32A: MULTIVARIABLE CALCULUS, FIRST QUARTER, FALL 2015

STEVEN HEILMAN

Contents

1. Introduction

As you have covered in your previous calculus classes, the subject of calculus has many applications. Here are some applications that are specific to multivariable calculus. As we have seen in the single variable setting, **optimization** can be quite useful. Often we want to optimize some function of more than one variable, and doing so requires tools from multivariable calculus. Such optimization procedures appear in **economics** and **engineering**. **Fluid mechanics** uses multivariable calculus to describe the flow of fluids through different geometries and along different surfaces. Within physics, the theory of **general relativity**

Date: November 9, 2015.

takes further many of the ideas that we will describe in this course, though this physical theory uses much more geometry than we will encounter. Within **medicine**, the mathematics of MRI and CT scans was discovered roughly forty years before these devices were invented. The relevant mathematics here uses both multivariable calculus and Fourier analysis.

1.1. **Plan for the Course.** We will first introduce vectors, and then discuss surfaces and build up our intuition for dealing with curves and surfaces in two and three dimensions. We then discuss functions of several variables and their derivatives. The course ends with optimization of functions of multiple variables. The ensuing courses discusses integration of functions of several variables.

2. Vectors

When we change from calculus in one dimension to calculus in two and three dimensions, we naturally need to deal with the extra dimensions. The most convenient way to deal with these extra dimensions is to introduce vectors. We will see that the operations of addition and subtraction on real numbers will extend to vectors. However, we cannot multiply two vectors like they are real numbers. And there are new things to consider for vectors that have no analogue for real numbers. For example, two vectors have an angle between them.

We begin by defining vectors in the plane.

2.1. **Vectors in the Plane.** We denote the plane as the set of all ordered pairs (x, y) where x and y are both real numbers. We use the notation $\mathbb{R}^2$ to denote the plane.

Definition 2.1. A **two-dimensional vector** v is a directed line segment. This line segment has a beginning point P and a terminal point Q, where $P, Q \in \mathbb{R}^2$. Suppose $P = (x_1, y_1)$ and if $Q = (x_2, y_2)$, where x_1, x_2, y_1, y_2 are real numbers. We define the **length** or **magnitude** of the vector v by

$$\|v\| = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}.$$

That is, by the Pythagorean Theorem, $\|v\|$ is the length of the hypotenuse of the right triangle, whose side lengths are $|x_1 - x_2|$ and $|y_1 - y_2|$.

Remark 2.2. Unless otherwise stated, all vectors from now on will have their beginning point P at the origin, so that $P = (0, 0)$. We then will write any two-dimensional vector v in the form $v = (x, y)$. Then

$$\|v\| = \sqrt{x^2 + y^2}.$$

Let λ be a real number. (We often refer to real numbers as **scalars**.) Define $\lambda v = (\lambda x, \lambda y)$. Note that

$$\|\lambda v\| = \sqrt{\lambda^2 x^2 + \lambda^2 y^2} = \sqrt{\lambda^2 (x^2 + y^2)} = |\lambda| \, \|v\| \, .$$

Definition 2.3. Two vectors $v, w \in \mathbb{R}^2$ are said to be **parallel** if there is some scalar $\lambda \in \mathbb{R}$ such that $v = \lambda w$.

We can add vectors together as follows.

Definition 2.4. Suppose $v = (x_1, y_1)$ and $w = (x_2, y_2)$. Define

$$v + w = (x_1 + x_2, y_1 + y_2).$$

Visually, the vector $v+w$ is obtained by taking the vector v, placing the beginning point of w at the endpoint of v, so that the resulting endpoint is that of $v+w$. We also define

$$v - w = (x_1 - x_2, y_1 - y_2).$$

Note that

$$v + (0,0) = (0,0) + v = v, \qquad v - v = (0,0).$$

Example 2.5.

$$(1,2) + 2(3,0) = (1,2) + (6,0) = (7,2).$$

Proposition 2.6. *For any vectors $u,v,w \in \mathbb{R}^2$ and for any scalar $\lambda \in \mathbb{R}$,*

$$v + w = w + v.$$

$$u + (v + w) = (u + v) + w.$$

$$\lambda(u + v) = (\lambda u) + (\lambda v).$$

Definition 2.7. A vector $v \in \mathbb{R}^2$ such that $\|v\| = 1$ is called a **unit vector**. If v is any nonzero vector, then the vector

$$e_v = \frac{v}{\|v\|}$$

is a unit vector which is parallel to v. To see that e_v is a unit vector, we use Remark 2.2 to get $\|e_v\| = \left\|\frac{v}{\|v\|}\right\| = \frac{\|v\|}{\|v\|} = 1$.

Example 2.8. Consider the vector $v = (1,2)$. Then $e_v = \frac{v}{\|v\|} = \frac{(1,2)}{\sqrt{5}} = (1/\sqrt{5}, 2/\sqrt{5})$ is a unit vector, which is parallel to v.

Remark 2.9. Some textbooks use the notation $\vec{i} = (1,0)$ and $\vec{j} = (0,1)$, so that a vector $v = (x,y) \in \mathbb{R}^2$ can be written as $v = x\vec{i} + y\vec{j}$. However, we will not use this notation.

Theorem 2.10 (**Triangle Inequality**). *For any vectors $v,w \in \mathbb{R}^2$, we have*

$$\|v + w\| \leq \|v\| + \|w\|.$$

2.2. **Vectors in Three Dimensions.** We denote three-dimensional space as the set of all ordered pairs (x,y,z) where x,y and z are all real numbers. We use the notation $\mathbb{R}^3$ to denote three-dimensional space.

When drawing vectors in the plane, we always use the convention that the x-axis is the horizontal axis, and the y-axis is the vertical axis. We will use a similar convention when plotting vectors in three-dimensional space $\mathbb{R}^3$. However, we now need to orient the axes using the right hand rule. Using your right hand, if the fingers curl from the x-axis to the y-axis, then the z-axis will point in the direction of your extended thumb.

In addition to coordinate axes, we also now have coordinate planes. The **xy-plane** is the set of all $(x,y,z) \in \mathbb{R}^3$ such that $z = 0$. The **xz-plane** is the set of all $(x,y,z) \in \mathbb{R}^3$ such that $y = 0$. The **yz-plane** is the set of all $(x,y,z) \in \mathbb{R}^3$ such that $x = 0$. These are the three coordinate planes. These planes divide $\mathbb{R}^3$ into eight regions, which we call **octants**. The octant where $x \geq 0, y \geq 0$ and $z \geq 0$ is called the **first octant**.

We can now derive a distance formula between points, using the Pythagorean Theorem, just as we did in the plane $\mathbb{R}^2$.

Theorem 2.11. *The distance from the point* $(x, y, z) \in \mathbb{R}^3$ *to the origin* $(0, 0, 0)$ *is*

$$\sqrt{x^2 + y^2 + z^2}.$$

Proof. Consider the point $(x, y, 0)$ in the xy-plane. From our distance formula in the plane, the distance from this point to the origin is $\sqrt{x^2 + y^2}$. Consider the right triangle with vertices $(0, 0, 0)$, $(x, y, 0)$ and (x, y, z). By the Pythagorean Theorem, the length of the hypotenuse (which is also the distance from (x, y, z) to the origin) is

$$\sqrt{(\sqrt{x^2 + y^2})^2 + z^2} = \sqrt{x^2 + y^2 + z^2}.$$

□

Remark 2.12. Consequently, the distance between any two points $(x_1, y_1, z_1) \in \mathbb{R}^3$ and $(x_2, y_2, z_2) \in \mathbb{R}^3$ is

$$\sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2 + (z_1 - z_2)^2}.$$

We now give a few examples of surfaces in $\mathbb{R}^3$.

Definition 2.13. A **sphere** of radius $r > 0$ and center $(a, b, c) \in \mathbb{R}^3$ is defined as the set of points of distance exactly r from the point (a, b, c). So, the sphere is the set of all $(x, y, z) \in \mathbb{R}^3$ such that

$$\sqrt{(x - a)^2 + (y - b)^2 + (z - c)^2} = r.$$

Equivalently, this sphere is described as the set of all $(x, y, z) \in \mathbb{R}^3$ such that

$$(x - a)^2 + (y - b)^2 + (z - c)^2 = r^2.$$

Example 2.14. The set of all (x, y, z) such that $x^2 + y^2 + z^2 \le r^2$ is the set of all points of distance at most r from the origin. This set is known as the **solid ball** of radius r.

Example 2.15. The set of all (x, y, z) such that $x^2 + y^2 + z^2 = r^2$ and $z \ge 0$ is the **upper hemisphere** of the sphere of radius r.

Remark 2.16. Note that when $z = 0$ and $c = 0$, the equation $(x-a)^2+(y-b)^2+(z-c)^2 = r^2$ reduces to the equation for a circle of radius R and center (a, b):

$$(x - a)^2 + (y - b)^2 = R^2.$$

That is, when $c = 0$, the sphere of radius R is cut by the plane $z = 0$, producing a circle.

Definition 2.17. A **cylinder** of radius $r > 0$ with displacement $(a, b, 0) \in \mathbb{R}^3$ is defined as the set of points $(x, y, z) \in \mathbb{R}^3$ such that

$$(x - a)^2 + (y - b)^2 = r^2.$$

Note that the variable z is unrestricted.

We can readily extend addition, length, etc. to the case of three-dimensional vectors.

Definition 2.18. A **three-dimensional vector** $v = (x, y, z)$ is a directed line segment from the origin $(0, 0, 0)$ to the point (x, y, z). We define the **length** or **magnitude** of the vector v by

$$\|v\| = \sqrt{x^2 + y^2 + z^2}.$$

Let $\lambda \in \mathbb{R}$. Define $\lambda v = (\lambda x, \lambda y, \lambda z)$. Note that

$$\|\lambda v\| = |\lambda| \, \|v\| \, .$$

Two vectors $v, w \in \mathbb{R}^3$ are said to be **parallel** if there is some scalar $\lambda \in \mathbb{R}$ such that $v = \lambda w$.

We can add vectors together as follows.

Definition 2.19. Suppose $v = (x_1, y_1, z_1)$ and $w = (x_2, y_2, z_2)$. Define

$$v + w = (x_1 + x_2, y_1 + y_2, z_1 + z_2).$$

$$v - w = (x_1 - x_2, y_1 - y_2, z_1 - z_2).$$

Note that

$$v + (0,0,0) = (0,0,0) + v = v, \qquad v - v = (0,0,0).$$

For any vectors $u, v, w \in \mathbb{R}^3$ and for any scalar $\lambda \in \mathbb{R}$,

$$v + w = w + v.$$

$$u + (v + w) = (u + v) + w.$$

$$\lambda(u + v) = (\lambda u) + (\lambda v).$$

Definition 2.20. A vector $v \in \mathbb{R}^3$ such that $\|v\| = 1$ is called a **unit vector**. If v is any nonzero vector, then the vector

$$e_v = \frac{v}{\|v\|}$$

is a unit vector which is parallel to v. To see that e_v is a unit vector, note that $\|e_v\| = \left\| \frac{v}{\|v\|} \right\| = \frac{\|v\|}{\|v\|} = 1$.

Example 2.21. Consider the vector $v = (1, 2)$. Then $e_v = \frac{v}{\|v\|} = \frac{(1,2)}{\sqrt{5}} = (1/\sqrt{5}, 2/\sqrt{5})$ is a unit vector, which is parallel to v.

Remark 2.22. Some textbooks use the notation $\vec{i} = (1,0,0)$, $\vec{j} = (0,1,0)$ and $\vec{k} = (0,0,1)$ so that a vector $v = (x, y, z) \in \mathbb{R}^3$ can be written as $v = x\vec{i} + y\vec{j} + z\vec{k}$. However, we will not use this notation.

Theorem 2.23 (**Triangle Inequality**)**.** *For any vectors $v, w \in \mathbb{R}^3$, we have*

$$\|v + w\| \leq \|v\| + \|w\| \, .$$

2.3. **Dot Product, Cross Product.** The dot product operation allows us to combine two vectors, producing a scalar. In some sense, the dot product measures how close two vectors are. In another sense, the dot product of two vectors measures both their length and their angle simultaneously. For now we will just define the dot product, but later on we will give more geometric meaning to it.

Definition 2.24 (**Dot Product**)**.** Let $v = (x_1, y_1)$ and let $w = (x_2, y_2)$ be vectors in $\mathbb{R}^2$. We define the **dot product** of v and w by

$$v \cdot w = x_1 x_2 + y_1 y_2.$$

Definition 2.25 (**Dot Product**)**.** Let $v = (x_1, y_1, z_1)$ and let $w = (x_2, y_2, z_2)$ be vectors in $\mathbb{R}^3$. We define the **dot product** of v and w by

$$v \cdot w = x_1x_2 + y_1y_2 + z_1z_2.$$

Note that $\sqrt{v \cdot v} = \sqrt{x_1^2 + y_1^2 + z_1^2} = \|v\|$.

Example 2.26.

$$(1, 2, 3) \cdot (0, 5, 7) = 0 \cdot 1 + 2 \cdot 5 + 3 \cdot 7 = 10 + 21 = 31.$$

Proposition 2.27 (**Properties of the dot product**)**.** *Let $u, v, w \in \mathbb{R}^3$ and let $\lambda \in \mathbb{R}$. Then*

- $(0, 0, 0) \cdot v = v \cdot (0, 0, 0) = 0$.
- $v \cdot w = w \cdot v$.
- $(\lambda v) \cdot w = v \cdot (\lambda w) = \lambda(v \cdot w)$.
- $u \cdot (v + w) = (u \cdot v) + (u \cdot w)$, *and* $(v + w) \cdot u = (v \cdot u) + (w \cdot u)$.

Definition 2.28. The **angle** $0 \leq \theta \leq \pi$ between two nonzero vectors v, w is defined to be

$$\theta = \cos^{-1}\left(\frac{v}{\|v\|} \cdot \frac{w}{\|w\|}\right).$$

That is, the dot product of the unit vector in the direction v with the unit vector in the direction w is the cosine of θ. Written another way,

$$v \cdot w = \|v\| \, \|w\| \cos\theta, \qquad \cos\theta = \frac{v}{\|v\|} \cdot \frac{w}{\|w\|}.$$

In this sense, the dot product of v and w contains information about the lengths of v and w, as well as the angle between these vectors.

Remark 2.29. Recall that the range of $\cos^{-1}$ is $[0, \pi]$, so our definition of θ automatically means that $\theta \in [0, \pi]$.

Definition 2.30. Two nonzero vectors v, w are said to be **perpendicular** or **orthogonal** if the angle θ between them is $\pi/2$. In this case, we write $v \perp w$. Since $\cos(\pi/2) = 0$, and since $v \cdot w = \|v\| \, \|w\| \cos\theta = 0$, we see that the vectors v, w are perpendicular if and only if $v \cdot w = 0$. (If $v \cdot w = 0$, then $\theta = \cos^{-1}(0)$, so $\theta = \pi/2$.)

Remark 2.31. If $v \cdot w > 0$, then $\theta < \pi/2$, that is, the angle between v and w is less than $\pi/2$. If $v \cdot w < 0$, then $\theta > \pi/2$, that is, the angle between v and w is more than $\pi/2$. Both assertions follow from the definition of θ.

We will often need to project one vector onto another. Suppose we have a vector v and we want to project it onto the nonzero vector w. That is, we want to write $v = \alpha w + n$ where $\alpha \in \mathbb{R}$ and n is perpendicular to w. In this case, we say that αw is the projection of v onto w, and we denote $\alpha w = \mathrm{proj}_w(v)$. To find α, note that since $n \perp w$,

$$v \cdot w = (\alpha w + n) \cdot w = \alpha(w \cdot w) + (n \cdot w) = \alpha(w \cdot w).$$

So, solving for α, we get

$$\alpha = \frac{v \cdot w}{w \cdot w}.$$

We therefore see that

$$\operatorname{proj}_w(v) = \left(\frac{v\cdot w}{w\cdot w}\right) w = \left(\frac{v\cdot w}{\|w\|^2}\right) w = \left(v\cdot\frac{w}{\|w\|}\right)\frac{w}{\|w\|}.$$

That is, $\operatorname{proj}_w(v)$ is given by taking the unit vector $w/\|w\|$, and multiplying it by the dot product of v against the unit vector $w/\|w\|$.

Example 2.32. Write the vector $v = (1, 2)$ as a sum of its projection onto $w = (2, 1)$, plus a vector perpendicular to w.

We know that

$$\operatorname{proj}_w(v) = \frac{v\cdot w}{w\cdot w} w = \frac{4}{5}(2, 1).$$

We want to write

$$v = \operatorname{proj}_w(v) + (v - \operatorname{proj}_w(v)) = (8/5, 4/5) + (-3/5, 6/5).$$

Note that $(8/5, 4/5)\cdot(-3/5, 6/5) = 0$.

Remark 2.33. Let v, w be vectors. As above, we can write $v = \operatorname{proj}_w(v)+n$, where $n\cdot w = 0$. Consider then the right triangle with edges formed by v, $\operatorname{proj}_w(v)$ and n. The hypotenuse of this triangle has length $\|v\|$. And one of its other edges has length

$$\|\operatorname{proj}_w(v)\| = \left|\frac{v\cdot w}{w\cdot w}\right| \|w\| = \frac{\|v\|\,\|w\|\cos\theta}{\|w\|^2}\|w\| = \|v\|\cos\theta.$$

That is, θ as defined above is the same as our "usual" definition of the angle θ between two edges of a triangle.

2.3.1. *Cross Product.* The cross product allows us to multiply two vectors v, w in $\mathbb{R}^3$, producing a third vector $v\times w$ in $\mathbb{R}^3$.

Warning 1: The cross product is **not** defined for vectors in $\mathbb{R}^2$.

Warning 2: The cross product is **not** commutative. That is, in general, we will have $v\times w \neq w\times v$. In fact, $v\times w = -(w\times v)$.

Definition 2.34 (Cross Product). Let $v = (x_1, y_1, z_1)\in\mathbb{R}^3$ and let $w = (x_2, y_2, z_2)\in\mathbb{R}^3$. We define the **cross product** $v\times w\in\mathbb{R}^3$ of v and w by

$$v\times w = (y_1z_2 - z_1y_2\,,\quad z_1x_2 - x_1z_2\,,\quad x_1y_2 - y_1x_2).$$

Thankfully, there is a nice way to remember this formula. In order to remember it, we need to discuss the determinant of a matrix.

Definition 2.35 (Determinant). Let $A = \begin{pmatrix} a & b\\ c & d\end{pmatrix}$ be a 2×2 matrix. We define

$$\det(A) = |A| = ad - bc.$$

Let $A = \begin{pmatrix} a & b & c\\ d & e & f\\ g & h & i\end{pmatrix}$ be a 3×3 matrix. We define

$$\begin{aligned}\det(A) = |A| &= a\cdot\det\begin{pmatrix} e & f\\ h & i\end{pmatrix} - b\cdot\det\begin{pmatrix} d & f\\ g & i\end{pmatrix} + c\cdot\det\begin{pmatrix} d & e\\ g & h\end{pmatrix}\\ &= a(ei - fh) + b(fg - di) + c(dh - eg).\end{aligned}$$

Example 2.36.

$$\det\begin{pmatrix}1 & 2\\ 3 & 4\end{pmatrix} = 1\cdot 4 - 2\cdot 3 = -2.$$

$$\begin{aligned}\det\begin{pmatrix}1 & 2 & 3\\ 0 & 0 & 4\\ -1 & 2 & 5\end{pmatrix} &= 1\cdot\det\begin{pmatrix}0 & 4\\ 2 & 5\end{pmatrix} - 2\cdot\det\begin{pmatrix}0 & 4\\ -1 & 5\end{pmatrix} + 3\cdot\det\begin{pmatrix}0 & 0\\ -1 & 2\end{pmatrix}\\ &= 1\cdot(0\cdot 5 - 4\cdot 2) - 2(0\cdot 5 - 4\cdot(-1)) + 3(0\cdot 2 - 0\cdot(-1))\\ &= -8 - 8 + 0 = -16.\end{aligned}$$

Remark 2.37. We can now write a way to remember the cross product formula. Let $v = (x_1, y_1, z_1) \in \mathbb{R}^3$ and let $w = (x_2, y_2, z_2) \in \mathbb{R}^3$. Then, informally,

$$\begin{aligned}v\times w &= \det\begin{pmatrix}(1,0,0) & (0,1,0) & (0,0,1)\\ x_1 & y_1 & z_1\\ x_2 & y_2 & z_2\end{pmatrix}\\ &= (1,0,0)\det\begin{pmatrix}y_1 & z_1\\ y_2 & z_2\end{pmatrix} - (0,1,0)\det\begin{pmatrix}x_1 & z_1\\ x_2 & z_2\end{pmatrix} + (0,0,1)\det\begin{pmatrix}x_1 & y_1\\ x_2 & y_2\end{pmatrix}\\ &= (y_1z_2 - z_1y_2,\quad z_1x_2 - x_1z_2,\quad x_1y_2 - y_1x_2).\end{aligned}$$

Proposition 2.38 (Properties of Cross Product). *Let $u, v, w \in \mathbb{R}^3$ be vectors, and let $a \in \mathbb{R}$. Then*

- $(av)\times w = v\times(aw) = a(v\times w)$.
- $u\times(v+w) = (u\times v) + (u\times w)$.
- $(v+w)\times u = (v\times u) + (w\times u)$.
- $v\times w = -(w\times v)$.
- $(0,0,0)\times u = (0,0,0)$.
- $v\times v = (0,0,0)$.
- $v\times w = (0,0,0)$ *if and only if v and w are parallel.*

Many of these properties follow readily from the definition of the cross product, Definition 2.34. We demonstrate that $v\times w = -(w\times v)$.

Proof. Let $v = (x_1, y_1, z_1) \in \mathbb{R}^3$ and let $w = (x_2, y_2, z_2) \in \mathbb{R}^3$. Then

$$\begin{aligned}v\times w &= (y_1z_2 - z_1y_2, z_1x_2 - x_1z_2, x_1y_2 - y_1x_2)\\ &= -(y_2z_1 - z_2y_1, z_2x_1 - x_2z_1, x_2y_1 - y_2x_1) = -w\times v\end{aligned}$$

□

Theorem 2.39 (Geometric Description of Cross Product). *Let $v, w \in \mathbb{R}^3$.*

(i) *$v\times w$ is orthogonal to v and to w.*
(ii) *$\|v\times w\| = \|v\|\,\|w\|\,|\sin\theta|$, where θ is the angle between v and w.*
(iii) *The ordered set of vectors $v, w, v\times w$ follows the right-hand rule.*

We will demonstrate these properties by example.

Example 2.40.

$$(1,0,0) \times (0,1,0) = \det \begin{pmatrix} (1,0,0) & (0,1,0) & (0,0,1) \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix} = (0,0,1).$$

Evidently, $(1,0,0) \cdot (0,0,1) = 0$ and $(0,1,0) \cdot (0,0,1) = 0$, so $(1,0,0) \times (0,1,0)$ is orthogonal to $(1,0,0)$ and to $(0,1,0)$.

$$\begin{aligned}(1,0,0) \times (3,1,\sqrt{2}) &= \det \begin{pmatrix} (1,0,0) & (0,1,0) & (0,0,1) \\ 1 & 0 & 0 \\ 3 & 1 & \sqrt{2} \end{pmatrix} \\ &= -\sqrt{2}(0,1,0) + (0,0,1) = (0,-\sqrt{2},1).\end{aligned}$$

Note that $\theta = \cos^{-1}[(1,0,0) \cdot (1/\sqrt{12})(3,1,2)] = \cos^{-1}(3/\sqrt{12}) = \cos^{-1}(\sqrt{3}/2) = \pi/6$. So, $\|(1,0,0) \times (3,1,\sqrt{2})\| = \|(0,-\sqrt{2},1)\| = \sqrt{3} = 1 \cdot \sqrt{12} \cdot \frac{1}{2} = \|(1,0,0)\| \, \|(3,1,\sqrt{2})\| \sin(\theta)$.

The curious reader can read proofs of properties (i) and (ii) below.

Proof of (i). Let $v = (x_1, y_1, z_1) \in \mathbb{R}^3$ and let $w = (x_2, y_2, z_2) \in \mathbb{R}^3$. Then

$$\begin{aligned} v \cdot (v \times w) &= (x_1, y_1, z_1) \cdot (y_1 z_2 - z_1 y_2, z_1 x_2 - x_1 z_2, x_1 y_2 - y_1 x_2) \\ &= x_1 y_1 z_2 - z_1 y_2 x_1 + y_1 z_1 x_2 - y_1 x_1 z_2 + z_1 x_1 y_2 - z_1 y_1 x_2 = 0. \end{aligned}$$

So, $v \times w$ is perpendicular to v. Similarly, $v \times w$ is perpendicular to w. □

Proof of (ii). We want to show $\|v \times w\| = \|v\| \, \|w\| \, |\sin \theta| = \|v\| \, \|w\| \sqrt{1 - \cos^2(\theta)}$. So, let's verify $\|v \times w\|^2 = \|v\|^2 \|w\|^2 (1 - \cos^2(\theta)) = \|v\|^2 \|w\|^2 - (v \cdot w)^2$.

$$\begin{aligned} \|v \times w\|^2 &= (y_1 z_2 - z_1 y_2)^2 + (z_1 x_2 - x_1 z_2)^2 + (x_1 y_2 - y_1 x_2)^2 \\ &= y_1^2 z_2^2 + z_1^2 y_2^2 - 2 y_1 y_2 z_1 z_2 + z_1^2 x_2^2 + x_1^2 z_2^2 - 2 x_1 x_2 z_1 z_2 + x_1^2 y_2^2 + y_1^2 x_2^2 - 2 x_1 x_2 y_1 y_2 \\ &= (x_1^2 + y_1^2 + z_1^2)(x_2^2 + y_2^2 + z_2^2) - (x_1 x_2 + y_1 y_2 + z_1 z_2)^2 \\ &= \|v\|^2 \|w\|^2 - (v \cdot w)^2. \end{aligned}$$

□

The cross product can also be related to volumes in the following way.

Theorem 2.41 (**Cross Product, Area, and Volume**). *Let u, v, w be vectors in $\mathbb{R}^3$.*

- *The parallelogram with edges v, w has area $\|v \times w\|$.*
- *The parallelepiped with edges u, v, w has volume $|u \cdot (v \times w)|$.*

Remark 2.42. Let v, w be vectors in $\mathbb{R}^2$. Write $v = (a, b)$ and $w = (c, d)$. Then the parallelogram defined by v, w has area

$$\|(a,b,0) \times (c,d,0)\| = \|(0,0,ad - bc)\| = |ad - bc| = \left| \det \begin{pmatrix} a & b \\ c & d \end{pmatrix} \right|.$$

Remark 2.43. Let $u = (x_1, y_1, z_1)$ let $v = (x_2, y_2, z_2)$ and let $w = (x_3, y_3, z_3)$ be vectors in $\mathbb{R}^3$. Then $|u \cdot (v \times w)|$ is equal to the determinant of the matrix with rows u, v, w. To see this, observe

$$\begin{aligned} u \cdot (v \times w) &= (x_1, y_1, z_1) \cdot \left(\det \begin{pmatrix} y_2 & z_2 \\ y_3 & z_3 \end{pmatrix}, -\det \begin{pmatrix} x_2 & z_2 \\ x_3 & z_3 \end{pmatrix}, \det \begin{pmatrix} x_2 & y_2 \\ x_3 & y_3 \end{pmatrix} \right) \\ &= x_1 \det \begin{pmatrix} y_2 & z_2 \\ y_3 & z_3 \end{pmatrix} - y_1 \det \begin{pmatrix} x_2 & z_2 \\ x_3 & z_3 \end{pmatrix} + z_1 \det \begin{pmatrix} x_2 & y_2 \\ x_3 & y_3 \end{pmatrix} \\ &= \det \begin{pmatrix} x_1 & x_2 & x_3 \\ y_1 & y_2 & y_3 \\ z_1 & z_2 & z_3 \end{pmatrix}. \end{aligned}$$

Example 2.44. The parallelogram defined by the vectors $(1, 2)$ and $(2, 1)$ has area equal to $\left| \det \begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix} \right| = |1 - 4| = 3$.

Example 2.45. The parallelepiped defined by the vectors $(1, 0, 0)$, $(0, 2, 1)$ and $(0, 0, 3)$ has volume $\left| \det \begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & 1 \\ 0 & 0 & 3 \end{pmatrix} \right| = 6$.

2.4. **Parametric Curves.** Let t be a real parameter. We will consider curves defined by functions of t. Consider the following function, whose input is a real parameter, and whose output is a vector in the plane.

$$s(t) = (\cos(t), \sin t), \qquad 0 \leq t \leq 2\pi.$$

The function $s(t)$ is called a **parametrization**. The function $s(t)$ defines a curve as follows. For each t with $0 \leq t \leq 2\pi$, we plot the point $s(t) = (\cos(t), \sin(t))$ in the plane. The set of all such points is then a curve in the plane. In fact, for this function $s(t)$, the resulting curve is the unit circle C, or the set of points (x, y) in the plane such that $x^2 + y^2 = 1$.

We write a general parametrization in the plane in the form

$$s(t) = (x(t), y(t)),$$

where $x(t)$ is a real-valued function of t, and $y(t)$ is a real-valued function of t.

Remark 2.46. We make a distinction between the parametrization s and the curve C itself. In particular, the parametrization is a function, but the curve C is a set of points. To see the difference, note that $s(t) = (\cos(t), \sin(t))$ with $0 \leq t \leq 4\pi$ is a parametrization that always lies in the unit circle C. However, s goes around the circle twice.

Example 2.47 (**Parametrizing Lines**)**.** Any line in the plane can be parametrized in the following way

$$s(t) = (a + bt, c + dt), \qquad a, b, c, d \in \mathbb{R}, \quad -\infty < t < \infty.$$

For example, let's describe the line that passes through the points $(0, 1)$ and $(2, 3)$. We know from single-variable calculus that this line has the equation $y = x + 1$. So, re-naming the parameter x as t, we see that $s(t) = (t, t + 1)$ gives the correct parametrization of this line, where $-\infty < t < \infty$

It is possible to find different parametrizations for the same line. For example, the parametrization $s_1(t) = (t^3, t^3 + 1)$ also parametrizes this line, where $-\infty < t < \infty$. Also, $s_2(t) = (t + 1, (t + 1) + 1)$ parametrizes this line.

We can also consider parametrized curves in three-dimensions. A general parametrization in $\mathbb{R}^3$ in the form

$$s(t) = (x(t), y(t), z(t)),$$

where $x(t)$ is a real-valued function of t, $y(t)$ is a real-valued function of t, and $z(t)$ is a real-valued function of t.

Example 2.48. Let $(x_0, y_0, z_0) \in \mathbb{R}^3$ and let $v = (a, b, c) \in \mathbb{R}^3$ be a vector. The line passing through (x_0, y_0, z_0) in the direction v can be parametrized as

$$s(t) = (x_0, y_0, z_0) + tv = (x_0 + ta, y_0 + tb, z_0 + tb), \quad -\infty < t < \infty.$$

Example 2.49 (Intersection of Lines). Suppose we parametrize two lines $s(t) = (t, 4t, 3+t)$ and $r(t) = (2 + t, t, 3 - t)$. Do these lines intersect?

If so, then we would have t_1, t_2 such that $s(t_1) = r(t_2)$. That is, we would have

$$t_1 = 2 + t_2, \qquad 4t_1 = t_2, \qquad 3 + t_1 = 3 - t_2.$$

The third equation says that $t_1 = -t_2$ and the second equation says $t_2 = 4t_1$. Together, they say that $t_1 = -4t_1$, so that $t_1 = 0$. Then $t_2 = -t_1 = 0$ as well. But then $0 = t_1 \neq 2 = 2 + t_2$, so it cannot occur that $s(t_1) = r(t_2)$. So, these lines do not intersect.

Note that two lines can either: coincide, intersect at a single point, or not intersect at all.

Example 2.50. Let $(x_0, y_0, z_0) \in \mathbb{R}^3$ and let $(x_1, y_1, z_1) \in \mathbb{R}^3$. The line passing through both of these points can be parametrized by

$$s(t) = t(x_1, y_1, z_1) + (1 - t)(x_0, y_0, z_0), \quad -\infty < t < \infty.$$

Note that $s(0) = (x_0, y_0, z_0)$, $s(1) = (x_1, y_1, z_1)$, and $s(1/2)$ is the midpoint between (x_0, y_0, z_0) and (x_1, y_1, z_1).

Example 2.51. An infinite helix can be parametrized by

$$s(t) = (\cos(t), \sin(t), t), \qquad -\infty < t < \infty.$$

Example 2.52 (Distance from a Point to a Line). Let u, v, w be fixed vectors. Let $s(t) = u + tv$ be a parametrization of a line through the origin. We define the distance of w from the line to be the minimum value of $\|w - s(t)\|$ over all t. Equivalently, the distance of w from the line is the minimum value of $\|w - s(t)\|^2$ over all t. Observe

$$\|w - s(t)\|^2 = (w - s(t)) \cdot (w - s(t)) = (w \cdot w) - 2(w \cdot s(t)) + (s(t) \cdot s(t))).$$

Taking the derivative in t, using the product rule, and setting this derivative equal to zero,

$$0 = \frac{d}{dt} \|w - s(t)\|^2 = -2(w \cdot s'(t)) + 2(s(t) \cdot s'(t)) = 2(s(t) - w) \cdot (s'(t)) = 2(s(t) - w) \cdot v.$$

That is, when $s(t)$ has minimal distance from w, we know that $s(t) - w$ is orthogonal to v. This fact allows us to find the minimum value of $\|s(t) - w\|$

Example 2.53. Find the distance of $w = (1, 0, 1)$ from the line $s(t) = (0, 1, 1) + t(1, 0, 0)$.

We need to find t such that $s(t) - (1, 0, 1) = (-1, 1, 0) + t(1, 0, 0)$ is orthogonal to $(2, 1, 0)$. That is, we need to solve for t when $(t - 1, 1, 0) \cdot (1, 0, 0) = 0$. That is, we need to find t such that $t - 1 = 0$. We therefore find that $t = 1$. So, when $s(t) = s(1) = (1, 1, 1)$, the distance $\|s(t) - w\|$ is minimized. For this t, we have $\|s(t) - w\| = \|(0, 1, 0)\| = 1$. So, the point $w = (1, 0, 1)$ has distance 1 from the line.

2.5. **Planes in Three Dimensions.**

Definition 2.54. Let $n = (a, b) \in \mathbb{R}^2$ be fixed and nonzero, let d be a constant, and let $(x, y) \in \mathbb{R}^2$ be a variable point. Recall that a **line** in the plane can be specified as the set of all points $(x, y) \in \mathbb{R}^2$ such that

$$ax + by = d.$$

Written another, way, a line is the set of all points $(x, y) \in \mathbb{R}^2$ such that

$$n \cdot (x, y) = d.$$

That is, a line is the set of all vectors (x, y) which have a constant dot product with the vector n. In the case $d = 0$, the line is the set of all vectors that are orthogonal to n. For this reason, we call n a **normal vector** to the line. (In the case $d \neq 0$, the line $n \cdot (x, y) = d$ is a translation of the line $n \cdot (x, y) = 0$.)

Definition 2.55. Let $n = (a, b, c) \in \mathbb{R}^3$ be fixed and nonzero, let d be a constant and let $v = (x, y, z) \in \mathbb{R}^3$ be a variable point. Then a **plane** is defined as the set of all points $(x, y, z) \in \mathbb{R}^3$ such that

$$ax + by + cz = d.$$

Written another, way, a plane is the set of all points $(x, y, z) \in \mathbb{R}^3$ such that

$$n \cdot (x, y, z) = d.$$

That is, a plane is the set of all vectors (x, y, z) which have a constant dot product with the vector n. In the case $d = 0$, the plane is the set of all vectors that are orthogonal to n. For this reason, we call n a **normal vector** to the plane. (In the case $d \neq 0$, the plane $n \cdot (x, y, z) = d$ is a translation of the plane $n \cdot (x, y, z) = 0$.)

Remark 2.56. A plane which passes through the point (x_0, y_0, z_0) with normal vector $n = (a, b, c)$ can be written as the set of all (x, y, z) such that

$$n \cdot (x - x_0, y - y_0, z - z_0) = 0.$$

Example 2.57. The plane which passes through $(0, 1, 2)$ with normal vector $(3, 0, 2)$ is described as the set of all (x, y, z) such that $(3, 0, 2) \cdot (x, y - 1, z - 2) = 0$.

Example 2.58 (**Plane through Three Points**)**.** Suppose we want to find the equation for a plane passing through $(1, 0, 0)$, $(0, 2, 0)$ and $(0, 1, 3)$. In order to find the equation of this plane, it is most convenient to find two vectors parallel to the plane, and to then take their cross product to find the normal vector. For example, $(0, 2, 0) - (1, 0, 0)$ is parallel to the plane, and so is $(0, 1, 3) - (1, 0, 0)$. (The plane will satisfy $n \cdot (x, y, z) = d$, so that $n \cdot (0, 2, 0) = d$ and $n \cdot (1, 0, 0) = d$, so that $n \cdot ((0, 2, 0) - (1, 0, 0)) = 0$. That is, $(0, 2, 0) - (1, 0, 0)$ is perpendicular to n. That is, $(0, 2, 0) - (1, 0, 0)$ is parallel to the plane. Similarly, $(0, 1, 3) - (1, 0, 0)$ is parallel to the plane.)

Now, to find a normal vector n, we compute

$$n = (-1, 2, 0) \times (-1, 1, 3) = (6, 3, -1 + 2) = (6, 3, 1).$$

So, the plane can be written as the set of all (x, y, z) such that $(6, 3, 1) \cdot (x, y, z) = d$ for some d. To find d, note that $(1, 0, 0)$ is in the plane, so $(6, 3, 1) \cdot (1, 0, 0) = 6 = d$. In conclusion, the plane is described as the set of all (x, y, z) where

$$(6, 3, 1) \cdot (x, y, z) = 6.$$

And we can verify that $(1, 0, 0)$, $(0, 2, 0)$ and $(0, 1, 3)$ all satisfy this equation.

Example 2.59 (Intersection of Two Planes). Let a, b, c, d, p, q, r, s be constants. What is the intersection of the planes $ax+by+cz = d$ and $px+qy+rz = s$? (Assume $(a, b, c) \neq (0, 0, 0)$ and $(p, q, r) \neq (0, 0, 0)$.)

We could solve for one variable and substitute that equation into the other equation, and this could give a (potentially complicated) algebraic description of the intersection. It is sometimes more convenient to first find a vector parallel to the line of intersection. This vector is given by $(a, b, c) \times (p, q, r)$. Note that if (x, y, z) is in the first plane, then $(x, y, z) + t(a, b, c) \times (p, q, r)$ is also in the first plane, since $(a, b, c) \cdot [(a, b, c) \times (p, q, r)] = 0$, so $(a, b, c) \cdot ((x, y, z) + t(a, b, c) \times (p, q, r)) = (a, b, c) \cdot (x, y, z) = d$. That is, the parametrized line $s(t) = (x, y, z) + t(a, b, c) \times (p, q, r)$ lies in the first plane. By similar reasoning, this line also lies in the second plane (if (x, y, z) is in the second plane too). So, the parametrized line $s(t) = (x, y, z) + t(a, b, c) \times (p, q, r)$ lies in both planes.

For example, consider the planes $x+y+2z = 1$ and $2x+y = 3$. The line of intersection is then parallel to $(1, 1, 2) \times (2, 1, 0) = (-2, 4, -1)$. Also, equating the planes gives one point in the intersection. That is, we have $x = 1-y-2z$, so $2(1-y-2z)+y = 3$, so $-y-4z = 1$. So, for example, the point $(0, 3, -1)$ is in the line of intersection. Since this line also is parallel to the vector $(-2, 4, -1)$, we can parametrize this line by $u(t) = (0, 3, -1) + t(-2, 4, -1)$.

Note that the two planes are parallel if and only if their normal vectors are parallel. So, the two planes are parallel if and only if $(a, b, c) \times (p, q, r) = (0, 0, 0)$.

Example 2.60 (Distance from a Point to a Plane). What is the distance of the point $w \in \mathbb{R}^3$ from the plane $ax + by + cz = 0$?

The distance from w to the plane $ax + by + cz = 0$ is the smallest value of $\|w - (x, y, z)\|$ where (x, y, z) ranges over all points in the plane $ax + by + cz = 0$. As in the case of a line, the point (x, y, z) where $\|w - (x, y, z)\|$ is smallest occurs when $w - (x, y, z)$ is perpendicular to the plane. We claim that $\|w - (x, y, z)\|$ is smallest when $w - (x, y, z) = \mathrm{proj}_{(a,b,c)}(w)$. To see this, consider the right triangle with edges $w - (x, y, z)$ and $\mathrm{proj}_{(a,b,c)}(w)$. Since this is a right triangle, the Pythagorean Theorem implies that the length of the hypotenuse is greater than the lengths of the other edges. That is, $\|w - (x, y, z)\| \geq \|\mathrm{proj}_{(a,b,c)}(w)\|$. In conclusion, the distance from w to the plane $ax + by + cz = 0$ is

$$\left\|\mathrm{proj}_{(a,b,c)}(w)\right\|.$$

For example, let's find the distance of $w = (1, 0, 2)$ to the plane $x + 2y + 3z = 0$. Since $(a, b, c) = (1, 2, 3)$, the distance is given by

$$|w \cdot (a, b, c)/\left\|a, b, c\right\|| = |(1, 0, 2) \cdot (1, 2, 3)| / \sqrt{14} = 7/\sqrt{14}.$$

Definition 2.61. The **angle** between two planes is the angle between their normal vectors. So, if the planes are $ax + by + cz = d$ and $px + qy + rz = s$, then their angle is

$$\cos^{-1}\left(\frac{|(a,b,c)\cdot(p,q,r)|}{\|(a,b,c)\|\,\|(p,q,r)\|}\right).$$

2.6. **Quadric Surfaces.** Let A, B, C, D, E, F be constants. Recall that a general conic section in the plane is a set of points $(x, y) \in \mathbb{R}^2$ satisfying

$$Ax^2 + By^2 + Cxy + Dx + Ey + F = 0.$$

A conic section of the following form is said to be in **standard position**.

$$Ax^2 + By^2 + Ey + F = 0.$$

For example, when $E = 0, A = 1, B = 1, F = -1$ corresponds to the unit circle $x^2 + y^2 = 1$, i.e. the circle centered at the origin with radius 1; $A = 1, B = 1/2, F = -1$ corresponds to an ellipse $x^2 + 2y^2 = 1$ with axis lengths 1 and $1/\sqrt{2}$; $A = 1, B = -1, F = -1$ corresponds to a hyperbola $x^2 - y^2 = 1$; and $y = x^2$ is a parabola, and so on.

The main non-degenerate conic sections are then: parabolas, hyperbolas, circles and ellipses. Some degenerate conic sections occur e.g. when $x^2 = y^2$, which is the union of two intersecting lines, or when $x^2 + y^2 = 0$, which is a single point $(x, y) = (0, 0)$.

Quadric surfaces generalize conic sections to three-dimensions. A general quadric surface is a set of points $(x, y, z) \in \mathbb{R}^3$ satisfying

$$Ax^2 + By^2 + Cz^2 + Dxy + Eyz + Fzx + Gx + Hy + Iz + J = 0.$$

The main non-degenerate quadric surfaces are: paraboloids, hyperboloids, spheres, ellipsoids, cones and cylinders. Some degenerate quadric surfaces occur e.g. when $x^2 = y^2$, which is the union of two intersecting planes, or when $x^2 + y^2 = 0$, which is a single line where $x = y = 0$ and $-\infty < z < \infty$, or when $x^2 + y^2 + z^2 = 0$, which is the single point $(0, 0, 0)$.

Below, we will focus on quadric surfaces in **standard position**.

Example 2.62. The case $A = B = C = 1$ and $J = -1$ with all other coefficients zero corresponds to the unit **sphere** $x^2 + y^2 + z^2 = 1$, i.e. the sphere centered at the origin with radius 1. More generally, if $r > 0$, then $x^2 + y^2 + z^2 = r^2$ defines the sphere of radius r centered at the origin.

Example 2.63. The case $A = 1, B = 2, C = 3$ and $J = -1$ with all other coefficients zero corresponds to an ellipsoid $x^2 + 2y^2 + 3z^2 = 1$, which is a surface which is somewhat similar to an ellipse. In particular, when $z = 0$ (which corresponds to intersection this ellipse with the xy-plane), we get $x^2 + 2y^2 = 1$, which is an ellipse. That is, the cross-section of the ellipsoid is an ellipse. A general **ellipsoid** is written as the set of all $(x, y, z) \in \mathbb{R}^3$ such that

$$\left(\frac{x}{a}\right)^2 + \left(\frac{y}{b}\right)^2 + \left(\frac{z}{c}\right)^2 = 1, \qquad a, b, c > 0.$$

An ellipsoid generally looks like a sphere that has been squished in certain directions.

Example 2.64. The case $A = 1, B = -1, C = 1$ and $J = -1$ with all other coefficients zero corresponds to a hyperboloid $x^2 - y^2 + z^2 = 1$, which is a surface which is somewhat similar to a hyperbola. In particular, when $z = 0$ (which corresponds to intersection this hyperboloid with the xy-plane), we get $x^2 - y^2 = 1$, which is a a hyperbola. That is, this cross-section of the hyperboloid is a hyperbola. However, another cross-section of the hyperboloid is an ellipse. When $y = 0$ (which corresponds to the intersection of this hyperboloid with the xz-plane), we get $x^2 + z^2 = 1$, which is an ellipse (in fact a circle).

A general **hyperboloid of one sheet** is written as the set of all $(x, y, z) \in \mathbb{R}^3$ such that

$$\left(\frac{x}{a}\right)^2 + \left(\frac{y}{b}\right)^2 = \left(\frac{z}{c}\right)^2 + 1, \qquad a, b, c > 0.$$

A general **hyperboloid of two sheets** is written as the set of all $(x, y, z) \in \mathbb{R}^3$ such that

$$\left(\frac{x}{a}\right)^2 + \left(\frac{y}{b}\right)^2 = \left(\frac{z}{c}\right)^2 - 1, \qquad a, b, c > 0.$$

The hyperboloid of one sheet is a single surface, while the hyperboloid of two sheets is two separate surfaces. To see the latter property, note that the hyperboloid of two sheets contains the points $(0, 0, c)$ and $(0, 0, -c)$, but it does not contain any point (x, y, z) with $-c < z < c$, since the left side of the defining equation would be nonnegative, while the right side would be negative.

Example 2.65. A **cone** is broadly defined as any surface such that, if (x, y, z) is in the cone, then so is $\lambda(x, y, z)$ for all $\lambda \geq 0$. For example, a general **elliptic cone** is the set of all (x, y, z) such that

$$\left(\frac{x}{a}\right)^2 + \left(\frac{y}{b}\right)^2 = \left(\frac{z}{c}\right)^2, \qquad a, b, c > 0.$$

Note that if (x, y, z) is in this elliptic cone, then so is $\lambda(x, y, z)$, for all $\lambda \geq 0$.

Example 2.66. A general **elliptic paraboloid** is the set of all (x, y, z) such that

$$z = \left(\frac{x}{a}\right)^2 + \left(\frac{y}{b}\right)^2, \qquad a, b > 0.$$

This paraboloid looks like a parabola that has been rotated around an axis. And a general **hyperbolic paraboloid** is the set of all (x, y, z) such that

$$z = \left(\frac{x}{a}\right)^2 - \left(\frac{y}{b}\right)^2, \qquad a, b > 0.$$

This last paraboloid looks like a saddle, or a pringle potato chip.

Example 2.67. A **cylinder** is broadly defined as any curve which is extruded across a perpendicular direction. For example, if C is curve in the xy-plane, then the set of all vertical lines passing through C is a cylinder. An example of an **elliptical cylinder** is the set of all (x, y, z) such that $x^2 + 4y^2 = 1$. Note that the z-coordinate is unrestricted, so that the ellipse with radii 1 and 2 is extruded along the z-axis.

An example of an **hyperbolic cylinder** is the set of all (x, y, z) such that $x^2 - y^2 = 1$.

An example of an **parabolic cylinder** is the set of all (x, y, z) such that $y = x^2$.

3. Vector-Valued Functions

In your previous calculus classes, we typically considered real-valued functions on the real line. That is, we considered functions $f : \mathbb{R} \to \mathbb{R}$. In multivariable calculus, we consider functions with more variables in the domain, and with more variables in the range. Later on we will deal with having more variables in the domain. For now, we consider functions with more variables in the range. That is, we will consider functions of the form

$$r(t) = (x(t), y(y))$$

and of the form

$$s(t) = (x(t), y(t), z(t))$$

where t is a real variable, and x, y, z are each real-valued functions of the real variable t. Since r has a real domain and a two-dimensional range, we write $r \colon \mathbb{R} \to \mathbb{R}^2$. Similarly, we write $s \colon \mathbb{R} \to \mathbb{R}^3$.

We have already seen functions like this in our discussion of parametrizations.

Example 3.1. Consider again the infinite helix

$$s(t) = (\cos t, \sin t, t), \qquad -\infty < t < \infty$$

Visualizing this parametrized curve can be made simpler by consider its projection onto the coordinate hyperplanes. The projection of the parametrization onto the xy-plane gives a parametrization of the unit circle

$$(\cos t, \sin t), \qquad -\infty < t < \infty.$$

The projection of the parametrization onto the xz-plane gives a parametrization of a cosine curve

$$(\cos t, t), \qquad -\infty < t < \infty.$$

And the projection of the parametrization onto the yz-plane gives a parametrization of a sine curve

$$(\sin t, t), \qquad -\infty < t < \infty.$$

Remark 3.2. Recall that there are many different ways to parametrize a given curve. For example, $r(t) = (\cos t, \sin t, t + 2\pi)$ also parametrizes the helix. And the function $h(t) = (\cos(t^3), \sin(t^3), t^3)$, $-\infty < t < \infty$ also parametrizes the helix. Note that $h(t) = s(t^3)$. That is, h is a **re-parametrization** of s.

Example 3.3 (**Intersection of Two Surfaces**)**.** We now describe how to parametrize the intersection of two surfaces in $\mathbb{R}^3$. For example, consider the intersection of the cylinder $x^2 + y^2 = 1$ with the paraboloid $x^2 - y^2 = z - 1$. In order to find the intersection, we substitute one formula into the other. Since $x^2 = 1 - y^2$, we then get $(1 - y^2) - y^2 = z - 1$, so that $2 - 2y^2 = z$. Since $x^2 + y^2 = 1$, we get $x^2 = 1 - y^2 = z/2$. So, using x as a parameter, we can parametrize the intersection as

$$s(x) = (x, \sqrt{1 - x^2}, 2x^2), \quad -1 \leq x \leq 1.$$

However, this parametrization only obtains part of the intersection, since whenever $x^2 + y^2 = 1$, we also have $x^2 + (-y)^2 = 1$. That is, the intersection could have both positive and negative values of y, whereas $s(x)$ only has nonnegative values of y. We therefore obtain the rest of the intersection with the parametrization

$$r(x) = (x, -\sqrt{1 - x^2}, 2x^2), \quad -1 \leq x \leq 1.$$

We could have also described this intersection with a single parametrization:

$$s(t) = (\cos(t), \sin(t), \cos^2 t - \sin^2 t + 1)., \qquad 0 \leq t < 2\pi$$

3.1. **Calculus for Vector-Valued Functions.** We now begin to extend the notions of calculus to the setting of vector-valued functions. As in single-variable calculus, we first discuss limits and continuity, and we then define derivatives. So, let us begin with limits.

Definition 3.4 (**Limit**)**.** Suppose $r\colon \mathbb{R} \to \mathbb{R}^3$ is a vector-valued function. Let $v \in \mathbb{R}^3$ be a vector, and let $t_0 \in \mathbb{R}$. We say that $r(t)$ approaches v as $t \to t_0$ if $\lim_{t\to t_0} \|r(t) - v\| = 0$. In this case, we write

$$\lim_{t\to t_0} r(t) = v.$$

Remark 3.5. Suppose $r\colon \mathbb{R} \to \mathbb{R}^3$ is a vector-valued function. Write $r(t) = (x(t), y(t), z(t))$. Let $v = (x, y, z) \in \mathbb{R}^3$ be a vector, and let $t_0 \in \mathbb{R}$. Then $\lim_{t\to t_0} \|r(t) - v\| = 0$ if and only if the following three limits exist: $\lim_{t\to t_0} x(t)$, $\lim_{t\to t_0} y(t)$, $\lim_{t\to t_0} z(t)$. In this case, we have

$$\lim_{t\to t_0} r(t) = \left(\lim_{t\to t_0} x(t),\ \lim_{t\to t_0} y(t),\ \lim_{t\to t_0} z(t)\right).$$

Example 3.6. Let $s(t) = (t-1, t^2, 2t-3)$. Then to compute $\lim_{t\to 0} s(t)$, we use

$$\lim_{t\to 0} s(t) = (\lim_{t\to 0}(t-1),\ \lim_{t\to 0} t^2,\ \lim_{t\to 0}(2t-3)) = (-1, 0, -3).$$

Definition 3.7 (**Continuity**)**.** Suppose $r\colon \mathbb{R} \to \mathbb{R}^3$ is a vector-valued function. We say that r is **continuous** at $t_0 \in \mathbb{R}$ if

$$r(t_0) = \lim_{t\to t_0} r(t).$$

From Remark 3.5, if $r(t) = (x(t), y(t), z(t))$, then r is continuous at t_0 if and only if: $x(t), y(t)$ and $z(t)$ are each continuous at t_0.

Definition 3.8 (**Derivative**)**.** Suppose $r\colon \mathbb{R} \to \mathbb{R}^3$ is a vector-valued function. We say that r is **differentiable** at $t \in \mathbb{R}$ if the following limit exists

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

We refer to $r'(t)$ as the **derivative** of r at t. If r is differentiable on its domain, we say that r is **differentiable**. Note that since r is a vector-valued function, the difference quotient is a limit of vectors, rather than a limit of numbers. Consequently, $r'\colon \mathbb{R} \to \mathbb{R}^3$.

Remark 3.9. We can then define higher order derivatives by iterating the first derivative. For example, if r' is differentiable, we define $r''(t) = (d/dt)r'(t)$. And if r'' is differentiable, we define $r'''(t) = (d/dt)r''(t)$, and so on.

Remark 3.10. Suppose $r\colon \mathbb{R} \to \mathbb{R}^3$ is a vector-valued function, where $r(t) = (x(t), y(t), z(t))$. Then $r(t)$ is differentiable if and only if: $x(t), y(t)$ and $z(t)$ are all differentiable. In this case, we have

$$r'(t) = (x'(t), y'(t), z'(t)).$$

Example 3.11. Let $r(t) = (t, t^2 - 1, t^{-1})$. Then $r'(t) = (1, 2t, -t^{-2})$, and $r''(t) = (0, 2, 2t^{-3})$.

Most of the usual rules of differentiation apply to vector-valued functions.

Proposition 3.12. *Let $s, r, \colon \mathbb{R} \to \mathbb{R}^3$, and let $f\colon \mathbb{R} \to \mathbb{R}$ all be differentiable functions. Let c be a constant.*

- $\frac{d}{dt}(s(t) + r(t)) = s'(t) + r'(t)$. ***(Sum Rule)***

- $\frac{d}{dt}(cs(t)) = cs'(t)$.
- $\frac{d}{dt}(f(t)s(t)) = f'(t)s(t) + f(t)s'(t)$. ***(Product Rule)***
- $\frac{d}{dt}s(f(t)) = s'(f(t))f'(t)$. ***(Chain Rule)***
- $\frac{d}{dt}(s(t) \cdot r(t)) = (s'(t) \cdot r(t)) + (s(t) \cdot r'(t))$ ***(Product Rule for Dot Product)***
- $\frac{d}{dt}(s(t) \times r(t)) = (s'(t) \times r(t)) + (s(t) \times r'(t))$ ***(Product Rule for Cross Product)***

Remark 3.13 (**Geometric Interpretation of Derivative**)**.** When we have a real-valued function $f: \mathbb{R} \to \mathbb{R}$, we recall that $f'(t)$ is the slope of the tangent line to f at the point t. This fact can be understood from the definition of the derivative itself

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

The difference quotient $[f(t+h) - f(t)]/h$ is the slope of the line which passes between the points $(t, f(t))$ and $(t+h, f(t+h))$ in the plane.

There is a similar interpretation for vector-valued functions $r: \mathbb{R} \to \mathbb{R}^3$. We have

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

The difference quotient is itself a line segment passing through the points $(t, r(t))$ and $(t+h, r(t+h))$. And as $h \to 0$, this line segment becomes a tangent vector to the parametrized curve $r(t)$.

Example 3.14. Let $r(t) = (t, t^2, 2+t)$. The tangent line to the parametrized curve $r(t)$ at $t = 2$ is given by the following parametrized line

$$h(t) = r(2) + t \cdot r'(2) = (2, 4, 4) + t(1, 4, 1), \quad -\infty < t < \infty.$$

Remark 3.15 (**Orthogonality of r and r' when $\|r\|$ is constant**)**.** Let $r: \mathbb{R} \to \mathbb{R}^3$. Assume that $\|r(t)\|$ is equal to a constant value for all t. Then

$$0 = \frac{d}{dt}\|r(t)\|^2 = \frac{d}{dt}(r(t) \cdot r(t)) = 2r(t) \cdot r'(t).$$

That is, $r(t)$ is orthogonal to $r'(t)$.

Definition 3.16 (**Integral**)**.** Let $r: \mathbb{R} \to \mathbb{R}^3$, where $r(t) = (x(t), y(t), z(t))$. Let $a < b$. We define the integral of r over $[a, b]$ to be the following vector

$$\int_a^b r(t)dt = \left(\int_a^b x(t)dt, \int_a^b y(t)dt, \int_a^b z(t)dt\right)$$

Example 3.17. Let $r(t) = (t, t^2, 1)$. Then $\int_0^1 r'(t)dt = (\int_0^1 1dt, \int_0^1 2tdt, \int_0^1 0dt) = (1, 1, 0)$.

Theorem 3.18 (**Fundamental Theorem of Calculus**)**.** *Let $a < b$. Let $r: [a, b] \to \mathbb{R}^3$ be differentiable. Assume that $r': [a, b] \to \mathbb{R}^3$ is continuous. Then*

$$\int_a^b r'(t)dt = r(b) - r(a).$$

Example 3.19. Let $r(t) = (t, t^2, 1)$. Then $\int_0^1 r'(t)dt = r(1) - r(0) = (1, 1, 1) - (0, 0, 1) = (1, 1, 0)$.

3.2. **Arc Length, Speed, Curvature.**

Definition 3.20 (**Arc Length**). Let $r\colon \mathbb{R} \to \mathbb{R}^3$. Write $r(t) = (x(t), y(t), z(t))$. Let $a < b$. We define the **arc length** of r from $t = a$ to $t = b$ by

$$\int_a^b \|r'(t)\|\,dt = \int_a^b \sqrt{(x'(t))^2 + (y'(t))^2 + (z'(t))^2}\,dt.$$

Example 3.21. Let's compute the arc length of the parametrization $r(t) = (\cos(t), \sin t, 0)$, $0 \leq t < 2\pi$, which parametrizes the unit circle. The length is

$$\int_0^{2\pi} \|r'(t)\|\,dt = \int_0^{2\pi} \sqrt{\cos^2 t + \sin^2 t}\,dt = \int_0^{2\pi} dt = 2\pi.$$

However, note that if we use a different domain, then we could "double-count" the unit circle's length. For example, if we use the same parametrization with $0 \leq t < 4\pi$, then we would compute a length of 4π. In this case, $r(t)$ would traverse the circle twice, resulting in a length of $2(2\pi) = 4\pi$. So, given a parametric curve, we can not necessarily associate its arc length with the length of the curve itself (unless $r(t_1) \neq r(t_2)$ whenever $t_1 \neq t_2$).

It is often more convenient to have a parametrization r such that $\|r'(t)\| = 1$. Starting with essentially any parametrization, we can re-parametrize it to achieve this condition, as we now show.

Definition 3.22 (**Arc Length Parametrization**). Let $r\colon [a,b] \to \mathbb{R}^3$ be any parametrized curve. Suppose $r'(t) \neq (0,0,0)$ for all t. Define $\ell(t) = \int_a^t \|r'(u)\|\,du$. Then $\ell'(t) = \|r'(t)\| > 0$ for all t, using the Fundamental Theorem of Calculus. Since $\ell\colon [a,b] \to \mathbb{R}$ is increasing, it has an inverse function ℓ^{-1}, which we denote by $\phi(s)$. Recall that $\phi'(s) = 1/(\ell'(\phi(s)))$. We define the **arc length parametrization** of r by $r_1(s) = r(\phi(s))$. This parametrization is also called the **unit speed parametrization** since we have $\|r_1'(s)\| = 1$. Observe, by the Chain Rule,

$$\frac{d}{ds} r_1(s) = \frac{d}{ds} r(\phi(s)) = \frac{dr}{dt}(\phi(s))\frac{d\phi(s)}{ds} = r'(\phi(s))\frac{1}{\ell'(\phi(s))} = \frac{r'(\phi(s))}{\|r'(\phi(s))\|}.$$

Therefore, $\|r_1'(s)\| = \|r'(\phi(s))\| \,/\, \|r'(\phi(s))\| = 1$, as desired.

Example 3.23. Consider $r(t) = (t, 2t, 1)$, $t \geq 0$. Then $r'(t) = (1, 2, 0)$, so we have $\ell(t) = \int_0^t \|r'(u)\|\,du = \int_0^t \sqrt{5}du = \sqrt{5}t$. So, $\phi(s) = s/\sqrt{5}$ (since $\ell(\phi(s)) = s$). Therefore, $r_1(s) = r(\phi(s)) = (s/\sqrt{5}, 2s/\sqrt{5}, 1)$ is an arc length parametrization for r.

Definition 3.24. Let $r\colon \mathbb{R} \to \mathbb{R}^3$. We say that a parametrization is **regular** if $r'(t) \neq (0,0,0)$ for all t. If r is regular, we define the **unit tangent vector** $T(t)$ by

$$T(t) = \frac{r'(t)}{\|r'(t)\|}.$$

Recall that $r'(t)$ is tangent to the curve r. So, we define $T(t)$ so that T is still tangent to the curve r, while $\|T(t)\| = \|r'(t)\| \,/\, \|r'(t)\| = 1$.

It is often convenient to know the unit-tangent vector of a curve. For example, the unit tangent vector can be used to measure how much a curve is "twisted around itself." The latter notion can be made precise with the introduction of curvature.

Definition 3.25 (**Curvature**). Let $r(s)$ be an arc-length parametrization of a curve. We define the **curvature** $\kappa(s)$ of r at s to be

$$\kappa(s) = \|r''(s)\| = \left\|(d^2/ds^2)r(s)\right\| = \|(d/ds)T(s)\|.$$

Example 3.26. The curvature of any straight line is always zero. To see this, let $r(s) = (x_0, y_0, z_0) + su$ be a parametrization of a line, where $u \in \mathbb{R}^3$ is a vector. If $\|u\| = 1$, then $\|r'(s)\| = \|u\| = 1$, so $r(s)$ is an arc-length parametrization. And $r''(s) = 0$, so $\kappa(s) = 0$.

Example 3.27. Let $R > 0$. The curvature of the circle of radius R is $1/R$. We use the parametrization $r(t) = (R\cos t, R\sin t)$, $0 \leq t < 2\pi$. Then $\|r'(t)\| = \sqrt{R^2\sin^2 t + R^2\cos^2 t} = R$. And $\ell(t) = \int_0^t \|r'(u)\|\, du = Rt$. So, the inverse of ℓ is $\phi(s) = s/R$ (since $\ell(\phi(s)) = s$). And the arc length parametrization of the circle of radius R is then $r_1(s) = r(\phi(s)) = (R\cos(s/R), R\sin(s/R))$. So, the curvature is

$$\kappa(s) = \|r_1''(s)\| = \left\|(-R^{-1}\cos(s/R), -R^{-1}\sin(s/R))\right\| = R^{-1}.$$

Proposition 3.28 (**Other Formulas for Curvature**). *Let $r\colon \mathbb{R} \to \mathbb{R}^3$. Then*

$$\kappa(t) = \frac{\|T'(t)\|}{\|r'(t)\|} = \frac{\|r'(t) \times r''(t)\|}{\|r'(t)\|^3}.$$

Proof. The first formula follows from the Chain rule. Let $\ell(t) = \int_0^t \|r'(u)\|\, du$. Let $\phi(s)$ be the inverse of $\ell(t)$. Recall that $r(\phi(s))$ is an arc length parametrization of r. Note that $T(\phi(\ell(t))) = T(t)$, so by the Chain rule

$$T'(t) = \frac{d}{dt}T(\phi(\ell(t))) = \frac{d}{d\ell}T(\phi(\ell(t)))\frac{d\ell(t)}{dt} = \frac{d}{d\ell}T(\phi(\ell(t)))\,\|r'(t)\| = \frac{d}{d\ell}T(t)\,\|r'(t)\|.$$

So, $\|T'(t)\| = \|(d/d\ell)T(t)\|\,\|r'(t)\| = \kappa(t)\,\|r'(t)\|$.

For the second formula, note that $T(t) = r'(t)/\|r'(t)\|$, so $r'(t) = T(t)\,\|r'(t)\|$, so $r''(t) = T'(t)\,\|r'(t)\| + T(t)(d/dt)\,\|r'(t)\|$. Then using $T(t) \times T(t) = (0,0,0)$,

$$r'(t) \times r''(t) = (T(t)\,\|r'(t)\|) \times (T'(t)\,\|r'(t)\| + T(t)(d/dt)\,\|r'(t)\|) = (T(t) \times T'(t))\,\|r'(t)\|^2.$$

Finally, recalling Remark 3.15, since $T(t)$ is of unit length, $T(t)\cdot T'(t) = 0$. So, using Theorem 2.39(ii), we get

$$\begin{aligned}\|r'(t) \times r''(t)\| &= \|T(t) \times T'(t)\|\,\|r'(t)\|^2 \\ &= \|T(t)\|\,\|T'(t)\|\,\|r'(t)\|^2 = \|T'(t)\|\,\|r'(t)\|^2 = \kappa(t)\,\|r'(t)\|^3.\end{aligned}$$

In the last line, we used the first formula: $\|T'(t)\| / \|r'(t)\| = \kappa(t)$. □

3.3. **Motion of Projectiles and Planets.** Let $r\colon \mathbb{R} \to \mathbb{R}^3$. We think of $r(t)$ as the position of an object at time t. We define the **velocity** $v(t)$ of the object at time t by

$$v(t) = r'(t).$$

Recall that the speed of the object is $\|v(t)\| = \|r'(t)\|$. And we define the **acceleration** $a(t)$ of the object at time t by

$$a(t) = v'(t) = r''(t).$$

Example 3.29 (**Projectile Motion**). Suppose a cannon launches a cannonball on the earth with initial position $r_0 \in \mathbb{R}^3$ and initial velocity $v_0 \in \mathbb{R}^3$. We treat the z-axis as the vertical direction, and we ignore air friction. Let g be the gravitational constant (so $g \approx 9.8$ in standard units). Once the projectile is in the air, the only acceleration acting upon it acceleration due to gravity. That is, $r''(t) = (0,0,-g)$. The initial conditions mean that $r(0) = r_0$ and $r'(0) = v(0) = v_0$. Integrating from 0 to t, and applying the Fundamental Theorem of Calculus (Theorem 3.18),

$$v(t) - v(0) = \int_0^t v'(p)dp = \int_0^t r''(p)dp = \int_0^t (0,0,-g) = (0,0,-gt).$$

That is $v(t) = v_0 + (0,0,-gt)$. Integrating again from 0 to t, we get

$$r(t) - r(0) = \int_0^t r'(p)dp = \int_0^t v(p)dp = \int_0^t (v_0 + (0,0,-gp))dp = tv_0 + (0,0,-gt^2/2).$$

That is, the trajectory of the projectile is given by the following parabola.

$$r(t) = r_0 + tv_0 + (0,0,-gt^2/2).$$

Example 3.30. Consider a particle moving in a circle $r(t) = (\cos t, \sin t)$, $t \geq 0$. Then $v(t) = r'(t) = (-\sin t, \cos t)$, and $a(t) = v'(t) = r''(t) = (-\cos t, -\sin t) = -r(t)$. That is, the acceleration of the particle always points towards the center of the circle.

We now describe Kepler's Laws for planetary motion. We have the following simplified assumptions. The sun is placed at the origin $(0,0,0)$. A single planet orbits the sun. Both the sun and the planet are treated as point masses. Suppose the planet is orbiting the sun with position $r(t)$ at time $t \geq 0$. Let $A(t)$ denote the area of the sector formed by the origin, $r(0), r(t)$ and the curve r itself. Let $T > 0$ be the time required for the planet to complete one full revolution, i.e. so that $r(T) = r(0)$.

Theorem 3.31 (**Kepler's Laws for Planetary Motion**).

- *The curve $r(t)$ is an ellipse with the sun at one focus.* ***(Law of Ellipses)***
- *$\frac{d}{dt}A(t)$ is a constant which does not depend on t. Consequently, $A(t_1) - A(t_2) = A(t_1 - t_2)$.* ***(Law of Equal Area in Equal Time)***
- *$T^2 = 4\pi^2 a^3/(GM)$.* ***(Law of the Period of Motion)****. Here a is the length of the semimajor axis of the ellipse, i.e. a is half the length of the longest segment contained in the ellipse, G is the Gravitational Constant, and M is the mass of the sun.*

Remark 3.32. Nowadays we know that the Law of Ellipses is not exactly right, but we cannot presently discuss this issue.

Before proving these laws, we derive a preliminary fact. From Newton's Law of Gravitation, if the planet has mass m, then the force of gravity acts on the planet in the direction $-r$ as follows:

$$mr''(t) = -\frac{GMm}{\|r(t)\|^2}\frac{r(t)}{\|r(t)\|} = -\frac{GMm}{\|r(t)\|^3}r(t).$$

That is, $r''(t)$ is parallel to $r(t)$. Consequently,

$$r(t) \times r''(t) = (0,0,0). \tag{1}$$

Proof of the second law. Let $h > 0$ be small. Then $A(t+h) - A(t)$ is approximately a triangle formed by the vectors $r(t)$ and $r(t+h)$. This triangle has area equal to half the area of the paralleogram formed by the vectors $r(t)$ and $r(t+h)$. This parallelogram is also defined by the vectors $r(t)$ and $r(t+h) - r(t)$. So, using Theorem 2.41, $A(t+h) - A(t) = \frac{1}{2} \|r(t) \times (r(t+h) - r(t))\|$. So,

$$\frac{d}{dt}A(t) = \lim_{h\to 0} \frac{A(t+h) - A(t)}{h} = \lim_{h\to 0} \frac{1}{2}\frac{1}{h} \|r(t) \times (r(t+h) - r(t))\|$$
$$= \lim_{h\to 0} \frac{1}{2} \left\| r(t) \times \frac{r(t+h) - r(t)}{h} \right\| = \frac{1}{2} \|r(t) \times r'(t)\|.$$

It remains to show that $\|r(t) \times r'(t)\|$ does not depend on t. In fact, we will show that $r(t) \times r'(t)$ is a constant vector. That is, we will show that $(d/dt)(r(t) \times r'(t)) = (0,0,0)$. Observe, using the product rule and (1),

$$\frac{d}{dt}(r(t) \times r'(t)) = (r(t) \times r''(t)) + (r'(t) \times r'(t)) = r(t) \times r''(t) = (0,0,0).$$

□

Remark 3.33. Since $r(t) \times r'(t)$ is orthogonal to both $r(t)$ and $r'(t)$, and $r(t) \times r'(t)$ is a constant vector, we conclude that $r(t)$ and $r'(t)$ are both always orthogonal to the fixed vector $r(t) \times r'(t)$ (and the latter vector does not depend on t, as we just verified). So, the orbit $r(t)$ is always contained in a plane.

Proof of the first law. Define

$$L(t) = \left(\frac{1}{GMm} r'(t) \times [r(t) \times r'(t)] \right) - \frac{r(t)}{\|r(t)\|}.$$

One can show that $\frac{d}{dt}L(t) = (0,0,0)$. Assuming this fact, we know that $L(t)$ is a constant vector, so let $D = \|L(t)\|$ be a constant which does not depend on t. Recall also that in our proof of the second law, $r(t) \times r'(t)$ was a constant vector as well. So let $p = (r(t) \times r'(t)) \cdot (r(t) \times r'(t))/(GMm)$ be a constant which also does not depend on t. If $\theta(t)$ denotes the angle between $r(t)$ and $L(t)$, we have, using the identity $u \cdot (v \times w) = w \cdot (u \times v)$,

$$\|r(t)\| D \cos\theta(t) = r(t) \cdot L(t) = \frac{1}{GMm} r(t) \cdot (r'(t) \times [r(t) \times r'(t)]) - \frac{r(t) \cdot r(t)}{\|r(t)\|}$$
$$= \frac{1}{GMm} [r(t) \times r'(t)] \cdot [r(t) \times r'(t)] - \frac{r(t) \cdot r(t)}{\|r(t)\|} = p - \|r(t)\|.$$

Solving for $\|r(t)\|$, we get

$$\|r(t)\| = \frac{p}{1 + D\cos\theta(t)}.$$

In the case $D = 0$, this says $\|r(t)\| = p$ so that $r(t)$ is a circle of radius p. Suppose then that $D \neq 0$. Since r is contained in a plane by Remark 3.33, and L is a constant vector, without loss of generality L points in the direction of the positive x axis, and $r(t)$ is contained in the xy-plane. Then, we can write $r(t) = (x, y, 0) = (x(t), y(t), 0) = (\|r(t)\| \cos\theta(t), \|r(t)\| \sin\theta(t), 0)$. We then have $\|r(t)\| (1 + D\cos\theta(t)) = p$, or $\|r(t)\| =$

$p - D\|r(t)\|\cos\theta(t) = p - Dx$. Squaring gives $x^2 + y^2 = \|r(t)\|^2 = (p - Dx)^2$, so that $x^2(1 - D^2) + 2pDx + y^2 = p^2$, and multiplying by $(1 - D^2)$ and completing the square,

$$(1 - D^2)^2\left(x + \frac{pD}{1 - D^2}\right)^2 + y^2(1 - D^2) = p^2 \qquad \text{if } D \neq 1.$$

(If $D = 1$, we get $2px + y^2 = p^2$.) This is the equation of a conic section, i.e. it could be a parabola ($D = 1$), hyperbola ($D > 1$), or ellipse ($0 \leq D < 1$). If the orbit is bounded, then this equation is an ellipse. Otherwise, it is a parabola or hyperbola, or a line ($p = 0$, $D = 1$), or a point ($p = 0$, $D \neq 1$). And indeed, some comets have these trajectories. □

Remark 3.34. A photon which passes by the sun also has a hyperbolic trajectory, which can be found using the above derivation. Photons have no mass, but they are still affected by the sun's gravity, a prediction of Einstein's theory of general relativity. However, the deflection angle of the photon will be twice that predicted by the above derivation. Also, it turns out that the planets do not travel in perfect ellipses, which is another prediction of general relativity.

4. Functions of Two or Three Variables

In a previous course, you discussed functions $f\colon \mathbb{R} \to \mathbb{R}$. We could then graph these functions in the xy-plane. In this class, we have been discussing functions $r\colon \mathbb{R} \to \mathbb{R}^2$ and $r\colon \mathbb{R} \to \mathbb{R}^3$. We can think of these functions as curves. We now consider functions with larger domains as well. For example, we will first consider functions $f\colon \mathbb{R}^2 \to \mathbb{R}$.

Example 4.1. Let $(x, y) \in \mathbb{R}^2$. Define $f\colon \mathbb{R}^2 \to \mathbb{R}$ by $f(x, y) = x^2 + y^2$. We can plot this function in three-dimensions by considering the value of f to be the z-coordinate. That is, we identify the function $f(x, y) = x^2 + y^2$ with the surface $z = f(x, y) = x^2 + y^2$. We recall that surface $z = x^2 + y^2$ is a paraboloid.

Example 4.2. Suppose we have a domain D in the plane, which we think of as a map of California. Given $(x, y) \in D$, we can think of $f(x, y)$ as the current temperature at the point (x, y). Then $f\colon D \to \mathbb{R}$.

Example 4.3. Suppose we have a domain D in three-dimensional space $\mathbb{R}^3$. For example, we can identify the unit ball $x^2 + y^2 + z^2 \leq 1$ with the earth. Given $(x, y, z) \in D$, we can think of $f(x, y, z)$ as the current temperature at the point (x, y, z) on the earth. Then $f\colon D \to \mathbb{R}$.

More generally, given any positive integer n, we can consider functions $f\colon \mathbb{R}^n \to \mathbb{R}$. We write a general point in $\mathbb{R}^n$ as $(x_1, x_2, \ldots, x_n) \in \mathbb{R}^n$. And at each such point $(x_1, \ldots, x_n) \in \mathbb{R}^n$, we let $f(x_1, \ldots, x_n)$ be some number.

Remark 4.4. Given a function $f\colon \mathbb{R}^2 \to \mathbb{R}$, we said we could identify the function f with the surface $z = f(x, y)$ in $\mathbb{R}^3$. However, visualizing functions $f\colon \mathbb{R}^3 \to \mathbb{R}$ is a bit more difficult. One way to visualize these functions is to think of $f(x, y, z)$ as a color intensity assigned to the point (x, y, z). For example, if $f(x, y, z)$ is a large number, we can think of (x, y, z) as having a light grey color. And if $f(x, y, z)$ is a very small number, we can think of (x, y, z) as having a dark grey color. Drawing a picture like this is perhaps impossible, but the visualization is perhaps more tractable. For example, if $f(x, y, z)$ is the temperature of the earth at the point (x, y, z) with $x^2 + y^2 + z^2 \leq 1$, we could think of $f(x, y, z)$ as a bright red color when the point (x, y, z) is very hot, and we could think of $f(x, y, z)$ as a dark blue color when the point (x, y, z) is very cold.

Example 4.5 (**Drawing Functions of Two Variables**)**.** Suppose we want to draw the function $f(x, y) = x^2 + y^2$. So, we want to draw the surface $z = x^2 + y^2$. It is sometimes convenient to draw the **level curves** of the function f. That is, we can maybe draw the curve $x^2 + y^2 = 1$ in the plane $z = 1$, and then draw the curve $x^2 + y^2 = 4$ in the plane $z = 4$. Doing so results in two circles. On each individual circle, the function f is constant (i.e. the surface has constant height). We can also draw the intersection of $z = x^2 + y^2$ with other hyperplanes. For example, in the plane $x = 0$, we could draw the hyperbola $z = y^2$; in the plane $x = 1$ we could draw the hyperbola $z = 1 + y^2$; in the plane $y = 1$, we could draw the parabola $z = x^2 + 1$, and so on. Drawing the intersection of $z = x^2 + y^2$ with several hyperplanes in this way gives a good sketch of the function.

Example 4.6 (**Level Surfaces in Three Variables**)**.** Consider the function $f(x, y, z) = x^2+y^2+z^2$. As we discussed above, this function is perhaps difficult to draw. However, we can still draw the **level surfaces** of the function. For example, the level surface $x^2 + y^2 + z^2 = 1$ is the unit sphere, and we can then draw this surface. And the level surface $x^2 + y^2 + z^2 = 4$ is the sphere of radius 2 centered at the origin, which can also be drawn, and so on. The function f is constant on any individual level surface.

Example 4.7 (**Domains**)**.** Multivariable domains can be a bit more complicated than single-variable domains. Suppose we have a function $f : D \to \mathbb{R}$ where D is a subset of $\mathbb{R}^2$ defined as the set of (x, y) such that $0 \leq x \leq 1$ and $0 \leq y \leq 1$. We can draw this domain by intersecting the domain corresponding to each condition. For example, the set $0 \leq x \leq 1$ is an infinite rectangle lying between the lines $x = 0$ and $x = 1$. And the set $0 \leq y \leq 1$ is an infinite rectangle lying between the lines $y = 0$ and $y = 1$. The intersection of these two rectangles is the unit square.

For another example, consider the domain D in $\mathbb{R}^3$ defined as the set of all (x, y, z) such that $x^2 + y^2 \leq 4$ and such that $y^2 + z^2 \geq 1$. The first condition gives an infinite solid cylinder, bounded by the cylinder $x^2 + y^2 = 4$. From this solid cylinder, we remove the cylinder $y^2 + z^2 < 1$. In summary, the domain D is an infinite solid cylinder with a hole drilled through it along the x-axis.

4.1. **Limits and Continuity.** In single-variable calculus, we learned a lot about functions by using the notions of continuity and derivatives. We can play a similar game now for multivariable functions. However, the continuity and derivative are now a bit different.

In this section, we only consider functions $f : \mathbb{R}^2 \to \mathbb{R}$. Treating functions with domain $\mathbb{R}^3$ can be done analogously.

Definition 4.8 (**Limit, Informal Definition**)**.** Let $(a, b) \in \mathbb{R}^2$ and let $f : \mathbb{R}^2 \to \mathbb{R}$. Let $L \in \mathbb{R}$. We say that $\lim_{(x,y)\to(a,b)} f(x, y) = L$ if, whenever (x, y) is close to (a, b), we know that $f(x, y)$ is close to L.

Definition 4.9 (**Limit, Formal Definition**)**.** Let $(a, b) \in \mathbb{R}^2$ and let $f : \mathbb{R}^2 \to \mathbb{R}$. Let $L \in \mathbb{R}$. We say that $\lim_{(x,y)\to(a,b)} f(x, y) = L$ if, given any $\varepsilon > 0$, there exists $\delta = \delta(\varepsilon) > 0$ such that, if (x, y) satisfies $0 < \|(x, y) - (a, b)\| < \delta$, then $|f(x, y) - L| < \varepsilon$.

Example 4.10. Let $f(x,y) = x + y$. Let $(a,b) = (1,2)$. Let $L = 3$. Let's verify from the formal definition that $\lim_{(x,y)\to(a,b)} f(x,y) = 3$. Let $\varepsilon > 0$. Then choose $\delta = \varepsilon/2$. Assume $0 < \|(x,y) - (1,2)\| < \delta$. Then $|x-1| < \delta$ and $|y-2| < \delta$. So, $|f(x,y) - L| = |x+y-3| = |(x-1)+(y-2)| \leq |x-1| + |y-2| < \delta + \delta = \varepsilon$. So, $\lim_{(x,y)\to(a,b)} f(x,y) = 3$.

Example 4.11. Let $f(x,y) = x^2/(x^2+y^2)$ We show that the following limit does not exist: $\lim_{(x,y)\to(0,0)} f(x,y)$. Consider $(x_n, y_n) = (0, 1/n)$, $n \geq 1$. Then $(x_n, y_n) \to (0,0)$ as $n \to \infty$, and $f(x_n, y_n) = 0$. However, if $(c_n, d_n) = (1/n, 0)$, $n \geq 1$, then $(c_{,n}, d_n) \to (0,0)$, but $f(c_n, d_n) = 1$. If $\lim_{(x,y)\to(0,0)} f(x,y)$ exists, then $\lim_{n\to\infty} f(x_n, y_n)$ would have to be equal to $\lim_{n\to\infty} f(c_n, d_n)$. So, $\lim_{(x,y)\to(0,0)} f(x,y)$ does not exist.

Proposition 4.12 (Limit Laws). *Let $f, g\colon \mathbb{R}^2 \to \mathbb{R}$. Assume that $\lim_{(x,y)\to(a,b)} f(x,y)$ exists and that $\lim_{(x,y)\to(a,b)} g(x,y)$ exists. Then*

- $\lim_{(x,y)\to(a,b)}[f(x,y) + g(x,y)] = (\lim_{(x,y)\to(a,b)} f(x,y)) + (\lim_{(x,y)\to(a,b)} g(x,y))$.
- *For any constant c,* $\lim_{(x,y)\to(a,b)}[cf(x,y)] = c \cdot \lim_{(x,y)\to(a,b)} f(x,y)$.
- $\lim_{(x,y)\to(a,b)}[f(x,y)g(x,y)] = (\lim_{(x,y)\to(a,b)} f(x,y))(\lim_{(x,y)\to(a,b)} g(x,y))$.
- *If* $\lim_{(x,y)\to(a,b)} g(x,y) \neq 0$, *then*
$$\lim_{(x,y)\to(a,b)} (f(x,y)/g(x,y)) = (\lim_{(x,y)\to(a,b)} f(x,y))/(\lim_{(x,y)\to(a,b)} g(x,y)).$$

Definition 4.13 (Continuity). Let $f\colon \mathbb{R}^2 \to \mathbb{R}$. We say that f is **continuous** at the point (a,b) if
$$f(a,b) = \lim_{(x,y)\to(a,b)} f(x,y).$$
We say that f is continuous on a domain D if f is continuous at (a,b) for every (a,b) in D.

Example 4.14. Let $f\colon \mathbb{R}^3 \to \mathbb{R}$, so that $f(x,y,z) = x^3 + xy^2 + zxy$. Then f is a polynomial, so f is continuous.

Let $g(x,y,z) = x\sin(z) + y$. Then g is continuous, and so is the product of the continuous functions $f(x,y,z) \cdot g(x,y,z)$.

Let $h(x,y,z) = x^4y^4z^2$. Let $F(x,y,z) = f(x,y,z)/g(x,y,z)$. Then F is a quotient of continuous functions, so F is continuous at any point (x,y,z) where $g(x,y,z) \neq 0$.

Theorem 4.15 (A Composition of Continuous Functions is Continuous). *Suppose $f\colon \mathbb{R}^2 \to \mathbb{R}$ is continuous and let $g\colon \mathbb{R} \to \mathbb{R}$ be continuous. Then the composition $g(f(x,y))$ is continuous.*

Example 4.16. The function $\ln(x^2+y^2)$ is continuous for all (x,y) with $(x,y) \neq (0,0)$. However, this function is discontinuous at $(x,y) = (0,0)$.

4.2. **Partial Derivatives.** Now that we have discussed continuity for multivariable functions, we now continue extending calculus to multivariable functions. Our next topic is now differentiability. For a single-variable function $f\colon \mathbb{R} \to \mathbb{R}$, we defined
$$f'(x) = \lim_{h\to 0} \frac{f(x+h) - f(x)}{h}.$$
And we interpreted $f'(x)$ as the rate of change of f, as x increases.

Now, consider a function $f\colon \mathbb{R}^2 \to \mathbb{R}$. Now that there are two variables in the domain, there are many more directions in which we can measure the rate of change of the function. For example, how much does the function change as x increases (but y is fixed)? How much does the function change as y increases (but x is fixed)? How much does the function change as y increases at a rate 3.5 times larger than the rate that x increases? And so on. We will begin by answering the first two questions, and we will save the other one for later.

Definition 4.17 (**Partial Derivatives**). Let $f: \mathbb{R}^2 \to \mathbb{R}$. Let $(a, b) \in \mathbb{R}$. We define the partial derivative of f at the point (a, b) in the x-direction by the following limit (if it exists):

$$\frac{\partial f}{\partial x}(a, b) = f_x(a, b) = \lim_{h \to 0} \frac{f(a + h, b) - f(a, b)}{h}.$$

That is, the partial derivative of f in the x-direction is just the derivative of $f(x, y)$ with respect to x, if we consider y to be fixed.

We similarly define the partial derivative of f at the point (a, b) in the y-direction by

$$\frac{\partial f}{\partial y}(a, b) = f_y(a, b) = \lim_{h \to 0} \frac{f(a, b + h) - f(a, b)}{h}.$$

Example 4.18. Let $f(x, y) = x^2y^3 + y^2 + x$. Then

$$\frac{\partial f}{\partial x} = 2xy^3 + 1, \qquad \frac{\partial f}{\partial y} = 3x^2y^2 + 2y.$$

Example 4.19. Let $f(x, y, z) = x^2 + y^2 + z^2$. Then

$$\frac{\partial f}{\partial x} = 2x, \qquad \frac{\partial f}{\partial y} = 2y \qquad \frac{\partial f}{\partial z} = 2z.$$

Remark 4.20. We can interpret a partial derivative in the same way we always interpret derivatives. For example, $\partial f/\partial x$ is the rate of change of f, as x increases (while other variables are held fixed).

Example 4.21. Since a partial derivative is essentially a one-variable derivative, it follows all of the usual rules such as the chain rule, product rule, quotient rule, etc.

Let $f(x, y) = \frac{x(x^2+1)^2}{y^2+1}$. Then

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

For single variable functions, after defining the first derivative $f'(x)$ of a function $f: \mathbb{R} \to \mathbb{R}$, we then defined $f''(x) = (d/dx)f'(x)$, $f'''(x) = (d/dx)f''(x)$, and so on. We will do something similar now for multivariable functions. However, now that we have more than one variable, there are many more derivatives to consider.

Definition 4.22 (**Iterated Partial Derivatives**). Let $f: \mathbb{R}^2 \to \mathbb{R}$. We define the second order partial derivatives of f in the following way.

$$\frac{\partial^2 f}{\partial x^2} = f_{xx} = \frac{\partial}{\partial x}\frac{\partial f}{\partial x}, \qquad \frac{\partial^2 f}{\partial y^2} = f_{yy} = \frac{\partial}{\partial y}\frac{\partial f}{\partial y}.$$

$$\frac{\partial^2 f}{\partial x \partial y} = f_{xy} = \frac{\partial}{\partial x}\frac{\partial f}{\partial y}, \qquad \frac{\partial^2 f}{\partial y \partial x} = f_{yx} = \frac{\partial}{\partial y}\frac{\partial f}{\partial x}.$$

Example 4.23. Let's compute some iterated partial derivatives of $f(x, y, z) = xyz + x^2$. We have

$$\frac{\partial^2 f}{\partial x^2} = \frac{\partial}{\partial x}\frac{\partial f}{\partial x} = \frac{\partial}{\partial x}(yz + 2x) = 2.$$

$$\frac{\partial^2 f}{\partial y^2} = \frac{\partial}{\partial y}\frac{\partial f}{\partial y} = \frac{\partial}{\partial y}(xz) = 0.$$

$$\frac{\partial^2 f}{\partial y \partial z} = \frac{\partial}{\partial y}\frac{\partial f}{\partial z} = \frac{\partial}{\partial y}(xy) = x, \qquad \frac{\partial^2 f}{\partial z \partial y} = \frac{\partial}{\partial z}\frac{\partial f}{\partial y} = \frac{\partial}{\partial z}(xz) = x$$

In the last example, we found that $f_{yz} = f_{zy}$. That is, the order of the iterated partial derivative does not matter. This observation is in fact generic.

Theorem 4.24 (Clairaut's Theorem). *Let $f\colon \mathbb{R}^2 \to \mathbb{R}$. Suppose f_{xy} and f_{yx} both exist and are continuous. Then $f_{xy} = f_{yx}$.*

4.3. **Differentiability and Tangent Planes.** Suppose $f\colon \mathbb{R} \to \mathbb{R}$. Fix $a \in \mathbb{R}$. Recall that the **linearization** of f at the point a is given by the following function of x.

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

This function is the linearization of f since L is a linear function of x, $L(a) = f(a)$ and $L'(a) = f'(a)$. We also refer to $L(x)$ as the **tangent line** to f at a. We can do something similar for function several variables.

Definition 4.25. Suppose $f\colon \mathbb{R}^2 \to \mathbb{R}$. Fix $(a, b) \in \mathbb{R}$. We define the **linearization** of f at the point (a, b) to be the following function of $(x, y) \in \mathbb{R}^2$.

$$L(x, y) = f(a, b) + (x - a)f_x(a, b) + (y - b)f_y(a, b) = f(a, b) + ((x, y) - (a, b)) \cdot \nabla f(a, b).$$

Here we defined the **gradient** of f at the point (a, b) to be the following vector in $\mathbb{R}^2$.

$$\nabla f(a, b) = (f_x(a, b), f_y(a, b)).$$

If $f\colon \mathbb{R}^3 \to \mathbb{R}$ and if $(a, b, c) \in \mathbb{R}$, we similarly define the **gradient** of f at (a, b, c) to be the vector in $\mathbb{R}^3$.

$$\nabla f(a, b, c) = (f_x(a, b, c), f_y(a, b, c), f_z(a, b, c)).$$

And we similarly define the **linearization** of f at the point (a, b, c) to be the following function of $(x, y, z) \in \mathbb{R}^3$.

$$L(x, y, z) = f(a, b, c) + ((x, y, z) - (a, b, c)) \cdot \nabla f(a, b, c).$$

Example 4.26. We find the linearization of the function $f(x, y) = x^2 + y^2$ at the point $(1, 3)$. The linearization is given by

$$\begin{aligned} L(x, y) &= f(1, 3) + ((x, y) - (1, 3)) \cdot \nabla f(1, 3) \\ &= 10 + ((x, y) - (1, 3)) \cdot (2, 6) = 10 + 2(x - 1) + 6(y - 3). \end{aligned}$$

Remark 4.27. Suppose $f\colon \mathbb{R}^2 \to \mathbb{R}$. The linearization L gives an equation for a plane $z = L(x, y)$. We say that the plane $z = L(x, y)$ is then the **tangent plane** to the surface $z = f(x, y)$ at the point (a, b). The tangent plane at (a, b) is a linear approximation to the surface $z = f(x, y)$ in the same way that a single-variable function $f\colon \mathbb{R} \to \mathbb{R}$ can have a tangent line.

Example 4.28. Continuing the previous example, the tangent plane to the paraboloid $z = x^2 + y^2$ at $(1, 3)$ is given by $z = 10 + 2(x - 1) + 6(y - 3)$.

Example 4.29. Let's find the tangent plane to the cone $z^2 = x^2 + y^2$, $z \geq 0$ at $(3, 4)$. Solving for z, we have $z = \sqrt{x^2 + y^2}$. Setting $f(x, y) = \sqrt{x^2 + y^2}$, we have $\nabla f(x, y) = (x(x^2 + y^2)^{-1/2}, y(x^2 + y^2)^{-1/2})$. So, $\nabla f(3, 4) = (3/5, 4/5)$. So, the tangent plane is given by $z = f(3, 4) + ((x, y) - (3, 4)) \cdot \nabla f(3, 4) = 5 + (x - 3)(3/5) + (y - 4)(4/5)$.

Note that the gradient of f is discontinuous at $(x, y) = (0, 0)$, and there is similarly no clear geometric meaning to the tangent plane at the tip of the cone.

Remark 4.30 (Linear Approximation). Recall that for a single-variable function $f : \mathbb{R} \to \mathbb{R}$, we have the heuristic $f(x) \approx L(x)$ when x is near a. We similarly have a heuristic for functions of multiple variables. For example, if $f : \mathbb{R}^2 \to \mathbb{R}$, then $f(x, y) \approx L(x, y)$ when (x, y) is near the point (a, b). This approximation holds since f and L agree to first order, just as in the single-variable case. Indeed, we have

$$L(a, b) = f(a, b), \qquad \nabla L(a, b) = (L_x(a, b), L_y(a, b)) = (f_x(a, b), f_y(a, b)) = \nabla f(a, b).$$

4.4. **Gradient and Directional Derivative.** Let $f : \mathbb{R} \to \mathbb{R}$. Recall that if the derivative of f is positive, then the function is increasing; if the derivative is negative, then the function is decreasing, and if the derivative is zero, then the function has a critical point. So, the derivative of f tells us how the function f is changing. There is an analogous thing to say for functions of two or more variables. However, we now have many variables, so a single derivative may not necessarily tell us what is going on. That is, we will need to look at a few different derivatives, but then some notion of positivity or negativity is then less straightforward.

Example 4.31. Consider the function $f(x, y) = x^2 - y^2$. If $y = 0$, then $f(x, y) = x^2$. That is, when we restrict the function to the line $y = 0$, we see that $f(x, y)$ has a local minimum at $x = 0$. However, if $x = 0$, then $f(x, y) = -y^2$. That is, when we restrict the function to the line $x = 0$, we see that $f(x, y)$ has a local *maximum* at $y = 0$. This behavior can be understood by looking at both derivatives of f. We have $f_x = 2x$, so that moving x towards zero will always decrease the value of f. And $f_y = -2y$, so moving y towards zero will always increase the value of f. This increasing/decreasing nature of f cannot be understood just by looking at a single derivative of f.

Let $f : \mathbb{R}^2 \to \mathbb{R}$ Recall that we defined the gradient of f at (a, b) by

$$\nabla f(a, b) = (f_x(a, b), f_y(a, b)).$$

Remark 4.32. The gradient vector $\nabla f(a, b)$ points in the direction in which the function f is increasing, at the point (a, b). If $\nabla f(a, b) = (0, 0)$, or if f_x or f_y is undefined, we say that (a, b) is a **critical point** of the function f.

Example 4.33. Let $f(x, y) = x^2 - y^2$ as in the previous example. Then $\nabla f(a, b) = (2a, -2b) = 2(a, -b)$. The vector $(a, -b)$ points in the direction in which f increases, at the point (a, b). If $(a, b) = (0, 0)$, then $\nabla f(a, b) = (0, 0)$, so $(0, 0)$ is a critical point of f. However, note that this point is neither a local maximum nor a local minimum of f. We will discuss critical points later on in the course.

Example 4.34. Let $f(x,y) = x^2 + y^2$. Then $\nabla f(a,b) = (2a, 2b)$. The vector (a,b) points in the direction in which f increases, at the point (a,b). That is, moving away from the origin always increases the value of f. If $(a,b) = (0,0)$, then $\nabla f(a,b) = (0,0)$, so $(0,0)$ is a critical point of f. In this case, the origin $(0,0)$ is the global minimum of the function f.

Proposition 4.35 (Properties of the Gradient). *Let $f, g\colon \mathbb{R}^2 \to \mathbb{R}$ and let $c \in \mathbb{R}$.*

- $\nabla(f+g) = (\nabla f) + (\nabla g)$
- $\nabla(cf) = c \cdot (\nabla f)$.
- $\nabla(fg) = f \cdot (\nabla g) + g \cdot (\nabla f)$. ***(Product Rule)***
- *Let $F\colon \mathbb{R} \to \mathbb{R}$. Then $\nabla F(f(x,y)) = F'(f(x,y)) \cdot \nabla f(x,y)$.* ***(Chain Rule)***

Example 4.36. Let $f(x,y) = x^2 + y^2$, let $g(x,y) = 2x$. Then at the point (a,b), we have $\nabla(f+g) = \nabla f + \nabla g = (2a, 2b) + (2,0) = (2a+2, 2b)$.

Let $h(x,y) = (x^2 + y^2)$ and let $F(t) = t^3$. Then at the point (a,b), we have $\nabla F(h(a,b)) = 3(h(a,b))^2 \nabla h(a,b) = 3(a^2+b^2)^2(2a, 2b)$.

Proposition 4.37 (Chain Rule for Vector-Valued Functions). *Let $f\colon \mathbb{R}^3 \to \mathbb{R}$. Let $r\colon \mathbb{R} \to \mathbb{R}^3$. Then*

$$\frac{d}{dt}(f(r(t))) = (\nabla f(r(t))) \cdot r'(t).$$

Proof Sketch. We use the linear approximation, $r(t+h) \approx r(t) + hr'(t)$. Also, $f((x,y,z)+v) \approx f(x,y,z) + v \cdot \nabla f(x,y,z)$. Substituting one approximation into the other,

$$\frac{d}{dt}f(r(t)) = \lim_{h\to 0}\frac{f(r(t+h)) - f(r(t))}{h} \approx \lim_{h\to 0}\frac{f(r(t) + hr'(t)) - f(r(t))}{h}$$
$$\approx \lim_{h\to 0}\frac{[f(r(t)) + hr'(t)\cdot \nabla f(r(t))] - f(r(t))}{h} = \lim_{h\to 0}\frac{hr'(t)\cdot \nabla f(r(t))}{h} = r'(t)\cdot \nabla f(r(t)).$$

□

Example 4.38. Let $f(x,y) = x^2 + y^2$. Let $s(t) = (t, t^3)$. Then $f(s(t)) = t^2 + t^6$. From the Chain Rule, we have $(d/dt)f(s(t)) = \nabla f(s(t)) \cdot s'(t) = (2t, 2t^3) \cdot (1, 3t^2) = 2t + 6t^5$.

4.5. **Directional Derivatives.** From the Chain Rule for vector-valued functions, we saw that, if $f\colon \mathbb{R}^3 \to \mathbb{R}$ and if $r\colon \mathbb{R} \to \mathbb{R}^3$, then

$$\frac{d}{dt}(f(r(t))) = \lim_{h\to 0}\frac{f(r(t) + hr'(t)) - f(r(t))}{h} = (\nabla f(r(t))) \cdot r'(t).$$

We can similarly allow $r'(t)$ to be any fixed vector $v \in \mathbb{R}^3$, and turn this limit into a definition.

Definition 4.39 (Directional Derivative). Let $f\colon \mathbb{R}^3 \to \mathbb{R}$. Let $(a,b,c) \in \mathbb{R}^3$ and let $v \in \mathbb{R}^3$. We define the derivative $D_v f(a,b,c)$ of f with respect to v at the point (a,b,c) by

$$D_v f(a,b,c) = \lim_{h\to 0}\frac{f((a,b,c)+hv) - f(a,b,c)}{h} = (\nabla f(a,b,c)) \cdot v.$$

If v is a unit vector, we call $D_v f(a,b,c)$ the **directional derivative** of f in the direction v at the point (a,b,c).

Remark 4.40. Let $f\colon \mathbb{R}^2 \to \mathbb{R}$ and let $v \in \mathbb{R}^2$. The directional derivative $D_v f(a,b)$ measures the rate of growth of f at the point (a,b) in the direction v. More specifically, $D_v f(a,b)$

is the slope of the tangent line to the surface $z = f(x, y)$ at the point $(x, y) = (a, b)$ which points in the direction $v = (v_1, v_2)$. To see this, consider the line

$$r(t) = (a, b, f(a, b)) + t(v_1, v_2, v \cdot \nabla f(a, b)).$$

Then this line is contained in the tangent plane to f at (a, b). That is, this line must satisfy $z = L(x, y) = f(a, b) + ((x, y) - (a, b)) \cdot \nabla f(a, b)$. Indeed, the z-component of r is $f(a, b) + tv \cdot \nabla f(a, b)$, while the other components satisfy $x = a + tv_1$ and $y = b + tv_2$, so

$$z = f(a, b) + tv \cdot \nabla f(a, b) = f(a, b) + ((x, y) - (a, b)) \cdot \nabla f(a, b).$$

In conclusion, the line $r(t)$ is tangent to the surface $z = f(x, y)$ at the point (a, b). And the slope of this line is the z-component of $r'(t)$, which is exactly

$$v \cdot \nabla f(a, b) = D_v f(a, b).$$

Example 4.41. Let $f(x, y) = x^2 + y^2$. Let $v = (1, 2)$. Then $D_v f(0, -1) = (\nabla f(0, -1)) \cdot (1, 2) = (0, -2) \cdot (1, 2) = -4$.

Directional derivatives allow us to better understand the information contained in the gradient itself. For example, suppose $v \in \mathbb{R}^2$ is a unit vector and let $(a, b) \in \mathbb{R}^2$. Then, if θ denotes the angle between $\nabla f(a, b)$ and v, we have

$$D_v f = \nabla f(a, b) \cdot v = \|\nabla f(a, b)\| \cos\theta.$$

That is, $D_v f$ is the most positive when $\theta = 0$, and $D_v f$ is most negative when $\theta = \pi$. In the case $\theta = 0$, v points in the same direction as $\nabla f(a, b)$. And in the case that $\theta = \pi$, v points in the opposite direction of $\nabla f(a, b)$.

Finally, in the case $\theta = \pi/2$, $D_v f = 0$. So, when v is perpendicular to $\nabla f(a, b)$, f is neither increasing nor decreasing in the direction v. We summarize these observations below.

Definition 4.42 (Optimization Interpretation of Gradient and Directional Derivatives). Let $f\colon \mathbb{R}^2 \to \mathbb{R}$, let $v \in \mathbb{R}^2$ be a unit vector, and let $(a, b) \in \mathbb{R}^2$. Assume $\nabla f(a, b) \neq (0, 0)$. Then

(i) $\nabla f(a, b)$ points in the direction of greatest increase of f. And $-\nabla f(a, b)$ points in the direction of greatest decrease of f.
(ii) $\nabla f(a, b)$ is orthogonal to the level curve of f at (a, b).

In the case $f\colon \mathbb{R}^3 \to \mathbb{R}$, then $\nabla f(a, b, c)$ is normal to the level surface of f at (a, b, c).

Proof of (ii). Let $r\colon \mathbb{R} \to \mathbb{R}^2$ be a level curve of f. That is, assume that $f(r(t))$ is constant as t varies, and assume that $r(0) = (a, b, c)$. Then using the Chain Rule,

$$0 = \frac{d}{dt} f(r(t)) = \nabla f(r(t)) \cdot r'(t).$$

That is, $r'(t)$ is orthogonal to $\nabla f(r(t))$ for any t (in particular for $t = 0$). And $r'(t)$ is tangent to the level curve $r(t)$. So, $\nabla f(r(t))$ is orthogonal to the level curve. □

Example 4.43 (Tangent Planes for Implicitly Defined Surfaces). Let $f\colon \mathbb{R}^3 \to \mathbb{R}$ and let d be a constant. Suppose we have a surface implicitly defined by $f(x, y, z) = d$. Then we can find both a normal vector and tangent plane to a given point (a, b, c) in the surface. We know that $\nabla f(a, b, c)$ is orthogonal to the surface $f(x, y, z) = d$. So, a normal vector at (a, b, c) is $\nabla f(a, b, c)$, and the tangent plane is given by the equation

$$((x, y, z) - (a, b, c)) \cdot \nabla f(a, b, c) = 0.$$

Remark 4.44. The above formula is almost identical to the formula for the linearization of a function $f\colon \mathbb{R}^3 \to \mathbb{R}$ from Definition 4.25. However, the formula for the tangent plane for the surface $g(x,y,z) = d$ is a bit different than the formula for the tangent plane $z = L(x,y)$ for a function $f\colon \mathbb{R}^2 \to \mathbb{R}$, as defined in Remark 4.27. However, the two formulas can be related in the following scenario. Suppose we can write g in the form $g(x,y,z) = -z + f(x,y)$, and if we consider the surface $g(x,y,z) = -z + f(x,y) = 0$, then the tangent plane for the implicitly defined surface $g(x,y,z) = 0$ becomes

$$\begin{aligned} 0 &= ((x,y,z) - (a,b,c)) \cdot \nabla g(a,b,c) = ((x,y,z) - (a,b,c)) \cdot (f_x(a,b), f_y(a,b), -1) \\ &= ((x,y) - (a,b)) \cdot \nabla f(a,b) + c - z = ((x,y) - (a,b)) \cdot \nabla f(a,b) - z + f(a,b) \end{aligned}$$

In the last line, we used $g(a,b,c) = 0 = -c + f(a,b)$, so $c = f(a,b)$. In summary, the implicitly defined tangent plane for g reduces to the formula for the tangent plane for f from Remark 4.27.

Example 4.45. Consider the sphere $x^2 + y^2 + z^2 = 1$. Let $f(x,y,z) = x^2 + y^2 + z^2$. The tangent plane to the point (a,b,c) in the sphere is given by

$$((x,y,z) - (a,b,c)) \cdot (2a, 2b, 2c) = 0.$$

Simplifying a bit, we have $(x,y,z) \cdot (a,b,c) = a^2 + b^2 + c^2 = 1$. The nice thing here is that we did not have to solve for z and take derivatives of the resulting expression.

Example 4.46. Consider the cone $z^2 = x^2 + y^2$. Let $f(x,y,z) = x^2 + y^2 - z^2$. The tangent plane to the point (a,b,c) in the cone is given by

$$((x,y,z) - (a,b,c)) \cdot (2a, 2b, -2c) = 0.$$

Simplifying a bit, we have $(x,y,z) \cdot (a,b,-c) = a^2 + b^2 - c^2 = 0$. Note that if $(a,b,c) = 0$, this equation no longer becomes a plane, but it becomes all of $\mathbb{R}^3$.

4.6. **Chain Rule.** Recall that if $f\colon \mathbb{R}^3 \to \mathbb{R}$ and if $r\colon \mathbb{R} \to \mathbb{R}^3$, then the Chain Rule says

$$\frac{d}{dt}(f(r(t))) = (\nabla f(r(t))) \cdot r'(t).$$

There is an even more general statement to make.

Theorem 4.47 (Chain Rule). *Let $f\colon \mathbb{R}^n \to \mathbb{R}$. We write $f(x_1, \ldots, x_n)$. Suppose each variable $x_1, \ldots, x_n$ is itself a function of variables $t_1, \ldots, t_m$. Then for any $1 \leq k \leq m$,*

$$\frac{\partial f}{\partial t_k} = \frac{\partial f}{\partial x_1}\frac{\partial x_1}{\partial t_k} + \frac{\partial f}{\partial x_2}\frac{\partial x_2}{\partial t_k} + \cdots + \frac{\partial f}{\partial x_n}\frac{\partial x_n}{\partial t_k}.$$

Alternatively, if we denote $\nabla f = (f_{x_1}, \ldots, f_{x_n})$, $r(t_k) = (x_1(t_1, \ldots, t_m), \ldots, x_n(t_1, \ldots, t_m))$, then we have

$$\frac{\partial f}{\partial t_k} = (\nabla f) \cdot r'.$$

Example 4.48. Let $f(x,y,z) = xy + z$. Suppose $x = s^3$, $y = st$ and $z = t^3$. We would like to compute $\partial f/\partial s$. We let $r(s) = (s^3, st, t^3)$. We then have

$$\begin{aligned} \frac{\partial f}{\partial s} &= \nabla f \cdot r' = (y, x, 1) \cdot (3s^2, t, 0) \\ &= (st, s^3, 1) \cdot (3s^2, t, 0) = 3s^3 t + ts^3 = 4s^3 t. \end{aligned}$$

We can verify this result directly. Note that $f(x,y,z) = xy + z = s^3 st + t^3$. So, $f_s = 4s^3 t$.

Example 4.49. Let $f(x,y,z) = xz + y^2$. Suppose $x = s^2$, $y = s/t$ and $z = t^2$. We would like to compute $\partial f/\partial t$. We let $r(t) = (s^2, s/t, t^2)$. We then have

$$\begin{aligned}\frac{\partial f}{\partial t} &= \nabla f \cdot r' = (z, 2y, x) \cdot (0, -st^{-2}, 2t)\\ &= (t^2, 2s/t, s^2) \cdot (0, -st^{-2}, 2t) = -2s^2t^{-3} + 2ts^2.\end{aligned}$$

We can verify this result directly. Note that $f(x,y,z) = xz + y^2 = s^2t^2 + s^2t^{-2}$. So, $f_t = 2s^2t - 2s^2t^{-3}$.

4.6.1. *Implicit Differentiation.* Recall that in single-variable calculus, if we have an equation of the form $f(x,y) = 0$, then we can treat y as an implicit function of x. That is, we write $f(x, y(x)) = 0$. And by the Chain Rule, we can differentiate this to get

$$f_x(x, y(x)) + f_y(x, y(x))\frac{dy}{dx} = 0.$$

Solving for dy/dx, we get $dy/dx = -f_x/f_y$ (assuming $f_y \neq 0$).

Example 4.50. Suppose $x^2 + y^2 - 1 = 0$, we define $f(x,y) = x^2 + y^2 - 1$, and we have $2x + 2y(dy/dx) = 0$, so that $dy/dx = -x/y$. And indeed, if we solved for y, we would get $y(x) = \pm\sqrt{1 - x^2}$, so that $dy/dx = -(\pm x/\sqrt{1 - x^2}) = -x/y(x)$.

We can do a similar thing with more variables, using our general Chain Rule. For example, suppose we have a function $f(x,y,z) = 0$ and we want to compute $\partial y/\partial x$. That is, we are thinking of y as an implicit function of x, but still treating z and x as independent variables. That is, $\partial z/\partial x = 0$. From the Chain Rule, we get

$$0 = f_x\frac{\partial x}{\partial x} + f_y\frac{\partial y}{\partial x} + f_z\frac{\partial z}{\partial x} = f_x + f_y\frac{\partial y}{\partial x}.$$

Solving for $\partial y/\partial x$, we get the same formula $\partial y/\partial x = -f_x/f_y$ (assuming $f_y \neq 0$).

Example 4.51. Suppose $x^2 + y + z^2x - 3 = 0$. We want to compute $\partial z/\partial x$. From the Chain Rule, we have $2x + 2zx(\partial z/\partial x) + z^2 = 0$. Solving for $\partial z/\partial x$, we get

$$\frac{\partial z}{\partial x} = \frac{-z^2 - 2z}{2zx}$$

5. Optimization

We have now built up enough machinery of differential calculus to begin optimizing functions of multiple variables. We have seen that the gradient of a function points in the direction of greatest increase, and the negative of the gradient of a function points in the direction of greatest decrease. We will exploit this and other properties of the gradient to optimize functions. We first recall optimization for a single variable.

Suppose $f\colon \mathbb{R} \to \mathbb{R}$. Recall that a critical point occurs at x when $f'(x) = 0$ or $f'(x)$ is undefined. If x is a critical point for f, then x may or may not be a local extremum of f. For example, if $f(x) = x^2$, then $f'(x) = 0$ occurs only at $x = 0$. And we know that $x = 0$ is a global minimum for f. That is, $f(x) \geq f(0)$ for all $x \in \mathbb{R}$. However, if $f(x) = x^3$, then $f'(x) = 0$ occurs only at $x = 0$. But $x = 0$ is not a local maximum or a local minimum. That is, no matter how close we look near $x = 0$, there are always points x_1, x_2 near zero such that $f(x_1) > f(0) > f(x_2)$. The issue here is that $f''(x)$ changes sign at $x = 0$, so that $x = 0$ is an inflection point of f. So, the first derivative can be helpful sometimes for optimizing functions, but not always. Lastly, recall that if $f''(x) > 0$, then f has a shape similar to $g(x) = x^2$ (f is concave up), and if $f''(x) < 0$, then f has a shape similar to $g(x) = -x^2$ (f is concave down). So, the second derivative of f can also tell us some things about local extrema of a function. Note that $f(x) = |x|$ has a global minimum at $x = 0$ even though $f'(0)$ is undefined. Note also that $f(x) = x^4$ has a global minimum at $x = 0$, while $f'(0) = f''(0) = 0$, so the value of the second derivative at a single point cannot always identify a local extremum.

Let's recall the second derivative test in one variable.

Example 5.1. Let $f\colon \mathbb{R} \to \mathbb{R}$. We say that c is a local maximum for f if all points x near c satisfy $f(x) \leq f(c)$. We say that c is a local minimum for f if all points x near c satisfy $f(x) \geq f(c)$. We say that c is a local extremum for f if c is either a local maximum or a local minimum of f.

(**Second Derivative Test on** $\mathbb{R}$)

- If $f'(c) = 0$ and $f''(c) > 0$ then c is a local minimum of f.
- If $f'(c) = 0$ and if $f''(c) < 0$, then c is a local maximum of f.
- If $f'(c) = 0$ and if $f''(c) = 0$, then this test is inconclusive. That is, c may or may not be a local extremum of f.

We will come up with an analogous test for functions of more variables. However, the test will be a bit more complicated than before. We begin with some definitions

Definition 5.2 (**Local Extremum**)**.** Let $f\colon \mathbb{R}^2 \to \mathbb{R}$.

- We say that f has a **local maximum** at (a, b) if $f(x, y) \leq f(a, b)$ for all (x, y) near (a, b). (Formally: there is some $t > 0$ such that, if $\|(a, b) - (x, y)\| < t$, then $f(x, y) \leq f(a, b)$.)
- We say that f has a **local minimum** at (a, b) if $f(x, y) \geq f(a, b)$ for all (x, y) near (a, b). (Formally: there is some $t > 0$ such that, if $\|(a, b) - (x, y)\| < t$, then $f(x, y) \leq f(a, b)$.)

We say that f has a **local extremum** at (a, b) if (a, b) is either a local maximum or a local minimum.

Example 5.3. $f(x, y) = x^2 + y^2$ has a local minimum at $(0, 0)$, since $f(x, y) \geq 0 = f(0, 0)$ for all $(x, y) \in \mathbb{R}^2$. Note that $\nabla f(0, 0) = (0, 0)$.

$g(x, y) = -x^2 - y^2$ has a local maximum at $(0, 0)$, since $g(x, y) \leq 0 = g(0, 0)$ for all $(x, y) \in \mathbb{R}^2$. Note that $\nabla g(0, 0) = (0, 0)$.

$h(x, y) = x^2 - y^2$ does not have a local extremum at $(0, 0)$. Still, we have $\nabla h(0, 0) = (0, 0)$.

Definition 5.4 (**Critical Point**)**.** Let $f\colon \mathbb{R}^2 \to \mathbb{R}$ and let $(a, b) \in \mathbb{R}^2$. We say that (a, b) is a **critical point** of f if either:

- $\nabla f(a, b) = (0, 0)$, or
- One component of $\nabla f(a, b)$ is undefined.

Example 5.5. The point $(0, 0)$ is the only critical point of $f(x, y) = x^2 + y^2$, of $g(x, y) = -x^2 - y^2$, and of $h(x, y) = x^2 - y^2$. The point $(0, 0)$ is a critical point of $f(x, y) = 1/x$.

Example 5.6. Let $f(x, y) = |x| + |y|$. Then (a, b) is a critical point of f whenever $a = 0$ or $b = 0$. The global minimum of f occurs at $(0, 0)$. Note that $\nabla f(0, 0)$ is undefined.

Definition 5.7 (**Saddle Point**). Let $f\colon \mathbb{R}^2 \to \mathbb{R}$ and let $(a,b) \in \mathbb{R}^2$. We say that (a,b) is a **saddle point** of f if (a,b) is a critical point of f and if (a,b) is not a local extremum of f. Formally: (a,b) is a saddle point of f if (a,b) is a critical point of f, and for any $t > 0$, there exist points (x_1,y_1) and (x_2,y_2) such that $\|(x_1,y_1)-(a,b)\| < t$, $\|(x_2,y_2)-(a,b)\| < t$, and such that $f(x_1,y_1) > f(a,b) > f(x_2,y_2)$.

Example 5.8. The point $(0,0)$ is a saddle point for $h(x,y) = x^2 - y^2$. We have $\nabla h(0,0) = (0,0)$, and for any $n > 1$, we have $h(1/n,0) > h(0,0) > h(0,1/n)$.

Theorem 5.9. *Let $f\colon \mathbb{R}^2 \to \mathbb{R}$. If f has a local extremum at (a,b), then (a,b) is a critical point of f.*

Proof Sketch. Assume that $\nabla f(a,b)$ exists. Using the linear approximation of f for points (x,y) near (a,b), we have

$$f(x,y) \approx f(a,b) + ((x,y)-(a,b)) \cdot \nabla f(a,b).$$

Let ε be a small number and choose $(x,y) = (a,b) + \varepsilon \nabla f(a,b)$, so that $f(x,y) \approx f(a,b) + \varepsilon \|\nabla f(a,b)\|$. If $\nabla f(a,b) \neq (0,0)$, then we can choose $\varepsilon > 0$ to get $f(x,y) > f(a,b)$. And we can choose $\varepsilon < 0$ so that $f(x,y) < f(a,b)$. Since (a,b) is a local extremum of f, we cannot find points near (a,b) that lie above and below $f(a,b)$. We conclude that $\nabla f(a,b) = (0,0)$, as desired. $\square$

From the previous Theorem, if we want to find the local extrema of a function, it suffices to find the critical points of f. Once we have found a critical point, we would like to identify it as a local maximum, local minimum, or saddle point. For functions of two variables, the following quantity allows such an identification.

Definition 5.10. Let $f\colon \mathbb{R}^2 \to \mathbb{R}$. Let $(a,b) \in \mathbb{R}^2$. Define the **discriminant** $D(a,b)$ of f at (a,b) by

$$D(a,b) = \det \begin{pmatrix} f_{xx}(a,b) & f_{xy}(a,b) \\ f_{yx}(a,b) & f_{yy}(a,b) \end{pmatrix} = f_{xx}(a,b) f_{yy}(a,b) - (f_{xy}(a,b))^2.$$

$D(a,b)$ is sometimes called the **Hessian** of f at (a,b).

Theorem 5.11 (**Second Derivative Test on $\mathbb{R}^2$**). *Let $f\colon \mathbb{R}^2 \to \mathbb{R}$. Let $(a,b) \in \mathbb{R}^2$. Assume that (a,b) is a critical point of f, and that the second derivatives f_{xx}, f_{xy}, f_{yy} exist and are continuous near (a,b). Then*

- *If $D > 0$ and if $f_{xx}(a,b) > 0$, then (a,b) is a local minimum of f.*
- *If $D > 0$ and if $f_{xx}(a,b) < 0$, then (a,b) is a local maximum of f.*
- *If $D < 0$, then (a,b) is a saddle point of f.*
- *If $D = 0$, then no conclusion can be drawn in general.*

Example 5.12. If $f(x,y) = x^2 + y^2$, then $D(0,0) = 4 > 0$. Since $f_{xx}(0,0) = 2 > 0$, we know that $(0,0)$ is a local minimum of f.

If $g(x,y) = -x^2 - y^2$, then $D(0,0) = 4 > 0$. Since $f_{xx}(0,0) = -2 < 0$, we know that $(0,0)$ is a local maximum of g.

If $h(x,y) = x^2 - y^2$, then $D(0,0) = -4 < 0$, so $(0,0)$ is a saddle point of h.

If $p(x,y) = x^4 + y^4$, then $D(0,0) = 0$, but $(0,0)$ is a global minimum of p. If $p(x,y) = -x^4 - y^4$, then $D(0,0) = 0$, but $(0,0)$ is a global maximum of p. If $p(x,y) = x^4 - y^4$, then $D(0,0) = 0$, but $(0,0)$ is saddle point of p.

Proof sketch of Theorem 5.11. For simplicity, assume $(a,b) = (0,0)$. It turns out that there are multivariable versions of Taylor's Theorem. In the present case, we have the following Taylor expansion near $(0,0)$. (Note the resemblance to the linear approximation of f.)

$$f(x,y) \approx f(0,0) + (x,y)\cdot\nabla f(0,0) + x^2 f_{xx}(0,0)/2 + y^2 f_{yy}(0,0)/2 + xy f_{xy}(0,0) + \cdots$$

If $\nabla f(0,0) = 0$, then the first-order terms go away, so that we have

$$f(x,y) \approx f(0,0) + x^2 f_{xx}(0,0)/2 + y^2 f_{yy}(0,0)/2 + xy f_{xy}(0,0) + \cdots$$

That is, the most significant part of the function is the quadratic term

$$q(x,y) = x^2 f_{xx}(0,0)/2 + y^2 f_{yy}(0,0)/2 + xy f_{xy}(0,0).$$

(Note that $q_{xx}(0,0) = f_{xx}(0,0)$, $q_{yy}(0,0) = f_{yy}(0,0)$ and $q_{xy}(0,0) = f_{xy}(0,0)$). To make things easier, we complete the square to get rid of the xy term.

$$q(x,y) = \frac{1}{2} f_{xx}(0,0)\left(x + y\frac{f_{xy}(0,0)}{f_{xx}(0,0)}\right)^2 + \frac{1}{2f_{xx}(0,0)} y^2\left(f_{yy}(0,0)f_{xx}(0,0) - (f_{xy}(0,0))^2\right).$$

If $D > 0$ and $f_{xx}(0,0) > 0$, then $q(x,y) > 0$, so $(0,0)$ is a local minimum. The other cases are handled similarly. □

Finding local extrema is nice, but it would be better to find the maximum or minimum values of a function on a given domain. Doing so requires more than just looking at the critical points of the function.

Definition 5.13. Let D be a domain in $\mathbb{R}^2$, and let $f\colon D \to \mathbb{R}$. The **global maximum** of f on D is the largest value of f on the domain D (if such a value exists). The **global minimum** of f on D is the smallest value of f on the domain D (if such a value exists).

Example 5.14. Let $f(x) = x$ on the domain $0 \leq x \leq 1$, so that $f\colon [0,1] \to \mathbb{R}$. The global maximum of f occurs at $x = 1$, and the global minimum of f occurs at $x = 0$. Note that $f'(x)$ is never zero, so we cannot identify the global maximum and minimum just by examining critical points. We also need to examine the boundary of the domain $0 \leq x \leq 1$, which in this case is exactly $x = 0$ and $x = 1$. Note also that if we consider f to have the domain $-\infty < x < \infty$, then f has no maximum or minimum value.

Example 5.15. Suppose we want to maximize $f(x) = (x^2-1)^2 = x^4 - 2x^2 + 1$, where $-\infty < x < \infty$. We see that $f'(x) = 4x(x^2-1)$, so that $f'(x) = 0$ when $x = 0, 1, -1$. Also, $f''(x) = 12x^2 - 4$, so $f''(0) < 0$. That is, $x = 0$ is a local maximum. However, f has no global maximum on $-\infty < x < \infty$, since $\lim_{x\to\infty} f(x) = \infty$. Also, the other two critical points are local minima. So, even though there is only one local maximum, this does not necessarily imply that a global maximum exists. On the other hand, we can come up with a condition on the domain of the function f such that, if this condition is satisfied, then we can find the global maximum or minimum of f with an algorithm.

Definition 5.16. Let D be a domain in $\mathbb{R}^2$. Let $(a,b) \in \mathbb{R}^2$. We say that (a,b) is a **boundary point** if, given any $t > 0$, the open disk $\{(x,y) \in \mathbb{R}^2\colon \|(x,y) - (a,b)\| < t\}$ contains at least one point in D, and it contains at least one point not in D. A domain D is **closed** if it contains all of its boundary points. A domain D is called **bounded** if there exists some $R > 0$ such that D is contained in the disk $\{(x,y) \in \mathbb{R}^2\colon \|(x,y) - (a,b)\| < R\}$.

Remark 5.17. These definitions can be applied to Euclidean space of any dimension.

Example 5.18. The closed interval $[0,1]$ is closed. Its boundary points are 0 and 1, and both points are contained in $[0,1]$.

The open interval $(0,1)$ is not closed. Its boundary points are 0 and 1, and both points are not contained in $(0,1)$.

Both of the intervals $[0,1]$ and $(0,1)$ are bounded. However, the interval $(0,\infty)$ is not bounded.

As we saw from the above examples, if the domain of a function is not closed, then the global maximum may not exist. Similarly, if the domain of a function is not bounded, then the global maximum may not exist. Fortunately, if the domain of a continuous function is both closed and bounded, then the global maximum does exist. The following theorem is proven is a proof-based calculus class, otherwise known as analysis. At UCLA, this Theorem is proven in Math 131A and in Math 131B.

Theorem 5.19 (**Extreme Value Theorem**). *Let D be a domain in $\mathbb{R}^2$. Assume that D is closed and bounded. Let $f\colon D\to\mathbb{R}$ be continuous. Then*

- *The function f achieves both its maximum and minimum values in D.*
- *The extreme values of f occur either at critical points of f in D, or on the boundary of D.*

We can then turn this Theorem into an algorithm for finding the extreme values of a function

Algorithm 1 (**Finding Extreme Values**). Let D be a domain in $\mathbb{R}^2$. Assume that D is closed and bounded. Let $f\colon D\to\mathbb{R}$ be continuous. To find the maximum and minimum values of f, we perform the following procedure:

(i) Find all critical points of f in D.
(ii) Find the maximum and minimum of f on the boundary points of D.
(iii) Among the points found in parts (i) and (ii), choose those points with the largest and smallest values of f.

The maximum and minimum values of f on D must occur at the points found in Step (iii).

Remark 5.20. Step (ii) requires care when D is a domain in $\mathbb{R}^3$, as we demonstrate below.

Example 5.21. Consider $f(x,y)=x+y$ on the domain D where $0\leq x\leq 1$ and $0\leq y\leq 1$. We have $\nabla f(x,y)=(1,1)$, so f has no critical points in D, and Step (i) is then complete. We now check the boundary of the domain. This boundary consists of four line segments, and we need to check each one separately. The first line segment is when $y=0$ and $0\leq x\leq 1$. In this case $f(x,y)=x$. We know $\partial f/\partial x=1\neq 0$ on this line, so the extrema on this line occur on its boundary, which is $x=0$ and $x=1$. So, we have added the points $(0,0)$ and $(1,0)$ to our list of points in Step (ii). The next line segment is when $y=1$ and $0\leq x\leq 1$. In this case $f(x,y)=x+1$. We know $\partial f/\partial x=1\neq 0$ on this line, so the extrema on this line occur on its boundary, which is $x=0$ and $x=1$. So, we have added the points $(0,1)$ and $(1,1)$ to our list of points in Step (ii). Similarly, we can check the third and fourth line segments ($x=0$, $0\leq y\leq 1$ and $x=1$, $0\leq y\leq 1$), which add the points $(0,0),(0,1)$ and $(1,0),(1,1)$ respectively to our list of points in Step (ii). However, we already added these points to our list, so we did not add any new points.

In conclusion, from Steps (i) and (ii), we found a list of all possible candidate extrema, and this list is: $(0,0),(0,1),(1,0)$ and $(1,1)$. So, it remains to check the values of f at these points. We have $f(0,0) = 0$, $f(1,0) = 1$, $f(0,1) = 1$ and $f(1,1) = 2$. The minimum and maximum of these values are the minimum and maximum of f on D. That is, the minimum value of f on D is 0, and it occurs only at $(0,0)$. And the maximum value of f on D is 2, and it occurs only at $(1,1)$.

Example 5.22. Consider $f(x,y) = x^2 + y^2 + z^2$ on the domain D where $-1 \leq x \leq 1$, $-1 \leq y \leq 1$ and $-1 \leq z \leq 1$. We have $\nabla f(x,y) = (2x, 2y, 2z)$, so f only has a critical point at $(0,0,0)$, and Step (i) is then complete. We now check the boundary of the domain. This boundary consists of six squares, and we need to check each one separately. The first square is $x = -1$, $-1 \leq y \leq 1$ and $-1 \leq z \leq 1$. In this case $f(x,y,z) = 1 + y^2 + z^2$. That is, we are confronted with a two-variable optimization of the function $g(y,z) = 1 + y^2 + z^2$ where $-1 \leq y \leq 1$ and $-1 \leq z \leq 1$. We therefore perform Algorithm 1 on the function g. We see that $\nabla g = (2y, 2z)$, so a critical point only occurs at $(y,z) = (0,0)$. We then need to check the boundary of the square, which consists of four line segments. The first line segment occurs when $y = 1$ and $-1 \leq z \leq 1$. On this line, we have $g(y,z) = 2 + z^2$. We have $\partial g/\partial z = 2z$, so that $z = 0$ is the only critical point of g on the line. We finally add the boundary points of the line segment to our list of points, where $z = 1, -1$.

In summary, so far we have added the following points to our list of potential extreme values: $(0,0,0)$, $(-1,0,0)$, $(-1,1,0)$, $(-1,1,1)$, $(-1,1,-1)$. We still have to examine the three other sides of the square $x = -1$, $-1 \leq y \leq 1$ and $-1 \leq z \leq 1$. And we have to examine the five other squares of the domain D. Once we do this procedure, we will get a list of points consisting of all possible triples of points from the set $\{-1, 0, 1\}$.

It remains to check the values of f at these points. We have $f(0,0,0) = 0$, $f(\pm 1, \pm 1, \pm 1) = 3$, and all other candidate extrema have value 1 or 2. So, the minimum value of f on D is 0, and it occurs only at $(0,0,0)$. And the maximum value of f on D is 3, and it occurs only at the set of eight points $(\pm 1, \pm 1, \pm 1)$.

5.1. **Lagrange Multipliers.** We have now reached the culmination of this chapter and of the course. As we have seen from the previous section, many optimization problems naturally require us to consider constraints. For example, when we tried to maximize the function $f(x,y,z) = x^2 + y^2 + z^2$ over the box $-1 \leq x \leq 1, -1 \leq y \leq 1, -1 \leq z \leq 1$, we were forced to optimize the function f on the boundary of the box. For example, we had to optimize the function f subject to the constraint $x = -1$, $-1 \leq y \leq 1$ and $-1 \leq z \leq 1$. This constraint was relatively easy to deal with, since we could just substitute the value $x = -1$ into the function f, and then optimize over y and z. However, for more general constraints, this substitution procedure **may not work**, or it could make things unnecessarily complicated.

Example 5.23. Consider the function $f(x,y) = x^2$ subject to the constraint $x+y = 1$. There may not be a way to substitute the constraint $x+y = 1$ into the function $f(x,y) = x^2$ to get a single variable function, which can be optimized. However, in this case, this procedure works: you could try to parametrize the line $x + y = 1$ as $r(t) = (t, 1-t)$, $t \in \mathbb{R}$, and then consider $f(x,y) = x^2 = t^2$ which has a critical point only at $t = 0$, which corresponds to the point $(0,1)$. And $(0,1)$ is the global minimum of f on $x + y = 1$. However, there is an automatic procedure which works in any dimension, and it does not require any parametrization.

Theorem 5.24 (**Lagrange Multipliers**). *Let $f \colon \mathbb{R}^2 \to \mathbb{R}$ and let $g \colon \mathbb{R}^2 \to \mathbb{R}$ be functions such that ∇f and ∇g exist and are continuous. Consider the problem of maximizing or minimizing f subject to the constraint $g(x,y) = 0$. If f has a local maximum or local* *minimum at the point (a,b) subject to the constraint $g(x,y) = 0$, and if $\nabla g(x,y) \neq (0,0)$, then there exists a real number λ such that*

$$\nabla f(a,b) = \lambda \nabla g(a,b).$$

Remark 5.25. The condition $\nabla f(a,b) = \lambda \nabla g(a,b)$ says that the vectors ∇f and ∇g are parallel.

Proof Sketch. Suppose $r(t)$ is a parametrization of the constraint curve $g(x,y) = 0$ where $r(0) = (a,b)$. Then $f(r(t))$ is a function of the real variable t which has a local maximum or minimum at $t = 0$. That is, $(d/dt)f(r(t)) = 0$ when $t = 0$. From the Chain Rule, this equation says

$$\nabla f(r(0)) \cdot r'(0) = 0.$$

That is, $r'(0)$ is perpendicular to $\nabla f(a,b) = \nabla f(r(0))$. Since $r'(0)$ is parallel to the curve $g(x,y) = 0$ and $\nabla g(a,b)$ is perpendicular to the curve $g(x,y) = 0$, we see that $r'(0)$ is perpendicular to $\nabla g(a,b)$. And since $r'(0)$ is also perpendicular to $\nabla f(a,b)$, we see that $\nabla f(a,b)$ is parallel to $\nabla g(a,b)$. $\square$

Remark 5.26. This Theorem also holds for functions $f\colon \mathbb{R}^3 \to \mathbb{R}$ and constraints $g\colon \mathbb{R}^3 \to \mathbb{R}$.

Example 5.27. Let's find the extreme values of the function $f(x,y) = x + 2y$ on the ellipse $x^2 + 2y^2 = 1$.

We use the method of Lagrange Multipliers, with $f(x,y) = x+2y$ and $g(x,y) = x^2+2y^2-1$. We need to solve for x, y and λ in the equation

$$\nabla f(x,y) = \lambda \nabla g(x,y).$$

That is, we need to solve for

$$(1,2) = \lambda(2x, 4y).$$

We have the system of three equations in three unknowns

$$\begin{cases} 1 = \lambda(2x) \\ 2 = \lambda(4y) \\ x^2 + 2y^2 = 1. \end{cases}$$

We have $2\lambda(2x) = 2 = \lambda(4y)$. So, if $\lambda \neq 0$, we have $4x = 4y$. That is, $x = y$. Substituting this into the last equation, we have

$$1 = 3x^2$$

Solving for x, we get $x = \pm 1/\sqrt{3}$, so that $y = \pm 1/\sqrt{3}$ as well. So, two candidate points for extreme values are $(1/\sqrt{3}, 1/\sqrt{3})$ and $(-1/\sqrt{3}, -1/\sqrt{3})$.

Note that the case $\lambda = 0$ cannot occur, since it would say that $(1,2) = (0,0)$, which is not true. So, the extreme values can only occur at $(1/\sqrt{3}, 1/\sqrt{3})$ and $(-1/\sqrt{3}, -1/\sqrt{3})$. By inspection, $f(1/\sqrt{3}, 1/\sqrt{3}) = 3/\sqrt{3} = \sqrt{3}$, and $f(-1/\sqrt{3}, -1/\sqrt{3}) = -3/\sqrt{3} = -\sqrt{3}$.

In conclusion, subject to the constraint $x^2 + 2y^2 = 1$, the function f has its global maximum of $\sqrt{3}$ at the point $(1/\sqrt{3}, 1/\sqrt{3})$ and its global minimum of $-\sqrt{3}$ at the point $(-1/\sqrt{3}, -1/\sqrt{3})$. (Note that $\nabla g(x,y) \neq (0,0)$ when $x^2 + 2y^2 = 1$, since if $x = 0$, then $y \neq 0$. So, the Lagrange Multiplier Theorem applies.)

Example 5.28. Let $a, b, c > 0$. Find the dimensions of the box of maximal volume, whose edges are parallel to the coordinate axes, which can be inscribed in the ellipsoid

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

Suppose the box B has vertices at the coordinates $(\pm x, \pm y, \pm z) \in \mathbb{R}^3$, where $x, y, z \geq 0$. Then the area of the box is $8xyz$. Also, since the vertices intersect the ellipsoid, we have $x^2/a^2 + y^2/b^2 + z^2/c^2 = 1$. To find the maximum area box, we therefore maximize the function $f(x, y, z) = xyz$ with the constraint $g(x, y, z) = x^2/a^2 + y^2/b^2 + z^2/c^2 - 1 = 0$, with $x, y, z, \geq 0$. Let

$$D = \{(x, y, z) \in \mathbb{R}^3 \colon x > 0, y > 0, z > 0\}.$$

Then

$$\nabla f = \begin{pmatrix} yz \\ xz \\ xy \end{pmatrix}, \qquad \nabla g = \begin{pmatrix} 2x/a^2 \\ 2y/b^2 \\ 2z/c^2 \end{pmatrix}.$$

(Since $\nabla g \neq (0, 0, 0)$ on D, the Lagrange Multiplier Theorem applies.)

We need to solve the equation $\nabla f = \lambda \nabla g$. This system of equations says

$$yz = \lambda 2x/a^2, \qquad xz = \lambda 2y/b^2, \qquad xy = \lambda 2z/c^2.$$

Since $\nabla f \neq (0, 0, 0)$ on D, we may assume that $\lambda \neq 0$. Substituting the first equation into the second gives $(yza^2/(2\lambda))z = 2\lambda y/b^2$, so $z^2 = 4\lambda^2/(a^2b^2)$. So, $z = 2\,|\lambda|\,/(ab)$. Substituting the second equation into the third gives $(xzb^2/(2\lambda))x = 2\lambda z/c^2$, so $x = 2\,|\lambda|\,/(bc)$. Similarly, $y = 2\,|\lambda|\,/(ac)$. Plugging these equalities for x, y, z into the condition $g(x, y, z) = 0$ shows that $12\lambda^2 = a^2b^2c^2$, so $2\sqrt{3}\,|\lambda| = abc$. So, the only critical point we have found in D is

$$(x, y, z) = (a/\sqrt{3}, b/\sqrt{3}, c/\sqrt{3}).$$

On the boundary of D, $f = 0$. Also, $f(a/\sqrt{3}, b/\sqrt{3}, c/\sqrt{3}) = abc3^{-3/2} > 0$. So, we have found the unique maximum of f. The box of maximal volume with edges parallel to the coordinate axes therefore has dimensions $2a/\sqrt{3}, 2b/\sqrt{3}, 2c/\sqrt{3}$.

Example 5.29. Suppose that we have a probability distribution on the set $\{1, \ldots, n\}$, i.e. we have a set of number $p_1, \ldots, p_n \in [0, 1]$ such that $\sum_{i=1}^n p_i = 1$. A fundamental quantity for a probability distribution is its *entropy*

$$f(p_1, \ldots, p_n) = -\sum_{i=1}^n p_i \log p_i\,.$$

(We extend the function $x \log x$ to 0 by continuity, so that $0 \log 0 = 0$.) The entropy of $p_1, \ldots, p_n$ measures the disorder or lack of information in p. This quantity is important in statistical physics.

We will try to maximize the entropy f subject to the constraint $\sum_{i=1}^n p_i = 1$. That is, we optimize f subject to the constraint $g(p_1, \ldots, p_n) = (\sum_{i=1}^n p_i) - 1 = 0$. Note that

$$\nabla f(p_1, \ldots, p_n) = \begin{pmatrix} -1 - \log p_1 \\ \vdots \\ -1 - \log p_n \end{pmatrix}, \qquad \nabla g(p_1, \ldots, p_n) = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}.$$

(Since $\nabla g \neq (0, \ldots, 0)$, the Lagrange Multiplier Theorem applies.)

We need to solve the equation $\nabla f(p_1,\ldots,p_n)=\lambda\nabla g(p_1,\ldots,p_n)$. This system of equations says

$$-1-\log p_1=\lambda,\ldots,-1-\log p_n=\lambda.$$

That is, $p_1=p_2=\cdots=p_n$. Since $g(p_1,\ldots,p_n)=0=(\sum_{i=1}^n p_i)-1=np_1-1$, we conclude that $p_1=1/n$, so that $p_1=\cdots=p_n=1/n$. That is, we have found only one candidate extreme value of f subject to the constraint $g(p_1,\ldots,p_n)=0$.

In order to optimize f, we now need to check the boundary of the domain. The boundary consists of $p_1,\ldots,p_n$ such that $\sum_{i=1}^n p_i=1$, and there is some p_j which is equal to 1 or 0, so that $p_j\log p_j=0$. So, when we are on the boundary of the domain, we have $n-1$ nonnegative numbers $p_1,\ldots,p_{j-1},p_{j+1},\ldots,p_n$ that add to 1. That is, we can exactly repeat the above calculation with $n-1$ replaced by n. We then see that the only critical point we find occurs when the $n-1$ numbers are equal to $1/(n-1)$. So, our next candidate extreme value of f occurs at such a point. We now need to check the boundary of this region, and so on. Iterating, we see that the only candidate extreme values occur when some of the numbers are zero, and the rest are equal to each other.

So, our candidate extreme values occur when $(p_1,\ldots,p_n)=(1/n,\ldots,1/n)$, $(p_1,\ldots,p_n)=(0,1/(n-1),\ldots,1/(n-1))$, $(p_1,\ldots,p_n)=(0,0,1/(n-2),\ldots,1/(n-2))$, and so on, along with any permutation of these points. By inspection, we have

$$f(1/n,\ldots,1/n)=\sum_{i=1}^n -(1/n)\log(1/n)=(n/n)\log n=\log n.$$

$$f(0,1/(n-1),\ldots,1/(n-1))=\sum_{i=1}^{n-1} -\frac{1}{n-1}\log\big(\frac{1}{n-1}\big)=\frac{n-1}{n-1}\log(n-1)=\log(n-1).$$

And so on. We therefore see that the maximum entropy occurs at the point $(1/n,\ldots,1/n)$. (And the minimum entropy occurs at the point $(0,\ldots,0,1)$.)

6. Appendix: Notation

$\mathbb{R}$ denotes the set of real numbers

$\in$ means "is an element of." For example, $2\in\mathbb{R}$ is read as "2 is an element of $\mathbb{R}$."

$\mathbb{R}^2=\{(x_1,x_2)\colon x_1\in\mathbb{R}\,\text{and}\,x_2\in\mathbb{R}\}$

$\mathbb{R}^3=\{(x_1,x_2,x_3)\colon x_1\in\mathbb{R}\,\text{and}\,x_2\in\mathbb{R}\,\text{and}\,x_3\in\mathbb{R}\}$

$f\colon A\to B$ means f is a function with domain A and range B. For example,

$f\colon\mathbb{R}^2\to\mathbb{R}$ means that f is a function with domain $\mathbb{R}^2$ and range $\mathbb{R}$

Let $v=(x_1,y_1,z_1)$ and let $w=(x_2,y_2,z_2)$ be vectors in Euclidean space $\mathbb{R}^3$.

$$
\begin{aligned}
v \cdot w &= x_1x_2 + y_1y_2 + z_1z_2 \text{ , the dot product of } v \text{ and } w \\
\|v\| &= \sqrt{x_1^2 + y_1^2 + z_1^2} \text{ , the length of the vector } v \\
v \times w &= (y_1z_2 - z_1y_2,\, z_1x_2 - x_1z_2,\, x_1y_2 - y_1x_2) \text{ , the cross product of } v \text{ and } w
\end{aligned}
$$

Let $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ be a 2×2 matrix. We define

$$\det(A) = |A| = ad - bc.$$

Let $A = \begin{pmatrix} a & b & c \\ d & e & f \\ g & h & i \end{pmatrix}$ be a 3×3 matrix. We define

$$\det(A) = |A| = a(ei - fh) + b(fg - di) + c(dh - eg).$$

Let $f\colon \mathbb{R}^2 \to \mathbb{R}$. Let $(a, b) \in \mathbb{R}^2$. Let $v \in \mathbb{R}^2$ be a vector.

$$
\begin{aligned}
\frac{\partial f}{\partial x}(a,b) &= f_x(a,b), \text{ denotes the partial derivative of } f \text{ in the } x\text{-direction} \\
\frac{\partial f}{\partial y}(a,b) &= f_y(a,b), \text{ denotes the partial derivative of } f \text{ in the } y\text{-direction} \\
\nabla f(a,b) &= (f_x(a,b), f_y(a,b)), \text{ denotes the gradient vector of } f \text{ at } (a,b) \\
D_v f(a,b) &= \nabla f(a,b) \cdot v, \text{ denotes the derivative of } f \text{ with respect to the direction } v \\
D(a,b) &= \det \begin{pmatrix} f_{xx}(a,b) & f_{xy}(a,b) \\ f_{yx}(a,b) & f_{yy}(a,b) \end{pmatrix} = f_{xx}(a,b) f_{yy}(a,b) - (f_{xy}(a,b))^2 \\
&\qquad \text{denotes the discriminant, or Hessian, of } f \text{ at } (a,b)
\end{aligned}
$$

UCLA Department of Mathematics, Los Angeles, CA 90095-1555
E-mail address: heilman@math.ucla.edu