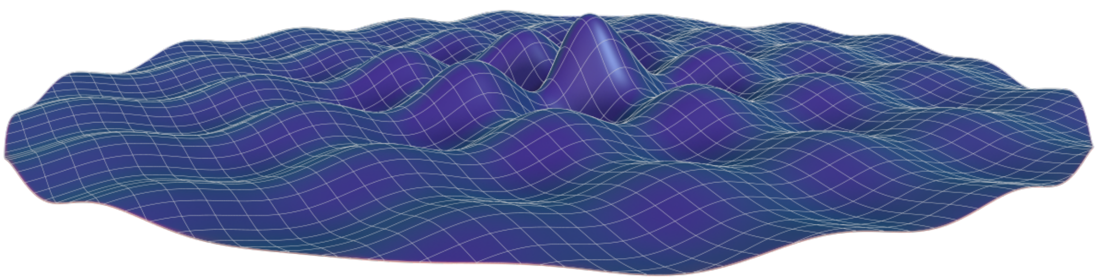
Multivariable Calculus

Jason Siefken



December 7, 2016

Contents

Introduction

Multivariable Calculus approaches the subject from a mathematical, but not overly technical, perspective. The key idea of calculus—chop things into little pieces and put them together again—is emphasized throughout.

Licensing

This book would not be possible without the long tradition of mathematical inquiry that came before. And like the ideas of mathematics, which are free for all to re-imagine, re-use, and re-purpose, so too is this book.

This book is a derived work of the Creative-Commons licensed *ISP Mathematics 281* multivariable calculus textbook by Leonard Evans of Northwestern University.

Contributors

This book is a collaborative effort. The following people have contributed to its creation:

◦ Dylaan Cornish ◦ Julia Dierksheide ◦ Maxwell Sigal ◦ Ieva Stakvileviciute ◦ Paul Strauss ◦

Chapter 1

Preliminaries

1.1 Mathematical Notation

Mathematics is a sophisticated and precise language, and we best not adventure into calculus without learning some basic words.

The most basic mathematical word is that of a *set*. A set is an unordered collection of distinct objects. We won't try and pin it down more exactly than this—our intuition about collections of objects will suffice[1]. We write a set with curly-braces $\{$ and $\}$ and list the objects inside. For instance

$$\{1,2,3\}.$$

This would be read aloud as "the set containing the elements 1, 2, and 3." The symbol $\in$ is used to specify that some object is an element of a set, and $\notin$ is used to specify it is not. For example,

$$3 \in \{1,2,3\} \qquad 4 \notin \{1,2,3\}.$$

Sets can contain mixtures of objects, including other sets. For example,

$$\{1,2,a,\{-70,\infty\},x\}$$

is a perfectly valid set.

It is tradition to use capital letters to name sets. So we might say $A = \{6,7,12\}$ or $X = \{7\}$. There is, however, a special set which already has its own name—the empty set. The *empty set* is the set containing no elements and is written $\emptyset$ or $\{\}$. Note that $\{\emptyset\}$ is *not* the empty set. It is the set containing the empty set! It is also traditional to call elements of a set *points* regardless of whether you consider them "point-like" objects.

Operations on Sets

If the set A contains all the elements that the set B does, we call A a *superset* and B a *subset*. We'll give this a formal definition.

[1] When you pursue more rigorous math, you rely on definitions to get yourself out of philosophical jams. For instance, with our definition of set, consider "the set of all sets that don't contain themselves." Such a set cannot exist! This is called *Russel's Paradox*, and shows that if we start talking about sets of sets, we may need more than intuition.

Definition 1.1.1 — Subset & Superset. The set B is a *subset* of the set A, written $B \subseteq A$ if for all $b \in B$ we also have $b \in A$. In this case, A is called a *superset* of B.[a]

[a] Some mathematicians use the symbol $\subset$ instead of $\subseteq$.

Some simple examples are $\{1,2,3\} \subseteq \{1,2,3,4\}$ and $\{1,2,3\} \subseteq \{1,2,3\}$. There's something funny about that last example, though. Those two sets are not only subsets/supersets of each other, they're *equal*. As surprising as it seems, we actually need to define what it means for two sets to be equal.

Definition 1.1.2 — Set Equality. The sets A and B are *equal*, written $A = B$ if $A \subseteq B$ and $B \subseteq A$.

Having a definition of equality to lean on will help us when we need to prove things about sets.

■ **Example 1.1** Let A be the set of numbers that can be expressed as $2n$ for some whole number n and let B be the set of numbers that can be expressed as $m+1$ where m is an odd whole number. We will show $A = B$.

First, let us show $A \subseteq B$. If $x \in A$ then $x = 2n$ for some whole number n. Therefore $x = 2n = 2(n-1)+1+1 = m+1$ where $m = 2(n+1)+1$ is, by definition, an odd number. Therefore $x \in B$.

Now we will show $B \subseteq A$. Let $x \in B$. By definition, $x = m+1$ for some odd m and so by the definition of oddness, $m = 2k+1$ for some whole number k. Thus

$$\begin{aligned} x = m+1 &= (2k+1)+1 = 2k+2 \\ &= 2(k+1) = 2n, \end{aligned}$$

where $n = k+1$. Thus, $x \in A$. Since $A \subseteq B$ and $B \subseteq A$, by definition $A = B$. ■

Set-builder Notation

Specifying sets by listing all their elements can be a hassle, and if there are an infinite number of elements, it's impossible! Fortunately, *set-builder notation* solves these problems. If X is a set, we can define a subset

$$Y = \{a \in X \,:\, \text{some rule involving } a\},$$

which is read "Y is the set of a in X *such that* some rule involving a is true." If X is intuitive, we may omit it and simply write $Y = \{a \,:\, \text{some rule involving } a\}$[2]. You may equivalently use "|" instead of ":", writing $Y = \{a \mid \text{some rule involving } a\}$.

■ **Example 1.2** The set $\mathbb{Z}$ is the set of integers (positive, negative, and zero whole numbers). To define E as the even integers, we could write

$$E = \{n \in \mathbb{Z} \,:\, n = 2k \text{ for some } k \in \mathbb{Z}\}.$$

[2] If you want to get technical, to make this notation unambiguous, you define a *universe of discourse*. That is, a set $\mathcal{U}$ containing every object you might want to talk about. Then $\{a \,:\, \text{some rule involving } a\}$ is short for $\{a \in \mathcal{U} \,:\, \text{some rule involving } a\}$

To define P as the set of positive integers, we could write

$$P = \{n \in \mathbb{Z} \,:\, n > 0\}.$$

■

There are also some common operations we can do with two sets.

Definition 1.1.3 — Intersections & Unions. Let A and B be sets. Then the *intersection* of A and B, written $A \cap B$, is defined by

$$A \cap B = \{x \,:\, x \in A \text{ and } x \in B\}.$$

The *union* of A and B, written $A \cup B$, is defined by

$$A \cup B = \{x \,:\, x \in A \text{ or } x \in B\}.$$

For example, if $A = \{1,2,3\}$ and $B = \{-1,0,1,2\}$, then $A \cap B = \{1,2\}$ and $A \cup B = \{-1,0,1,2,3\}$. Set unions and intersections are *associative*, which means it doesn't matter how you apply parentheses to an expression involving just unions or just intersections. For example $(A \cup B) \cup C = A \cup (B \cup C)$, which means we can give an unambiguous meaning to an expression like $A \cup B \cup C$ (just put the parenthesis wherever you like). But watch out, $(A \cup B) \cap C$ means something different than $A \cup (B \cap C)$!

Definition 1.1.4 — Set Subtraction. For sets A and B, the *set-wise difference* between A and B, written $A \backslash B$, is the set

$$A \backslash B = \{x \,:\, x \in A \text{ and } x \notin B\}.$$

Definition 1.1.5 — Cardinality. For a set A, the *cardinality* of A, written $|A|$ is the number of elements in A. If A contains infinitely many elements, we write $|A| = \infty$.

Let's define some notation for common sets.

$$\begin{aligned}
\emptyset &= \{\}, \text{ the empty set} \\
\mathbb{N} &= \{0,1,2,3,\ldots\} = \{\text{natural numbers}\} \\
\mathbb{Z} &= \{\ldots,-3,-2,-1,0,1,2,3,\ldots\} = \{\text{integers}\} \\
\mathbb{Q} &= \{\text{rational numbers}\} \\
\mathbb{R} &= \{\text{real numbers}\} \\
\mathbb{R}^n &= \{\text{vectors in } n\text{-dimensional Euclidean space}\}
\end{aligned}$$

Besides unions, there's another way to join sets together: *products*.

Definition 1.1.6 — Cartesian Product. Given two sets A and B, the *Cartesian product* (sometimes shortened to *product*) of the sets A and B is written $A \times B$ and defined to be

$$A \times B = \{(a,b) \,:\, a \in A \text{ and } b \in B\}.$$

The Cartesian product of two sets is the set of all ordered pairs of elements from those sets. For example,

$$\{1,2\} \times \{1,2,3\} = \{(1,1),(1,2),(1,3),(2,1),(2,2),(2,3)\}.$$

You can repeat this operation more than once. $\mathbb{R}\times\mathbb{R}\times\mathbb{R}$ is the set of all triples of real numbers. Borrowing notation from numbers, if you take the Cartesian product of a set with itself some number of times, you can represent it with an exponent. Thus, $\mathbb{R}\times\mathbb{R}\times\mathbb{R}$ can be written as $\mathbb{R}^3$, which is a set we've seen before[3].

Functions

You're probably used to seeing functions like $f(x)=x^2$, but it's worth reviewing some of the concepts and terminology associated with functions.

Definition 1.1.7 — Function. A *function* with *domain* the set A and *co-domain* the set B is an object that associates every point in the set A with *exactly one* point in the set B.

If a function f has domain A and co-domain[4] B, we notate this by writing $f:A\to B$. If we want to further specify what the function f actually is, we need to express how f associates each point in A to a point in B. This can be done with an equation. For example, we could define the function $f:\mathbb{R}\to\mathbb{R}$ by

$$f(x)=2x,$$

which says that each real number gets associated to its double. We can notate the same thing using a special type of arrow: "$\mapsto$". Now we might write

$$f:\mathbb{R}\to\mathbb{R} \text{ where } x\mapsto 2x,$$

which is read "f is a function from $\mathbb{R}$ to $\mathbb{R}$ where $x\in\mathbb{R}$ gets mapped to $2x$."

Note that every point in the co-domain of a function doesn't need to get mapped to. For example $g:\mathbb{R}\to\mathbb{R}$ given by $g(x)=x^2$ outputs only non-negative numbers, but it is still valid to specify $\mathbb{R}$ as the co-domain. However, if we wanted to make a point of it, we are perfectly justified in writing $g:\mathbb{R}\to[0,\infty)$ when defining g.

Many common math operations give rise to functions. For example, $f(x)=\sqrt{x}$ is the familiar square root function. Sometimes, when we wish to talk about a function for which notation already exists, we will put a " $\cdot$ " where we would normally put a variable. Thus, we might say, "$\sqrt{\cdot}$ is the square root function."[5]

Definition 1.1.8 — Range. The *range* of a function $f:A\to B$ is the set of all outputs of f. That is

$$\operatorname{range} f=\{y\in B\ :\ y=f(x) \text{ for some } x\in A\}.$$

Definition 1.1.9 — Image. Let $f:A\to B$ be a function. The *image* of a set $X\subseteq A$, written $f(X)$ is defined by

$$f(X)=\{y\in B\ :\ y=f(x) \text{ for some } x\in X\}.$$

We see that if $f:A\to B$, $\operatorname{range} f=f(A)$. In words, the range of f is the image of its domain. This language will become useful when we think of functions as transformations that

[3] If you're scratching your head saying, "I thought $\mathbb{R}^3$ was vectors in 3-dimensional space. How do we know that's the same thing as triples of real numbers?" your mind is keen. This is a theorem of linear algebra.

[4] Some people use the word *range* interchangeably with co-domain.

[5] Since $\sqrt{x}$ is "the square root of the quantity x," it is technically a quantity and not a function. This is why we write $\cdot$ instead of x when we want to refer to the square root *function*.

move or bend space. If $f : \mathbb{R}^2 \to \mathbb{R}^2$ is a function that warps the Cartesian plane, then the image of X under f could be visualized by painting X on the Cartesian plane, warping the whole plane, and then looking at the resulting, painted shape.

Closely related to images, we have the idea of *restriction*. Suppose $f : \mathbb{R}^2 \to \mathbb{R}$ is defined by $f(x, y) = xy$, but we were only really interested in f on the unit circle, $\mathcal{C}$. In this case, we might say f attains a maximum on $\mathcal{C}$, or f *restricted to* $\mathcal{C}$ attains a maximum, even though f itself is unbounded. This idea comes up often enough to deserve its own notation.

Definition 1.1.10 — Restriction. If $f : A \to B$ and $X \subseteq A$, the *restriction* of f to X is written $f\big|_X$ and is defined to be the function $g : X \to A$ where $x \mapsto f(x)$.

The last important function-related ideas for us are function composition and inverses. Given two functions $f : A \to B$ and $g : B \to C$, we can *compose* g and f to get a new function.

Definition 1.1.11 — Composition. Given two functions $f : A \to B$ and $g : B \to C$, the *composition* of g and f, written $g \circ f$, is the function $h : A \to C$ where $x \mapsto g(f(x))$.

Note that the composition $g \circ f$ has the domain of f and the co-domain of g. When a point is fed into $g \circ f$, it moves from $A \to B \to C$. The composition $g \circ f$ only makes sense because the outputs of f are allowed as inputs to g. If we wrote $f \circ g$, it wouldn't mean much, because g outputs points in C and f has no idea what to do with points in C.[6]

Inverses relate to composition and the *identity function*, the function that does nothing to its inputs.

Definition 1.1.12 — Identity Function. The *identity function* $\mathrm{id} : A \to A$ is defined by the relation

$$\mathrm{id}(x) = x$$

for all $x \in A$.

Notice that for every set, that set is the domain of an identity function. Since the domain and co-domain of a function are part of its definition, we don't want to confuse them. After all, $f : \{0, 1\} \to \{0, 1\}$ given by $f(x) = x^2$ is a different function from $f : \mathbb{R} \to \mathbb{R}$ given by $f(x) = x^2$. For the special case of the identity function, we sometimes write the domain of the function as a subscript. That is, for $\mathrm{id} : A \to A$ we'd write id_A so it doesn't get confused with $\mathrm{id} : B \to B$, which we'd write id_B.

Definition 1.1.13 — Inverse Function. Let $f : A \to B$ be a function. If there exists a function $g : B \to A$ such that

$$f \circ g = \mathrm{id}_B \qquad \text{and} \qquad g \circ f = \mathrm{id}_A,$$

we say f is *invertible* and we call g the *inverse* of f. If f is invertible, we notate its inverse by f^{-1}.

Inverses can be tricky some times. For example, consider $f(x) = x^2$ and $g(x) = \sqrt{x}$. Here $g \circ f(x) = \sqrt{x^2} = |x|$ and $f \circ g(x) = \sqrt{x}^2 = x$. What's the deal? Well, it's all about domains.

[6] It seems a little backward to write $f : A \to B$, $g : B \to C$ and then write $g \circ f$ instead of $f \circ g$. You can thank Euler for that. He decided to write functions with their input on the right instead of the left. If we wrote functions backwards, like $((x)f)g$ for "g of f of x," they we could just *follow the arrows* and life would be simpler.

$f : \mathbb{R} \to [0,\infty)$ and $g : [0,\infty) \to [0,\infty)$. So, the domain of $g \circ f$ is $\mathbb{R}$ and the domain of $f \circ g$ is $[0,\infty)$. The domains are different, and indeed f is not invertible. However, g *is* invertible, and $g^{-1} = f\big|_{[0,\infty)}$. If we only input non-negative numbers into f, then f exactly undoes what g did. This subtle domain trickery can cause us a lot of headaches if we're not used to thinking carefully, and many of our favorite functions that we're used to calling "inverse functions" are actually only inverses when paired with specific domains.

1.2 Proof

Mathematics has the highest standard of proof of any field. In the Platonic ideal of mathematics, we start from some basic assumptions, called *axioms*, that we have all agreed upon. Then from those axioms, using the rules of logic, we deduce *theorems*. Every single mathematical statement we make can be traced back from theorem to theorem and eventually to our initial axioms.

This is contrary to other disciplines, like physics. In physics, based on observation, we construct *laws*. Laws in physics are like axioms in mathematics, but they have an important difference—they can be disproven by observation. A mathematical axiom can never be disproven. One can certainly argue that an axiom is not *useful* or not *interesting*, but you cannot say it's *wrong*[7]. Of course, as human practitioners, we may misuse logic and be wrong ourselves, but that is no fault of the axioms.

But now, let's deviate from philosophical perfection and visit reality. In reality, *mathematics is a human pursuit to understand relationships between ideas and their consequences*. The key there is that *humans* do mathematics to *understand* relationships. If a theorem in math can ultimately be reduced to logical statements about axioms, but the argument is 100000 steps long, it doesn't help a human understand why something is true. Instead, a shorter argument that skips over some steps is more useful to us. And, indeed, most of our mathematics to date skips over some steps[8].

We call a correct mathematical argument a *proof*. A proof starts from a set of assumptions, and following the rules of logic, arrives at a conclusion. Strictly speaking, a proof doesn't need to make sense or show motivation, applications, or examples. It just has to be a sequence of correct logical steps. However, for us, as humans studying mathematics, we prove things for two reasons: to understand why things are true and to avoid making mistakes.

Reconciling these two goals can be very hard for a novice mathematician. If you include *all* the steps, it won't help with understanding, but if you don't include enough steps, the argument may not be convincing and might contain mistakes. Even professionals struggle to balance these competing goals, and how you balance those goals depends on your audience—if

[7] There are multiple ways to axiomatize geometry. In Euclidean geometry every pair of lines either coincides, intersects in exactly one place, or does not intersect. In spherical geometry, every pair of lines either coincides or intersects in exactly two places. Euclidean geometry is useful when your space looks flat. Spherical geometry is useful when your space is the surface of a sphere (like the Earth). Is one of these more *right* than the other? They're certainly contradictory.

[8] There are some projects to prove all of mathematics directly from the axioms using computer assistance. They've made progress, but there are still theorems in calculus that have not been reduced to the axioms. We believe that they *could be* reduced to the axioms, but no one has taken the time to do so.

you're trying to convince your math professor of something your proof will need to have more detail than if you were trying to convince your friend (mathematicians are very skeptical!).

Enough talk, let's go through a 2000-year-old example of a proof.

Theorem 1.2.1 There is no rational number p/q such that $(p/q)^2 = 2$.

Proof. If p/q is a rational number, it can be expressed in lowest terms. Suppose p/q is in lowest terms and $(p/q)^2 = 2$. Then $p^2 = 2q^2$ and so p^2 is even. Since p^2 is even, it must be that p is even, and so by definition, $p = 2m$ for some integer m. Now,

$$\frac{p^2}{q^2} = \frac{(2m)^2}{q^2} = \frac{4m^2}{q^2} = 2,$$

with the last equality following by assumption. Multiplying both sides by q^2 and dividing by 2 we arrive at the equation

$$2m^2 = q^2,$$

and so q^2 is even which means q is even. By definition, this means $q = 2n$ for some integer n. But now,

$$\frac{p}{q} = \frac{2m}{2n}$$

is not in lowest terms! This is a contradiction and so it cannot be that $(p/q)^2 = 2$. ■

This is nearly identical to the argument the ancient Greeks gave. It's elegant, beautiful, and convincing. But, if we look closer, it does skip some steps. For example, it relies on the fact that there is such a thing as *lowest terms*. This is something that would need to be proven—a priori, the conclusion of the proof could be that the assumption that p/q could be in lowest terms is false.

You will not, overnight, become a master at understanding what steps you can leave out and what steps you must show. However, with feedback, you'll get better. For a detailed guide on writing good proofs, please see Appendix A.

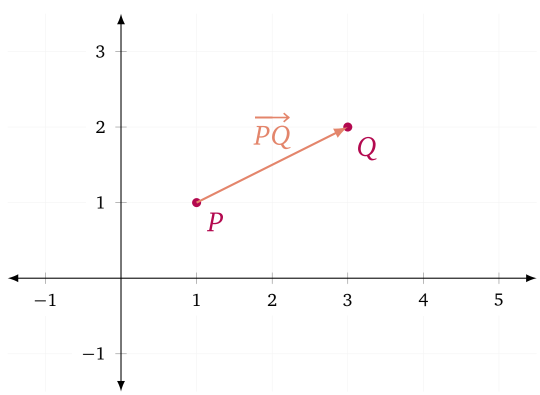
Chapter 2

Vectors

A *vector* is a quantity which is characterized by a *magnitude* and a *direction*. Many quantities are best described by vectors rather than numbers. For example, when driving a car, it may be sufficient to know your speed, which can be described by a single number, but the motion of an airplane must be described by a vector quantity—velocity—which takes into account its direction as well as its speed.

Ordinary numerical quantities are called *scalars* when we want to emphasize that they are not vectors.

Whereas numbers allow us to specify relationships between single quantities (put in twice as much flour as sugar), vectors will allow us to specify relationships between geometric objects in space[1]. If we have two points, $P = (1,1)$ and $Q = (3,2)$, we specify the *displacement* from P to Q as a vector.



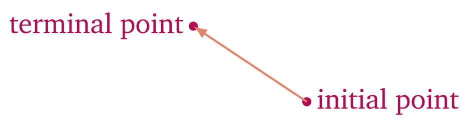
We notate the displacement vector form P to Q by $\overrightarrow{PQ}$. The magnitude of $\overrightarrow{PQ}$ is given by the Pythagorean theorem to be $\sqrt{5}$ and its direction is specified by the directed line segment from P to Q.

[1] Though in this book we will treat vectors as intertwined with Euclidean space, they are much more general. For instance, someone's internet browsing habits could be described by a vector—the topics they find most interesting might be the "direction" and the amount of time they browse might be the "magnitude."

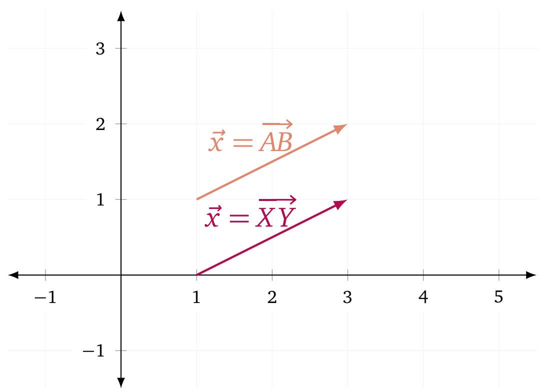
2.1 Vector Notation

There are many ways to represent vector quantities in writing. If we have two points, P and Q, we write $\overrightarrow{PQ}$ to represent the vector from P to Q. Absent of points, bold-faced letters or a letter with an arrow over it are the most common typographical representations of vectors. For example, $\vec{a}$ or $\mathbf{a}$ may both be used to represent the vector quantity named "a." In this book we will use $\vec{a}$ to represent a vector. The notation $||\vec{a}||$ represents the magnitude of the vector $\vec{a}$, which is sometimes called the *norm* of $\vec{a}$.

Graphically, vectors are represented as directed line segments (a line segment with an arrow at one end). The endpoints of the segment are called the *initial point* (the base) and the *terminal point* (the tip) of the vector.



Let $A = (1,1)$, $B = (3,2)$, $X = (1,0)$, and $Y = (3,1)$ and consider the vectors $\vec{a} = \overrightarrow{AB}$ and $\vec{x} = \overrightarrow{XY}$. Are these the same or different vectors? If we drew them as directed line segments, the drawings would be distinct. However, both $\vec{a}$ and $\vec{x}$ have equivalent magnitudes and directions. Thus, $\vec{a}$ and $\vec{x}$ are *equivalent*, and we would be justified writing $\vec{a} = \vec{x}$.



Alternatively, we could consider the *rooted vector* $\vec{a}$ rooted at the point A. In this terminology, $\vec{a}$ rooted at A is *different* than $\vec{a}$ rooted at X. This idea of rooted vectors will occasionally be useful, but our primary study will be unrooted vectors.

Vectors and Points

The distinction between vectors and points is sometimes nebulous because they are so closely related to each other. A *point* in Euclidean space specifies an absolute position whereas a vector specifies a magnitude and direction. However, given a point P, one associates P with the vector $\vec{p} = \overrightarrow{OP}$, where O is the origin. Similarly, we associate the vector $\vec{v}$ with the point V so that $\overrightarrow{OV} = \vec{v}$. Thus, we have a way to unambiguously go back and forth between vectors and points[2]. As such, *we will treat vectors and points as interchangeable*.

[2] Mathematically, we say there is an *isomorphism* between vectors and points.

2.2 Vector Arithmetic

Vectors provide a natural way to give directions. For example, suppose $\hat{\mathbf{x}}$ points one mile eastwards and $\hat{\mathbf{y}}$ points one mile northwards. Now, if you were standing at the origin and wanted to move to a location 3 miles east and 2 miles north, you might say: "Walk 3 times the length of $\hat{\mathbf{x}}$ in the $\hat{\mathbf{x}}$ direction and 2 times the length of $\hat{\mathbf{y}}$ in the $\hat{\mathbf{y}}$ direction." Mathematically, we express this as

$$3\hat{\mathbf{x}} + 2\hat{\mathbf{y}}.$$

Of course, we've incidentally described a new vector. Namely, let P be the point at 3-east and 2-north. Then

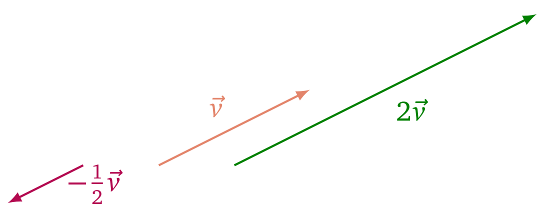
$$\overrightarrow{OP} = 3\hat{\mathbf{x}} + 2\hat{\mathbf{y}}.$$

If the vector $\vec{r}$ points north but has a length of 10 miles, we have a similar formula:

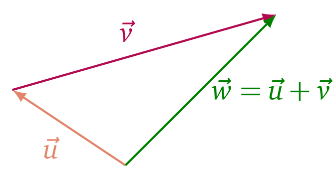
$$\overrightarrow{OP} = 3\hat{\mathbf{x}} + \tfrac{1}{5}\vec{r},$$

and we have the relationship $\vec{r} = 10\hat{\mathbf{y}}$. Our notation here is very suggestive. Indeed, if we could make sense of what $\alpha\vec{v}$ is for any scalar α and vector $\vec{v}$, and we could make sense of what $\vec{v}+\vec{w}$ means for any vectors $\vec{v}$ and $\vec{w}$, we would be able to do algebra with vectors. We might even say we have *an algebra of vectors*.

Intuitively, for a vector $\vec{v}$ and a scalar $\alpha > 0$, the vector $\vec{w} = \alpha\vec{v}$ should point in the same direction as $\vec{v}$ but have magnitude scaled up by α. That is, $\|\vec{w}\| = \alpha\|\vec{v}\|$. Similarly, $-\vec{v}$ should be the vector of the same length as $\vec{v}$ but pointing in the exact opposite direction.



For two vectors $\vec{u}$ and $\vec{v}$, the sum $\vec{w} = \vec{u}+\vec{v}$ should be the displacement vector created by first displacing along $\vec{u}$ and then displacing along $\vec{v}$.



Now, there is one snag. What should $\vec{v}+(-\vec{v})$ be? Well, first we displace along $\vec{v}$ and then we displace in the exact opposite direction by the same amount. So, we have gone nowhere. This corresponds to a displacement with zero magnitude. But, what direction did we displace? Here we make a philosophical stand.

Definition 2.2.1 — Zero Vector. The *zero vector*, notated as $\vec{0}$, is the vector with no magnitude.

We will be pragmatic about the direction of the zero vector and say, *the zero vector does not have a well-defined direction*[3]. That means sometimes we consider the zero vector to point in every direction and sometimes we consider it to point in no directions. It depends on our mood—but we must never talk about *the* direction of the zero vector, since it's not defined.

We need the zero vector if we are to make precise mathematical sense of vector arithmetic. Further along this line of thinking, we can define precisely how vector arithmetic should behave. Specifically, if $\vec{u}$, $\vec{v}$, $\vec{w}$ are vectors and α and β are scalars, the following conditions should be satisfied:

$$(\vec{u}+\vec{v})+\vec{w}=\vec{u}+(\vec{v}+\vec{w}) \qquad \text{(Associativity)}$$
$$\vec{u}+\vec{v}=\vec{v}+\vec{u} \qquad \text{(Commutativity)}$$
$$\alpha(\vec{u}+\vec{v})=\alpha\vec{u}+\alpha\vec{v} \qquad \text{(Distributivity)}$$

and

$$(\alpha\beta)\vec{v}=\alpha(\beta\vec{v}) \qquad \text{(Associativity II)}$$
$$(\alpha+\beta)\vec{v}=\alpha\vec{v}+\beta\vec{v} \qquad \text{(Distributivity II)}$$

Indeed, if we intuitively think about vectors in flat (Euclidean) space, all of these properties are satisfied[4]. From now on, these properties of vector operations will be considered the *laws (or axioms) of vector arithmetic*.

We'll be talking about these vector operations (scalar multiplication and vector addition) a lot. So much so that the concept is worth naming.

Definition 2.2.2 — Linear Combination. A *linear combination* of the vectors $\vec{v}_1,\dots\vec{v}_n$ is any vector expressible as

$$\alpha_1\vec{v}_1+\cdots+\alpha_n\vec{v}_n,$$

where $\alpha_1,\alpha_2,\dots,\alpha_n$ are scalars.

We've given laws for linear combinations of vectors, but what about for magnitudes of vectors? We'd like the magnitude (or norm) of a vector to obey the following laws.

$$||\vec{v}||\geq 0 \qquad \text{(Non-negativity)}$$
$$||\vec{v}||=0 \text{ only when} \vec{v}=\vec{0} \qquad \text{(Definiteness)}$$
$$||\alpha\vec{v}||=|\alpha|||\vec{v}|| \qquad \text{(Homogeneity)}$$
$$||\vec{v}+\vec{w}||\leq||\vec{v}||+||\vec{w}|| \qquad \text{(Triangle Inequality)}$$

for all $\vec{v}$, $\vec{w}$, and scalars α. Any function on vectors satisfying those four properties is called a *norm*, and our usual notion of length in three-dimensional space indeed obeys those properties[5].

[3] In the mathematically precise definition of vector, the idea of "magnitude" and "direction" are dropped. Instead, a set of vectors is defined to be a set over which you can reasonably define addition and scalar multiplication.

[4] If we deviate from flat space, some of these rules are no longer respected. Consider moving 100 miles north then 100 miles east on a sphere. Is this the same as moving 100 miles east and then 100 miles north?

[5] The Euclidean norm comes from the Pythagorean theorem $a^2+b^2=c^2$. However, by changing the exponent, we have a whole family of norms coming from the equations $|a|^p+|b|^p=|c|^p$.

Homogeneity is a particularly special property of a norm. It allows us to easily create *unit vectors*.

Definition 2.2.3 — Unit Vector. A *unit vector* is a vector $\vec{u}$ satisfying $||\vec{u}|| = 1$.

Unit vectors are handy because if $\vec{u}$ is a unit vector, then $k\vec{u}$ has length $|k|$. Further, we can always turn a vector into a unit vector.

■ **Example 2.1** The vector $\vec{v}/||\vec{v}||$ is always a unit vector in the direction of $\vec{v}$. Computing,

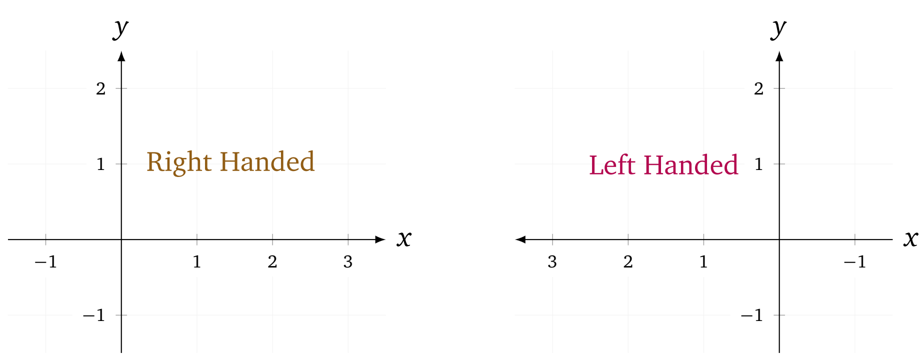
$$\left\|\frac{\vec{v}}{||\vec{v}||}\right\| = \left|\frac{1}{||\vec{v}||}\right| ||\vec{v}|| = \frac{1}{||\vec{v}||}||\vec{v}|| = 1.$$

■

2.3 Coordinates

Recall that a coordinate system in the plane is specified by choosing an origin O and then choosing two perpendicular axes meeting at the origin. These axes are chosen in some order so that we know which axis (usually the x-axis) comes first and which (usually the y-axis) second. Note that there are many different coordinate systems which could be used although we often draw pictures as if there were only one.

In physics, one often has to think carefully about the coordinate system because choosing one appropriately may greatly simplify the analysis. Note that axes for coordinate systems are usually drawn with *right-hand orientation*, where the right angle from the positive x-axis to the positive y-axis is in the counter-clockwise direction. However, it would be equally valid to use the *left-hand orientation* in which that angle is in the clockwise direction. One can easily switch the orientation of a coordinate system by reversing one of the axes[6].



For any coordinate system, there are special vectors associated with it. For the plane, the vector pointing one unit along the positive x-axis is called $\hat{\mathbf{x}}$ and the vector pointing one unit along the positive y-axis is called $\hat{\mathbf{y}}$. The vectors $\hat{\mathbf{x}}$ and $\hat{\mathbf{y}}$ are called the *standard basis* vectors for $\mathbb{R}^2$.

[6] The concept of orientation is quite fascinating and it arises in mathematics, physics, chemistry, and even biology in many interesting ways. Note that almost all of us base our intuitive concept of orientation on our inborn notion of "right" versus "left".

Notice that every point (or vector) in the plane can be represented as a linear combination of $\hat{\mathbf{x}}$ and $\hat{\mathbf{y}}$, and the vector $\alpha\hat{\mathbf{x}}+\beta\hat{\mathbf{y}}$ is the vector $\overrightarrow{OP}$ where $P=(\alpha,\beta)$. Now, to state an intuitive fact: if $\vec{w}$ is a vector in the plane, *there is only one way to write a vector as a linear combination of* $\hat{\mathbf{x}}$ *and* $\hat{\mathbf{y}}$. This means, if $\vec{w}=\alpha\hat{\mathbf{x}}+\beta\hat{\mathbf{y}}$, the pair (α,β) captures all information[7] about $\vec{w}$.

For a vector $\vec{w}=\alpha\hat{\mathbf{x}}+\beta\hat{\mathbf{y}}$, we call the pair (α,β) the *components* of the vector $\vec{w}$. There are many equivalent notations used to represent components.

(α,β)	parenthesis
$\langle\alpha,\beta\rangle$	angle brackets
$\begin{bmatrix}\alpha & \beta\end{bmatrix}$	square brackets in a row (a row matrix)
$\begin{bmatrix}\alpha \\ \beta\end{bmatrix}$	square brackets in a column (a column matrix)

Given what we now know about representing vectors and their equivalency with points, we can now dissect the notation $\mathbb{R}^2$. On the one hand, $\mathbb{R}^2$ is the set of vectors in two-dimensional Euclidean space. On the other hand $\mathbb{R}^2=\mathbb{R}\times\mathbb{R}$ is the set of all pairs of real numbers. Via the use of coordinates, we know these concepts represent the same thing! Further, since vectors in $\mathbb{R}^2$ are equivalent to their representation in coordinates, we will often write

$$\vec{v}=(\alpha,\beta)$$

as a shorthand for $\vec{v}=\alpha\hat{\mathbf{x}}+\beta\hat{\mathbf{y}}$.

Breaking vectors into components, and in particular, viewing vectors as linear combinations of the standard basis vectors, allows us to solve problems that were difficult before. For instance, suppose we have vectors $\vec{v}$ and $\vec{w}$. How can we compute $\|\vec{v}+\vec{w}\|$? With components, it's easy.

■ **Example 2.2** Suppose $\vec{v}=\alpha_1\hat{\mathbf{x}}+\beta_1\hat{\mathbf{y}}$ and $\vec{w}=\alpha_2\hat{\mathbf{x}}+\beta_2\hat{\mathbf{y}}$. By the laws of vector arithmetic we have

$$\vec{v}+\vec{w}=(\alpha_1\hat{\mathbf{x}}+\beta_1\hat{\mathbf{y}})+(\alpha_2\hat{\mathbf{x}}+\beta_2\hat{\mathbf{y}})=(\alpha_1+\alpha_2)\hat{\mathbf{x}}+(\beta_1+\beta_2)\hat{\mathbf{y}}.$$

Now, since $\hat{\mathbf{x}}$ and $\hat{\mathbf{y}}$ are orthogonal to each other, the Pythagorean theorem gives

$$\|\vec{v}+\vec{w}\|=\sqrt{(\alpha_1+\alpha_2)^2+(\beta_1+\beta_2)^2}.$$

■

Writing things in terms of the standard basis allowed us to make easy work of computing $\|\vec{v}+\vec{w}\|$ in Example 2.2. We can use the laws of vector arithmetic to produce rules for working with components.

The rules are are likely familiar:

$$\begin{bmatrix}a \\ b\end{bmatrix}+\begin{bmatrix}\alpha \\ \beta\end{bmatrix}=\begin{bmatrix}a+\alpha \\ b+\beta\end{bmatrix} \qquad \text{and} \qquad \alpha\begin{bmatrix}a \\ b\end{bmatrix}=\begin{bmatrix}\alpha a \\ \alpha b\end{bmatrix}.$$

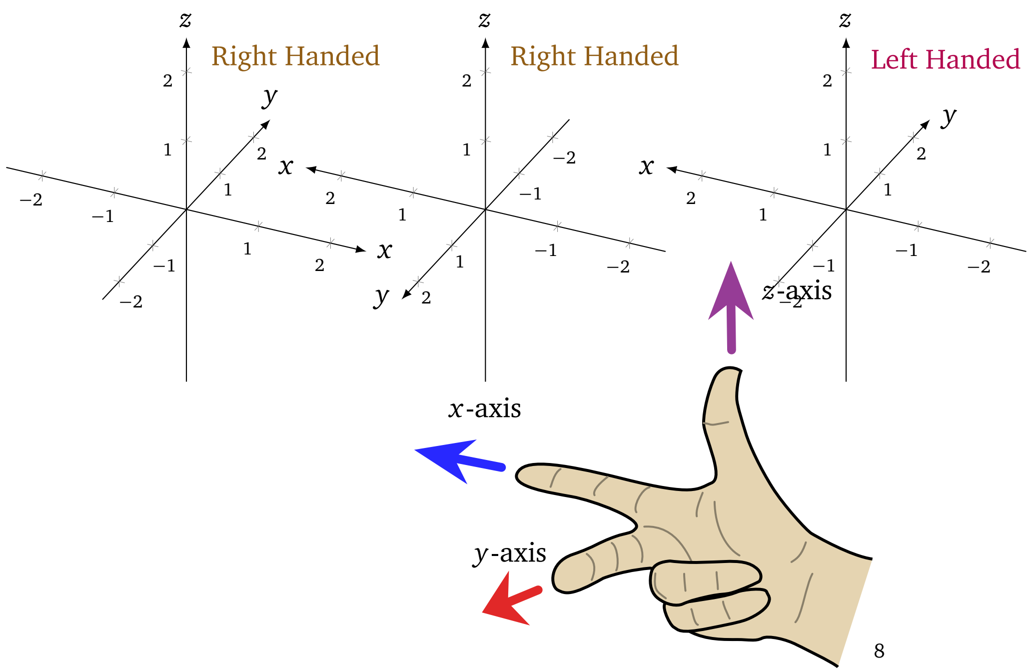
[7] Maybe you already knew this because the point (α,β) is described by the pair of numbers (α,β), duh! But consider, what would we do if we didn't know about coordinates at all? One approach is to *define* coordinates in terms of vectors, which is really what we're doing.

> **Exercise 2.1** Prove the rules for adding the component representation of vectors and multiplying the component representation of vectors directly from the laws of vector arithmetic.

Armed with these rules, we will be able to tackle sophisticated vector problems.

Three-dimensional Coordinates

In three-dimensional space, the story is very similar. Again, we imagine three perpendicular axes, the x, y, and z axes. To draw consistent pictures, we have a notion of a right-handed three-dimensional coordinate system given by the *right-hand rule*.



[8]

We now have three standard basis vectors, $\hat{\mathbf{x}}$, $\hat{\mathbf{y}}$, and $\hat{\mathbf{z}}$, each pointing one unit in the positive direction of their respective axes. Any vector in three-dimensional space can be represented in exactly one way as a linear combination $\alpha\hat{\mathbf{x}}+\beta\hat{\mathbf{y}}+\gamma\hat{\mathbf{z}}$. Thus, vectors in three-dimensional space, notated $\mathbb{R}^3$, are synonymous with triplets (α,β,γ) of real numbers. With some clever geometry, we deduce

$$\|\alpha\hat{\mathbf{x}}+\beta\hat{\mathbf{y}}+\gamma\hat{\mathbf{z}}\| = \sqrt{\alpha^2+\beta^2+\gamma^2}.$$

Historically, three-dimensional space has been studied a lot and there are several notations for the standard basis vectors still in use.

The following is a non-exhaustive list.

$\hat{\mathbf{x}}$	$\hat{\mathbf{y}}$	$\hat{\mathbf{z}}$
$\hat{\imath}$	$\hat{\jmath}$	$\hat{k}$
$\mathbf{i}$	$\mathbf{j}$	$\mathbf{k}$
$\vec{e}_1$	$\vec{e}_2$	$\vec{e}_3$

[8] Image credit: Acdx, from Wikipedia https://en.wikipedia.org/wiki/Cross_product

Keep these notations in the back of your mind. You might see them in other classes.

Higher dimensions

One can't progress very far in the study of science and mathematics without encountering a need for higher dimensional "vectors." For example, physicists have known since Einstein that the physical universe is best thought of as a four-dimensional entity called spacetime in which time plays a role close to that of the three spatial coordinates. Since we don't have any way to deal with $\mathbb{R}^n$ intuitively, we must proceed by analogy with two and three dimensions. The easiest way to proceed is to generalize the idea of a standard basis. From there, we can represent vectors in $\mathbb{R}^n$ as n-tuples of real numbers. We then define

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

We've now unified our theory of vectors across all integer dimensions $n > 0$. The case $n = 1$ yields "geometry" on a line, the cases $n = 2$ and $n = 3$ geometry in the plane and in space, and the case $n = 4$ yields the geometry of "4-vectors" which are used in the special theory of relativity. Larger values of n are used in a variety of contexts, some of which we shall encounter later.

Exercises for 2.3

1. Find $||a||$, $5\vec{a} - 2\vec{b}$, and $-3\vec{b}$ for each of the following vector pairs.
   a) $\vec{a} = 2\hat{\mathbf{x}} + 3\hat{\mathbf{y}}$, $\vec{b} = 4\hat{\mathbf{x}} - 9\hat{\mathbf{y}}$
   b) $\vec{a} = (1, 2, -1)$, $\vec{b} = (2, -1, 0)$
2. Let $P = (7, 2, 9)$ and $Q = (-2, 1, 4)$. Find $\overrightarrow{PQ}$ as a linear combination of $\hat{\mathbf{x}}$, $\hat{\mathbf{y}}$, and $\hat{\mathbf{z}}$.

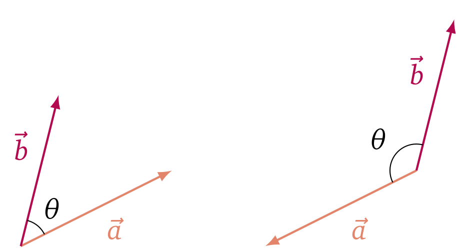
2.4 Dot Products & Projections

Dot Product

Let $\vec{a}$ and $\vec{b}$ be vectors. We assume they are placed so their tails coincide. Let θ denote the *smaller* of the two angles between them, so $0 \le \theta \le \pi$. The *dot product* of $\vec{a}$ and $\vec{b}$ is defined to be

$$\vec{a} \cdot \vec{b} = ||\vec{a}||||\vec{b}|| \cos\theta.$$

We will call this the *geometric definition of the dot product*. The dot product is also sometimes called the *scalar product* because the result is a scalar. Note that $\vec{a} \cdot \vec{b} = 0$ when either $\vec{a}$ or $\vec{b}$ is zero or, more interestingly, if their directions are perpendicular.



Algebraically, we can define the dot product in terms of components:

$$\begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix} \cdot \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix} = a_1 b_1 + a_2 b_2 + \cdots + a_n b_n.$$

We will call this the *algebraic definition of the dot product*[9].

By switching between algebraic and geometric definitions, we can use the dot product to find quantities that are otherwise difficult to find.

■ **Example 2.3** Find the angle between the vectors $\vec{v} = (1,2,3)$ and $\vec{w} = (1,1,-2)$.

From the algebraic definition of the dot product, we know

$$\vec{v} \cdot \vec{w} = 1(1) + 2(1) + 3(-2) = -3.$$

From the geometric definition, we know

$$\vec{v} \cdot \vec{w} = \|\vec{v}\|\|\vec{w}\| \cos\theta = \sqrt{14}\sqrt{6}\cos\theta = \sqrt{21}\cos\theta.$$

Equating the two definitions of $\vec{v} \cdot \vec{w}$, we see

$$\cos\theta = \frac{-3}{\sqrt{21}}$$

and so $\theta = \arccos(-3/\sqrt{21})$. ■

Recall that for vectors $\vec{a}$ and $\vec{b}$, the relationship $\vec{a} \cdot \vec{b} = 0$ can hold for two reasons: (i) either $\vec{a} = \vec{0}$, $\vec{b} = \vec{0}$, or both or (ii) $\vec{a}$ and $\vec{b}$ meet at 90°. Thus, the dot product can be used to tell if two vectors are perpendicular. There is some strangeness with the zero vector here, but it turns out this strangeness simplifies our lives mathematically.

Definition 2.4.1 — Orthogonal. The vectors $\vec{u}$ and $\vec{v}$ are *orthogonal* if $\vec{u} \cdot \vec{v} = 0$.

The definition of orthogonal encapsulates both the idea of two vectors being perpendicular and the idea of one of them being $\vec{0}$.

Before we continue, let's pin down the idea of one vector pointing in the *direction* of another. There are many ways we could define this idea, but we'll go with this one.

Definition 2.4.2 The vector $\vec{u}$ points in the *direction* of the vector $\vec{v}$ if $k\vec{u} = \vec{v}$ for some scalar k.

A simple example is that $2\hat{\mathbf{x}}$ points in the direction of $\hat{\mathbf{x}}$ since $\frac{1}{2}(2\hat{\mathbf{x}}) = \hat{\mathbf{x}}$. However, nothing in the definition says the scalar needs to be positive, so $-\hat{\mathbf{x}}$ also points in the direction $\hat{\mathbf{x}}$.

[9] Philosophically, every object should have only one definition from which equivalent characterizations can be deduced as theorems. If you're bothered, pick your favorite definition to be the "true" definition and consider the other definition a theorem.

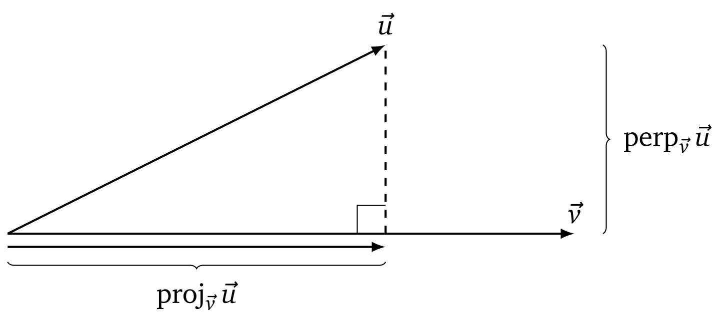
Projection

Another common vector operation is *projection*. Projection measures how much a vector points in the direction of another. This quantity is encoded as a vector. We make this definition mathematically precise as follows.

Definition 2.4.3 — Projection. For a vector $\vec{u}$ and a non-zero vector $\vec{v}$, the *projection* of $\vec{u}$ onto $\vec{v}$ is written as $\text{proj}_{\vec{v}}\,\vec{u}$ and is a vector in the direction of $\vec{v}$ with the property that $\vec{u}-\text{proj}_{\vec{v}}\,\vec{u}$ is orthogonal to $\vec{v}$.

The vector $\vec{u}-\text{proj}_{\vec{v}}\,\vec{u}$ is called the *perpendicular component* of the projection of $\vec{u}$ onto $\vec{v}$ and is notated $\text{perp}_{\vec{v}}\,\vec{u}$.

We can visualize projections with the following diagram.



From the picture, it appears that $\vec{u}$, $\text{proj}_{\vec{v}}\,\vec{u}$, and $\text{perp}_{\vec{v}}\,\vec{u}$ form a right triangle. Of course, we shouldn't trust the picture. We should verify this mathematically.

Theorem 2.4.1 If $\vec{u}$ and $\vec{v}$ are non-zero vectors, then $\vec{v}$, $\text{proj}_{\vec{v}}\,\vec{u}$, and $\text{perp}_{\vec{v}}\,\vec{u}$ form a (possibly degenerate) right triangle.

Proof. We need to verify that the sides $\text{proj}_{\vec{v}}\,\vec{u}$ and $\text{perp}_{\vec{v}}\,\vec{u}$ meet at a right angle and that the hypotenuse $\vec{u}$ meets the sides. That is, $\text{perp}_{\vec{v}}\,\vec{u}+\text{proj}_{\vec{v}}\,\vec{u}=\vec{u}$.

By the definition of projection, $\text{perp}_{\vec{v}}\,\vec{u}=\vec{u}-\text{proj}_{\vec{v}}\,\vec{u}$ is orthogonal to $\vec{v}$. Since $\text{proj}_{\vec{v}}\,\vec{u}$ points in the direction of $\vec{v}$, we have $\text{proj}_{\vec{v}}\,\vec{u}=k\vec{v}$ and so $\text{perp}_{\vec{v}}\,\vec{u}$ is orthogonal to $\text{proj}_{\vec{v}}\,\vec{u}$.

Finally, consider

$$\text{perp}_{\vec{v}}\,\vec{u}+\text{proj}_{\vec{v}}\,\vec{u}=(\vec{u}-\text{proj}_{\vec{v}}\,\vec{u})+\text{proj}_{\vec{v}}\,\vec{u}=\vec{u},$$

so indeed the vectors form a right triangle. ■

Now that we've proved $\vec{u}$, $\text{proj}_{\vec{v}}\,\vec{u}$, and $\text{perp}_{\vec{v}}\,\vec{u}$ form a right triangle, we are free to use trigonometry to compute projections. If θ is the angle between $\vec{u}$ and $\vec{v}$ and $0\le\theta\le\pi/2$, we know $\|\text{proj}_{\vec{v}}\,\vec{u}\|=\|\vec{u}\|\cos\theta$. This means

$$\text{proj}_{\vec{v}}\,\vec{u}=k\vec{v}=\|\vec{u}\|\cos\theta\,\frac{\vec{v}}{\|\vec{v}\|}$$

(Recall that $\vec{v}/||\vec{v}||$ is a unit vector in the direction of $\vec{v}$). But $\cos\theta$ appears in the formula for the dot product. Solving for $\cos\theta$ in the dot product formula, we see $\cos\theta = \frac{\vec{u}\cdot\vec{v}}{||\vec{u}||||\vec{v}||}$. Thus,

$$\text{proj}_{\vec{v}}\,\vec{u} = ||\vec{u}||\cos\theta\,\vec{v} = \frac{\vec{u}\cdot\vec{v}}{||\vec{u}||||\vec{v}||}\left(\frac{\vec{v}}{||\vec{v}||}\right) = \frac{\vec{u}\cdot\vec{v}}{||\vec{v}||^2}\vec{v}.$$

Upon close inspection, we see $||\vec{v}||^2 = \vec{v}\cdot\vec{v}$ (since $\cos 0 = 1$) and so we finally arrive at the formula

$$\text{proj}_{\vec{v}}\,\vec{u} = \frac{\vec{u}\cdot\vec{v}}{\vec{v}\cdot\vec{v}}\vec{v}$$

Incredibly, if we use the algebraic definition of the dot product, we can compute a projection without computing cosine of anything!

Exercises for 2.4

2.5 The Cross Product

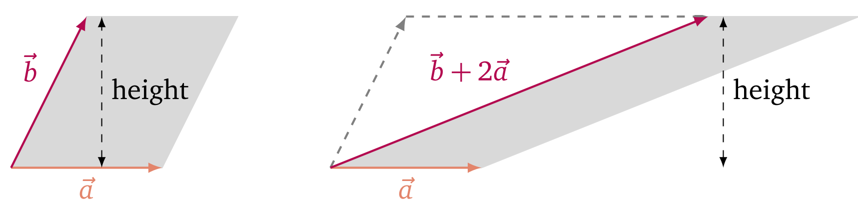
For vectors $\vec{a}$ and $\vec{b}$, the dot product $\vec{a}\cdot\vec{b}$ measures how close $\vec{a}$ and $\vec{b}$ are to being orthogonal. In contrast, the *cross product* of $\vec{a}$ and $\vec{b}$, written $\vec{a}\times\vec{b}$ will measure the *area* of the parallelogram whose sides are given by $\vec{a}$ and $\vec{b}$.

Let's explore this idea. Since the cross product is a *product*, we will demand it follow reasonable distribution laws[10]:

$$\begin{aligned}
\vec{a}\times(\vec{b}+\vec{c}) &= \vec{a}\times\vec{b}+\vec{a}\times\vec{c}\\
(\vec{a}+\vec{b})\times\vec{c} &= \vec{a}\times\vec{c}+\vec{b}\times\vec{c}\\
(\alpha\vec{a})\times\vec{b} &= \alpha(\vec{a}\times\vec{b})\\
\vec{a}\times(\alpha\vec{b}) &= \alpha(\vec{a}\times\vec{b})
\end{aligned}$$

for vectors $\vec{a}$, $\vec{b}$, $\vec{c}$ and scalars α.

Now, suppose $\vec{a}\times\vec{b}$ indeed encapsulates the area of the parallelogram with sides $\vec{a}$ and $\vec{b}$. If we slide the tip of $\vec{b}$ parallel to the vector $\vec{a}$, we should not change the area. Thus the cross product of $\vec{a}$ and $\vec{b}$ should be the same as that of $\vec{a}$ and $\vec{b}+\alpha\vec{a}$.



Using this invariance along with our distributive rules, we now see

$$\vec{a}\times\vec{b} = \vec{a}\times(\vec{b}+\alpha\vec{a}) = \vec{a}\times\vec{b}+\alpha(\vec{a}\times\vec{a}),$$

[10] The technical term for satisfying these laws is *bilinearity*.

and so $\vec{a} \times \vec{a} = 0$. We can apply this newly-found fact to the vector $\vec{a} + \vec{b}$ to deduce

$$\begin{aligned} 0 = (\vec{a} + \vec{b}) \times (\vec{a} + \vec{b}) &= \vec{a} \times \vec{a} + \vec{a} \times \vec{b} + \vec{b} \times \vec{a} + \vec{b} \times \vec{b} \\ &= 0 + \vec{a} \times \vec{b} + \vec{b} \times \vec{a} + 0 \\ &= \vec{a} \times \vec{b} + \vec{b} \times \vec{a}, \end{aligned}$$

and so

$$\vec{a} \times \vec{b} = -\vec{b} \times \vec{a}.$$

Products with this property are called *anti-commutative*. Now for an incredible fact of the universe: the result of the cross product of two vectors in $\mathbb{R}^3$ can be represented by another vector in $\mathbb{R}^3$ whose magnitude corresponds to the area of the parallelogram with sides $\vec{a}$ and $\vec{b}$.[11] Using trigonometry, we deduce

$$\|\vec{a} \times \vec{b}\| = \|\vec{a}\|\|\vec{b}\| \sin\theta$$

where $0 \le \theta \le \pi$ is smaller of the two angles between $\vec{a}$ and $\vec{b}$. What remains to be seen is what direction $\vec{a} \times \vec{b}$ points in. For this, we use the standard basis for $\mathbb{R}^3$ as a launching point. Recall $\hat{\mathbf{x}}$, $\hat{\mathbf{y}}$, and $\hat{\mathbf{z}}$ are all unit vectors and all orthogonal to each other. Thus, any of them cross each other must result in a unit vector. By convention,

$$\begin{aligned} \hat{\mathbf{x}} \times \hat{\mathbf{y}} &= \hat{\mathbf{z}}, \\ \hat{\mathbf{y}} \times \hat{\mathbf{z}} &= \hat{\mathbf{x}} \\ \hat{\mathbf{z}} \times \hat{\mathbf{x}} &= \hat{\mathbf{y}}. \end{aligned}$$

Let $\vec{a} = a_x\hat{\mathbf{x}} + a_y\hat{\mathbf{y}} + a_z\hat{\mathbf{z}}$ and $\vec{b} = b_x\hat{\mathbf{x}} + b_y\hat{\mathbf{y}} + b_z\hat{\mathbf{z}}$. Using the distributive laws of the cross product we see,

$$\begin{aligned} \vec{a} \times \vec{b} = (a_x\hat{\mathbf{x}} + a_y\hat{\mathbf{y}} + a_z\hat{\mathbf{z}}) \times (b_x\hat{\mathbf{x}} + b_y\hat{\mathbf{y}} + b_z\hat{\mathbf{z}}) \\ = \quad & a_x b_x \hat{\mathbf{x}} \times \hat{\mathbf{x}} + a_x b_y \hat{\mathbf{x}} \times \hat{\mathbf{y}} + a_x b_z \hat{\mathbf{x}} \times \hat{\mathbf{z}} \\ & + a_y b_x \hat{\mathbf{y}} \times \hat{\mathbf{x}} + a_y b_y \hat{\mathbf{y}} \times \hat{\mathbf{y}} + a_y b_z \hat{\mathbf{y}} \times \hat{\mathbf{z}} \\ & + a_z b_x \hat{\mathbf{z}} \times \hat{\mathbf{x}} + a_z b_y \hat{\mathbf{z}} \times \hat{\mathbf{y}} + a_z b_z \hat{\mathbf{z}} \times \hat{\mathbf{z}} \\ = \quad & \vec{0} + a_x b_y \hat{\mathbf{z}} - a_x b_z \hat{\mathbf{y}} \\ & - a_y b_x \hat{\mathbf{z}} + \vec{0} + a_y b_z \hat{\mathbf{x}} \\ & + a_z b_x \hat{\mathbf{y}} - a_z b_y \hat{\mathbf{x}} + \vec{0} \end{aligned}$$

so

$$\vec{a} \times \vec{b} = (a_y b_z - a_z b_y)\hat{\mathbf{x}} - (a_x b_z - a_z b_x)\hat{\mathbf{y}} + (a_x b_y - a_y b_x)\hat{\mathbf{z}}.$$

[11] This is *only* true in $\mathbb{R}^3$. In $\mathbb{R}^4$ a product that produces area-like quantities does exist, but the output cannot be described by a vector. In higher dimensions, the cross product is called the *wedge product*.

Exercise 2.2 Verify that $\|\vec{a}\times\vec{b}\| = \|\vec{a}\|\|\vec{b}\|\sin\theta$. (Hint: you can use $\vec{a}\cdot\vec{b} = \|\vec{a}\|\|\vec{b}\|\cos\theta$ to solve for θ and the proceed using components.)

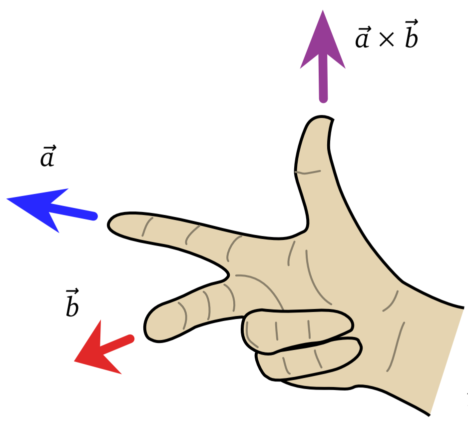
Now that we know what the cross product is and how to compute it, let's explore some of its incredible properties. First,

$$(\vec{a}\times\vec{b})\cdot\vec{a} = (a_y b_z - a_z b_y)a_x - (a_x b_z - a_z b_x)a_y + (a_x b_y - a_y b_x)a_z = 0$$

and

$$(\vec{a}\times\vec{b})\cdot\vec{b} = (a_y b_z - a_z b_y)b_x - (a_x b_z - a_z b_x)b_y + (a_x b_y - a_y b_x)b_z = 0.$$

Thus, $\vec{a}\times\vec{b}$ is orthogonal to both $\vec{a}$ and $\vec{b}$. Just based on this property, since the length of $\vec{a}\times\vec{b}$ is fixed, $\vec{a}\times\vec{b}$ can be one of two vectors in space. If we investigate further, we'll see that $\vec{a}\times\vec{b}$ is the vector that satisfies the *right-hand rule*.



[12]

A vector that encodes area, points orthogonally to others, and obeys the right-hand rule is handy indeed, and the cross product will be a useful tool for solving many problems.

Exercises for 2.5

2.6 Lines and Planes

With a handle on vectors, we can now use them to describe some common geometric objects: lines and planes.

Lines

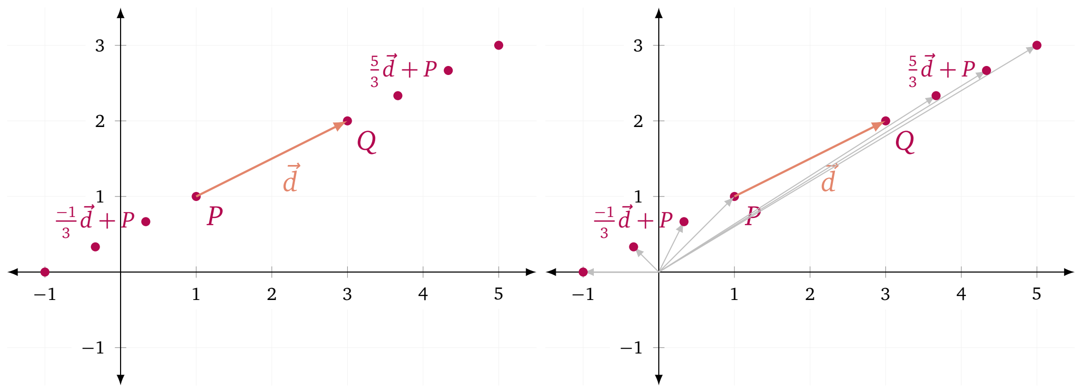
Consider for a moment the line ℓ through the points P and Q. If $P,Q\in\mathbb{R}^2$, we could describe this line in $y = mx + b$ form (provided it isn't a vertical line), but if $P,Q\in\mathbb{R}^3$ it's much harder to describe ℓ with an equation. Using vectors provides an easier way.

Let $\vec{d} = \overrightarrow{PQ}$ and consider the set of points (or vectors)

$$\vec{x} = t\vec{d} + P$$

[12] Image credit: Acdx, from Wikipedia https://en.wikipedia.org/wiki/Cross_product

for $t \in \mathbb{R}$. Geometrically, this is the set of all points we get by starting at P and displacing by some multiple of the direction $\vec{d}$. This is a line!



Note that sometimes when we draw pictures of vectors, drawing them as line segments is illuminating. Sometimes, however, drawing them as line segments can make it hard to see what's going on, and it is better to draw each vector as a dot.

Call this line ℓ. In set-builder notation, we would write

$$\ell = \{\vec{x} : \vec{x} = t\vec{d} + P \text{ for some } t \in \mathbb{R}\}.$$

Notice that in set-builder notation we write "for some $t \in \mathbb{R}$." Make sure you understand why replacing "for some $t \in \mathbb{R}$" with "for all $t \in \mathbb{R}$" would be incorrect.

Writing lines with set-builder notation all the time can be overkill, so we will allow ourselves to describe lines in a shorthand called *vector form*[13].

Definition 2.6.1 — Vector form of a Line. A line ℓ is described in *vector form* if there are two vectors $\vec{d} \neq \vec{0}$ and $\vec{p}$ so that

$$\vec{x} = t\vec{d} + \vec{p}$$

satisfies $\vec{x} \in \ell$ for all $t \in \mathbb{R}$. In this case we call $\vec{d}$ the *direction* of ℓ and the equation $\vec{x} = t\vec{d} + \vec{p}$ the *vector equation* or *vector form* of ℓ.

Note that if $\vec{x} = t\vec{d} + \vec{p}$ is the vector equation of a line ℓ, by setting $t = 0$ we necessarily have $\vec{p} \in \ell$.

The direction of a line is easily obtained by finding the displacement vector between two points on the line. Thus, given a line in another form, computing its vector form is straightforward.

■ **Example 2.4** Find vector form of the line ℓ in $\mathbb{R}^2$ with equation $y = 2x + 3$. First, we find two points on the line. By guess-and-check we see $P = (0,3)$ and $Q = (1,5)$ are on ℓ. Thus, a direction vector for ℓ is given by

$$\vec{d} = (1,5) - (0,3) = (1,2).$$

[13] $y = mx + b$ form of a line is also shorthand. The line ℓ described by the equation $y = mx + b$ is actually the set $\{(x,y) \in \mathbb{R}^2 : y = mx + b\}$.

We may now write the vector equation of ℓ as

$$\vec{x} = t\vec{d} + P$$

or, in components,

$$\begin{bmatrix} x \\ y \end{bmatrix} = t\begin{bmatrix} 1 \\ 2 \end{bmatrix} + \begin{bmatrix} 0 \\ 3 \end{bmatrix}.$$

■

The downside of writing lines in vector form is that there are multiple direction vectors and multiple points for every line. Thus, merely by looking at the vector equation for two lines, it can be hard to tell if they're equal.

For example,

$$\begin{bmatrix} x \\ y \end{bmatrix} = t\begin{bmatrix} 1 \\ 2 \end{bmatrix} + \begin{bmatrix} 0 \\ 3 \end{bmatrix}, \qquad \begin{bmatrix} x \\ y \end{bmatrix} = t\begin{bmatrix} 2 \\ 4 \end{bmatrix} + \begin{bmatrix} 0 \\ 3 \end{bmatrix}, \quad \text{and} \quad \begin{bmatrix} x \\ y \end{bmatrix} = t\begin{bmatrix} 1 \\ 2 \end{bmatrix} + \begin{bmatrix} 1 \\ 5 \end{bmatrix}$$

all represent the same line. In the second equation, the direction is parallel but scaled, and in the third equation, a different point on the line was chosen.

In vector form, the variable t is called the *parameter variable*. It is an instance of a *dummy variable*; that is, it is mostly there as a placeholder. Remember, vector form is shorthand for set-builder notation.

Let $\vec{d}_1, \vec{d}_2 \neq \vec{0}$ and $\vec{p}_1, \vec{p}_2$ be vectors and define the lines

$$\ell_1 = \{\vec{x} \,:\, \vec{x} = t\vec{d}_1 + \vec{p}_1 \text{ for some } t \in \mathbb{R}\}$$

$$\ell_2 = \{\vec{x} \,:\, \vec{x} = t\vec{d}_2 + \vec{p}_2 \text{ for some } t \in \mathbb{R}\}.$$

These lines have vector equations $\vec{x} = t\vec{d}_1 + \vec{p}_1$ and $\vec{x} = t\vec{d}_2 + \vec{p}_2$. However, declaring that $\ell_1 = \ell_2$ if and only if $t\vec{d}_1 + \vec{p}_1 = t\vec{d}_2 + \vec{p}_2$ does *not* make sense. Instead $\ell_1 = \ell_2$ if $\ell_1 \subseteq \ell_2$ and $\ell_2 \subseteq \ell_1$. If $\vec{x} \in \ell_1$ then $\vec{x} = t\vec{d}_1 + \vec{p}_1$ for some $t \in \mathbb{R}$. If $\vec{x} \in \ell_2$ then $\vec{x} = t\vec{d}_2 + \vec{p}_2$ for some *possibly different* $t \in \mathbb{R}$. This can get confusing really quickly. The easiest solution is to use different parameter variables if we want to compare lines in vector form.

■ **Example 2.5** Determine if the lines ℓ_1 and ℓ_2, represented in vector form by the equations

$$\vec{x} = t\begin{bmatrix} 1 \\ 1 \end{bmatrix} + \begin{bmatrix} 2 \\ 1 \end{bmatrix} \qquad \text{and} \qquad \vec{x} = t\begin{bmatrix} 2 \\ 2 \end{bmatrix} + \begin{bmatrix} 4 \\ 3 \end{bmatrix}$$

are the same line. To determine this, we need to figure out if $\vec{x} \in \ell_1$ implies $\vec{x} \in \ell_2$ and if $\vec{x} \in \ell_2$ implies $\vec{x} \in \ell_1$.

If $\vec{x} \in \ell_1$, then $\vec{x} = t\begin{bmatrix} 1 \\ 1 \end{bmatrix} + \begin{bmatrix} 2 \\ 1 \end{bmatrix}$ for some $t \in \mathbb{R}$. If $\vec{x} \in \ell_2$, then $\vec{x} = s\begin{bmatrix} 2 \\ 2 \end{bmatrix} + \begin{bmatrix} 4 \\ 3 \end{bmatrix}$ for some $s \in \mathbb{R}$. Thus if

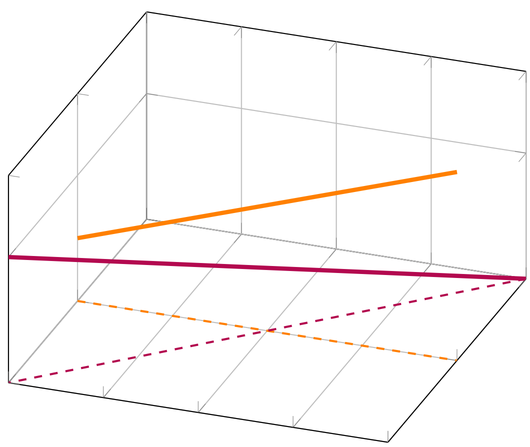
$$t\begin{bmatrix} 1 \\ 1 \end{bmatrix} + \begin{bmatrix} 2 \\ 1 \end{bmatrix} = \vec{x} = s\begin{bmatrix} 2 \\ 2 \end{bmatrix} + \begin{bmatrix} 4 \\ 3 \end{bmatrix}$$

always has a solution, $\ell_1 = \ell_2$. Moving everything to one side we see

$$\begin{aligned}\vec{0} = \begin{bmatrix}4\\3\end{bmatrix} - \begin{bmatrix}2\\1\end{bmatrix} + s\begin{bmatrix}2\\2\end{bmatrix} - t\begin{bmatrix}1\\1\end{bmatrix} &= \begin{bmatrix}2\\2\end{bmatrix} + s\begin{bmatrix}2\\2\end{bmatrix} - t\begin{bmatrix}1\\1\end{bmatrix}\\ &= (s+1)\begin{bmatrix}2\\2\end{bmatrix} - \tfrac{t}{2}\begin{bmatrix}2\\2\end{bmatrix}\\ &= (s+1-\tfrac{t}{2})\begin{bmatrix}2\\2\end{bmatrix}.\end{aligned}$$

This has a solution whenever $0 = s+1-t/2$. Since for every $t \in \mathbb{R}$ we can find an $s \in \mathbb{R}$ and for every $s \in \mathbb{R}$ we can find a $t \in \mathbb{R}$ satisfying this equation, we know $\ell_1 = \ell_2$. ■

The geometry of lines in space is a bit more complicated than that of lines in the plane. Lines in the plane either intersect or are parallel. In space, we have to be a bit more careful about what we mean by "parallel lines," since lines with entirely different directions can still fail to intersect[14].



■ **Example 2.6** Consider the lines described by

$$\vec{x} = t(1,3,-2) + (1,2,1)$$
$$\vec{x} = t(-2,-6,4) + (3,1,0).$$

They have parallel directions since $(-2,-6,4) = -2(1,3,-2)$. Hence, in this case, we say the lines are *parallel*. (How can we be sure the lines are not the same?) ■

■ **Example 2.7** Consider the lines described by

$$\vec{x} = t(1,3,-2) + (1,2,1)$$
$$\vec{x} = t(0,2,3) + (0,3,9).$$

They are not parallel because neither of the direction vectors is a multiple of the other. They may or may not intersect. (If they don't, we say the lines are *skew*.) How can we find out?

[14] Recall that in Euclidean geometry two lines are defined to be parallel if they coincide or never intersect.

Mirroring our earlier approach, we can set their equations equal and see if we can solve for the point of intersection *after ensuring we give their parametric variables different names*. We'll keep one parametric variable named t and name the other one s. Thus, we want

$$\vec{x} = t(1,3,-2)+(1,2,1) = s(0,2,3)+(0,3,9),$$

which after collecting terms yields

$$(t+1, 3t+2, -2t+1) = (0, 2s+3, 3s+9).$$

Picking out the components yields three equations

$$\begin{aligned} t+1 &= 0 \\ 3t+2 &= 2s+3 \\ -2t+1 &= 3s+9 \end{aligned}$$

in 2 unknowns s and t. This is an *overdetermined* system, and it may or may not have a consistent solution. The first two equations yield $t=-1$ and $s=-2$. Putting these values in the last equation yields $(-2)(-1)+1 = 3(-2)+9$, which is indeed true. Hence, the equations are consistent, and the lines intersect. To find the point of intersection, put $t=-1$ in the equation for the first line (or $s=-2$ in that for the second) to obtain $(0,-1,3)$. ■

Planes

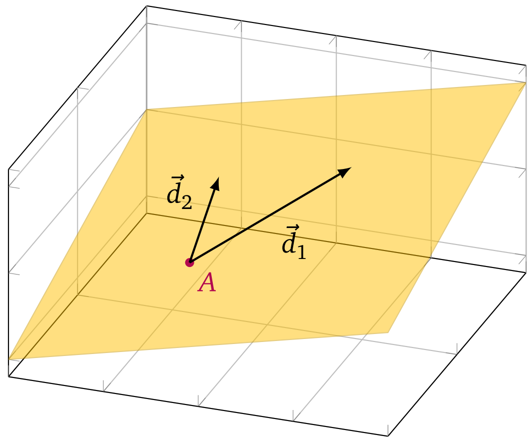
Any two distinct points define a line. To define a plane, we need three points. But there's a caveat: the three points cannot be on the same line, otherwise they'd define a line and not a plane. Let $A, B, C \in \mathbb{R}^3$ be three points that are not collinear and let $\mathcal{P}$ be the plane that passes through A, B, and C.

Just like lines, planes have direction vectors. For $\mathcal{P}$, both $\vec{d}_1 = \overrightarrow{AB}$ and $\vec{d}_2 = \overrightarrow{AC}$ are direction vectors for $\mathcal{P}$. Of course, $\vec{d}_1$, $\vec{d}_2$ and their multiples are not the only direction vectors for $\mathcal{P}$. There are infinitely many more, including $\vec{d}_1 + \vec{d}_2$, and $\vec{d}_1 - 7\vec{d}_2$, and so on. However, since a plane is a *two*-dimensional object, we only need two different direction vectors to describe it.

Again like lines, planes have a vector form. $\mathcal{P}$ can be written in vector form as

$$\begin{bmatrix} x \\ y \\ z \end{bmatrix} = t\vec{d}_1 + s\vec{d}_2 + A.$$

Vector form of $\mathcal{P}$ is not unique. Any two different directions in $\mathcal{P}$ suffice for defining $\mathcal{P}$ in vector form.

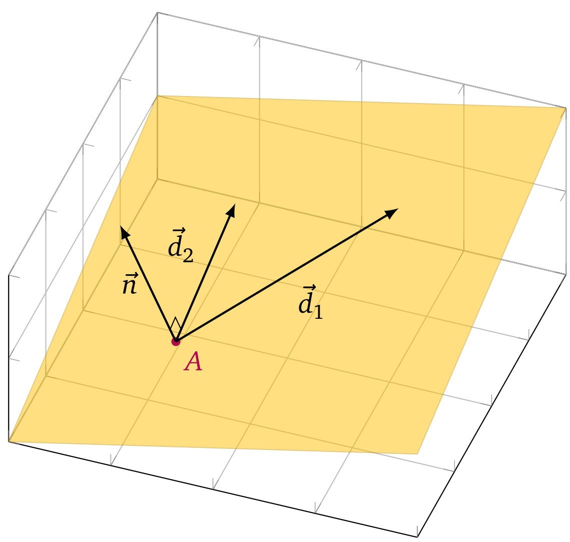


Definition 2.6.2 — Vector form of a plane. The plane $\mathcal{P}$ is described in *vector form* if there are three vectors $\vec{d}_1$, $\vec{d}_2$, and $\vec{p}$ where $\vec{d}_1, \vec{d}_2 \neq \vec{0}$ point in different directions and

$$\vec{x} = t\vec{d}_1 + s\vec{d}_2 + \vec{p}$$

satisfies $\vec{x} \in \mathcal{P}$ for all scalars $t, s \in \mathbb{R}$. The vectors $\vec{d}_1$ and $\vec{d}_2$ are called *direction vectors* for the plane $\mathcal{P}$.

Since we will commonly be working in $\mathbb{R}^3$ there is another way to define a plane. Given any vector $\vec{n} \in \mathbb{R}^3$, we can consider the set $\mathcal{Q} \subseteq \mathbb{R}^3$ of vectors orthogonal to $\vec{n}$. If $\vec{n} = \vec{0}$, then $\mathcal{Q} = \mathbb{R}^3$. Otherwise, $\mathcal{Q}$ is a plane through the origin. In this case, $\vec{n}$ is called the *normal vector* of the plane $\mathcal{Q}$.



Definition 2.6.3 — Normal form of a plane. The plane $\mathcal{P}$ is described in *normal form* if for

some $\vec{n}$ and $\vec{p}$, the equation

$$\vec{n}\cdot(\vec{x}-\vec{p})=0$$

if and only if $\vec{x}\in\mathcal{P}$. Equivalently, $\mathcal{P}$ is described in normal form if for some $\vec{n}$ and scalar $\alpha\in\mathbb{R}$ the equation

$$\vec{n}\cdot\vec{x}=\alpha$$

is satisfied if and only if $\vec{x}\in\mathcal{P}$. In either case, the vector $\vec{n}$ is call a *normal vector* for $\mathcal{P}$.

Normal form of a plane only exists in $\mathbb{R}^3$, but it is often useful[15]. The equivalence of the two ways to write a normal form of a plane is straight forward.

$$\vec{n}\cdot(\vec{x}-\vec{p})=0$$

if and only if

$$\vec{n}\cdot\vec{x}=\vec{n}\cdot\vec{p}=\alpha.$$

Since $\vec{n}$ and $\vec{p}$ are fixed, α is a constant. Expanding normal form in terms of components we see

$$\vec{n}\cdot(\vec{x}-\vec{p})=\vec{n}\cdot\vec{x}-\alpha=n_xx+n_yy+n_zz-\alpha=0$$

and so

$$n_xx+n_yy+n_zz=\alpha \tag{2.1}$$

is another way to write a plane. Equation (2.1) is sometimes called *scalar form* of a plane. For us, it will not be important to distinguish between scalar and normal form.

It should be noted that like vector form of a plane, normal form of a plane is not unique. For example, the plane described by $\vec{n}\cdot(\vec{x}-\vec{p})=0$ is the same as the plane $(2\vec{n})\cdot(\vec{x}-\vec{p})=0$.

■ **Example 2.8** Find vector form and normal form of the plane $\mathcal{P}$ passing through the point $A=(1,0,0)$, $B=(0,1,0)$ and $C=(0,0,1)$.

To find vector form of $\mathcal{P}$, we need a point on the plane and two direction vectors. We have three points on the plain, so we can obtain two direction vectors by subtracting these points in different ways. Let

$$\vec{d}_1=\overrightarrow{AB}=\begin{bmatrix}-1\\1\\0\end{bmatrix}\qquad \vec{d}_2=\overrightarrow{AC}=\begin{bmatrix}-1\\0\\1\end{bmatrix}.$$

Using the point A, we may now write vector form of $\mathcal{P}$ as

$$\begin{bmatrix}x\\y\\z\end{bmatrix}=t\begin{bmatrix}-1\\1\\0\end{bmatrix}+s\begin{bmatrix}-1\\0\\1\end{bmatrix}+\begin{bmatrix}1\\0\\0\end{bmatrix}.$$

To write normal form we need to find a normal vector to $\mathcal{P}$. By symmetry, we can see that $\vec{n}=(1,1,1)$ is a normal vector to $\mathcal{P}$. If we weren't so insightful, we could also compute $\vec{d}_1\times\vec{d}_2=(1,1,1)$ to find a normal vector. Now, we may express $\mathcal{P}$ in normal form as

$$\begin{bmatrix}1\\1\\1\end{bmatrix}\cdot\left(\begin{bmatrix}x\\y\\z\end{bmatrix}-\begin{bmatrix}1\\0\\0\end{bmatrix}\right)=0$$

[15] Just like $y=mx+b$ form of a line only exists in $\mathbb{R}^2$.

or equivalently,

$$x+y+z=1.$$

■

■ **Example 2.9** Find the line $\mathcal{P}_1 \cap \mathcal{P}_2$ where $\mathcal{P}_1$ is the plane given by the equation

$$x+y+z=2$$

and $\mathcal{P}_2$ is the plane given by the equation

$$2x-y+z=0.$$

Let $\ell = \mathcal{P}_1 \cap \mathcal{P}_2$. Since $\ell \subseteq \mathcal{P}_1$ and $\ell \subseteq \mathcal{P}_2$, every direction vector for ℓ is also a direction vector for $\mathcal{P}_1$ and $\mathcal{P}_2$.

Let $\vec{n}_1 = (1,1,1)$ be a normal vector for $\mathcal{P}_2$ and $\vec{n}_2 = (2,-1,1)$ be a normal vector for $\mathcal{P}_2$. If $\vec{d}$ is a direction vector for ℓ, then $\vec{n}_1 \cdot \vec{d} = 0$ and $\vec{n}_2 \cdot \vec{d} = 0$. Thus,

$$\vec{d} = \vec{n}_1 \times \vec{n}_2 = \begin{bmatrix} 2 \\ 1 \\ -3 \end{bmatrix}$$

is a direction vector for ℓ. By guess and check we find that $\vec{p} = (0,1,1)$ satisfies $\vec{p} \in \mathcal{P}_1$ and $\vec{p} \in \mathcal{P}_2$ and so $\vec{p} \in \ell$. Thus, we may write ℓ in vector form as

$$\begin{bmatrix} x \\ y \\ z \end{bmatrix} = t \begin{bmatrix} 2 \\ 1 \\ -3 \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix}.$$

■

Exercises for 2.6

Chapter 3

Parameterization

Parameterization is a mouthful, but the fundamental idea of a parameterization is to describe one object in terms of another. For example, consider the line ℓ described by the equation $y = 2x$. By its nature, ℓ is a set. Using set-builder notation, we could write

$$\ell = \left\{ \begin{bmatrix} x \\ y \end{bmatrix} \ : \ y = 2x \right\}.$$

But, we could also write ℓ in vector form as

$$\begin{bmatrix} x \\ y \end{bmatrix} = t \begin{bmatrix} 1 \\ 2 \end{bmatrix}.$$

Writing ℓ in vector form shows a pairing between scalars $t \in \mathbb{R}$ and points on ℓ. In many ways, ℓ is the same as $\mathbb{R}$—it's just sitting in two-dimensional space instead of being on its own.

Taking a more technical viewpoint, we may consider ℓ to be the range of a vector-valued function. Define $\vec{p}(t) = t \begin{bmatrix} 1 \\ 2 \end{bmatrix}$. Then,

$$\ell = \mathrm{range}(\vec{p}) = \{\vec{x} \ : \ \vec{x} = \vec{p}(t) \text{ for some } t \in \mathbb{R}\}.$$

Now we have something special. The function $\vec{p} : \mathbb{R} \to \mathbb{R}^2$ has domain $\mathbb{R}$ and outputs every point on the line ℓ exactly once. In other words, we've described ℓ in terms of $\mathbb{R}$ and $\vec{p}$. We could make a further assertion that *anything that you could learn by studying ℓ, you could learn by studying $\mathbb{R}$ and $\vec{p}$.*

However, there are other ways to create functions that describe ℓ. For example, consider $\vec{q} : \mathbb{R} \to \mathbb{R}^2$ where $\vec{q}(t) = 2t\vec{d}$. Again, $\ell = \mathrm{range}(\vec{q})$ and so everything we could possibly learn about ℓ, we could learn by studying $\mathbb{R}$ and $\vec{q}$. We call both $\vec{p}$ and $\vec{q}$ *parameterizations of ℓ by $\mathbb{R}$.*

Definition 3.0.1 — Parameterization. A *parameterization* of an object X by an object Y is a continuous function $p : Y \to X$ with the added conditions that p is one-to-one[a] and $\mathrm{range}(p) = Y$. In this case p is called a *parameterization* and Y is called the *parameter*.

[a] Sometimes we will drop the requirement that a parameterization be one-to-one, but for now we'll be strict about it.

This definition is fairly abstract, which will come in handy later. For now, we will think of X as being some curve in $\mathbb{R}^n$ and Y as being an interval of real numbers.

■ **Example 3.1 — A Circle.** Let $\mathcal{C} \subseteq \mathbb{R}^2$ be the unit circle centered at the origin. We can parameterize $\mathcal{C}$ by angles in $[0,2\pi)$. Consider the function $\vec{p} : [0,2\pi) \to \mathcal{C}$ defined by

$$\vec{p}(\theta) = \begin{bmatrix} \cos\theta \\ \sin\theta \end{bmatrix}.$$

Here, $\vec{p}$ traces out $\mathcal{C}$ starting at the point $(1,0)$ and moving counter clockwise as the parameter θ increases. ■

■ **Example 3.2 — A Circle Again.** Let $\mathcal{C} \subseteq \mathbb{R}^2$ be the unit circle centered at the origin. We will parameterize $\mathcal{C}$ by the interval $[0,1)$. Here we might imagine that our parameter $t \in [0,1)$ represents a point that is t-percentage around the circle.

Recall $\vec{p}(\theta) = \begin{bmatrix} \cos\theta \\ \sin\theta \end{bmatrix}$, which parameterizes $\mathcal{C}$ based on angles.

Now, consider the function $w(t) = 2\pi t$. w inputs numbers in $[0,1)$ and outputs angles in $[0,2\pi)$. We should now be able to use w to parameterize $\mathcal{C}$ in the desired way. After all, if we convert $[0,1)$ to $[0,2\pi)$ to $\mathcal{C}$, we win!

Let the parameterization $\vec{q} : [0,1) \to \mathcal{C}$ be defined as $\vec{q} = \vec{p} \circ w$. Explicitly,

$$\vec{q}(t) = \vec{p} \circ w(t) = \vec{p}(2\pi t) = \begin{bmatrix} \sin 2\pi t \\ \cos 2\pi t \end{bmatrix}.$$

■

Exercise 3.1 Parameterize the unit circle $\mathcal{C} \subseteq \mathbb{R}^2$ by the interval $[1/2,1)$.

Exercise 3.2 Let ℓ be the line segment connecting $(0,0)$ and $(1,1)$. Explain why $\vec{p} : [-1,1] \to \ell$ given by $p(t) = (t^2, t^2)$ is *not* a parameterization.

3.1 Speed and Velocity of a Parameterization

In our day-to-day life, almost without thinking, we make a comparison between real numbers and time. Time has a forwards and backwards, which we equate to the real number's increasing and decreasing. We might even say we parameterize *time* by the real numbers. Thus, if $\vec{p} : [a,b] \to \mathcal{S}$ is a parameterization of the curve $\mathcal{S}$ by the interval $[a,b]$, we could think of $\vec{p}$ as describing the motion of a particle—at time $t \in [a,b]$ the particle is at $\vec{p}(t)$.

Interpreting parameterizations in this way, the *speed* of a parameterization should be the rate of change of distance with respect to time and the *velocity* of a parameterization should be the rate of change of displacement with respect to time.

Suppose $\vec{p} : [a,b] \to \mathcal{S}$ is a parameterization of $\mathcal{S}$ and $t \in [a,b]$ represents time. The *displacement* of $\vec{p}$ from time t to time $t+\Delta t$ is $\vec{p}(t+\Delta t) - \vec{p}(t)$ and the change in *distance* is $\|\vec{p}(t+\Delta t) - \vec{p}(t)\|$. Thus, if Δt is small, the velocity at time t can be approximated by

$$\text{velocity}\,\vec{p}(t) \approx \frac{\vec{p}(t+\Delta t) - \vec{p}(t)}{\Delta t}$$

and the speed[1] by

$$\operatorname{speed} \vec{p}(t) \approx \frac{\|\vec{p}(t+\Delta t)-\vec{p}(t)\|}{|\Delta t|}.$$

Taking limits, we arrive at exact rates of change, which leads us to the following definitions.

Definition 3.1.1 — Speed. Let $\vec{p}:[a,b]\to\mathcal{S}$ be a parameterization of $\mathcal{S}$. The *speed* of $\vec{p}$ at the time $t\in[a,b]$ is

$$\operatorname{speed} \vec{p}(t) = \lim_{\Delta t\to 0} \frac{\|\vec{p}(t+\Delta t)-\vec{p}(t)\|}{|\Delta t|}.$$

Definition 3.1.2 — Velocity. Let $\vec{p}:[a,b]\to\mathcal{S}$ be a parameterization of $\mathcal{S}$. The *velocity* of $\vec{p}$ at the time $t\in[a,b]$ is

$$\operatorname{velocity} \vec{p}(t) = \lim_{\Delta t\to 0} \frac{\vec{p}(t+\Delta t)-\vec{p}(t)}{\Delta t}.$$

Both the definition of speed and the definition of velocity look a lot like the definition of the derivative. In fact, if $\vec{p}$ were a scalar valued function, the velocity of $\vec{p}$ would be exactly the derivative of $\vec{p}$. For this reason, we will define a notation similar to that of the derivative you're familiar with. From now on, the following notations mean the same thing:

$$\operatorname{velocity} \vec{p}(t) = \vec{p}\,'(t) = \frac{\mathrm{d}}{\mathrm{d}t}\vec{p}(t) = \frac{\mathrm{d}\vec{p}}{\mathrm{d}t}(t).$$

Let's try to use our new definition. Let $\vec{r}(t) = \begin{bmatrix}\cos t\\ \sin t\end{bmatrix}$. Now,

$$\begin{aligned}\operatorname{velocity} \vec{r}(t) &= \lim_{\Delta t\to 0} \frac{\begin{bmatrix}\cos(t+\Delta t)\\ \sin(t+\Delta t)\end{bmatrix} - \begin{bmatrix}\cos t\\ \sin t\end{bmatrix}}{\Delta t}\\ &= \lim_{\Delta t\to 0} \begin{bmatrix}\dfrac{\cos(t+\Delta t)-\cos t}{\Delta t}\\ \dfrac{\sin(t+\Delta t)-\sin t}{\Delta t}\end{bmatrix}.\end{aligned}$$

At this point, we should pause. We don't know how to take limits of vectors. Fortunately the rule is simple enough—to take a limit of a vector, take the limit of each of its components[2].

[1] Recall that speed is always positive; if a particle is moving with speed 2 and we then ran the particle back in time, it would still move at speed 2, so speed is not distance/Δt, it is distance/$|\Delta t|$.

[2] As intuitive as it sounds, this rule actually has a proof which relies on the definition of limit and the continuity of $\|\cdot\|$.

Thus we see

$$\begin{aligned}
\text{velocity}\,\vec{r}(t) &= \lim_{\Delta t\to 0}\begin{bmatrix} \dfrac{\cos(t+\Delta t)-\cos t}{\Delta t} \\ \dfrac{\sin(t+\Delta t)-\sin t}{\Delta t}\end{bmatrix} \\
&= \begin{bmatrix} \lim\limits_{\Delta t\to 0} \dfrac{\cos(t+\Delta t)-\cos t}{\Delta t} \\ \lim\limits_{\Delta t\to 0} \dfrac{\sin(t+\Delta t)-\sin t}{\Delta t}\end{bmatrix} \\
&= \begin{bmatrix} \cos'(t) \\ \sin'(t) \end{bmatrix} = \begin{bmatrix} -\sin t \\ \cos t \end{bmatrix}.
\end{aligned}$$

Our use of the notation $\vec{r}\,'(t)$ for velocity $\vec{r}(t)$ seems further justified.

Speed also appears to be a derivative. From physics, we know that speed is the magnitude of velocity. We can prove it mathematically.

Theorem 3.1.1 For a parameterization $\vec{p}:\mathbb{R}\to\mathbb{R}^n$ where velocity $\vec{p}(t)$ exists, we have

$$\text{speed}\,\vec{p}(t) = \|\text{velocity}\,\vec{p}(t)\| = \|\vec{p}\,'(t)\|.$$

Proof. The proof relies on the continuity of $\|\cdot\|$. Since $\|\cdot\|$ is continuous, we may freely move limits in and out. Thus

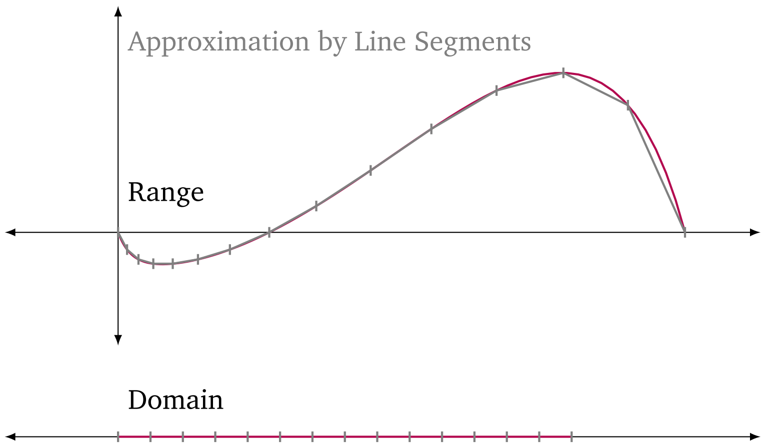
$$\begin{aligned}
\text{speed}\,\vec{p}(t) &= \lim_{\Delta t\to 0}\frac{\|\vec{p}(t+\Delta t)-\vec{p}(t)\|}{|\Delta t|} \\
&= \lim_{\Delta t\to 0}\left\|\frac{\vec{p}(t+\Delta t)-\vec{p}(t)}{\Delta t}\right\| \\
&= \left\|\lim_{\Delta t\to 0}\frac{\vec{p}(t+\Delta t)-\vec{p}(t)}{\Delta t}\right\| = \|\text{velocity}\,\vec{p}(t)\|.
\end{aligned}$$

■

Arc-length

Let $\mathcal{S}\subseteq\mathbb{R}^n$ be a curve parameterized by $\vec{p}:[a,b]\to\mathbb{R}^n$. The *arc-length* of $\mathcal{S}$ should be the length of $\mathcal{S}$ if you somehow untwisted $\mathcal{S}$ into a straight line without stretching anything. One of the big ideas of calculus is that we can handle curvy things by chopping them up into little pieces, computing for each piece, and then adding them back together. We use the same principle to define arc-length.

In essence, we will divide our curve $\mathcal{S}$ into many tiny line segments, add up the lengths of those line segments and take a limit as our line segments get tinier. A parameterization provides us with a way to do this. Since parameterizations are continuous, if we chop the domain of the parameterization into tiny pieces, we will have chopped the range into tiny pieces.



To find the arc-length of a curve, we will approximate each tiny piece with a straight line segment connecting the endpoints. We then add up the lengths of all tiny segments and take a limit as our segment's length goes to zero.

Definition 3.1.3 — Arc-length. Let $\mathcal{S} \subseteq \mathbb{R}^n$ be a curve parameterized by $\vec{p} : [a, b] \to \mathcal{S}$. The *arc-length* of $\mathcal{S}$ is

$$\operatorname{arc\,len} \mathcal{S} = \lim_{\Delta t \to 0^+} \sum_{i=1}^{\frac{b-a}{\Delta t}} \|\vec{p}(a + (i-1)\Delta t) - \vec{p}(a + i\Delta t)\|.$$

There's something unsatisfying about this definition, though. We used a parameterization of $\mathcal{S}$ to compute the arc-length of $\mathcal{S}$. But $\mathcal{S} \subseteq \mathbb{R}^n$ is a curve regardless of whether or not it has a parameterization, and if you use a different parameterization, you should get the same arc length for $\mathcal{S}$. If you're worried about this, good! You're thinking carefully! We won't show it here, but in fact no matter what parameterization you use for a curve, this definition will always produce the same arc length.

There's another reason we might be unhappy with this definition. Limits of sums are hard to compute! However, the sum involved in arc-length looks very close to a Riemann sum. If we can rewrite it exactly as a Riemann sum, we can replace it with an integral. With some superficial manipulation we see

$$\begin{aligned} \operatorname{arc\,len} \mathcal{S} &= \lim_{\Delta t \to 0^+} \sum_{i=1}^{\frac{b-a}{\Delta t}} \|\vec{p}(a + (i-1)\Delta t) - \vec{p}(a + i\Delta t)\| \\ &= \lim_{\Delta t \to 0^+} \sum_{i=1}^{\frac{b-a}{\Delta t}} \frac{\|\vec{p}(a + (i-1)\Delta t) - \vec{p}(a + i\Delta t)\|}{\Delta t} \Delta t \\ &= \int_a^b \operatorname{speed} \vec{p}(t)\,\mathrm{d}t. \end{aligned}$$

For the last equality, we noticed that $\lim_{\Delta t \to 0^+} \frac{\|\vec{p}(t) - \vec{p}(t+\Delta t)\|}{\Delta t} = \operatorname{speed} \vec{p}(t)$, which involved switching a limit and an infinite sum. In order to do this rigorously, we need a mathematical proof that it is logically valid. Such a proof is the subject of a course in *real analysis*, and won't be covered here, but it's always good to keep track of what you've actually proved and what you've been told is true[3].

Speed is easy to calculate, and we have a better handle on calculating integrals than we do limits of sums, so now we have a chance of calculating arc-length.

■ **Example 3.3** We shall find the length of the parabola with equation $y = x^2$ on the interval $-1 \le x \le 1$. A parametric representation of the parabola is $\vec{p}(t) = (t, t^2)$ where $-1 \le t \le 1$. Now,

$$\frac{\mathrm{d}\vec{p}}{\mathrm{d}t} = \begin{bmatrix} 1 \\ 2t \end{bmatrix},$$

so speed $\vec{p}(t) = \|\vec{p}\,'(t)\| = \sqrt{1+4t^2}$. Hence,

$$L = \int_{-1}^{1} \sqrt{1+4t^2}dt = \sqrt{5} + \frac{1}{4}\ln\frac{\sqrt{5}+2}{\sqrt{5}-2}.$$

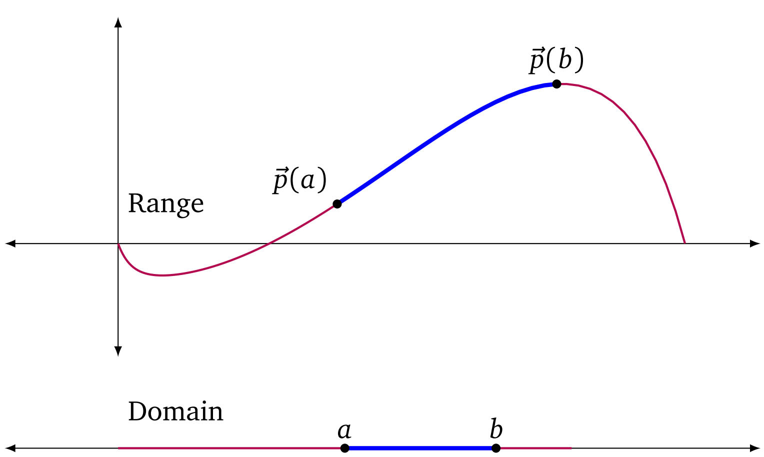
■

As you can see from Example 3.3, the integrals involved in computing arc length can be difficult. In fact, most of them don't have an elementary form, which means in the real world we often approximate arc length directly from the Riemann sum rather than calculate it exactly.

Exercises for 3.1

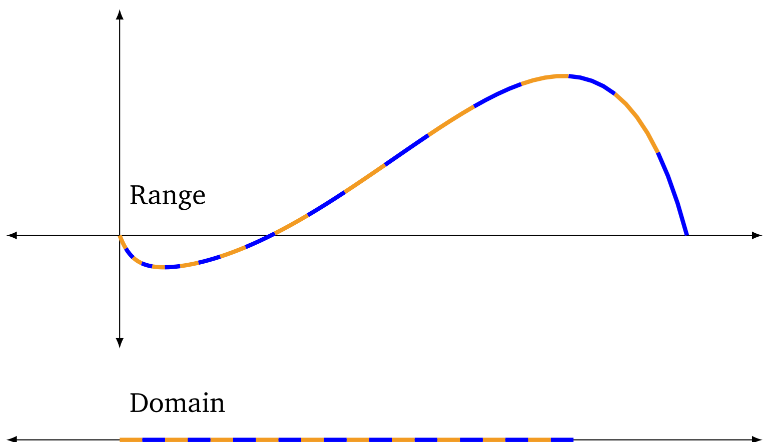
3.2 Arc-length Parameterization

Recall a parameterization is a relation between two objects. If a curve $\mathcal{S} \subseteq \mathbb{R}^n$ is parameterized by $\mathbb{R}$, it means there is a continuous, one-to-one function $\vec{p} : \mathbb{R} \to \mathcal{S}$. This function can be thought of as a map from $\mathbb{R}$ to $\mathcal{S}$. Any interval $[a, b] \subseteq \mathbb{R}$ corresponds to a segment $\vec{p}([a, b]) \subseteq \mathcal{S}$.



[3] In order to prove that swapping the limit and sum is valid, we actually need extra assumptions on $\vec{p}$. If we make $\vec{p}$ *differentiable* rather than merely *continuous*, we can prove that the swap is valid.

Alternatively, we may think of $\vec{p} : \mathbb{R} \to \mathcal{S}$ as a function that picks up the real line, stretches, twists, and warps it, and sticks it into $\mathbb{R}^n$ in the shape of $\mathcal{S}$. In this sense, not all parameterizations are created equally. Some significantly stretch and warp and others barely do at all.



The least stretchy type of parameterization is called an *arc-length parameterization*.

Before we define arc-length parameterization, let's introduce some notation. If $\mathcal{S} \subseteq \mathbb{R}^n$ is a curve parameterized by $\vec{p} : \mathbb{R} \to \mathcal{S}$, then

$$\operatorname{arc\,len} \vec{p}\Big|_a^b = \text{ arc length of } \mathcal{S} \text{ between } \vec{p}(a) \text{ and } \vec{p}(b).$$

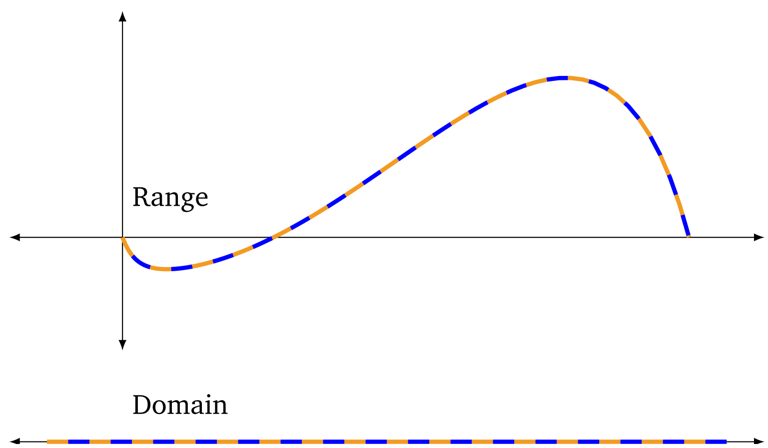
We might read $\operatorname{arc\,len} \vec{p}\Big|_a^b$ as the "arc length of the curve traced by $\vec{p}(t)$ from $t = a$ to $t = b$." Using notion for the *image* of a set, we can also write

$$\operatorname{arc\,len} \vec{p}\Big|_a^b = \operatorname{arc\,len} \vec{p}([a, b]).$$

Definition 3.2.1 Let $\mathcal{S} \subseteq \mathbb{R}^n$ be a curve and $\vec{p} : \mathbb{R} \to \mathcal{S}$ be a parameterization. The parameterization $\vec{p}$ is called an *arc-length parameterization* if for $b \geq a$,

$$\operatorname{arc\,len} \vec{p}\Big|_a^b = b - a.$$

In plain language, if $\vec{p}$ is an arc-length parameterization, then the distance traveled in parameter space is the same as the distance traveled along the curve.



■ **Example 3.4** Find an arc-length parameterization of the line ℓ which passes through the origin and has a direction vector $\vec{d} = \hat{\mathbf{x}} + 2\hat{\mathbf{y}} + \hat{\mathbf{z}}$.

Let's name our parameterization $\vec{p}$ and see if we can build up a formula for $\vec{p}$. Since ℓ passes through the origin, let's have $\vec{p}(0) = \vec{0}$. Now, $\vec{p}(1)$ has to be a point on ℓ that is distance 1 from $\vec{0}$. There are two such points, so we'll arbitrarily declare $\vec{p}(1) = \frac{1}{\sqrt{6}}(1,2,1)$. The point $\vec{p}(2)$ must be distance 1 from $\vec{p}(1)$ and distance 2 from $\vec{p}(0)$. Since parameterizations aren't allowed to double-back, we only have one choice. Namely, $\vec{p}(2) = \frac{2}{\sqrt{6}}(1,2,1)$. Noticing the pattern, let's guess

$$\vec{p}(t) = \frac{t}{\sqrt{6}} \begin{bmatrix} 1 \\ 2 \\ 1 \end{bmatrix}.$$

Now we'll verify that $\vec{p}$ is an arc-length parameterization. Since ℓ is a straight line this is easy:

$$\operatorname{arc\,len} \vec{p}\,\Big|_a^b = \|\vec{p}(b) - \vec{p}(a)\| = \left|\tfrac{b-a}{\sqrt{6}}\right| \left\| \begin{bmatrix} 1 \\ 2 \\ 1 \end{bmatrix} \right\| = |b-a|,$$

and so $\vec{p}$ is an arc-length parameterization. ■

Arc-length parameterizations can also be characterized by their speed. For an arc-length parameterization, the distance traveled in parameter space must be equal to the distance traveled along the curve. Therefore, an arc-length parameterization must also move at speed 1.

Theorem 3.2.1 Let $\mathcal{S} \subseteq \mathbb{R}^n$ be a curve and $\vec{p} : \mathbb{R} \to \mathcal{S}$ a parameterization. The parameterization $\vec{p}$ is an arc-length parameterization if and only if $\operatorname{speed} \vec{p} = 1$.

Proof. Suppose $\vec{p}$ is a parameterization of $\mathcal{S}$ and $\operatorname{speed} \vec{p} = 1$. Then

$$\operatorname{arc\,len} \vec{p}\,\Big|_a^b = \int_a^b \operatorname{speed} \vec{p}(t)\,\mathrm{d}t = \int_a^b 1\,\mathrm{d}t = b - a.$$

Now, suppose that $\vec{p}$ satisfies $\operatorname{arclen}\vec{p}\Big|_a^b = b - a$. We then know if $\Delta t > 0$ is small

$$\frac{\|\vec{p}(t+\Delta t)-\vec{p}(t)\|}{\Delta t} \approx \frac{\operatorname{arclen}\vec{p}\Big|_t^{t+\Delta t}}{\Delta t} = \frac{(t+\Delta t)-t}{\Delta t} = 1.$$

After taking a limit as $\Delta t \to 0$, "$\approx$" will turn into "$=$".

To make this argument completely rigorous, we will have to do a little bit of mathematical gymnastics. Since the shortest distance between two points is a straight line, what we really know is

$$\|\vec{p}(t+\Delta t)-\vec{p}(t)\| \leq \operatorname{arclen}\vec{p}\Big|_t^{t+\Delta t} = \Delta t$$

for all Δt. Now, taking a limit as $\Delta t \to 0$, we deduce $\operatorname{speed}\vec{p}(t) \leq 1$. But,

$$b-a = \operatorname{arclen}\vec{p}\Big|_a^b = \int_a^b \operatorname{speed}\vec{p}(t)\,\mathrm{d}t \leq \int_a^b 1\mathrm{d}t = b-a,$$

and so if $\operatorname{speed}\vec{p}$ is a continuous function, $\operatorname{speed}\vec{p}(t) = 1$ for all t. Even if speed is not continuous, we can argue that $\operatorname{speed}\vec{p}$ is *essentially* 1.[4] ■

Now we have a new way of thinking about arc-length parameterizations and a new way of creating them. Instead of working with arc length directly, we can attempt to manipulate *speed*.

■ **Example 3.5** Find an arc-length parameterization of $\mathcal{C}$, the circle of radius 2 centered at the origin.

We already know a parameterization of $\mathcal{C}$. Namely,

$$\vec{r}(t) = \begin{bmatrix} 2\cos t \\ 2\sin t \end{bmatrix}.$$

However, $\operatorname{speed}\vec{r}(t) = \|\vec{r}\,'(t)\| = \sqrt{(-2\sin t)^2 + (2\cos t)^2} = 2$. If we could somehow *slow down* time, we could slow down the speed to be 1, giving an arc-length parameterization.

Let $w(t) = t/2$. The function w stretches time by a factor of 2. Define

$$\vec{p}(t) = \vec{r}\circ w(t) = \begin{bmatrix} 2\cos t/2 \\ 2\sin t/2 \end{bmatrix}.$$

Now,

$$\begin{aligned}\operatorname{speed}\vec{p}(t) = \|\vec{p}\,'(t)\| = \|(\vec{r}\circ w)'(t)\| &= \|(\vec{r}\,'\circ w(t))w'(t)\| \\ &= |w'(t)|\|\vec{r}\,'\circ w(t)\| = \tfrac{1}{2}2 = 1.\end{aligned}$$

In this computation we made judicious use of the chain rule, however we could have computed directly from our formula for $\vec{p}$. Now, since $\operatorname{speed}\vec{p}(t) = 1$, the function $\vec{p}$ is an arc-length parameterization. ■

[4] Here, the word *essentially* is a technical term coming from real analysis.

Exercise 3.3 Let $\mathcal{C}$ and $\vec{r}$ be as in Example 3.5. The function $\vec{q}(t) = \frac{1}{2}\vec{r}(t)$ has speed 1 and so is an arc-length parameterization of something. Explain why $\vec{q}$ is *not* an arc-length parameterization of $\mathcal{C}$.

In Example 3.5, we adjusted the speed of a parameterization by warping time. Suppose $\mathcal{S} \subseteq \mathbb{R}^n$ is a curve parameterized by $\vec{p} : \mathbb{R} \to \mathcal{S}$ and let $w : \mathbb{R} \to \mathbb{R}$ be a parameterization of $\mathbb{R}$. In other words, w stretches or squishes (or flips) $\mathbb{R}$ by varying amounts. Now consider

$$\vec{r} = \vec{p} \circ w.$$

The function $\vec{r}$ has domain $\mathbb{R}$ and range $\mathcal{S}$. Further, since $\vec{p}$ and w are both one-to-one and continuous, we know $\vec{r}$ is one-to-one and continuous. Thus, $\vec{r}$ is a parameterization of $\mathcal{S}$. And, by using the chain rule,

$$\vec{r}\,'(t) = (\vec{p} \circ w)'(t) = w'(t)\big[\vec{p}\,' \circ \vec{w}(t)\big],$$

so

$$\text{speed}\,\vec{r} \text{ at time } t = |w'(t)|\big[\,\text{speed}\,\vec{p} \text{ at time } w(t)\big].$$

It is important to note that we must compose $\vec{p}$ and w in order to get a parameterization of $\mathcal{S}$. See Exercise 3.3 for an example of why you cannot multiply $\vec{p}$ and w.

With the idea of stretching time in the back of our minds, let's work through a hypothetical example.

Suppose $\mathcal{S} \subseteq \mathbb{R}^n$ is a curve parameterized by $\vec{p} : \mathbb{R} \to \mathcal{S}$ and we've computed the following table of values.

t	$\operatorname{arc\,len}\vec{p}\Big\vert_0^t$
0	0
1	2
2	3.5
3	6
4	7

XXX Figure

We'd like to find a time-stretching function $w : \mathbb{R} \to \mathbb{R}$ so that $\vec{r} = \vec{p} \circ w$ is an arc-length parameterization of $\mathcal{S}$. In other words, we need

$$\operatorname{arc\,len}\vec{r}\Big|_0^t = \operatorname{arc\,len}(\vec{p} \circ w)\Big|_0^t = t.$$

In particular, $\operatorname{arc\,len}\vec{r}\Big|_0^0 = 0$ (this we get for free) and $\operatorname{arc\,len}\vec{r}\Big|_0^2 = 2$. From the table, we know that the arc length from $\vec{p}(0)$ to $\vec{p}(1)$ is 2, and so we need $w(2) = 1$. This way $\vec{r}(2) = \vec{p}(w(2)) = \vec{p}(1)$. Continuing in this way, we get the following table of values for w.

t	$w(t)$
0	0
2	1
3.5	2
6	3
7	4

The function w is just the inverse of the function $\operatorname{arc\,len}\vec{p}\Big|_0^t$! This also makes sense from a purely algebraic perspective. Consider

$$x = \operatorname{arc\,len}\vec{r}\Big|_0^x = \operatorname{arc\,len}(\vec{p}\circ w)\Big|_0^x = \operatorname{arc\,len}(\vec{p})\Big|_0^{w(x)}.$$

Replacing x with $w^{-1}(t)$, we see

$$w^{-1}(t) = \operatorname{arc\,len}\vec{p}\Big|_0^{w\circ w^{-1}(t)} = \operatorname{arc\,len}\vec{p}\Big|_0^t.$$

This means the inverse of w is $\operatorname{arc\,len}\vec{p}\Big|_0^t$ and so the inverse of $\operatorname{arc\,len}\vec{p}\Big|_0^t$ must be w.[5]

We now have a concrete way to find an arc-length parameterization.

■ **Example 3.6** Let $\mathcal{R}$ be the ray parameterized by $\vec{p} : [0,\infty) \to \mathbb{R}^2$ where $\vec{p}(t) = (t^2, 2t^2)$. Find an arc-length parameterization of $\mathcal{R}$.

We'll start by finding the arc-length function for $\vec{p}$.

$$\begin{aligned} a(t) = \operatorname{arc\,len}\vec{p}\Big|_0^t &= \int_0^t \|\vec{p}\,'(x)\|\mathrm{d}x \\ &= \int_0^t \sqrt{(2x)^2+(4x)^2}\mathrm{d}x = t^2\sqrt{5}. \end{aligned}$$

Now define $w(t) = a^{-1}(t)$. Since $t \geq 0$, the inverse of a is well defined and is given by

$$w(t) = a^{-1}(t) = \sqrt{t/\sqrt{5}}.$$

Now, define $\vec{r}(t) = \vec{p}\circ w(t) = (t/\sqrt{5}, 2t/\sqrt{5})$. The parameterization $\vec{r}$ is an arc-length parameterization!

Of course we didn't need to go through all this work in this case. If all we wanted was to find an arc-length parameterization of $\mathcal{R}$ we could create one directly using geometry, since we know how to parameterize lines and rays. ■

Explicit Arc-length Parameterizations

The idea of arc-length parameterization is very important. However, for most curves, we're hopeless in finding a formula for the arc-length parameterization. We can for a hand full of curves, like a circle, a line, a helix, but even something as simple as an ellipse cannot be arc-length parameterized with elementary functions.

Why is it so hard? Well, integrals are hard in general. Most formulas cannot be integrated in closed form. Integrals involving square roots are even harder to evaluate. And, even if you manage to integrate to find the arc-length function, you still have to invert that function[6]. So, if finding formulas for arc-length parameterizations is so hard, why do we bother with them at all? The answer is that the *idea* of an arc-length parameterization is incredibly useful. Its mere existence will aid our thinking. And, in many of the problems we will be solving, the arc-length parameterization will somehow get canceled out and we won't ever need to find a formula for it.

[5] Recall that for an invertible function f, we have $(f^{-1})^{-1} = f$.

[6] If you don't believe me, go ahead and try to find the inverse of the function $f(x) = xe^x$.

Exercises for 3.2

3.3 Acceleration and Curvature

In the Newtonian mechanics of one-dimensional motion, acceleration is the second derivative of position with respect to time. In $\mathbb{R}^n$ we define it in the same way.

Definition 3.3.1 — Acceleration. Let $\vec{p} : \mathbb{R} \to \mathcal{S}$ be a parameterization of $\mathcal{S}$. The *acceleration* of $\vec{p}$ is

$$\operatorname{accel}\vec{p}(t) = (\operatorname{velocity}\vec{p})'(t) = \vec{p}\,''(t).$$

Just like velocity, acceleration is a *vector*. Let's consider two examples. Define

$$\vec{l}(t) = \begin{bmatrix} t^2 \\ t^2 \end{bmatrix} \qquad \text{and} \qquad \vec{c}(t) = \begin{bmatrix} \cos t \\ \sin t \end{bmatrix}.$$

Here $\vec{l}$ parameterizes a ray and $\vec{c}$ a circle. Further, $\vec{l}$ traces along the ray faster and faster, whereas $\vec{c}$ has constant speed as it traces the circle. We compute

$$\operatorname{accel}\vec{l}(t) = \begin{bmatrix} 2 \\ 2 \end{bmatrix} \qquad \text{and} \qquad \operatorname{accel}\vec{c}(t) = \begin{bmatrix} -\cos t \\ -\sin t \end{bmatrix},$$

and see that the acceleration of $\vec{l}$ is constant whereas the acceleration of $\vec{c}$ is not. Further, $\|\operatorname{accel}\vec{l}(t)\| = 2$ and $\|\operatorname{accel}\vec{c}(t)\| = 1$, and so the magnitude of the acceleration of both $\vec{l}$ and $\vec{c}$ is constant.

In the past, you might have distinguished linear acceleration (running faster and faster along a straight line) from centripetal acceleration (the acceleration you experience by moving at a constant speed around a circle). With vectors, these two types of acceleration are unified into a single vector.

In the previous example, $\vec{l}$ had purely linear acceleration and the acceleration vector pointed tangent to the path $\vec{l}$ traced. Analyzing $\vec{c}$, we see $\vec{c}$ had purely centripetal acceleration and the acceleration was orthogonal to the curve it traced. What happens if we mix the two types of acceleration?

Consider

$$\vec{r}(t) = \begin{bmatrix} \cos t^2 \\ \sin t^2 \end{bmatrix}.$$

Computing,

$$\operatorname{accel}\vec{r}(t) = \begin{bmatrix} -2(\sin t^2 + 2t^2\cos t^2) \\ 2(\cos t^2 - 2t^2\sin t^2) \end{bmatrix}.$$

There is no clear relationship between the curve $\vec{r}$ traces and the acceleration vector for $\vec{r}$. To see this relationship, we need to decompose $\operatorname{accel}\vec{r}$ into its tangential and normal components.

Definition 3.3.2 — Tangential and Normal Acceleration. Let $\vec{p} : \mathbb{R} \to \mathcal{S}$ be a parameterization of $\mathcal{S}$. Then $\operatorname{accel}\vec{p}$ can be written as

$$\operatorname{accel}\vec{p}(t) = \vec{a}_T(t) + \vec{a}_N(t)$$

where $\vec{a}_T(t)$ is tangent to $\mathcal{S}$ at the point $\vec{p}(t)$ and $\vec{a}_N(t)$ is orthogonal to $\mathcal{S}$ at the point $\vec{p}(t)$. In this case, $\vec{a}_T(t)$ is called the *tangential component of the acceleration* and $\vec{a}_N(t)$ is called the *normal component of the acceleration* of $\vec{p}$.

■ **Example 3.7** Let $\vec{r}(t) = (\cos t^2, \sin t^2)$. Find the tangential and normal components of the acceleration of $\vec{r}$.

Earlier we computed

$$\operatorname{accel}\vec{r}(t) = \begin{bmatrix} -2(\sin t^2 + 2t^2\cos t^2) \\ 2(\cos t^2 - 2t^2 \sin t^2) \end{bmatrix}.$$

We can use projections to split accel $\vec{r}$ into its tangential and normal components.

Recall $\vec{r}\,'(t)$ is tangent to the curve $\vec{r}$ traces at the point $\vec{r}(t)$. Thus,

$$\vec{a}_T = \operatorname{proj}_{\vec{r}\,'(t)} \operatorname{accel}\vec{r}(t) = \begin{bmatrix} -2\sin t^2 \\ 2\cos t^2 \end{bmatrix},$$

and

$$\vec{a}_N = \operatorname{accel}\vec{r}(t) - \vec{a}_T(t) = \begin{bmatrix} -4t^2\cos t^2 \\ -4t^2 \sin t^2 \end{bmatrix}.$$

■

Suppose now that $\vec{r} : \mathbb{R} \to \mathcal{S}$ is an arc-length parameterization of $\mathcal{S}$. Since the speed of $\vec{r}$ is constant, it is intuitive that the tangential component of the acceleration of $\vec{r}$ is zero. Equipped with our knowledge of vectors, it won't be so hard to prove our intuition. But, it will be helpful to establish the product rule for dot products.

Exercise 3.4 Let $\vec{a}(t) = (a_x(t), a_y(t), a_z(t))$ and $\vec{b}(t) = (b_x(t), b_y(t), b_z(t))$ be parameterizations. Establish the *product rule for dot products*. That is, show that

$$\left[\vec{a}(t)\cdot\vec{b}(t)\right]' = \vec{a}\,'(t)\cdot\vec{b}(t) + \vec{a}(t)\cdot\vec{b}\,'(t).$$

Theorem 3.3.1 If $\vec{p} : \mathbb{R} \to \mathcal{S}$ is an arc-length parameterization of $\mathcal{S}$ then accel $\vec{p}$ is always orthogonal to $\mathcal{S}$. Equivalently,

$$\vec{p}\,''(t)\cdot\vec{p}\,'(t) = 0.$$

Proof. Since $\vec{p}$ is an arc-length parameterization, we know

$$\sqrt{\vec{p}\,'(t)\cdot\vec{p}\,'(t)} = \|\vec{p}\,'(t)\| = 1.$$

Squaring both sides we get the relationship

$$\vec{p}\,'(t)\cdot\vec{p}\,'(t) = 1. \tag{3.1}$$

Now we may take the derivative of both sides of Equation (3.1) and apply the product rule for dot products to find

$$0 = [\vec{p}\,'\cdot\vec{p}\,']'(t) = \vec{p}\,''(t)\cdot\vec{p}\,'(t) + \vec{p}\,'(t)\cdot\vec{p}\,''(t) = 2\vec{p}\,''(t)\cdot\vec{p}\,'(t),$$

and so $\vec{p}\,''(t)$ and $\vec{p}\,'(t)$ are orthogonal. ■

If you examine the proof of theorem 3.3.1 closely, you'll notice that we didn't actually need the speed of our parameterization to be 1. The proof works just as well if the speed is some other constant.

Curvature

For any curve $\mathcal{S}$ there are infinitely many choices of parameterizations, but in some sense, there is only one arc-length parameterization. An arc-length parameterization of $\mathcal{S}$ is uniquely determined by a direction (forwards or backwards along $\mathcal{S}$) and a starting position. Thus, we might think of an arc-length parameterization as *intrinsic* to a curve.

The *curvature* of a curve is a measure of how sharply a curve bends or twists. Curvature is another property of a curve—you don't need a parameterization to define curvature—but it is a lot easier to define with reference to an arc-length parameterization.

Definition 3.3.3 — Curvature. Let $\mathcal{S} \subseteq \mathbb{R}^n$ be a curve and let $\vec{p} : \mathbb{R} \to \mathcal{S}$ be an arc-length parameterization of $\mathcal{S}$. The *curvature* of $\mathcal{S}$ at the point $\vec{p}(t)$ is

$$\|\operatorname{accel} \vec{p}(t)\| = \|\vec{p}\,''(t)\|.$$

This definition of curvature can be made intuitive. If $\vec{p} : \mathbb{R} \to \mathcal{S}$ is an arc-length parameterization, all velocity vectors are unit length. Therefore, all acceleration of $\vec{p}$ must come from the velocity vectors changing direction (and not changing length). If a curve has a sharp bend (high curvature), the velocity vectors with rapidly change direction. If a curve is generally flat (low curvature), the velocity vectors hardly change direction at all.

XXX Figure

Curvature can be hard to calculate exactly, but it isn't so hard to eyeball.

■ **Example 3.8** Estimate the curvature at various points of the parabola $y = x^2$.

XXX Finish and include numerics ■

Exercises for 3.3

3.4 Line Integrals

3.5 Multi-dimensional Parameterizations

Chapter 4

Multi-variable Functions

We want to develop the calculus necessary to discuss functions of many variables. We shall start with functions $f(x,y)$ of two independent variables and functions $f(x,y,z)$ of three independent variables. However, in general, we need to consider functions $f(x_1,x_2,\ldots,x_n)$ of any number of independent variables. Recall, $\mathbb{R}^n$ stands for the set of all n-tuples $(x_1,x_2,\ldots,x_n)$ with real entries x_i. We shall use the old fashioned term *locus* to denote the set of all points satisfying some equation or condition.

4.1 Graphing in Many Variables

We shall encounter equations involving two, three, or more variables. As you know, an equation of the form

$$f(x,y) = c$$

may be viewed as defining a curve in the plane. For example, $ax + by = c$ defines a line in $\mathbb{R}^2$, while $x^2 + y^2 = r^2$ defines a circle of radius r centered at the origin. Similarly, an equation involving three variables

$$f(x,y,z) = c$$

may be thought of as defining a *surface* in space. We saw previously that the locus in $\mathbb{R}^3$ of a linear equation

$$ax + by + cz = d$$

(where not all a, b, and c are zero) is a plane. If we use more complicated equations, we get more complicated surfaces.

■ **Example 4.1** The equation

$$x^2 + y^2 + z^2 = r^2$$

may be rewritten $\|\vec{r}\| = \sqrt{x^2 + y^2 + z^2} = R$, so it asserts that the point with position vector $\vec{r}$ is at distance r from the origin. Hence, the locus of all such points is a *sphere* of radius r centered at the origin. ■

■ **Example 4.2** Consider the locus of the equation

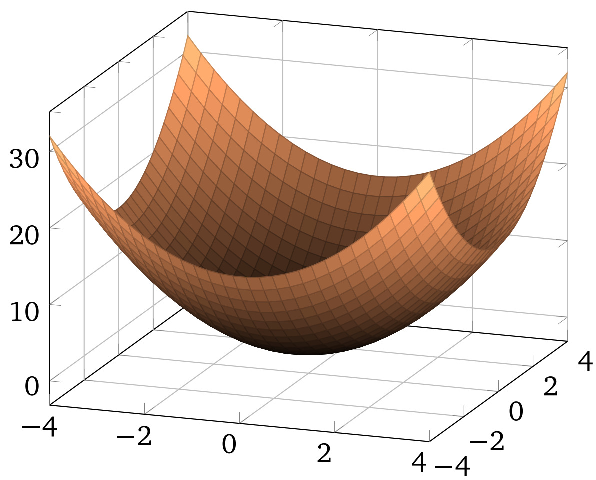
$$x^2 + 2x + y^2 - 4y + z^2 = 20.$$

This is also a sphere, but one not centered at the origin. To see this, *complete the squares* for the terms involving x and y.

$$x^2 + 2x + 1 + y^2 - 4y + 4 + z^2 = 10 + 1 + 4 = 25$$
$$(x+1)^2 + (y-2)^2 + z^2 = 5^2.$$

This asserts that the point with position vector $\vec{r} = (x, y, z)$ is 5 units from the point $(-1, 2, 0)$, i.e., it lies on a sphere of radius 5 centered at $(-1, 2, 0)$. ■

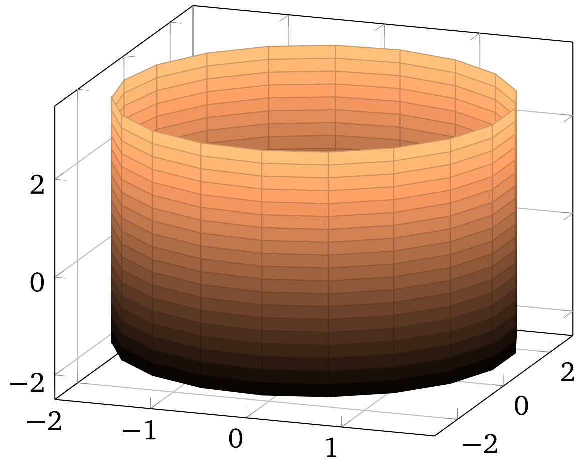
■ **Example 4.3** Consider the locus of the equation $z = x^2 + y^2$ (which could also be written $x^2 + y^2 - z = 0$). To see what this looks like, we consider its intersection with various planes. Its intersection with the yz-plane is obtained by setting $x = 0$ to get $z = y^2$. This is a parabola in the yz-plane. Similarly, its intersection with the xz-plane is the parabola given by $z = x^2$. To fill in the picture, consider intersections with planes parallel to the xy-plane. Any such plane has the equation $z = h$, so the intersection has the equation $x^2 + y^2 = h = (\sqrt{h})^2$, which you should recognize as a circle of radius $\sqrt{h}$, if $h > 0$. Note that the circle is centered at $(0, 0, h)$ on the z-axis since it lies in the plane $z = h$. If $z = h = 0$, then the circle reduces to a single point, and for $z = h < 0$, there is no locus. The surface is "bowl" shaped and is called a *circular paraboloid*.



■

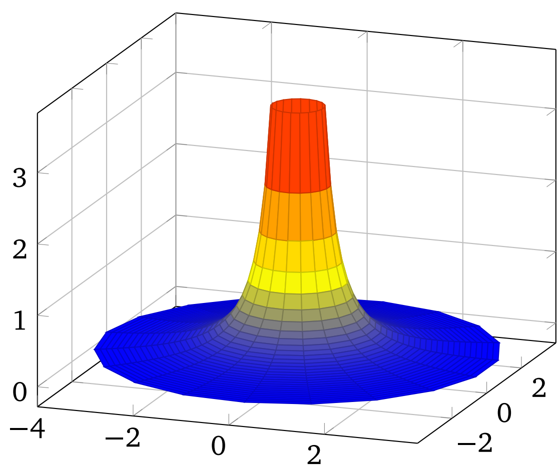
Graphing a surface in $\mathbb{R}^3$ by sketching its traces on various planes is a useful strategy. In order to be good at it, you need to know the basics of plane analytic geometry so you can recognize the resulting curves. In particular, you should be familiar with the elementary facts concerning *conic sections*, i.e., ellipses, hyperbolas, and parabolas. Edwards and Penney, 3rd Edition, Chapter 10 is a good reference for this material.

■ **Example 4.4** Consider the locus in space of $\dfrac{x^2}{4} + \dfrac{y^2}{9} = 1$. Its intersection with a plane $z = h$ parallel to the xy-plane is an ellipse centered on the z-axis and with semi-minor and semi-major axes 2 and 3, respectively. The surface is a *cylinder* perpendicular to the xy-plane with elliptical cross sections. Note that the locus *in space* is not just the ellipse in the xy-plane with the same equation.



XXX Figure ■

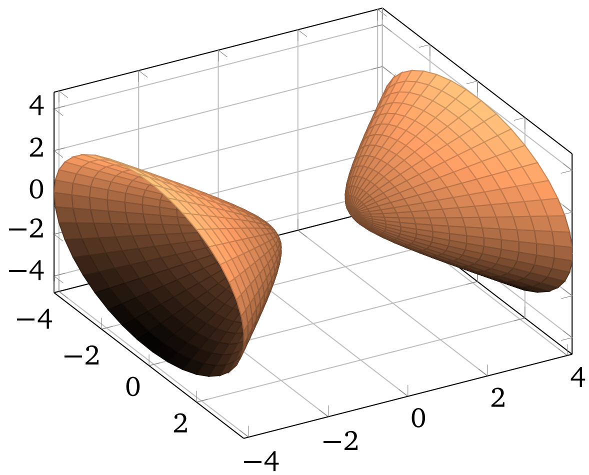
■ **Example 4.5** Consider the locus in space of the equation $z = \dfrac{1}{x^2+y^2}$. Its intersection with the plane $z = h$ (for $h > 0$) is the circle with equation $x^2 + y^2 = 1/h = (\sqrt{1/h})^2$. The surface does not intersect the xy-plane itself ($z = 0$) nor any plane below the xy-plane. Its intersection with the xz-plane ($y = 0$) is the curve $z = 1/x^2$, which is asymptotic to the x-axis and to the positive z-axis. Similarly, for its intersection with the yz-plane, the surface flattens out and approaches the xy-plane as $r = \sqrt{x^2+y^2} \to \infty$. It approaches the positive z-axis as $r \to 0$.



■

■ **Example 4.6** Consider the locus in space of the equation $yz = 1$. Its intersection with a plane parallel to the yz-plane ($x = d$) is a hyperbola asymptotic to the y and z axes. The surface is perpendicular to the yz-plane. Such a surface is also called a *cylinder*, although it doesn't close upon itself as the elliptical cylinder considered above. ■

■ **Example 4.7** Consider the locus of the equation $x^2+z^2=y^2-1$. For each plane parallel to the xz-plane ($y=c$), the intersection is a circle $x^2+z^2=c^2-1=\left(\sqrt{c^2-1}\right)^2$ centered on the y-axis, at least of $c^2>1$. For $y=c=\pm 1$, the locus is a point, and for $-1<y=c<1$, the locus is empty. In addition, the intersection of the surface with the xy-plane ($z=0$) is the hyperbola with equation $x^2-y^2=-1$, and similarly for its intersection with the yz-plane. The surface comes in two pieces which open up as "bowls" centered on the positive and negative y-axes. The surface is called a *hyperboloid of 2 sheets*.



■

Graphing Functions

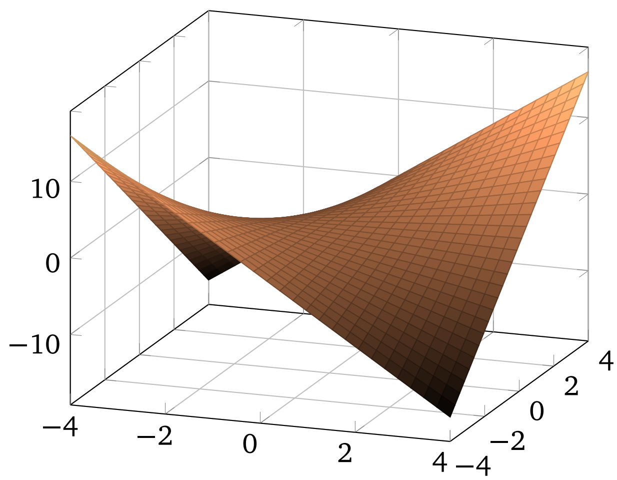
For a scalar function f of one independent variable, the *graph of the function* is the set of all points in $\mathbb{R}^2$ of the form $(x, f(x))$ for x in the domain of the function. (The domain of a function is the set of values of the independent variable for which the function is defined). In other words, it is the locus of the equation $y=f(x)$. It is generally a curve in the plane.

We can define a similar notion for a scalar function f of two independent variables. The graph is the set of points in $\mathbb{R}^3$ of the form $(x,y,f(x,y))$ for (x,y), a point in the domain of the function. In other words, it is the locus of the equation $z=f(x,y)$, and it is generally a surface in space. The graph of a function is often useful in understanding the function.

We have already encountered several examples of graphs of functions. For example, the locus of $z=x^2+y^2$ is the graph of the function f defined by $f(x,y)=x^2+y^2$. Similarly, the locus of $z=1/(x^2+y^2)$ is the graph of the function f defined by $f(x,y)=1/(x^2+y^2)$ for $(x,y)\neq(0,0)$. Note that in the first case there does not need to be a restriction on the domain of the function, but in the second case $(0,0)$ was omitted.

In some of the other examples, the locus of the equation cannot be considered the graph of a function. For example, the equation $x^2+y^2+z^2=R^2$ cannot be solved uniquely for z in terms of (x,y). Indeed, we have $z=\pm\sqrt{R^2-x^2-y^2}$, so that two possible functions are suggested. $z=f_1(x,y)=\sqrt{R^2-x^2-y^2}$ defines a function with graph the *top hemisphere*, while $z=f_2(x,y)=-\sqrt{R^2-x^2-y^2}$ yields the lower hemisphere. (Note that for either of the functions, the relevant domain is the set of points on or inside the circle $x^2 + y^2 = R^2$. For points outside that circle, the expression inside the square root is negative, and since we are only talking about functions assuming real values, such points must be excluded).

■ **Example 4.8** Let $f(x, y) = xy$ for all (x, y) in $\mathbb{R}^2$. The graph is the locus of the equation $z = xy$. We can sketch it by considering traces on various planes. Its intersection with a plane parallel to the xy-plane ($z =$ constant) is a hyperbola asymptotic to lines parallel to the x and y axes. For $z > 0$, the hyperbola is in the first and third quadrants of the plane, but for $z < 0$, it is in the second and fourth quadrants. For $z = 0$, the equation is $xy = 0$ with locus consisting of the x-axis ($y = 0$) and the y-axis ($x = 0$). Thus, the graph intersects the xy-plane in two straight lines. The surface is generally shaped like an "infinite saddle." It is called a *hyperbolic paraboloid*. It is clear where the term "hyperbolic" comes from. Can you see any parabolas? (Hint: Try planes perpendicular to the xy-plane with equations of the form $y = mx$).



■

■ **Example 4.9** Let $f(x, y) = x/y$ for $y \neq 0$. Thus, the domain of this function consists of all points (x, y) not on the x-axis ($y = 0$). The trace in the plane $y = c, c \neq 0$ is the line $z = (1/c)x$ with slope $1/c$. Similarly, the trace in the plane $z = c, c \neq 0$ is the line $y = (1/c)x$. Finally, the trace in the plane $x = c$, is the hyperbola $z = c/x$. Even with this information you will have some trouble visualizing the graph. However, the equation $z = x/y$ can be rewritten $yz = x$. By permuting the variables, you should see that the locus of $yz = x$ is similar to the saddle-shaped surface we just described but oriented differently in space. However, the saddle is not quite the graph of the function since it contains the z-axis ($y = x = 0$) but the graph of the function does not. In general, the graph of a function, since it consists of points of the form $(x, y, f(x, y))$, cannot contain points with the same values for x and y but different values for z. In other words, any line parallel to the z-axis can intersect such a graph at most once. ■

Sketching graphs of functions, or more generally loci of equations in x, y, and z, is not easy. One approach drawn from the study of topography is to interpret the equation $z = f(x, y)$ as giving the *elevation* of the surface, viewed as a hilly terrain, above a reference plane. (Negative elevation $f(x, y)$ is interpreted to mean that the surface dips below the reference plane.) For each possible elevation c, the intersection of the plane $z = c$ with the graph yields a curve

$f(x,y) = c$. This curve is called a *level curve*, and we draw a 2-dimensional map of the graph by sketching the level curves and labeling each by the appropriate elevation c. Of course, there are generally infinitely many level curves since there are infinitely many possible values of z, but we select some subset to help us understand the topography of the surface.

XXX Figure

■ **Example 4.10** The level curves of the surface $z = xy$ have equations $xy = c$ for various c. They form a family of hyperbolas, each with two branches. For $c > 0$, these hyperbolas fill the first and third quadrants, and for $c < 0$ they fill the second and fourth quadrants. For $c = 0$ the x and y axes together constitute the level "curve." See the diagram.

You can see that the region around the origin $(0,0)$ is like a "mountain pass" with the topography rising in the first and third quadrants and dropping off in the second and fourth quadrants. In general, a point where the graph behaves this way is called a *saddle point*. Saddle points indicate the added complexity which can arise when one goes from functions of one variable to functions of two or more variables. At such points, the function can be considered as having a maximum or a minimum depending on where you look. ■

Quadric Surfaces

One important class of surfaces are those defined by quadratic equations. These are analogs in three dimensions of conics in two dimensions. They are called *quadric surfaces*. We describe here *some* of the possibilities. You can verify the pictures by using the methods described above.

Consider first equations of the form

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

If all the signs are positive, the surface is called an *ellipsoid*. Planes perpendicular to one of the coordinate axes intersect it in ellipses (if they intersect at all). However, at the extremes these ellipses degenerate into the points $(\pm a, 0, 0), (0, \pm b, 0)$, and $(0, 0, \pm c)$.

XXX Figure

If exactly one of the signs are negative, the surface is called a *hyperboloid of one sheet*. It is centered on one axis (the one associated with the negative coefficient in the equation) and it opens up in both positive and negative directions along that axis. Its intersection with planes perpendicular to that axis are ellipses. Its intersections with planes perpendicular to the other axes are hyperbolas.

If exactly two of the signs are negative, the surface is called a *hyperboloid of two sheets*. It is centered on one axis (associated with the positive coefficient). For example, suppose the equation is

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

For $y < -b$ or $y > b$, the graph intersects a plane perpendicular to the y-axis in an ellipse. For $y = \pm b$, the intersection is the point $(0, \pm b, 0)$. (These two points are called vertices of the surface.) For $-b < y < b$, there is no intersection with a plane perpendicular to the y-axis.

The above surfaces are called *central quadrics*. Note that for the hyperboloids, with equations in standard form as above, the number of sheets is the same as the number of minus signs.

Consider next equations of the form

$$z = \pm\frac{x^2}{a^2} \pm \frac{y^2}{b^2}$$

(or similar equations obtained by permuting x, y and z).

If both signs are the same, the surface is called an *elliptic paraboloid*. If both signs are positive, it is centered on the positive z-axis and its intersections with planes perpendicular to the positive z-axis are a family of similar ellipses which increase in size as z increases. If both signs are negative, the situation is similar, but the surface lies below the x, y plane.

If the signs are different, the surface is called a *hyperbolic paraboloid*. Its intersection with planes perpendicular to the z-axis are hyperbolas asymptotic to the lines in those planes parallel to the lines $x/a = \pm y/b$. Its intersection with the xy-plane is just those two lines. The surface has a saddle point at the origin.

The locus of the equation $z = cxy, c \neq 0$ is also a hyperbolic paraboloid, but rotated so it intersects the xy-plane in the x and y axes.

XXX Figure

Finally, we should note that many so called "degenerate conics" are loci of quadratic equations. For example, consider

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

which may be solved to obtain

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

The locus is a double cone with elliptical cross sections and vertex at the origin.

Generalizations

In general, we will want to study functions of any number of independent variables. For example, we may define the graph of a scalar valued function f of three independent variables to be the set of all points in $\mathbb{R}^4$ of the form $(x, y, z, f(x, y, z))$. Such an object should be considered a three dimensional subset of $\mathbb{R}^4$, and it is certainly not easy to visualize. It is more useful to consider the analogs of level curves for such functions. Namely, for each possible value c attained by the function, we may consider the locus in $\mathbb{R}^3$ of the equation $f(x, y, z) = c$. This is generally a surface called a *level surface* for the function.

■ **Example 4.11** For $f(x, y, z) = x^2 + y^2 + z^2$, the level surfaces are concentric spheres centered at the origin if $c > 0$. For $c = 0$ the level 'surface' is not really a surface at all; it just consists of the point at the origin. (What if $c < 0$?)

For $f(x, y, z) = x^2 + y^2 - z^2$, the level surfaces are either hyperboloids of one sheet if $c > 0$ or hyperboloids of two sheets if $c < 0$. (What if $c = 0$?) ■

For functions of four or more variables, geometric interpretations are even harder to come by. If $f(x_1, x_2, \ldots, x_n)$ denotes a function of n variables, the locus in $\mathbb{R}^n$ of the equation $f(x_1, x_2, \ldots, x_n) = c$ is called a level set, but one doesn't ordinarily try to visualize it geometrically.

Instead of talking about many independent variables, it is useful to think instead of a single independent variable which is a *vector*, i.e., an element of $\mathbb{R}^n$ for some n. In the case $n = 2, 3$, we usually write $\vec{r} = (x, y)$ or $\vec{r} = (x, y, z)$ so $f(x, y)$ or $f(x, y, z)$ would be written simply $f(\vec{r})$. If $n > 3$, then one often denotes the variables $x_1, x_2, \ldots, x_n$ and denotes the vector (i.e., element of $\mathbb{R}^n$) by $\vec{x} = (x_1, x_2, \ldots, x_n)$. Then $f(x_1, x_2, \ldots, x_n)$ becomes simply $f(\vec{x})$. The case of a function of a single real variable can be subsumed in this formalism by allowing the case $n = 1$. That is, we consider a scalar x to be just a vector of dimension 1, i.e., an element of $\mathbb{R}^1$.

When we talked about kinematics, we considered *vector valued* functions $\vec{r}(t)$ of a single independent variable. Thus we see that it makes sense to consider in general functions of a vector variable which can also assume vector values. We indicate this by the notation $f : \mathbb{R}^n \to \mathbb{R}^m$. That shall mean that the domain of the function f is a subset of $\mathbb{R}^n$ while the set of values is a subset of $\mathbb{R}^m$. Thus, $n = m = 1$ would yield a scalar function of one variable, $n = 2, m = 1$ a scalar function of two variables, and $n = 1, m = 3$ a vector valued function of one scalar variable. We shall have occasion to consider several other special cases in detail.

There is one slightly non-standard aspect to the above notation. In ordinary usage in mathematics, "$f : \mathbb{R}^n \to \mathbb{R}^m$" means that $\mathbb{R}^n$ is the entire domain of the function f, whereas we are taking it to mean that the domain is some subset. We do this mostly to save writing since usually the domain will be almost all of $\mathbb{R}^n$ or at least some significant chunk of it. What we want to make clear by the notation is the dimensionality of both the independent and dependent variables.

Exercises for 4.1

You are encouraged to make use of the available computer software (e.g., Maple, Mathematica, etc.) to help you picture the graphs in the following problems.

1. State the largest possible domain for the function

   a) $f(x, y) = e^{x^2 - y^2}$

   b) $f(x, y) = \ln(y^2 - x^2 - 2)$

   c) $f(x, y) = \dfrac{x^2 - y^2}{x - y}$

   d) $f(x, y, z) = \dfrac{1}{xyz}$

   e) $f(x, y, z) = \dfrac{1}{\sqrt{z^2 - x^2 - y^2}}$

2. Describe the graph of the function described by

   a) $f(x, y) = 5$

   b) $f(x, y) = 2x - y$

c) $f(x,y) = 1 - x^2 - y^2$
d) $f(x,y) = 4 - \sqrt{x^2 + y^2}$
e) $f(x,y) = \sqrt{24 - 4x^2 - 6y^2}$

3. Sketch selected level curves for the functions given by

   a) $f(x,y) = x + y$
   b) $f(x,y) = x^2 + 9y^2$
   c) $f(x,y) = x - y^2$
   d) $f(x,y) = x - y^3$
   e) $f(x,y) = x^2 + y^2 + 4x + 2y + 9$

4. Describe selected level surfaces for the functions given by

   a) $f(x,y,z) = x^2 + y^2 - z$
   b) $f(x,y,z) = x^2 + y^2 + z^2 + 2x - 2y + 4z$
   c) $f(x,y,z) = z^2 - x^2 - y^2$

5. Describe the quadric surfaces which are loci in $\mathbb{R}^3$ of the following equations.

   a) $x^2 + y^2 = 16$
   b) $z^2 = 49x^2 + y^2$
   c) $z = 25 - x^2 - y^2$
   d) $x = 4y^2 - z^2$
   e) $4x^2 + y^2 + 9z^2 = 36$
   f) $x^2 + y^2 - 4z^2 = 4$
   g) $9x^2 + 4y^2 - z^2 = 36$
   h) $9x^2 - 4y^2 - z^2 = 36$

6. Describe the traces of the following functions in the given planes

   a) $z = xy$, in horizontal planes $z = c$
   b) $z = x^2 + 9y^2$ in vertical planes $x = c$ or $y = c$
   c) $z = x^2 + 9y^2$ in horizontal planes $z = c$

7. Describe the intersection of the cone $x^2 + y^2 = z^2$ with the plane $z = x + 1$.

8. Let $\vec{f} : \mathbb{R}^2 \to \mathbb{R}^2$.

   a) Try to invent a definition of ‘graph’ for such a function. For what n would it be a subset of $\mathbb{R}^n$?
   b) Try to invent a definition of ‘level set’ for such a function. For what n would it be a subset of $\mathbb{R}^n$?

4.2 Rates of Change and the Directional Derivative

For a function $f : \mathbb{R} \to \mathbb{R}$ of one variable, we have the idea of the *rate of change* of f. We might ask, "rate of change with respect to what?" For a function of one variable, this question has an obvious answer—the rate of change with respect to the only thing that varies! If $g : \mathbb{R}^n \to \mathbb{R}$ is a multi-variable function, the question of "what is the rate of change of g?" has a less obvious answer. After all, there are multiple variables and infinitely many directions in the domain.

Consider the function $g : \mathbb{R}^2 \to \mathbb{R}$ where $g(x, y) = (x - 2)^2 + y^2$. Let $\vec{u}$ be a unit vector. Now, starting at $(0, 0)$, we might ask approximately how much g changes when we move α units away from $(0, 0)$ in the direction $\vec{u}$.

XXX Figure

Since $\vec{u}$ is a unit vector, the displacement in the domain is $\alpha\vec{u}$, and so the exact change in g from $\vec{0}$ to $\alpha\vec{u}$ is $g(\alpha\vec{u}) - g(\vec{0})$. The approximate *rate* of change is then $\frac{g(\alpha\vec{u})-g(\vec{0})}{\alpha}$. Taking a limit as $\alpha \to 0$ will give us an instantaneous rate of change—one worth name.

Definition 4.2.1 — Rate of change with respect to distance. For a function $f : \mathbb{R}^n \to \mathbb{R}$, the *rate of change with respect to distance* of f in the direction $\vec{v}$ at the point $\vec{a}$ is

$$\lim_{h\to 0^+} \frac{f(\vec{a} + h\vec{v}) - f(\vec{a})}{||h\vec{v}||}.$$

We must divide by $||h\vec{v}||$ in the definition of rate of change with respect to distance because $||h\vec{v}||$ is exactly how far we've displaced. Of course, if we were forward-thinking enough to pick $\vec{v}$ to be a unit vector, then $||h\vec{v}|| = |h|$.

We can think of the rate of change with respect to distance in another way. Suppose that $f : \mathbb{R}^n \to \mathbb{R}$ and that $\vec{p} : \mathbb{R} \to \mathbb{R}^n$ is the arc-length parameterization of the with direction vector $\vec{v}$ and where $\vec{p}(0) = \vec{a}$. Then $f \circ \vec{p} : \mathbb{R} \to \mathbb{R}$ is a single-variable function and the rate of change of $\vec{f}$ at $\vec{a}$ in the direction $\vec{v}$ is just $(f \circ \vec{p})'(0)$. Think for a moment about why this is.

The rate of change of a function with respect to distance is, in some sense, the most natural notion of "rate of change" for a multivariable function. (If you're handed a function, what would you measure the rate of change against if not distance?) However, it turns out not to be the most useful notion of rate of change. For that, we introduce the *directional derivative*.

Definition 4.2.2 — Directional Derivative. For a function $f : \mathbb{R}^n \to \mathbb{R}$, the *directional derivative* of f at the point $\vec{a}$ in the direction $\vec{v}$ is

$$D_{\vec{a}} f(\vec{v}) = \lim_{h\to 0} \frac{f(\vec{a} + h\vec{v}) - f(\vec{a})}{h}.$$

The directional derivative $D_{\vec{a}} f(\vec{v})$ corresponds to the rate of change of f at the point $\vec{a}$ if you moved in the direction of $\vec{v}$ with speed $||\vec{v}||$. This seems like a less intuitive notion than the rate of change with respect to distance, but its virtues will soon become clear[1].

■ **Example 4.12** Let $f : \mathbb{R}^2 \to \mathbb{R}$ be given by $f(x, y) = (x - 2)^2 - y^2$. Compute $D_{\vec{0}} f(1, 2)$.

[1] For starters, $D_{\vec{a}} f$ is *linear*. Namely, if f is differentiable, then $D_{\vec{a}} f(\vec{u} + \vec{v}) = D_{\vec{a}} f(\vec{u}) + D_{\vec{a}} f(\vec{v})$. The same cannot be said for the rate of change with respect to distance.

Since we have no further theory to lean on, we must use the definition of the directional derivative directly.

$$D_{\vec{0}} f(1,2) = \lim_{h\to 0} \frac{f(\vec{0}+h(1,2)) - f(\vec{0})}{h} = \lim_{h\to 0} \frac{((h-2)^2 - (2h)^2) - 4}{h}$$
$$= \lim_{h\to 0} \frac{(-3h^2 - 4h + 4) - 4}{h} = -4.$$

■

The directional derivative can be used to approximate a function. Recall that if $g : \mathbb{R} \to \mathbb{R}$ is a differentiable function of one variable, then

$$g(a+h) \approx f(a) + hg'(a).$$

Similarly, for $f : \mathbb{R}^2 \to \mathbb{R}$ we have[2]

$$f(\vec{a} + h\vec{v}) \approx f(\vec{a}) + hD_{\vec{a}} f(\vec{v}). \tag{4.1}$$

The similarity in these two formulas is one reason why directional derivatives are more useful than rates of change with respect to distance[3].

Exercise 4.1 Create the analogous equation to Equation (4.1) but approximate $f(\vec{a}+h\vec{v})$ using the rate of change with respect to distance instead of the directional derivative.

Partial Derivatives

There are some particularly common directional derivatives. Namely, those in the directions of the standard basis. Let $f : \mathbb{R}^2 \to \mathbb{R}$ and let $\vec{a} = (a_x, a_y)$. Now,

$$D_{\vec{a}} f(\hat{\mathbf{x}}) = \lim_{h\to 0} \frac{f(\vec{a} + h\hat{\mathbf{x}}) - f(\vec{a})}{h} = \lim_{h\to 0} \frac{f(a_x + h, a_y) - f(a_x, a_y)}{h}.$$

This limit looks quite similar to a one-dimensional derivative. However, f does not have a one-dimensional domain. Accordingly we call $D_{\vec{a}} f(\hat{\mathbf{x}})$ a *partial derivative* and we have a special notation for it.

Definition 4.2.3 For a function $f : \mathbb{R}^n \to \mathbb{R}$, let x_i denote the ith input variable. The *partial derivative* of f at $\vec{a} = (a_1, \ldots, a_n)$ with respect to x_i is notated $\frac{\partial f}{\partial x_i}(\vec{a})$ and is defined to be

$$\frac{\partial f}{\partial x_i}(\vec{a}) = \lim_{h\to 0} \frac{f(a_1, \ldots, a_{i-1}, a_i + h, a_{i+1}, \ldots, a_n) - f(\vec{a})}{h}.$$

Sometimes we'll write $\frac{\partial f}{\partial x_i}$, omitting the point where the partial derivative is taken, if we want to save space, or if we want to view a partial derivative as a function of where it is taken. For functions whose domains are $\mathbb{R}^2$ or $\mathbb{R}^3$, we often use x, y, z instead of x_1, x_2, x_3 to denote the first, second, or third variable. It is not important how you write your variable as long as it is clear to you and your reader which variable is changing.

[2] Assuming f has directional derivatives, of course.

[3] If we think about vectors in $\mathbb{R}^1$, the number 1 is actually a vector. Therefore, for $g : \mathbb{R} \to \mathbb{R}$, we have $g'(1) = D_a g(1)$. Try writing out the limit expression. It's the same as the usual definition of derivative!

Tangent Vectors to Surfaces

When we had a curve $\mathcal{S}$ parameterized by $\vec{p}:\mathbb{R}\to\mathcal{S}$, we could find tangent vectors to $\mathcal{S}$ by computing the velocity vectors of $\vec{p}$. In particular, $\vec{p}\,'(t_0)$ is a tangent vector to $\mathcal{S}$ at the point $\vec{p}(y_0)$. We will find tangent vectors to surfaces in a similar way—by parameterizing a path on the surface.

Suppose $f:\mathbb{R}^2\to\mathbb{R}$ is a function and let $\mathcal{S}$ be the surface given by $z=f(x,y)$. Formally,

$$\mathcal{S}=\left\{\begin{bmatrix}x\\y\\z\end{bmatrix} : z=f(x,y)\right\}.$$

Since $\mathcal{S}$ is the graph of a function whose domain is $\mathbb{R}^2$ we have a natural parameterization of $\mathcal{S}$ coming from f. That is,

$$\vec{p}:\mathbb{R}^2\to\mathcal{S}\qquad\text{where}\qquad \vec{p}(x,y)=\begin{bmatrix}x\\y\\f(x,y)\end{bmatrix}$$

is a parameterization of $\mathcal{S}$.

Let $\vec{a}=(a_x,a_y,f(a_x,a_y))\in\mathcal{S}$ and consider the function $\vec{r}_x:\mathbb{R}\to\mathbb{R}^2$ given by $\vec{r}_x(t)=(a_x+t,a_y)$. The function $\vec{r}_x$ starts at the point (a_x,a_y) in the xy-plane and moves parallel to the x-axis with speed one. Since we have parameterized $\mathcal{S}$ by the xy-plane, by composing $\vec{p}$ and $\vec{r}_x$, we will get a parameterization of a curve on $\mathcal{S}$. That is, $\vec{c}=\vec{p}\circ\vec{r}_x$ parameterizes the curve on the surface $\mathcal{S}$ that when viewed from above looks like the line $y=a_y$.

XXX Figure

Further, $\vec{c}(0)=(a_x,a_y,f(a_x,a_y))=\vec{a}$. Thus, if we want a tangent vector to $\mathcal{S}$ at $\vec{a}$, all we need to do is compute $\vec{c}\,'(0)$.

Computing,

$$\vec{c}\,'(0)=(\vec{p}\circ\vec{r}_x)'(0)=\left(\begin{bmatrix}a_x+t\\a_y\\f(a_x+t,a_y)\end{bmatrix}'\text{ at } t=0\right)=\begin{bmatrix}1\\0\\\frac{\partial f}{\partial x}(a_x,a_y)\end{bmatrix}.$$

Of course, this was just one choice of a curve along $\mathcal{S}$. We could have just as easily used $\vec{p}\circ\vec{r}_y$ where $\vec{r}_y(t)=(a_x,a_y+t)$ as a parameterization of a curve. We could have parameterized paths that weren't parallel to the x or y axes. For instance $\vec{p}\circ\vec{r}_d$ where $\vec{r}_d(t)=(a_x+td_x,a_y+td_y)$ for non-zero constants d_x,d_y.

■ **Example 4.13** Find at least three tangent vectors at the point $\vec{a}=(1,2,12)$ to the surface $\mathcal{S}$ defined by $z=f(x,y)$ where $f(x,y)=(x-3)^2+y^3$.

Let $\vec{r}_a(t)=(1+t,2)$, $\vec{r}_b(t)=(1,2+t)$, and $\vec{r}_c(t)=(1+t,2+t)$. Further, let $\vec{p}:\mathbb{R}^2\to\mathbb{R}^3$ be defined by $\vec{p}(x,y)=(x,y,f(x,y))$. Since $\vec{p}$ is a parameterization of $\mathcal{S}$, $\vec{p}\circ\vec{r}_a$, $\vec{p}\circ\vec{r}_b$, and $\vec{p}\circ\vec{r}_c$ all parameterize paths in $\mathcal{S}$. Further,

$$\vec{p}\circ\vec{r}_a(0)=\vec{p}\circ\vec{r}_b(0)=\vec{p}\circ\vec{r}_c(0)=\vec{a},$$

so the velocity vectors at $t=0$ of these parameterizations will give tangent vectors to $\mathcal{S}$ at $\vec{a}$.

Computing, we see

$$(\vec{p}\circ\vec{r}_a)'(0) = \begin{bmatrix} 1 \\ 0 \\ \frac{\partial f}{\partial x}(1,2) \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ -4 \end{bmatrix},$$

$$(\vec{p}\circ\vec{r}_b)'(0) = \begin{bmatrix} 0 \\ 1 \\ \frac{\partial f}{\partial y}(1,2) \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \\ 12 \end{bmatrix},$$

and

$$(\vec{p}\circ\vec{r}_c)'(0) = \begin{bmatrix} 1 \\ 1 \\ D_{(1,2)}f(1,1) \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 8 \end{bmatrix}$$

are all tangent vectors to $\mathcal{S}$ at $\vec{a}$. ∎

Exercises for 4.2

4.3 Tangent Planes and Differentiability

Suppose $f:\mathbb{R}^2\to\mathbb{R}$ is a function and consider the surface $\mathcal{S}\subseteq\mathbb{R}^3$ given by the equation $z=f(x,y)$. If f is a "nice" function, like a polynomial, the surface $\mathcal{S}$ will be smooth. Visually, at each point on $\mathcal{S}$ we could imagine a tangent plane—the analog of a tangent line to a curve.

XXX Figure

More formally, fix a point $\vec{a}\in\mathcal{S}$ and consider the set $\mathcal{P}$ of all tangent vectors to $\mathcal{S}$ at $\vec{a}$. If $\mathcal{S}$ is a smooth surface, these tangent vectors will lie in a single plane. In other words, $\mathcal{P}$ is a plane, and in this case we call it the *tangent plane* to $\mathcal{S}$ at $\vec{a}$. (Fantastically, if $\mathcal{P}$ is a plane, it can be fully described by the point $\vec{a}$ and *two* non-parallel tangent vectors. We'll make use of this idea later.)

XXX Figure

Tangent planes to surfaces, like tangent lines to curves, have the property that they are a very good approximation to a surface at their point of tangency. We will leverage this idea to define what it means to be differentiable for a multivariable function.

Directional Derivatives and Planes

A plane $\mathcal{P}$ is a *set* and is distinct from any equations or parameterizations. However, if $\mathcal{P}\subseteq\mathbb{R}^3$ is not a vertical plane (i.e., one for which $\hat{\mathbf{z}}$ is a direction vector), then it can be represented uniquely by an equation of the form $z=\alpha x+\beta y+c$. In other words, a non-vertical plane in $\mathbb{R}^3$ is the graph of a function $f:\mathbb{R}^2\to\mathbb{R}$ which takes the form $f(x,y)=\alpha x+\beta y+c$. Functions of this form are known as *affine functions*, and they often show up in the context of *linear approximations*[4].

[4] In many contexts in calculus, the term "linear approximation" is used where "affine approximation" would be more accurate. Sometimes, though, it's easiest not to fight the tide of established culture.

Definition 4.3.1 — Affine Function. A function $f : \mathbb{R}^n \to \mathbb{R}$ is an *affine function* if it can be expressed as
$$f(x_1, x_2, \dots, x_n) = \alpha_1 x_1 + \alpha_2 x_2 + \dots + \alpha_n x_n + c$$
for some $\alpha_1, \dots, \alpha_n, c \in \mathbb{R}$. Equivalently, $f : \mathbb{R}^n \to \mathbb{R}$ is an affine function if it can be expressed as
$$f(\vec{x}) = \vec{\alpha} \cdot \vec{x} + c$$
for some $\vec{\alpha} \in \mathbb{R}^n$ and $c \in \mathbb{R}$.

Directional derivatives of affine functions take a particularly nice form. Let $f : \mathbb{R}^2 \to \mathbb{R}$ be an affine function. That is, $f(x, y) = \alpha x + \beta y + c$ for some $\alpha, \beta, c \in \mathbb{R}$. Let $\vec{a} = (a_x, a_y)$ and $\vec{v} = (v_x, v_y)$. Now,

$$\begin{aligned}
D_{\vec{a}} f(\vec{v}) &= \lim_{h\to 0} \frac{f(\vec{a} + h\vec{v}) - f(\vec{a})}{h} = \lim_{h\to 0} \frac{f(a_x + hv_x, a_y + hv_y) - f(a_x, a_y)}{h} \\
&= \lim_{h\to 0} \frac{\alpha(a_x + hv_x) + \beta(a_y + hv_y) + c - (\alpha a_x + \beta a_y + c)}{h} \\
&= \lim_{h\to 0} \frac{\alpha h v_x + \beta h v_y}{h} = \alpha v_x + \beta v_y \\
&= \begin{bmatrix} \alpha \\ \beta \end{bmatrix} \cdot \vec{v}.
\end{aligned}$$

We have written a formula for the directional derivative of f in an arbitrary direction! There's more.

$$\alpha = \begin{bmatrix} \alpha \\ \beta \end{bmatrix} \cdot \hat{\mathbf{x}} = \frac{\partial f}{\partial x}(\vec{a}) \qquad \text{and} \qquad \beta = \begin{bmatrix} \alpha \\ \beta \end{bmatrix} \cdot \hat{\mathbf{y}} = \frac{\partial f}{\partial y}(\vec{a})$$

and so

$$D_{\vec{a}} f(\vec{v}) = \begin{bmatrix} \frac{\partial f}{\partial x}(\vec{a}) \\ \frac{\partial f}{\partial y}(\vec{a}) \end{bmatrix} \cdot \vec{v}.$$

The vector $\begin{bmatrix} \frac{\partial f}{\partial x}(\vec{a}) \\ \frac{\partial f}{\partial y}(\vec{a}) \end{bmatrix}$ comes up often. It is called the *gradient*.

Definition 4.3.2 — Gradient. For a function $f : \mathbb{R}^n \to \mathbb{R}$, the *gradient* of f at the point a is written $\nabla f(\vec{a})$ and is defined to be the vector

$$\nabla f(\vec{a}) = \left(\tfrac{\partial f}{\partial x_1}(\vec{a}), \dots, \tfrac{\partial f}{\partial x_n}(\vec{a}) \right).$$

Using the notation of *gradient*, we conclude (using the same $f(x, y) = \alpha x + \beta y + c$ from earlier) that
$$D_{\vec{a}} f(\vec{v}) = \nabla f(\vec{a}) \cdot \vec{v}. \tag{4.2}$$
All directional derivatives of affine functions can be expressed as a dot product with the gradient. So far, we have only proven this for affine function $f : \mathbb{R}^2 \to \mathbb{R}$, but it is true for affine functions $g : \mathbb{R}^n \to \mathbb{R}$ as well. This fact is important enough to write down as a theorem.

Theorem 4.3.1 If $f : \mathbb{R}^n \to \mathbb{R}$ is an affine function, then $D_{\vec{a}} f(\vec{v}) = \nabla f(\vec{a}) \cdot \vec{v}$ for all $\vec{a}, \vec{v} \in \mathbb{R}^n$.

Note that at the moment, the gradient is just a bag of derivatives—it doesn't yet have meaning for us beyond a convenient notational trick.

So far the rule in Equation (4.2) only applies to affine functions. It would be nice if it worked for other functions. So, in true mathematical fashion, we will define a class of functions based on whether or not Equation (4.2) applies[5].

Linear Approximations

Recall from single variable calculus, if $f : \mathbb{R} \to \mathbb{R}$ was a differentiable function, a *linear approximation* of f at the point $a \in \mathbb{R}$ was a function $L : \mathbb{R} \to \mathbb{R}$ whose graph is the tangent line to $y = f(x)$ at the point $(a, f(a))$.

XXX Figure

Further, L always takes the form $L(x) = \alpha(x - a) + c$. In fact, if f' exists, we can be more specific:

$$L(x) = f'(a)(x - a) + f(a).$$

The approximation L is quite good around the point a, so we have

$$f(a + \Delta x) \approx L(a + \Delta x) = f'(a)\Delta x + f(a)$$

if Δx is small. It's a little unclear though what "$\approx$" exactly means. In your single-variable calculus class, you may or may not have made this precise. Visually, the line given by $y = L(x) = f'(a)(x - a) + f(a)$ is distinguished from all other lines passing through the point $(a, f(a))$ because it is *tangent*. Analytically, the linear approximation $L(x) = f'(a)(x-a)+f(a)$ is distinguished from all other linear approximations to f at a because it satisfies the property

$$\lim_{\Delta x \to 0} \frac{f(a + \Delta x) - L(a + \Delta x)}{\Delta x} = 0. \tag{4.3}$$

Exercise 4.2 Let $L(x) = \alpha(x - a) + c$ and let $f : \mathbb{R} \to \mathbb{R}$ be a differentiable function. Show that if L satisfies Equation (4.3) then it must be the case that $c = f(a)$ and $\alpha = f'(a)$.

Equation (4.3) encapsulates such a strong idea that it could, in fact, be used (and sometimes is used) to define the derivative of a single-variable function. We will use the multi-dimensional analog of Equation (4.3) to define differentiability.

Definition 4.3.3 — Differentiable. The function $f : \mathbb{R}^n \to \mathbb{R}$ is *differentiable* at $\vec{a} = (a_1, \ldots, a_n)$ if there exists an affine function $p(x_1, \ldots, x_n) = c + \sum \alpha_i (x_i - a_i) = c + \vec{\alpha} \cdot (\vec{x} - \vec{a})$ so that

$$f(\vec{a}) = p(\vec{a}) \qquad \text{and} \qquad \lim_{\vec{u} \to 0} \frac{f(\vec{a} + \vec{u}) - p(\vec{a} + \vec{u})}{\|\vec{u}\|} = 0.$$

[5] For technical reasons we won't do exactly this. If we considered the set of functions for which Equation (4.2) holds at every point, we'd never have a reasonable chain rule for multivariable functions. Instead, we'll use the slightly stronger notation of *linear approximations*.

The definition of differentiability can be interpreted in two equivalent ways. One is that a function is differentiable at a point $\vec{a}$ if there exists a "good" linear approximation to the function at $\vec{a}$ (where "good" is defined to mean that it satisfies the limit). Alternatively, one could think about the graph of a function as a surface. A function f is then differentiable at $\vec{a}$ if there exists a tangent plane to the surface given by the graph of f at the point $(\vec{a}, f(\vec{a}))$. Being able to switch back between these two perspectives will aid your intuition.

However, there's a big problem with this definition of differentiability. It involves a multivariable limit[6]. Multivariable limits are much more subtle than single-variable limits, and they are covered in the next section. But before we do that, we will further explore the consequences of differentiability.

Consequences of Differentiability

Recall, directional derivatives of affine functions can be expressed as dot products with the gradient vector. This is true in general for differentiable functions.

Theorem 4.3.2 If $f : \mathbb{R}^n \to \mathbb{R}$ is differentiable at the point $\vec{a}$, then

$$D_{\vec{a}} f(\vec{v}) = \nabla f(\vec{a}) \cdot \vec{v}$$

for all $\vec{v} \in \mathbb{R}^n$.

Proof. Let $f : \mathbb{R}^n \to \mathbb{R}$ be a function that is differentiable at the point $\vec{a}$, and let $L : \mathbb{R}^n \to \mathbb{R}$ be the unique affine function such that $f(\vec{a}) = L(\vec{a})$ and

$$\lim_{\vec{u}\to\vec{0}} \frac{f(\vec{a}+\vec{u}) - L(a+\vec{u})}{||\vec{u}||} = 0. \tag{4.4}$$

By Theorem 4.3.1, $D_{\vec{a}} L(\vec{v}) = \nabla L(\vec{a}) \cdot \vec{v}$ for all $\vec{v}$.

Now, let us recall a fact about limits. If h, g are functions and $\lim_{x\to a} h(x)$ exists and $\lim_{x\to a}(h(x) + g(x))$ exists, then we must have $\lim_{x\to a} g(x)$ exists and

$$\lim_{x\to a} \Big(h(x) + g(x)\Big) = \lim_{x\to a} h(x) + \lim_{x\to a} g(x).$$

Putting it all together, Equation (4.4) implies for any $\vec{v}$,

$$0 = 0||\vec{v}|| = ||\vec{v}|| \lim_{h\to 0} \frac{f(\vec{a}+h\vec{v}) - L(a+h\vec{v})}{||h\vec{v}||} = \lim_{h\to 0} \frac{f(\vec{a}+h\vec{v}) - L(a+h\vec{v})}{h}$$

exists. In particular, this shows that $D_{\vec{a}}(f - L)(\vec{v}) = 0$. Applying our fact about limits, since $D_{\vec{a}} L(\vec{v})$ exists and $D_{\vec{a}}(f - L)(\vec{v})$ exists, we must have that $D_{\vec{a}} f(\vec{v})$ exists and that

$$D_{\vec{a}} f(\vec{v}) - D_{\vec{a}} L(\vec{v}) = 0.$$

In other words,

$$D_{\vec{a}} f(\vec{v}) = D_{\vec{a}} L(\vec{v}) = \nabla L(\vec{a}) \cdot \vec{v}. \tag{4.5}$$

[6] If you already know how to evaluate multi-dimensional limits, this isn't much of a problem.

To finish off the proof, notice that

$$\frac{\partial L}{\partial x}(\vec{a}) = D_{\vec{a}}L(\hat{\mathbf{x}}) = D_{\vec{a}}f(\hat{\mathbf{x}}) = \frac{\partial f}{\partial x}(\vec{a}).$$

Repeating this argument with $\hat{\mathbf{y}}$, $\hat{\mathbf{z}}$, etc. shows that the partial derivatives of f at $\vec{a}$ must equal the partial derivatives of L at $\vec{a}$. Therefore,

$$\nabla L(\vec{a}) = \nabla f(\vec{a}).$$

Substituting $\nabla L(\vec{a})$ with $\nabla f(\vec{a})$ in Equation (4.5) completes the proof. ■

Theorem 4.3.2 shows that if a function is differentiable, directional derivatives can be computed with dot products instead of limits!

■ **Example 4.14** Find the directional derivative of the function $f : \mathbb{R}^3 \to \mathbb{R}$ given by $f(x,y,z) = x^2 - z^3$ at the point $\vec{a} = (1,1,1)$ in the direction $\vec{v} = (1,2,3)$.

Since f is a polynomial it is differentiable. Computing,

$$\nabla f(1,1,1) = \begin{bmatrix} 2 \\ 0 \\ -3 \end{bmatrix}$$

and so

$$D_{\vec{a}}f(\vec{v}) = \nabla f(\vec{a}) \cdot \vec{v} = -7.$$

■

Geometric Interpretation of the Gradient

Suppose $f : \mathbb{R}^n \to \mathbb{R}$ is a differentiable function and $\vec{u}$ is a unit vector. We know that

$$D_{\vec{a}}f(\vec{u}) = \nabla f(\vec{a}) \cdot \vec{u} = \|\nabla f(\vec{a})\|\|\vec{u}\| \cos\theta = \|\nabla f(\vec{a})\| \cos\theta,$$

where θ is the angle between $\vec{u}$ and $\nabla f(\vec{a})$. Since $\cos\theta \le 1$, we see that $D_{\vec{a}}f(\vec{u})$ is maximized when $\theta = 0$. In other words, *the largest directional derivative of f occurs in the direction of the gradient of f*. Thus, the gradient may be interpreted as pointing in the direction of greatest change.

XXX Figure

Continuing this further, if $D_{\vec{a}}f(\vec{u}) = 0$ then one of two things must occur. Either $\nabla f(\vec{a}) = \vec{0}$ or $\theta = 90^\circ$. Therefore, if $\nabla f(\vec{a}) \neq \vec{0}$, the direction in which the function doesn't change at all is orthogonal to the gradient.

We have another word for paths on which a function doesn't change–level curves. If we plot level curves for a function, the gradient of the function must always be orthogonal to the level curves.

XXX Figure

Exercises for 4.3

4.4 Multidimensional Limits and Continuity

Most users of mathematics don't worry about things that might go wrong with the functions they use to represent physical quantities. They tend to assume that functions are differentiable when derivatives are called for (except possibly for a finite set of isolated points), and they assume all functions which need to be integrated are continuous so the integrals will exist. For much of the period during which Calculus was developed (during the 17th and 18th centuries), mathematicians also did not bother themselves with such matters. Unfortunately, during the 19th century, mathematicians discovered that general functions could behave in unexpected and subtle ways, so they began to devote much more time to careful formulation of definitions and careful proofs in analysis. This is an aspect of mathematics which is covered in courses in real analysis, so we won't devote much time to it in this course. (You may have noticed that we didn't worry about the existence of derivatives in our discussion of velocity and acceleration). However, for functions of several variables, lack of rigor can be more troublesome than in the one variable case, so we briefly devote some attention to such questions. In this section, we shall discuss the concepts of *limit* and *continuity* for functions $f : \mathbb{R}^2 \to \mathbb{R}$. The big step, it turns out, is going from one independent variable to two. Once you understand that, going to three or more independent variables introduces few additional difficulties.

Let $\vec{r}_0 = (x_0, y_0)$ be (the position vector of) a point in the domain of the function f. We want to define the concept to be expressed symbolically

$$\lim_{\vec{r}\to\vec{r}_0} f(\vec{r}) = L \qquad \text{or} \qquad \lim_{(x,y)\to(x_0,y_0)} f(x,y) = L.$$

We start with two examples which illustrate the concept and some differences from the single variable case.

■ **Example 4.15** Let $f(x,y) = x^2 + 2y^2$, and consider the nature of the graph of f near the point $(1,2)$. As we saw in the previous section, the graph is an elliptic paraboloid, the locus of $z = x^2 + 2y^2$. In particular, the surface is quite smooth, and if (x,y) is a point in the domain *close to* $(1,2)$, then $f(x,y)$ will be very close to the value of the function there, $f(1,2) = 1^2 + 2(2^2) = 9$. Thus, it makes sense to assert that

$$\lim_{(x,y)\to(1,2)} x^2 + 2y^2 = 9.$$

XXX Figure ■

In Example 4.15, the limit was determined simply by evaluating the function at the desired point. You may remember that in the single variable case, you cannot always do that. For example, putting $x = 0$ in $\sin x/x$ yields the meaningless expression $0/0$, but $\lim_{x\to 0} \sin x/x$ is known to be 1. Usually, it requires some ingenuity to find such examples in the single variable case, but the next example shows that fairly simple formulas can lead to unexpected difficulties for functions of two or more variables.

■ **Example 4.16** Let

$$f(x,y) = \frac{x^2 - y^2}{x^2 + y^2} \qquad \text{for } (x,y) \neq (0,0).$$

What does the graph of this function look like in the vicinity of the point $(0,0)$? (Since, $(0,0)$ is not in the domain of the function, it does not make sense to talk about $f(0,0)$, but we can still seek a "limit.") The easiest way to answer this question is to switch to polar coordinates. Using $x = r\cos\theta$, $y = r\sin\theta$, we find

$$f(\vec{r}) = f(x,y) = \frac{r^2\cos^2\theta - r^2\sin^2\theta}{r^2\cos^2\theta + r^2\sin^2\theta} = \cos^2\theta - \sin^2\theta = \cos 2\theta.$$

Thus, $f(\vec{r}) = f(x,y)$ is independent of the polar coordinate r and depends only on θ. As $r = \|\vec{r}\| \to 0$ with θ fixed, $f(\vec{r})$ is constant, and equal to $\cos 2\theta$, so, if it 'approaches' a limit, that limit would have to be $\cos 2\theta$. Unfortunately, $\cos 2\theta$ varies between -1 and 1, so it does not make sense to say $f(\vec{r})$ has a limit as $\vec{r} \to \vec{0}$. You can get some idea of what the graph looks like by studying the level curves which are pictured in the diagram. For each value of θ, the function is constant, so the level curves consist of rays emanating from the origin, as indicated. On any such ray, the graph is at some constant height z with z taking on *every value* between -1 and $+1$. ■

In general, the statement

$$\lim_{\vec{r}\to\vec{r}_0} f(\vec{r}) = L$$

will be taken to mean that $f(\vec{r})$ is *close to* L whenever $\vec{r}$ is *close to* $\vec{r}_0$. To make this completely rigorous, we will give an "ε-δ definition" of limit.

Definition 4.4.1 — Limit. Let $f : \mathbb{R}^n \to \mathbb{R}$ be a function. For a point $\vec{a}$, we say *the limit of f as $\vec{x}$ approaches $\vec{a}$ is $L \in \mathbb{R}$*, written

$$\lim_{\vec{x}\to\vec{a}} f(\vec{x}) = L,$$

if for all numbers $\varepsilon > 0$ there exists a number $\delta > 0$ so that whenever

$$0 < \|\vec{x} - \vec{a}\| < \delta \qquad \text{we have} \qquad |f(\vec{x}) - L| < \varepsilon.$$

In this statement, $0 < \|\vec{x} - \vec{a}\| < \delta$ asserts that the distance from $\vec{x}$ to $\vec{a}$ is less than δ. Since δ is thought of as small, the inequality makes precise the meaning of "$\vec{x}$ is close to $\vec{a}$, but not equal to it." Similarly, $|f(\vec{x}) - L| < \epsilon$ catches the meaning of "$f(\vec{x})$ is close to L." Note that we never consider the case $\vec{x} = \vec{a}$, so the value of $f(\vec{r}_0)$ is not relevant in checking the limit as $\vec{x} \to \vec{a}$. (It is not even necessary that $f(\vec{r})$ be well defined at $\vec{x} = \vec{a}$.)

Limits for functions of several variables behave formally much the same as limits for functions of one variable. Thus, you may calculate the limit of a sum by taking the sum of the limits, and similarly for products and quotients (except that for quotients the limit of the denominator should not be zero). The understanding you gained of these matters in the single variable case should be an adequate guide to what to expect for several variables. If you never really understood all this before, we won't enlighten you much here. You will have to wait for a course in real analysis for real understanding.

Continuity

In Example 4.15, the limit was determined simply by evaluating the function at the point. This is certainly not always possible because the value of the function may be irrelevant or there may be no meaningful way to attach a value. Functions for which it is always possible to find the limit this way are called *continuous*. (This is the same notion as for functions of a single scalar variable). More precisely, we say that f is continuous at a point $\vec{r}_0$ if the point is in its domain (i.e., $f(\vec{r}_0)$ is defined) and

$$\lim_{\vec{r}\to\vec{r}_0} f(\vec{r}) = f(\vec{r}_0).$$

Points at which this fails are called *discontinuities* or sometimes *singularities*. (The latter term is also sometimes reserved for less serious kinds of mathematical pathology.) It sometimes happens, that a function f has a well defined limit L at a point $\vec{r}_0$ which does not happen to be in the domain of the function, i.e., $f(\vec{r}_0)$ is not defined. (In the single variable case, $\sin x/x$ at $x = 0$ is a good example). Then we can extend the domain of the function to include the point $\vec{r}_0$ by defining $f(\vec{r}_0) = L$. Thus the original function had a discontinuity, but it can be eliminated simply by extending the definition of the function. In this case, the discontinuity is called *removable*. As Example 4.16 shows, there are functions with discontinuities which cannot be defined away no matter what you try.

A function without discontinuities is called continuous. Continuous functions have graphs which look reasonably smooth. They don't have big holes or sudden jumps, but as we shall see later, they can still look pretty bizarre. Usually, just knowing that a function is continuous won't be enough to make it a good candidate to represent a physical quantity. We shall also want to be able to take derivatives and do the usual things one does in differential calculus, but as you might expect, this is somewhat more involved than it is in the single variable case.

Exercises for 4.4

4.5 Polynomial Approximation

When we try to model the real world with mathematics, we often encounter very complicated functions. The more complicated a function, the more difficult it is to analyze, so a common mathematical strategy is to substitute complicated functions for simple ones.

Among the simplest mathematical functions are *polynomials*, and in the multivariable context, this means polynomials of multiple variables. Before we dive into the details of polynomial approximations, let us introduce some terminology.

Definition 4.5.1 — Polynomial. Let $x_1, x_2, \ldots, x_n$ denote variables. A *degree-m term* is a product of the form

$$x_1^{\alpha_1} x_2^{\alpha_2} \cdots x_n^{\alpha_n}$$

where $\alpha_i \geq 0$ and $\sum_{i=1}^{n} \alpha_i = m$.

A *degree-m polynomial* is a linear combination of terms of degree at most m. A *degree-0 polynomial* is constant.

In the definition of polynomial, we always interpret x_i^0 as 1. You're intimately familiar with polynomials of a single variable. For instance, $x^3 - 2x$ is a degree-3 polynomial. However, when it comes to the multivariable context, $x^2 y - 2x$ and $xyz - 2x$ are also degree-3 polynomials.

Affine functions are all degree-1 polynomials, and the definition of $f : \mathbb{R}^n \to \mathbb{R}$ being differentiable depended on being able to find an affine function that approximated f very well. There is a similar concept for higher degree polynomial functions.

Definition 4.5.2 — Taylor Approximation. Given a continuous function $f : \mathbb{R}^n \to \mathbb{R}$, a degree m polynomial p, and a point $\vec{a} \in \mathbb{R}^n$, we call p the *degree-m Taylor approximation of f at $\vec{a}$* if

$$\lim_{\vec{r}\to\vec{0}} \frac{f(\vec{a}+\vec{r})-p(\vec{a}+\vec{r})}{||\vec{r}||^m} = 0.$$

Note that if $||\vec{r}|| < 1$, then $||\vec{r}|| > ||\vec{r}||^2 > \cdots > ||\vec{r}||^m$. Thus,

$$\frac{f(\vec{a}+\vec{r})-p(\vec{a}+\vec{r})}{||\vec{r}||^m} < \cdots < \frac{f(\vec{a}+\vec{r})-p(\vec{a}+\vec{r})}{||\vec{r}||},$$

and so having a degree-m Taylor approximation is a bigger demand that merely being differentiable.

■ **Example 4.17** Let $f(x,y) = (x-2)^2+y^2$. Find the degree 1, 2, and 3 Taylor approximations to f at $(0,0)$.

On the one hand, since f is itself a polynomial, we know it is differentiable. Therefore it has an affine approximation coming from the definition of differentiability which coincides with a degree-1 Taylor approximation. Further f is a degree-2 polynomial and so it is its own degree-2 and degree-3 Taylor approximation.

On the other hand, let's imagine we weren't so clever. Let $p_1(x,y) = ax+by+c$ be a degree-1 polynomial. If p is a degree-1 Taylor approximation to f at $\vec{0}$, then

$$\lim_{\vec{r}\to\vec{0}} \frac{f(\vec{0}+\vec{r})-p_1(\vec{0}+\vec{r})}{||\vec{r}||} = 0.$$

Let r_x and r_y stand for the x and y components of $\vec{r}$. Expanding, we see

$$\begin{aligned}\lim_{\vec{r}\to\vec{0}} \frac{f(\vec{0}+\vec{r})-p_1(\vec{0}+\vec{r})}{||\vec{r}||} &= \lim_{\vec{r}\to\vec{0}} \frac{(r_x-2)^2+r_y^2-(ar_x+br_y+c)}{\sqrt{r_x^2+r_y^2}} \\ &= \lim_{\vec{r}\to\vec{0}} \frac{r_x^2-4r_x+4+r_y^2-ar_x-br_y-c}{\sqrt{r_x^2+r_y^2}} = 0\end{aligned}$$

if and only if $a = -4$, $b = 0$, and $c = 4$. Thus $p_1(x,y) = -4x+4$ is the unique degree-1 Taylor approximation to f.

We can use a similar process to find the degree-2 Taylor approximation to f. Let $p_3(x,y) = ax^2+by^2+cxy+dx+ey+h$ be a degree-2 polynomial. Then

$$\begin{aligned}\lim_{\vec{r}\to\vec{0}} \frac{f(\vec{0}+\vec{r})-p_3(\vec{0}+\vec{r})}{||\vec{r}||^2} &= \lim_{\vec{r}\to\vec{0}} \frac{(r_x-2)^2+r_y^2-(ar_x^2+br_y^2+cr_xr_y+dr_x+er_y+h)}{r_x^2+r_y^2} \\ &= \lim_{\vec{r}\to\vec{0}} \frac{r_x^2-4r_x+4+r_y^2-(ar_x^2+br_y^2+cr_xr_y+dr_x+er_y+h)}{r_x^2+r_y^2} = 0\end{aligned}$$

if and only if $a=1$, $b=1$, $c=0$, $d=-4$, $e=0$, and $h=4$. Therefore the degree-2 Taylor approximation of f is $p_3(x,y)=x^2-4x+4+y^2=(x-2)^2+y^2$, which is just f itself.

We could continue on to use the limit definition to find a degree-3 Taylor approximation to f, or we could notice that $f(x,y)-p_2(x,y)=0$ for all x and y. Since $\lim_{t\to 0} 0/t=0$ and p_2 satisfies the definition of a degree-3 polynomial (none of its terms exceed degree 3), p_2 must be a degree-3 Taylor approximation of f. Further, if we used the limit definition, we would conclude that p_2 is actually the *unique* degree-3 Taylor approximation of f. ■

There's a faster way to find degree-m Taylor approximations than taking a limit each time. We can rely on higher-order partial derivatives.

Higher-order Partial Derivatives

Recall that if $f:\mathbb{R}^n\to\mathbb{R}$ is a differentiable function, then $\frac{\partial f}{\partial x_i}(\vec{a})$ is a well-defined number for all $\vec{a}\in\mathbb{R}^n$. Thus, we may view $\frac{\partial f}{\partial x_i}(\vec{a})$ as a function of $\vec{a}$. From this perspective,

$$\frac{\partial f}{\partial x_i}:\mathbb{R}^n\to\mathbb{R}$$

is just another multivariable function. If this function is differentiable, we may consider

$$\frac{\partial \frac{\partial f}{\partial x_i}}{\partial x_j}(\vec{a}),$$

and so on. A function is called an *mth order partial derivative* if this process has been repeated m times. The nesting fraction notation can get pretty awkward, pretty quickly, so, modeling after the notation from single-variable calculus, we write

$$\frac{\partial^2 f}{\partial x_j \partial x_i}(\vec{a})=\frac{\partial \frac{\partial f}{\partial x_i}}{\partial x_j}(\vec{a}).$$

The expression $\frac{\partial^2 f}{\partial x_i \partial x_j}(\vec{a})$ is an *iterated partial derivative*. If $x_i\neq x_j$, it is called a *mixed partial derivative*. If $x_i=x_j$, we further simplify our notation and write

$$\frac{\partial^2 f}{\partial x_i^2}(\vec{a}).$$

■ **Example 4.18** Compute $\frac{\partial^2 f}{\partial x \partial y}(x,y)$ where $f(x,y)=e^{x-2xy}$.

XXX Finish ■

Strictly speaking, $\frac{\partial^2 f}{\partial x_j \partial x_i}(\vec{a})$ means take a derivative with respect to x_i first and then with respect to x_j, and a priori, it could be that $\frac{\partial^2 f}{\partial x_j \partial x_i}(\vec{a})\neq\frac{\partial^2 f}{\partial x_i \partial x_j}(\vec{a})$. However, if our function has a degree-2 Taylor approximation, these two quantities are always equal.

Theorem 4.5.1 — Clairaut's Theorem. If $f : \mathbb{R}^n \to \mathbb{R}$ has a degree-2 Taylor approximation at $\vec{a} \in \mathbb{R}^n$, then

$$\frac{\partial^2 f}{\partial x_j \partial x_i}(\vec{a}) = \frac{\partial^2 f}{\partial x_i \partial x_j}(\vec{a})$$

for all choices of variables x_i, x_j.

Clairaut's Theorem will apply to most functions we encounter, but some fairly reasonable looking functions can fail.

■ **Example 4.19** Consider $f(x,y) = \begin{cases} \frac{xy(x^2-y^2)}{x^2+y^2} & \text{if } (x,y) \neq (0,0) \\ 0 & \text{if } (x,y) = (0,0) \end{cases}$.

XXX Example where mixed partials aren't equal ■

Higher order partial derivatives are interesting in their own right, but the following theorem makes them useful for finding Taylor approximations.

Theorem 4.5.2 Suppose $f : \mathbb{R}^n \to \mathbb{R}$ has a degree-m Taylor approximation $p : \mathbb{R}^n \to \mathbb{R}$ at $\vec{a} \in \mathbb{R}^n$. Then, every ith-order partial derivative of f at the point $\vec{a}$ agrees with every ith-order partial derivative of p at the point $\vec{a}$ for $i \leq m$. Symbolically,

$$\frac{\partial^i f}{\partial x_{j_1} \cdots \partial x_{j_i}}(\vec{a}) = \frac{\partial^i p}{\partial x_{j_1} \cdots \partial x_{j_i}}(\vec{a})$$

for $i \leq m$.

Theorem 4.5.2 is not difficult to prove, but it is a notational nightmare, so we won't prove it here. Theorem 4.5.2 also simplifies the task of finding a Taylor approximation (provided we know one already exists).

■ **Example 4.20** XXX Finish ■

Uses of Polynomial Approximations

As we have seen previously if $f : \mathbb{R}^n \to \mathbb{R}$ has a degree-1 Taylor approximation at $\vec{a}$, that approximation captures all information about directional derivatives of f at $\vec{a}$ (we used this fact to prove $D_{\vec{a}} f(\vec{v}) = \nabla f(\vec{a}) \cdot \vec{v}$). Similarly, a degree-$m$ Taylor approximation to f at $\vec{a}$ captures all m-th order derivative information. This statement can be made precise with a theorem.

Theorem 4.5.3 Let $f : \mathbb{R}^n \to \mathbb{R}$ be a function and let $p : \mathbb{R}^n \to \mathbb{R}$ be the degree-m Taylor approximation to f at $\vec{a}$. If $\vec{r} : \mathbb{R} \to \mathbb{R}^n$ is an m-times differentiable parameterization of a curve with $\vec{p}(0) = \vec{a}$, then

$$\begin{gathered}(f \circ \vec{r})'(0) = (p \circ \vec{r})'(0) \\ (f \circ \vec{r})''(0) = (p \circ \vec{r})''(0) \\ \vdots \\ (f \circ \vec{r})^{(m)}(0) = (p \circ \vec{r})^{(m)}(0).\end{gathered}$$

That is, the first m derivatives of $f \circ \vec{r}$ and $p \circ \vec{r}$ agree at 0.

The implications of Theorem 4.5.3 are far reaching. You can use polynomial approximations to study curvature, concavity, local minimums and maximums, etc.—and we will.

4.6 Optimization

Optimization is a fancy word for finding minimums and maximums. If $f : \mathbb{R} \to \mathbb{R}$ is a function of one variable, the procedure should be familiar: find critical points of f and check whether they are minimums, maximums, or neither; then, check the boundary. For a multivariable function, the procedure is similar.

Let's consider an example. Let $f : \mathbb{R}^2 \to \mathbb{R}$ be defined by $f(x, y) = 4 - x^2 - y^2$. This example is nice because we already know where the maximum occurs!

XXX Figure

A global maximum of 4 for f occurs at $(0, 0)$. Now, suppose $\vec{p} : \mathbb{R} \to \mathbb{R}^2$, given by

$$\vec{p}(t) = t\vec{d} + \vec{0}$$

is the parameterization of a line passing through $(0, 0)$. Since 4 is a global maximum of f achieved at $(0, 0)$ and $\vec{p}$ passes through $(0, 0)$, the function $f \circ \vec{p}$ must attain a global maximum of 4 exactly when $\vec{p}(t) = \vec{0}$.

XXX Figure

Since $f \circ \vec{p} : \mathbb{R} \to \mathbb{R}$ is a function of one variable, all of its maximums must occur at either a critical point or a boundary point of the domain. Since the domain is unbounded and $(f \circ \vec{p})'$ always exists, we know a maximum of $f \circ \vec{p}$ must occur when

$$(f \circ \vec{p})'(t) = 0. \tag{4.6}$$

If both f and $\vec{p}$ were single-variable functions, we could apply the chain rule to Equation (4.6). Unfortunately they are not. So, before we continue, let's see if we can figure out a multivariable chain rule.

The Multivariable Chain Rule

In mathematics, when you don't know the answer, a technique[7] is to replace your objects of study with something simpler. In differential calculus, the simplest thing you could hope for is something flat (a line, a plane, etc.). These are precisely characterized by affine functions.

[7] In math, some things are called tricks and some techniques. The difference is: a technique is used more than once.

Suppose $f : \mathbb{R}^2 \to \mathbb{R}$ is an affine function and $\vec{p} : \mathbb{R} \to \mathbb{R}^2$ is a differentiable parameterization of a curve. Since f is an affine function, it can be written as $f(x,y) = \alpha_x x + \alpha_y y + c$ or alternatively as $f(\vec{x}) = \vec{\alpha} \cdot \vec{x} + c$. Let $p_x : \mathbb{R} \to \mathbb{R}$ and $p_y : \mathbb{R} \to \mathbb{R}$ be the components of $\vec{p}$. Now we have

$$f \circ \vec{p}(t) = f(\vec{p}(t)) = f\big(p_x(t), p_y(t)\big) = \alpha_x p_x(t) + \alpha_y p_y(t) + c.$$

This is a function of one variable so we can take a derivative as usual:

$$(f \circ \vec{p})'(t) = \alpha_x p_x'(t) + \alpha_y p_y'(t) = \vec{\alpha} \cdot \vec{p}\,'(t).$$

But, since f is an affine function, $\nabla f(\vec{x}) = \vec{\alpha}$ for all $\vec{x}$, and so we in fact have

$$(f \circ \vec{p})'(t) = \nabla f(\vec{x}) \cdot \vec{p}\,'(t). \tag{4.7}$$

For an affine function, ∇f is constant, so we don't need to specify what $\vec{x}$ is for Equation (4.7) to make sense. However, without knowing what $\vec{x}$ should be, it's difficult to see the *meaning* behind Equation (4.7).

Let's consider a more general case. Let $f : \mathbb{R}^2 \to \mathbb{R}$ be a function that is differentiable at the point $\vec{a}$ and let $\vec{p} : \mathbb{R} \to \mathbb{R}^2$ be the parameterization of a curve so that $\vec{p}(t_0) = \vec{a}$. Since f is differentiable at $\vec{a}$, we know there is an affine function $L : \mathbb{R}^2 \to \mathbb{R}$ that approximates f very well near the point $\vec{a}$. This means,

$$f \circ \vec{p}(t) \approx L \circ \vec{p}(t) \qquad \text{when} \qquad t \approx t_0.$$

From Equation (4.7), we know $(L \circ \vec{p})'(t) = \nabla L(\vec{x}) \cdot \vec{p}\,'(t)$ for any $\vec{x}$. Further, $\nabla L(\vec{a}) = \nabla f(\vec{a})$ and $\vec{a} = \vec{p}(t_0)$. Putting this all together,

$$\begin{aligned}(L \circ \vec{p})'(t_0) &= \nabla L(\vec{a}) \cdot \vec{p}\,'(t_0) \\ &= \nabla f(\vec{p}(t_0)) \cdot \vec{p}\,'(t_0) = D_{\vec{p}(t_0)} f(\vec{p}\,'(t_0)).\end{aligned}$$

That equation is a mouthful, but the right hand side involves only f and $\vec{p}$. If we could show that $(L \circ \vec{p})'(t_0) = (f \circ \vec{p})'(t_0)$, we would have a formula for the multivariable chain rule. Fortunately, we have the following theorem.

Theorem 4.6.1 Let $f : \mathbb{R}^n \to \mathbb{R}$ be a function that is differentiable at $\vec{a}$ and let $L_{\vec{a}} : \mathbb{R}^n \to \mathbb{R}$ be the affine function coming from the definition of differentiability. Let $\vec{p} : \mathbb{R} \to \mathbb{R}^n$ satisfy $\vec{p}(t_0) = \vec{a}$. Then,

$$(f \circ \vec{p})'(t_0) = (L_{\vec{a}} \circ \vec{p})'(t_0).$$

We will not prove Theorem 4.6.1 in the general context. However, if $\vec{p}$ parameterizes a line, the proof is straightforward. Suppose $\vec{p}$ parameterizes a line, and for simplicity, assume $t_0 = 0$. Then $\vec{p}(t) = \vec{a} + t\vec{v}$ for some $\vec{v}$. Now,

$$(f \circ \vec{p})'(0) = D_{\vec{a}} f(\vec{v}) \qquad \text{and} \qquad (L_{\vec{a}} \circ \vec{p})'(0) = D_{\vec{a}} L_{\vec{a}}(\vec{v}),$$

and we have previously established that $D_{\vec{a}} f(\vec{v}) = D_{\vec{a}} L_{\vec{a}}(\vec{v})$. Proving Theorem 4.6.1 in the general case amounts to showing that you can replace $\vec{p}$ by *its* affine approximation, which parameterizes a line.

We may now write down the multivariable chain rule.

Theorem 4.6.2 — Multivariable Chain Rule. If $f : \mathbb{R}^n \to \mathbb{R}$ is differentiable at $\vec{a}$ and $\vec{p} : \mathbb{R} \to \mathbb{R}^n$ is differentiable and satisfies $\vec{p}(t_0) = \vec{a}$, then

$$(f \circ \vec{p})'(t_0) = D_{\vec{p}(t_0)} f\left(\vec{p}\,'(t_0)\right).$$

Equivalently,

$$(f \circ \vec{p})'(t_0) = \nabla f\left(\vec{p}(t_0)\right) \cdot \vec{p}\,'(t_0).$$

The Multivariable First-Derivative Test

We can now develop the multivariable first derivative test for local extrema. Recall that if $f : \mathbb{R}^n \to \mathbb{R}$ and $\vec{p} : \mathbb{R} \to \mathbb{R}^n$ is a parameterization such that f attains a local maximum at $\vec{p}(0)$, then

$$(f \circ \vec{p})'(0) = 0.$$

Using the multivariable chain rule we see

$$(f \circ \vec{p})'(0) = D_{\vec{p}(0)} f\left(\vec{p}'(0)\right) = 0.$$

This is true for any parameterization that passes through a local extreme. Since we can pass through the local extreme from any direction, we have the following theorem.

Theorem 4.6.3 — Multivariable First Derivative Test. If $f : \mathbb{R}^n \to \mathbb{R}$ is differentiable at $\vec{a} \in \mathbb{R}^n$ and f attains a local maximum or local minimum at $\vec{a}$ then

$$D_{\vec{a}} f(\vec{v}) = 0$$

for all vectors $\vec{v}$. Equivalently,

$$\nabla f(\vec{a}) = \vec{0}.$$

Note that Theorem 4.6.3 does not say that if $\nabla f(\vec{a}) = \vec{0}$ then f *must* have a local max or min at $\vec{a}$. It only says that if you have a local max or min at $\vec{a}$, then you have $\nabla f(\vec{a}) = \vec{0}$.

This is just like the single-variable first derivative test. If $g : \mathbb{R} \to \mathbb{R}$ and $g'(a) = 0$, it need not be the case that g has a local max or min at a.

XXX Figure

Definition 4.6.1 — Critical Point. For a function $f : \mathbb{R}^n \to \mathbb{R}$, the point $\vec{a} \in \mathbb{R}^n$ is called a *critical point* if f is not differentiable at $\vec{a}$ or if $\nabla f(\vec{a}) = \vec{0}$.

■ **Example 4.21** Let $f : \mathbb{R}^2 \to \mathbb{R}$ be given by $f(x, y) = 4 - x^2 - y^2$. Find all critical points of f.

Computing, $\nabla f(x, y) = (2x, -2y)$. Since f is differentiable everywhere, the only critical points of f must occur when $\nabla f(x, y) = (0, 0)$. Thus the only critical points are at $(x, y) = (0, 0)$. ■

■ **Example 4.22** XXX Finish (Find Local Max) ■

■ **Example 4.23** XXX Finish Find (Local saddle point) ■

The Multivariable Second-Derivative Test

Inspired by the single-variable second-derivative test, we will create a multivariable second-derivative test. First, a reminder about the single-variable case.

Theorem 4.6.4 — Single-variable Second-Derivative Test. Suppose $f : \mathbb{R} \to \mathbb{R}$ is a twice differentiable function and $f'(a) = 0$. Then, if $f''(a) > 0$, it must be that $f(a)$ is a local minimum and if $f''(a) < 0$, it must be that $f(a)$ is a local maximum. If $f''(a) = 0$, nothing can be concluded about whether $f(a)$ is a local maximum, local minimum, or neither.

Now, imagine $f : \mathbb{R}^n \to \mathbb{R}$ has a local maximum at $\vec{a}$. Let $\vec{u}$ be a unit vector and let ℓ be the line with direction $\vec{u}$ passing through $\vec{a}$. Then, ℓ has a parameterization $\vec{p}(t) = a + t\vec{u}$. The function $f \circ p$ can be thought of as the restriction of f to ℓ. Since f has a local max at $\vec{a}$, we know $f \circ \vec{p}$ has a local max at 0 (since $\vec{p}(0) = \vec{a}$), and we would expect $f \circ \vec{p}$ to be a concave-down function.

XXX Figure

Thus, we expect $(f \circ \vec{p})''(0) < 0$. We call $(f \circ \vec{p})''(0)$ a *second order directional derivative in the direction $\vec{u}$*. This inspires the multivariable second derivative test.

Theorem 4.6.5 — Multivarialbe Second-Derivative Test. Suppose $f : \mathbb{R}^n \to \mathbb{R}$ is differentiable at $\vec{a}$ and $D_{\vec{a}}f(\vec{v}) = 0$ for all $\vec{v}$. Let $\vec{p}_{\vec{v}}(t) = \vec{a} + t\vec{v}$. If

$$(f \circ \vec{p}_{\vec{v}})''(0) < 0$$

for all $\vec{v}$, then f has a local maximum at $\vec{a}$ and if

$$(f \circ \vec{p}_{\vec{v}})''(0) > 0$$

for all $\vec{v}$, then f has a local minimum at $\vec{a}$. Further, if there are $\vec{v}_1$ and $\vec{v}_2$ such that

$$(f \circ \vec{p}_{\vec{v}_1})''(0) < 0 < (f \circ \vec{p}_{\vec{v}_2})''(0),$$

then f has neither a local minimum or a local maximum at $\vec{a}$. In any other case, $f(\vec{a})$ may have a local minimum, a local maximum, or neither.

The multivariable second-derivative test is more complicated because more things can happen in higher dimensions. Consider the following examples.

■ **Example 4.24** $f(x, y) = x^2 + y^2$

XXX Finish ■

■ **Example 4.25** $f(x, y) = x^2 - y^2$

XXX Finish ■

■ **Example 4.26** $f(x, y) = x^3 + y^3$

XXX Finish ■

For complicated functions, using the multivariable second-derivative test involves a lot of computation. Fortunately, we can use Theorem 4.5.3 to replace our function with a degree-2 Taylor approximation at a critical point. Since all second-order directional derivatives of our function and our Taylor approximation will agree, we can instead use the Taylor approximation to determine whether we have a local minimum, local maximum, or neither.

■ **Example 4.27** $f(x,y) = \frac{1}{1+x^2+y^2}$

XXX Finish ■

Lagrange Multipliers

We've just explored how to find local minimums and maximums on an unconstrained domain, but what if we wish to restrict the domain.

■ **Example 4.28** Let $f(x,y) = (x-2)^2 + (y-3)^2$ and let $\mathcal{D}$ be the disk of radius one centered at $\vec{0}$. We wish to find the minimum of $f\big|_{\mathcal{D}}$. That is, we wish to find the minimum of f restricted to the disk $\mathcal{D}$.

Computing, the only critical point of f occurs at $(2,3)$ which is outside of $\mathcal{D}$. We can conclude that $f\big|_{\mathcal{D}}$ does not have a minimum on the interior of $\mathcal{D}$, but $f\big|_{\mathcal{D}}$ still may have a minimum on the boundary of $\mathcal{D}$.

The boundary of $\mathcal{D}$ is the unit circle which can be parameterized by $\vec{r}(t) = (\cos t, \sin t)$. Thus, we need to find the minimum of

$$\begin{aligned} f \circ \vec{r}(t) = (\cos t - 2)^2 + (\sin t - 3)^2 &= \cos^2 t + \sin^2 t - 6\sin t - 4\cos t + 13 \\ &= -6\sin t - 4\cos t + 14. \end{aligned}$$

Since this is a single-variable differentiable function, we need only find the places where $(f \circ \vec{r})'(t) = 0$. In other words, we need to solve

$$4\sin t - 6\cos t = 0.$$

This simplifies to $\tan t = 3/2$ and so the equation is satisfies when $t = t_1 = \arctan\frac{3}{2}$ or $t = t_2 = \pi + \arctan\frac{3}{2}$.

Evaluating, $f \circ \vec{r}(t_1) = 14 - 2\sqrt{13}$ and $f \circ \vec{r}(t_2) = 14 + 2\sqrt{13}$, and so $f\big|_{\mathcal{D}}$ attains a minimum of $14 - 2\sqrt{13}$ at $(\cos t_1, \sin t_1) = (\frac{2}{\sqrt{13}}, \frac{3}{\sqrt{13}})$. ■

Example 4.29 involved complicated, but doable, trigonometry. But, we can make our lives much easier by thinking geometrically for a while.

Let $f : \mathbb{R}^n \to \mathbb{R}$ and suppose $\mathcal{R}$ is a region whose boundary is parameterized by $\vec{r} : \mathbb{R} \to \mathbb{R}^n$. To find the extrema of $f\big|_{\mathcal{R}}$, we first check for critical points in $\mathcal{R}$, and then we need to check for minimums and maximums along the boundary of $\mathcal{R}$. Since $\vec{r}$ parameterizes the boundary of $\mathcal{R}$, we know

$$0 = (f \circ \vec{r})'(t) = D_{\vec{r}(t)} f\left(\vec{r}\,'(t)\right) = \nabla f(\vec{r}(t)) \cdot \vec{r}\,'(t)$$

whenever $\vec{r}(t)$ is at a local minimum or local maximum along the boundary. Now, observe that $\vec{r}\,'(t)$ is one of many vectors tangent to the boundary of $\mathcal{R}$ at the point $\vec{r}(t)$. Let $\vec{T}_{\vec{r}(t)}$ be an arbitrary tangent vector to the boundary of $\mathcal{R}$. We now have that

$$0 = D_{\vec{r}(t)} f\left(\vec{T}_{\vec{r}(t)}\right) = \nabla f(\vec{r}(t)) \cdot \vec{T}_{\vec{r}(t)}$$

whenever $\vec{r}(t)$ is at a local minimum or local maximum along the boundary.

XXX Figure

We're going to generalize slightly more. Suppose $\vec{T}:\mathbb{R}^n\to\mathbb{R}^n$ is a function such that if $\vec{a}$ is on the boundary of $\mathcal{R}$, we have $\vec{T}(\vec{a})$ is a tangent vector to the boundary of $\mathcal{R}$. We won't care about how $\vec{T}(\vec{a})$ is defined if $\vec{a}$ is not on the boundary of $\mathcal{R}$.

XXX Figure of such a function

Now, if $f\big|_{\mathcal{R}}$ has a local extreme at some point $\vec{a}$ on the boundary of $\mathcal{R}$, we know

$$0=\nabla f(\vec{a})\cdot\vec{T}(\vec{a}).$$

This gives us another way to approach the problem of finding extrema of $f\big|_{\mathcal{R}}$.

■ **Example 4.29** Let $f(x,y)=(x-2)^2+(y-3)^2$ and let $\mathcal{D}$ be the disk of radius one centered at $\vec{0}$. Find the minimum of $f\big|_{\mathcal{D}}$.

As before, we know that $f\big|_{\mathcal{D}}$ attains its minimum on the boundary of $\mathcal{D}$ which can be parameterized by $\vec{r}(t)=(\cos t,\sin t)$. Now, consider the function $\vec{T}(x,y)=(-y,x)$. $\vec{T}$ has the property that $\vec{T}(\vec{r}(t))=\vec{r}\,'(t)$. Thus, if

$$0=(f\circ\vec{r})'(t)$$

we must have

$$0=\nabla f(\vec{a})\cdot\vec{T}(\vec{a})$$

for some $\vec{a}=\vec{r}(t)$ on the boundary of $\mathcal{D}$. Let $\vec{a}=(x,y)$. Solving, we see

$$\begin{aligned}0=\nabla f(x,y)\cdot\vec{T}(x,y)&=\begin{bmatrix}2x-4\\2y-6\end{bmatrix}\cdot\begin{bmatrix}-y\\x\end{bmatrix}\\&=-2xy+4y+2xy-6x=-6x+4y\end{aligned}$$

and so $y=\frac{3}{2}x$. Finally, since $\vec{a}=(x,y)$ is on the unit circle we must have $x^2+y^2=1$ and so $(x,y)=(\frac{2}{\sqrt{13}},\frac{3}{\sqrt{13}})$ or $(x,y)=(-\frac{2}{\sqrt{13}},-\frac{3}{\sqrt{13}})$. Checking each of these points, we see $f\big|_{\mathcal{D}}$ attains a minimum at $(x,y)=(\frac{2}{\sqrt{13}},\frac{3}{\sqrt{13}})$. ■

Now, suppose we wish to find the extrema of $f\big|_{\mathcal{C}}$ and $\mathcal{C}$ can be expressed as a level curve of some function $h:\mathbb{R}^n\to\mathbb{R}$. Without loss of generality, we may assume that $\mathcal{C}=\{\vec{x}:h(\vec{x})=0\}$. (Make sure you understand why we can always assume the level curve takes the form $h(\vec{x})=0$.) Again, we have the same geometric intuition as before—we are looking for a place where $\nabla f\cdot\vec{T}=0$ where $\vec{T}$ is a tangent vector to $\mathcal{C}$.

Since $\mathcal{C}$ is a level set of h, we know every tangent vector to $\mathcal{C}$ must be orthogonal to ∇h. Thus we have $\nabla h\cdot\vec{T}=0$ and $\nabla f\cdot\vec{T}=0$, at a local extreme. Since this is true for any $\vec{T}$ that is a tangent vector at a local extreme, we must have that both ∇h and ∇f are normal vectors for the tangent line/tangent plane to $\mathcal{C}$ at the local extreme. Thus, ∇f and ∇h are parallel and so either $\nabla f=0$ or $\nabla h=0$ or $\nabla f=\lambda\nabla h$ for some $\lambda\neq 0$. This observation leads to the method of *Lagrange Multipliers*.

■ **Example 4.30** Let $f(x,y)=(x-2)^2+(y-3)^2$ and let $\mathcal{D}$ be the disk of radius one centered at $\vec{0}$. Find the minimum of $f\big|_{\mathcal{D}}$.

As before, we know that $f\big|_{\mathcal{D}}$ attains its minimum on the boundary of $\mathcal{D}$. Let $h(x,y) = x^2+y^2-1$. Then the boundary of $\mathcal{D}$ is precisely the level curve $h(x,y)=0$. We therefore seek to find all solutions to the system

$$\begin{cases} \nabla f(x,y) = \lambda \nabla h(x,y) \\ h(x,y) = 0 \end{cases} .$$

Computing ∇f and ∇h our system becomes

$$\begin{cases} 2x-4 = 2\lambda x \\ 2y-6 = 2\lambda y \\ x^2+y^2 = 1 \end{cases} .$$

We don't actually care about what λ is, so we can quickly eliminate it from the equations. Assuming that $x \neq 0$ and $y \neq 0$ we obtain

$$\frac{2x-4}{2x} = \lambda = \frac{2y-6}{2y}$$

and so $y = \frac{3}{2}x$. Since we also have $x^2+y^2=1$ we have candidate local extremes at $(x,y) = (\frac{2}{\sqrt{13}}, \frac{3}{\sqrt{13}})$ or $(x,y) = (-\frac{2}{\sqrt{13}}, -\frac{3}{\sqrt{13}})$. Now we must handle the case where $x=0$ or $y=0$. These cases give us candidate points of $(x,y)=(0,\pm 1)$ or $(x,y)=(\pm 1, 0)$. Testing these six points we see that a local minimum occurs at $(x,y) = (\frac{2}{\sqrt{13}}, \frac{3}{\sqrt{13}})$. ■

Exercises for 4.6

Chapter 5

Integrals

5.1 Multiple Integrals

Fundamentally, an integral is the result of chopping a region up into tiny pieces, adding all the pieces up again, and taking a limit as the size of the tiny pieces goes to zero. Thus far, the domain of this procedure—the thing we chop into tiny pieces—has been a line or curve. We will now consider integrating over multi-dimensional domains.

Consider the motivating example of finding the area of a region in the plane. Let $\mathcal{R} \subseteq \mathbb{R}^2$ be the region below the line $y = 1$ and above the curve $y = x^2$. We wish to find the area of $\mathcal{R}$.

Using the usual calculus strategy, we will chop $\mathcal{R}$ up into little rectangles of width Δx and height Δy. Then

$$\text{area of } \mathcal{R} \approx \sum_{\text{tiny rectangles}} \text{area of tiny rectangle}$$

and

$$\text{area of } \mathcal{R} = \lim_{\Delta x, \Delta y \to 0} \sum_{\text{tiny rectangles}} \text{area of tiny rectangle.}$$

XXX Figure

Using integral notation, we would write

$$\text{area of } \mathcal{R} = \int_{\mathcal{R}} \mathrm{d}A.$$

Here $\mathrm{d}A$ represents a "tiny area," the subscript $\mathcal{R}$ represents the region of integration, and the integral sign means we're adding things up. In this case, we're finding area, so $\mathrm{d}A = 1\mathrm{d}A$ is exactly what we're adding up. In other situations we'll be adding up more complicated functions.

This is all well and good, but how do we actually *find* the area. To do this, we'll need to convert $\int_{\mathcal{R}} \mathrm{d}A$ into a more traditional-looking integral—one that we know how to evaluate.

Let's write down our sum more carefully. We need to sum over all tiny rectangles that fit inside $\mathcal{R}$. To do so, we can take a systematic approach: let's sum all the rectangles in a column first and then sum up all the columns. The lower left corner of all rectangles in a single column share a common x-coordinate. Consider the column with lower left corner at $(x_0, ?)$. Counting, we see there are approximately $(1-x_0^2)/\Delta y$ such rectangles. Further, there are approximately $2/\Delta x$ columns. Therefore,

$$\text{area of } \mathcal{R} \approx \sum_{i=1}^{2/\Delta x} \sum_{j=1}^{(1-(-1+i\Delta x)^2)/\Delta y} \Delta y \Delta x.$$

That sum is really hard to parse, so we'll write it another way.

$$\text{area of } \mathcal{R} \approx \sum_{x_0=-1,-1+\Delta x,-1+2\Delta x,\dots,1} \sum_{y_0=x_0^2,x_0^2+\Delta y,x_0^2+2\Delta y,1} \Delta y \Delta x.$$

This is still hard to read, but it's looking more like an integral. The inner sum is adding up things from $y_0 = x_0^2$ to $y_0 = 1$ and the outer sum is adding up things from $x_0 = -1$ to $x_0 = 1$. Upon taking a limit, this directly translates to an integral, giving

$$\text{area of } \mathcal{R} = \int_{\mathcal{R}} \mathrm{d}A = \int_{x=-1}^{x=1} \int_{y=x^2}^{y=1} \mathrm{d}y\,\mathrm{d}x. \tag{5.1}$$

Now, we should take a moment to make sure we understand what we've just written. The right side of Equation (5.1) is an *iterated integral*. That is,

$$\int_{x=-1}^{x=1} \int_{y=x^2}^{y=1} \mathrm{d}y\,\mathrm{d}x = \int_{x=-1}^{x=1} \left(\int_{y=x^2}^{y=1} \mathrm{d}y \right) \mathrm{d}x$$

and so the integral with respect to y *must* be done before the integral with respect to x. To be clear, $\mathrm{d}y$ and $\mathrm{d}x$ are *not* being multiplied. However, Δy and Δx *were* being multiplied in our sum expression. What happened? The answer is some slight of hand. The full thought process should look like

$$\lim_{\Delta x,\Delta y\to 0} \sum_{x_i} \sum_{y_i} \Delta y \Delta x = \lim_{\Delta x,\Delta y\to 0} \sum_{x_i} \left(\sum_{y_i} \Delta y \right) \Delta x = \int_{x=-1}^{x=1} \left(\int_{y=x^2}^{y=1} \mathrm{d}y \right) \mathrm{d}x.$$

Now we can evaluate this iterated integral to conclude

$$\text{area of } \mathcal{R} = \int_{\mathcal{R}} \mathrm{d}A = \int_{x=-1}^{x=1} \left(\int_{y=x^2}^{y=1} \mathrm{d}y \right) \mathrm{d}x = \tfrac{4}{3}.$$

But, there was another way we could have divided up our original sum. We could have summed along rows first and then summed up each row. Working from this approach, we see

$$\lim_{\Delta x,\Delta y\to 0} \sum_{y_i} \sum_{x_i} \Delta x \Delta y = \lim_{\Delta x,\Delta y\to 0} \sum_{y_i} \left(\sum_{x_i} \Delta x \right) \Delta y = \int_{y=0}^{y=1} \left(\int_{x=-\sqrt{y}}^{x=\sqrt{y}} \mathrm{d}x \right) \mathrm{d}y.$$

Computing this integral, we again get $4/3$, as expected.

Integral Notation

5.2 The Volume Form

5.3 Surface Integrals

Chapter 6

Vector Fields

Appendix A

Proofs

Below are some guidelines to help you write proofs. The following rules apply whenever you write a proof[1].

1. **The burden of communication lies on you, not on your reader.** It is your job to explain your thoughts; it is not your reader's job to guess them from a few hints. You are trying to convince a skeptical reader who doesn't believe you, so you need to argue with airtight logic in crystal clear language; otherwise the reader will continue to doubt. If you didn't write something on the paper, then (a) you didn't communicate it, (b) the reader didn't learn it, and (c) the grader has to assume you didn't know it in the first place.

2. **Tell the reader what you're proving.** The reader doesn't necessarily know or remember what "Theorem 2.13" is. Even a professor grading a stack of papers might lose track from time to time. Therefore, the statement you are proving should be on the same page as the beginning of your proof. For an exam this won't be a problem, of course, but on your homework, recopy the claim you are proving. This has the additional advantage that when you study for exams by reviewing your homework, you won't have to flip back in the notes/textbook to know what you were proving.

3. **Use English words.** Although there will usually be equations or mathematical statements in your proofs, use English sentences to connect them and display their logical relationships. If you look in your notes/textbook, you'll see that each proof consists mostly of English words.

4. **Use complete sentences.** If you wrote a history essay in sentence fragments, the reader would not understand what you meant; likewise, in mathematics you must use complete sentences with verbs to convey your logical train of thought.

   Some complete sentences can be written purely in mathematical symbols, such as equations (e.g., $a^3 = b^{-1}$), inequalities (e.g., $x < 5$), and other relations (like $5 \mid 10$ or $7 \in \mathbb{Z}$). These statements usually express a relationship between two mathematical *objects*, like

[1] This list is an adaptation of *The Elements of Style for Proofs* written by Anders Hendrickson of St. Norbert College and modified by Dana Ernst of Northern Arizona University.

numbers or sets. However, it is considered bad style to begin a sentence with symbols. A common phrase to use to avoid starting a sentence with mathematical symbols is "We see that..."

5. **Show the logical connections among your sentences.** Use phrases like "Therefore" or "because" or "if..., then..." or "if and only if" to connect your sentences.

6. **Know the difference between statements and objects.** A mathematical object is a *thing*, a noun, such as a group, an element, a vector space, a number, an ordered pair, etc. Objects either exist or don't exist. Statements, on the other hand, are mathematical *sentences*: they can be true or false.

   When you see or write a cluster of math symbols, be sure you know whether it's an object (e.g., "x^2+3") or a statement (e.g., "$x^2+3<7$"). One way to tell is that every mathematical statement includes a verb, such as $=$, $\leq$, "divides", etc.

7. **"$=$" means equals.** Don't write $A=B$ unless you mean that A actually equals B. This rule seems obvious, but there is a great temptation to be sloppy. In calculus, for example, some people might write $f(x)=x^2=2x$ (which is false), when they really mean that "if $f(x)=x^2$, then $f'(x)=2x$."

8. **Don't interchange $=$ and $\implies$.** The equals sign connects two *objects*, as in "$x^2=b$"; the symbol "$\implies$" is an abbreviation for "implies" and connects two *statements*, as in "$ab=a \implies b=1$." You should avoid using $\implies$ in your formal write-ups.

9. **Say exactly what you mean.** Just as the $=$ is sometimes abused, so too people sometimes write $A \in B$ when they mean $A \subseteq B$, or write $a_{ij} \in A$ when they mean that a_{ij} is an entry in matrix A. Mathematics is a very precise language, and there is a way to say exactly what you mean; find it and use it.

10. **Don't write anything unproven.** Every statement on your paper should be something you *know* to be true. The reader expects your proof to be a series of statements, each proven by the statements that came before it. If you ever need to write something you don't yet know is true, you *must* preface it with words like "assume," "suppose," or "if" (if you are temporarily assuming it), or with words like "we need to show that" or "we claim that" (if it is your goal). Otherwise the reader will think they have missed part of your proof.

11. **Write strings of equalities (or inequalities) in the proper order.** When your reader sees something like

$$A=B\leq C=D,$$

    he/she expects to understand easily why $A=B$, why $B\leq C$, and why $C=D$, and he/she expects the *point* of the entire line to be the more complicated fact that $A\leq D$. For example, if you were computing the distance d of the point $(12,5)$ from the origin, you could write

$$d=\sqrt{12^2+5^2}=13.$$

In this string of equalities, the first equals sign is true by the Pythagorean theorem, the second is just arithmetic, and the *point* is that the first item equals the last item: $d = 13$.

A common error is to write strings of equations in the wrong order. For example, if you were to write "$\sqrt{12^2+5^2} = 13 = d$", your reader would understand the first equals sign, would be baffled as to how we know $d = 13$, and would be utterly perplexed as to why you wanted or needed to go through 13 to prove that $\sqrt{12^2+5^2} = d$.

12. **Avoid circularity.** Be sure that no step in your proof makes use of the conclusion!

13. **Don't write the proof backwards.** Beginning students often attempt to write "proofs" like the following, which attempts to prove that $\tan^2(x) = \sec^2(x) - 1$:

$$\begin{aligned}
\tan^2(x) &= \sec^2(x) - 1 \\
\left(\frac{\sin(x)}{\cos(x)}\right)^2 &= \frac{1}{\cos^2(x)} - 1 \\
\frac{\sin^2(x)}{\cos^2(x)} &= \frac{1-\cos^2(x)}{\cos^2(x)} \\
\sin^2(x) &= 1 - \cos^2(x) \\
\sin^2(x) + \cos^2(x) &= 1 \\
1 &= 1
\end{aligned}$$

Notice what has happened here: the student *started* with the conclusion, and deduced the true statement "$1 = 1$." In other words, he/she has proved "If $\tan^2(x) = \sec^2(x) - 1$, then $1 = 1$," which is true but highly uninteresting.

Now this isn't a bad way of *finding* a proof. Working backwards from your goal often is a good strategy *on your scratch paper*, but when it's time to *write* your proof, you have to start with the hypotheses and work to the conclusion.

14. **Be concise.** Most students err by writing their proofs too short, so that the reader can't understand their logic. It is nevertheless quite possible to be too wordy, and if you find yourself writing a full-page essay, it's probably because you don't have a proof, but just an intuition. When you find a way to turn that intuition into a formal proof, it will be much shorter.

15. **Introduce every symbol you use.** If you use the letter "k," the reader should know exactly what k is. Good phrases for introducing symbols include "Let $n \in \mathbb{N}$," "Let k be the least integer such that...," "For every real number a...," and "Suppose that X is a counterexample."

16. **Use appropriate quantifiers (once).** When you introduce a variable $x \in S$, it must be clear to your reader whether you mean "for all $x \in S$" or just "for some $x \in S$." If you just say something like "$y = x^2$ where $x \in S$," the word "where" doesn't indicate whether you mean "for all" or "some."

Phrases indicating the quantifier "for all" include "Let $x \in S$"; "for all $x \in S$"; "for every $x \in S$"; "for each $x \in S$"; etc. Phrases indicating the quantifier "some" (or "there exists") include "for some $x \in S$"; "there exists an $x \in S$"; "for a suitable choice of $x \in S$"; etc.

On the other hand, don't introduce a variable more than once! Once you have said "Let $x \in S$," the letter x has its meaning defined. You don't *need* to say "for all $x \in S$" again, and you definitely should *not* say "let $x \in S$" again.

17. **Use a symbol to mean only one thing.** Once you use the letter x once, its meaning is fixed for the duration of your proof. You cannot use x to mean anything else.

18. **Don't "prove by example."** Most problems ask you to prove that something is true "for all"—You *cannot* prove this by giving a single example, or even a hundred. Your answer will need to be a logical argument that holds for *every example there possibly could be*.

19. **Write "Let $x = \ldots$," not "Let $\cdots = x$."** When you have an existing expression, say a^2, and you want to give it a new, simpler name like b, you should write "Let $b = a^2$," which means, "Let the new symbol b mean a^2." This convention makes it clear to the reader that b is the brand-new symbol and a^2 is the old expression he/she already understands.

    If you were to write it backwards, saying "Let $a^2 = b$," then your startled reader would ask, "What if $a^2 \neq b$?"

20. **Make your counterexamples concrete and specific.** Proofs need to be entirely general, but counterexamples should be concrete. When you provide an example or counterexample, make it as specific as possible. For a set, for example, you must name its elements, and for a function, you must give its rule. Do not say things like "θ could be one-to-one but not onto"; instead, provide an actual function θ that *is* one-to-one but not onto.

21. **Don't include examples in proofs.** Including an example very rarely adds anything to your proof. If your logic is sound, then it doesn't need an example to back it up. If your logic is bad, a dozen examples won't help it (see rule 18). There are only two valid reasons to include an example in a proof: if it is a *counterexample* disproving something, or if you are performing complicated manipulations in a general setting and the example is just to help the reader understand what you are saying.

22. **Use scratch paper.** Finding your proof will be a long, potentially messy process, full of false starts and dead ends. Do all that on scratch paper until you find a real proof, and only then break out your clean paper to write your final proof carefully. *Do not hand in your scratch work!*

    Only sentences that actually contribute to your proof should be part of the proof. Do not just perform a "brain dump," throwing everything you know onto the paper before showing the logical steps that prove the conclusion. *That is what scratch paper is for.*

Appendix B

Creative Commons License Fulltext

Creative Commons Legal Code

Attribution-ShareAlike 4.0 International

Official translations of this license are available in other languages.

Creative Commons Corporation ("Creative Commons") is not a law firm and does not provide legal services or legal advice. Distribution of Creative Commons public licenses does not create a lawyer-client or other relationship. Creative Commons makes its licenses and related information available on an as-is basis. Creative Commons gives no warranties regarding its licenses, any material licensed under their terms and conditions, or any related information. Creative Commons disclaims all liability for damages resulting from their use to the fullest extent possible.

Using Creative Commons Public Licenses

Creative Commons public licenses provide a standard set of terms and conditions that creators and other rights holders may use to share original works of authorship and other material subject to copyright and certain other rights specified in the public license below. The following considerations are for informational purposes only, are not exhaustive, and do not form part of our licenses.

> **Considerations for licensors:** Our public licenses are intended for use by those authorized to give the public permission to use material in ways otherwise restricted by copyright and certain other rights. Our licenses are irrevocable. Licensors should read and understand the terms and conditions of the license they choose before applying it. Licensors should also secure all rights necessary before applying our licenses so that the public can reuse the material as expected. Licensors should clearly mark any material not subject to the license. This includes other CC-licensed material, or material used under an exception or limitation to copyright. More considerations for licensors.
>
> **Considerations for the public:** By using one of our public licenses, a licensor grants the public permission to use the licensed material under specified terms and conditions. If the licensors permission is not necessary for any reason—for example, because of any applicable exception or limitation to copyrightthen that use is

not regulated by the license. Our licenses grant only permissions under copyright and certain other rights that a licensor has authority to grant. Use of the licensed material may still be restricted for other reasons, including because others have copyright or other rights in the material. A licensor may make special requests, such as asking that all changes be marked or described. Although not required by our licenses, you are encouraged to respect those requests where reasonable. More considerations for the public.

Creative Commons Attribution-ShareAlike 4.0 International Public License

By exercising the Licensed Rights (defined below), You accept and agree to be bound by the terms and conditions of this Creative Commons Attribution-ShareAlike 4.0 International Public License ("Public License"). To the extent this Public License may be interpreted as a contract, You are granted the Licensed Rights in consideration of Your acceptance of these terms and conditions, and the Licensor grants You such rights in consideration of benefits the Licensor receives from making the Licensed Material available under these terms and conditions.

Section 1 Definitions.

a. **Adapted Material** means material subject to Copyright and Similar Rights that is derived from or based upon the Licensed Material and in which the Licensed Material is translated, altered, arranged, transformed, or otherwise modified in a manner requiring permission under the Copyright and Similar Rights held by the Licensor. For purposes of this Public License, where the Licensed Material is a musical work, performance, or sound recording, Adapted Material is always produced where the Licensed Material is synched in timed relation with a moving image.

b. **Adapter's License** means the license You apply to Your Copyright and Similar Rights in Your contributions to Adapted Material in accordance with the terms and conditions of this Public License.

c. **BY-SA Compatible License** means a license listed at http://creativecommons.org/compatiblelicenses, approved by Creative Commons as essentially the equivalent of this Public License.

d. **Copyright and Similar Rights** means copyright and/or similar rights closely related to copyright including, without limitation, performance, broadcast, sound recording, and Sui Generis Database Rights, without regard to how the rights are labeled or categorized. For purposes of this Public License, the rights specified in Section 2(b)(1)-(2) are not Copyright and Similar Rights.

e. **Effective Technological Measures** means those measures that, in the absence of proper authority, may not be circumvented under laws fulfilling obligations under Article 11 of the WIPO Copyright Treaty adopted on December 20, 1996, and/or similar international agreements.

f. **Exceptions and Limitations** means fair use, fair dealing, and/or any other exception or limitation to Copyright and Similar Rights that applies to Your use of the Licensed Material.

g. **License Elements** means the license attributes listed in the name of a Creative Commons Public License. The License Elements of this Public License are Attribution and ShareAlike.

h. **Licensed Material** means the artistic or literary work, database, or other material to which the Licensor applied this Public License.

i. **Licensed Rights** means the rights granted to You subject to the terms and conditions of this Public License, which are limited to all Copyright and Similar Rights that apply to Your use of the Licensed Material and that the Licensor has authority to license.

j. **Licensor** means the individual(s) or entity(ies) granting rights under this Public License.

k. **Share** means to provide material to the public by any means or process that requires permission under the Licensed Rights, such as reproduction, public display, public performance, distribution, dissemination, communication, or importation, and to make material available to the public including in ways that members of the public may access the material from a place and at a time individually chosen by them.

l. **Sui Generis Database Rights** means rights other than copyright resulting from Directive 96/9/EC of the European Parliament and of the Council of 11 March 1996 on the legal protection of databases, as amended and/or succeeded, as well as other essentially equivalent rights anywhere in the world.

m. **You** means the individual or entity exercising the Licensed Rights under this Public License. **Your** has a corresponding meaning.

Section 2 Scope.

a. **License grant**.

1. Subject to the terms and conditions of this Public License, the Licensor hereby grants You a worldwide, royalty-free, non-sublicensable, non-exclusive, irrevocable license to exercise the Licensed Rights in the Licensed Material to:
   A. reproduce and Share the Licensed Material, in whole or in part; and
   B. produce, reproduce, and Share Adapted Material.
2. Exceptions and Limitations. For the avoidance of doubt, where Exceptions and Limitations apply to Your use, this Public License does not apply, and You do not need to comply with its terms and conditions.
3. Term. The term of this Public License is specified in Section 6(a).
4. Media and formats; technical modifications allowed. The Licensor authorizes You to exercise the Licensed Rights in all media and formats whether now known or hereafter created, and to make technical modifications necessary to do so. The Licensor waives and/or agrees not to assert any right or authority to forbid You from making technical modifications necessary to exercise the Licensed Rights, including technical modifications necessary to circumvent Effective Technological Measures. For purposes of this Public License, simply making modifications authorized by this Section 2(a)(4) never produces Adapted Material.

5. Downstream recipients.
   A. Offer from the Licensor Licensed Material. Every recipient of the Licensed Material automatically receives an offer from the Licensor to exercise the Licensed Rights under the terms and conditions of this Public License.
   B. Additional offer from the Licensor Adapted Material. Every recipient of Adapted Material from You automatically receives an offer from the Licensor to exercise the Licensed Rights in the Adapted Material under the conditions of the Adapters License You apply.
   C. No downstream restrictions. You may not offer or impose any additional or different terms or conditions on, or apply any Effective Technological Measures to, the Licensed Material if doing so restricts exercise of the Licensed Rights by any recipient of the Licensed Material.
6. No endorsement. Nothing in this Public License constitutes or may be construed as permission to assert or imply that You are, or that Your use of the Licensed Material is, connected with, or sponsored, endorsed, or granted official status by, the Licensor or others designated to receive attribution as provided in Section 3(a)(1)(A)(i).

Other rights.

b. 1. Moral rights, such as the right of integrity, are not licensed under this Public License, nor are publicity, privacy, and/or other similar personality rights; however, to the extent possible, the Licensor waives and/or agrees not to assert any such rights held by the Licensor to the limited extent necessary to allow You to exercise the Licensed Rights, but not otherwise.
2. Patent and trademark rights are not licensed under this Public License.
3. To the extent possible, the Licensor waives any right to collect royalties from You for the exercise of the Licensed Rights, whether directly or through a collecting society under any voluntary or waivable statutory or compulsory licensing scheme. In all other cases the Licensor expressly reserves any right to collect such royalties.

Section 3 License Conditions.
Your exercise of the Licensed Rights is expressly made subject to the following conditions.

a. **Attribution**.

1. If You Share the Licensed Material (including in modified form), You must:
   A. retain the following if it is supplied by the Licensor with the Licensed Material:
      i. identification of the creator(s) of the Licensed Material and any others designated to receive attribution, in any reasonable manner requested by the Licensor (including by pseudonym if designated);
      ii. a copyright notice;
      iii. a notice that refers to this Public License;
      iv. a notice that refers to the disclaimer of warranties;

v. a URI or hyperlink to the Licensed Material to the extent reasonably practicable;

B. indicate if You modified the Licensed Material and retain an indication of any previous modifications; and

C. indicate the Licensed Material is licensed under this Public License, and include the text of, or the URI or hyperlink to, this Public License.

2. You may satisfy the conditions in Section 3(a)(1) in any reasonable manner based on the medium, means, and context in which You Share the Licensed Material. For example, it may be reasonable to satisfy the conditions by providing a URI or hyperlink to a resource that includes the required information.

3. If requested by the Licensor, You must remove any of the information required by Section 3(a)(1)(A) to the extent reasonably practicable.

b. **ShareAlike**.

In addition to the conditions in Section 3(a), if You Share Adapted Material You produce, the following conditions also apply.

1. The Adapters License You apply must be a Creative Commons license with the same License Elements, this version or later, or a BY-SA Compatible License.

2. You must include the text of, or the URI or hyperlink to, the Adapter's License You apply. You may satisfy this condition in any reasonable manner based on the medium, means, and context in which You Share Adapted Material.

3. You may not offer or impose any additional or different terms or conditions on, or apply any Effective Technological Measures to, Adapted Material that restrict exercise of the rights granted under the Adapter's License You apply.

Section 4 Sui Generis Database Rights.

Where the Licensed Rights include Sui Generis Database Rights that apply to Your use of the Licensed Material:

a. for the avoidance of doubt, Section 2(a)(1) grants You the right to extract, reuse, reproduce, and Share all or a substantial portion of the contents of the database;

b. if You include all or a substantial portion of the database contents in a database in which You have Sui Generis Database Rights, then the database in which You have Sui Generis Database Rights (but not its individual contents) is Adapted Material, including for purposes of Section 3(b); and

c. You must comply with the conditions in Section 3(a) if You Share all or a substantial portion of the contents of the database.

For the avoidance of doubt, this Section 4 supplements and does not replace Your obligations under this Public License where the Licensed Rights include other Copyright and Similar Rights.

Section 5 Disclaimer of Warranties and Limitation of Liability.

a. **Unless otherwise separately undertaken by the Licensor, to the extent possible, the Licensor offers the Licensed Material as-is and as-available, and makes no representations or warranties of any kind concerning the Licensed Material, whether express, implied, statutory, or other. This includes, without limitation, warranties of title, merchantability, fitness for a particular purpose, non-infringement, absence of latent or other defects, accuracy, or the presence or absence of errors, whether or not known or discoverable. Where disclaimers of warranties are not allowed in full or in part, this disclaimer may not apply to You.**

b. **To the extent possible, in no event will the Licensor be liable to You on any legal theory (including, without limitation, negligence) or otherwise for any direct, special, indirect, incidental, consequential, punitive, exemplary, or other losses, costs, expenses, or damages arising out of this Public License or use of the Licensed Material, even if the Licensor has been advised of the possibility of such losses, costs, expenses, or damages. Where a limitation of liability is not allowed in full or in part, this limitation may not apply to You.**

a. The disclaimer of warranties and limitation of liability provided above shall be interpreted in a manner that, to the extent possible, most closely approximates an absolute disclaimer and waiver of all liability.

Section 6 Term and Termination.

a. This Public License applies for the term of the Copyright and Similar Rights licensed here. However, if You fail to comply with this Public License, then Your rights under this Public License terminate automatically.

b. Where Your right to use the Licensed Material has terminated under Section 6(a), it reinstates:

   1. automatically as of the date the violation is cured, provided it is cured within 30 days of Your discovery of the violation; or
   2. upon express reinstatement by the Licensor.

   For the avoidance of doubt, this Section 6(b) does not affect any right the Licensor may have to seek remedies for Your violations of this Public License.

c. For the avoidance of doubt, the Licensor may also offer the Licensed Material under separate terms or conditions or stop distributing the Licensed Material at any time; however, doing so will not terminate this Public License.

d. Sections 1, 5, 6, 7, and 8 survive termination of this Public License.

Section 7 Other Terms and Conditions.

a. The Licensor shall not be bound by any additional or different terms or conditions communicated by You unless expressly agreed.

b. Any arrangements, understandings, or agreements regarding the Licensed Material not stated herein are separate from and independent of the terms and conditions of this Public License.

Section 8 Interpretation.

a. For the avoidance of doubt, this Public License does not, and shall not be interpreted to, reduce, limit, restrict, or impose conditions on any use of the Licensed Material that could lawfully be made without permission under this Public License.

b. To the extent possible, if any provision of this Public License is deemed unenforceable, it shall be automatically reformed to the minimum extent necessary to make it enforceable. If the provision cannot be reformed, it shall be severed from this Public License without affecting the enforceability of the remaining terms and conditions.

c. No term or condition of this Public License will be waived and no failure to comply consented to unless expressly agreed to by the Licensor.

d. Nothing in this Public License constitutes or may be interpreted as a limitation upon, or waiver of, any privileges and immunities that apply to the Licensor or You, including from the legal processes of any jurisdiction or authority.

Creative Commons is not a party to its public licenses. Notwithstanding, Creative Commons may elect to apply one of its public licenses to material it publishes and in those instances will be considered the Licensor. The text of the Creative Commons public licenses is dedicated to the public domain under the CC0 Public Domain Dedication. Except for the limited purpose of indicating that material is shared under a Creative Commons public license or as otherwise permitted by the Creative Commons policies published at creativecommons.org/policies, Creative Commons does not authorize the use of the trademark Creative Commons or any other trademark or logo of Creative Commons without its prior written consent including, without limitation, in connection with any unauthorized modifications to any of its public licenses or any other arrangements, understandings, or agreements concerning use of licensed material. For the avoidance of doubt, this paragraph does not form part of the public licenses.

Creative Commons may be contacted at http://creativecommons.org.