Consequently,

$$\alpha \geq n+1 \qquad \text{for all } n \in \mathbb{N},$$

since $n+1 \in \mathbb{N}$ if $n \in \mathbb{N}$. But then

$$\alpha - 1 \geq n \qquad \text{for all } n \in \mathbb{N},$$

which says that $\alpha - 1$ is also an upper bound for $\mathbb{N}$. Since $\alpha - 1 < \alpha$, this contradicts the fact that α is the *least* upper bound for $\mathbb{N}$. □

Corollary C.11. For every $\epsilon > 0$ there exists $n \in \mathbb{N}$ such that $\frac{1}{n} < \epsilon$.

Proof. (By contradiction.) Suppose otherwise. Then $\frac{1}{n} \geq \epsilon$ for all $n \in \mathbb{N}$, and therefore $n \leq \frac{1}{\epsilon}$ for all $n \in \mathbb{N}$, which says that $\frac{1}{\epsilon}$ is an upper bound for $\mathbb{N}$, contradicting Theorem C.10. □

Example C.12. Consider the set $A = \left\{\frac{1}{n} : n \in \mathbb{N}\right\}$. Then $\sup A = 1$, since $\frac{1}{n} \leq 1$ for all $n \in \mathbb{N}$ and if $y < 1$, then $\frac{1}{1} = 1 > 1$, so y is not an upper bound. Note that $\sup A \in A$.

We have $\inf A = 0$. Since $\frac{1}{n} > 0$ for all $n \in \mathbb{N}$, 0 is a lower bound of A. If $\epsilon > 0$, then by Cor.C.11 there exists $n \in \mathbb{N}$ such that $\frac{1}{n} < \epsilon$, which shows that no positive number ϵ is a lower bound for A. Note that $\inf A = 0 \notin A$.

Proposition C.13. (a) Suppose $A \neq \emptyset$ is a subset of $\mathbb{R}$ that is bounded below. Let $-A = \{-x : x \in A\}$. Then $-A \neq \emptyset$ is bounded above, and $\inf A = -\sup(-A)$.

(b) If $A \neq \emptyset$ is bounded below, then the set B of all lower bounds of A is nonempty, bounded above, and $\sup B = \inf A$.

Proof. [Finish.] □

Exercise C.1. Find $\sup A$ and $\inf A$ (if they exist) for each the following subset of $\mathbb{R}$: $A = \left\{\frac{1}{n} : n \in \mathbb{Z} \text{ and } n \neq 0\right\}$.

Solution. By exactly the same argument as above, $\sup A = 1$. The fact that $\inf A = -1$ follows from Proposition C.13. [Finish.] □

C.2 Continuous maps OLD

Definition C.14 (Lipschitz continuity). A function $f : \mathbb{R}^n \to \mathbb{R}^k$ is *Lipschitz continuous* if there exists a positive real number K such that

$$||f(x_1) - f(x_2)||_{\mathbb{R}^k} \leq K||x_1 - x_2||_{\mathbb{R}^n}$$

for all $x_1, x_2 \in \mathbb{R}^n$. The real number K is said to be a *Lipschitz constant* for the function f.

Lemma C.15. If f is Lipschitz continuous, then it is uniformly continuous.

Proof. Given $\epsilon > 0$, choose $\delta = \frac{\epsilon}{K}$. Then

$$||f(x_1) - f(x_2)||_{\mathbb{R}^k} \leq K\frac{\epsilon}{K} = \epsilon$$

whenever $||x_1 - x_2||_{\mathbb{R}^n} < \delta$. □

Theorem C.16. Let $T : V \to W$ be a linear map between normed linear spaces. The following are equivalent:

(i) T is bounded

(ii) T is uniformly continuous

(iii) T is continuous at $0 \in V$

Proof. To see that (i) implies (ii), if T is bounded then

$$||T(\mathbf{v} - \mathbf{w})||_W = ||T\mathbf{v} - T\mathbf{w}||_W \leq C||\mathbf{v} - \mathbf{w}||_V,$$

which says that T is Lipschitz continuous, and therefore is uniformly continuous. To see that (ii) implies (iii), just take $\mathbf{w} = 0$ above. To prove that (iii) implies (i), if T is not bounded then choose $\mathbf{v}_n \in V$ such that $||\mathbf{v}_n||_V = 1$ and $||T\mathbf{v}_n||_W > n$. Then $\mathbf{w}_n \equiv \mathbf{v}_n/n$ satisfies $\lim_{n\to\infty} \mathbf{w}_n = 0$ but $||T\mathbf{w}_n||_W > 1$, so the sequence $(T\mathbf{w}_n)$ does not converge to $0 \in W$. Hence, T is not continuous at $0 \in V$. □

C.3 Continuous maps

Definition C.17 (Bounded linear map). A *linear* map $T : \mathbb{R}^n \to \mathbb{R}^m$ is said to be *bounded* if there exists $C \geq 0$ such that [24]

$$||T(\mathbf{x})|| \leq C||\mathbf{x}||, \qquad \text{for all } \mathbf{x} \in \mathbb{R}^n. \tag{C.3}$$

Note that this definition differs from the usual definition of a bounded function, which says that $T : \mathbb{R}^n \to \mathbb{R}^m$ is bounded if its image is a bounded subset of $\mathbb{R}^m$; that is, if there exists C such that $||T(\mathbf{x})|| \leq C$ for all $\mathbf{x} \in \mathbb{R}^n$. Since T is linear, it can't be bounded in this sense, since for all $\lambda \in \mathbb{R}$, we have $||T(\lambda\mathbf{x})|| = |\lambda|\,||T(\mathbf{x})||$. What this *does* say, is that applying T to a vector $\mathbf{x} \in \mathbb{R}^n$ produces a vector in $\mathbb{R}^m$ whose length is never increased by more than a factor of C. Thus, the image of a bounded set under a bounded linear map T is always bounded.

It therefore makes sense to define the "size" of a bounded linear map T to be the "maximum" factor by which T "lengthens" vectors (which corresponds to the smallest value of C for which (C.3) holds for all $\mathbf{x} \in \mathbb{R}^n$). The precise definition is the following.

Definition C.18 (Operator norm). Let $T : \mathbb{R}^n \to \mathbb{R}^m$ be a bounded linear operator. Define

$$||T||_{\text{op}} = \inf\{C \geq 0 : ||T(\mathbf{x})|| \leq C||\mathbf{x}|| \text{ for all } \mathbf{x} \in \mathbb{R}^n\}.$$

Note that the infimum exists, because this set is a nonempty subset of $\mathbb{R}$ which is bounded below (by zero).

The number $||T||_{\text{op}}$ is called the *operator norm* of T, which we will simply denote by $||T||$.

[Example?]

Proposition C.19 (Properties of the operator norm). Let $T, T_1, T_2 : \mathbb{R}^n \to \mathbb{R}^m$ be bounded linear operators and let $c \in \mathbb{R}$. Then

(i) $||T|| \geq 0$ and $||T|| = 0$ if and only if $T = 0$ (i.e., if and only if T is the zero map);

(ii) $||cT|| = |c|\,||T||$;

(iii) $||T_1 + T_2|| \leq ||T_1|| + ||T_2||$;

(iv) The following inequality is an immediate consequence of the definition:

$$||T(\mathbf{x})|| \leq ||T||\,||\mathbf{x}|| \qquad \text{for every } \mathbf{x} \in \mathbb{R}^n.$$

Proof. [Finish.] □

The next theorem shows that the bounded linear maps are precisely the continuous ones.

Theorem C.20. Let $T : \mathbb{R}^n \to \mathbb{R}^m$ be a bounded linear map. Then T is bounded if and only if it is continuous.

Proof. Suppose T is bounded. Then there exists $C > 0$ such that

$$||T(\mathbf{x})|| \leq C||\mathbf{x}||, \qquad \text{for all } \mathbf{x} \in \mathbb{R}^n.$$

[24] Note that $||T\mathbf{x}||$ is computed with respect to the distance function of $\mathbb{R}^m$ while $||\mathbf{x}||$ is computed with respect to the distance function of $\mathbb{R}^n$. Since this is obvious from context, we will use the same notation.

Given any $\epsilon > 0$, choose $\delta = \frac{\epsilon}{C}$. Then $||\mathbf{x} - \mathbf{y}|| < \delta$ implies

$$\begin{aligned} ||T(\mathbf{x} - \mathbf{y})|| &\leq C||\mathbf{x} - \mathbf{y}|| \\ &< C\delta \\ &= C\frac{\epsilon}{C} \\ &= \epsilon, \end{aligned}$$

which shows that T is continuous.

Conversely, suppose that T is continuous. In particular, T is continuous at the zero vector, so

$$||T(\mathbf{x})|| = ||T(\mathbf{x}) - T(\mathbf{0})|| \leq 1$$

for all vectors $\mathbf{x} \in \mathbb{R}^n$ with $||\mathbf{x}|| \leq \delta$. Thus, for all nonzero vectors $\mathbf{v} \in \mathbb{R}^n$,

$$\begin{aligned} ||T(\mathbf{v})|| &= \left|\left| \frac{||\mathbf{v}||}{\delta} T(\delta \frac{\mathbf{v}}{||\mathbf{v}||}) \right|\right| \\ &= \frac{||\mathbf{v}||}{\delta} \left|\left| T(\delta \frac{\mathbf{v}}{||\mathbf{v}||}) \right|\right| \\ &\leq \frac{||\mathbf{v}||}{\delta} \cdot 1 \\ &= \frac{1}{\delta}||\mathbf{v}||, \end{aligned}$$

which proves that T is bounded. □

We now show that every linear map from $\mathbb{R}^n$ to $\mathbb{R}^m$ is continuous. We will need the following lemma.

Lemma C.21. For any vector $\mathbf{v} = \sum_{i=1}^n c_i \mathbf{v}_i \in \mathbb{R}^n$,

$$||\mathbf{v}|| \geq \frac{1}{n} \sum_{i=1}^n |c_i|.$$

Proof.

$$||\mathbf{v}|| = \sqrt{\sum_{i=1}^n c_i^2} \geq |c_i|$$

for each $i = 1, \ldots, n$. Therefore

$$\underbrace{||\mathbf{v}|| + \cdots + ||\mathbf{v}||}_{n \text{ times}} = n||\mathbf{v}|| \geq \sum_{i=1}^n |c_i|$$

and therefore

$$||\mathbf{v}|| \geq \frac{1}{n} \sum_{i=1}^n |c_i|.$$

□

Theorem C.22. Every linear map $T : \mathbb{R}^n \to \mathbb{R}^m$ is bounded.

Proof. Let $B = \{\mathbf{v}_i\}_{i=1}^n$ be an ordered basis for $\mathbb{R}^n$. Then for any vector $\mathbf{v}$ there exist unique scalars $\{c_i\}_{i=1}^n$ such that

$$\mathbf{v} = \sum_{i=1}^{n} c_i \mathbf{v}_i.$$

Then

$$\begin{aligned} ||T(\mathbf{v})|| &= \left|\left| T(\sum_{i=1}^{n} c_i \mathbf{v}_i) \right|\right| \\ &= \left|\left| \sum_{i=1}^{n} c_i T(\mathbf{v}_i) \right|\right| \\ &= \sum_{i=1}^{n} |c_i| \ ||T(\mathbf{v}_i)|| \end{aligned}$$

Let $C = \max\{T(\mathbf{v}_i)\}_{i=1}^n$. Then

$$\begin{aligned} ||T(\mathbf{v})|| &\leq C \sum_{i=1}^{n} |c_i| \\ &\leq Cn||\mathbf{v}||, \end{aligned}$$

(where we have used Lemma C.21) which shows that T is bounded. □