Math 291-2: MENU Linear Algebra and Multivariable Calculus
Northwestern University, Winter 2022

Aaron Peterson

Last updated: February 28, 2022

Contents

Lecture 1: The Dot Product

Learning Objectives:

- Establish and exploit the fundamental algebraic properties of the dot product.
- Explore the relationship between the dot product and transposes.

Welcome to MATH 291-2! Last quarter we explored the basic objects of linear algebra (vectors and matrices, systems of linear equations, and general vector spaces and linear transformations) and the rich connections between them. This quarter we will complete our study of linear algebra by investigating algebraic constructions (the dot product and determinant) that encode geometric information (angles and volume, respectively). This will culminate in the notion of diagonalization, which allows us to study certain linear transformations by choosing a basis with respect to which the transformation becomes particularly easy to understand. Around the middle of the quarter we will end out study of linear algebra and begin our study of multivariable calculus. Armed with a rich set of ideas from linear algebra, though, we will see that in many of the important ideas of calculus (especially the ideas surrounding differentiation and integration) are merely applications of linear algebra in the context of limits. Without further ado, let's begin!

Notation Change

In MATH 291-1, to be consistent with our linear algebra textbook, we used the arrow notation $\vec{v}$ denote a vector in $\mathbb{K}^n$. Our new textbook (which we will use for multivariable calculus) uses bold notation $\boldsymbol{v}$ to denote vectors. I will use the arrow notation in all printed materials for the rest of the year.

The Dot Product

We begin this quarter by introducing an algebraic operation on $\mathbb{R}^n$ that encapsulates the elementary geometric properties of *length* and *angle*.

Definition 1. Let $\vec{x}, \vec{y} \in \mathbb{R}^n$. The **dot product** of $\vec{x}$ and $\vec{y}$ is defined by

$$\vec{x} \cdot \vec{y} = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} \cdot \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \stackrel{def}{=} \sum_{k=1}^{n} x_k y_k = x_1 y_1 + \cdots + x_n y_n.$$

Remark 1. Because the dot product is defined for vectors in $\mathbb{R}^n$, we will sometimes call it the **real** dot product. On your homework you will investigate a dot product for vectors in $\mathbb{C}^n$ (called the **complex** or **Hermitian** dot product). As with many of the results last quarter, the algebraic properties and results of the real and complex dot products will be very similar, but there is geometric meaning for the real dot product that is not exactly shared by the complex dot product.

Summation Notation

This quarter it will be crucial to have a convenient and compact way to notate sums that follow patterns. To this end we will use **summation notation** (also sometimes called **sigma notation**). In particular, suppose that for each $k = 1, \ldots, n$ we have $a_k \in \mathbb{K}$. Then we define

$$\sum_{k=1}^{n} a_k \stackrel{def}{=} a_1 + a_2 + \cdots + a_n.$$

Here k is an **index**: an additional variable used to keep track of the terms in the sum. The index has no meaning outside of the summation notation.

The notation $\sum_{k=1}^{n} a_k$ can be read in words as "the sum of a_k for $k = 1$ to $k = n$". For example, we would have

$$\sum_{k=1}^{17} k^2 = 1^2 + 2^2 + \cdots + (16)^2 + (17)^2$$

and

$$\begin{aligned}
&\sum_{j=0}^{4} \frac{(-1)^j}{j!}(x-a)^j \\
&\quad = \frac{(-1)^0}{0!}(x-a)^0 + \frac{(-1)^1}{1!}(x-a)^1 + \frac{(-1)^2}{2!}(x-a)^2 + \frac{(-1)^3}{3!}(x-a)^3 + \frac{(-1)^4}{4!}(x-a)^4 \\
&\quad = 1 - (x-a) + \frac{1}{2}(x-a)^2 - \frac{1}{6}(x-a)^3 + \frac{1}{24}(x-a)^4.
\end{aligned}$$

As the second example indicates, the exact symbol we use for the index variable is not important, nor must the index always start counting at 1. The **summands** (i.e. the numbers being added together) can contain variables other than the index variable, and the index variable is only useful for keeping track of the pattern that the summands follow.
Summation notation is linear, in the sense that if $a_1, \ldots, a_n, b_1, \ldots, b_n \in \mathbb{K}$ and $\lambda \in \mathbb{K}$, then

$$\sum_{k=1}^{n} (\lambda a_k) = \lambda \sum_{k=1}^{n} a_k \quad \text{and} \quad \sum_{k=1}^{n} (a_k + b_k) = \sum_{k=1}^{n} a_k + \sum_{k=1}^{n} b_k.$$

The first property follows from the distributativity of multiplcation over addition:

$$\sum_{k=1}^{n} (\lambda a_k) = \lambda a_1 + \cdots + \lambda a_n = \lambda(a_1 + \cdots + a_n) = \lambda \sum_{k=1}^{n} a_k.$$

The second property follows from both associativity and commutativity of addition:

$$\sum_{k=1}^{n} (a_k + b_k) = (a_1 + b_1) + (a_2 + b_2) + \cdots + (a_n + b_n) = (a_1 + \cdots + a_n) + (b_1 + \cdots + b_n) = \sum_{k=1}^{n} a_k + \sum_{k=1}^{n} b_k.$$

Note that the proofs of these two results require induction arguments (which I leave to you!).

We will shortly see exactly how the dot product is linked to geometry, but first we summarize its basic properties.

Proposition 1 (Properties of $\cdot$)**.** The (real) dot product satisfies the following properties.

(i) (Positive Definite) For every $\vec{x} \in \mathbb{R}^n$, $\vec{x} \cdot \vec{x} \geq 0$ with $\vec{x} \cdot \vec{x} = 0$ if, and only if, $\vec{x} = \vec{0}$.

(ii) (Distributivity Over Addition) For every $\vec{x}, \vec{y}, \vec{z} \in \mathbb{R}^n$,

$$(\vec{x} + \vec{y}) \cdot \vec{z} = \vec{x} \cdot \vec{z} + \vec{y} \cdot \vec{z} \quad \text{and} \quad \vec{x} \cdot (\vec{y} + \vec{z}) = \vec{x} \cdot \vec{y} + \vec{x} \cdot \vec{z},$$

(iii) (Distributivity Over Scalar Multiplication) For every $\vec{x}, \vec{y} \in \mathbb{R}^n$ and $\lambda \in \mathbb{R}$,

$$(\lambda \vec{x}) \cdot \vec{y} = \lambda(\vec{x} \cdot \vec{y}) = \vec{x} \cdot (\lambda \vec{y}).$$

(iv) (Commutative) For every $\vec{x}, \vec{y} \in \mathbb{R}^n$, $\vec{x} \cdot \vec{y} = \vec{y} \cdot \vec{x}$.

Remark 2. Property (iii) is sometimes also called **homogeneity**, and propert (iv) is sometimes called **symmetry**. These terms are often used when discussing **inner products**, which generalize the dot product to more general vector spaces.

Proof. Let $\vec{x}, \vec{y}, \vec{z} \in \mathbb{R}^n$ and $\lambda \in \mathbb{R}$.

We start with (i). Because $x_k^2 \geq 0$ for each $1 \leq k \leq n$, $\vec{x} \cdot \vec{x} = x_1^2 + \cdots + x_n^2 \geq 0$. Suppose $\vec{x} \cdot \vec{x} = 0$. Let $1 \leq k \leq n$. Then

$$0 \leq x_k^2 \leq x_1^2 + \cdots + x_n^2 = \vec{x} \cdot \vec{x} = 0,$$

so that $x_k^2 = 0$ and therefore $x_k = 0$. It follows that $\vec{x} = \vec{0}$. On the other hand, if $\vec{x} = \vec{0}$ then $\vec{x} \cdot \vec{x} = 0^2 + \cdots + 0^2 = 0$.

For (iv), the commutativity of real number multiplication gives

$$\vec{x} \cdot \vec{y} = \sum_{k=1}^{n} x_k y_k = \sum_{k=1}^{n} y_k x_k = \vec{y} \cdot \vec{x}.$$

For (ii) and (iii), first note that by the distributivity of real number multiplication over addition and the associativity of real number multiplication,

$$(\vec{x} + \vec{y}) \cdot \vec{z} = \sum_{k=1}^{n} (x_k + y_k) z_k = \sum_{k=1}^{n} (x_k z_k + y_k z_k) = \sum_{k=1}^{n} x_k z_k + \sum_{k=1}^{n} y_k z_k = \vec{x} \cdot \vec{z} + \vec{y} \cdot \vec{z},$$

and

$$(\lambda \vec{x}) \cdot \vec{z} = \sum_{k=1}^{n} (\lambda x_k) z_k = \lambda \sum_{k=1}^{n} x_k z_k = \lambda(\vec{x} \cdot \vec{z}).$$

But then we can apply these results along with (iv) to get

$$\vec{x} \cdot (\vec{y} + \vec{z}) = (\vec{y} + \vec{z}) \cdot \vec{x} = \vec{y} \cdot \vec{x} + \vec{z} \cdot \vec{x} = \vec{x} \cdot \vec{y} + \vec{x} \cdot \vec{z}$$

and

$$\vec{x} \cdot (\lambda \vec{y}) = (\lambda \vec{y}) \cdot \vec{x} = \lambda(\vec{y} \cdot \vec{x}) = \lambda(\vec{x} \cdot \vec{y}),$$

completing the proof. □

Remark 3. There is another way to understand the dot product. The first formulas in properties (ii) and (iii) indicate that for fixed $\vec{y} \in \mathbb{R}^n$, the map $T : \mathbb{R}^n \to \mathbb{R}$, $T(\vec{x}) \stackrel{def}{=} \vec{x} \cdot \vec{y}$ is a linear transformation. Because fixing the second input ($\vec{y}$) of the dot product results in a map that is linear in the first input, the dot product is sometimes called **linear in the first argument**[1]. The real dot product is also **linear in the second argument**, in the sense that fixing $\vec{x} \in \mathbb{R}^n$ results in $T : \mathbb{R}^n \to \mathbb{R}$, $T(\vec{y}) \stackrel{def}{=} \vec{x} \cdot \vec{y}$ being a linear transformation. In this sense, the real dot product (as a function from $\mathbb{R}^n \times \mathbb{R}^n$ to $\mathbb{R}$ is **multilinear**. We will not need this terminology now, but we will see it again when we discuss determinants.

Remark 4. The standard (**Hermitian**) dot product $\cdot$ on $\mathbb{C}^n$ is defined by

$$\vec{z} \cdot \vec{w} \stackrel{def}{=} \sum_{k=1}^{n} z_k \overline{w_k} = z_1 \overline{w_1} + \cdots + z_n \overline{w_n}.$$

The properties of the Hermitian dot product are very similar to those of the real dot product, except that we must replace symmetry with conjugate symmetry $\vec{w} \cdot \vec{z} = \overline{\vec{z} \cdot \vec{w}}$, which further forces the complex dot product to be not linear in the second argument, but *anti-linear*. You will explore this in your homework.

The Dot Product and Length

The relationship between the dot product and length is immediate, since for each $\vec{x} \in \mathbb{R}^n$ we have

$$\|\vec{x}\| = \sqrt{x_1^2 + \cdots + x_n^2} = \sqrt{\vec{x} \cdot \vec{x}}.$$

Remark 5. Because $\vec{x} + (\vec{y} - \vec{x}) = \vec{y}$, $\vec{y} - \vec{x}$ is the vector that starts at the endpoint of $\vec{x}$ and ends at the endpoint of $\vec{y}$. Therefore the length of this vector, $\|\vec{y} - \vec{x}\|$ can be interpreted as the **distance** from (the endpoint of) $\vec{x}$ to (the endpoint of) $\vec{y}$.

Remark 6. As you saw last quarter, for each $\vec{x} \in \mathbb{R}^n$ and $\lambda \in \mathbb{R}$ we have

$$\|\lambda\vec{x}\| = \sqrt{(\lambda\vec{x}) \cdot (\lambda\vec{x})} = \sqrt{\lambda^2(\vec{x} \cdot \vec{x})} = |\lambda|\sqrt{\vec{x} \cdot \vec{x}} = |\lambda|\|\vec{x}\|,$$

so that multiplying a vector $\vec{x}$ by a scalar λ multiplies that length of $\vec{x}$ by $|\lambda|$.

Remark 7. Let $\vec{a}, \vec{b} \in \mathbb{R}^n$ and $c, d \in \mathbb{R}$. In computations, it will be helpful to note that

$$\|c\vec{a} + d\vec{b}\|^2 = c^2\|\vec{a}\|^2 + 2cd(\vec{a} \cdot \vec{b}) + d^2\|\vec{b}\|^2,$$

which is proved by expanding the left-hand-side using the commutativity and distributivity of the dot product:

$$\begin{aligned}
\|c\vec{a} + d\vec{b}\|^2 &= (c\vec{a} + d\vec{b}) \cdot (c\vec{a} + d\vec{b}) \\
&= (c\vec{a}) \cdot (c\vec{a}) + (c\vec{a}) \cdot (d\vec{b}) + (d\vec{b}) \cdot (c\vec{a}) + (d\vec{b}) \cdot (d\vec{b}) \\
&= c^2(\vec{a} \cdot \vec{a}) + cd(\vec{a} \cdot \vec{b}) + dc(\vec{b} \cdot \vec{a}) + d^2(\vec{b} \cdot \vec{b}) \\
&= c^2\|\vec{a}\|^2 + 2cd(\vec{a} \cdot \vec{b}) + d^2\|\vec{b}\|^2.
\end{aligned}$$

In the next couple days we will investigate more sophisticated geometric properties of the norm, but this will require first exploring the dot product.

[1] "Argument" is another name for the input of a function. Since the dot product has two inputs, we call these the two arguments of the function.

The Dot Product and Cancellation

Remark 8. The dot product does not enjoy a "cancellation" property as simple as the one enjoyed by real number multiplication, since (for example) $\left[\begin{smallmatrix}1\\0\end{smallmatrix}\right] \cdot \left[\begin{smallmatrix}0\\1\end{smallmatrix}\right] = 0$ and $\left[\begin{smallmatrix}0\\0\end{smallmatrix}\right] \cdot \left[\begin{smallmatrix}0\\1\end{smallmatrix}\right]$ even though $\left[\begin{smallmatrix}1\\0\end{smallmatrix}\right] \neq \left[\begin{smallmatrix}0\\0\end{smallmatrix}\right]$. The following result captures the appropriate cancellation property for the dot product.

Proposition 2 (Cancellation Property for $\cdot$)**.** Let $\vec{x}, \vec{y} \in \mathbb{R}^n$. Then the following are equivalent.

(i) $\vec{x} = \vec{y}$.

(ii) For every $\vec{z} \in \mathbb{R}^n$, $\vec{x} \cdot \vec{z} = \vec{y} \cdot \vec{z}$.

(iii) For every $\vec{z} \in \mathbb{R}^n$, $\vec{z} \cdot \vec{x} = \vec{z} \cdot \vec{y}$.

Proof. The implication (i)⇒(ii) is immediate. The implication (ii)⇒(iii) follows from the observation that if (ii) holds and if $\vec{z} \in \mathbb{R}^n$, then $\vec{z} \cdot \vec{x} = \vec{x} \cdot \vec{z} = \vec{y} \cdot \vec{z} = \vec{z} \cdot \vec{y}$. Suppose that (iii) holds. Taking $\vec{z} \stackrel{def}{=} \vec{x} - \vec{y}$, the distributive property of $\cdot$ gives

$$0 = \vec{z} \cdot \vec{x} - \vec{z} \cdot \vec{y} = \vec{z} \cdot (\vec{x} - \vec{y}) = (\vec{x} - \vec{y}) \cdot (\vec{x} - \vec{y}),$$

so the fact that $\cdot$ is positive-definite implies that $\vec{x} - \vec{y} = \vec{0}$, and therefore $\vec{x} = \vec{y}$. □

The Dot Product and Transposes

It is particularly fruitful to interpret the dot product in terms of matrix transposition and matrix products.

Remark 9. As a first pass, we note that if $\vec{x}, \vec{y} \in \mathbb{R}^n$, then considering $\vec{x}, \vec{y}$ to be in $M_{n\times 1}(\mathbb{R})$ gives

$$\vec{x} \cdot \vec{y} = x_1 y_1 + \cdots + x_n y_n = \begin{bmatrix} x_1 & \cdots & x_n \end{bmatrix} \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \vec{x}^T \vec{y}.$$

The connection between the dot product and transpose goes much deeper than this. Indeed, the importance of transposes is largely due to the relationship between the dot product and linear transformations. To explain this connection, we first recall a lemma (proved on your homework last quarter).

Lemma 1. Let $A \in M_{m\times n}(\mathbb{K})$ and $B \in M_{n\times p}(\mathbb{K})$. Then $(AB)^T = B^T A^T$.

Proof. For each $1 \leq j \leq m$, let $\vec{a}_j \stackrel{def}{=} \begin{bmatrix} a_{j,1} & a_{j,2} & \cdots & a_{j,n} \end{bmatrix} \in M_{1\times n}(\mathbb{R})$ denote the matrix that forms that j-th row of A. Similarly, for $1 \leq k \leq p$ denote by $\vec{b}_k \in M_{n\times 1}(\mathbb{R})$ the k-th column of B.

Let $1 \leq j \leq m$ and $1 \leq k \leq p$. Then

$$(kj - th \text{ entry of } (AB)^T) = (jk - th \text{ entry of } AB) = \vec{a}_j \vec{b}_k = \vec{a}_j^T \cdot \vec{b}_k,$$

while

$$(kj - th \text{ entry of } B^T A^T) = \vec{b}_k^T \vec{a}_j^T = \vec{b}_k \cdot \vec{a}_j^T = \vec{a}_j^T \cdot \vec{b}_k.$$

We conclude that $(AB)^T = B^T A^T$. □

Theorem 1 (Transposes and $\cdot$)**.** Let $A \in M_{m\times n}(\mathbb{R})$. Then $A^T \in M_{n\times m}(\mathbb{R})$ is the unique matrix that satisfies

$$(A\vec{x}) \cdot \vec{y} = \vec{x} \cdot (A^T\vec{y}) \quad \text{for every } \vec{x} \in \mathbb{R}^n \text{ and } \vec{y} \in \mathbb{R}^m. \tag{1}$$

Proof. Let $\vec{x} \in \mathbb{R}^n$ and $\vec{y} \in \mathbb{R}^m$. Then the lemma shows that

$$(A\vec{x}) \cdot \vec{y} = (A\vec{x})^T\vec{y} = (\vec{x}^T A^T)\vec{y} = \vec{x}^T(A^T\vec{y}) = \vec{x} \cdot (A^T\vec{y}).$$

For uniqueness, suppose that $B \in M_{n\times m}(\mathbb{R})$ also satisfies (1). Then for every $\vec{y} \in \mathbb{R}^m$, for every $\vec{x} \in \mathbb{R}^n$ we have

$$\vec{x} \cdot (B\vec{y}) = (A\vec{x}) \cdot \vec{y} = \vec{x} \cdot (A^T\vec{y}).$$

The Cancellation Property for $\cdot$ therefore shows that $B\vec{y} = A^T\vec{y}$. Because $B\vec{y} = A^T\vec{y}$ for every $\vec{y} \in \mathbb{R}^m$, $B = A^T$. □

Lecture 2: Orthogonality

Learning Objectives:

- Relate the dot product to notions of length and angle.
- Link orthogonality to the notion of projection.

The most important notion captured by the dot product is orthogonality.

Definition 2. Let $\vec{x}, \vec{y} \in \mathbb{R}^n$. We say $\vec{x}$ and $\vec{y}$ are **orthogonal** if $\vec{x} \cdot \vec{y} = 0$.

The name "orthogonal" is intended to generalize the notion of "perpendicular." To illustrate this, This is made clear in the following proposition, which is a vector version of the Pythagorean Theorem.

Proposition 3 (Pythagorean Theorem)**.** Let $\vec{x}, \vec{y} \in \mathbb{R}^n$. Then the following are equivalent.

(i) $\vec{x} \cdot \vec{y} = 0$

(ii) $\|\vec{x}\|^2 + \|\vec{y}\|^2 = \|\vec{x} + \vec{y}\|^2$

In the case where $n = 2$ and $\vec{x}, \vec{y} \neq \vec{0}$, these conditions are also equivalent to saying that $\vec{x}$ and $\vec{y}$ are perpendicular.

Proof. The equivalence of (i) and (ii) follows immediately from the equation

$$\begin{aligned} \|\vec{x} + \vec{y}\|^2 - \|\vec{x}\|^2 - \|\vec{y}\|^2 &= (\vec{x} + \vec{y}) \cdot (\vec{x} + \vec{y}) - \|\vec{x}\|^2 - \|\vec{y}\|^2 \\ &= \vec{x} \cdot \vec{x} + \vec{x} \cdot \vec{y} + \vec{y} \cdot \vec{x} + \vec{y} \cdot \vec{y} - \|\vec{x}\|^2 - \|\vec{y}\|^2 \\ &= 2(\vec{x} \cdot \vec{y}). \end{aligned}$$

You will prove the final claim on your homework. □

Remark 10. In your homework, you will finish linking the dot product to angles by showing that (at least for nonzero vectors in $\mathbb{R}^2$), if $\theta \in [0, \pi]$ is the angle formed by $\vec{x}$ and $\vec{y}$ then

$$\vec{x} \cdot \vec{y} = \|\vec{x}\| \|\vec{y}\| \cos(\theta), \quad \text{or rather} \quad \theta = \arccos\Big(\frac{\vec{x} \cdot \vec{y}}{\|\vec{x}\| \|\vec{y}\|}\Big).$$

We will prove that when $n \geq 3$ and $\vec{x}, \vec{y} \in \mathbb{R}^n$ are nonzero, then we still have

$$-1 \leq \frac{\vec{x} \cdot \vec{y}}{\|\vec{x}\| \|\vec{y}\|} \leq 1,$$

and therefore $\theta \stackrel{def}{=} \arccos\left(\frac{\vec{x} \cdot \vec{y}}{\|\vec{x}\| \|\vec{y}\|}\right)$ can actually be taken as the *definition* of the angle between $\vec{x}$ and $\vec{y}$. We will revisit this when we have the machinery to do so.

Remark 11. The previous proposition is called the Pythagorean Theorem because, interpreting $\vec{x}$, $\vec{y}$, and $\vec{x}+\vec{y}$ as sides of a triangle (with $\vec{x}+\vec{y}$ forming the hypotenuse), then the theorem says that this triangle is right if, and only if, the sum of the squares of the lengths of the legs of the triangle is equal to the square of the length of the hypotenuse.

The dot product also allows us to analyze the projection of a vector $\vec{x}$ onto a nonzero vector $\vec{y}$, which can be thought of as the vector in $\text{span}(\vec{y})$ that is closest to $\vec{x}$.

Definition 3. Let $\vec{x}, \vec{y} \in \mathbb{R}^n$ with $\vec{y} \neq \vec{0}$. Define the **orthogonal projection** of $\vec{x}$ onto $\vec{y}$, $\text{proj}_{\vec{y}}(\vec{x})$, by

$$\text{proj}_{\vec{y}}(\vec{x}) \stackrel{def}{=} \frac{(\vec{x}\cdot\vec{y})}{\|\vec{y}\|^2}\vec{y}.$$

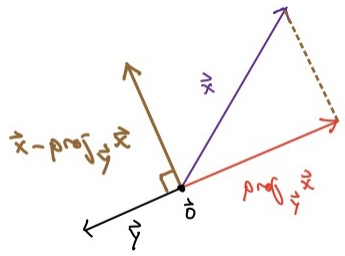
The following proposition justifies the geometric intuition behind orthogonal projections.

Proposition 4 (Characterizations of Projection)**.** Let $\vec{x}, \vec{y} \in \mathbb{R}^n$ with $\vec{y} \neq \vec{0}$.

(i) $\text{proj}_{\vec{y}}(\vec{x})$ is the unique vector in $\text{span}(\vec{y})$ such that $\vec{x} - \text{proj}_{\vec{y}}(\vec{x})$ is orthogonal to $\vec{y}$.

(ii) For every $\vec{u} \in \text{span}(\vec{y})$, $\|\vec{x}-\vec{u}\| \geq \|\vec{x}-\text{proj}_{\vec{y}}(\vec{x})\|$, with equality if, and only if, $\vec{u} = \text{proj}_{\vec{y}}(\vec{x})$.

Remark 12. To visualize the claims in this proposition, consider generic vectors $\vec{y} \neq \vec{0}$ and $\vec{x}$ shown in the following picture.



Part (ii) says that $\text{proj}_{\vec{y}}(\vec{x})$ is the vector in $\text{span}(\vec{y})$ that is closest to $\vec{x}$. In the figure, this means that the dotted line shown (which connects the endpoint of $\vec{x}$ to the endpoint of the red vector in $\text{span}(\vec{y})$) is shortest when the red vector is $\text{proj}_{\vec{y}}(\vec{x})$.

On the other hand, part (i) says that if you subtract $\text{proj}_{\vec{y}}(\vec{x})$ from $\vec{x}$, then the resulting vector (shown in brown) is orthogonal to $\vec{y}$. In other words, the vector that is parallel to the dotted brown line is orthogonal to $\vec{y}$.

Proof of Proposition. For (i), note first that

$$\left(\vec{x} - \text{proj}_{\vec{y}}(\vec{x})\right)\cdot\vec{y} = \vec{x}\cdot\vec{y} - \frac{(\vec{x}\cdot\vec{y})}{\|\vec{y}\|^2}(\vec{y}\cdot\vec{y}) = \vec{x}\cdot\vec{y} - \vec{x}\cdot\vec{y} = 0,$$

so that $\vec{x} - \text{proj}_{\vec{y}}(\vec{x})$ is orthogonal to $\vec{y}$. Let $\vec{v} \in \text{span}(\vec{y})$ and suppose that $0 = (\vec{x}-\vec{v})\cdot\vec{y}$. Choose $c \in \mathbb{R}$ with $\vec{v} = c\vec{y}$. Then $0 = (\vec{x}-c\vec{y})\cdot\vec{y} = \vec{x}\cdot\vec{y} - c(\vec{y}\cdot\vec{y})$, so that $c = \frac{\vec{x}\cdot\vec{y}}{\|\vec{y}\|^2}$ and $\vec{v} = \frac{(\vec{x}\cdot\vec{y})}{\|\vec{y}\|^2}\vec{y} = \text{proj}_{\vec{y}}(\vec{x})$. This shows uniqueness, and (i) is proved.

You will prove (ii) on your first discussion worksheet. □

Remark 13. Let $\vec{y} \in \mathbb{R}^n$ with $\vec{y} \neq \vec{0}$. Then $\text{proj}_{\vec{y}} : \mathbb{R}^n \to \mathbb{R}^n$ is a linear transformation, since for every $\vec{x}, \vec{z} \in \mathbb{R}^n$ and $\lambda \in \mathbb{R}$,

$$\text{proj}_{\vec{y}}(\lambda\vec{x} + \vec{z}) = \frac{((\lambda\vec{x} + \vec{z}) \cdot \vec{y})}{\|\vec{y}\|^2}\vec{y} = \lambda\frac{(\vec{x} \cdot \vec{y})}{\|\vec{y}\|^2}\vec{y} + \frac{(\vec{z} \cdot \vec{y})}{\|\vec{y}\|^2}\vec{y} = \lambda\text{proj}_{\vec{y}}(\vec{x}) + \text{proj}_{\vec{y}}(\vec{z}).$$

Orthogonal Bases

We now turn our attention to the consequences of working with orthogonal *sets* of vectors.

Definition 4. Let $\vec{v}_1, \ldots, \vec{v}_m \in \mathbb{R}^n$. We say $\vec{v}_1, \ldots, \vec{v}_m$ is an **orthogonal** set if $\vec{v}_i \cdot \vec{v}_j = 0$ for every i, j with $i \neq j$.

We say the set $\vec{v}_1, \ldots, \vec{v}_m$ is **orthonormal** if $\vec{v}_1, \ldots, \vec{v}_m$ is an orthogonal set and $\|\vec{v}_i\| = 1$ for each $1 \leq i \leq m$. In other words, $\vec{v}_1, \ldots, \vec{v}_m$ is an orthonormal set if, for every i, j betwen 1 and m,

$$\vec{v}_i \cdot \vec{v}_j = \begin{cases} 1 & \text{if } i = j, \\ 0 & \text{if } i \neq j. \end{cases}$$

Example 1. $\vec{e}_1, \ldots, \vec{e}_m$ is an orthonormal set in $\mathbb{R}^n$ whenever $m \leq n$.

Remark 14. The empty set of vectors is (vacuously) orthonormal.

The following theorem gives a first hint as to why orthonormal sets of vectors are so nice.

Theorem 2 (Coefficients of Orthogonal Combinations)**.** Let $\vec{v}_1, \ldots, \vec{v}_m$ be an orthogonal set of nonzero vectors in $\mathbb{R}^n$, and let $\vec{x} \in \mathbb{R}^n$. If $c_1, \ldots, c_m \in \mathbb{R}$ satisfy $\vec{x} = c_1\vec{v}_1 + \cdots + c_m\vec{v}_m$, then $c_k = \dfrac{\vec{x} \cdot \vec{v}_k}{\|\vec{v}_k\|^2}$ for each $1 \leq k \leq m$.

Proof. Suppose that $c_1, \ldots, c_m \in \mathbb{R}$ satisfy

$$\vec{x} = c_1\vec{v}_1 + \cdots + c_m\vec{v}_m.$$

Let $1 \leq k \leq m$. Then $\vec{v}_k \cdot \vec{v}_j = 0$ if $j \neq k$, so that

$$\begin{aligned} \vec{x} \cdot \vec{v}_k &= (c_1\vec{v}_1 + \cdots + c_m\vec{v}_m) \cdot \vec{v}_k \\ &= c_1(\vec{v}_1 \cdot \vec{v}_k) + \cdots + c_k(\vec{v}_k \cdot \vec{v}_k) + \cdots + c_m(\vec{v}_m \cdot \vec{v}_k) \\ &= 0 + \cdots + 0 + c_k\|\vec{v}_k\|^2 + 0 + \cdots + 0 \\ &= c_k\|\vec{v}_k\|^2. \end{aligned}$$

Because $\|\vec{v}_k\| > 0$, $c_k = \frac{\vec{x} \cdot \vec{v}_k}{\|\vec{v}_k\|^2}$. □

Corollary 1. Every orthogonal set of nonzero vectors is linearly independent.

Proof. Let $\vec{v}_1, \ldots, \vec{v}_m \in \mathbb{R}^n$ be an orthogonal set of nonzero vectors, and suppose that $c_1, \ldots, c_m \in \mathbb{R}$ satisfy $\vec{0} = c_1\vec{v}_1 + \cdots + c_m\vec{v}_m$. By the Coefficients of Orthogonal Combinations Theorem, for each $1 \leq k \leq m$ we have $c_k = \dfrac{\vec{0} \cdot \vec{v}_k}{\|\vec{v}_k\|^2} = 0$. □

Remark 15. Note that the previous corollary implies that orthonormal sets of nonzero vectors are linearly independent.

Corollary 2. Let $V \subseteq \mathbb{R}^n$ be a subspace, and suppose that $\vec{v}_1, \ldots, \vec{v}_m \in V$ is an orthogonal set of nonzero vectors. Then $\vec{v}_1, \ldots, \vec{v}_m$ is a basis for V if, and only if, $m = \dim(V)$.

Proof. This follows immediately from the previous corollary and a result from MATH 291-1 (Theorem 26 in the MATH 291-1 lecture notes). □

Last quarter we explored the properties of bases, and their utility for describing subspaces. As it turns out, orthogonal bases (i.e. bases that are also orthogonal sets) are even better to work with, because of the Coefficients of Orthogonal Combinations Theorem. In general, the idea is that if bases allow us to put a coordinate system into a vector space (with the span of each basis vector acting as a coordinate axis), then an orthogonal basis ensures that each of these coordinate axes is perpendicular to the others. Therefore we expect whatever intuition we have for the standard coordinate system to transfer over to coordinate systems imposted by orthogonal bases. Here is one example of this phenomenon.

Proposition 5. Let V be a subspace of $\mathbb{R}^n$, and suppose that V has an orthogonal basis $\vec{v}_1, \ldots, \vec{v}_m$. Then for each $\vec{x} \in V$,

$$\vec{x} = \text{proj}_{\vec{v}_1}(\vec{x}) + \cdots + \text{proj}_{\vec{v}_m}(\vec{x}) = \left(\frac{\vec{x} \cdot \vec{v}_1}{\|\vec{v}_1\|^2}\right)\vec{v}_1 + \cdots + \left(\frac{\vec{x} \cdot \vec{v}_m}{\|\vec{v}_m\|^2}\right)\vec{v}_m.$$

Proof. This follows immediately from the Coefficients of Orthogonal Combinations Theorem and the assumption that $\vec{v}_1, \ldots, \vec{v}_m$ is a basis for V. □

Lecture 3: More Orthogonality

Learning Objectives:

- Transform a basis into an orthogonal or orthonormal basis using the Gram-Schmidt Process.
- Compute the orthogonal projection of a vector onto a subspace.

One crucial step in our study of bases in MATH 291-1 was the proof (as a corollary of the Constructing Bases Theorem) that every finite-dimensional vector space has a basis. It would be helpful to know that every subspace of $\mathbb{R}^n$ has not only a basis, but an orthonormal basis. This is true, and is a consequence of a computational algorithm known as the Gram-Schmidt process.

Intuitively, the Gram-Schmidt process turns a linearly independent set of vectors into an orthonormal set of vectors without changing the span of the set. It does this by adjusting each vector in the set one at a time. Here is the result.

Theorem 3 (Gram-Schmidt)**.** Suppose that $\vec{v}_1, \ldots, \vec{v}_m \in \mathbb{R}^n$ is a linearly independent set and let $V \stackrel{def}{=} \text{span}(\vec{v}_1, \ldots, \vec{v}_m)$. Define $\vec{b}_1, \ldots, \vec{b}_m$ recursively by

$$\vec{b}_1 \stackrel{def}{=} \vec{v}_1, \quad \vec{b}_k \stackrel{def}{=} \vec{v}_k - \text{proj}_{\vec{b}_1}(\vec{v}_k) - \text{proj}_{\vec{b}_2}(\vec{v}_k) - \cdots - \text{proj}_{\vec{b}_{k-1}}(\vec{v}_k), \quad 2 \leq k \leq m.$$

Then $\vec{b}_1, \ldots, \vec{b}_m$ is an orthogonal basis for V.
For $1 \leq k \leq m$ define $\vec{u}_k \stackrel{def}{=} \frac{1}{\|\vec{b}_k\|}\vec{b}_k$, then $\vec{u}_1, \ldots, \vec{u}_m$ is an orthonormal basis for V.

Proof. We proceed by induction. For the base case, note that $\vec{b}_1$ is a linearly independent (and vacuously orthogonal) list of vectors and $\text{span}(\vec{b}_1) = \text{span}(\vec{v}_1)$.

Let $m \in \mathbb{N}$ with $m \geq 2$, suppose that $\vec{v}_1, \ldots, \vec{v}_m$ is a linearly independent set, and that $\vec{b}_1, \ldots, \vec{b}_{m-1}$ is an orthogonal basis for $\text{span}(\vec{v}_1, \ldots, \vec{v}_{m-1})$. Throughout the proof, for $1 \leq j \leq m-1$ write $d_j \stackrel{def}{=} \frac{\vec{v}_m \cdot \vec{b}_j}{\|\vec{b}_j\|^2}$.

We first claim that $\text{span}(\vec{b}_1, \ldots, \vec{b}_m) \subseteq \text{span}(\vec{v}_1, \ldots, \vec{v}_m)$. To see this, let $\vec{u} \in \text{span}(\vec{b}_1, \ldots, \vec{b}_m)$, and choose $c_1, \ldots, c_m \in \mathbb{R}$ such that $\vec{u} = c_1\vec{b}_1 + \cdots + c_{m-1}\vec{b}_{m-1} + c_m\vec{b}_m$. Since $\vec{b}_m = \vec{v}_m - d_1\vec{b}_1 - \cdots - d_{m-1}\vec{b}_{m-1}$, we have

$$\vec{u} = (c_1 - d_1)\vec{b}_1 + \cdots + (c_{m-1} - d_{m-1})\vec{b}_{m-1} + c_m\vec{v}_m.$$

By the induction hypothesis, $(c_1 - d_1)\vec{b}_1 + \cdots + (c_{m-1} - d_{m-1})\vec{b}_{m-1} \in \text{span}(\vec{v}_1, \ldots, \vec{v}_{m-1})$ and therefore there are $a_1, \ldots, a_{m-1} \in \mathbb{R}$ with $(c_1 - d_1)\vec{b}_1 + \cdots + (c_{m-1} - d_{m-1})\vec{b}_{m-1} = a_1\vec{v}_1 + \cdots + a_{m-1}\vec{v}_{m-1}$. But then

$$\begin{aligned}
\vec{u} &= c_1\vec{b}_1 + \cdots + c_{m-1}\vec{b}_{m-1} + c_m\vec{b}_m \\
&= (c_1 - d_1)\vec{b}_1 + \cdots + (c_{m-1} - d_{m-1})\vec{b}_{m-1} + c_m\vec{v}_m \\
&= a_1\vec{v}_1 + \cdots + a_{m-1}\vec{v}_{m-1} + c_m\vec{v}_m \\
&\in \text{span}(\vec{v}_1, \ldots, \vec{v}_m),
\end{aligned}$$

proving the claim.

Now suppose additionally that $\vec{0} = c_1\vec{b}_1 + \cdots + c_m\vec{b}_m$. Because $\vec{v}_1, \ldots, \vec{v}_m$ is a linearly independent set, $a_1 = \cdots = a_{m-1} = c_m = 0$. Therefore $\vec{0} = c_1\vec{b}_1 + \cdots + c_{m-1}\vec{b}_{m-1}$ and the induction hypothesis implies that $c_1 = \cdots = c_{m-1} = 0$. Therefore $\vec{b}_1, \ldots, \vec{b}_m$ is a linearly independent set. Because $\dim(V) = m$, $\vec{b}_1, \ldots, \vec{b}_m$ is also a basis for V.

The only remaining this to check is that the set $\vec{b}_1, \ldots, \vec{b}_m$ is orthogonal. Because $\vec{b}_1, \ldots, \vec{b}_{m-1}$ is an orthogonal list and, for each $1 \leq \ell \leq m-1$,

$$\vec{b}_\ell \cdot \vec{b}_m = \vec{b}_\ell \cdot (\vec{v}_m - d_1\vec{b}_1 - \cdots - d_{m-1}\vec{b}_{m-1}) = \vec{b}_\ell \cdot \vec{v}_m - d_\ell(\vec{b}_\ell \cdot \vec{b}_\ell) = \vec{b}_\ell \cdot \vec{v}_m - \frac{(\vec{v}_m \cdot \vec{b}_\ell)}{\|\vec{b}_\ell\|^2}\|\vec{b}_\ell\|^2 = 0,$$

the Principle of Mathematical Induction concludes the proof of the first claim.

Because $\vec{b}_1, \ldots, \vec{b}_m$ is an orthogonal basis for V, it follows immediately that for each $1 \leq j, k \leq m$ we have

$$\vec{u}_j \cdot \vec{u}_k = \frac{\vec{b}_j \cdot \vec{b}_k}{\|\vec{b}_j\|\|\vec{b}_k\|} = \begin{cases} 0 & \text{if } j \neq k, \\ 1 & \text{if } j = k. \end{cases}$$

□

Corollary 3. Let V be a subspace of $\mathbb{R}^n$. Then V has an orthonormal basis.

Proof. By the Constructing Bases Theorem, V has a basis. If V is trivial, then this basis is the empty set (which is already orthonormal). If V is not trivial, then we apply the Gram-Schmidt Process to a basis for V to obtain an orthonormal basis. □

Remark 16. At its heart, the Gram-Schmidt theorem is really an algorithm (called the **Gram-Schmidt Process**) for producing an orthonormal basis for a subspace $V \subseteq \mathbb{R}^n$ as follows:

(i) Produce a basis $\vec{v}_1, \ldots, \vec{v}_m$ for V.

(ii) Produce an *orthogonal* basis $\vec{b}_1, \ldots, \vec{b}_m$ for V by computing (in order):

$$\begin{aligned}
\vec{b}_1 &\stackrel{def}{=} \vec{v}_1 \\
\vec{b}_2 &\stackrel{def}{=} \vec{v}_2 - \text{proj}_{\vec{b}_1}(\vec{v}_2) \\
\vec{b}_3 &\stackrel{def}{=} \vec{v}_3 - \text{proj}_{\vec{b}_1}(\vec{v}_3) - \text{proj}_{\vec{b}_2}(\vec{v}_3) \\
&\vdots \\
\vec{b}_m &\stackrel{def}{=} \vec{v}_m - \text{proj}_{\vec{b}_1}(\vec{v}_m) - \cdots - \text{proj}_{\vec{b}_{m-1}}(\vec{v}_m).
\end{aligned}$$

(iii) Produce an *orthonormal* basis $\vec{u}_1, \ldots, \vec{u}_m$ for V by taking $\vec{u}_k = \frac{1}{\|\vec{b}_k\|}\vec{b}_k$ for $1 \leq k \leq m$.

We will soon be able to develop a better understanding of exactly why the Gram-Schmidt Process, but for now we illustrate the algorithm with a few examples.

Example 2. Consider the basis for $\mathbb{R}^3$ given by

$$\vec{v}_1 = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}, \quad \vec{v}_2 = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}, \quad \vec{v}_3 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}.$$

We apply the Gram-Schmidt Process to produce an orthonormal basis for $\mathbb{R}^3$. We execute step (ii) in the algorithm to obtain

$$\begin{aligned}
\vec{b}_1 &\stackrel{def}{=} \vec{v}_1 = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}, \\
\vec{b}_2 &\stackrel{def}{=} \vec{v}_2 - \text{proj}_{\vec{b}_1}(\vec{v}_2) = \vec{v}_2 - \left(\frac{\vec{v}_2 \cdot \vec{b}_1}{\|\vec{b}_1\|^2} \right) \vec{b}_1 = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} - \frac{2}{3} \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 1/3 \\ 1/3 \\ -2/3 \end{bmatrix}, \\
\vec{b}_3 &\stackrel{def}{=} \vec{v}_3 - \text{proj}_{\vec{b}_1}(\vec{v}_3) - \text{proj}_{\vec{b}_2}(\vec{v}_3) = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} - \frac{1}{3} \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} - \frac{1/3}{6/9} \begin{bmatrix} 1/3 \\ 1/3 \\ -2/3 \end{bmatrix} = \begin{bmatrix} 1/2 \\ -1/2 \\ 0 \end{bmatrix}.
\end{aligned}$$

The final step (iii) turns this orthogonal basis for $\mathbb{R}^3$ into an orthonormal basis for $\mathbb{R}^3$:

$$\vec{u}_1 = \frac{1}{\|\vec{b}_1\|} \vec{b}_1 = \begin{bmatrix} 1/\sqrt{3} \\ 1/\sqrt{3} \\ 1/\sqrt{3} \end{bmatrix}, \quad \vec{u}_2 = \frac{1}{\|\vec{b}_2\|} \vec{b}_2 = \begin{bmatrix} 1/\sqrt{6} \\ 1/\sqrt{6} \\ -2/\sqrt{6} \end{bmatrix}, \quad \vec{u}_3 = \frac{1}{\|\vec{b}_3\|} \vec{b}_3 = \begin{bmatrix} 1/\sqrt{2} \\ -1/\sqrt{2} \\ 0 \end{bmatrix}.$$

Remark 17. Note in the previous example that if we had instead applied the Gram-Schmidt Process to the vectors in $\vec{v}_1, \vec{v}_2, \vec{v}_3$ in the order

$$\vec{v}_3 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \quad \vec{v}_2 = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}, \quad \vec{v}_1 = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix},$$

then we would have ended up with $\vec{u}_1 = \vec{e}_1$, $\vec{u}_2 = \vec{e}_2$, $\vec{u}_3 = \vec{e}_3$. This is a different orthonormal basis for $\mathbb{R}^3$ than the one that we ended up with, and illustrates an important point: the orthonormal basis produced by the Gram-Schmidt process depends not only on the basis you start with, but even on the order in which the vectors in that basis are listed. Although this is not extremely important in theory, it is useful to know that the result of the Gram-Schmidt is not necessarily unique.

Example 3. Produce an orthonormal basis for null(A), where $A = \left[\begin{smallmatrix} 1 & 3 & 6 & -2 & 3 \\ 2 & 4 & 8 & -4 & 4 \end{smallmatrix}\right] \in M_{2\times 5}(\mathbb{R})$.

Our first goal is to find *some* basis for null(A). Once we have this, we will be able to apply the Gram-Schmidt process to produce an *orthonormal* basis for null(A).

We first compute that

$$\text{rref}(A) = \left[\begin{smallmatrix} 1 & 0 & 0 & -2 & 0 \\ 0 & 1 & 2 & 0 & 1 \end{smallmatrix}\right],$$

whence it follows (from a homework problem last quarter) that a basis for null(A) is

$$\vec{v}_1 \stackrel{def}{=} \begin{bmatrix} 0 \\ -1 \\ 0 \\ 0 \\ 1 \end{bmatrix}, \quad \vec{v}_2 \stackrel{def}{=} \begin{bmatrix} 2 \\ 0 \\ 0 \\ 1 \\ 0 \end{bmatrix}, \quad \vec{v}_3 \stackrel{def}{=} \begin{bmatrix} 0 \\ -2 \\ 1 \\ 0 \\ 0 \end{bmatrix}.$$

Applying step (ii) of Gram-Schmidt, we obtain

$$\vec{b}_1 \stackrel{def}{=} \vec{v}_1 = \begin{bmatrix} 0 \\ -1 \\ 0 \\ 0 \\ 1 \end{bmatrix},$$

$$\vec{b}_2 \stackrel{def}{=} \vec{v}_2 - \text{proj}_{\vec{b}_1}(\vec{v}_2) = \begin{bmatrix} 2 \\ 0 \\ 0 \\ 1 \\ 0 \end{bmatrix} - \frac{0}{2}\begin{bmatrix} 0 \\ -1 \\ 0 \\ 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 2 \\ 0 \\ 0 \\ 1 \\ 0 \end{bmatrix},$$

$$\vec{b}_3 \stackrel{def}{=} \vec{v}_3 - \text{proj}_{\vec{b}_1}(\vec{v}_3) - \text{proj}_{\vec{b}_2}(\vec{v}_3) = \begin{bmatrix} 0 \\ -2 \\ 1 \\ 0 \\ 0 \end{bmatrix} - \frac{2}{2}\begin{bmatrix} 0 \\ -1 \\ 0 \\ 0 \\ 1 \end{bmatrix} - \frac{0}{5}\begin{bmatrix} 2 \\ 0 \\ 0 \\ 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ -1 \\ 1 \\ 0 \\ -1 \end{bmatrix}.$$

Performing step (iii) of the process, we obtain

$$\vec{u}_1 = \begin{bmatrix} 0 \\ -1/\sqrt{2} \\ 0 \\ 0 \\ 1/\sqrt{2} \end{bmatrix}, \quad \vec{u}_2 = \begin{bmatrix} 2/\sqrt{5} \\ 0 \\ 0 \\ 1/\sqrt{5} \\ 0 \end{bmatrix}, \quad \vec{u}_3 = \begin{bmatrix} 0 \\ -1/\sqrt{3} \\ 1/\sqrt{3} \\ 0 \\ -1/\sqrt{3} \end{bmatrix},$$

which forms an orthonormal basis for null(A).

General Orthogonal Projections

As an application the existence of orthonormal bases, we can show that the notion of orthogonal projection extends to general subspaces of $\mathbb{R}^n$. We start with a definition.

Definition 5. For a subspace $V \subseteq \mathbb{R}^n$, define $V^\perp$, the **orthogonal complement of** V, by

$$V^\perp \stackrel{def}{=} \{\vec{w} \ : \ \vec{w} \cdot \vec{v} = 0 \text{ for every } \vec{v} \in V\}.$$

Remark 18. For a subspace V, $V^\perp$ consists of all vectors in $\mathbb{R}^n$ that are orthogonal to all vectors in V.

Remark 19. Let V be a subspace of $\mathbb{R}^n$. Then $V^\perp$ is a subspace of $\mathbb{R}^n$. This follows immediately from the Subspace Criteria, since $\vec{0} \in V^\perp$ and if $\vec{x}, \vec{y} \in V^\perp$ and $\lambda \in \mathbb{R}$, then for every $\vec{v} \in V$ we have

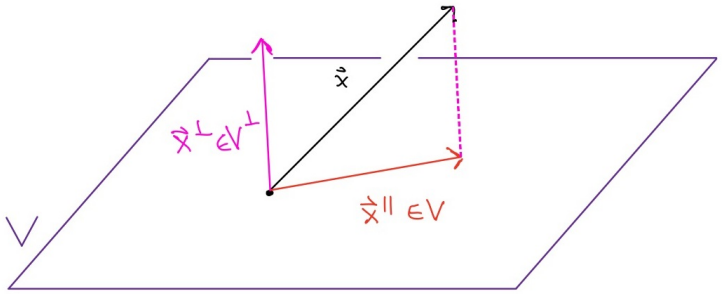
$$(\lambda\vec{x} + \vec{y}) \cdot \vec{v} = \lambda(\vec{x} \cdot \vec{v}) + \vec{y} \cdot \vec{v} = 0.$$

Remark 20. For nonzero $\vec{y} \in \mathbb{R}^n$ and for $\vec{x} \in \mathbb{R}^n$, we know that $\text{proj}_{\vec{y}}(\vec{x}) \in \text{span}(\vec{y})$ and that since $(\vec{x} - \text{proj}_{\vec{y}}(\vec{x})) \cdot (c\vec{y}) = 0$ for each $c \in \mathbb{R}$, $\vec{x} - \text{proj}_{\vec{y}}(\vec{x}) \in [\text{span}(\vec{y})]^\perp$. Therefore we can write $\vec{x}$ as the sum of a vector in span$(\vec{y})$ and a vector in span$(\vec{y})^\perp$ as $\vec{x} = \text{proj}_{\vec{y}}(\vec{x}) + (\vec{x} - \text{proj}_{\vec{y}}(\vec{x}))$. We can decompose vectors in a similar way relative to general subspaces of $\mathbb{R}^n$, not only for subspaces that are the span of a single vector. This is the content of the following theorem.

Theorem 4 (Orthogonal Decomposition)**.** Let $V \subseteq \mathbb{R}^n$ be a subspace, and let $\vec{x} \in \mathbb{R}^n$. There exist unique vectors $\vec{x}^{\parallel} \in V$ and $\vec{x}^{\perp} \in V^{\perp}$ with $\vec{x} = \vec{x}^{\parallel} + \vec{x}^{\perp}$. The vector $\vec{x}^{\parallel}$ is called the **orthogonal projection** of $\vec{x}$ onto V, and denoted $\text{proj}_V(\vec{x})$.

Moreover, if $\vec{b}_1, \ldots, \vec{b}_m$ is an orthogonal basis for V, then

$$\text{proj}_V(\vec{x}) = \text{proj}_{\vec{b}_1}(\vec{x}) + \cdots + \text{proj}_{\vec{b}_m}(\vec{x}).$$



Remark 21. Note that because the vector $\vec{x}^{\parallel}$ is unique, $\text{proj}_V(\vec{x})$ does not depend on which orthogonal basis we take for V.

Proof. We start by showing existence. If V is trivial, then take $\vec{x}^{\parallel} = \vec{0}$ and $\vec{x}^{\perp} = \vec{x}$. Since $\vec{x}^{\parallel} = \vec{0} \in V$, and since $\vec{x} \cdot \vec{0} = 0$ so that $\vec{x}^{\perp} = \vec{x} \in V^{\perp}$, this is a decomposition of the type promised by the theorem. Suppose that V is non-trivial. Fix an orthogonal basis $\vec{b}_1, \ldots, \vec{b}_m$ for V, and define $\vec{x}^{\parallel}$ by the formula at the end of the theorem. Because

$$\text{proj}_{\vec{b}_k}(\vec{x}) = \frac{(\vec{x} \cdot \vec{b}_k)}{\|\vec{b}_k\|^2}\vec{b}_k \in V$$

for each $1 \leq k \leq m$ and since V is closed under linear combinations, $\vec{x}^{\parallel} \in V$. To conclude the proof of existence of the decomposition, we therefore need only show that $\vec{x}^{\perp} \overset{def}{=} \vec{x} - \vec{x}^{\parallel} \in V^{\perp}$.

Let $\vec{v} \in V$. By writing $\vec{v} = c_1\vec{b}_1 + \cdots + c_m\vec{b}_m$, we use orthogonality of $\vec{b}_1, \ldots, \vec{b}_m$ to see that

$$\begin{aligned}(\vec{x} - \vec{x}^{\parallel}) \cdot \vec{v} &= (\vec{x} - \frac{(\vec{x} \cdot \vec{b}_1)}{\|\vec{b}_1\|^2}\vec{b}_1 - \cdots - \frac{(\vec{x} \cdot \vec{b}_m)}{\|\vec{b}_m\|}\vec{b}_m) \cdot (c_1\vec{b}_1 + \cdots + c_m\vec{b}_m) \\ &= c_1(\vec{x} \cdot \vec{b}_1) + \cdots + c_m(\vec{x} \cdot \vec{b}_m) - (\vec{x} \cdot \vec{b}_1)c_1 - \cdots - (\vec{x} \cdot \vec{b}_m)c_m \\ &= 0.\end{aligned}$$

Therefore $\vec{x}^{\perp} \in V^{\perp}$, and the existence of the orthogonal decomposition of $\vec{x}$ relative to V is complete.

We now show the the orthogonal decomposition of $\vec{x}$ relative to V is unique. Suppose that $\vec{x} = \vec{x}_0^{\parallel} + \vec{x}_0^{\perp}$ with $\vec{x}_0^{\parallel} \in V$ and $\vec{x}_0^{\perp} \in V^{\perp}$. Then $\vec{x}_0^{\parallel} - \vec{x}^{\parallel} \in V$ and $\vec{x}_0^{\perp} - \vec{x}^{\perp} \in V^{\perp}$ and

$$\vec{0} = \vec{x} - \vec{x} = (\vec{x}_0^{\parallel} - \vec{x}^{\parallel}) + (\vec{x}_0^{\perp} - \vec{x}^{\perp}).$$

Taking the dot product of both sides with $\vec{x}_0^{\parallel} - \vec{x}^{\parallel}$ yields

$$0 = \vec{0} \cdot (\vec{x}_0^{\parallel} - \vec{x}^{\parallel}) = (\vec{x}_0^{\parallel} - \vec{x}^{\parallel}) \cdot (\vec{x}_0^{\parallel} - \vec{x}^{\parallel}) + \underbrace{(\vec{x}_0^{\perp} - \vec{x}^{\perp})}_{\in V^{\perp}} \cdot \underbrace{(\vec{x}_0^{\parallel} - \vec{x}^{\parallel})}_{\in V} = (\vec{x}_0^{\parallel} - \vec{x}^{\parallel}) \cdot (\vec{x}_0^{\parallel} - \vec{x}^{\parallel}).$$

Because $\cdot$ is positive definite, we have $\vec{x}_0^{\parallel} - \vec{x}^{\parallel} = \vec{0}$, so that $\vec{x}_0^{\parallel} = \vec{x}^{\parallel}$. If we had instead taken the dot product of both sides with $\vec{x}_0^{\perp} - \vec{x}^{\perp}$, the result would be that $\vec{x}_0^{\perp} = \vec{x}^{\perp}$. Therefore, the vectors $\vec{x}^{\parallel}$ and $\vec{x}^{\perp}$ satisfying the conclusion of the theorem are unique, and we are done. □

Remark 22. Note that if $\{\vec{0}\}$ is the trivial subspace of $\mathbb{R}^n$ and if $\vec{x} \in \mathbb{R}^n$, then in the course of the proof we showed that $\text{proj}_{\{\vec{0}\}}(\vec{x}) = \vec{0}$.

Remark 23. Let $V \subseteq \mathbb{R}^n$. Because $\text{proj}_V : \mathbb{R}^n \to \mathbb{R}^n$ can be written as $\text{proj}_V = \text{proj}_{\vec{b}_1} + \cdots + \text{proj}_{\vec{b}_m}$, and because the sum of linear transformations is linear, proj_V is linear.

Remark 24. Fix a nontrivial subspace V of $\mathbb{R}^n$, and let $\vec{u}_1, \ldots, \vec{u}_m$ be an orthonormal basis for V. Then for $\vec{x} \in \mathbb{R}^n$ we can write

$$\begin{aligned}
\text{proj}_V \vec{x} &= \text{proj}_{\vec{u}_1} \vec{x} + \cdots + \text{proj}_{\vec{u}_m} \vec{x} \\
&= (\vec{x} \cdot \vec{u}_1)\vec{u}_1 + \cdots + (\vec{x} \cdot \vec{u}_m)\vec{u}_m \\
&= \begin{bmatrix} \vec{u}_1 & \cdots & \vec{u}_m \end{bmatrix} \begin{bmatrix} \vec{u}_1^T \vec{x} \\ \vdots \\ \vec{u}_m^T \vec{x} \end{bmatrix} \\
&= \begin{bmatrix} \vec{u}_1 & \cdots & \vec{u}_m \end{bmatrix} \begin{bmatrix} \vec{u}_1^T \\ \vdots \\ \vec{u}_m^T \end{bmatrix} \vec{x} \\
&= \begin{bmatrix} \vec{u}_1 & \cdots & \vec{u}_m \end{bmatrix} \begin{bmatrix} \vec{u}_1 & \cdots & \vec{u}_m \end{bmatrix}^T \vec{x}.
\end{aligned}$$

In other words, if $Q \in M_{n \times m}(\mathbb{R})$ is the matrix with columns $\vec{u}_1, \ldots, \vec{u}_m$, then $\text{proj}_V(\vec{x}) = (QQ^T)\vec{x}$. This furnishes another proof that $\text{proj}_V : \mathbb{R}^n \to \mathbb{R}^n$ is linear with matrix QQ^T.

Remark 25. We are now in a position to understand more generally the idea behind the Gram-Schmidt Process. Let $\vec{v}_1, \ldots, \vec{v}_m \in \mathbb{R}^n$, and suppose that $\vec{v}_1, \ldots, \vec{v}_m$ is a linearly independent set. For each $1 \leq k \leq m$, define $V_k \stackrel{def}{=} \text{span}(\vec{v}_1, \ldots, \vec{v}_k)$. Then each V_k is a subspace, and $V_1 \subset V_2 \subset V_3 \subset \cdots \subset V_m$. Then the orthogonal set of vectors $\vec{b}_1, \ldots, \vec{b}_m$ defined by the Gram-Schmidt process can be viewed as

$$\vec{b}_k \stackrel{def}{=} \vec{v}_k - \text{proj}_{V_{k-1}}(\vec{v}_k).$$

That is, each $\vec{b}_k$ is designed to be the vector $\vec{v}_k^{\perp}$ in the orthogonal decomposition of $\vec{v}_k$ relative to V_{k-1}, which is in the orthogonal complement of $V_{k-1} = \text{span}(\vec{v}_1, \ldots, \vec{v}_{k-1})$ (and therefore orthogonal to the previous vectors in the list). The proof of that theorem could be reinterpreted using this framework, but we could not have used this argument to prove the Gram-Schmidt Process because our treatment of orthogonal projections relied on the existence of orthonormal bases, which was a consequence of the Gram-Schmidt Process!

Lecture 4: Even More Orthogonality

Learning Objectives:

- Determine when a matrix or transformation is orthogonal.
- Apply several equivalent characterizations of orthogonal matrices and transformations.

Recall from the first quarter that, once we understood the linear structure of a vector space, the next natural question to ask is which functions between vectors spaces preserve that linear structure. (The answer was "linear functions".) Now that we understand how the dot product captures the geometric notions of length and orthogonality (and, more generally, angle) in $\mathbb{R}^n$, we might ask ourselves which linear transformations $T : \mathbb{R}^n \to \mathbb{R}^n$ preserve this geometric structure (in an appropriate sense). To this end, we make a definition.

Definition 6. A linear transformation $T : \mathbb{R}^n \to \mathbb{R}^n$ is called **orthogonal** if $\|T(\vec{x})\| = \|\vec{x}\|$ for every $\vec{x} \in \mathbb{R}^n$.

Remark 26. The choice of the moniker **orthogonal** here is not obvious, since the defining condition of an orthogonal transformation seems to have a lot more to do with the transformation preserving the lengths of vectors (i.e. $\|T(\vec{x})\| = \|\vec{x}\|$) rather than with preserving orthogonality between vectors (i.e. $T(\vec{x})$ and $T(\vec{y})$ are orthogonal exactly when $\vec{x}$ and $\vec{y}$ are orthogonal). Perhaps surprisingly, these conditions turn out to be equivalent (as we will presently show, with some help from you on your homework).

To justify that our definition of orthogonal transformation is the correct one, we have the following theorem.

Theorem 5 (Orthogonal Transformations)**.** Let $T : \mathbb{R}^n \to \mathbb{R}^n$, $T(\vec{x}) = A\vec{x}$ be linear. Then the following are equivalent.

(a) $\|T(\vec{x})\| = \|\vec{x}\|$ for every $\vec{x} \in \mathbb{R}^n$.

(b) $T(\vec{x}) \cdot T(\vec{y}) = \vec{x} \cdot \vec{y}$ for every $\vec{x}, \vec{y} \in \mathbb{R}^n$.

(c) $T(\vec{e}_1), \ldots, T(\vec{e}_n)$ forms an orthonormal basis for $\mathbb{R}^n$.

(d) The columns of A form an orthonormal basis for $\mathbb{R}^n$.

(e) $A^T A = I_n$

(f) A is invertible with $A^{-1} = A^T$.

Proof. (a)⇒(b): This is one of your homework problems this week. The idea is to first show that $\vec{x} \cdot \vec{y}$ can be computed in terms of $\|\vec{x}\|^2$, $\|\vec{y}\|^2$, and $\|\vec{x} + \vec{y}\|^2$ (as we've seen in the proof of a past result...).

(b)⇒(c): Suppose (b) holds. Note that for each $1 \leq i, j \leq n$,

$$T(\vec{e}_i) \cdot T(\vec{e}_j) = \vec{e}_i \cdot \vec{e}_j = \begin{cases} 1 & \text{if } i = j, \\ 0 & \text{if } i \neq j. \end{cases}$$

Therefore $T(\vec{e}_1), \ldots, T(\vec{e}_n)$ is an orthonormal set in $\mathbb{R}^n$. Because $\dim(\mathbb{R}^n) = n$ and orthonormal sets are linearly independent, $T(\vec{e}_1), \ldots, T(\vec{e}_n)$ is an orthonormal basis for $\mathbb{R}^n$.

(c)⇒(d): This is immediate because the columns of A are $T(\vec{e}_1), T(\vec{e}_2), \ldots, T(\vec{e}_n)$.

(d)⇒(e): Suppose that the columns $\vec{a}_1, \ldots, \vec{a}_n$ of A form an orthonormal basis for $\mathbb{R}^n$. Then

$$(ij)\text{-th entry of } A^T A = \vec{a}_i^T \vec{a}_j = \vec{a}_i \cdot \vec{a}_j = \begin{cases} 1 & \text{if } i = j, \\ 0 & \text{if } i \neq j. \end{cases}$$

Therefore $A^T A = I_n$.

(e)⇒(f): Suppose $A^T A = I_n$. Since $A, A^T \in M_{n\times n}(\mathbb{R})$, the Invertibility Theorem implies that A is invertible and $A^{-1} = A^T$.

(f)⇒(a): Suppose that (f) holds, and let $\vec{x} \in \mathbb{R}^n$. Then we have

$$\|T(\vec{x})\| = \|A\vec{x}\| = \sqrt{(A\vec{x}) \cdot (A\vec{x})} = \sqrt{\vec{x} \cdot (A^T A\vec{x})} = \sqrt{\vec{x} \cdot (I_n\vec{x})} = \sqrt{\vec{x} \cdot \vec{x}} = \|\vec{x}\|.$$

□

Definition 7. Let $A \in M_{n\times n}(\mathbb{R})$. If $A^T A = I_n$, then we call A an **orthogonal matrix**.

Remark 27. Note that (by the equivalence of (d) and (e)) a real square matrix A is orthogonal if and only if its columns form an orthonormal set.

Remark 28. The equivalence of (a) and (e) can be interpreted as follows: A linear transformation $T : \mathbb{R}^n \to \mathbb{R}^n$ is orthogonal if and only if its standard matrix is orthogonal.

Orthogonality of a transformation or matrix is quite natural, in the sense that it plays nicely many several of the concepts we've discussed so far this year. Here are a few examples.

Corollary 4. The inverse and transpose of an orthogonal matrix is also orthogonal, and the product of orthogonal matrices is also orthogonal.

Proof. Let $A, B \in M_{n\times n}(\mathbb{R})$ be orthogonal matrices.

Since $A^T = A^{-1}$, $(A^T)^T A^T = AA^T = I_n$. Therefore A^T is orthogonal (and therefore so is $A^{-1} = A^T$).

Now note that $AB \in M_{n\times n}(\mathbb{R})$ and

$$(AB)^T AB = B^T A^T AB = B^T I_n B = B^T B = I_n,$$

so that AB is orthogonal. □

Example 4. Let $T : \mathbb{R}^2 \to \mathbb{R}^2$ be an orthogonal transformation, and let A be its matrix. Then the columns of A are orthonormal, so there exist $a, b \in \mathbb{R}$ with $a^2 + b^2 = 1$ such that either

$$A = \begin{bmatrix} a & -b \\ b & a \end{bmatrix} \quad \text{or} \quad A = \begin{bmatrix} a & b \\ b & -a \end{bmatrix}.$$

As you saw in MATH 291-1, in the first case T represents counterclockwise rotation about the origin by an angle θ that satisfies $a = \cos(\theta)$ and $b = \sin(\theta)$. In the second case, T represents orthogonal reflection across a line that passes through the origin. Therefore, the only orthogonal transformations of $\mathbb{R}^2$ are rotations and reflections.

A suitable generalization of this is true in higher dimensions, but we won't be able to prove this until after we investigate determinants and diagonalization.

Example 5. The above results carry over to linear transformations $T : \mathbb{C}^n \to \mathbb{C}^n$, although the terminology and a few of the details change because we are working with the complex dot product. In particular, a complex square matrix $U \in M_{n\times n}(\mathbb{C})$ is called **unitary** if $U^*U = I_n$, where $U^* = \overline{U^T}$ is the conjugate transpose of U. The complex analog of the Orthogonal Transformations Theorem is as follows.

Theorem 6 (Unitary Transformations)**.** Let $T : \mathbb{C}^n \to \mathbb{C}^n$, $T(\vec{z}) = U\vec{z}$. Then the following are equivalent.

(a) $\|T(\vec{z})\| = \|\vec{z}\|$ for every $\vec{z} \in \mathbb{C}^n$.

(b) $T(\vec{z}) \cdot T(\vec{w}) = \vec{z} \cdot \vec{w}$ for every $\vec{z}, \vec{w} \in \mathbb{C}^n$.

(c) $T(\vec{e}_1), \ldots, T(\vec{e}_n)$ forms an orthonormal basis for $\mathbb{C}^n$.

(d) The columns of U form an orthonormal basis for $\mathbb{C}^n$.

(e) $U^*U = I_n$

(f) U is invertible with $U^{-1} = U^*$.

I leave it to you to verify the details.

Example 6. [2]For each $a, b, c, d, f \in \mathbb{C}$ and $\lambda \in \mathbb{R}$, define

$$U = \begin{bmatrix} a & 0 & \lambda i/\sqrt{2} \\ b & d & f \\ c & 0 & -\lambda i/\sqrt{2} \end{bmatrix}.$$

Determine which values of a, b, c, d, f, λ result in U being unitary.

Let $\vec{x}, \vec{y}, \vec{z}$ denote the columns of U, and suppose that $\vec{x}, \vec{y}, \vec{z}$ is an orthonormal set.
Because $1 = \vec{y} \cdot \vec{y} = d\bar{d} = |d|^2$, we must have $|d| = 1$.
Because $0 = \vec{y} \cdot \vec{z} = d\bar{f}$ and $d \neq 0$, we must have $\bar{f} = 0$ and therefore $f = 0$.

[2]This example is due to Santiago Cañez.

Because λ is real and therefore

$$1 = \vec{z}\cdot\vec{z} = \frac{\lambda i}{\sqrt{2}}\frac{\overline{\lambda i}}{\sqrt{2}} + \frac{(-\lambda i)}{\sqrt{2}}\frac{\overline{(-\lambda i)}}{\sqrt{2}} = \frac{\lambda i}{\sqrt{2}}\frac{(-\lambda i)}{\sqrt{2}} + \frac{(-\lambda i)}{\sqrt{2}}\frac{\lambda i}{\sqrt{2}} = \frac{\lambda^2}{2} + \frac{\lambda^2}{2} = \lambda^2,$$

we must have $\lambda = 1$ or $\lambda = -1$ (again, because $\lambda \in \mathbb{R}$).

Because $0 = \vec{x}\cdot\vec{y} = b\bar{d}$ and $\bar{d} \neq 0$, we must have $b = 0$.

Because

$$0 = \vec{x}\cdot\vec{z} = a\frac{\overline{\lambda i}}{\sqrt{2}} + c\frac{\overline{(-\lambda i)}}{\sqrt{2}} = a\frac{(-\lambda i)}{\sqrt{2}} + c\frac{(\lambda i)}{\sqrt{2}} = \frac{\lambda i}{\sqrt{2}}(c-a)$$

and $\lambda \neq 0$, $c - a = 0$ and therefore $c = a$.

Finally, we note that since $1 = \vec{x}\cdot\vec{x} = a\bar{a} + a\bar{a} = 2|a|^2$, we must have $|a| = \frac{1}{\sqrt{2}}$.

Therefore, we must have

$$U = \begin{bmatrix} a & 0 & \pm i/\sqrt{2} \\ 0 & d & 0 \\ a & 0 & \mp i/\sqrt{2} \end{bmatrix}, \quad \text{where } |a| = \frac{1}{\sqrt{2}} \text{ and } |d| = 1.$$

We can verify that matrices of this form are unitary by either checking that the columns are orthonormal directly, or by simply noting that

$$\begin{aligned} U^*U &= \begin{bmatrix} \bar{a} & 0 & \bar{a} \\ 0 & \bar{d} & 0 \\ \mp i/\sqrt{2} & 0 & \pm i/\sqrt{2} \end{bmatrix} \begin{bmatrix} a & 0 & \pm i/\sqrt{2} \\ 0 & d & 0 \\ a & 0 & \mp i/\sqrt{2} \end{bmatrix} \\ &= \begin{bmatrix} 2|a|^2 & 0 & \pm(\bar{a}i/\sqrt{2}) \mp (\bar{a}i/\sqrt{2}) \\ 0 & |d|^2 & 0 \\ \mp(\bar{a}i/\sqrt{2}) \pm (\bar{a}i/\sqrt{2}) & 0 & -(i^2)/2 - (i^2)/2 \end{bmatrix} \\ &= \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}. \end{aligned}$$

Lecture 5: Determinants

Learning Objectives:

- Establish the existence and uniqueness of determinant functions using their fundamental properties.
- Compute the determinant of a matrix using the pattern definition of the determinant.

While the dot product accounts for (or, if we are not working in $\mathbb{R}^2$ or $\mathbb{R}^3$, define) the geometric notions of length and angle (and the related notion of orthogonality), the **determinant** can be used to account for the dimension-appropriate notion of *volume* in $\mathbb{R}^n$ (i.e. length in $\mathbb{R}^1$, area in $\mathbb{R}^2$, volume in $\mathbb{R}^3$), actually giving us a way to define the notion of volume when $n \geq 4$. The determinant $\det(A)$ of a square matrix is a scalar that one computes in terms of the entries of the matrix. That is, the determinant will be a function $\det : M_{n\times n}(\mathbb{K}) \to \mathbb{K}$ whose inputs are $n \times n$ matrices (with entries in $\mathbb{K}$) and whose outputs are scalars. For example, recall that last quarter we saw the determinant of 2×2 matrices when we were studying invertibility:

$$\det\Big(\begin{bmatrix} a & b \\ c & d \end{bmatrix} \Big) = ad - bc.$$

Notation 1. If $D : M_{n\times n}(\mathbb{K}) \to \mathbb{K}$ is a function, then for $A = \begin{bmatrix} \vec{a}_1 & \vec{a}_2 & \cdots & \vec{a}_n \end{bmatrix} \in M_{n\times n}(\mathbb{K})$ we will write

$$D(\vec{a}_1, \vec{a}_2, \ldots, \vec{a}_n) \stackrel{def}{=} D(A),$$

where we list the columns of the matrix as inputs to D rather than the matrix as a whole. This will be useful when describing the algebraic properties of the determinant, which are expressed in terms of the columns of a matrix. We will use these notations interchangeably.

Last quarter we saw that if $A \in M_{2\times 2}(\mathbb{K})$, then A is invertible if, and only if, $\det(A) \neq 0$. The same result with hold for $n \times n$ matrices with entries in $\mathbb{K}$, and this is our primary motivation for studying determinants. However, the geometric interpretation of the determinant (in terms of volume) is equally important and (when we get there) will give us another layer of intuition for understanding linear independence and span.

Like for the dot product, the determinant of a matrix will have nice algebraic properties. The explicit formula for the determinant (of an $n \times n$ matrix A in terms of the entries of A) is complicated to write down explicitly. Instead, it is sometimes helpful to define determinants in terms of their most important algebraic properties.

Definition 8. A function $D : M_{n\times n}(\mathbb{K}) \to \mathbb{K}$ is called a **determinant function** if for every $A = \begin{bmatrix}\vec{a}_1 & \cdots & \vec{a}_n\end{bmatrix} \in M_{n\times n}(\mathbb{K})$, the following hold.

(i) (Multilinearity) For each $1 \le j \le n$, the map

$$T : \mathbb{K}^n \to \mathbb{K}, \quad T(\vec{x}) = D(\vec{a}_1, \ldots, \vec{a}_{j-1}, \vec{x}, \vec{a}_{j+1}, \ldots, \vec{a}_n)$$

is linear.

(ii) (Alternating) If $B \in M_{n\times n}(\mathbb{K})$ is the matrix obtained from A by swapping two of its columns, then $D(B) = -D(A)$.

(iii) $D(I_n) = 1$

Remark 29. On your homework you will prove that $D : M_{n\times n}(\mathbb{K}) \to \mathbb{K}$ is alternating if, and only if, $D(A) = 0$ whenever two columns of A are equal.

Remark 30. Note that $\det : M_{2\times 2}(\mathbb{K}) \to \mathbb{K}$, $\det(\left[\begin{smallmatrix} a & b \\ c & d\end{smallmatrix}\right]) = ad - bc$ satisfies these properties.

(i) (Multilinearity) For each $\vec{x}, \vec{y} \in \mathbb{K}^2$ and $\lambda \in \mathbb{K}$,

$$\begin{aligned}\det\Big(\begin{bmatrix} x_1 + \lambda y_1 & b \\ x_2 + \lambda y_2 & d\end{bmatrix}\Big) &= (x_1 + \lambda y_1)d - (x_2 + \lambda y_2)b \\ &= x_1 d - x_2 b + \lambda(y_1 d - y_2 b) \\ &= \det\Big(\begin{bmatrix} x_1 & b \\ x_2 & d\end{bmatrix}\Big) + \lambda\det\Big(\begin{bmatrix} y_1 & b \\ y_2 & d\end{bmatrix}\Big).\end{aligned}$$

Similarly, $\det(\left[\begin{smallmatrix} a & x_1+\lambda y_1 \\ c & x_2+\lambda y_2\end{smallmatrix}\right]) = \det(\left[\begin{smallmatrix} a & x_1 \\ c & x_2\end{smallmatrix}\right]) + \lambda\det(\left[\begin{smallmatrix} a & y_1 \\ c & y_2\end{smallmatrix}\right])$.

(ii) (Alternating) $\det(\left[\begin{smallmatrix} b & a \\ d & c\end{smallmatrix}\right]) = bc - ad = -(ad - bc) = -\det(\left[\begin{smallmatrix} a & b \\ c & d\end{smallmatrix}\right])$.

(iii) $\det(I_2) = \det(\left[\begin{smallmatrix} 1 & 0 \\ 0 & 1\end{smallmatrix}\right]) = 1$.

The remarkable fact is that these three properties completely characterize the determinant for $n \times n$ matrices in the following sense.

Theorem 7. For each $n \in \mathbb{N}$ there is a unique determinant function $\det : M_{n\times n}(\mathbb{K}) \to \mathbb{K}$.

The proof of this result is quite long, and many of the steps are repetitious. Therefore we will include the entire arugment in the notes, but only cover a portion of the argument in class.

Proof of Uniqueness of Determinants

Suppose that $D : M_{n\times n}(\mathbb{K}) \to \mathbb{K}$ is a multilinear, alternating function that satisfies $D(I_n) = 1$. We will find a formula for $D(A)$ that is completely determined from the entries of A and the fact that D is multilinear, alternating, and satisfies $D(I_n) = 1$. This will show that the value of $D(A)$ does not depend on D, but only on the properties on determinant functions and the entries of A.

Let $A = \begin{bmatrix}\vec{a}_1 & \cdots & \vec{a}_n\end{bmatrix} \in M_{n\times n}(\mathbb{K})$. Then for each $1 \le j \le n$, $\vec{a}_j = a_{1,j}\vec{e}_1 + \cdots + a_{n,j}\vec{e}_n$. Because D is multilinear, we can apply the multlinearity of D a total of n times to see that

$$\begin{aligned} D(A) &= \sum_{i_1=1}^{n} a_{i_1,1} D(\vec{e}_{i_1}, \vec{a}_2, \ldots, \vec{a}_n) \\ &= \sum_{i_1,i_2=1}^{n} a_{i_1,1}a_{i_2,2} D(\vec{e}_{i_1}, \vec{e}_{i_2}, \ldots, \vec{a}_n) \\ &\vdots \\ &= \sum_{i_1,i_2,\ldots,i_n=1}^{n} a_{i_1,1}a_{i_2,2}\cdots a_{i_n,n} D(\vec{e}_{i_1}, \vec{e}_{i_2}, \ldots, \vec{e}_{i_n}). \end{aligned} \tag{2}$$

Notation

Here we write $\sum_{i_1,i_2=1}^{n} \text{☺}_{i_1,i_2}$ as shorthand for $\sum_{i_2=1}^{n}\sum_{i_1=1}^{n} \text{☺}_{i_1,i_2}$. Both of these indicate that we are adding the terms $\text{☺}_{i_1,i_2}$ for $1 \le i_1 \le n$ and $1 \le i_2 \le n$ (i.e. n^2-many terms).
More generally (see (3)) we will want to use summation notation to refers to sums where the summands are not indexed by the integers, but by a different set of objects. That is, we want to consider a sum where each summand corresponds to a different element of an easy-to-describe set. For example, in (3) we write

$$\sum_{\{i_1,i_2\ldots,i_n\}=\{1,2,\ldots,n\}} a_{i_1,1}a_{i_2,2}\cdots a_{i_n,n} D(\vec{e}_{i_1}, \vec{e}_{i_2}, \ldots, \vec{e}_{i_n})$$

to indicate that we are adding up summands of the form

$$a_{i_1,1}a_{i_2,2}\cdots a_{i_n,n} D(\vec{e}_{i_1}, \vec{e}_{i_2}, \ldots, \vec{e}_{i_n}),$$

and that there is one summand for each rearrangement $i_1, \ldots, i_n$ of the numbers $1, 2, \ldots, n$. To capture this idea, we use the notation "$\{i_1, i_2, \ldots, i_n\} = \{1, 2, \ldots, n\}$" to indicate that we are taking each of $i_1, \ldots, i_n$ to be a natural number between 1 and n, with the additional requirement that, as sets, $\{i_1, i_2, \ldots, i_n\} = \{1, 2, \ldots, n\}$, so that each of the numbers 1 through n appears exactly once in the list $i_1, \ldots, i_n$. There are more examples later, so do not hesitate to ask if the notation is confusion!

But because D is alternating, we can throw out any term in the sum where $\vec{e}_{i_k} = \vec{e}_{i_\ell}$ for any $k \ne \ell$. Therefore (2) simplifies to

$$\begin{aligned} D(A) &= \sum_{i_1,i_2,\ldots,i_n=1}^{n} a_{i_1,1}a_{i_2,2}\cdots a_{i_n,n} D(\vec{e}_{i_1}, \vec{e}_{i_2}, \ldots, \vec{e}_{i_n}) \\ &= \sum_{\{i_1,i_2\ldots,i_n\}=\{1,2,\ldots,n\}} a_{i_1,1}a_{i_2,2}\cdots a_{i_n,n} D(\vec{e}_{i_1}, \vec{e}_{i_2}, \ldots, \vec{e}_{i_n}). \end{aligned} \tag{3}$$

Let's think about what each term in the sum represents. First, note that $i_1, \ldots, i_n$ is a rearrangement of the numbers $1, 2, \ldots, n$. Therefore, no two of the entries $a_{i_1,1}, a_{i_2,2}, \ldots, a_{i_n,n}$ appear in the same row of A or (since $1, \ldots, n$ is also a rearrangment of $1, \ldots, n$) in the same column of A. Moreover, by using the alternating property of D we see that $D(\vec{e}_{i_1}, \ldots, \vec{e}_{i_n}) = \pm D(I_n) = \pm 1$, where the sign depends on how many pairs of columns we need to swap in order to rearrange $\vec{e}_{i_1}, \ldots, \vec{e}_{i_n}$ into $\vec{e}_1, \ldots, \vec{e}_n$.

Let's make some definition to capture these ideas.

Definition 9. An $n \times n$ **pattern** is a set of n pairs of indices $P = \{(i_1, j_1), \ldots, (i_n, j_n)\}$ such that $i_1, \ldots, i_n$ and $j_1, \ldots, j_n$ are each rearrangements of $1, \ldots, n$.

We say two elements (i_k, j_k) and (i_ℓ, j_ℓ) form an **inversion** in P if $i_k < i_\ell$ while $j_k > j_\ell$. We define the **signature of** P, denoted $\operatorname{sgn}(P)$, to be

$$\operatorname{sgn}(P) \stackrel{def}{=} (-1)^{\#\text{ of inversions of } P}.$$

If A is an $n \times n$ matrix, then we define the **product of** P **in** A, denoted $\operatorname{prod} P(A)$, by $\operatorname{prod} P(A) \stackrel{def}{=} a_{i_1,j_1} a_{i_2,j_2} \cdots a_{i_n,j_n}$.

Example 7. The 3×3 pattern $P = \{(2,1), (1,2), (3,3)\}$, corresponding to the entries in a 3×3 matrix marked by asterisks below, has one inversion (formed by $(2,1)$ and $(1,2)$).

$$\begin{bmatrix} & * & \\ * & & \\ & & * \end{bmatrix}$$

The 3×3 pattern $P = \{(3,1), (2,2), (1,3)\}$, corresponding to the entries in a 3×3 matrix marked by asterisks below, has three inversions (formed by $(3,1)$ and $(2,2)$, $(3,1)$ and $(1,3)$, and $(2,2)$ and $(1,3)$).

$$\begin{bmatrix} & & * \\ & * & \\ * & & \end{bmatrix}$$

The 3×3 pattern $P = \{(1,1), (2,2), (3,3)\}$ corresponding to the entries in a 3×3 matrix marked by asterisks below, has zero inversions.

$$\begin{bmatrix} * & & \\ & * & \\ & & * \end{bmatrix}.$$

Remark 31. By relabeling the indices in a pattern P so that $j_1 = 1,\ j_2 = 2, \ldots, j_n = n$, we see that P can be uniquely represented in the form $P = \{(i'_1, 1), \ldots, (i'_n, n)\}$. Similarly, by instead relabeling P so that $i_1 = 1, \ldots, i_n = n$, we see that P can be uniquely represented in the form $P = \{(1, j'_1), \ldots, (n, j'_n)\}$.

Note that the relevance of $\operatorname{prod} P(A)$ to (3) is clear, since the expression $a_{i_1,1} a_{i_2,2} \cdots a_{i_n,n} = \operatorname{prod} P(A)$, where $P = \{(i_1, 1), \ldots, (i_n, n)\}$.

Perhaps less obviously, the number of inversions of P can be used to characterize whether $D(\vec{e}_{i_1}, \ldots, \vec{e}_{i_n})$ is 1 or -1, as the next proposition shows.

Proposition 6. Let $P = \{(i_1, 1), \ldots, (i_n, n)\}$ be an $n \times n$ pattern. Then

$$D(\vec{e}_{i_1}, \ldots, \vec{e}_{i_n}) = \operatorname{sgn}(P).$$

This result follows by simply swapping pairs of columns on the left-hand-side until we obtain the identity matrix. Every swap will introduce a factor of (-1) because D is alternating. The trick is to do this in such a way that we can relate the number of swaps needed with the number of inversions of P.

Proof. The proposition follows from the following claim when $m = n$.

Claim: Let $1 \leq m \leq n$, let $i_1, \ldots, i_m$ be a rearrangement of $1, \ldots, m$ and let $P = \{(i_1, 1), \ldots, (i_m, m), (m+1, m+1), \ldots, (n, n)\}$. Then

$$D(\vec{e}_{i_1}, \ldots, \vec{e}_{i_m}, \vec{e}_{m+1}, \ldots, \vec{e}_n) = \operatorname{sgn}(P).$$

We prove this via induction on m. Assume $m = 1$. Because i_1 is a rearrangement of 1 we must have $i_1 = 1$ and therefore $P = \{(1,1), (2,2), \ldots, (n,n)\}$ has no inversions. But because D is a determinant function we have

$$D(\vec{e}_1, \ldots, \vec{e}_n) = D(I_n) = 1 = (-1)^0 = \operatorname{sgn}(P).$$

Now suppose that we have shown the result for $1 \leq m < n$, and suppose that $i_1, \ldots, i_{m+1}$ is a rearrangement of $1, \ldots, m+1$. Choose $1 \leq k \leq m+1$ such that $i_k = m+1$. Note that P has exactly $m+1-k$ inversions involving $(i_k, k) = (m+1, k)$ because $(m+1, k)$ is inverted with each entry $(i_{k+1}, k+1), \ldots, (i_{m+1}, m+1)$ because $m+1 > i_{k+1}, \ldots, i_{m+1}$ while $k < k+1, \ldots, m+1$, while $(m+1, k)$ does not form an inversion with any of the entries $(i_1, 1), \ldots, (i_{k-1}, k-1)$ or $(m+2, m+2), \ldots, (n, n)$. Moreover, all other inversions of P involve only entries in $\{(i_1, 1), \ldots, (i_{k-1}, k-1), (i_{k+1}, k+1), \ldots, (i_{m+1}, m+1)\}$.

Now consider the pattern

$$P' = \{(i'_1, 1), \ldots, (i'_m, m), (m+1, m+1), \ldots, (n, n)\},$$

where $i'_\ell = i_\ell$ if $\ell < k$, and $i'_\ell = i_{\ell+1}$ if $k \leq \ell \leq m$. Then note that

$$\#\text{ of inversions in } P' = (\#\text{ of inversions in } P) - (m+1-k),$$

and (by making $m+1-k$ swaps of consecutive pairs of columns) we have

$$D(\vec{e}_{i_1}, \ldots, \vec{e}_{i_{m+1}}, \vec{e}_{m+2}, \ldots, \vec{e}_n) = (-1)^{m+1-k} D(\vec{e}_{i'_1}, \ldots, \vec{e}_{i'_m}, \vec{e}_{m+1}, \ldots, \vec{e}_n).$$

Because $i'_1, \ldots, i'_m$ is a rearrangement of $1, \ldots, m$, we apply the induction hypothesis to see that

$$\begin{aligned} D(\vec{e}_{i_1}, \ldots, \vec{e}_{i_{m+1}}, \vec{e}_{m+2}, \ldots, \vec{e}_n) &= (-1)^{m+1-k} D(\vec{e}_{i'_1}, \ldots, \vec{e}_{i'_m}, \vec{e}_{m+1}, \ldots, \vec{e}_n) \\ &= (-1)^{m+1-k+(\#\text{ of inversions in } P')} \\ &= (-1)^{\#\text{ of inversions in } P} \\ &= \operatorname{sgn}(P). \end{aligned}$$

By The Principle of Mathematical Induction, the proof is complete. □

We can therefore recast (3) in terms of patterns as

$$\begin{aligned} D(A) &= \sum_{\{i_1, i_2 \ldots, i_n\} = \{1,2,\ldots,n\}} a_{i_1,1} a_{i_2,2} \cdots a_{i_n,n} D(\vec{e}_{i_1}, \vec{e}_{i_2}, \ldots, \vec{e}_{i_n}) \\ &= \sum_{n \times n \text{ patterns } P} (\operatorname{prod} P(A))(\operatorname{sgn}(P)) \end{aligned} \tag{4}$$

Because every determinant function on $n \times n$ matrices must be given by the formula in (4), we see that the formula in (4) is the only possible determinant function. We still need to verify that the formula in (4) actually *is* a determinant function. But we will check this now.

Existence of Determinants

We must show that the formula (4) actually gives a multilinear, alternating function with $\det(I_n) = 1$.

Lemma 2. The function

$$\det : M_{n\times n}(\mathbb{K}) \to \mathbb{K}, \qquad \det(A) \stackrel{def}{=} \sum_{n\times n \text{ patterns } P} (\operatorname{prod} P(A))(\operatorname{sgn}(P)) \tag{5}$$

is a determinant function.

Proof. We first show multilinearity. Suppose that $1 \le j \le n$, let $\vec{a}_1, \ldots, \vec{x}, \vec{y}, \ldots, \vec{a}_n \in \mathbb{K}^n$, and let $\lambda \in \mathbb{K}$. Then for each pattern $P = \{(i_1, 1), \ldots, (i_n, n)\}$ we have

$$\begin{aligned}
&\operatorname{prod} P(\vec{a}_1, \ldots, \vec{x}+\lambda\vec{y}, \ldots, \vec{a}_n) \\
&\qquad = a_{i_1,1} \cdots (x_{i_j} + \lambda y_{i_j}) \cdots a_{i_n,n} \\
&\qquad = a_{i_1,1} \cdots x_{i_j} \cdots a_{i_n,n} + \lambda(a_{i_1,1} \cdots y_{i_j} \cdots a_{i_n,n}) \\
&\qquad = \operatorname{prod} P(\vec{a}_1, \ldots, \vec{x}, \ldots, \vec{a}_n) + \lambda \operatorname{prod} P(\vec{a}_1, \ldots, \vec{y}, \ldots, \vec{a}_n),
\end{aligned}$$

whence it follows that

$$\begin{aligned}
&\det(\vec{a}_1, \ldots, \vec{x} + \lambda\vec{y}, \ldots, \vec{a}_n) \\
&\quad = \sum_{n\times n \text{ patterns } P} \Big(\operatorname{prod} P(\vec{a}_1, \ldots, \vec{x}, \ldots, \vec{a}_n) + \lambda \operatorname{prod} P(\vec{a}_1, \ldots, \vec{y}, \ldots, \vec{a}_n)\Big)(\operatorname{sgn}(P)) \\
&\quad = \sum_{n\times n \text{ patterns } P} (\operatorname{prod} P(\vec{a}_1, \ldots, \vec{x}, \ldots, \vec{a}_n))(\operatorname{sgn}(P)) \\
&\qquad\qquad + \lambda \sum_{n\times n \text{ patterns } P} (\operatorname{prod} P(\vec{a}_1, \ldots, \vec{y}, \ldots, \vec{a}_n))\,(\operatorname{sgn}(P)) \\
&\quad = \det(\vec{a}_1, \ldots, \vec{x}, \ldots, \vec{a}_n) + \lambda \det(\vec{a}_1, \ldots, \vec{y}, \ldots, \vec{a}_n).
\end{aligned}$$

For alternating, we start with the special case where we swap two adjacent columns. (Here we can assume that $n \ge 2$, since the alternating condition is vacuous in the case where $n = 1$.) Let $A \in M_{n\times n}(\mathbb{K})$, fix $1 \le k \le n-1$, and suppose that B is the matrix obtained by swapping columns k and $k+1$ of A. In particular, if $\vec{a}_1, \ldots, \vec{a}_k, \vec{a}_{k+1}, \vec{a}_n$ are the columns of A, then $\vec{a}_1, \ldots, \vec{a}_{k+1}, \vec{a}_k, \ldots, \vec{a}_n$ are the columns of B. Consider the $n \times n$ pattern $P = \{(i_1, 1), \ldots, (i_k, k), (i_{k+1}, k+1), \ldots, (i_n, n)\}$, and consider the 'swapped' pattern $P^s = \{(i_1, 1), \ldots, (i_{k+1}, k), (i_k, k+1), \ldots, (i_n, n)\}$. Because $i_k \ne i_{k+1}$, $P^s \ne P$.

Note that

$$\operatorname{prod} P(B) = \underbrace{b_{i_1,1}}_{a_{i_1,1}} \cdots \underbrace{b_{i_k,k}}_{a_{i_k,k+1}} \underbrace{b_{i_{k+1},k+1}}_{a_{i_{k+1},k}} \cdots \underbrace{b_{i_n,n}}_{a_{i_n,n}} = \operatorname{prod} P^s(A).$$

Also note that (i_{k+1}, k) and $(i_k, k+1)$ are inverted exactly when $i_{k+1} > i_k$, which happens exactly when (i_k, k) and $(i_{k+1}, k+1)$ are *not* inverted. Any inversions involving some (i_j, j) for $j \ne k, k+1$ in P are still present in P^s, and therefore the number of inversions of P^s and P differ by exactly 1. It follows that $\operatorname{sgn}(P^s) = -\operatorname{sgn}(P)$. Therefore we have

$$(\operatorname{prod} P(B))(\operatorname{sgn}(P)) + (\operatorname{prod} P^s(B))(\operatorname{sgn}(P^s)) = -(\operatorname{prod} P^s(A))(\operatorname{sgn}(P^s)) - (\operatorname{prod} P(A))(\operatorname{sgn}(P)).$$

Now note that since $(P^s)^s = P$, the swapping operation partitions the space of all $n \times n$ patterns into pairs of patterns, where each pair consists of a partition P and its swapped version P^s. It follows that

$$\begin{aligned}\det(B) &= \sum_{n\times n \text{ patterns } P} (\text{prod}(P(B)))(\text{sgn}(P)) \\ &= -\sum_{n\times n \text{ patterns } P} (\text{prod}(P(A)))(\text{sgn}(P)) = -\det(A).\end{aligned}$$

For the general case where B is the matrix obtained from A by swapping column k with column ℓ (where $k < \ell$), then we can transform A into B by making $(\ell - k) + (\ell - k - 1) = 2(\ell - k) - 1$ swaps of adjacent columns (first by moving the column k until it becomes column ℓ, and then taking the former column ℓ (which is not the $(\ell - 1)$-st column) and moving it so that it becomes the k-th column). Therefore $\det(B) = (-1)^{2(\ell-j)-1}\det(A) = -\det(A)$.

Finally, we note that the only $n\times n$ pattern P for which $\text{prod}\, P(I_n) \neq 0$ is $P = \{(1,1),(2,2),\ldots,(n,n)\}$. Therefore we have

$$\det(I_n) = \text{prod}(P(I_n))(\text{sgn}(P)) = 1(-1)^0 = 1.$$

□

Therefore, there is exactly one determinant function $\det : M_{n\times n}(\mathbb{K}) \to \mathbb{K}$, and it is given by the formula (5). This completes the proof of Theorem 7.

Lecture 6: More Determinants

Learning Objectives:

- Explore ways to compute the determinant of particular matrices.
- Show that the determinant of a matrix is the same as the determinant of its transpose.

At the start of today we finished out proof of Theorem 7. The formula (5) can be useful to compute the determinant of matrices with particularly simple structures, as the following example indicates.

Example 8. Let $A = \left[a_{j,k}\right] \in M_{n\times n}(\mathbb{K})$ be a triangular matrix. Then $\det(A) = a_{11}a_{22}\cdots a_{nn}$.

Suppose that A is upper-triangular (the proof when A is lower-triangular is similar). Let $P = \{(i_1, 1), \ldots, (i_n, n)\}$ be an $n \times n$ pattern. Suppose that $\text{prod}P(A) \neq 0$. We show by induction that $i_j = j$ for each $1 \leq j \leq n$. First note that if $i_1 > 1$, then since $a_{i_1,1} = 0$ we have $\text{prod}P(A) = 0$. Therefore $i_1 = 1$. Now suppose that $1 \leq k < n$ and that $i_j = j$ for each $1 \leq j \leq k$. Because $i_j = j$ for each $1 \leq j \leq k$, $i_{k+1} \geq k + 1$. If $i_{k+1} > k + 1$, then since $a_{i_{k+1},k+1} = 0$ we would have $\text{prod}P(A) = 0$. Therefore $i_{k+1} = k + 1$. By the Principle of Mathematical Induction, $i_j = j$ for each $1 \leq j \leq n$.

Therefore $P = P_0 = \{(1, 1), (2, 2), \ldots, (n, n)\}$. Because P_0 has no inversions, $\text{sgn}(P_0) = 1$. It follows that

$$\det(A) = \sum_{n\times n \text{ patterns } P} (\text{prod}\, P(A))(\text{sgn}(P)) = (\text{prod}\, P_0(A))(\text{sgn}(P_0)) = a_{11}a_{22}\cdots a_{nn}.$$

Remark 32. The last example shows that the determinant of an upper- or lower-triangular matrix is exactly the product of the entries on the main diagonal. Because diagonal matrices are also both upper- and lower-triangular, the same is true for diagonal matrices.

The Determinant and Transposes

One fundamental property of the determinant is that $\det(A^T) = \det(A)$. This turns out to be very easy to prove.

Theorem 8 (Determinants and Transposes)**.** For every $A \in M_{n\times n}(\mathbb{K})$, $\det(A^T) = \det(A)$.

Proof. For each $n\times n$ pattern $P = \{(i_1, j_1), \ldots, (i_n, j_n)\}$, define the transpose pattern $P^T \stackrel{def}{=} \{(j_1, i_1), \ldots, (j_n, i_n)\}$. Note that $(P^T)^T = P$ for each pattern P, and therefore the transpose operation is a bijection on the set of $n \times n$ patterns.

Note also that if (i_k, j_k) and (i_ℓ, j_ℓ) are inverted in P if and only if (j_k, i_k) and (j_ℓ, i_ℓ) are inverted in P^T. Therefore P^T and P have the same number of inversions, so that $\text{sgn}(P^T) = \text{sgn}(P)$.

Finally, note that

$$\text{prod}\, P^T(A^T) = a_{i_1,j_1}\cdots a_{i_n,j_n} = \text{prod}\, P(A).$$

We therefore see that

$$\det(A^T) = \sum_{n\times n \text{ patterns } P} (\text{prod}\, P^T(A^T))(\text{sgn}(P^T)) = \sum_{n\times n \text{ patterns } P} (\text{prod}\, P(A))(\text{sgn}(P)) = \det(A).$$

□

Lecture 7: Even More Determinants

Learning Objectives:

- Explore ways to compute the determinant of particular matrices.
- Investigate and exploit the relationship between determinants and elementary operations.
- Characterize invertibility of a matrix using the determinant.
- Compute the determinant of a product of square matrices.

Although the pattern formula for the determinant is useful (especially for proving theoretical results), it is not the version that is used most often in computations. Indeed, on your homework you will prove that we can compute the determinant of an $n \times n$ matrix A in terms of determinants of related $(n-1) \times (n-1)$ matrices called the **cofactors** of A.

To be precise, let $A = \begin{bmatrix} a_{i,j} \end{bmatrix} \in M_{n\times n}(\mathbb{K})$. The matrix $A_{i,j} \in M_{(n-1)\times(n-1)}(\mathbb{K})$ obtained by deleting the i-th row and j-th column of A is called the i,j-th **cofactor** of A. That is, $A_{i,j}$ is the matrix obtained by deleting the row and column of A containing $a_{i,j}$. Then we have the following result.

Theorem 9 (Cofactor Expansion)**.** Let $A \in M_{n\times n}(\mathbb{K})$. Then for each $1 \leq i \leq n$ we have

$$\det(A) = (-1)^{i+1}a_{i,1}\det(A_{i,1}) + (-1)^{i+2}a_{i,2}\det(A_{i,2}) + \cdots + (-1)^{i+n}a_{i,n}\det(A_{i,n}), \tag{6}$$

and for each $1 \leq j \leq n$ we have

$$\det(A) = (-1)^{1+j}a_{1,j}\det(A_{1,j}) + (-1)^{2+j}a_{2,j}\det(A_{2,j}) + \cdots + (-1)^{n+j}a_{n,j}\det(A_{n,j}). \tag{7}$$

Remark 33. Formula (6) is called the **cofactor expansion of** $\det(A)$ **along the** i**-th row**, and formula (7) is called the **cofactor expansion of** $\det(A)$ **along the** j**-th column**.

Proof. You will prove (6) on your homework by showing that the forumula on the right-hand-side of (6) is a determinant function, and therefore must be equal to *the* determinant function.

Formula (7) will then follow by combining the Determinants and Transposes Theorem with (6):

$$\begin{aligned}
\det(A) &= \det(A^T) \\
&= (-1)^{j+1}(a^T)_{j,1}\det((A^T)_{j,1}) + (-1)^{j+2}(a^T)_{j,2}\det((A^T)_{j,2}) + \cdots + (-1)^{j+n}(a^T)_{j,n}\det((A^T)_{j,n}) \\
&= (-1)^{1+j}a_{1,j}\det((A_{1,j})^T) + (-1)^{2+j}a_{2,j}\det((A_{2,j})^T) + \cdots + (-1)^{n+j}a_{n,j}\det((A_{n,j})^T) \\
&= (-1)^{1+j}a_{1,j}\det(A_{1,j}) + (-1)^{2+j}a_{2,j}\det(A_{2,j}) + \cdots + (-1)^{n+j}a_{n,j}\det(A_{n,j}).
\end{aligned}$$

□

We illustrate how to use the cofactor expansions to compute determinants.

Example 9. Note that $\det([a]) = a\det(I_1) = a$ for every $[a] \in M_{1\times 1}(\mathbb{K})$. Let's compute the determinants of 2×2 and 3×3 matrices as well. The formula for 2×2 matrices is

$$\begin{aligned}\det\left(\begin{bmatrix} a & b \\ c & d \end{bmatrix}\right) &= (-1)^{1+1} a \det(\begin{bmatrix} d \end{bmatrix}) + (-1)^{1+2} b \det(\begin{bmatrix} c \end{bmatrix}) && \text{(expand along row 1)} \\ &= ad - bc,\end{aligned}$$

which agrees with what you saw last quarter. For 3×3 matrices, we have (first expanding along column 2 just for illustration, and then applying our formula for 2×2 matrices)

$$\begin{aligned}\det\left(\begin{bmatrix} a & b & c \\ d & e & f \\ g & h & i \end{bmatrix}\right) &= (-1)^{1+2} b \det\left(\begin{bmatrix} d & f \\ g & i \end{bmatrix}\right) + (-1)^{2+2} e \det\left(\begin{bmatrix} a & c \\ g & i \end{bmatrix}\right) + (-1)^{3+2} h \det\left(\begin{bmatrix} a & c \\ d & f \end{bmatrix}\right) \\ &= -b(di - fg) + e(ai - gc) - h(af - cd) \\ &= aei + bfg + cdh - gec - hfa - idb.\end{aligned}$$

Determinants and Elementary Operations

The alternating and multilinear properties of the determinant—combined with with the relationship between determinants and transposes—allows us to study how elementary operations on rows and columns affect (or not) the determinant of a matrix. In particular, we have the following result.

Theorem 10 (Determinants and Elementary Operations)**.** Let $A \in M_{n\times n}(\mathbb{K})$.

(a) If B is obtained by swapping two columns (resp. rows) of A, then $\det(B) = -\det(A)$.

(b) If B is obtained by adding to one column (resp. row) of A a linear combination of the other columns (resp. rows), then $\det(B) = \det(A)$.

(c) If B is the matrix obtained from A by multiplying a single column (or row) of A by $\lambda \in \mathbb{K}$, then $\det(B) = \lambda\det(A)$.

Proof. It suffices to explain the results for columns, since the analogous results for rows follows by first writing $\det(B) = \det(B^T)$, then applying the result for columns, and then writing $\det(A^T) = \det(A)$.

For columns, (a) and (c) are immediate from (respectively) the alternating and multilinearity properties of the determinant. For (b), suppose that $A = \begin{bmatrix} \vec{a}_1 & \dots & \vec{a}_n \end{bmatrix}$. Without loss of generality we will assume that we are adding a linear combination of $\vec{a}_2, \dots, \vec{a}_n$ to the first column. Let $c_2, \dots, c_n \in \mathbb{K}$, and note that if the first column of B is $\vec{a}_1 + c_2\vec{a}_2 + \cdots + c_n\vec{a}_n$,

$$\begin{aligned}\det(B) &= \det(\vec{a}_1 + c_2\vec{a}_2 + \cdots + c_n\vec{a}_n, \vec{a}_2, \dots, \vec{a}_n) \\ &= \det(\vec{a}_1, \vec{a}_2, \dots, \vec{a}_n) + \sum_{k=2}^{n} c_k \det(\vec{a}_k, \vec{a}_2, \dots, \vec{a}_n) \\ &= \det(\vec{a}_1, \vec{a}_2, \dots, \vec{a}_n) + \sum_{k=2}^{n} c_k 0 \\ &= \det(A),\end{aligned}$$

where in the second step we used multilinearity of the determinant, and in the third step we used the alternating property of the determinant. □

Example 10. In practice, it is sometimes helpful to use elementary row operations to reduce a matrix to, say, an upper-triangular matrix before computing its determinant. For example, we have

$$\begin{aligned}
\det\left(\begin{bmatrix} 0 & -7 & 5 & 3 \\ 1 & 1 & 2 & 1 \\ 1 & 1 & 2 & -1 \\ 1 & 1 & 0 & 3 \end{bmatrix}\right) &= \det\left(\begin{bmatrix} 0 & -7 & 5 & 3 \\ 1 & 1 & 2 & 1 \\ 0 & 0 & 0 & -2 \\ 0 & 0 & -2 & 2 \end{bmatrix}\right) && (R_3 \to R_3 - R_2,\ R_4 \to R_4 - R_2) \\
&= (-1)^2 \det\left(\begin{bmatrix} 1 & 1 & 2 & 1 \\ 0 & -7 & 5 & 3 \\ 0 & 0 & -2 & 2 \\ 0 & 0 & 0 & -2 \end{bmatrix}\right) && (R_1 \leftrightarrow R_2,\ R_3 \leftrightarrow R_4) \\
&= (-1)^2 \cdot 1 \cdot (-7) \cdot (-2) \cdot (-2) && (\text{Upper-Triangular}) \\
&= -28.
\end{aligned}$$

We apply the computational observations in the previous example to show that determinants are tied to invertibility.

Theorem 11 (Invertibility, cont'd.)**.** Let $A \in M_{n\times n}(\mathbb{K})$. Then A is invertible if, and only if, $\det(A) \neq 0$.

Proof. Let $B = \text{rref}(A)$. Suppose that in the application of Gauss-Jordan elimination transforming A into B we make m row swaps and divide the rows of A by nonzero scalars $\lambda_1, \ldots, \lambda_k$. Then we have

$$\det(A) = (-1)^m \lambda_1 \cdots \lambda_k \det(B) = (-1)^m \lambda_1 \cdots \lambda_k b_{11} b_{22} \cdots b_{nn},$$

where in the final step we used that $B = \text{rref}(A)$ is upper triangular. If A is invertible then the Invertibility Theorem implies that $B = I_n$, so that $b_{11} = \cdots = b_{nn} = 1$ and therefore $\det(A) \neq 0$. On the other hand, if A is not invertible then the Invertibility Theorem implies that the columns of A do not span $\mathbb{K}^n$, so that $b_{nn} = 0$ and therefore $\det(A) = 0$. □

Determinants and Products

The determinant plays very well with matrix products, in the following sense.

Theorem 12 (Determinants and Products)**.** For every $A, B \in M_{n\times n}(\mathbb{K})$, $\det(AB) = \det(A)\det(B)$.

Proof. Although your book proves this result using elementary row operations, we will given an elegant proof using the uniqueness of the determinant function. First note that if $A \in M_{n\times n}(\mathbb{K})$ is not invertible and $B \in M_{n\times n}(\mathbb{K})$ then, by a homework problem from last quarter, AB is not invertible and therefore $\det(AB) = 0 = 0\det(B) = \det(A)\det(B)$.

Now let $A \in M_{n\times n}(\mathbb{K})$ and suppose that A is invertible, so that $\det(A) \neq 0$. Define

$$D : M_{n\times n}(\mathbb{K}) \to \mathbb{K}, \qquad D(B) \stackrel{def}{=} \frac{\det(AB)}{\det(A)}.$$

We claim that D is a determinant function. First note that $D(I_n) = \dfrac{\det(AI_n)}{\det(A)} = \dfrac{\det(A)}{\det(A)} = 1$.

Let $B = \begin{bmatrix} \vec{b}_1 & \cdots & \vec{b}_n \end{bmatrix} \in M_{n\times n}(\mathbb{K})$. Suppose that two of the columns of B are equal. Then two of the columns $A\vec{b}_1, \ldots, A\vec{b}_n$ of AB are equal, so that $D(B) = \dfrac{\det(AB)}{\det(A)} = \dfrac{0}{\det(A)} = 0$.

Finally, we turn to multilinearity. Suppose that $1 \leq j \leq n$, and that $\vec{x}, \vec{y} \in \mathbb{K}^n$, and that $\lambda \in \mathbb{K}$. Then by the multilinearity of the determinant,

$$\begin{aligned} D(\vec{b}_1, \ldots, \vec{x} + \lambda\vec{y}, \ldots, \vec{b}_n) &= \frac{\det(A\vec{b}_1, \ldots, A(\vec{x} + \lambda\vec{y}), \ldots, A\vec{b}_n)}{\det(A)} \\ &= \frac{\det(A\vec{b}_1, \ldots, A\vec{x} + \lambda A\vec{y}, \ldots, A\vec{b}_n)}{\det(A)} \\ &= \frac{\det(A\vec{b}_1, \ldots, A\vec{x}, \ldots, A\vec{b}_n)}{\det(A)} + \lambda\frac{\det(A\vec{b}_1, \ldots, A\vec{y}, \ldots, A\vec{b}_n)}{\det(A)} \\ &= D(\vec{b}_1, \ldots, \vec{x}, \ldots, \vec{b}_n) + \lambda D(\vec{b}_1, \ldots, \vec{y}, \ldots, \vec{b}_n). \end{aligned}$$

We conclude that D is a determinant function, and therefore that $D = \det$ by uniqueness of the determinant function. But then for each $B \in M_{n\times n}(\mathbb{K})$ we have $\det(AB) = \det(A)D(B) = \det(A)\det(B)$, and the proof is complete. □

As an immediate consequence, we see that the original motivation for the determinant (as a scalar associated with 2×2 matrices that determined when they are invertible) generalizes to determinants for matrices of all sizes.

Corollary 5. Let $A \in M_{n\times n}(\mathbb{K})$. If A is invertible, then $\det(A^{-1}) = (\det(A))^{-1}$.

Proof. We simply note that

$$1 = \det(I_n) = \det(A^{-1}A) = \det(A^{-1})\det(A).$$

□

Corollary 6. Let $A \in M_{n\times n}(\mathbb{R})$. If A is orthogonal, then $|\det(A)| = 1$, so that $\det(A) = 1$ or $\det(A) = -1$.

Proof. We simply note that

$$1 = \det(I_n) = \det(A^TA) = \det(A^T)\det(A) = (\det(A))^2.$$

□

Remark 34. We already mentioned that if $A \in M_{n\times n}(\mathbb{R})$ is orthogonal, then the linear transformation described by A can be characterized as either a rotation or a reflection. It turns out that the determinant is the simplest way to distinguish between these two cases: if A is orthogonal and $\det(A) = 1$ then A is a rotation, while if A is orthogonal and $\det(A) = -1$ then A is a reflection.

Lecture 8: Determinants and Volume

Learning Objectives:

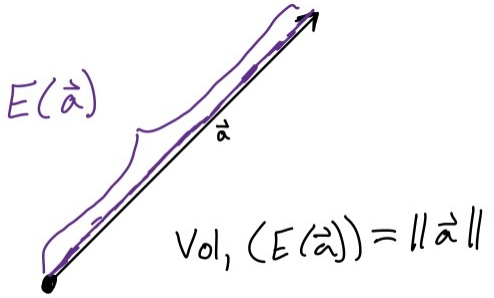
- Relate the determinant to volume and, by extension, the effect of linear transformations on volume of regions.

We finally turn to the connection between determinants and volumes. To understand this, we introduce a natural geometric region associated to a collection of vectors in $\mathbb{R}^n$.

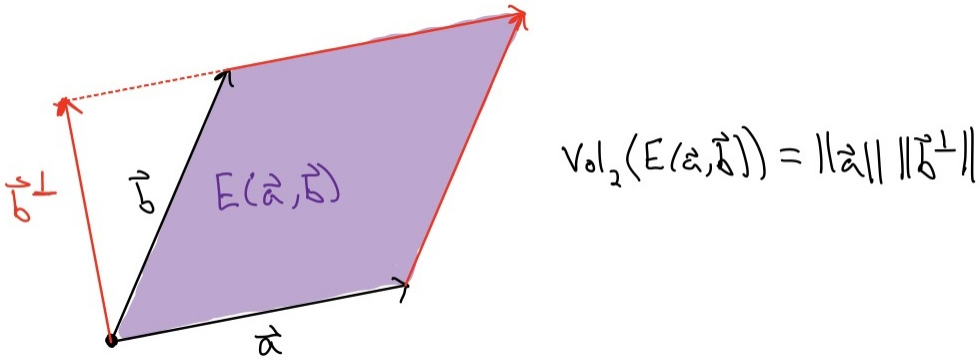
Definition 10. A collection of k vectors $\vec{v}_1, \ldots, \vec{v}_k \in \mathbb{R}^n$ determines a **parallelotope** $E(\vec{v}_1, \ldots, \vec{v}_k)$ defined by

$$E(\vec{v}_1, \ldots, \vec{v}_k) \stackrel{def}{=} \{c_1\vec{v}_1 + \cdots + c_k\vec{v}_k \ : \ c_1, \ldots, c_k \in [0,1]\}.$$

Example 11. When $m = 1$, the parallelotope determined by a single vector $\vec{a} \in \mathbb{R}^n$ is the line segment connecting the initial and terminal points of $\vec{a}$. The length ("1-volume") of this parallelotope is $\|\vec{a}\|$:

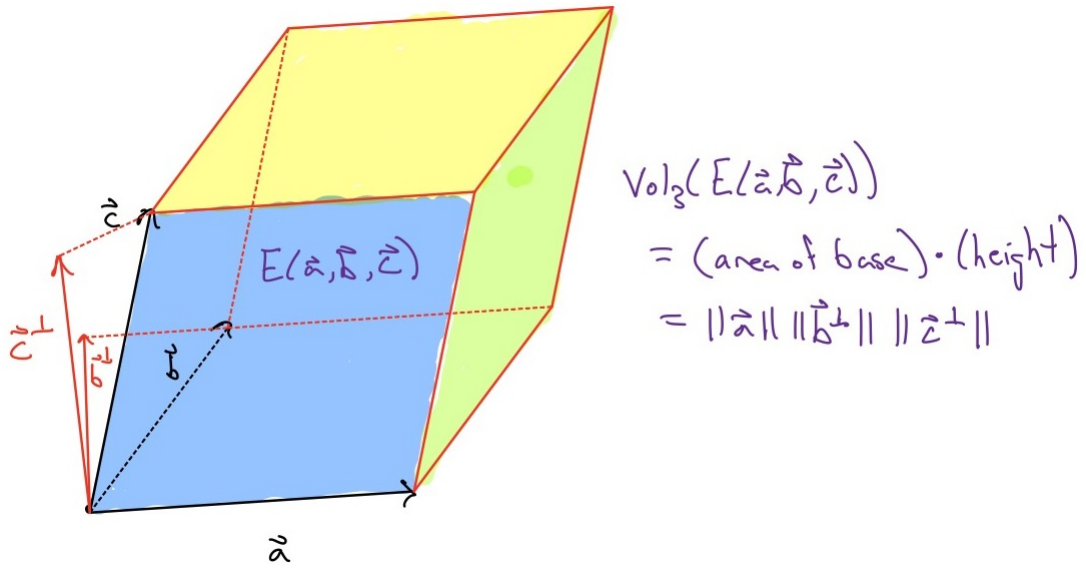


When $m = 2$ (and $n \geq 2$), the parallelotope determined by a pair of vectors $\vec{a}, \vec{b} \in \mathbb{R}^n$ is a parallelogram with one vertex at $\vec{0}$ and adjacent sides parallel to (and the same length as) $\vec{a}$ and $\vec{b}$. The area ("2-volume") of this parallelotope is the area of this parallelogram, which can be computed as $\|\vec{a}\|\|\vec{b}^{\perp}\|$, where $\vec{b}^{\perp} = \vec{b} - \text{proj}_{\text{span}(\vec{a})}\vec{b}$:



When $m = 3$ (and $n \geq 3$), the parallelotope determined by a trio of vectors $\vec{a}, \vec{b}, \vec{c} \in \mathbb{R}^n$ is a parallelepiped with one vertex at $\vec{0}$ and three adjacent edges parallel to (and the same length as) $\vec{a}$, $\vec{b}$, and $\vec{c}$. The volume

("3-volume") of this parallelotope can be computed using as "(area of base)x(height)", where the base is the parallelogram determined by two of the vectors (say $\vec{a}$ and $\vec{b}$), and the height is $\vec{c}^{\perp} = \vec{c} - \text{proj}_{\text{span}(\vec{a},\vec{b})}\vec{c}$. Therefore, if $\vec{b}^{\perp} = \vec{b} - \text{proj}_{\text{span}(\vec{a})}\vec{b}$, then the volume is $\|\vec{a}\|\|\vec{b}^{\perp}\|\|\vec{c}^{\perp}\|$:



We can use the ideas in the previous examples to define a reasonable notion of m-volume of a collection of vectors in $\mathbb{R}^n$.

Definition 11. Let $\vec{v}_1, \ldots, \vec{v}_m \in \mathbb{R}^n$. For each $k = 2, 3, \ldots, m$, define

$$\vec{v}_k^{\perp} \stackrel{def}{=} \vec{v}_k - \text{proj}_{\text{span}(\vec{v}_1,\ldots,\vec{v}_{k-1})}(\vec{v}_k).$$

Then the m-volume of the parallelotope $E(\vec{v}_1, \ldots, \vec{v}_m)$ determined by $\vec{v}_1, \ldots, \vec{v}_m$ is defined by

$$\text{Vol}_m(E(\vec{v}_1, \ldots, \vec{v}_m)) \stackrel{def}{=} \|\vec{v}_1\|\|\vec{v}_2^{\perp}\| \cdots \|\vec{v}_m^{\perp}\|.$$

Remark 35. Note that the definition of m-volume for parallelotopes determined by m vectors in $\mathbb{R}^n$ does not require that $m \leq n$, as suggested by the opening examples. Note, though, that if $m > n$ then the list $\vec{v}_1, \ldots, \vec{v}_m$ is linearly dependent, and therefore there exists k with $\vec{v}_k \in \text{span}(\vec{v}_1, \ldots, \vec{v}_{k-1})$, so that $\vec{v}_k^{\perp} = \vec{0}$, and therefore $\text{Vol}_m(E(\vec{v}_1, \ldots, \vec{v}_m)) = 0$.

Our main theorem connecting the determinant to volume is as follows.

Theorem 13 (Determinant and Volume)**.** Let $A = \begin{bmatrix} \vec{v}_1 & \cdots & \vec{v}_m \end{bmatrix} \in M_{n\times m}(\mathbb{R})$. Then

$$\text{Vol}_m(E(\vec{v}_1, \ldots, \vec{v}_m)) = \sqrt{\det(A^T A)}.$$

In particular, when $m = n$ we have $\text{Vol}_n(E(\vec{v}_1, \ldots, \vec{v}_n)) = |\det(A)|$.

Proof. Let B be the matrix with columns $\vec{v}_1, \vec{v}_2^{\perp}, \ldots, \vec{v}_m^{\perp}$. Because

$$\vec{v}_k^{\perp} = \vec{v}_k + (\text{linear combination of } \vec{v}_1, \ldots, \vec{v}_{k-1}) \quad \text{for each } 2 \leq k \leq m,$$

there is an upper triangular matrix $U \in M_{m\times m}(\mathbb{R})$ with $u_{k,k} = 1$ for each $1 \leq k \leq m$, and such that $B = AU$. Indeed, $U = \left[u_{j,k}\right]$ satisfies, for each $1 \leq k \leq m$,

$$\vec{v}_k^{\perp} = A\vec{u}_k = u_{1,k}\vec{v}_1 + \cdots + u_{k-1,k}\vec{v}_{k-1} + 1\vec{v}_k + 0\vec{v}_{k+1} + \cdots + 0\vec{v}_m.$$

Because $\det(U) = u_{1,1}\cdots u_{m,m} = 1 \neq 0$, U is invertible and therefore $A = BU^{-1}$. On the other hand, note that

$$B^TB = \begin{bmatrix} \vec{v}_1^T \\ (\vec{v}_2^{\perp})^T \\ \vdots \\ (\vec{v}_m^{\perp})^T \end{bmatrix} \begin{bmatrix} \vec{v}_1 & \vec{v}_2^{\perp} & \cdots & \vec{v}_m^{\perp} \end{bmatrix} = \begin{bmatrix} \|\vec{v}_1\|^2 & 0 & \cdots & 0 \\ 0 & \|\vec{v}_2^{\perp}\|^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \|\vec{v}_m^{\perp}\|^2 \end{bmatrix}.$$

We therefore have

$$\begin{aligned} \sqrt{\det(A^TA)} &= \sqrt{\det((U^{-1})^TB^TBU^{-1})} \\ &= \sqrt{\det((U^{-1})^T)\det(B^TB)\det(U^{-1})} \\ &= \sqrt{\det((U^{-1}))\det(B^TB)\det(U^{-1})} \\ &= \sqrt{(\det(U))^{-1}\det(B^TB)(\det(U))^{-1}} \\ &= \sqrt{\det(B^TB)} \\ &= \sqrt{\|\vec{v}_1\|^2\|\vec{v}_2^{\perp}\|^2\cdots\|\vec{v}_m^{\perp}\|^2} \\ &= \|\vec{v}_1\|\|\vec{v}_2^{\perp}\|\cdots\|\vec{v}_m^{\perp}\| \\ &= \mathrm{Vol}_m(E(\vec{v}_1,\ldots,\vec{v}_m)). \end{aligned}$$

In the case where $m = n$, then $A, A^T \in M_{n\times n}(\mathbb{R})$, so that

$$\sqrt{\det(A^TA)} = \sqrt{\det(A^T)\det(A)} = \sqrt{(\det(A))^2} = |\det(A)|.$$

□

The previous theorem can be extrapolated into a geometric interpretation of how linear transformations expand (or compress) volumes.

Corollary 7. Let $T : \mathbb{R}^n \to \mathbb{R}^n$ be linear with standard matrix $A \in M_{n\times n}(\mathbb{R})$. If $P = E(\vec{b}_1,\ldots,\vec{b}_n)$ for $\vec{b}_1,\ldots,\vec{b}_n \in \mathbb{R}^n$, then $T(P)$ is a parallelotope and

$$\mathrm{Vol}_n(T(P)) = |\det(A)|\mathrm{Vol}_n(P).$$

Proof. First note that

$$\begin{aligned} T(P) &= \{T(c_1\vec{b}_1 + \cdots + c_n\vec{b}_n) \ : \ c_1,\ldots,c_n \in [0,1]\} \\ &= \{c_1T(\vec{b}_1) + \cdots + c_nT(\vec{b}_n) \ : \ c_1,\ldots,c_n \in [0,1]\} \\ &= \{c_1A\vec{b}_1 + \cdots + c_nA\vec{b}_n \ : \ c_1,\ldots,c_n \in [0,1]\} \\ &= E(A\vec{b}_1,\ldots,A\vec{b}_n) \end{aligned}$$

is the parallelotope determined by $A\vec{b}_1,\ldots,A\vec{b}_n$.

Let B be the matrix with columns $\vec{b}_1, \ldots, \vec{b}_n$. Then

$$\begin{aligned}
\text{Vol}_n(T(P)) &= \text{Vol}_n(E(A\vec{b}_1, \ldots, A\vec{b}_n)) \\
&= |\det(A\vec{b}_1, \ldots, A\vec{b}_n)| \\
&= |\det(AB)| \\
&= |\det(A)\det(B)| \\
&= |\det(A)||\det(B)| \\
&= |\det(A)|\text{Vol}_n(E(\vec{b}_1, \ldots, \vec{b}_n)) \\
&= |\det(A)|\text{Vol}_n(P).
\end{aligned}$$

□

Definition 12. The number $|\det(A)|$ is called the **expansion factor** of the linear map $T : \mathbb{R}^n \to \mathbb{R}^n$.

Remark 36. Once we have a reasonable way to measure n-volume of general (but reasonable) regions in $\mathbb{R}^n$ (e.g. integration), then we will also see that if $\Omega \subseteq \mathbb{R}^n$ is such a region then $\text{Vol}_n(T(\Omega)) = |\det(A)|\text{Vol}_n(\Omega)$. In other words, the expansion factor of a linear map exactly captures the effect of the linear map on the volumes of regions.

In light of the previous two results, we can add two lines to the Invertibility Theorem. We'll also include the business about the determinant just to show where it should go.

Theorem 14 (Invertibility, cont'd.)**.** Let $A = \begin{bmatrix} \vec{a}_1 & \cdots & \vec{a}_n \end{bmatrix} \in M_{n\times n}(\mathbb{K})$, and let $T : \mathbb{K}^n \to \mathbb{K}^n$ be the linear transformation $T(\vec{x}) = A\vec{x}$. Then the following are equivalent.

(a-1) A is invertible.

⋮

(a-11) $\det(A) \neq 0$.

(a-12) (When $\mathbb{K} = \mathbb{R}$) $\text{Vol}_n(E(\vec{a}_1, \ldots, \vec{a}_n)) > 0$.

(a-13) (When $\mathbb{K} = \mathbb{R}$) There is a parallelotope $P \overset{def}{=} E(\vec{b}_1, \ldots, \vec{b}_n) \subset \mathbb{R}^n$ such that $\text{Vol}_n(T(P)) > 0$.

⋮

Proof. We have already shown that (a-1) and (a-11) are equivalent. Suppose that $\mathbb{K} = \mathbb{R}$. Because $\text{Vol}_n(E(\vec{a}_1, \ldots, \vec{a}_n)) = |\det(A)|$, (a-11) holds if, and only if, (a-12) holds. Note also that taking $P = E(\vec{e}_1, \ldots, \vec{e}_n)$ gives $\text{Vol}_n(T(P)) = |\det(A)|\text{Vol}_n(P) = |\det(A)|$, and therefore (a-13) holds if, and only if, (a-11) holds. □

Lecture 9: Eigenvalues and Eigenvectors

Learning Objectives:

- Explain what it means for a linear transformation (or square matrix) to be diagonalizable.
- Compute eigenvalues and eigenvectors for given linear transformations.
- Characterize invertibility in terms of eigenvalues.

Determinants of Linear Transformations

As one final result from determinants, we show that we can actually define the determinant of a linear map $T : V \to V$ from a finite-dimensional vector space V to itself.

We start by recalling some definitions and results from the end of last quarter. We quickly summarize some of the definitions and results surrounding coordinates and change of bases, but you should read the end of last quarter's notes for a full review!

Let V be a finite-dimensional vector space over $\mathbb{K}$, let $T : V \to V$ be linear, let $\mathscr{B} = (v_1, \ldots, v_n)$ and $\mathscr{C} = (w_1, \ldots, w_n)$ be bases for V.

- For each $v \in V$, $\left[v\right]_{\mathscr{B}} \stackrel{def}{=} \begin{bmatrix} c_1 \\ \vdots \\ c_n \end{bmatrix} \in \mathbb{K}^n$ is the $\mathscr{B}$-coordinate vector of v, where the entries $c_1, \ldots, c_n$ are the unique scalars satisfying $v = c_1 v_1 + \cdots + c_n v_n$.
  In the special case where $V = \mathbb{K}^n$ then we have $\vec{v} = \begin{bmatrix} \vec{v}_1 & \cdots & \vec{v}_n \end{bmatrix} \left[\vec{v}\right]_{\mathscr{B}}$.
- The map $\left[\cdot\right]_{\mathscr{B}} : V \to \mathbb{K}^n$ is a linear isomorphism.
- The matrix $\left[T\right]_{\mathscr{B}} = \begin{bmatrix} \left[T(v_1)\right]_{\mathscr{B}} & \cdots & \left[T(v_n)\right]_{\mathscr{B}} \end{bmatrix} \in M_{n\times n}(\mathbb{K})$ is the unique matrix satisfying $\left[T\right]_{\mathscr{B}} \left[v\right]_{\mathscr{B}} = \left[T(v)\right]_{\mathscr{B}}$ for every $v \in V$. It is the matrix of $\left[\cdot\right]_{\mathscr{B}} \circ T \circ \left[\cdot\right]_{\mathscr{B}}^{-1} : \mathbb{K}^n \to \mathbb{K}^n$, and is called the matrix of T relative to $\mathscr{B}$.
- For every $v \in V$ we have $\left[v\right]_{\mathscr{C}} = S_{\mathscr{C}\leftarrow\mathscr{B}} \left[v\right]_{\mathscr{B}}$, where $S_{\mathscr{C}\leftarrow\mathscr{B}} = \begin{bmatrix} \left[v_1\right]_{\mathscr{C}} & \cdots & \left[v_n\right]_{\mathscr{C}} \end{bmatrix}$. The change-of-basis matrix $S_{\mathscr{C}\leftarrow\mathscr{B}}$ is invertible and its inverse is $S_{\mathscr{B}\leftarrow\mathscr{C}}$.

Definition 13. Recall that $A, B \in M_{n\times n}(\mathbb{K})$ are **similar** if there is an invertible matrix $S \in M_{n\times n}(\mathbb{K})$ with $AS = SB$ (or equivalently $SA = BS$ or $A = S^{-1}BS$ or $B = S^{-1}AS$).

Remark 37. Suppose that $A, B \in M_{n\times n}(\mathbb{K})$ are similar. Then A and B share many properties (left as an exercise last quarter): their column spaces have the same dimension, their nullspaces have the same dimension, they have the same rank, A^k and B^k are similar for every $k \in \mathbb{N}$, A is nilpotent if and only if B is nilpotent, and A is invertible if and only if B is invertible (and, in this case, A^{-1} and B^{-1} are similar).

We can now add to this list that $\det(A) = \det(B)$. To see this, let $S \in M_{n\times n}(\mathbb{K})$ be an invertible matrix with $A = S^{-1}BS$. Then since S is invertible,

$$\det(A) = \det(S^{-1})\det(B)\det(S) = (\det(S))^{-1}\det(B)\det(S) = \det(B).$$

There is a simple explanation of why similar matrices share so many properties (which also explains why we care about similarity in the first place): similar matrices represent the same linear transformation in different coordinate systems, in the following sense.

Theorem 15 (Characterization of Similarity)**.** Let V be a finite-dimensional vector space over $\mathbb{K}$ with $n = \dim(V)$, let $T : V \to V$ be a linear transformation, let $\mathscr{B}$ be a basis for V, and let $C \in M_{n\times n}(\mathbb{K})$. Then C and $[T]_{\mathscr{B}}$ are similar if, and only if, there is a basis $\mathscr{C}$ for V such that $C = [T]_{\mathscr{C}}$.

Proof. Suppose first that there is a basis $\mathscr{C}$ for V such that $C = [T]_{\mathscr{C}}$. Then (as shown last quarter) $\left[T\right]_{\mathscr{C}} S_{\mathscr{C}\leftarrow\mathscr{B}} = S_{\mathscr{C}\leftarrow\mathscr{B}} \left[T\right]_{\mathscr{B}}$, so that $C = \left[T\right]_{\mathscr{C}}$ and $\left[T\right]_{\mathscr{B}}$ are similar.

Suppose that C and $[T]_{\mathscr{B}}$ are similar, so that there is an invertible matrix S such that $C = S^{-1}[T]_{\mathscr{B}}S$. Let $\vec{s}_1, \ldots, \vec{s}_n$ be the columns of S. Because S is invertible, $\vec{s}_1, \ldots, \vec{s}_n$ is a basis for $\mathbb{K}^n$. Because $\left[\cdot\right]_{\mathscr{B}} : V \to \mathbb{K}^n$ is an isomorphism, there is a basis $\mathscr{C} = (w_1, \ldots, w_n)$ for V such that $[w_k]_{\mathscr{B}} = \vec{s}_k$ for each $1 \leq k \leq n$. Then for each $1 \leq k \leq n$ we have $S_{\mathscr{C}\leftarrow\mathscr{B}}\vec{s}_k = S_{\mathscr{C}\leftarrow\mathscr{B}} \left[w_k\right]_{\mathscr{B}} = \left[w_k\right]_{\mathscr{C}} = \vec{e}_k$ and therefore $S_{\mathscr{C}\leftarrow\mathscr{B}}S = I_n$. It follows that $S = S^{-1}_{\mathscr{C}\leftarrow\mathscr{B}} = S_{\mathscr{B}\leftarrow\mathscr{C}}$ and $S^{-1} = S_{\mathscr{C}\leftarrow\mathscr{B}}$, so that

$$C = S^{-1} \left[T\right]_{\mathscr{B}} S = S_{\mathscr{C}\leftarrow\mathscr{B}} \left[T\right]_{\mathscr{B}} S_{\mathscr{B}\leftarrow\mathscr{C}} = \left[T\right]_{\mathscr{C}}.$$

□

We can therefore make the following definition.

Definition 14. Let V be a finite-dimensional vector space over $\mathbb{K}$, and let $T : V \to V$ be linear. Define $\det(T) \stackrel{def}{=} \det(\left[T\right]_{\mathscr{B}})$, where $\mathscr{B}$ is any basis for V.

Remark 38. Note that $\det(T)$ is well-defined because, by the Characterization of Similarity, $\det(\left[T\right]_{\mathscr{B}})$ will have the same value no matter which basis $\mathscr{B}$ for V we use (and therefore does not depend on our choice of basis).

On your homework you will practice computing the determinants of various linear transformations from a finite-dimensional vector space over $\mathbb{K}$ to itself.

Diagonalizability

At this point we now understand how the dot product and the determinant capture the geometric notions of angles, lengths, and (dimension-appropriate) volume. As a final lead-up to our last major linear algebra result of the year, we build on our investigation of linear transformations and bases from last quarter by considering the following central question: if V is a finite-dimensional vector space over $\mathbb{K}$ and $T : V \to V$ is linear, is there a basis $\mathscr{B}$ for V that makes studying T particularly simple, and (if so) how do we find this basis? There are many interpretations for what "particularly simple" means, but we will mean that the matrix $\left[T\right]_{\mathscr{B}}$ of T relative to $\mathscr{B}$ is diagonal. To answer this we will need results covered about coordinates last quarter, so please review as needed! The material is also in Chapter 4 of your book, but your book's treatment is a little light for our purposes.

We start with a definition.

Definition 15. Let V be a finite-dimensional vector space over $\mathbb{K}$, and let $T : V \to V$ be a linear transformation. We call T **diagonalizable** if there is a basis $\mathscr{B}$ for V such that $\left[T\right]_{\mathscr{B}} \in M_{n\times n}(\mathbb{K})$ is diagonal. We **diagonalize** T by producing such a basis for V.

Although diagonalizability is a concept that applies to linear transformations, it will be helpful to have an algebraic notion of diagonalizability for matrices.

Definition 16. Let $A \in M_{n\times n}(\mathbb{K})$. We call A **diagonalizable** if it is similar to a diagonal matrix. We **diagonalize** A by producing matrices $D, S \in M_{n\times n}(\mathbb{K})$ with D diagonal, S invertible, and $AS = SD$.

Just as for invertibility, the notions of diagonalizability for matrices and linear transformations are compatible in the sense that a linear transformation is diagonalizable if, and only if, its matrix with respect to any particular (and therefore every!) basis is diagonalizable.

Proposition 7. Let V be a finite-dimensional vector space over $\mathbb{K}$, let $T : V \to V$ be a linear transformation, and let $\mathscr{B}$ be a basis for V. Then T is diagonalizable if, and only if, $\left[T\right]_{\mathscr{B}}$ is diagonalizable.

Proof. Suppose T is diagonalizable. There is a basis $\mathscr{C}$ for V such that $[T]_{\mathscr{C}}$ is diagonal. Then $[T]_{\mathscr{B}}$ and $[T]_{\mathscr{C}}$ are similar, so that $[T]_{\mathscr{B}}$ is diagonalizable.

Suppose $[T]_{\mathscr{B}}$ is diagonalizable. Let $C \in M_{n\times n}(\mathbb{K})$ be a diagonal matrix that is similar to $[T]_{\mathscr{B}}$. By the previous theorem there is a basis $\mathscr{C}$ of V such that $[T]_{\mathscr{C}} = C$ is diagonal, so that T is diagonalizable. □

Therefore, going forward, we will often conflate diagonalizability of a matrix with diagonalizability of any linear transformation that it represents. To study diagonalizability, we need to determine when a linear transformation (or matrix) behaves like a diagonal matrix. To facilitate this we introduce the notions of eigenvalues and eigenvectors.

Eigenvalues and Eigenvectors

Remark 39. Diagonalizable linear maps have a particularly accessible geometric description. Suppose V is a finite-dimensional vector space over $\mathbb{K}$, and that $T : V \to V$ is a diagonalizable linear transformation. Then for some basis $\mathscr{B} = (v_1, \ldots, v_n)$ for V we must have $[T]_{\mathscr{B}}$ diagonal. If we write

$$[T]_{\mathscr{B}} = \begin{bmatrix} \lambda_1 & 0 & 0 & \cdots & 0 \\ 0 & \lambda_2 & 0 & \cdots & 0 \\ 0 & 0 & \lambda_3 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \lambda_n \end{bmatrix},$$

then for each $1 \leq k \leq n$ we have

$$[T(v_k)]_{\mathscr{B}} = [T]_{\mathscr{B}}[v_k]_{\mathscr{B}} = [T]_{\mathscr{B}}\vec{e}_k = \lambda_k\vec{e}_k = \lambda_k[v_k]_{\mathscr{B}} = [\lambda_k v_k]_{\mathscr{B}}, \quad \text{so that } T(v_k) = \lambda_k v_k.$$

In particular, T simply scales each of the basis vectors $v_1, \ldots, v_n$ (perhaps by different amounts)!

In light of the above remark, diagonalizing $T : V \to V$ boils down to finding a basis $v_1, \ldots, v_n$ for V such that there are scalars $\lambda_1, \ldots, \lambda_n \in \mathbb{K}$ with $T(\vec{v}_k) = \lambda_k \vec{v}_k$ for $1 \leq k \leq n$. (Of course, we can only hope do this if such a basis actually exists.) With this in mind, we make a few definitions.

Definition 17. Let V be a finite-dimensional vector space over $\mathbb{K}$, and let $T : V \to V$ be linear. We call $v \in V$ an **eigenvector** of T if $v \neq 0$ and if there is a scalar $\lambda \in \mathbb{K}$ with $T(v) = \lambda v$. Similarly, $\lambda \in \mathbb{K}$ is called an **eigenvalue** of T if there is $v \in V$ with $v \neq 0$ such that $T(v) = \lambda v$. In either case, λ is called the **eigenvalue associated to** v.
An **eigenbasis** for T is a basis $\mathscr{B} = (v_1, \ldots, v_n)$ for V such that each v_k is an eigenvector of T.

Remark 40. Note that if $v_1, \ldots, v_n$ is an eigenbasis for $T : V \to V$, then there is no restriction on how the eigenvalues $\lambda_1, \ldots, \lambda_n$ associated to $v_1, \ldots, v_n$ are related to each other.

Remark 41. We will also talk about eigenvalues, eigenvectors, and eigenbases for a (square) matrix $A \in M_{n\times n}(\mathbb{K})$, with the understanding that we are referring to the linear transformation $\mathbb{K}^n \to \mathbb{K}^n$ represented by A. In particular, if $A \in M_{n\times n}(\mathbb{K})$, then we call $\vec{v} \in \mathbb{K}^n$ an **eigenvector** of A if $\vec{v} \neq \vec{0}$ and if there is a scalar $\lambda \in \mathbb{K}$ with $A\vec{v} = \lambda\vec{v}$. Similarly, $\lambda \in \mathbb{K}$ is called an **eigenvalue** of A if there is $\vec{v} \in \mathbb{K}^n$ with $\vec{v} \neq \vec{0}$ such that $A\vec{v} = \lambda\vec{v}$. In either case, λ is called the **eigenvalue associated to** $\vec{v}$. An **eigenbasis** for A is a basis $\mathscr{B} = (\vec{v}_1, \ldots, \vec{v}_n)$ for $\mathbb{K}^n$ such that each $\vec{v}_k$ is an eigenvector of A.

In light of Remark 39, the relationship between eigenvectors and diagonalizability is straightforward. Here, $\mathscr{E} = (\vec{e}_1, \ldots, \vec{e}_n)$ denotes the standard basis of $\mathbb{K}^n$.

Theorem 16 (Diagonalization and Eigenbases)**.** Let V be a finite-dimensional vector space over $\mathbb{K}$, let $n = \dim(V)$, let $\mathscr{B} = (v_1, \ldots, v_n)$ be a basis for V, and let $T : V \to V$ be linear. Then the following are equivalent.

(i) $\left[T\right]_{\mathscr{B}}$ is diagonal.

(ii) $\mathscr{B}$ is an eigenbasis for T.

Moreover, if (i) or (ii) (and therefore both) hold, then

$$\left[T\right]_{\mathscr{B}} = \begin{bmatrix} \lambda_1 & 0 & 0 & \cdots & 0 \\ 0 & \lambda_2 & 0 & \cdots & 0 \\ 0 & 0 & \lambda_3 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \lambda_n \end{bmatrix}$$

where, for each $1 \leq k \leq n$, λ_k is the eigenvalue of T corresponding to the eigenvector v_k.
In the special case where $T : \mathbb{K}^n \to \mathbb{K}^n$, $T(\vec{x}) = A\vec{x}$, then we have $A = S\left[T\right]_{\mathscr{B}} S^{-1}$, where $S = S_{\mathscr{E}\leftarrow\mathscr{B}} = \begin{bmatrix} \vec{v}_1 & \cdots & \vec{v}_n \end{bmatrix}$.

Proof. On your homework!

□

Summary of Coordinate Results

Let V be a finite-dimensional vector space over $\mathbb{K}$, let $T : V \to V$ be linear, let $\mathscr{B} = (v_1, \ldots, v_n)$ and $\mathscr{C} = (w_1, \ldots, w_n)$ be bases for V.

- For each $v \in V$, $\left[v\right]_{\mathscr{B}} \stackrel{def}{=} \begin{bmatrix} c_1 \\ \vdots \\ c_n \end{bmatrix} \in \mathbb{K}^n$ is the **$\mathscr{B}$-coordinate vector** of v, where the entries $c_1, \ldots, c_n$ are the unique scalars satisfying $v = c_1 v_1 + \cdots + c_n v_n$.
  In the special case where $V = \mathbb{K}^n$ then we have $\vec{v} = \begin{bmatrix} \vec{v}_1 & \cdots & \vec{v}_n \end{bmatrix} \left[\vec{v}\right]_{\mathscr{B}}$.
- The map $\left[\cdot\right]_{\mathscr{B}} : V \to \mathbb{K}^n$ is a linear isomorphism.
- The matrix $\left[T\right]_{\mathscr{B}} = \begin{bmatrix} \left[T(v_1)\right]_{\mathscr{B}} & \cdots & \left[T(v_n)\right]_{\mathscr{B}} \end{bmatrix} \in M_{n\times n}(\mathbb{K})$ is the unique matrix satisfying $\left[T\right]_{\mathscr{B}} \left[v\right]_{\mathscr{B}} = \left[T(v)\right]_{\mathscr{B}}$ for every $v \in V$. It is the matrix of $\left[\cdot\right]_{\mathscr{B}} \circ T \circ \left[\cdot\right]_{\mathscr{B}}^{-1} : \mathbb{K}^n \to \mathbb{K}^n$, and is called the **matrix of T relative to $\mathscr{B}$**.
- For every $v \in V$ we have $\left[v\right]_{\mathscr{C}} = S_{\mathscr{C}\leftarrow\mathscr{B}} \left[v\right]_{\mathscr{B}}$, where $S_{\mathscr{C}\leftarrow\mathscr{B}} = \begin{bmatrix} \left[v_1\right]_{\mathscr{C}} & \cdots & \left[v_n\right]_{\mathscr{C}} \end{bmatrix}$. The change-of-basis matrix $S_{\mathscr{C}\leftarrow\mathscr{B}}$ is invertible and its inverse is $S_{\mathscr{B}\leftarrow\mathscr{C}}$.

Definition 18. $A, B \in M_{n\times n}(\mathbb{K})$ are **similar** if there is an invertible matrix $S \in M_{n\times n}(\mathbb{K})$ with $AS = SB$ (or equivalently $SA = BS$ or $A = S^{-1}BS$ or $B = S^{-1}AS$).

Exercise from MATH 291-1: Suppose that $A, B \in M_{n\times n}(\mathbb{K})$ are similar. Then $\dim(\text{col}(A)) = \dim(\text{col}(B))$, $\dim(\text{null}(A)) = \dim(\text{null}(B))$, $\text{rank}(A) = \text{rank}(B)$, A^k and B^k are similar for each $k \in \mathbb{N}$, A is nilpotent if and only if B is nilpotent, and A is invertible if and only if B is invertible (and, in this case, A^{-1} and B^{-1} are similar).

Proposition. Suppose $A, B \in M_{n\times n}(\mathbb{K})$ are similar. Then $\det(A) = \det(B)$.

Theorem (Characterization of Similarity)**.** Let V be a finite-dimensional vector space over $\mathbb{K}$ with $n = \dim(V)$, let $T : V \to V$ be a linear transformation, let $\mathscr{B}$ be a basis for V, and let $C \in M_{n\times n}(\mathbb{K})$. Then C and $[T]_{\mathscr{B}}$ are similar if, and only if, there is a basis $\mathscr{C}$ for V such that $C = [T]_{\mathscr{C}}$.

Definition 19. Let V be a finite-dimensional vector space over $\mathbb{K}$, and let $T : V \to V$ be linear. Define $\det(T) \stackrel{def}{=} \det(\big[T\big]_{\mathscr{B}})$, where $\mathscr{B}$ is any basis for V.

Lecture 10: More Eigenvalues and Eigenvectors

Learning Objectives:

- Compute the eigenvalues and eigenvectors of a given linear transformation.
- Characterize invertibility in terms of eigenvalues.
- Compute the eigenspace of an operator corresponding to an eigenvalue λ, and compute the geometric multiplicity of λ.

Let's start today on a hopeful note by diagonalizing some linear maps.

Example 12. In discussion you saw that $A = \begin{bmatrix} 7/2 & -9/2 \\ 3/2 & -5/2 \end{bmatrix}$ satisfies

$$A\begin{bmatrix} 3 \\ 1 \end{bmatrix} = 2\begin{bmatrix} 3 \\ 1 \end{bmatrix} \quad \text{and} \quad A\begin{bmatrix} 1 \\ 1 \end{bmatrix} = -1\begin{bmatrix} 1 \\ 1 \end{bmatrix}.$$

Therefore $\begin{bmatrix} 3 \\ 1 \end{bmatrix}$ and $\begin{bmatrix} 1 \\ 1 \end{bmatrix}$ are eigenvectors of A associated to the eigenvalues 2 and -1 (respectively). Moreover, if $\mathscr{B}$ is the basis for $\mathbb{K}^2$ consisting of $\begin{bmatrix} 3 \\ 1 \end{bmatrix}$ and $\begin{bmatrix} 1 \\ 1 \end{bmatrix}$, then $\mathscr{B}$ is an eigenbasis for A and we have

$$\underbrace{\begin{bmatrix} 7/2 & -9/2 \\ 3/2 & -5/2 \end{bmatrix}}_{A} = \underbrace{\begin{bmatrix} 3 & 1 \\ 1 & 1 \end{bmatrix}}_{S}\underbrace{\begin{bmatrix} 2 & 0 \\ 0 & -1 \end{bmatrix}}_{[A]_{\mathscr{B}}}\underbrace{\begin{bmatrix} 3 & 1 \\ 1 & 1 \end{bmatrix}^{-1}}_{S^{-1}}, \quad \text{or} \quad A = S[A]_{\mathscr{B}}S^{-1}.$$

Example 13. Let $H : P_n(\mathbb{R}) \to P_n(\mathbb{R})$ be $H(p(x)) = xp'(x)$. Then note that $H(1) = 0 = 01$, and that for each $1 \leq k \leq n$ we have $H(x^k) = x(kx^{k-1}) = kx^k$, so that x^k is an eigenvector of H with associated eigenvalue k. Therefore the standard basis $\mathscr{E} = (1, x, x^2, \ldots, x^n)$ is an eigenbasis for H, and we see that since $[H(x^k)]_{\mathscr{E}} = k\vec{e}_{k+1}$ for $1 \leq k \leq n$,

$$[H]_{\mathscr{E}} = \begin{bmatrix} 0 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & n \end{bmatrix}$$

is diagonal. Therefore H is diagonalizable.

The next examples illustrate the types of wild things that can happen with eigenvalues and eigenvectors.

Example 14. In this example, we show that the underlying field can affect the existence of eigenvalues and eigenvectors. To see this, let's find the eigenvalues and eigenvectors of $A \stackrel{def}{=} \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} \in M_{2\times 2}(\mathbb{K})$.

Let $\lambda \in \mathbb{K}$, and suppose that λ is an eigenvalue of T. Then there is $\begin{bmatrix} a \\ b \end{bmatrix} \neq \vec{0}$ such that

$$\begin{bmatrix} -b \\ a \end{bmatrix} = A \begin{bmatrix} a \\ b \end{bmatrix} = \lambda \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} \lambda a \\ \lambda b \end{bmatrix}, \quad \text{so that} \quad b = -\lambda a \quad \text{and} \quad a = \lambda b.$$

But then we must have $b = -\lambda a = -\lambda^2 b$ and $a = \lambda b = -\lambda^2 a$, so that $(1+\lambda^2)a = 0$ and $(1+\lambda^2)b = 0$. Because $a \neq 0$ or $b \neq 0$, we must have $1 + \lambda^2 = 0$. Our ability to solve this equation depends on whether $\mathbb{K} = \mathbb{R}$ or $\mathbb{K} = \mathbb{C}$. In particular:

Case 1: $\mathbb{K} = \mathbb{R}$. Because there are no real numbers λ with $1 + \lambda^2 = 0$, we conclude that A has no eigenvalues (and, therefore no eigenvectors). Note that in this case A is the matrix of $R_{\frac{\pi}{2}} : \mathbb{R}^2 \to \mathbb{R}^2$, the counterclockwise rotation of the plane by $\frac{\pi}{2}$ radians, which makes this conclusion unsurprising.

Case 2: $\mathbb{K} = \mathbb{C}$. We solve $1 + \lambda^2 = 0$ to see that $\lambda = i$ or $\lambda = -i$ are the only two possible eigenvalues of A. To find all possible eigenvectors, suppose first that $\vec{v} = \left[\begin{smallmatrix} a \\ b \end{smallmatrix}\right]$ is an eigenvector of A with associated eigenvalue i. Then (by the above computation) we have $a = ib$, so that $\vec{v} = b\left[\begin{smallmatrix} i \\ 1 \end{smallmatrix}\right]$. Since $\vec{v} \neq \vec{0}$, $b \neq 0$. Therefore the only possible eigenvectors of A with associated eigenvalue i are $b\left[\begin{smallmatrix} i \\ 1 \end{smallmatrix}\right]$ for $b \in \mathbb{C}$ with $b \neq 0$. A similar computation shows that the only possible eigenvectors of A with associated eigenvalue $-i$ have the form $b\left[\begin{smallmatrix} 1 \\ i \end{smallmatrix}\right]$ for $b \in \mathbb{C}$ with $b \neq 0$.

Of course, we still need to prove that i and $-i$ are actually eigenvalues of A, and that the vectors we produced are eigenvectors of A associated to those eignevalues. To this end, let $b \in \mathbb{C}$ with $b \neq 0$. Then

$$A\left(b \begin{bmatrix} i \\ 1 \end{bmatrix}\right) = b \begin{bmatrix} -1 \\ i \end{bmatrix} = ib \begin{bmatrix} i \\ 1 \end{bmatrix} \quad \text{and} \quad A\left(b \begin{bmatrix} 1 \\ i \end{bmatrix}\right) = b \begin{bmatrix} -i \\ 1 \end{bmatrix} = -ib \begin{bmatrix} 1 \\ i \end{bmatrix},$$

so that $i, -i$ are indeed eigenvalues of A and $b \begin{bmatrix} i \\ 1 \end{bmatrix}$ and $b \begin{bmatrix} 1 \\ i \end{bmatrix}$ are eigenvectors of A with associated eigenvalues i and $-i$ (respectively).

Things become very interesting in the infinite-dimensional case, as the next couple examples show.

Example 15. Let $P(\mathbb{R})$ denote the vector space of polynomials over $\mathbb{R}$, and let $D : P(\mathbb{R}) \to P(\mathbb{R})$ be $D(p) = p'$. Note that if $p(x) = c$ is a nonconstant, nonzero polynomial then $D(p) = 0 = 0p$, so that p is an eigenvector of D with associated eigenvalue 0. Because $\deg(p') = \deg(p) - 1$ if p is a nonconstant polynomial, D does not have any other eigenvectors.

Example 16. Recall that $C^1(\mathbb{R}, \mathbb{R})$ is the vector space of all continuously differentiable functions over $\mathbb{R}$, and let $D : C^1(\mathbb{R}, \mathbb{R}) \to C^1(\mathbb{R}, \mathbb{R})$ be differentiation $D(f) = f'$. Let $\lambda \in \mathbb{R}$. Then $D(e^{\lambda x}) = \lambda e^{\lambda x}$ and $e^{\lambda x}$ is not the zero function, so λ is an eigenvalue of D and $e^{\lambda x}$ is an eigenvector of D with associated eigenvalue λ. Note how the behavior of D is markedly different on $C^1(\mathbb{R}, \mathbb{R})$ than on its subspace $P(\mathbb{R})$.

Eigenspaces

Diagonalization is not a process of luck; there is a systematic way that one can analyze a given linear map to determine its eigenvalues, its eigenvectors, and then decide whether there is an eigenbasis for the map. We start by exploring the easiest part of this process: determining the eigenvectors of a linear map given its eigenvalues.

We start with a (whimsically named) proposition.

Proposition 8 (Eigencharacterization)**.** Let V be a vector space over $\mathbb{K}$, let $T : V \to V$ be linear, and let $\lambda \in \mathbb{K}$ and $v \in V$. Then the following are equivalent.

(a) λ is an eigenvalue of T.

(b) $\ker(T - \lambda I)$ is not trivial.

(c) (if V is finite-dimensional) $\det(T - \lambda I) = 0$.

Moreover, v is an eigenvector of T with associated eigenvalue λ if, and only if, $v \in \ker(T - \lambda I)$ and $v \neq 0$.

Proof. We start by showing that (a) and (b) are equivalent. Suppose λ is an eigenvalue of T. Choose $w \in V$ with $w \neq 0$ such that $T(w) = \lambda w = \lambda I(w)$. Then $(T - \lambda I)(w) = 0$, so that $w \in \ker(T - \lambda I)$. Since $w \neq 0$, $\ker(T - \lambda I)$ is not trivial. Conversely, if $\ker(T - \lambda I)$ is not trivial then there is a nonzero vector w with $0 = (T - \lambda I)(w) = T(w) - \lambda I(w) = T(w) - \lambda w$, so that $T(w) = \lambda w$. Because $w \neq 0$, λ is an eigenvalue of T.

If V is finite-dimensional (and $\mathscr{B}$ is any basis for V), then (c) is equivalent to the statement that $\det([T - \lambda I]_{\mathscr{B}}) = 0$, which is equivalent to the statement that $[T - \lambda I]_{\mathscr{B}}$ is not invertible, which (by a result last quarter) is equivalent to the statement that $T - \lambda I$ is not invertible, which (by the finite-dimensionality of V) is equivalent to (b).

The "moreover" statement follows from almost the identical argument used to prove (a)⇔(b) above. □

As an immediate corollary, we obtain the following.

Corollary 8. Let V be a vector space over $\mathbb{K}$, and suppose that $T : V \to V$ is linear.

(a) 0 is an eigenvalue of T if, and only if, T is not injective.

(b) If V is finite dimensional, then 0 is an eigenvalue of T if, and only if, T is not invertible.

Proof. The first claim follows immediately from the Eigencharacterization Proposition and a result from last quarter because $\ker(T - 0I) = \ker(T)$. The second claim then follows immediately by a result from last quarter. □

Of course, the interpretation when $V = \mathbb{K}^n$ gives one more addition to the Invertibility Theorem.

Theorem 17 (Invertibility, cont'd.)**.** Let $A = \begin{bmatrix} \vec{a}_1 & \cdots & \vec{a}_n \end{bmatrix} \in M_{n\times n}(\mathbb{K})$, and let $T : \mathbb{K}^n \to \mathbb{K}^n$ be the linear transformation $T(\vec{x}) = A\vec{x}$. Then the following are equivalent.

(a-1) A is invertible.

$\vdots$

(c-13) 0 is not an eigenvalue of T.

Proof. The equivalence of (c-13) with (c-6) (that T is injective) is immediately from the previous corollary. □

The Eigencharacterization Proposition implies that $\ker(T - \lambda I)$ is a subspace of V consisting of 0 and all eigenvectors of T with associated eigenvalue λ. In light of this, we make the following definition.

Definition 20. Let V be a vector space over $\mathbb{K}$, let $T : V \to V$ be linear, and suppose that $\lambda \in \mathbb{K}$ is an eigenvalue of T. The **eigenspace** of T corresponding to λ, E_λ, is

$$E_\lambda \stackrel{def}{=} \ker(T - \lambda I).$$

Example 17. The matrix $A = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$ has only a single eigenvalue 1. (We will prove this later.) Taking this for granted, note that

$$E_1 = \ker(A - 1I_2) = \ker(\text{rref}(A - I_2)) = \ker\left(\text{rref}\begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}\right) = \ker\left(\begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}\right) = \text{span}(\vec{e}_1).$$

So, every vector of the form $c\vec{e}_1$ (with $c \neq 0$) is an eigenvector of A with associated eigenvalue 1, and A has no other eigenvectors.

Because any pair of eigenvectors $\vec{v}_1, \vec{v}_2$ for A must lie in the (one-dimensional) subspace E_1, we see that every such pair must be linearly dependent, and therefore A does not have an eigenbasis (and is therefore not diagonalizable).

Example 18. The symmetric matrix $A = \begin{bmatrix} -1 & 2 & -2 \\ 2 & 2 & 1 \\ -2 & 1 & 2 \end{bmatrix}$ has eigenvalues 3 and -3. (We will prove this later.) Taking this for granted, compute bases for the eigenspaces E_3 and E_{-3}.

We can therefore compute that the eigenspaces are

$$\begin{aligned} E_3 &= \text{null}(A - 3I_3) \\ &= \text{null}(\text{rref}(A - 3I_3)) \\ &= \text{null}\left(\text{rref}\begin{bmatrix} -4 & 2 & -2 \\ 2 & -1 & 1 \\ -2 & 1 & -1 \end{bmatrix}\right) = \text{null}\left(\begin{bmatrix} 1 & -\frac{1}{2} & \frac{1}{2} \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}\right) = \text{span}\left(\begin{bmatrix} \frac{1}{2} \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} -\frac{1}{2} \\ 0 \\ 1 \end{bmatrix}\right) \end{aligned}$$

and

$$\begin{aligned} E_{-3} &= \text{null}(A-(-3)I_3) \\ &= \text{null}(\text{rref}(A+3I_3)) \\ &= \text{null}\left(\text{rref}\begin{bmatrix} 2 & 2 & -2 \\ 2 & 5 & 1 \\ -2 & 1 & 5 \end{bmatrix}\right) \\ &= \text{null}\left(\begin{bmatrix} 1 & 0 & -2 \\ 0 & 1 & 1 \\ 0 & 0 & 0 \end{bmatrix}\right) \\ &= \text{span}\left(\begin{bmatrix} 2 \\ -1 \\ 1 \end{bmatrix}\right) \end{aligned}$$

For fun, note that the set of vectors $\mathscr{B} = \left(\begin{bmatrix} 1/2 \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} -1/2 \\ 0 \\ 1 \end{bmatrix}, \begin{bmatrix} 2 \\ -1 \\ 1 \end{bmatrix}\right)$ is linearly independent because (replacing $R_2 \to R_2 - 2R_1$ and then $R_3 \to R_3 - R_2$)

$$\det\left(\begin{bmatrix} 1/2 & -1/2 & 2 \\ 1 & 0 & -1 \\ 0 & 1 & 1 \end{bmatrix}\right) = \det\left(\begin{bmatrix} 1/2 & -1/2 & 2 \\ 0 & 1 & -5 \\ 0 & 1 & 1 \end{bmatrix}\right) = \det\left(\begin{bmatrix} 1/2 & -1/2 & 2 \\ 0 & 1 & -5 \\ 0 & 0 & 6 \end{bmatrix}\right) = 3 \neq 0.$$

Therefore $\mathscr{B}$ is a basis for $\mathbb{R}^3$ consisting of eigenvectors of A, and is therefore an eigenbasis for A. We conclude that A is diagonalizable and

$$A = \begin{bmatrix} 1/2 & -1/2 & 2 \\ 1 & 0 & -1 \\ 0 & 1 & 1 \end{bmatrix} \begin{bmatrix} 3 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & -3 \end{bmatrix} \begin{bmatrix} 1/2 & -1/2 & 2 \\ 1 & 0 & -1 \\ 0 & 1 & 1 \end{bmatrix}^{-1}.$$

The previous two examples drive home the important fact that the dimension of each eigenspace controls how many eigenvectors (associated to that eigenvalue) that the eigenspace could contribute towards an eigenbases. Let's give this dimension a name.

Definition 21. Let V be a finite-dimensional vector space over $\mathbb{K}$, let $T : V \to V$ be linear, and suppose $\lambda \in \mathbb{K}$ is an eigenvalue of T. The **geometric multiplicity** of λ, denoted $\text{gemu}(\lambda)$, is

$$\text{gemu}(\lambda) \stackrel{def}{=} \dim(E_\lambda) = \text{nullity}(T - \lambda I).$$

Remark 42. Note that if λ is an eigenvalue of T then $\ker(T - \lambda I)$ is nontrivial, and therefore has dimension at least 1. It follows that $\text{gemu}(\lambda) \geq 1$.

We now know how to find a basis for the eigenspaces of a linear transformation, and therefore to compute the dimension of the space (which is the geometric multiplicity of the eigenvalue). This will be a crucial part of both the theory and practice of diagonalization.

Invertibility Theorem (cont'd)

Theorem 18 (Invertibility, cont'd.)**.** Let $A = \begin{bmatrix} \vec{a}_1 & \cdots & \vec{a}_n \end{bmatrix} \in M_{n\times n}(\mathbb{K})$, and let $T : \mathbb{K}^n \to \mathbb{K}^n$ be the linear transformation $T(\vec{x}) = A\vec{x}$. Then the following are equivalent.

(a-1) A is invertible.

$\vdots$

(a-11) $\det(A) \neq 0$.

(a-12) (When $\mathbb{K} = \mathbb{R}$) $\mathrm{Vol}_n(E(\vec{a}_1, \ldots, \vec{a}_n)) > 0$.

(a-13) (When $\mathbb{K} = \mathbb{R}$) There is a parallelotope $P \overset{def}{=} E(\vec{b}_1, \ldots, \vec{b}_n) \subset \mathbb{R}^n$ such that $\mathrm{Vol}_n(T(P)) > 0$.

$\vdots$

(c-13) 0 is not an eigenvalue of T.

Proof. We have already shown that (a-1) and (a-11) are equivalent. Suppose that $\mathbb{K} = \mathbb{R}$. Because $\mathrm{Vol}_n(E(\vec{a}_1, \ldots, \vec{a}_n)) = |\det(A)|$, (a-11) holds if, and only if, (a-12) holds. Note also that taking $P = E(\vec{e}_1, \ldots, \vec{e}_n)$ gives $\mathrm{Vol}_n(T(P)) = |\det(A)|\mathrm{Vol}_n(P) = |\det(A)|$, and therefore (a-13) holds if, and only if, (a-11) holds.

The equivalence of (c-13) with (c-6) (i.e. that T is injective) is immediate from the Eigencharacterization Proposition. □

Lecture 11: Even More Eigenvalues and Eigenvectors

Learning Objectives:

- Compute eigenvalues of a given linear transformation on a finite-dimensional vector space.
- Compute the algebraic multiplicity of an eigenvalue of a linear operator.
- Link the eigenvalues of an complex operator to the determinant of the operator.

The Characteristic Polynomial

We now turn to a slightly more difficult problem: finding the eigenvalues of a linear map to begin with. To make things more tractable we will restrict ourselves to the finite-dimensional case.

Definition 22. Let V be a finite-dimensional vector space over $\mathbb{K}$, and let $T : V \to V$ be linear. The equation $\det(T - \lambda I) = 0$ is called the **characteristic equation of** T.

Recall that in the Eigencharacterization Proposition, we proved that $\lambda \in \mathbb{K}$ is an eigenvalue of a linear map $T : V \to V$ exactly when $\det(T - \lambda I) = 0$. Therefore, the solutions of the characteristic equation of T are exactly the eigenvalues of T.

Example 19. If $A, B \in M_{n\times n}(\mathbb{K})$ are similar, then A and B have the same eigenvalues.

To prove this, note that if S is an invertible matrix such that $A = S^{-1}BS$, then

$$A - \lambda I_n = S^{-1}BS - S^{-1}(\lambda I_n)S = S^{-1}(B - \lambda I_n)S,$$

so that $A - \lambda I_n$ and $B - \lambda I_n$ are similar and therefore $\det(A - \lambda I_n) = \det(B - \lambda I_n)$. Therefore the characteristic equations of A and B are the same, so the eigenvalues of A and B are the same. (Note that we have *not* shown that A and B have the same *eigenvectors*.)

Computing the eigenvalues of a linear operator T (on a finite-dimensional vector space) is therefore equivalent to solving its characteristic equation $\det(T - \lambda I) = 0$. Let's see some examples.

Example 20. For the matrix $A = \begin{bmatrix} 7/2 & -9/2 \\ 3/2 & -5/2 \end{bmatrix}$,

$$\begin{aligned}\det(A - \lambda I_2) &= \det\left(\begin{bmatrix} (7/2) - \lambda & -9/2 \\ 3/2 & -(5/2) - \lambda \end{bmatrix}\right) \\ &= \Big(\frac{7}{2} - \lambda\Big)\Big(-\frac{5}{2} - \lambda\Big) - \frac{3}{2}\Big(-\frac{9}{2}\Big) = \lambda^2 - \lambda - 2 = (\lambda - 2)(\lambda + 1),\end{aligned}$$

so that $\det(A - \lambda I_2) = 0$ exactly when $\lambda = 2, -1$. Note that these are exactly the eigenvalues that we identified earlier.

Example 21. For the symmetric matrix $A = \begin{bmatrix} -1 & 2 & -2 \\ 2 & 2 & 1 \\ -2 & 1 & 2 \end{bmatrix}$,

$$\det(A - \lambda I_3) = \det\left(\begin{bmatrix} -1-\lambda & 2 & -2 \\ 2 & 2-\lambda & 1 \\ -2 & 1 & 2-\lambda \end{bmatrix}\right) = \cdots = -(\lambda-3)^2(\lambda+3),$$

so that the eigenvalues of A are exactly 3 and -3. The "$\cdots$" step is left to you to check, and can be done using any of your favorite methods for computing determinants.

Example 22. Suppose that $A \in M_{n\times n}(\mathbb{K})$ is a triangular matrix. Then the eigenvalues of A are exactly the entries of A on the main diagonal.

To see why, note that since A is triangular with entries $a_{1,1}, \ldots, a_{n,n}$ along the main diagonal and λI_n is diagonal with each entry along the main diagonal equal to λ, $A - \lambda I_n$ is triangular with entries $a_{1,1} - \lambda, \ldots, a_{n,n} - \lambda$ along the main diagonal, so that

$$\det(A - \lambda I_n) = (a_{1,1} - \lambda) \cdots (a_{n,n} - \lambda).$$

Therefore the eigenvalues of A (i.e. the solutions of $\det(A - \lambda I_n) = 0$) are exactly $a_{1,1}, a_{2,2}, \ldots, a_{n,n}$.

It was no coincidence in the previous examples that $\det(A - \lambda I_n)$ was a polynomial of degree n.

Proposition 9. If $A \in M_{n\times n}(\mathbb{K})$ then $\det(A - \lambda I_n)$ is a polynomial in λ of degree n.

Proof. Note that $A - \lambda I_n$ has the same entries as A off of the main diagonal, and the entries on its main diagonal are $a_{1,1} - \lambda, \ldots, a_{n,n} - \lambda$.

For the pattern $P = \{(1,1), \ldots, (n,n)\}$,

$$\operatorname{prod} P(A - \lambda I_n) = (a_{1,1} - \lambda) \cdots (a_{n,n} - \lambda)$$

is a polynomial in λ of degree n. On the other hand, for any other pattern $P' \neq P$ then $\operatorname{prod} P'(A - \lambda I_n)$ is a product of n entries of $A - \lambda I_n$ with no more than $n-2$ entries on the main diagonal[3] from the main diagonal, so that $\operatorname{prod} P'(A - \lambda I_n)$ is a polynomial in λ of degree no more than $n-2$. It follows that

$$\begin{aligned} \det(A - \lambda I_n) &= \operatorname{prod} P(A - \lambda I_n) + \sum_{n\times n \text{ patterns } P' \neq P} (\operatorname{prod} P'(A - \lambda I_n))(\operatorname{sgn}(P)) \\ &= \underbrace{(a_{1,1} - \lambda) \cdots (a_{n,n} - \lambda)}_{\text{degree } n} + (\text{polynomial in } \lambda \text{ with degree } \leq n-2) \end{aligned}$$

is a polynomial of degree n. □

Remark 43. Note that, in the proof of the previous proposition, the coefficient of the λ^n term is $(-1)^n$. We will use this fact later.

[3]Note that if a pattern P consisted of $n-1$ entries on the main diagonal, then the remaining entry in the pattern could only be the remaining entry on the main diagonal.

In light of the previous remarks, we make a definition.

Definition 23. Let V be a finite-dimensional vector space over $\mathbb{K}$, and let $T : V \to V$ be linear. The function $p : \mathbb{K} \to \mathbb{K}$, $p(\lambda) \stackrel{def}{=} \det(T - \lambda I)$ is called the **characteristic polynomial of** T.

Remark 44. Therefore, for a linear map $T : V \to V$ on a finite-dimensional vector space over $\mathbb{K}$, $\lambda \in \mathbb{K}$ is an eigenvalue of T if, and only if, λ is a root of the characteristic polynomial of T.

The fact that the characteristic polynomial of an operator is a polynomial allows us to analyze the eigenvalues of a linear transformation using basic properties of polynomials. We start by giving a definition.

Definition 24. Let V be a finite-dimensional vector space over $\mathbb{K}$, and let $T : V \to V$ be linear. The **algebraic multiplicity** of an eigenvalue λ_0 of T, denoted $\text{almu}(\lambda_0)$, is the multiplicity[4] of λ_0 as a root of the characteristic polynomial $p(\lambda)$ of T.

Remark 45. Note that if λ is an eigenvalue of T, then λ is a root of the characteristic polynomial p of T, so that $\text{almu}(\lambda) \geq 1$.

Fundamental results about the number of roots a polynomial might (or might not) possess immediately give some results about the algebraic multiplicity of eigenvalues.

Proposition 10. Let V be a finite-dimensional vector space over $\mathbb{K}$, $T : V \to V$ is linear, and suppose that $\lambda_1, \ldots, \lambda_m$ are the distinct eigenvalues of T. Then $m \leq \dim(V)$ and

$$\text{almu}(\lambda_1) + \cdots + \text{almu}(\lambda_m) \leq \dim(V),$$

with equality guaranteed if $\mathbb{K} = \mathbb{C}$.

Proof. $\det(T - \lambda I)$ is a polynomial of degree at most $\dim(V)$, and therefore (by polynomial division and uniqueness of factorization) has no more than $\dim(V)$ roots (counted according to algebraic multiplicity).

If $\mathbb{K} = \mathbb{C}$, then the Fundamental Theorem of Algebra implies that $\det(T - \lambda I)$ factors completely into linear factors

$$\det(T - \lambda I) = c(\lambda_1 - \lambda)^{r_1} \cdots (\lambda_k - \lambda)^{r_m},$$

where $\lambda_1, \ldots, \lambda_k$ are the distinct roots of the polynomial, $r_1 + \cdots + r_k = n$, and $c \in \mathbb{C}$. □

Remark 46. Note that, in the last proof, the coefficient of λ^n is $c(-1)^n$. Since (by a previous remark) we know that the coefficient of λ^n is $(-1)^n$, we can conclude that $c = 1$.

If $\mathbb{K} = \mathbb{C}$, then we can even say a little more about the relationship between eigenvalues and the determinant.

Proposition 11. Let V be a finite-dimensional vector space over $\mathbb{C}$, $T : V \to V$ be linear, and suppose that $\lambda_1, \ldots, \lambda_m$ are the distinct eigenvalues of T. Then

$$\det(T) = \lambda_1^{\text{almu}(\lambda_1)} \cdots \lambda_m^{\text{almu}(\lambda_m)}.$$

[4]That is, $\text{almu}(\lambda_0)$ is the largest natural number k for which $(\lambda_0 - \lambda)^k$ divides $p(\lambda)$.

Proof. By the observation in the proof of the previous proposition (and the remark after the proof), for each $\lambda \in \mathbb{C}$ we have

$$\det(T - \lambda I) = (\lambda_1 - \lambda)^{\text{almu}(\lambda_1)} \cdots (\lambda_m - \lambda)^{\text{almu}(\lambda_m)},$$

Setting $\lambda = 0$ gives $\det(T) = \lambda_1^{\text{almu}(\lambda_1)} \cdots \lambda_m^{\text{almu}(\lambda_m)}$. □

Lecture 12: Diagonalization

Learning Objectives:

- Relate the geometric multiplicity of an eigenvalue to the algebraic multiplicity of an eigenvalue.
- Describe the relationship between distinct eigenspaces and linear independence.
- Characterize diagoonalizability in terms of geometric multiplicities.

The notions of algebraic and geometric multiplicity are critical to understanding when it is possible to diagonalize a linear map. One of the crucial facts in this process is the (perhaps surprising) relationship between them captured in the following theorem.

Theorem 19 (Multiplicity Comparison)**.** Let V be a finite-dimensional vector space over $\mathbb{K}$, and suppose that $T : V \to V$ is linear with eigenvalue $\lambda_0 \in \mathbb{K}$. Then $\text{gemu}(\lambda_0) \leq \text{almu}(\lambda_0)$.

Proof. Suppose that $\text{gemu}(\lambda_0) = \ell$. Let $\vec{v}_1, \ldots, \vec{v}_\ell$ be a basis for E_{λ_ℓ}. By the Constructing Bases Theorem, there are $\vec{u}_1, \ldots, \vec{u}_p$ such that $\mathscr{B} = (\vec{v}_1, \ldots, \vec{v}_\ell, \vec{u}_1, \ldots, \vec{u}_p)$ is a basis for V. Then because $T(\vec{v}_k) = \lambda_0 \vec{v}_k$ for each $1 \leq k \leq \ell$, the matrix for $[T]_{\mathscr{B}}$ is a block matrix of the form

$$[T]_{\mathscr{B}} = \begin{bmatrix} \lambda_0 & 0 & 0 & \cdots & 0 & * & \cdots & * \\ 0 & \lambda_0 & 0 & \cdots & 0 & * & \cdots & * \\ 0 & 0 & \lambda_0 & \cdots & 0 & * & \cdots & * \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \lambda_0 & * & \cdots & * \\ 0 & 0 & 0 & \cdots & 0 & * & \cdots & * \\ 0 & 0 & 0 & \cdots & 0 & * & \cdots & * \\ 0 & 0 & 0 & \cdots & 0 & * & \cdots & * \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & * & \cdots & * \end{bmatrix} = \begin{bmatrix} \lambda_0 I_\ell & B_{\ell \times p} \\ 0_{p \times \ell} & C_{p \times p} \end{bmatrix}$$

Therefore

$$\begin{aligned} \det(T - \lambda I) = \det([T]_{\mathscr{B}} - \lambda I_n) &= \det\left(\begin{bmatrix} (\lambda_0 - \lambda) I_\ell & B_{\ell \times p} \\ 0_{p \times \ell} & C_{p \times p} - \lambda I_p \end{bmatrix}\right) \\ &= (\lambda_0 - \lambda)^\ell \det(C_{p \times p} - \lambda I_p), \end{aligned}$$

where in the last step follows from m applications of the cofactor expansion along the first column. Therefore $\ell \leq \text{almu}(\lambda_0)$, and the result is proved. □

Remark 47. By the previous theorem, it therefore follows that if V is a finite-dimensional vector space over $\mathbb{K}$, $T : V \to V$ is linear, and $\lambda_1, \ldots, \lambda_m$ are the distinct eigenvalues of T, then

$$m \leq \text{gemu}(\lambda_1) + \cdots + \text{gemu}(\lambda_m) \leq \text{almu}(\lambda_1) + \cdots + \text{almu}(\lambda_m) \leq \dim(V).$$

We now turn to the remaining central question of diagonalization: how do we attempt to produce an eigenbasis, given that we know how to compute eigenvalues and that we know how to compute bases for eigenspaces? The following result—that combining linearly independent sets from different eigenspaces results in a linearly independent set—is the last major piece of the puzzle.

Theorem 20 (Independent Eigenspaces)**.** Let V be a finite-dimensional vector space over $\mathbb{K}$, let $T : V \to V$ be linear, and suppose that $\lambda_1, \ldots, \lambda_m$ are distinct eigenvalues of T. For each $1 \leq k \leq m$, let $1 \leq r_k \leq \text{gemu}(\lambda_k)$ and suppose that $\vec{v}_{1,k}, \ldots, \vec{v}_{r_k,k}$ is a linearly independent set in E_{λ_k}. Then the concatenation

$$\underbrace{\vec{v}_{1,1}, \ldots, \vec{v}_{r_1,1}}_{\text{from } E_{\lambda_1}}, \underbrace{\vec{v}_{1,2}, \ldots, \vec{v}_{r_2,2}}_{\text{from } E_{\lambda_2}}, \ldots, \underbrace{\vec{v}_{1,m}, \ldots, \vec{v}_{r_m,m}}_{\text{from } E_{\lambda_m}}$$

is a linearly independent set in V.

Proof. Suppose $c_{1,1}, \cdots, c_{r_1,1}, \ldots, c_{1,m}, \ldots, c_{r_m,m} \in \mathbb{K}$ satisfy

$$\vec{0} = c_{1,1}\vec{v}_{1,1} + \cdots + c_{r_1,1}\vec{v}_{r_1,1} + c_{1,2}\vec{v}_{1,2} + \cdots + c_{r_2,2}\vec{v}_{r_2,2} + \cdots + c_{1,m}\vec{v}_{1,m} + \cdots + c_{r_m,m}\vec{v}_{r_m,m}.$$

For $1 \leq k \leq m$, write $\vec{u}_k \stackrel{def}{=} c_{1,k}\vec{v}_{1,k} + \cdots + c_{r_k,k}\vec{v}_{r_k,k}$. Then $\vec{0} = \vec{u}_1 + \cdots + \vec{u}_m$, and $\vec{u}_k \in E_{\lambda_k}$ for each $1 \leq k \leq m$.

Note that for each $0 \leq j \leq m-1$ and $1 \leq k \leq m$, $T^j(\vec{u}_k) = \lambda_k^j \vec{u}_k$. Therefore we have

$$\vec{0} = T^j(\vec{0}) = T^j(\vec{u}_1 + \cdots \vec{u}_m) = \lambda_1^j\vec{u}_1 + \cdots + \lambda_m^j\vec{u}_m = \begin{bmatrix}\vec{u}_1 & \vec{u}_2 & \cdots & \vec{u}_m\end{bmatrix} \begin{bmatrix}\lambda_1^j \\ \lambda_2^j \\ \vdots \\ \lambda_m^j\end{bmatrix} \quad \text{for each } 0 \leq j \leq m-1.$$

This system of m equations can be expressed as a single matrix equation as

$$0_{n\times m} = \begin{bmatrix}\vec{u}_1 & \vec{u}_2 & \cdots & \vec{u}_m\end{bmatrix} \underbrace{\begin{bmatrix}1 & \lambda_1 & \lambda_1^2 & \cdots & \lambda_1^{m-1} \\ 1 & \lambda_2 & \lambda_2^2 & \cdots & \lambda_2^{m-1} \\ 1 & \lambda_3 & \lambda_3^2 & \cdots & \lambda_3^{m-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & \lambda_m & \lambda_m^2 & \cdots & \lambda_m^{m-1}\end{bmatrix}}_{A}.$$

By the homework problem on Vandermonde determinants and the assumption that $\lambda_1, \ldots, \lambda_m$ are distinct we have

$$\det(A) = \det(A^T) = \det\left(\begin{bmatrix}1 & 1 & 1 & \cdots & 1 \\ \lambda_1 & \lambda_2 & \lambda_3 & \cdots & \lambda_m \\ \lambda_1^2 & \lambda_2^2 & \lambda_3^2 & \cdots & \lambda_m^2 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \lambda_1^{m-1} & \lambda_2^{m-1} & \lambda_3^{m-1} & \cdots & \lambda_m^{m-1}\end{bmatrix}\right) = \prod_{1\leq i<j\leq m}(\lambda_j - \lambda_i) \neq 0.$$

Therefore A is invertible, so that

$$\begin{bmatrix}\vec{u}_1 & \vec{u}_2 & \cdots & \vec{u}_m\end{bmatrix} = 0_{n\times m}A^{-1} = 0_{n\times m},$$

whence it follows that $\vec{u}_1 = \cdots = \vec{u}_m = \vec{0}$.

Therefore, for each $1 \leq k \leq m$, $\vec{0} = c_{1,k}\vec{v}_{1,k} + \cdots + c_{r_k,k}\vec{v}_{r_k,k}$ and therefore (by linear independence) $c_{1,k} = \cdots = c_{r_k,k} = 0$. This concludes the proof. □

We can now say exactly when a linear transformation is diagonalizable.

Theorem 21 (Diagonalizability)**.** Let V be a finite-dimensional vector space over $\mathbb{K}$, let $n = \dim(V)$, and let $T : V \to V$ be linear. Suppose that $\lambda_1, \ldots, \lambda_m$ are the distinct eigenvalues of T. Then T is diagonalizable if, and only if,

$$\text{gemu}(\lambda_1) + \cdots + \text{gemu}(\lambda_m) = n.$$

In this case one can obtain an eigenbasis for T by concatenating bases for each of $E_{\lambda_1}, \ldots, E_{\lambda_m}$.

Proof. If T is diagonalizable then T has an eigenbasis $\vec{v}_1, \ldots, \vec{v}_n$. Sorting these eigenvectors into those that lie in E_{λ_1}, those in E_{λ_2}, an so one, and using the fact that any subset of a linearly independent set is linearly independent, we have

$$n \leq \text{gemu}(\lambda_1) + \cdots + \text{gemu}(\lambda_m),$$

and therefore $\text{gemu}(\lambda_1) + \cdots + \text{gemu}(\lambda_m) = n$ by Remark 47.

On the other hand, if $\text{gemu}(\lambda_1) + \cdots + \text{gemu}(\lambda_m) = n$, then by the previous corollary we can concatenate bases for $E_{\lambda_1}, \ldots, E_{\lambda_m}$ to obtain a linearly independent set of n eigenvectors of T, which must therefore also be a basis for V because $\dim(V) = n$. □

Algorithm for Diagonalization

This previous corollary provides a road-map for determining when a linear map $T : V \to V$ from a finite-dimensional vector space over $\mathbb{K}$ is diagonalizable, and actually diagonalizing T:

(1) Compute the eigenvalues of T by finding the roots of $\det(T - \lambda I) = 0$.

(2) For each eigenvalue λ that you found in (1), compute the dimension of the eigenspace $E_\lambda = \ker(T - \lambda I)$ (i.e. compute $\text{gemu}(\lambda)$).

(3) If the sum of the geometric multiplicities is equal to $\dim(V)$, then T is diagonalizable and you can produce an eigenbasis for T by simply producing bases for each eigenspace, and then concatenating these bases.

Lecture 13: The Spectral Theorem

Learning Objectives:

- Determine necessary and sufficient conditions under which a real square matrix has an orthonormal eigenbasis.

In the last lecture we determined that $T : V \to V$ (or, equivalently, $A \in M_{n\times n}(\mathbb{K})$) is diagonalizable if and only if the sum of the geometric multiplicities of its eigenvalues is n. Eigenbases are particularly nice to work with because diagonal matrices are easy to work with. Here is one example.

Example 23. If $A \in M_{n\times n}(\mathbb{K})$ is diagonalizable then there is an invertible matrix S and a diagonal matrix D such that $A = SDS^{-1}$. But then for every $m \in \mathbb{N}$,

$$A^m = (SDS^{-1})^m = (SDS^{-1})(SDS^{-1})\cdots(SDS^{-1}) = SD^mS^{-1},$$

so we can compute a power of A by simply computing the analogous power of D (and then multiplying D by S and S^{-1}).

Today we finish our study of diagonalization by inspecting when we can not only find an eigenbasis for a linear transformation $T : \mathbb{R}^n \to \mathbb{R}^n$, but the conditions under which we can take the eigenbasis for T to be orthonormal. Geometrically, this is exactly the situation where we can diagonalize T be simply rotating the standard coordinate axes in $\mathbb{R}^n$ (while keeping them perpendicular to each other). By the homework result where you explored the relationship between diagonalizability of a linear map and diagonalizability of its standard matrix, we can make the following definition.

Definition 25. Let $T : \mathbb{R}^n \to \mathbb{R}^n$ be linear with standard matrix $A \in M_{n\times n}(\mathbb{R})$. We say T is **orthogonally diagonalizable** if T has an orthonormal eigenbasis. Equivalently, we say that A is **orthogonally diagonalizable** if there is an orthogonal matrix $S \in M_{n\times n}(\mathbb{R})$ and a diagonal matrix D such that $A = SDS^{-1} = SDS^T$.

Example 24. We saw that the symmetric matrix $A = \begin{bmatrix} -1 & 2 & -2 \\ 2 & 2 & 1 \\ -2 & 1 & 2 \end{bmatrix}$ has eigenvalues 3 and -3, and that the associated eigenspaces can be described as

$$E_3 = \ker(A - 3I_3) = \operatorname{span}\left(\begin{bmatrix} \frac{1}{2} \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} -\frac{1}{2} \\ 0 \\ 1 \end{bmatrix}\right) \quad \text{and} \quad E_{-3} = \ker(A - (-3)I_3) = \operatorname{span}\left(\begin{bmatrix} 2 \\ -1 \\ 1 \end{bmatrix}\right).$$

The eigenbasis $\begin{bmatrix} 1/2 \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} -1/2 \\ 0 \\ 1 \end{bmatrix}, \begin{bmatrix} 2 \\ -1 \\ 1 \end{bmatrix}$ is not an orthonormal eigenbasis. None of the vectors have norm 1 (which is easy to correct), and $\begin{bmatrix} 2 \\ -1 \\ 1 \end{bmatrix}$ orthogonal to both of $\begin{bmatrix} 1/2 \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} -1/2 \\ 0 \\ 1 \end{bmatrix}$, but this last pair of vectors is not orthogonal. Luckily, this last pair of vectors is a basis for the subspace E_3, so we can replace that pair of vectors with an orthonormal basis for E_3 using the Gram-Schmidt Process. If we perform the

Gram-Schmidt Process on this pair of eigenvectors we get the orthonormal basis $\begin{bmatrix} 1/\sqrt{5} \\ 2/\sqrt{5} \\ 0 \end{bmatrix}, \begin{bmatrix} -2/\sqrt{30} \\ 1/\sqrt{30} \\ 5/\sqrt{30} \end{bmatrix}$ for E_3. We can also perform the Gram-Schmidt process on the basis $\begin{bmatrix} 2 \\ -1 \\ 1 \end{bmatrix}$ for E_{-3} to get the orthonormal basis $\begin{bmatrix} 2/\sqrt{6} \\ -1/\sqrt{6} \\ 1/\sqrt{6} \end{bmatrix}$ for E_{-3}. We therefore see that

$$\begin{bmatrix} 1/\sqrt{5} \\ 2/\sqrt{5} \\ 0 \end{bmatrix}, \begin{bmatrix} -2/\sqrt{30} \\ 1/\sqrt{30} \\ 5/\sqrt{30} \end{bmatrix}, \begin{bmatrix} 2/\sqrt{6} \\ -1/\sqrt{6} \\ 1/\sqrt{6} \end{bmatrix}$$

is an orthonormal eigenbasis for A. Therefore A is orthogonally diagonalizable with

$$A = \begin{bmatrix} 1/\sqrt{5} & -2/\sqrt{30} & 2/\sqrt{6} \\ 2/\sqrt{5} & 1/\sqrt{30} & -1/\sqrt{6} \\ 0 & 5/\sqrt{30} & 1/\sqrt{6} \end{bmatrix} \begin{bmatrix} 3 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & -3 \end{bmatrix} \begin{bmatrix} 1/\sqrt{5} & -2/\sqrt{30} & 2/\sqrt{6} \\ 2/\sqrt{5} & 1/\sqrt{30} & -1/\sqrt{6} \\ 0 & 5/\sqrt{30} & 1/\sqrt{6} \end{bmatrix}^{-1}.$$

One feature of the last example—that vectors in different eigenspaces were orthogonal to each other—is not an accident. It is a general property of symmetric matrices.

Proposition 12 (Orthogonal Eigenspaces)**.** Let $A \in M_{n\times n}(\mathbb{R})$ by symmetric, suppose that $\lambda_1 \neq \lambda_2$ are eigenvalues of A and $\vec{x}_1 \in E_{\lambda_1}$ and $\vec{x}_2 \in E_{\lambda_2}$. Then $\vec{x}_1 \cdot \vec{x}_2 = 0$.

Proof. Note that

$$\lambda_1(\vec{x}_1 \cdot \vec{x}_2) = (\lambda_1\vec{x}_1) \cdot \vec{x}_2 = (A\vec{x}_1) \cdot \vec{x}_2 = \vec{x}_1 \cdot (A^T\vec{x}_2) = \vec{x}_1 \cdot (A\vec{x}_2) = \vec{x}_1 \cdot (\lambda_2\vec{x}_2) = \lambda_2(\vec{x}_1 \cdot \vec{x}_2),$$

so that $(\lambda_1 - \lambda_2)(\vec{x}_1 \cdot \vec{x}_2) = 0$. Because $\lambda_1 \neq \lambda_2$, $\vec{x}_1 \cdot \vec{x}_2 = 0$. □

Therefore the symmetry of the matrix A in the previous example had something to do with our ability to orthogonally diagonalize it. Perhaps surprisingly, when $\mathbb{K} = \mathbb{R}$ the ability to orthogonally diagonalize $A \in M_{n\times n}(\mathbb{R})$ is equivalent to A being symmetric. This remarkable fact is captured in one of the most celebrated theorems in mathematics: the Spectral Theorem.

Theorem 22 (Spectral Theorem)**.** Let $T : \mathbb{R}^n \to \mathbb{R}^n$ be linear with standard matrix $A \in M_{n\times n}(\mathbb{R})$. Then T is orthogonally diagonalizable if, and only if, A is symmetric.

To start, we need to know that every symmetric real matrix has at least one (real) eigenvalue.

Lemma 3 (Eigenvalues of Symmetric Matrices)**.** Let $A \in M_{n\times n}(\mathbb{R})$ by symmetric. Then A has a (real) eigenvalue.

Proof. Let $T : \mathbb{R}^n \to \mathbb{R}^n$ be the linear transformation $T(\vec{x}) = A\vec{x}$, and let $T_{\mathbb{C}} : \mathbb{C}^n \to \mathbb{C}^n$ be the linear transformation $T(\vec{z}) = A\vec{z}$. Let $p : \mathbb{R} \to \mathbb{R}$ be the characteristic polynomial of T, and let $p_{\mathbb{C}} : \mathbb{C} \to \mathbb{C}$ be the characteristic polynomial of $T_{\mathbb{C}}$. Then $p_{\mathbb{C}}(\lambda) = \det(A - \lambda I_n) = p(\lambda)$ whenever $\lambda \in \mathbb{R}$.

By a past result, there is $\lambda_0 \in \mathbb{C}$ with $p_{\mathbb{C}}(\lambda_0) = 0$. We are done when we show that $\lambda_0 \in \mathbb{R}$, since then $p(\lambda_0) = 0$.

Let $\vec{z} = \vec{u} + i\vec{v}$ (with $\vec{u}, \vec{v} \in \mathbb{R}^n$) be an eigenvector of $T_{\mathbb{C}}$ with associated eigenvalue λ_0. By one of your homework problems, $\overline{z} = \vec{u} - i\vec{v}$ is an eigenvector of $T_{\mathbb{C}}$ with associated eigenvalue $\overline{\lambda_0}$. But then we have

$$\overline{\vec{z}}^T A\vec{z} = \overline{\vec{z}}^T(\lambda_0\vec{z}) = \lambda_0(\overline{\vec{z}}^T\vec{z}) = \lambda_0\|\vec{z}\|^2$$

and (because $A = A^T = \overline{A^T}$ since A is symmetric and real)

$$\overline{\vec{z}}^T A\vec{z} = \overline{\vec{z}}^T\overline{A^T}\vec{z} = \overline{(A\vec{z})}^T\vec{z} = \overline{(\lambda_0\vec{z})}^T\vec{z} = \overline{\lambda_0}\overline{\vec{z}}^T\vec{z} = \overline{\lambda_0}\|\vec{z}\|^2.$$

Therefore $0 = (\lambda_0 - \overline{\lambda_0})\|\vec{z}\|^2$, so that $2i\text{Im}(\lambda_0) = \lambda_0 - \overline{\lambda_0} = 0$ since $\|\vec{z}\| \neq 0$. Therefore $\text{Im}(\lambda_0) = 0$, so that $\lambda_0 \in \mathbb{R}$ and the proof is complete. □

We are now ready to prove the Spectral Theorem.

Proof of Spectral Theorem: Using Matrices. First suppose that $\mathscr{B}$ is an orthonormal eigenbasis for A, and let S be the matrix with columns given by the vectors in $\mathscr{B}$. Then there is a diagonal matrix D with $A = SDS^{-1} = SDS^T$. But then we have $A^T = (SDS^T)^T = (S^T)^T D^T S^T = SDS^T = A$, so that A is symmetric.

We prove the other direction via induction. If $n = 1$, then $S = [1]$ is an orthogonal matrix such that $S^{-1}AS = [1][a][1] = [a]$ is diagonal.

Now let $n \in \mathbb{N}$ and assume the result holds for $A \in M_{n\times n}(\mathbb{R})$. Suppose $T : \mathbb{R}^{n+1} \to \mathbb{R}^{n+1}$ is linear with symmetric standard matrix $A \in M_{(n+1)\times(n+1)}(\mathbb{R})$. By the Eigenvalues of Symmetric Matrices lemma, A has an eigenvalue λ. Let $\vec{v}_0$ be an eigenvector of A with associated eigenvalue λ with (perhaps after scaling) $\|\vec{v}_0\| = 1$. By the Extending Bases Theorem and the Gram-Schmidt Process, there is an orthonormal basis $\vec{v}_0, \vec{v}_1, \ldots, \vec{v}_n$ for $\mathbb{R}^{n+1}$ that extends $\vec{v}_0$. Let $P \stackrel{def}{=} \begin{bmatrix}\vec{v}_0 & \vec{v}_1 & \cdots & \vec{v}_n\end{bmatrix} \in M_{(n+1)\times(n+1)}(\mathbb{R})$. Then P is orthogonal. The first column of P^TAP is $P^TAP\vec{e}_1 = P^TA\vec{v}_0 = P^T(\lambda\vec{v}_0) = \lambda P^T\vec{v}_0 = \lambda\vec{e}_1$. Moreover, P^TAP is symmetric because A is symmetric: $(P^TAP)^T = P^TA^TP = P^TAP$. Therefore there is a symmetric $B \in M_{n\times n}(\mathbb{R})$ with

$$P^TAP = \begin{bmatrix}\lambda & 0_{1\times n} \\ 0_{n\times 1} & B\end{bmatrix}.$$

By the induction hypothesis, there is an orthogonal $Q \in M_{n\times n}(\mathbb{R})$ with $Q^TBQ = D$ diagonal. Let $R = \begin{bmatrix}1 & 0_{1\times n} \\ 0_{n\times 1} & Q\end{bmatrix}$. Then R is orthogonal, since $R^TR = \begin{bmatrix}1 & 0_{1\times n} \\ 0_{n\times 1} & Q^TQ\end{bmatrix} = I_{n+1}$. Moreover,

$$R^TP^TAPR = \begin{bmatrix}1 & 0_{1\times n} \\ 0_{n\times 1} & Q^T\end{bmatrix}\begin{bmatrix}\lambda & 0_{1\times n} \\ 0_{n\times 1} & B\end{bmatrix}\begin{bmatrix}1 & 0_{1\times n} \\ 0_{n\times 1} & Q\end{bmatrix} = \begin{bmatrix}\lambda & 0_{1\times n} \\ 0_{n\times 1} & Q^TBQ\end{bmatrix} = \begin{bmatrix}\lambda & 0_{1\times n} \\ 0_{n\times 1} & D\end{bmatrix}$$

is diagonal. But since P and R are orthogonal matrices, PR an orthogonal matrix and $(PR)^{-1} = (PR)^T = R^TP^T$. Therefore A is orthogonally diagonalizable. By the Principle of Mathematical Induction, the proof is complete. □

Proof of Spectral Theorem: Using Coordinates. First suppose that $\mathscr{B}$ is an orthonormal eigenbasis for A, and let S be the matrix with columns given by the vectors in $\mathscr{B}$. Then there is a diagonal matrix D with $A = SDS^{-1} = SDS^T$. But then we have $A^T = (SDS^T)^T = (S^T)^T D^T S^T = SDS^T = A$, so that A is symmetric.

We prove the other direction via induction. If $n = 1$, then $[1]$ is an orthonormal eigenbasis for T.

Now let $n \in \mathbb{N}$ and assume the result holds for linear maps from $\mathbb{R}^n$ to $\mathbb{R}^n$. Suppose $T : \mathbb{R}^{n+1} \to \mathbb{R}^{n+1}$ is linear with symmetric standard matrix $A \in M_{(n+1)\times(n+1)}(\mathbb{R})$.

By the Eigenvalues of Symmetric Matrices lemma, A has an eigenvalue λ. Let $\vec{u}_0$ be an eigenvector of A with associated eigenvalue λ with (perhaps after scaling) $\|\vec{u}_0\| = 1$. Let $V = \text{span}(\vec{u}_0)^\perp$. By Problem 1 on Discussion 1, $\dim(V) = n$. Note that since $A\vec{u}_0) \in \text{span}(\vec{u}_0)$, $\text{span}(\vec{u}_0)$ is invariant under A. Because A is symmetric, Exercise 4 on Homework 1 implies that V is invariant under A as well. Define $T' : V \to V$ by $T'(\vec{x}) \stackrel{def}{=} A\vec{x}$. Let $\mathscr{B} = (\vec{v}_1, \ldots, \vec{v}_n)$ be an orthonormal basis for V, and let $Q = \begin{bmatrix} \vec{v}_1 & \cdots & \vec{v}_n \end{bmatrix} \in M_{(n+1)\times n}(\mathbb{R})$. Note that $Q^TQ = \begin{bmatrix} \vec{v}_j^T \vec{v}_k \end{bmatrix} = I_n$, and that for each $\vec{v} \in V$ we have $\vec{v} = Q\left[\vec{v}\right]_{\mathscr{B}}$ and $Q^T\vec{v} = \left[\vec{v}\right]_{\mathscr{B}}$.

We first show that $\left[T'\right]_{\mathscr{B}} \in M_{n\times n}(\mathbb{R})$ is symmetric. Because A is symmetric, it is sufficient to show that $\left[T'\right]_{\mathscr{B}} = Q^TAQ$. But the equality $Q^TAQ = \left[T'\right]_{\mathscr{B}}$ follows by noting, for each $1 \le k \le n$, that

$$(Q^TAQ)\vec{e}_k = Q^TAQ\left[\vec{v}_k\right]_{\mathscr{B}} = Q^TA\vec{v}_k = Q^TT'(\vec{v}_k) = \left[T'(\vec{v}_k)\right]_{\mathscr{B}} = \left[T'\right]_{\mathscr{B}}\left[\vec{v}_k\right]_{\mathscr{B}} = \left[T'\right]_{\mathscr{B}}\vec{e}_k.$$

By the induction hypothesis, $\left[T'\right]_{\mathscr{B}}$ has an orthonormal eigenbasis $\vec{u}'_1, \ldots, \vec{u}'_n$. For each $1 \le k \le n$, let $\vec{u}_k \stackrel{def}{=} Q\vec{u}'_k \in V = \text{span}(\vec{u}_0)^\perp$, and let λ_k be the eigenvalue of $\left[T'\right]_{\mathscr{B}}$ associated to $\vec{u}'_k$. We claim that $\vec{u}_0, \vec{u}_1, \ldots, \vec{u}_n$ is an orthonormal eigenbasis for T.

First note that $\vec{u}_0 \cdot \vec{u}_k = 0$ if $k \neq 0$ (since $\vec{u}_k \in \text{span}(\vec{u}_0)^\perp$) and $\vec{u}_0 \cdot \vec{u}_0 = \|\vec{u}_0\|^2 = 1$. For $1 \le j, k \le n$,

$$\vec{u}_j \cdot \vec{u}_k = (Q\vec{u}'_j) \cdot (Q\vec{u}'_k) = (Q^TQ\vec{u}'_j) \cdot \vec{u}'_k = \vec{u}'_j \cdot \vec{u}'_k = \begin{cases} 0 & \text{if } j \neq k, \\ 1 & \text{if } j = k. \end{cases}$$

Therefore $\vec{u}_0, \vec{u}_1, \ldots, \vec{u}_n$ is an orthonormal set of $n+1$ vectors in $\mathbb{R}^{n+1}$, hence an orthonormal basis.

Finally, $T(\vec{u}_0) = \lambda_0\vec{u}_0$, and for $1 \le k \le n$ we have $\vec{u}_k \in V$ so that

$$\left[T'(\vec{u}_k)\right]_{\mathscr{B}} = \left[T'\right]_{\mathscr{B}}\left[\vec{u}_k\right]_{\mathscr{B}} = \left[T'\right]_{\mathscr{B}}\vec{u}'_k = \lambda_k\vec{u}'_k = \left[\lambda_k\vec{u}_k\right]_{\mathscr{B}}, \text{ so that } T(\vec{u}_k) = T'(\vec{u}_k) = \lambda_k\vec{u}_k.$$

Therefore $\vec{u}_0, \vec{u}_1, \ldots, \vec{u}_n$ is an orthonormal eigenbasis for T, and the Principle of Mathematical Induction completes the proof. $\square$

Lecture 14: Quadratic Forms

Learning Objectives:

- Determine the matrix of a quadratic form.
- Analyze the definiteness of a quadratic form.
- Diagonalize a quadratic form.

Last time we learned the Spectral Theorem, which says that a necessary and sufficient condition that $A \in M_{n\times n}(\mathbb{R})$ be orthogonally diagonalizable is that A is symmetric. You will see many applications of this theorem on your homework, but today we investigate one important application with consequences throughout mathematics: quadratic forms. A quadratic form on $\mathbb{R}^n$ is a second-degree polynomial in $x_1, \ldots, x_n$ where the first-order and constant terms are all 0.

Definition 26. A function $q : \mathbb{R}^n \to \mathbb{R}$ is a **quadratic form** if for $1 \leq i, j \leq n$ there are scalars $b_{i,j} \in \mathbb{R}$ with

$$q(\vec{x}) = \sum_{i=1}^{n}\sum_{j=1}^{n} b_{i,j}x_i x_j.$$

The first link between quadratic forms and linear algebra is that each quadratic form can be expressed in the form $\vec{x} \cdot (B\vec{x})$ for a (not unique if $n \geq 2$) square matrix B.

Example 25. The function $q : \mathbb{R}^2 \to \mathbb{R}$, $q(\vec{x}) = -3x^2 - 4xy - 6x^2$ is a quadratic form. Note that we can write

$$q(\vec{x}) = \begin{bmatrix} x \\ y \end{bmatrix} \cdot \begin{bmatrix} -3x - 4y \\ -6y \end{bmatrix} = \vec{x} \cdot \left(\begin{bmatrix} -3 & -4 \\ 0 & -6 \end{bmatrix} \vec{x} \right)$$

for every $\vec{x} \in \mathbb{R}^2$.

Example 26. The function $q : \mathbb{R}^3 \to \mathbb{R}$, $q(\vec{x}) = -x^2 + 2y^2 + 2z^2 + 4xy - 4xz + 2yz$ is a quadratic form. Note that we can write

$$q(\vec{x}) = \begin{bmatrix} x \\ y \\ z \end{bmatrix} \cdot \begin{bmatrix} -x + 4y - 4z \\ 2y + 2z \\ 2z \end{bmatrix} = \vec{x} \cdot \left(\begin{bmatrix} -1 & 4 & -4 \\ 0 & 2 & 2 \\ 0 & 0 & 2 \end{bmatrix} \vec{x} \right).$$

We will use quadratic forms later in the year to approximate scalar-valued functions[5]. For a given quadratic form q on $\mathbb{R}^n$ (with $n \geq 2$) there are infinitely many matrices B such that $q(\vec{x}) = \vec{x} \cdot (B\vec{x})$. However, there is exactly one *symmetric* matrix with this property.

Theorem 23 (Matrix of a Quadratic Form)**.** Let $q : \mathbb{R}^n \to \mathbb{R}$ be a quadratic form. There is a unique symmetric matrix $A \in M_{n\times n}(\mathbb{R})$ (called the **matrix of** q) with $q(\vec{x}) = \vec{x} \cdot (A\vec{x})$.

Moreover, if $B \in M_{n\times n}(\mathbb{R})$ is any matrix that satisfies $q(\vec{x}) = \vec{x} \cdot (B\vec{x})$, then $A = \frac{1}{2}(B^T + B)$.

[5]Indeed, they will be used to establish the multivariable version of the second derivative test.

Proof. We first show that there is such a symmetric matrix A. Note that there is at least one (not necessarily symmetric) matrix B such that $q(\vec{x}) = \vec{x} \cdot (B\vec{x})$ for every $\vec{x} \in \mathbb{R}^n$, for if $B = [b_{ij}] \in M_{n\times n}(\mathbb{R})$ then, by linearity of $\cdot$ and writing $\vec{x} = \sum_{k=1}^{n} x_k \vec{e}_k$,

$$q(\vec{x}) = \sum_{i=1}^{n}\sum_{j=1}^{n} b_{ij}x_i x_j = \sum_{i=1}^{n}\sum_{j=1}^{n} x_i x_j(\vec{e}_i \cdot (B\vec{e}_j)) = \sum_{i=1}^{n} x_i(\vec{e}_i \cdot (B\vec{x})) = \vec{x} \cdot (B\vec{x}) \quad \text{for every } \vec{x} \in \mathbb{R}^n.$$

On the other hand, suppose that B is any matrix satisfying $q(\vec{x}) = \vec{x} \cdot (B\vec{x})$ for every $\vec{x} \in \mathbb{R}^n$. Then for every $\vec{x} \in \mathbb{R}^n$ we have

$$q(\vec{x}) = \vec{x} \cdot (B\vec{x}) = \frac{1}{2}\Big((B\vec{x}) \cdot \vec{x} + \vec{x} \cdot (B\vec{x})\Big) = \frac{1}{2}\Big(\vec{x} \cdot (B^T\vec{x}) + \vec{x} \cdot (B\vec{x})\Big) = \vec{x} \cdot (\frac{1}{2}(B^T + B)\vec{x}).$$

Because $\Big(\frac{1}{2}(B^T+B)\Big)^T = \frac{1}{2}(B^T+B)$, there is at least one symmetric matrix A satisfying the conclusion of the theorem.

It only remains to show that A is unique. Suppose that $A, C \in M_{n\times n}(\mathbb{R})$ are symmetric and $q(\vec{x}) = \vec{x} \cdot (A\vec{x})$ and $q(\vec{x}) = \vec{x} \cdot (C\vec{x})$ for all $\vec{x} \in \mathbb{R}^n$. Then $S = A - C$ is symmetric and $\vec{x} \cdot (S\vec{x}) = \vec{x} \cdot (A\vec{x}) - \vec{x} \cdot (C\vec{x}) = 0$ for all $\vec{x} \in \mathbb{R}^n$. By the Spectral Theorem, S has an orthonormal eigenbasis $\vec{u}_1, \ldots, \vec{u}_n$. For $1 \le k \le n$, let λ_k be the eigenvalue of S associated to the eigenvector $\vec{u}_k$. Then $0 = \vec{u}_k \cdot (S\vec{u}_k) = \vec{u}_k \cdot (\lambda_k\vec{u}_k) = \lambda_k(\vec{u}_k \cdot \vec{u}_k) = \lambda_k$. Therefore the only eigenvalue of S is 0. By one of your homework problems, $S = 0I_n = 0_{n\times n}$, and therefore $C = A$.

□

Example 27. Find the matrix of the quadratic form $q : \mathbb{R}^2 \to \mathbb{R}$, $q(\vec{x}) = -3x^2 - 4xy - 6y^2$.

Because we can write $q(\vec{x}) = \vec{x} \cdot \left(\begin{bmatrix} -3 & -4 \\ 0 & -6 \end{bmatrix}\vec{x}\right)$, the matrix of q is

$$A = \frac{1}{2}\Big(\begin{bmatrix} -3 & -4 \\ 0 & -6 \end{bmatrix}^T + \begin{bmatrix} -3 & -4 \\ 0 & -6 \end{bmatrix}\Big) = \frac{1}{2}\begin{bmatrix} -6 & -4 \\ -4 & -12 \end{bmatrix} = \begin{bmatrix} -3 & -2 \\ -2 & -6 \end{bmatrix}.$$

Therefore we can also write $q(\vec{x}) = \vec{x} \cdot (A\vec{x})$.

Example 28. Find the matrix of the quadratic form $q : \mathbb{R}^3 \to \mathbb{R}$, $q(\vec{x}) = -x^2+2y^2+2z^2+4xy-4xz+2yz$.

Because we can write $q(\vec{x}) = \vec{x} \cdot \left(\begin{bmatrix} -1 & 4 & -4 \\ 0 & 2 & 2 \\ 0 & 0 & 2 \end{bmatrix}\vec{x}\right)$, the matrix of q is

$$A = \frac{1}{2}\left(\begin{bmatrix} -1 & 4 & -4 \\ 0 & 2 & 2 \\ 0 & 0 & 2 \end{bmatrix}^T + \begin{bmatrix} -1 & 4 & -4 \\ 0 & 2 & 2 \\ 0 & 0 & 2 \end{bmatrix}\right) = \frac{1}{2}\begin{bmatrix} -2 & 4 & -4 \\ 4 & 4 & 2 \\ -4 & 2 & 4 \end{bmatrix} = \begin{bmatrix} -1 & 2 & -2 \\ 2 & 2 & 1 \\ -2 & 1 & 2 \end{bmatrix}.$$

Therefore we also have $q(\vec{x}) = \vec{x} \cdot (A\vec{x})$.

Example 29. Sketch the graph of the equation $-3x^2 - 4xy - 6y^2 = -1$.

We are attempting to understand the equation

$$-1 = -3x^2 - 4xy - 6y^2 = \vec{x} \cdot \left(\begin{bmatrix} -3 & -2 \\ -2 & -6 \end{bmatrix} \vec{x} \right) = q(\vec{x}).$$

This equation is complicated using the standard coordinate system, but because the matrix $A = \begin{bmatrix} -3 & -2 \\ -2 & -6 \end{bmatrix}$ is symmetric we know (via the Spectral Theorem) that A has an orthonormal eigenbasis. Working with this eigenbasis will make the graph of this equation easier to understand.

First note that the characteristic equation of A is $0 = \lambda^2 + 9\lambda + 14 = (\lambda + 7)(\lambda + 2)$, so that the eigenvalues of A are $\lambda = -7, -2$. The eigenspaces are

$$E_{-7} = \ker(A + 7I_2) = \ker\left(\begin{bmatrix} 4 & -2 \\ -2 & 1 \end{bmatrix} \right) = \operatorname{span}\left(\begin{bmatrix} 1 \\ 2 \end{bmatrix} \right),$$

$$E_{-2} = \ker(A + 2I_2) = \ker\left(\begin{bmatrix} -1 & -2 \\ -2 & -4 \end{bmatrix} \right) = \operatorname{span}\left(\begin{bmatrix} -2 \\ 1 \end{bmatrix} \right).$$

Therefore $\mathscr{B} = (\vec{u}_1, \vec{u}_2) = (\begin{bmatrix} 1/\sqrt{5} \\ 2/\sqrt{5} \end{bmatrix}, \begin{bmatrix} -2/\sqrt{5} \\ 1/\sqrt{5} \end{bmatrix})$ is an orthonormal eigenbasis for A.

Because $\vec{x} = \underbrace{(\vec{x} \cdot \vec{u}_1)}_{c_1} \vec{u}_1 + \underbrace{(\vec{x} \cdot \vec{u}_2)}_{c_2} \vec{u}_2$ for every $\vec{x} \in \mathbb{R}^2$,

$$\begin{aligned} q(\vec{x}) &= (c_1\vec{u}_1 + c_2\vec{u}_2) \cdot (A(c_1\vec{u}_1 + c_2\vec{u}_2)) \\ &= (c_1\vec{u}_1 + c_2\vec{u}_2) \cdot (-7c_1\vec{u}_1 - 2c_2\vec{u}_2) \\ &= -7c_1^2 - 2c_2^2. \end{aligned}$$

So, the curve $q(\vec{x}) = -1$ is the same as the curve (in c_1, c_2 coordinates) $-7c_1^2 - 2c_2^2 = -1$, or rather

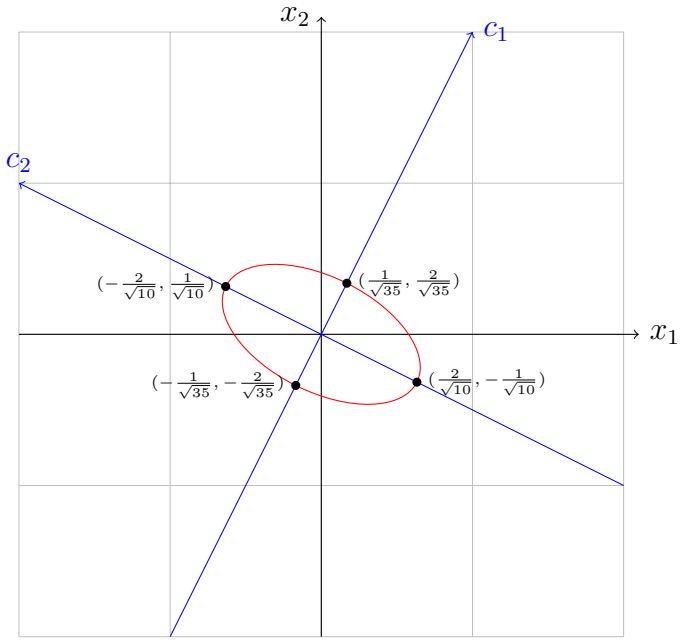
$$7c_1^2 + 2c_2^2 = 1.$$

This is the equation of an ellipse, which we can easily plot in $\mathbb{R}^2$ if we superimpose our c_1, c_2 coordinate system. The ellipse intersects the c_1-axis at the points $[\vec{x}]_{\mathscr{B}} = \begin{bmatrix} c_1 \\ c_2 \end{bmatrix} = \begin{bmatrix} \pm 1/\sqrt{7} \\ 0 \end{bmatrix}$, which corresponds to the points

$$\vec{x} = \pm\frac{1}{\sqrt{7}}\vec{u}_1 + 0\vec{u}_2 = \pm \begin{bmatrix} 1/\sqrt{35} \\ 2/\sqrt{35} \end{bmatrix},$$

an intersects the c_2-axis at the points $[\vec{x}]_{\mathscr{B}} = \begin{bmatrix} c_1 \\ c_2 \end{bmatrix} = \begin{bmatrix} 0 \\ \pm 1/\sqrt{2} \end{bmatrix}$, which corresponds to the points

$$\vec{x} = 0\vec{u}_1 \pm \frac{1}{\sqrt{2}}\vec{u}_2 = \pm \begin{bmatrix} -2/\sqrt{10} \\ 1/\sqrt{10} \end{bmatrix}.$$



The c_1 and c_2 axes (i.e. the spans of $\vec{u}_1$ and $\vec{u}_2$) are the **principal axes** of the ellipse. In the n-dimensional case, we typically only talk about principal axes if all of the eigenspaces of A are 1-dimensional (i.e. if A has n distinct eigenvalues).

The crucial observation in the previous example—that we can simplify our understanding of a quadratic form q by orthogonally diagonalizing its matrix—generalizes to higher dimensions.

Theorem 24 (Diagonalizing a Quadratic Form)**.** Let $q : \mathbb{R}^n \to \mathbb{R}$, and let A be the matrix of q. Then there is an orthonormal eigenasis $\vec{u}_1, \ldots, \vec{u}_n$ for A such that

$$q(\vec{x}) = \lambda_1(\vec{x} \cdot \vec{u}_1)^2 + \cdots + \lambda_n(\vec{x} \cdot \vec{u}_n)^2 \quad \text{for every } \vec{x} \in \mathbb{R}^n,$$

where λ_k is the eigenvalue associated to $\vec{u}_k$ for each $1 \leq k \leq n$.

Remark 48. Note that if $\mathscr{B} = (\vec{u}_1, \ldots, \vec{u}_n)$, then $\vec{x} = (\vec{x} \cdot \vec{u}_1)\vec{u}_1 + \cdots + (\vec{x} \cdot \vec{u}_n)\vec{u}_n$, and therefore $[\vec{x}]_{\mathscr{B}} = \begin{bmatrix} \vec{x} \cdot \vec{u}_1 \\ \vdots \\ \vec{x} \cdot \vec{u}_n \end{bmatrix}$. Therefore we can simplify the equation in the statement of the theorem by saying that if $[\vec{x}]_{\mathscr{B}} = \begin{bmatrix} c_1 \\ \vdots \\ c_n \end{bmatrix}$, then $q(\vec{x}) = \lambda_1 c_1^2 + \cdots + \lambda_n c_n^2$. In other words, there is an orthonormal coordinate system for $\mathbb{R}^n$ that transforms the quadratic form q into one with a diagonal matrix!

Proof. Let A be the matrix for q. By the Spectral Theorem, A has an orthonormal eigenbasis $\vec{u}_1, \ldots, \vec{u}_n$. For each $1 \leq k \leq n$, let λ_k be the eigenvalue of A associated to $\vec{u}_k$. Then for each $\vec{x} \in \mathbb{R}^n$ we have

$$\begin{aligned} q(\vec{x}) &= \vec{x} \cdot (A\vec{x}) \\ &= \vec{x} \cdot (A((\vec{x} \cdot \vec{u}_1)\vec{u}_1 + \cdots + (\vec{x} \cdot \vec{u}_n)\vec{u}_n)) \\ &= \vec{x} \cdot (\lambda_1(\vec{x} \cdot \vec{u}_1)\vec{u}_1 + \cdots + \lambda_n(\vec{x} \cdot \vec{u}_n)\vec{u}_n) \\ &= \lambda_1(\vec{x} \cdot \vec{u}_1)^2 + \cdots + \lambda_n(\vec{x} \cdot \vec{u}_n)^2, \end{aligned}$$

where we used the fact that $\vec{u}_1, \ldots, \vec{u}_n$ is orthonormal to express $\vec{x}$ as a linear combination of $\vec{u}_1, \ldots, \vec{u}_n$ and to compute the resulting dot product. □

The eigenvalues of the matrix of a quadratic form q tell us a lot about the behavior of q as a function from $\mathbb{R}^n$ to $\mathbb{R}$, especially in terms of whether the values of $q(\vec{x})$ for $\vec{x} \neq \vec{0}$ are positive, negative, or a mixture. Later on we will use this information to determine whether $q(\vec{0}) = 0$ is a local maximum value of q, a local minimum value of q, or neither. To capture this, we make the following definition.

Definition 27. Let $q : \mathbb{R}^n \to \mathbb{R}$ be a quadratic form with matrix $A \in M_{n\times n}(\mathbb{R})$.

(i) q is **positive definite** (resp. **positive semi-definite**) if all of the eigenvalues of A are positive (resp. non-negative).

(ii) q is **negative definite** (resp. **negative semi-definite**) if all of the eigenvalues of A are negative (resp. non-positive).

(iii) q is **indefinite** if A has at least one positive and one negative eigenvalue.

Example 30. The quadratic form $q : \mathbb{R}^3 \to \mathbb{R}$ given by

$$q(\vec{x}) = \vec{x} \cdot \left(\begin{bmatrix} -1 & 2 & -2 \\ 2 & 2 & 1 \\ -2 & 1 & 2 \end{bmatrix} \vec{x} \right)$$

is indefinite because the eigenvalues of its matrix are 3 and -3. If $\mathscr{B} = (\vec{u}_1, \vec{u}_2, \vec{u}_3)$ is the orthonormal eigenbasis

$$\vec{u}_1 = \begin{bmatrix} 1/\sqrt{5} \\ 2/\sqrt{5} \\ 0 \end{bmatrix}, \quad \vec{u}_2 = \begin{bmatrix} -2/\sqrt{30} \\ 1/\sqrt{30} \\ 5/\sqrt{30} \end{bmatrix}, \quad \vec{u}_3 = \begin{bmatrix} 2/\sqrt{6} \\ -1/\sqrt{6} \\ 1/\sqrt{6} \end{bmatrix}$$

for the matrix of q computed in a past example, then

$$q(\vec{x}) = 3(\vec{x} \cdot \vec{u}_1)^2 + 3(\vec{x} \cdot \vec{u}_2)^2 - 3(\vec{x} \cdot \vec{u}_3)^2 = 3c_1^2 + 3c_2^2 - 3c_3^2$$

for every $\vec{x} \in \mathbb{R}^3$, where $[\vec{x}]_{\mathscr{B}} = \begin{bmatrix} c_1 \\ c_2 \\ c_3 \end{bmatrix}$.

Example 31. The quadratic form $q : \mathbb{R}^2 \to \mathbb{R}$ given by

$$q(\vec{x}) = \vec{x} \cdot \left(\begin{bmatrix} -3 & -2 \\ -2 & -6 \end{bmatrix} \vec{x} \right)$$

is negative definite because the eigenvalues of the matrix of q are -7 and -2.

Lecture 15: Quadric Surfaces

Learning Objectives:

- Recognize and name various quadric surfaces.
- Sketch the graphs of quadric surfaces by analyzing their cross-sections.

Today we pivot from our study of linear algebra to our study of multivariable calculus by studying a class of objects that are important to both topics: the quadric surfaces.

Definition 28. A **quadric surface** is a subset of $\mathbb{R}^3$ described by an equation of the form[6]

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

where $A, B, C, D, E, F, G, H, I, J$ are constants (and at least one of A,B,C,D,E, or F are nonzero).

Remark 49. Our general approach to studying quadric surfaces is as follows. We are being somewhat general here because a precise description of this approach would be unwieldy.

(i) Choose an orthonormal basis $\mathscr{B} = (\vec{u}_1, \vec{u}_2, \vec{u}_3)$ for $\mathbb{R}^3$ such that, with respect to the new coordinates $[\vec{x}]_{\mathscr{B}} = \begin{bmatrix} r \\ s \\ t \end{bmatrix}$ relative to this basis the equation for the quadric surface becomes (perhaps after rearranging the lower order terms)

$$A'r^2 + B's^2 + C't^2 + G'r + H's + I't + J' = 0.$$

Note that this change of basis essentially amounts to rotating (or reflecting) space.

(ii) Complete the square in each variable (if necessary) to simplify the equation further. This has the effect of replacing the equation defining the surface with a simpler equation, but where r, s, and t are replaced by $r - a$, $s - b$, and $t - c$ for suitable constants a, b, c. This allows us to view the surface as a translated version (centered at (a, b, c)) of a simpler surface centered at $(0, 0, 0)$.

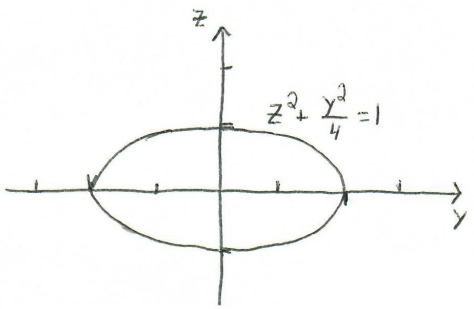
In this way (by changing coordinates, and perhaps "translating" the surface by completing the square) we can reduce the equation for a quadric surface to one of a small handful of standard examples. Pages 94 and 95 of your book (Colley) list different basic surfaces that appear, but the exhaustive list is longer still. You should learn the names of these surfaces by heart. Rather than memorize all of the facts about these surfaces, we will instead focus on analyzing them using their **sections** (i.e. their intersections with various planes).

Remark 50. Back in 2004, a mathematician named Jonathan Rogness created an excellent online visualization tool for quadric surfaces. This tool was updated by Nathan Dunfield in 2016. The tool shows all of the standard quadric surfaces, and allows you to play around with sections and coefficients. You can access the tool on the math department server at the following link: `http://www.math.northwestern.edu/~aaron/quadrics/index.html`.

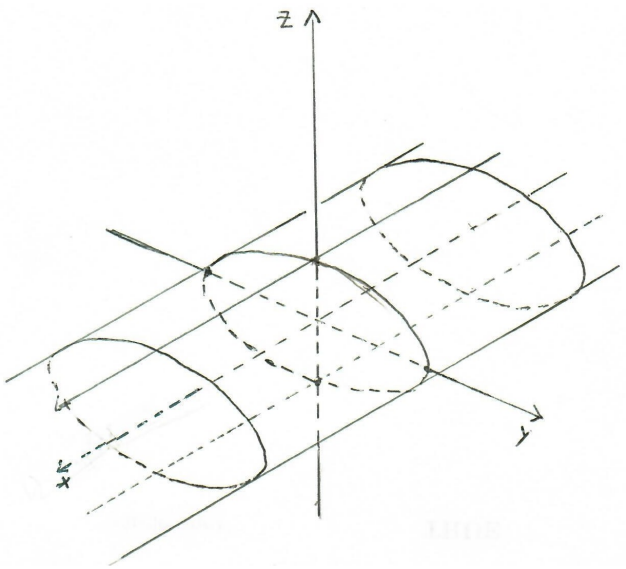
[6] Some of these "quadric surfaces" are not actually surfaces in the rigorous sense. For example, the "quadric surface" $x^2 + y^2 + z^2 = 0$ is exactly the single point $(0, 0, 0)$. Nevertheless, to avoid a proliferation of language we will use the name "quadric surface" to refer to any object of the type described in the definition.

Example 32. Sketch the graph of the quadric surface defined by $\frac{y^2}{4} + z^2 = 1$.

This equation does not involve the variable x. If the point $(0, y, z)$ in the yz-plane satisfies the equation, then all points of the form (x, y, z) satisfy the equation; each such point $(0, y, z)$ therefore determines a line parallel to the x-axis that lies on the surface and intersects the yz-plane at $(0, y, z)$. The graph of $\frac{y^2}{4} + z^2 = 1$ is an ellipse in the yz-plane, centered at the origin, with minor axis 1 in the y-direction and minor axis 2 in the z-direction:



Therefore the graph of $\frac{y^2}{4} + z^2 = 1$ in $\mathbb{R}^3$ consists of all lines parallel to the x-axis that intersect the yz-plane through points on the ellipse above. Here is the sketch:



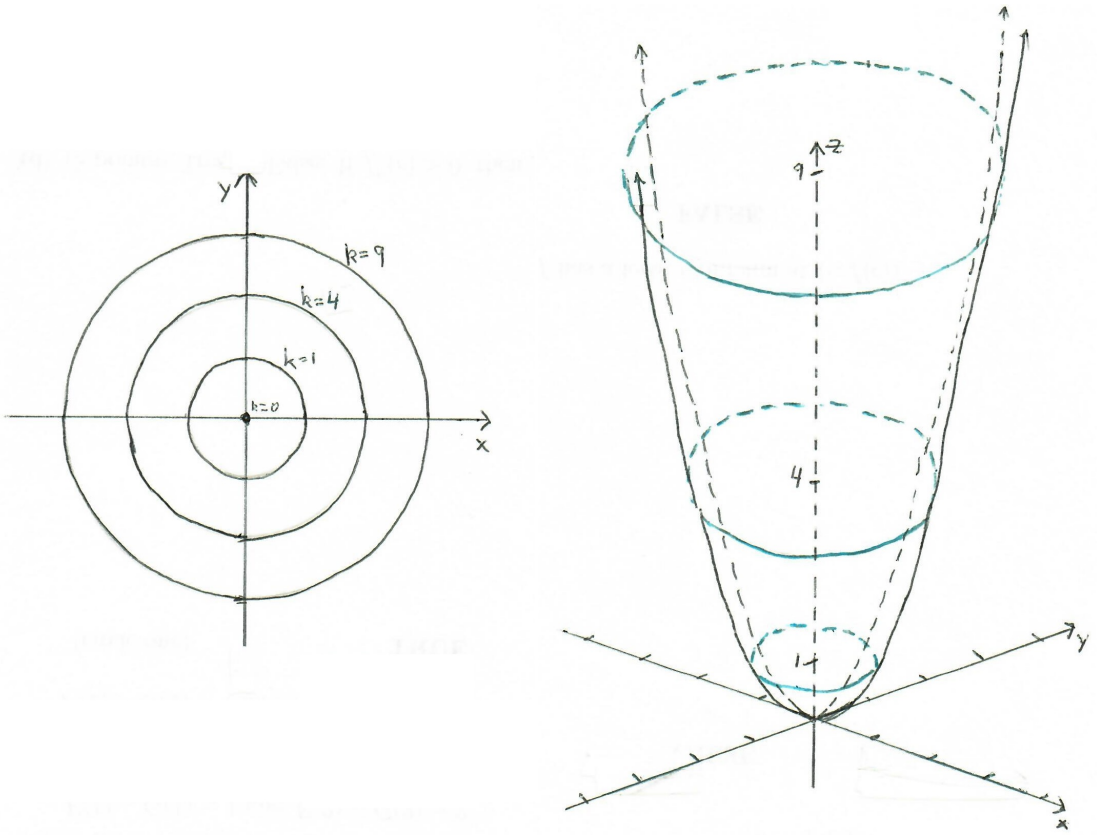
The graph of an equation like the one above (i.e. one that omits at least one of the variables) is called a **cylinder**. Because the sections of this surface with planes of the form $x = k$ for some number k (here x is important because it is the 'missing variable' in the equation $\frac{y^2}{4} + z^2 = 1$) are the ellipse $\frac{y^2}{4} + z^2 = 1$, we can give this the more descriptive name of **elliptic cylinder**.

Taking sections of a surface with planes of the form $x = k$, $y = k$, or $z = k$ is a great technique for visualizing the graph of the surface.

Example 33. Sketch the surface described by $-x^2 - y^2 + z = 0$.

To get an idea of what the graph of this surface looks like, let's use sections. If we rewrite the equation as $z = x^2 + y^2$, then for fixed z this almost looks like the equation of a circle in the xy-plane with radius $\sqrt{z}$. Let's therefore look at the z-sections of this surface. We intersect the surface with planes of the form $z = k$ for various values of k.

The intersection of the surface with the plane $z = k$ gives the curve $x^2 + y^2 = k$. No x and y solve this equation if $k < 0$. If $k \geq 0$, then this equation describes the circle (in the plane $z = k$) of radius $\sqrt{k}$ centered at $(0, 0, k)$. In particular, when $k = 0$ this is only the point $(0, 0, 0)$. When $k = 1$ this is the circle $x^2 + y^2 = 1$ (in the plane $z = 1$). When $k = 4$ this is the circle $x^2 + y^2 = 4$ (in the plane $z = 4$). When $k = 9$, this is the circle $x^2 + y^2 = 9$ (in the plane $z = 9$). We sketch these circles below, and then use these to piece together the graph:



Alternatively, we could look at sections with respect to the other variables. Indeed, the sections of the surface with planes of the form $y = k$ have the form $z = x^2 + k^2$, which are just vertically-shifted versions of the parabola $z = x^2$. When $k = 0$ we have the parabola $z = x^2$. When $k = 1$ or $k = -1$ we have the parabola $z = x^2 + 1$. When $k = 2$ or $k = -2$ we have $z = x^2 + 4$, etc. We could keep computing these until we are confident that we understand what the surface looks like. In class we used the software above to see these sections superimposed on the surface.

This is an example of an **elliptic paraboloid**. Paraboloids are the surface analogues of parabolas. Your book also has a worked out example of an **hyperbolic paraboloid** using the same process.

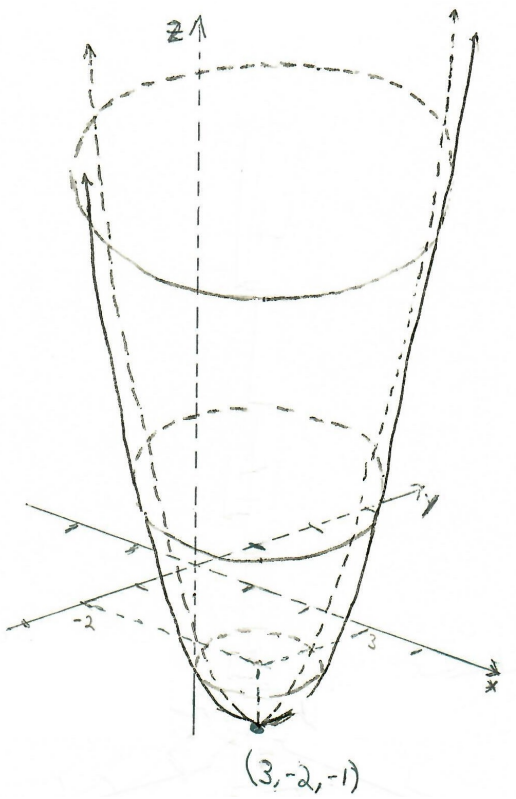
Elliptic and hyperbolic paraboloids all have equations of the form $Ax^2 + By^2 + Cz = 0$ (or perhaps with the roles of x, y, z interchanged, like $Az^2 + By^2 + Cx = 0$). The size of the constants A, B, C have the effect of 'stretching' or 'compressing' the surface. For example, holding A and B the same but increasing the size C has the effect of 'flattening' the paraboloid in the z-direction. Similarly, holding B and C the same but increasing the size of A has the effect of 'flattening' the paraboloid in the x-direction. This can be determined from the equation itself, but you can get a lot of intuition for this by playing around with the online application linked to above. (The same goes for the other quadric surfaces!)

Example 34. The quadric surface $z - x^2 + 6x - y^2 - 4y = 12$ looks similar to the previous example,

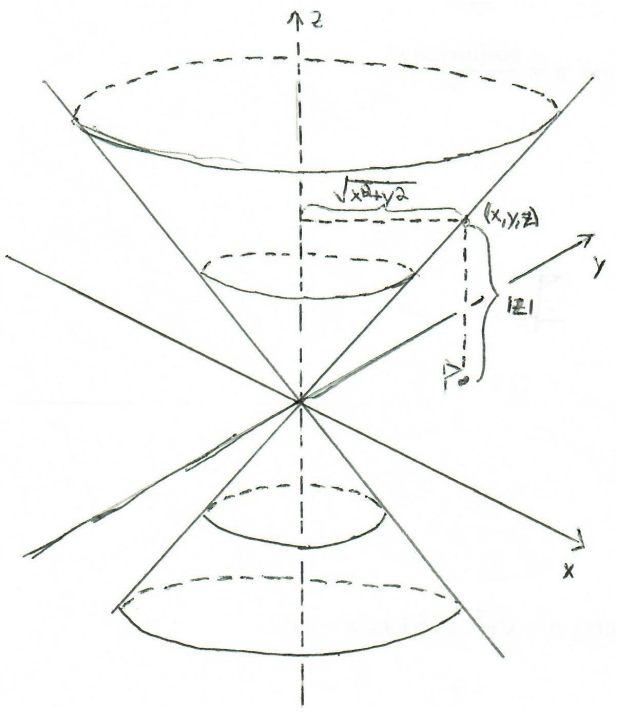
but it includes terms like $6x$ and $-4y$. These can be absorbed by completing the square as follows:

$$\begin{aligned}
12 &= z - x^2 + 6x - y^2 - 4y \\
&= z - (x^2 - 6x) - (y^2 + 4y) \\
&= z - (x^2 - 6x + 9 - 9) - (y^2 + 4y + 4 - 4) \\
&= z - ((x-3)^2 - 9) - ((y+2)^2 - 4) \\
&= z - (x-3)^2 + 9 - (y+2)^2 + 4 \\
&= (z+1) - (x-3)^2 - (y+2)^2 + 12
\end{aligned}$$

or rather $0 = (z+1) - (x-3)^2 - (y+2)^2$. The graph of this is the same as the one from the previous example, except it is shifted 1 unit in the negative z-direction, 3 units in the positive x-direction, and 2 units in the negative y direction. Here is a sketch:



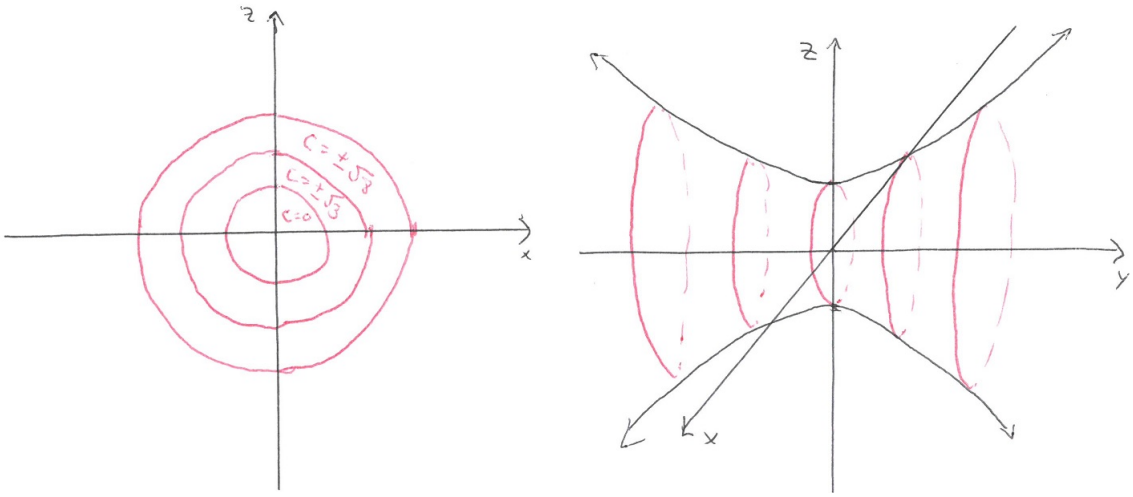
Example 35. The equation $z^2 = x^2 + y^2$ can be solved for z as $|z| = \sqrt{x^2 + y^2}$. Since $\sqrt{x^2 + y^2}$ is the distance from (x, y, z) to the z-axis, and since $|z|$ is the distance from (x, y, z) to the xy-plane, this surface consists of all points whose distance from the xy-plane is the same as the distance from the z-axis. In other words, this is a **double cone**:



Alternatively, we can analyze this using sections: Note that the sections of this surface in planes of the form $z = k$ all have the form $k^2 = x^2 + y^2$, which is exactly the circle centered at $(0, 0, k)$ of radius $|k|$ (note that k might be negative here). We've sketched a few of these circles in the picture above.

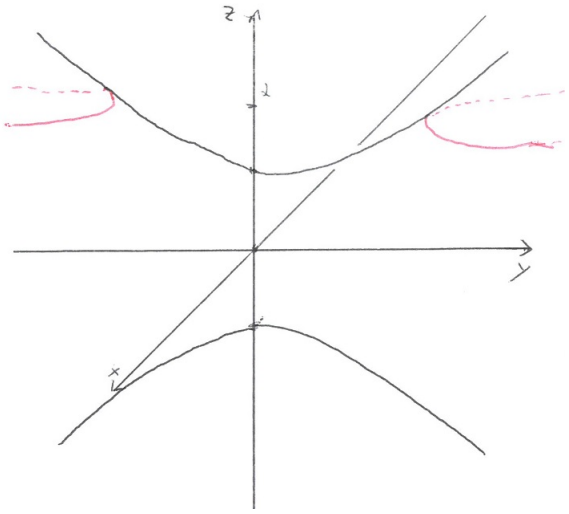
Example 36. Describe the surface given by the equation $x^2 - y^2 + z^2 = 1$.

For this surface, the easiest sections to think about are in planes of the form $y = k$; these sections have the form $x^2 + z^2 = 1 + k^2$ when $y = k$, which are again circles! We sketch several of these below, and use them to construct the graph of the full equation.



As before, we can also find the z-sections of the graph: the intersection of the graph with the plane $z = k$ yields $x^2 - y^2 = 1 - k^2$, which is a hyperbola in the plane $z = k$ which has different behavior for different values of k. If $|k| < 1$, then the hyperbola intersects the xz-plane at $(\pm\sqrt{1-k^2}, 0, k)$ and has no yz-plane intercepts. If $|k| = 1$, then this is just $(x - y)(x + y) = 0$, which are the two crossed lines $x = y$ and $x = -y$ in the plane $z = k$. If $|k| > 1$, then this hyperbola does not intersect the xz-plane,

but rather intersects the yz-plane in the points $(0, \pm\sqrt{k^2-1}, k)$. We sketch one of the z-sections ($z = 2$) below:



This surface is the second quadric analogue of the hyperbola; because it is in one piece, we call it a **one-sheeted hyperboloid**.

Example 37. Describe the surface whose equation is $-2xy - 2yz - 2xz = 1$.

The left-hand side is a quadratic form with matrix $\begin{bmatrix} 0 & -1 & -1 \\ -1 & 0 & -1 \\ -1 & -1 & 0 \end{bmatrix}$. This matrix has eigenvalues $1, 1, -2$ (repeated according to algebraic multiplicity, and therefore geometric multiplicity because the matrix is symmetric), so by choosing coordinates relative to an orthonormal eigenbasis of this matrix, the equation for our surface becomes

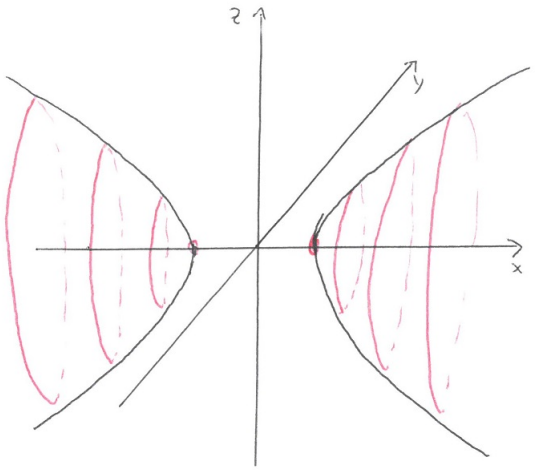
$$c_1^2 + c_2^2 - 2c_3^2 = 1,$$

which describes a one-sheeted hyperboloid. So, our surface is a (rotated) one-sheeted hyperboloid.

Example 38. Let's use sections to sketch the graph of the equation $4x^2 - y^2 - z^2 = 1$.

This equation looks similar to the equation of the one-sheeted hyperboloid $x^2 - y^2 + z^2 = 1$ that we looked at in a previous example. As we shall see, though, this surface is actually quite different (besides the permutation of variables).

The x-sections of the graph are seen to be circles: The intersection of the graph with the plane $x = k$ yields $4k^2 - 1 = y^2 + z^2$, which is the circle with radius $\sqrt{4k^2-1}$ (which of course only makes sense if $|k| \geq \frac{1}{2}$) in the plane $x = k$. We sketch these sections below on the graph.



Alternatively, we see that the sections of this surface in planes of the form $z = k$ have the form $4x^2 - y^2 = k^2 + 1$, which is the equation of a hyperbola.

This is a **two-sheeted hyperboloid**, and is one of the quadric analogues of hyperbolas.

Lecture 16: Functions of Several Variables

Learning Objectives:

- Wield the basic definitions of multivariable functions.
- Compute the domain of a function.
- Analyze the graph of a function, and sketch or describe it via level sets.

Example 39 (Warm-Up)**.** Describe the surface whose equation is $-2xy - 2yz - 2xz = 1$.

The left-hand side is a quadratic form with matrix $\begin{bmatrix} 0 & -1 & -1 \\ -1 & 0 & -1 \\ -1 & -1 & 0 \end{bmatrix}$. This matrix has eigenvalues $1, 1, -2$ (repeated according to algebraic multiplicity, and therefore geometric multiplicity because the matrix is symmetric), so by choosing coordinates relative to an orthonormal eigenbasis of this matrix, the equation for our surface becomes

$$c_1^2 + c_2^2 - 2c_3^2 = 1,$$

which describes a one-sheeted hyperboloid. So, our surface is a (rotated) one-sheeted hyperboloid.

Functions of Several Variables

We now embark on the multivariable calculus portion of the course by recalling some terms used to describe properties of functions.

Remark 51. As mentioned at the beginning of MATH 291-1, in this stage of the course we will often need to consider $\vec{x} \in \mathbb{R}^n$ as both a point (a location in n-dimensional space) and as a vector. We will often use the notation $\vec{x} = (x_1, \ldots, x_n)$ and $\vec{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} = x_1\vec{e}_1 + \cdots + x_n\vec{e}_n$ interchangeably to avoid complicating the notation, but in a more careful treatment of this material (say, in a differential geometry course) it is helpful to make a systemic difference between these two ways of viewing objects. When helpful, we will use the language 'point' and 'vector' to highlight the important interpretation of $\vec{x}$ in a given context.

Remark 52. Your book uses (and we will use) the notation $\vec{i}$, $\vec{j}$, and $\vec{k}$ for the standard basis vectors $\vec{e}_1$, $\vec{e}_2$, and $\vec{e}_3$ in $\mathbb{R}^3$ (and sometimes $\vec{i}, \vec{j}$ for $\vec{e}_1, \vec{e}_2$ in $\mathbb{R}^2$). This notation is typical in the sciences, and we will adopt it for specific examples or for results that are stated only in terms of $\mathbb{R}^3$ or $\mathbb{R}^2$.

Example 40. Let $Q : \mathbb{R}^3 \to \mathbb{R}$ be defined by $Q(x, y, z) = x^2 + y^2 - 2z^2$. The **domain** (i.e. the set of inputs) of Q is all of $\mathbb{R}^3$, and the **codomain** (i.e. the space in which the outputs live) is $\mathbb{R}$.

The **range** (or **image**) of Q (i.e. the set of all outputs of Q) is also $\mathbb{R}$. To show this, suppose that $t \in \mathbb{R}$. If $t \geq 0$, then we have $Q(\sqrt{t}, 0, 0) = t$. If $t \leq 0$, then $Q(0, 0, \sqrt{-t/2}) = -2(\sqrt{-t/2})^2 = t$. Therefore t is in the range of Q. This is because $Q(0, 0, z) = -2z^2$ takes on all possible non-positive numbers as outputs, while $Q(x, 0, 0) = x^2$ assumes all non-negative numbers as outputs. Because the range of Q is equal to $\mathbb{R}$, Q is surjective. The term *onto* is sometimes used interchangeably with surjectivity; here, we would say that Q maps $\mathbb{R}^3$ **onto** $\mathbb{R}$.

However, because $Q(-1,0,0) = Q(1,0,0)$, and therefore we can find two different inputs to Q which yield the same output, Q is not injective. (The adjective **one-to-one** is used interchangeably with injective.)

Remark 53. If the outputs of a function are real numbers, then we typically call that function **scalar-valued**. Functions whose outputs are vectors or points in $\mathbb{R}^m$ (for $m > 1$) are typically called **vector-valued**.

Remark 54. If a (scalar- or vector-valued) function f is defined by a formula using inputs from $\mathbb{R}^n$, then (unless specified otherwise) the domain of f is understood to be the largest subset Ω of points in $\mathbb{R}^n$ for which the formula f is defined.

Example 41. Define[7] $\vec{N} : \mathbb{R}^n \backslash \{\vec{0}\} \to \mathbb{R}^n$ by $\vec{N}(\vec{x}) \stackrel{def}{=} \frac{1}{\|\vec{x}\|}\vec{x}$. Its image is the $(n-1)-$dimensional **unit sphere** $S^{n-1} = \{\vec{x} \in \mathbb{R}^n \ : \ \|\vec{x}\| = 1\}$ in $\mathbb{R}^n$. $\vec{N}$ does not map $\mathbb{R}^n \backslash \{\vec{0}\}$ onto $\mathbb{R}^n$, and since $\vec{N}(\lambda\vec{x}) = \vec{N}(\vec{x})$ for every nonzero $\vec{x} \in \mathbb{R}^n$ and $\lambda > 0$, $\vec{N}$ is not one-to-one.

Note that we can write $\vec{N}(\vec{x}) = (N_1(\vec{x}), \ldots, N_n(\vec{x}))$ or $N(\vec{x}) = N_1(\vec{x})\vec{e}_1 + \cdots + N_n(\vec{x})\vec{e}_n$, where $N_j(\vec{x}) = \frac{x_j}{\|\vec{x}\|}$. Therefore we can think of $N_j : \mathbb{R}^n \backslash \{\vec{0}\} \to \mathbb{R}$ as a scalar-valued function such that $N_j(\vec{x})$ is the j-th coordinate of $\vec{N}(\vec{x})$.

Definition 29. Let $\Omega \subseteq \mathbb{R}^n$ and let $\vec{f} : \Omega \to \mathbb{R}^m$ be a vector-valued function. The scalar-valued functions $f_1, \ldots, f_m : \Omega \to \mathbb{R}$ that satisfy $\vec{f}(\vec{x}) = (f_1(\vec{x}), \ldots, f_m(\vec{x}))$ are called the **component functions** of $\vec{f}$., and are related to $\vec{f}$ by the formula $f_j(\vec{x}) = \vec{e}_j \cdot \vec{f}(\vec{x})$ for each $1 \leq j \leq m$ and $\vec{x} \in \Omega$.

Visualizing Scalar-Valued Functions

For $\Omega \subseteq \mathbb{R}^n$, it is difficult to visualize a function $\vec{f} : \Omega \to \mathbb{R}^m$ when both $n > 1$ and $m > 1$. In this course we will only discuss how to visualize scalar-valued functions (i.e. the case where $m = 1$) or vector-valued functions of a single variable (i.e. the case where $n = 1$ and $m > 1$). For scalar-valued functions, the typical way that we visualize functions is through their graphs.

Definition 30. Let $\Omega \subseteq \mathbb{R}^n$. The **graph** of $f : \Omega \to \mathbb{R}$ is the collection of points

$$\{(x_1, \ldots, x_n, x_{n+1}) \ : \ x_{n+1} = f(x_1, \ldots, x_n)\} \subset \mathbb{R}^{n+1}.$$

When $n = 1$ the graph of f is a subset of $\mathbb{R}^2$, and you have lots of experience dealing with such graphs from single-variable calculus. When $n \geq 2$ we can analyze the graph of f by investigating the *level sets* of f.

Definition 31. Let $\Omega \subseteq \mathbb{R}^n$, and $f : \Omega \to \mathbb{R}$. The **level set** of f at height $k \in \mathbb{R}$ is the set

$$\{\vec{x} \in \Omega \ : \ f(\vec{x}) = k\}.$$

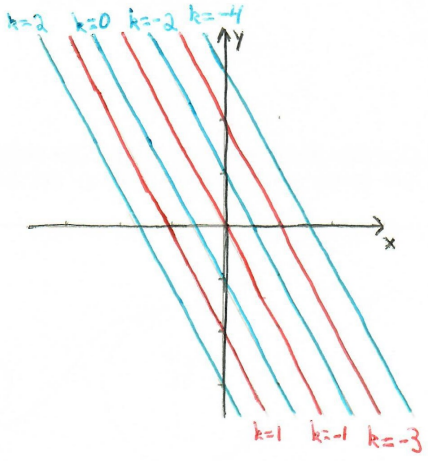
[7]Here, the notation $A \backslash B$ denotes the sets of objects in A that are *not* in B. For example, $\mathbb{R}^n \backslash \{\vec{0}\}$ denotes the collection of nonzero vectors in $\mathbb{R}^n$.

Remark 55. Note that the level of of f at height k is exactly the x_{n+1}-section of the graph of f, which is the intersection of the graph of f with the hyperplane $x_{n+1} = k$.

Remark 56. Note that the level set (at level k) of a function $f : \Omega \to \mathbb{R}$ is a subset of $\mathbb{R}^n$ instead of $\mathbb{R}^{n+1}$, and is therefore (strictly speaking) easier to visualize than the graph of f. When $n = 2$ we sometimes use the term **level curves** (in cases where $k = f(x, y)$ describes a curve), while for $n = 3$ we sometimes use the term **level surfaces** (in cases where $k = f(x, y, z)$ describes a surface).

Level curves are familiar to us from topographic maps, which superimpose elevation curves over a traditional 'flat' map to give a sense of altitude. More precisely, if $f(x, y)$ denotes the altitude of the location with longitude x and latitude y, then the level curves of $f(x, y)$ are exactly the elevation curves that appear on the topographical map. Here is an example of such a map of Lake Michigan[8]:

Example 42. The level curves of $f(x, y) = -1 - 2x - y$ all have the form $k = -1 - 2x - y$, or rather $y = -2x - 1 - k$. These are all lines in the xy-plane with slope -2 and y-intercept $-1 - k$. For each of these, we interpret k as the 'height' of the points on the graph $z = f(x, y)$ above the corresponding level curve in the xy-plane (where $k < 0$ means that the points are actually *below* the xy-plane):



[8]This image was taken from `http://www.carvedlakeart.com/`.

The graph of the function $f(x, y) = -1 - 2x - y$ (which is the graph of $z = -1 - 2x - y$, or rather $2x + y + z = -1$) is a plane. Here the level curves are the (projections onto the xy-plane of the) lines of intersection of the planes $z = k$ with the graph of $z = f(x, y)$ at various values of k.

In general, the graph of a function of the form $f(x, y) = a + bx + cy$ (for real numbers a, b, c) is a plane. Functions of this form are called **affine**, but we will see a more general definition of this term later.

Lecture 17: Topology of $\mathbb{R}^n$

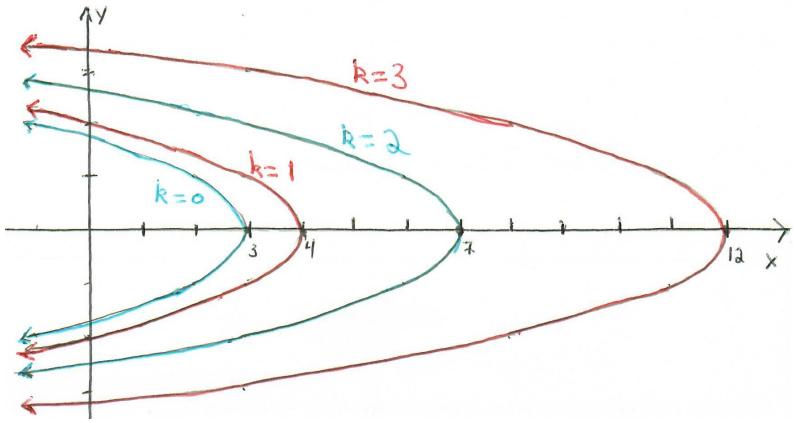
Learning Objectives:

- Work with the basic topological notions of open, closed, boundary, bounded, and compact.

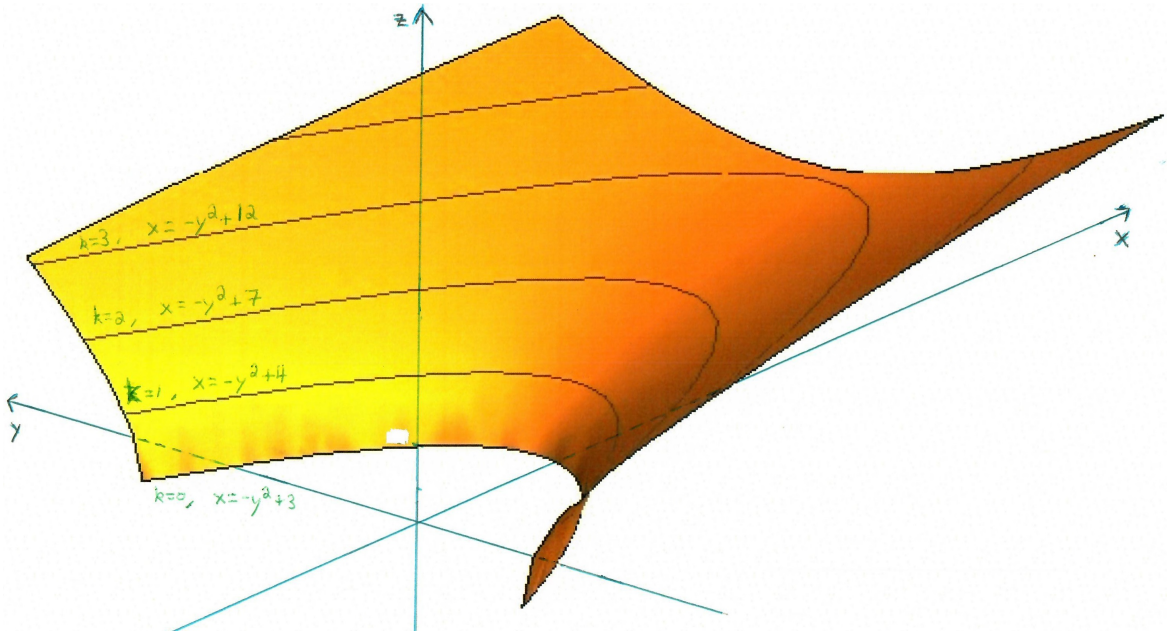
Example 43. Use level curves to sketch the graph of $f(x,y) = \sqrt{x+y^2-3}$.

The level curves of the graph of $f(x,y) = \sqrt{x+y^2-3}$ all have the form $k = \sqrt{x+y^2-3}$ for real numbers k.

When $k < 0$ there are no points (x,y) in the plane that satisfy the equation $k = \sqrt{x+y^2-3}$ (since the output of the square root function is always nonnegative). When $k \geq 0$ the level curve at $z = k$ satisfies $k^2 = x + y^2 - 3$, or rather $x = -y^2 + (k^2+3)$. This is a parabola in the xy-plane that is symmetric about the x-axis, opens in the negative x-direction, and has vertex $(k^2+3, 0)$. We sketch several of these below:



We can use these sketches to visualize the graph of f:



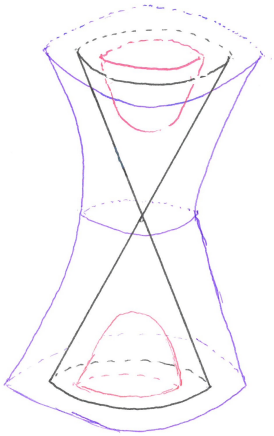
Example 44. The level surfaces of the function $f(x, y, z) = x^2 + y^2 - z^2$ are familiar to us, and actually give us a very nice understanding of some quadric surfaces.

The level surface with $k = 0$ ($x^2+y^2-z^2 = 0$) has the equation $z^2 = x^2+y^2$, or rather $z = \pm\sqrt{x^2 + y^2}$. This is the (double) cone centered around the z-axis.

The level surface with $k > 0$ ($x^2 + y^2 - z^2 = k$) can be written as $z^2 + k = x^2 + y^2$ (for $k > 0$), which describes a one-sheeted hyperboloid (which is symmetric about the z-axis). Larger values of k cause this hyperboloid to expand (or perhaps 'inflate' would be better).

The level surface with $k = -c < 0$ (with $c > 0$) we be written as $z^2 = x^2 + y^2 + c$, which is the equation of a two-sheeted hyperboloid (symmetric about the z-axis). The two sheets of this hyperboloid get farther apart as k becomes more negative.

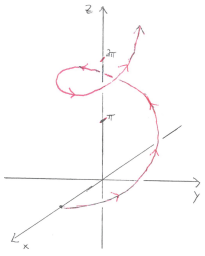
In other words, the two-sheeted hyperboloid, (double) cone, and one-sheeted hyperboloid are all level surfaces of the same function! We sketch these below:



Visualizing Vector-Valued Functions of a Single Variable

If $\Omega \subseteq \mathbb{R}$, for vector valued function $\vec{f} : \Omega \to \mathbb{R}^m$ of a single variable we can instead visualize $\vec{f}$ by interpreting $\vec{f}(t)$ as the location of a particle in $\mathbb{R}^m$ at time t. In this case we think that $\vec{f}$ traces out a path in $\mathbb{R}^m$, which we can visualize by sketching (with perhaps arrows to indicate the direction in which the particle travels).

Example 45. Consider[9] $\vec{f} : [0, +\infty) \to \mathbb{R}^3$, $\vec{f}(t) = (\cos(t), \sin(t), t)$. The projection of this curve onto the xy-plane is the circle $x^2+y^2 = 1$, and so the particle is rotates around the z-axis in a counterclockwise direction (when viewed from above). The fact that the z-coordinate is t tells us that the height of the particle above the xy-plane at time t is t, so the particle is steadily 'gaining altitude' as time passes. This path traces out a **helix**.



[9]This function is an example of a **path** in $\mathbb{R}^n$, which we will define more rigorously once we have access to the term **continuity**, which is defined in terms of limits.

Example 46. The function $\vec{g} : [0, +\infty) \to \mathbb{R}^3$, $\vec{g}(t) = (\cos(t^3), \sin(t^3), t^3)$ traces out the same curve as in the previous example, but is a different function than $\vec{f}$. Therefore you should be cautious when making inferences based on pictures.

Topology of $\mathbb{R}^n$

Throughout the first half of MATH 291 we have been concerned with the *algebraic* structure of $\mathbb{R}^n$ as a vector space. In the calculus portion of the course, though, we will be much more interested in the *topological* structure of $\mathbb{R}^n$. Topology is the mathematical study of qualitative structure, and we will frame most topological notions in terms of special subsets of $\mathbb{R}^n$ called *open* sets and *closed* sets. Our discussion today only introduces these ideas at a convenient level for our purposes, but you will see a more detailed treatment in MATH 321 or 344.

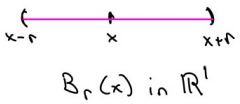
Open Sets

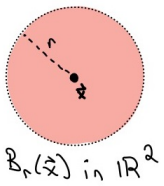
Intuitively, an open set $U \subseteq \mathbb{R}^n$ is a set that insulates each point in the set from $U^c = \{\vec{x} \ : \ \vec{x} \notin U\}$. The definition of open set is framed in terms of open balls.

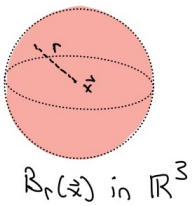
Definition 32. Let $\vec{x} \in \mathbb{R}^n$ and $r > 0$. We define the **open ball** of radius r centered at $\vec{x}$ to be

$$B_r(\vec{x}) \stackrel{def}{=} \{\vec{y} \in \mathbb{R}^n \ : \ \|\vec{y} - \vec{x}\| < r\}.$$

In other words, $B_r(\vec{x})$ consists of exactly the points in $\mathbb{R}^n$ whose distance to $\vec{x}$ is less than r. In $\mathbb{R}^1$ this would be the interval $(x - r, x + r)$. In $\mathbb{R}^2$ this is the disk of radius r centered at $\vec{x}$ (but excluding the boundary circle). In $\mathbb{R}^3$ this is the region enclosed by the sphere of radius r centered at $\vec{x}$ (but excluding the sphere itself).







Open balls form a basis[10] for a more general definition of open sets.

Definition 33. Let $U \subseteq \mathbb{R}^n$. We say that U is **open** if for every $\vec{x} \in U$ there exists $r > 0$ such that $B_r(\vec{x}) \subseteq U$.

That is, U is open if every point in U is the center of a ball that is completely contained in U. This is the sense in which open sets "insulate" their points from the complement of U.

[10]There is a pun here, but to truly appreciate it you should take MATH 344: Introduction to Topology!

Lecture 18: More Topology of $\mathbb{R}^n$

Learning Objectives:

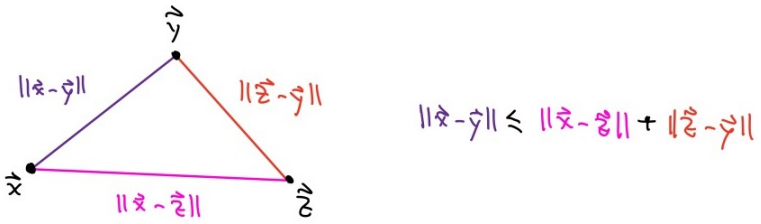
- Work with the basic topological notions of open, closed, boundary, bounded, and compact.

Example 47. For $\vec{p} \in \mathbb{R}^n$ and $r > 0$, the open ball $B_r(\vec{p})$ is open in the more general sense.

To prove this, we will need the **triangle inequality**: for every $\vec{x}, \vec{y}, \vec{z} \in \mathbb{R}^n$,

$$\|\vec{x} - \vec{y}\| \leq \|\vec{x} - \vec{z}\| + \|\vec{z} - \vec{y}\|.$$

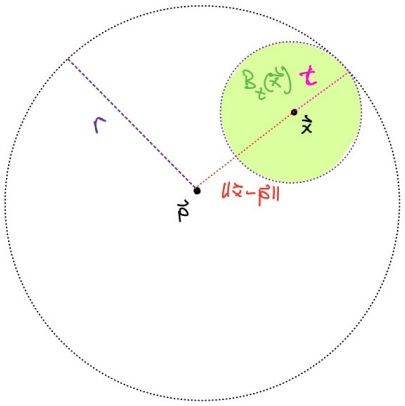
You will prove this on your homework, but the intuition is that if $\vec{x}, \vec{y}, \vec{z}$ form the vertices of a triangle, then the length of the side determined by $\vec{x}$ and $\vec{z}$ is no more than the sum of the lengths of the remaining sides.



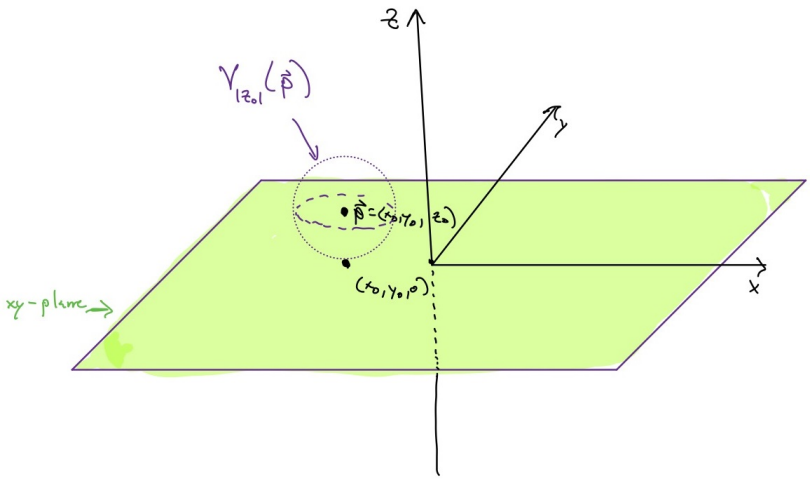
Let's proceed with the example. Let $\vec{x} \in B_r(\vec{p})$. Because $\|\vec{p} - \vec{x}\| < r$ we have $t \stackrel{def}{=} r - \|\vec{p} - \vec{x}\| > 0$. We claim that $B_t(\vec{x}) \subseteq B_r(\vec{p})$. To show this, let $\vec{q} \in B_t(\vec{x})$. Then

$$\|\vec{p} - \vec{q}\| \leq \|\vec{p} - \vec{x}\| + \|\vec{x} - \vec{q}\| < \|\vec{p} - \vec{x}\| + t = r,$$

so that $\vec{q} \in B_r(\vec{p})$. This shows that $B_t(\vec{x}) \subseteq B_r(\vec{p})$, and therefore $B_r(\vec{p})$ is open.



Example 48. Let $U = \{(x, y, z) \ : \ z \neq 0\}$ be the $\mathbb{R}^3$ with the xy-plane removed open. Then U is open.



To see why, take $\vec{p} = (x_0, y_0, z_0) \in U$. Then $z_0 \neq 0$, and we expect that $B_{|z_0|}(\vec{p}) \subseteq U$. To show this, let $\vec{x} = (x, y, z) \in B_{|z_0|}(\vec{p})$. Then we have

$$|z - z_0| = \sqrt{(z - z_0)^2} \leq \sqrt{(x - x_0)^2 + (y - y_0)^2 + (z - z_0)^2} = \|\vec{x} - \vec{p}\| < |z_0|,$$

so that since $|z_0| \leq |z_0 - z| + |z|$ we have $|z| \geq |z_0| - |z - z_0| > 0$, and therefore $z \neq 0$ so that $\vec{x} \in U$. This shows that $B_{|z_0|}(\vec{p}) \subseteq U$, so that U is open.

Boundary

To distinguish open from non-open sets, we need a way to describe sets where not every point in the set is "insulated". To facilitate this, we define the boundary of a set (which should describe points on the "edge" of a set).

Definition 34. Let $S \subseteq \mathbb{R}^n$. We say that $\vec{p} \in \mathbb{R}^n$ is a **boundary point** of S if for every $r > 0$,

$$B_r(\vec{p}) \cap S \neq \emptyset \quad \text{and} \quad B_r(\vec{p}) \cap S^c \neq \emptyset.$$

In other words $\vec{p}$ is a boundary point of S is every ball centered at $\vec{p}$ contains at least one point in S and at least one point not in S.
The **boundary** of S, ∂S, is the collection of all boundary points of S. That is,

$$\partial S = \{\vec{p} \in \mathbb{R}^n \ : \ \vec{p} \text{ is a boundary point of } S\}.$$

Example 49. The boundary of $U = \{(x, y, z) \ : \ z \neq 0\}$ is exactly the xy-plane. Note that for each point $\vec{p} = (x, y, 0)$ and each $r > 0$ the ball $B_r(\vec{p})$ contains at least one point in U (i.e. $(x, y, \frac{r}{2})$) and at least one point not in U (i.e. $\vec{p} = (x, y, 0)$ itself!). On the other hand, if $\vec{p} = (x, y, z)$ with $z \neq 0$, then we showed that $B_{|z|}(\vec{p})$ contains no points on the xy-plane.

Example 50. If $K = \{(x, y, z) \ : \ z = 0\}$ is the xy-plane, then $\partial K = K$.

Closed Sets

Definition 35. Let $S \subseteq \mathbb{R}^n$. We say that S is **closed** if $\partial S \subseteq S$. That is, if S contains its boundary.

Example 51. Let $K = \{(x, y, z) \ : \ z = 0\}$ be the xy-plane. Because $\partial K = K \subseteq K$, K is closed.

Example 52. Consider the interval $A = [-1, 1) \subset \mathbb{R}$. Then A is not open, because there is no open ball centered at -1 that is contained in A. On the other hand, $\partial A = \{-1, 1\}$, and since $1 \notin A$ we see that A does not contain all of its boundary, and is therefore not closed.

On the other hand, the set $U = \mathbb{R}^n$ is both open and (because it does not have any boundary points, and therefore $\partial U = \emptyset \subset U$) closed.

These examples illustrate an important point: *Sets are not doors. A door must be either open or closed, but a set could be neither (or both)!*

Bounded Sets

A set is bounded if it does not contain points arbitrarily far away from the origin.

Definition 36. A set $S \subseteq \mathbb{R}^n$ is **bounded** if there is $M > 0$ such that $\|\vec{x}\| \leq M$ for every $\vec{x} \in S$.

In other words, S is bounded if $S \subseteq B_M(\mathbf{0})$ for some $M > 0$. So, bounded sets are exactly the sets that lie inside of some large ball centered at the origin.

Example 53. The solid ellipsoid E defined by

$$E = \{(x, y, z) \in \mathbb{R}^3 \ : \ 3x^2 + 2y^2 + \frac{1}{2}z^2 \leq 1\}$$

is bounded. To see this, note that for each $\vec{x} = (x, y, z) \in E$ we have

$$\|\vec{x}\|^2 = x^2 + y^2 + z^2 = 2\Big(\frac{1}{2}x^2 + \frac{1}{2}y^2 + \frac{1}{2}z^2\Big) \leq 2\Big(3x^2 + 2y^2 + \frac{1}{2}z^2\Big) = 2,$$

so that $E \subseteq B_{\sqrt{2}}(\vec{0})$.

Example 54. The xy-plane $K = \{(x, y, z) \ : \ z = 0\}$ is not bounded, since for every $M > 0$ the point $(M + 1, 0, 0) \in K$ and $\|(M + 1, 0, 0)\| = M + 1 > M$.

Compactness

There are some notions of "smallness" for subsets of $\mathbb{R}^n$, but the most important topological notion of "smallness" is compactness.

Definition 37. $K \subseteq \mathbb{R}^n$ is called **compact** if it is closed and bounded.

You will see more general definitions of compactness in later courses, but this definition is useful for our purposes. Compact sets will be very important for us when studying properties of functions. To hint vaguely at what we will see: unbounded sets and sets that are missing a boundary point allow functions on those sets to be "poorly behaved" at inputs that become unbounded or approach a boundary point. Therefore compact sets allow neither of these opportunities for mischief.

Example 55. $B_r(\vec{x})$ is not compact for any $\vec{x} \in \mathbb{R}^n$ or $r > 0$. Although this set is bounded, it is not closed.

Example 56. The solid ellipsoid E defined by $3x^2 + 2y^2 + \frac{1}{2}z^2 \leq 1$ is compact. We showed above that E is bounded.

To show that E is closed, we can show that the boundary of E is exactly the ellipsoid S given by $3x^2 + 2y^2 + \frac{1}{2}z^2 = 1$. This will show that $\partial E = S \subseteq E$, and E is closed. Because E is closed and bounded, E is compact. Here is an informal argument: if $\vec{x} \in S$, then each ball $B_r(\vec{x})$ contains a point in E (i.e. $\vec{x}$) and a point not in E (i.e. $\vec{y}$ for $\vec{y}$ very close to $\vec{x}$ but outside of the solid ellipsoid E). Therefore $S \subseteq \partial E$. On the other hand, if $\vec{x} \in \partial E$ then $\vec{x}$ is not in the open region enclosed by S (since then a small ball centered at $\vec{x}$ would be contained in E, and therefore would have empty intersection with E^c), and $\vec{x}$ could not be in the open region exterior to S (since then a small ball centered at $\vec{x}$ would be contained in E^c, and therefore would have empty intersection with E). Therefore $\partial E \subseteq S$, so that $\partial E = S \subseteq E$ and E is closed.

Here is a more formal argument that $\partial E = S$. Let $(x, y, z) \in S$ be a point on the ellipsoid, so that $3x^2 + 2y^2 + \frac{1}{2}z^2 = 1$. Let $r > 0$. Then $(x, y, z) \in B_r(x, y, z) \cap E$. so $B_r(x, y, z) \cap E \neq \emptyset$. On the other hand, let $t \in (1, 1 + \frac{r}{2})$. Then

$$\|(tx, ty, tz) - (x, y, z)\| = |t-1|\|(x, y, z)\|0 = (t-1)(x^2+y^2+z^2) \leq 2(t-1)(3x^2+2y^2+\frac{1}{2}z^2) = 2(t-1) < r,$$

so $(tx, ty, tz) \in B_r(x, y, z)$. On the other hand,

$$3(tx)^2 + 2(ty)^2 + \frac{1}{2}(tz)^2 = t^2(3x^2 + 2y^2 + \frac{1}{2}z^2) > 1^2(3x^2 + 2y^2 + \frac{1}{2}z^2) = 1,$$

so that $(tx, ty, tz) \in E^c$. Therefore $B_r(x, y, z) \cap E^c \neq \emptyset$. This shows that $S \subseteq \partial E$.

To show that $\partial E \subseteq S$ (i.e. "$\vec{x} \in \partial E$ implies that $\vec{x} \in S$") it suffices to show the contrapositive: $S^c \subseteq (\partial E)^c$ (i.e. "$\vec{x} \notin S$ implies that $\vec{x} \notin \partial E$"). Suppose that $(x, y, z) \notin S$, so $3x^2 + y^2 + \frac{1}{2}z^2 \neq 1$. Suppose $t \stackrel{def}{=} 3x^2 + 2y^2 + \frac{1}{2}z^2 < 1$ (the proof in the case > 1 will be similar). For $r \stackrel{def}{=} \frac{1-t}{18(\|(x,y,z)\|+1)} > 0$, let $(x', y', z') \in B_r(x, y, z)$. Note that $r = \frac{1-t}{18(\|(x,y,z)\|+1)} \leq \frac{1-t}{18} \leq \frac{1}{18} < 1$. We have

$$\begin{aligned}
3(x')^2 + 2(y')^2 + \frac{1}{2}(z')^2 &= 3(x' - x + x)^2 + 2(y' - y + y)^2 + \frac{1}{2}(z' - z + z)^2 \\
&= 3(x' - x)(x' + x) + 2(y' - y)(y' - y) + \frac{1}{2}(z' - z)(z' + z) + 3x^2 + 2y^2 + \frac{1}{2}z^2 \\
&= 3(x' - x)(x' + x) + 2(y' - y)(y' - y) + \frac{1}{2}(z' - z)(z' + z) + t.
\end{aligned}$$

But note that

$$|x' - x| \leq \sqrt{(x'-x)^2 + (y'-y)^2 + (z'-z)^2} < r$$

and

$$|x' + x| = |x' - x + 2x| \leq |x' - x| + 2|x| < 2|x| + r \leq 2\|(x, y, z)\| + r,$$

and similarly $|y' - y| < r$, $|z' - z| < r$ and $|y' + y| < 2\|(x, y, z)\| + r$ and $|z' + z| < 2\|(x, y, z)\| + r$. Therefore we have (using the fact that $r < 1$)

$$\begin{aligned}
3(x')^2 + 2(y')^2 + \frac{1}{2}(z')^2 &= 3(x' - x)(x' + x) + 2(y' - y)(y' - y) + \frac{1}{2}(z' - z)(z' + z) + t \\
&\leq 3|x' - x||x' + x| + 2|y' - y||y' - y| + \frac{1}{2}|z' - z||z' + z| + t \\
&\leq 3r(2\|(x, y, z)\| + r) + 2r(2\|(x, y, z)\| + r) + \frac{1}{2}r(2\|(x, y, z)\| + r) + t \\
&< 3r(2\|(x, y, z)\| + r) + 3r(2\|(x, y, z)\| + r) + 3r(2\|(x, y, z)\| + r) + t \\
&< 3r(2\|(x, y, z)\| + 2) + 3r(2\|(x, y, z)\| + 2) + 3r(2\|(x, y, z)\| + 2) + t \\
&= 18r(\|(x, y, z)\| + 1) + t \\
&< 18\frac{1 - t}{18(\|(x, y, z)\| + 1)}(\|(x, y, z)\| + 1) + t \\
&= 1,
\end{aligned}$$

so that $(x', y', z') \in E$, and therefore $B_r(x, y, z) \cap E^c = \emptyset$, so that $(x, y, z) \notin \partial E$.

Lecture 19: Limits

Learning Objectives:

- Show the existence of the limit of a function of several variables.

The motivation for multivariable limits is the same for single variable limits. For a function $\vec{f} : \Omega \to \mathbb{R}^m$ (where $\Omega \subseteq \mathbb{R}^n$) we want the expression "$\lim_{\vec{x}\to\vec{a}} \vec{f}(\vec{a}) = \vec{L}$" to mean that as $\vec{x}$ "approaches" $\vec{a}$, the corresponding outputs $\vec{f}(\vec{x})$ "approach" $\vec{L}$. Here we want $\vec{a}$ to be a *limit point* of Ω, which means that every ball $B_r(\vec{a})$ centered at $\vec{a}$ contains at least one point in Ω other than $\vec{a}$. More precisely,

Definition 38. $\vec{a} \in \mathbb{R}^n$ is a **limit point** of $\Omega \subseteq \mathbb{R}^n$ if for every $r > 0$, $(B_r(\vec{a})\backslash\{a\}) \cap \Omega \neq \emptyset$.

The notion of limit is intended to capture the behavior of the values of $\vec{f}(\vec{x})$ when $\vec{x}$ is close to (but not necessarily equal to) $\vec{a}$. Our formal definition should capture the notion that, to ensure that $\vec{f}(\vec{x})$ is close to $\vec{L}$, it suffices to take $\vec{x}$ "close enough" to $\vec{a}$. Hence we frame our definition as a game: we should be able to ensure that $\|\vec{f}(\vec{x}) - \vec{L}\|$ is smaller than any given positive number (i.e. "$\vec{f}(\vec{x})$ is as close as we'd like to $\vec{L}$"), provided only that $\|\vec{x} - \vec{a}\|$ is sufficiently small (and $\vec{x} \neq \vec{a}$). This leads us to the following definition.

Definition 39. Let $\Omega \subseteq \mathbb{R}^n$ and $\vec{f} : \Omega \to \mathbb{R}^m$, and suppose that $\vec{a}$ is a limit point of Ω. We say that $\lim_{\vec{x}\to\vec{a}} \vec{f}(\vec{x}) = \vec{L}$ if for every $\epsilon > 0$ there exists $\delta > 0$ such that for every $\vec{x} \in \Omega$ with $0 < \|\vec{x} - \vec{a}\| < \delta$, we have $\|\vec{f}(\vec{x}) - \vec{L}\| < \epsilon$.

Remark 57. To interpret this definition in terms of the original discussion, note that condition that $\|\vec{f}(\vec{x}) - \vec{L}\| < \epsilon$ is exactly that $\vec{f}(\vec{x})$ is within ϵ of $\vec{L}$. When ϵ is small, we therefore consider $\vec{f}(\vec{x})$ to be close to $\vec{L}$. The condition that there is $\delta > 0$ such that $\|\vec{f}(\vec{x}) - \vec{L}\| < \epsilon$ provided that $0 < \|\vec{x} - \vec{a}\| < \delta$ can be translated as "to ensure that $\vec{f}(\vec{x})$ is within ϵ of $\vec{L}$, it suffices to take $\vec{x} \neq \vec{a}$ in a ball centered at $\vec{a}$ with (probably small) radius δ." Because for each ϵ we must be able to produce such a δ, this says that no matter how close (ϵ) we want to ensure that $\vec{f}(\vec{x})$ is to $\vec{L}$, all we need to do is ensure that $\vec{x}$ is sufficiently close (within δ) to $\vec{a}$.

Remark 58. We can also state this definition more concisely using some of our topological notions. For one piece of notation, we will write $B_r(\vec{a})^* \stackrel{def}{=} \{\vec{x} \ : \ 0 < \|\vec{x} - \vec{a}\| < r\} = B_r(\vec{a}) - \{\vec{a}\}$ to denote the **punctured ball** of radius r centered at $\vec{a}$. With this notation, the definition of limit becomes:

Definition 40. Let $\Omega \subseteq \mathbb{R}^n$ and $\vec{f} : \Omega \to \mathbb{R}^m$, and suppose that $\vec{a}$ is a limit point of Ω. We say that $\lim_{\vec{x}\to\vec{a}} \vec{f}(\vec{x}) = \vec{L}$ if for every $\epsilon > 0$ there exists $\delta > 0$ such that for every $\vec{x} \in \Omega \cap B^*_\delta(\vec{a})$, $\vec{f}(\vec{x}) \in B_\epsilon(\vec{L})$.

Remark 59. The definition of limit involves nested quantifiers, and can be expressed in the form

$$(\forall \epsilon > 0)(\exists \delta > 0)(\forall \vec{x} \in \Omega)\Big(0 < \|\vec{x} - \vec{a}\| < \delta \;\Rightarrow\; \|\vec{f}(\vec{x}) - \vec{L}\| < \epsilon\Big). \tag{8}$$

Therefore a proof that $\lim_{\vec{x}\to\vec{a}} \vec{f}(\vec{x}) = \vec{L}$ should have the following structure:

Let $\epsilon > 0$. Take $\delta = ...$ Suppose $\vec{x} \in \Omega$ satisfies $0 < \|\vec{x} - \vec{a}\| < \delta$. Then (give an argument showing that) $\|\vec{f}(\vec{x}) - \vec{L}\| < \epsilon$.

It is entirely appropriate to do some scratchwork to determine what δ should be in terms of ϵ and $\vec{a}$. When you go to write up your proof, though, it should be structured like the one above. The argument might look a lot like you scratchwork, and will often be very short and clean.

Example 57. Show that $\lim_{(x,y)\to(1,-2)} (2x + 3y - 1) = -5$.

Scratchwork. Given $\epsilon > 0$, we want to chose $\delta > 0$ such that if

$$\|(x, y) - (1, -2)\| = \sqrt{(x-1)^2 + (y+2)^2} < \delta,$$

then $|2x + 3y - 1 - (-5)| < \epsilon$. But

$$|2x + 3y - 1 - (-5)| = |2(x-1) + 3(y+2)| \le 2|x-1| + 3|y+2|,$$

and since $|x-1| = \sqrt{(x-1)^2} \le \sqrt{(x-1)^2 + (y+2)^2} < \delta$ (and, similarly $|y+2| < \delta$) we have

$$|2x + 3y - 1 - (-5)| \le 2|x-1| + 3|y+2| < 5\delta.$$

Therefore, if we take δ so that $5\delta \le \epsilon$, then the proof should go through.

Proof. Let $\epsilon > 0$. Define $\delta \stackrel{def}{=} \frac{\epsilon}{5}$. Let $(x, y) \in \mathbb{R}^2$ with $0 < \|(x, y) - (1, -2)\| < \delta$. Then

$$|2x + 3y - 1 - (-5)| = |2(x-1) + 3(y+2)| \le 2|x-1| + 3|y+2| \le 5\sqrt{(x-1)^2 + (y+2)^2} < 5\frac{\epsilon}{5} = \epsilon.$$

There are many facts about limits that we will take for granted in this course. All of these can be proved rigorously proved using the definition of limit, and are certainly within your power, but the complete picture takes a lot of time to develop (and you will do so when you take real analysis).

Theorem 25 (Properties of Limits)**.** Let $\Omega \subseteq \mathbb{R}^n$, let $\vec{f}, \vec{g} : \Omega \to \mathbb{R}^m$, and assume $\vec{a}$ is a limit point of Ω. If $\lim_{\vec{x}\to\vec{a}} \vec{f}(\vec{x}) = \vec{L}$ and $\lim_{\vec{x}\to\vec{a}} \vec{g}(\vec{x}) = \vec{M}$, then the following hold.

(a) (Uniqueness) If $\lim_{\vec{x}\to\vec{a}} \vec{f}(\vec{x}) = \vec{K}$, then $\vec{K} = \vec{L}$.

(b) (Linearity) For every $\lambda \in \mathbb{R}$, $\lim_{\vec{x}\to\vec{a}} (\vec{f}(\vec{x}) + \lambda\vec{g}(\vec{x})) = \vec{L} + \lambda\vec{M}$.

(c) (Products) If $h : \Omega \to \mathbb{R}$ and $C \in \mathbb{R}$ and $\lim_{\vec{x}\to\vec{a}} h(\vec{x}) = C$, then $\lim_{\vec{x}\to\vec{a}} h(\vec{x})\vec{f}(\vec{x}) = C\vec{L}$.

(d) (Scalar-Valued Quotients) If $m = 1$ and $M \neq 0$, then $\lim_{\vec{x}\to\vec{a}} \dfrac{f(\vec{x})}{g(\vec{x})} = \dfrac{L}{M}$.

One important (and perhaps expected) property of limits is that the limit of a vector-valued function is intimately connected to the limits of its component functions, in the following sense. We'll include the proof here so that you get a sense of the flavor of all of these sorts of proofs.

Theorem 26 (Limits and Component Functions)**.** Let $\Omega \subseteq \mathbb{R}^n$, let $\vec{f} : \Omega \to \mathbb{R}^m$, $\vec{f}(\vec{x}) = (f_1(\vec{x}), \ldots, f_m(\vec{x}))$, let $\vec{a}$ be a limit point of Ω, and let $\vec{L} = (L_1, \ldots, L_m) \in \mathbb{R}^m$. Then $\lim_{\vec{x}\to\vec{a}} \vec{f}(\vec{x}) = \vec{L}$ if, and only if, $\lim_{\vec{x}\to\vec{a}} f_j(\vec{x}) = L_j$ for each $1 \leq j \leq m$.

Proof. Suppose that $\lim_{\vec{x}\to\vec{a}} \vec{f}(\vec{x}) = \vec{L}$. Let $1 \leq j \leq m$. Let $\epsilon > 0$. Choose $\delta > 0$ such that $\|\vec{f}(\vec{x}) - \vec{L}\| < \epsilon$ whenever $\vec{x} \in \Omega$ and $0 < \|\vec{x} - \vec{a}\| < \delta$. Let $\vec{x} \in \Omega$ with $0 < \|\vec{x} - \vec{a}\| < \delta$. Then

$$|f_j(\vec{x}) - L_j| = \sqrt{(f_j(\vec{x}) - L_j)^2} \leq \sqrt{(f_1(\vec{x}) - L_1)^2 + \cdots + (f_m(\vec{x}) - L_m)^2} = \|\vec{f}(\vec{x}) - \vec{L}\| < \epsilon.$$

Now suppose that $\lim_{\vec{x}\to\vec{a}} f_j(\vec{x}) = L_j$ for each $1 \leq j \leq m$. Let $\epsilon > 0$. For each $1 \leq j \leq m$, choose $\delta_j > 0$ such that $|f_j(\vec{x}) - L_j| < \frac{\epsilon}{m}$ whenever $\vec{x} \in \Omega$ and $0 < \|\vec{x} - \vec{a}\| < \delta_j$. Let $\delta = \min(\delta_1, \ldots, \delta_m)$. Let $\vec{x} \in \Omega$ with $0 < \|\vec{x} - \vec{a}\| < \delta$. Then the triangle inequality implies that

$$\begin{aligned} \|\vec{f}(\vec{x}) - \vec{L}\| &= \|(f_1(\vec{x}) - L_1)\vec{e}_1 + \cdots + (f_m(\vec{x}) - L_m)\vec{e}_m\| \\ &\leq |f_1(\vec{x}) - L_1|\|\vec{e}_1\| + \cdots + |f_m(\vec{x}) - L_m|\|\vec{e}_m\| \\ &< \frac{\epsilon}{m} + \cdots + \frac{\epsilon}{m} = \epsilon. \end{aligned}$$

□

Lecture 20: More Limits

Learning Objectives:

- Show the nonexistence of the limit of a function of several variables.

Nonexistence of Limits

Limits do not always exist. In single-variable calculus, one can sometimes argue that $\lim_{x\to a} f(x)$ fails to exist by inspecting the one-sided limits. That is, if $\lim_{x\to a} f(x)$ exists then both $\lim_{x\to a^-} f(x)$ and $\lim_{x\to a^+} f(x)$ exist and are equal to $\lim_{x\to a} f(x)$. Therefore if $\lim_{x\to a^-} f(x)$ and $\lim_{x\to a^+} f(x)$ both exist but are not equal to one another (or if one of these one-sided limits fails to exist altogether), then $\lim_{x\to a} f(x)$ does not exist.

One way of understanding this result is that the limit of f (say L) exists as $x \to a$ only if $f(x) \to L$ as x approaches a, *no matter how x does so.* Because there are only two "approach paths" (i.e. "from the left" and "from the right"), one can check the existence of $\lim_{x\to a} f(x)$ by inspecting the (two) one-sided limits.

In the multivariable case the situation is much more complicated. In particular, if $\vec{x}, \vec{a} \in \mathbb{R}^n$ then there are infinitely many paths along which $\vec{x}$ could approach $\vec{a}$. Nevertheless, we can generalize the one-variable idea with the following result, which essentially computes the limit of a function along a path.

Theorem 27 (Testing Limits)**.** Let $\Omega \subseteq \mathbb{R}^n$, $\vec{a}$ a limit point of Ω, let $\vec{f} : \Omega \to \mathbb{R}^m$, and suppose that $I \subseteq \mathbb{R}$ is an interval containing a number a. Assume that $\vec{x} : I \to \mathbb{R}^n$ is a function with $\vec{x}(t) \in \Omega$ and $\vec{x}(t) \neq \vec{a}$ for all $t \neq a$, such that $\lim_{t\to a} \vec{x}(t) = \vec{a}$. If $\lim_{\vec{x}\to\vec{a}} \vec{f}(\vec{x}) = \vec{L}$, then $\lim_{t\to a} \vec{f}(\vec{x}(t)) = \vec{L}$.

Proof. Suppose that $\lim_{\vec{x}\to\vec{a}} \vec{f}(\vec{x}) = \vec{L}$. Let $\epsilon > 0$. Choose $\delta' > 0$ such that if $0 < \|\vec{x} - \vec{a}\| < \delta'$ then $\|\vec{f}(\vec{x}) - \vec{L}\| < \epsilon$. Choose $\delta > 0$ such that if $t \in I$ and $0 < |t - a| < \delta$ then $\|\vec{x}(t) - \vec{a}\| < \delta'$. Because $\vec{x}(t) \neq \vec{a}$ for all $t \neq a$, $0 < \|\vec{x}(t) - \vec{a}\|$ for all $t \in I$. Let $t \in I$ with $0 < |t - a| < \delta$. Then $0 < \|\vec{x}(t) - \vec{a}\| < \delta'$, so that $\|\vec{f}(\vec{x}(t)) - \vec{L}\| < \epsilon$. □

This gives the following result.

Corollary 9 (Nonexistence Theorem)**.** Let $\Omega \subseteq \mathbb{R}^n$, $\vec{a}$ a limit point of Ω, let $\vec{f} : \Omega \to \mathbb{R}^m$. Assume $I, J \subseteq \mathbb{R}$ are intervals, $a \in I$ and $b \in J$, and if there exist two functions $\vec{x} : I \to \mathbb{R}^n$ and $\vec{y} : J \to \mathbb{R}^n$ with $\vec{x}(t) \in \Omega$ and $\vec{x}(t) \neq \vec{a}$ for each $t \neq a$, $\vec{y}(s) \in \Omega$ and $\vec{y}(s) \neq \vec{a}$ for each $s \neq a$, with $\lim_{t\to a} \vec{x}(t) = \vec{a} = \lim_{s\to b} \vec{y}(s)$. Then $\lim_{\vec{x}\to\vec{a}} \vec{f}(\vec{x})$ fails to exist if either

(a) $\lim_{t\to a} \vec{f}(\vec{x}(t))$ does not exist (or is infinite), or

(b) $\lim_{t\to a} \vec{f}(\vec{x}(t)) = \vec{B}$ and $\lim_{s\to b} \vec{f}(\vec{y}(s)) = \vec{C}$ with $\vec{B} \neq \vec{C}$.

Proof. By the Testing Limits Theorem, if $\lim_{\vec{x}\to\vec{a}} \vec{f}(\vec{x})$ existed, then it must equal $\vec{B}$ and it must equal $\vec{C}$. But $\vec{B} \neq \vec{C}$, and therefore $\lim_{\vec{x}\to\vec{a}} \vec{f}(\vec{x})$ does not exist. □

In the Nonexistence Theorem, $\vec{x}(t)$ describes a path in $\mathbb{R}^n$ that approaches $\vec{a}$ as $t \to a$. Similarly, $\vec{y}(s)$ also describes a path in $\mathbb{R}^n$ that approaches $\vec{a}$ as $s \to b$. The statement that

$$\lim_{t\to a} \vec{f}(\vec{x}(t)) = \vec{B} \quad \text{and} \quad \lim_{s\to b} \vec{f}(\vec{y}(s)) = \vec{C} \quad \text{with} \quad \vec{B} \neq \vec{C}$$

indicates that, if we compute the limit of $\vec{f}(\vec{x})$ along the paths described by $\vec{x}(t)$ and $\vec{y}(s)$ (respectively), we get *different* limits $\vec{B}$ and $\vec{C}$. This is exactly the same result as in the single-variable case, but with the twist that in the single-variable case there are really only two different ways for x to approach a. In the multivariable setting there are infinitely many different approach paths!

Example 58. Show that $\lim_{(x,y)\to(0,0)} \frac{x^2}{\sqrt{x^4+y^2}}$ does not exist.

To do this, we produce two paths approaching $(0,0)$ along which the limit of $\frac{x^2}{\sqrt{x^4+y^2}}$ does not exist. For this example, let's try to use the coordinate axes.

$$\text{Along the } x-\text{axis}: \quad \lim_{\substack{(x,y)\to(0,0)\\(x,y)=(t,0)}} \frac{x^2}{\sqrt{x^4+y^2}} = \lim_{t\to 0} \frac{t^2}{\sqrt{t^4+0^2}} = \lim_{t\to 0} \frac{t^2}{t^2} = 1.$$

$$\text{Along the } y-\text{axis}: \quad \lim_{\substack{(x,y)\to(0,0)\\(x,y)=(0,t)}} \frac{x^2}{\sqrt{x^4+y^2}} = \lim_{t\to 0} \frac{0^2}{\sqrt{0^4+t^2}} = \lim_{t\to 0} 0 = 0.$$

Since the limits of f along these two curves do not agree, we conclude that $\lim_{(x,y)\to(0,0)} \frac{x^2}{\sqrt{x^4+y^2}}$ does not exist.

Looking along curves can sometimes be misleading if you are not careful, as the next examples illustrate.

Example 59. Determine whether or not $\lim_{(x,y)\to(0,0)} \frac{x^2y}{x^4+y^2}$ exists.

This is a particularly sticky problem. Taking the limit along the y-axis gives

$$\text{Along } x = 0: \quad \lim_{\substack{(x,y)\to(0,0)\\(x,y)=(0,t)}} \frac{x^2y}{x^4+y^2} = \lim_{t\to 0} \frac{0}{0+t^2} = 0,$$

and along the x-axis gives

$$\text{Along } y = 0: \quad \lim_{\substack{(x,y)\to(0,0)\\(x,y)=(t,0)}} \frac{x^2y}{x^4+y^2} = \lim_{t\to 0} \frac{0}{t^4+0} = 0.$$

The other lines through the origin tell a similar story. Indeed, taking the limit along any line of the form $y = mx$ (for $m \neq 0$) gives

$$\text{Along } y = mx: \quad \lim_{\substack{(x,y)\to(0,0)\\(x,y)=(t,mt)}} \frac{x^2y}{x^4+y^2} = \lim_{t\to 0} \frac{mt^3}{t^4+m^2t^2} = \lim_{t\to 0} \frac{mt}{t^2+m^2} = 0.$$

Therefore, the limit of $\frac{x^2y}{x^4+y^2}$ as $(x,y) \to (0,0)$ along any line through the origin is 0. This does *not* prove that the full limit is 0, though; remember that we need to check *every possible* way to approach $(0,0)$.

Approaching along the parabola $y = x^2$ (chosen so that the denominator will simplify to $x^4 + (x^2)^2 = 2x^4$), we see that

$$\text{Along } y = x^2: \qquad \lim_{\substack{(x,y)\to(0,0)\\(x,y)=(t,t^2)}} \frac{x^2y}{x^4+y^2} = \lim_{t\to 0} \frac{t^2(t^2)}{t^4+(t^2)^2} = \lim_{t\to 0} \frac{t^4}{2t^4} = \frac{1}{2} \neq 0.$$

Since $\frac{x^2y}{x^4+y^2}$ approaches different values as $(x,y) \to (0,0)$ along two different paths (say $y = 0$ and $y = x^2$), we conclude that the limit $\lim_{(x,y)\to(0,0)} \frac{x^2y}{x^4+y^2}$ does not exist.

Example 60. Determine whether or not $\lim_{(x,y)\to(0,0)} \dfrac{x^2+y^2}{\sqrt{x^4+y^4}}$ exists.

Rather than guess and check by computing the limit along lines, we should be able to get some insight into the problem by switching to polar coordinates. (The polar coordinates of a point $(x,y) \in \mathbb{R}^2 - \{(0,0)\}$ are (r,θ), where $r = \sqrt{x^2+y^2}$ is the distance from (x,y) to the origin, and θ is the angle we'd need to rotate the positive x-axis by so that it passes through (x,y). In particular, $(x,y) = (r\cos(\theta), r\sin(\theta))$.)

By making the change $x = r\cos(\theta)$ and $y = r\sin(\theta)$, we see that taking the limit as $(x,y) \to (0,0)$ is equivalent to taking the limit as $r \to 0+$ (where θ is allowed to range through $\mathbb{R}$)), so that

$$\begin{aligned}
\lim_{(x,y)\to(0,0)} \frac{x^2+y^2}{\sqrt{x^4+y^4}} &= \lim_{r\to 0^+} \frac{r^2\cos^2(\theta) + r^2\sin^2(\theta)}{\sqrt{r^4\cos^4(\theta) + r^4\sin^4(\theta)}} \\
&= \lim_{r\to 0^+} \frac{r^2}{r^2\sqrt{\cos^4(\theta)+\sin^4(\theta)}} \\
&= \lim_{r\to 0^+} \frac{1}{\sqrt{\cos^4(\theta)+\sin^4(\theta)}}.
\end{aligned}$$

Therefore it appears that the value of the limit will depend on θ. Since taking $r \to 0+$ while keeping θ fixed is equivalent to allowing $(x,y) \to (0,0)$ along the ray $(r\cos(\theta), r\sin(\theta)$, (for $r > 0$), we therefore expect the limit to fail to exist. To show this, let's choose two particular values of θ that will yield different results.

$$\text{Along the ray } \theta = 0: \qquad \lim_{\substack{(x,y)\to(0,0)\\(x,y)=(r,0)}} \frac{x^2+y^2}{\sqrt{x^4+y^4}} = \lim_{r\to 0+} \frac{1}{\sqrt{1^4}} = \lim_{r\to 0+} 1 = 1.$$

$$\text{Along the ray } \theta = \frac{\pi}{4}: \qquad \lim_{\substack{(x,y)\to(0,0)\\(x,y)=(\frac{r}{\sqrt{2}},\frac{r}{\sqrt{2}})}} \frac{x^2+y^2}{\sqrt{x^4+y^4}} = \lim_{r\to 0+} \frac{1}{\sqrt{\frac{1}{4}+\frac{1}{4}}} = \lim_{r\to 0+} \sqrt{2} = \sqrt{2}.$$

Since $\frac{x^2+y^2}{\sqrt{x^4+y^4}}$ approaches different values as $(x,y) \to (0,0)$ along two different paths, we conclude that the limit $\lim_{(x,y)\to(0,0)} \dfrac{x^2+y^2}{\sqrt{x^4+y^4}}$ does not exist.

Here is an example in $\mathbb{R}^3$ that illustrates a similar phenomenon.

Example 61. Determine whether or not $\lim\limits_{(x,y,z)\to(0,0,0)} \dfrac{xy^2z}{x^4+y^8+z^2}$ exists.

Approaching along any line of the form (ta, tb, tc) with $c = 0$ and at least one of $a, b \neq 0$ produces a limit of 0, since

$$\text{Along } (ta,tb,0): \quad \lim_{\substack{(x,y,z)\to(0,0,0)\\(x,y,z)=(ta,tb,0)}} \frac{xy^2z}{x^4+y^8+z^2} = \lim_{t\to 0} \frac{0}{t^4a^4+t^8b^8} = 0.$$

On the other hand, along any such line for which $c \neq 0$ we have

$$\text{Along } (ta,tb,tc): \quad \lim_{\substack{(x,y,z)\to(0,0,0)\\(x,y,z)=(ta,tb,tc)}} \frac{xy^2z}{x^4+y^8+z^2} = \lim_{t\to 0} \frac{t^4ab^2c}{t^4a^4+t^8b^8+t^2c^2} = \lim_{t\to 0} \frac{t^2ab^2c}{t^2a^4+t^6b^8+c^2} = 0.$$

This justifies the claim that the values of $\dfrac{xy^2z}{x^4+y^8+z^2}$ approach 0 as $(x, y, z) \to (0, 0, 0)$ along any line through the origin.

On the other hand, if we instead approach $(0, 0, 0)$ along the curve parametrized by $\vec{r}(t) = (t^2, t, t^4)$, then

$$\text{Along } (t^2,t,t^4): \quad \lim_{\substack{(x,y,z)\to(0,0,0)\\(x,y,z)=(t^2,t,t^4)}} \frac{xy^2z}{x^4+y^8+z^2} = \lim_{t\to 0} \frac{t^2t^2t^4}{(t^2)^4+t^8+(t^4)^2} = \lim_{t\to 0} \frac{t^8}{3t^8} = \frac{1}{3} \neq 0.$$

Since we have found two paths (say, along the x-axis and along the path (t^2, t, t^4)) approaching $(0, 0, 0)$ along which $\dfrac{xy^2z}{x^4+y^8+z^2}$ approaches different values, we conclude that the limit $\lim\limits_{(x,y,z)\to(0,0,0)} \dfrac{xy^2z}{x^4+y^8+z^2}$ does not exist.

Lecture 21: Continuity

Learning Objectives:

- Determine whether a function is continuous at a point in its domain.
- Apply the Extreme Value Theorem to infer the existence of maximum and minimum values of a continuous function on a compact subset of its domain.

We start with an example that we weren't able to address last time.

Example 62. Determine whether or not $\lim\limits_{(x,y)\to(0,0)} \dfrac{x^2+y^2}{\sqrt{x^4+y^4}}$ exists.

Rather than guess and check by computing the limit along lines, we should be able to get some insight into the problem by switching to polar coordinates. (The polar coordinates of a point $(x,y) \in \mathbb{R}^2 - \{(0,0)\}$ are (r,θ), where $r = \sqrt{x^2+y^2}$ is the distance from (x,y) to the origin, and θ is the angle we'd need to rotate the positive x-axis by so that it passes through (x,y). In particular, $(x,y) = (r\cos(\theta), r\sin(\theta))$.)

By making the change $x = r\cos(\theta)$ and $y = r\sin(\theta)$, we see that taking the limit as $(x,y) \to (0,0)$ is equivalent to taking the limit as $r \to 0+$ (where θ is allowed to range through $\mathbb{R}$)), so that

$$\begin{aligned}\lim_{(x,y)\to(0,0)} \frac{x^2+y^2}{\sqrt{x^4+y^4}} &= \lim_{r\to 0^+} \frac{r^2\cos^2(\theta)+r^2\sin^2(\theta)}{\sqrt{r^4\cos^4(\theta)+r^4\sin^4(\theta)}} \\ &= \lim_{r\to 0^+} \frac{r^2}{r^2\sqrt{\cos^4(\theta)+\sin^4(\theta)}} \\ &= \lim_{r\to 0^+} \frac{1}{\sqrt{\cos^4(\theta)+\sin^4(\theta)}}.\end{aligned}$$

Therefore it appears that the value of the limit will depend on θ. Since taking $r \to 0+$ while keeping θ fixed is equivalent to allowing $(x,y) \to (0,0)$ along the ray $(r\cos(\theta), r\sin(\theta))$, (for $r > 0$), we therefore expect the limit to fail to exist. To show this, let's choose two particular values of θ that will yield different results.

$$\text{Along the ray } \theta = 0: \quad \lim_{\substack{(x,y)\to(0,0)\\(x,y)=(r,0)}} \frac{x^2+y^2}{\sqrt{x^4+y^4}} = \lim_{r\to 0+} \frac{1}{\sqrt{1^4}} = \lim_{r\to 0+} 1 = 1.$$

$$\text{Along the ray } \theta = \frac{\pi}{4}: \quad \lim_{\substack{(x,y)\to(0,0)\\(x,y)=(\frac{r}{\sqrt{2}},\frac{r}{\sqrt{2}})}} \frac{x^2+y^2}{\sqrt{x^4+y^4}} = \lim_{r\to 0+} \frac{1}{\sqrt{\frac{1}{4}+\frac{1}{4}}} = \lim_{r\to 0+} \sqrt{2} = \sqrt{2}.$$

Since $\frac{x^2+y^2}{\sqrt{x^4+y^4}}$ approaches different values as $(x,y) \to (0,0)$ along two different paths, we conclude that the limit $\lim\limits_{(x,y)\to(0,0)} \dfrac{x^2+y^2}{\sqrt{x^4+y^4}}$ does not exist.

Continuity

Just as in single-variable calculus, the notion of continuity for multivariable functions captures the (very nice) situation when the behavior of $\vec{f}(\vec{x})$ *near* $\vec{a}$ agrees with the value of $\vec{f}$ *at* $\vec{a}$, in the following sense[11]

Definition 41. Let $\Omega \subseteq \mathbb{R}^n$ be open and $\vec{f} : \Omega \to \mathbb{R}^m$. We say that $\vec{f}$ is **continuous** at $\vec{a} \in \Omega$ if

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

Note that we can only ask whether a function is continuous or not at a point in its domain; we do not ask about continuity at of a function at points outside of its domain. There are several layers to the previous definition. In particular, we say that $\vec{f}$ is continuous at $\vec{a}$ if both $\lim_{\vec{x}\to\vec{a}} \vec{f}(\vec{x})$ exists and $\lim_{\vec{x}\to\vec{a}} \vec{f}(\vec{x}) = \vec{f}(\vec{a})$. If either of these statements fails, then $\vec{f}$ will fail to be continuous at $\vec{a}$.

Example 63. Let $\vec{v} \in \mathbb{R}^n$ and $c \in \mathbb{R}$. Then

$$f : \mathbb{R}^n \to \mathbb{R}, \quad f(\vec{x}) \stackrel{def}{=} \vec{v} \cdot \vec{x} + c = v_1x_1 + \cdots + v_nx_n + c$$

is continuous at each point in $\mathbb{R}^n$.

To prove this, we must show that $\lim_{\vec{x}\to\vec{a}} f(\vec{x}) = f(\vec{a})$ for every $\vec{a} \in \mathbb{R}^n$.
Let $\vec{a} \in \mathbb{R}^n$ and let $\epsilon > 0$. Choose $\delta = \frac{\epsilon}{\|\vec{v}\|+1}$. Then for every $\vec{x} \in \mathbb{R}^n$ with $0 < \|\vec{x} - \vec{a}\| < \delta$, we have

$$|f(\vec{x}) - f(\vec{a})| = |\vec{v} \cdot \vec{x} + c - \vec{v} \cdot \vec{a} - c| = |\vec{v} \cdot (\vec{x} - \vec{a})| \leq \|\vec{v}\|\|\vec{x} - \vec{a}\| \leq \frac{\|\vec{v}\|}{\|\vec{v}\| + 1}\epsilon < \epsilon.$$

Here, we chose $\delta = \frac{1}{\|\vec{v}\|+1}\epsilon$ instead of $\frac{1}{\|\vec{v}\|}\epsilon$ so that we didn't need to separately consider the case where $\vec{v} = \vec{0}$.

Just as was the case for limits, the standard results about sums, products, multiples, quotients, and compositions of continuous functions apply. In particular, we have the following result[12], proved using the limit definition of continuity.

Theorem 28 (Building Continuous Functions)**.** Let $\Omega \subseteq \mathbb{R}^n$ be open, let $\vec{f} : \Omega \to \mathbb{R}^m$ and $\vec{g} : \Omega \to \mathbb{R}^m$, and suppose $\vec{a} \in \Omega$. If $\vec{f}$ and $\vec{g}$ are each continuous at $\vec{a}$, then

(i) for each $\lambda \in \mathbb{R}$, $\vec{f} + \lambda\vec{g}$ is continuous at $\vec{a}$;

(ii) if $h : \Omega \to \mathbb{R}$ is continuous at $\vec{a}$, then $h\vec{f}$ is continuous at $\vec{a}$ and (provided that $h(\vec{a}) \neq 0$) so is $\frac{1}{h}\vec{f}$;

(iii) if $W \subseteq \mathbb{R}^m$ is open, $\vec{f}(\vec{a}) \in W$, and if $\vec{h} : W \subseteq \mathbb{R}^m \to \mathbb{R}^p$ is continuous at $\vec{f}(\vec{a})$, then $\vec{h} \circ \vec{f} : \Omega \to \mathbb{R}^p$ is continuous at $\vec{a}$.

[11] Here we are restricting ourselves to open sets in order to guarantee that every element of the set is a limit point of the set. Because you will deal with the full generality of limits and continuity in your real analysis course, we will consider only this special case here to avoid obstructing our progress to the higher-level results and applications.

[12] Here we are implicitly using the fact that the intersection of two open sets is open. You will prove this fact on your homework this week.

There is also the expected link between continuity of a vector-valued function and continuity of its component functions.

Theorem 29 (Continuity and Component Functions)**.** Let $\Omega \subseteq \mathbb{R}^n$ be open, let $\vec{a} \in \Omega$, and suppose that $\vec{f} : \Omega \to \mathbb{R}^m$ has component functions $\vec{f} = (f_1, \ldots, f_m)$. Then $\vec{f}$ is continuous at $\vec{a}$ if, and only if, f_j is continuous at $\vec{a}$ for each $1 \leq j \leq m$.

Proof. $\vec{f}$ is continuous at $\vec{a}$ exactly when $\lim_{\vec{x}\to\vec{a}} \vec{f}(\vec{x}) = \vec{f}(\vec{a})$. By the Limits and Component Functions Theorem, this happens exactly when $\lim_{\vec{x}\to\vec{a}} f_j(\vec{x}) = f_j(\vec{a})$ for each $1 \leq j \leq m$, which is equivalent to saying that f_j is continuous at $\vec{a}$ for each $1 \leq j \leq m$. □

Remark 60. As a consequence of the previous theorems, scalar-valued functions that are built from polynomials, rational functions, root functions, trigonometric functions, exponential functions, and logarithms using sums, product, quotients, scalar multiplication, and composition are continuous on their domains. For example, the function $f(x, y, z) = 3\sqrt{\sin(xy) + 2z}$ is continuous throughout its domain. Moreover, a vector-valued function whose component functions have this structure is also continuous throughout its domain. We will use these facts going forward without proof.

Example 64. The function $\|\cdot\| : \mathbb{R}^n \to \mathbb{R}$, $\|\vec{x}\| = \sqrt{x_1^2 + \cdots + x_n^2}$ is continuous throughout $\mathbb{R}^n$, being the composition of a continuous function $(t \mapsto \sqrt{t},\ t \geq 0)$ and a polynomial $(\vec{x} \to x_1^2 + \cdots + x_n^2)$, with domain $\mathbb{R}^n$ because $x_1^2 + \cdots + x_n^2 \geq 0$.

Example 65. Does there exist a number $c \in \mathbb{R}$ such that

$$f : \mathbb{R}^2 \to \mathbb{R}, \quad f(x, y) \stackrel{def}{=} \begin{cases} \frac{x^2+y^2}{\sqrt{x^4+y^4}} & \text{if } (x, y) \neq (0, 0), \\ c & \text{if } (x, y) = (0, 0) \end{cases}$$

is continuous at every point in $\mathbb{R}^2$?

Because the expression $\frac{x^2+y^2}{\sqrt{x^4+y^2}}$ is the quotient of a polynomial and the square root of another polynomial, and is defined for every $(x, y) \neq (0, 0)$, f is continuous at every $(x, y) \in \mathbb{R}^2 - \{(0, 0)\}$. It therefore suffices to check whether f is continuous at $(0, 0)$. But we showed earlier that $\lim_{(x,y)\to(0,0)} f(x, y)$ does not exist. Therefore f cannot be continuous at $(0, 0)$, regardless of the choice of c. The answer to the question is "no".

One *extremely* important property of continuous scalar-valued functions is that, when restricted to a compact subset of their domains, they obtain maximum and minimum values. This is known as the Extreme Value Theorem, and is formally stated as follows.

Theorem 30 (Extreme Value Theorem)**.** Let $\Omega \subseteq \mathbb{R}^n$ be open, let $f : \Omega \to \mathbb{R}$ be continuous, and suppose that $K \subset U$ is compact. Then f attains maximum and minimum values on K, in the sense that there exist $\vec{a}_{min}, \vec{a}_{max} \in K$ with

$$f(\vec{a}_{min}) \leq f(\vec{x}) \leq f(\vec{a}_{max}) \quad \text{for every } \vec{x} \in K.$$

The rigorous proof of the Extreme Value Theorem is beyond the scope of the course because it uses properties of compact sets that require a bit of work to establish. We will use this theorem later on in the course when we discuss optimization.

Lecture 22: Differentiability

Learning Objectives:

- Determine when the partial derivatives of a scalar-valued function exist, and compute them.
- Interpret differentiability of a function of several variables in terms of affine approximation.
- Infer that differentiable functions are continuous.
- Show that the matrix of partial derivatives of a function at a point is the only possible derivative of the function.

Example 66. Let $A \in M_{m\times n}(\mathbb{R})$, $\vec{b} \in \mathbb{R}^n$, and $\vec{c} \in \mathbb{R}^m$. The **affine function**

$$\vec{f} : \mathbb{R}^n \to \mathbb{R}^m, \quad \vec{f}(\vec{x}) \stackrel{def}{=} A(\vec{x} - \vec{b}) + \vec{c}$$

is continuous at each point in $\mathbb{R}^n$.

To prove this, note that if the rows of A are $\vec{r}_1^T, \ldots, \vec{r}_m^T$ (where $\vec{r}_1, \ldots, \vec{r}_m \in \mathbb{R}^n$), then for each $1 \leq j \leq m$ the j-th component function of $\vec{f}$ is

$$f_j(\vec{x}) = \vec{r}_j \cdot (\vec{x} - \vec{b}) + c_j = \vec{r}_j \cdot \vec{x} + (c_j - \vec{r}_j \cdot \vec{b}),$$

which we proved last time was continuous at each point of $\mathbb{R}^n$. Therefore the Continuity and Component Functions theorem implies that $\vec{f}$ is continuous at each point in $\mathbb{R}^n$.

We now turn to the biggest topic of the multivariable differential calculus portion of the course: differentiation. Recall that a single-variable function $f : I \to \mathbb{R}$ (where $I \subseteq \mathbb{R}$ is an interval) is differentiable a point $a \in I$ if

$$\lim_{x\to a} \frac{f(x) - f(a)}{x - a} \quad \text{or, equivalently,} \quad \lim_{h\to 0} \frac{f(a+h) - f(a)}{h}$$

exists. The value of this limit is called the **derivative** of f, and denoted as $f'(a)$. You saw many interpretations of the derivative in your single-variable calculus course: as the instantaneous rate of change of f at a, as slope of line tangent to the graph of f at $(a, f(a))$, and perhaps as part of the linearization $L(x) = f(a) + f'(a)(x - a)$ of f at a (which is an affine function that one can use to approximate $f(x)$ for x near a). You also saw many properties of derivatives (differentiation rules, differentiability implies continuity, etc.) Our goal is to appropriately generalize these ideas and results to the multivariable setting. Perhaps surprisingly, this is one instance where the jump from single-variable to several variables actually illuminates more clearly how one should think.

Crucial to our discussion is the notion of *partial derivatives*, which can be viewed as a naïve extension of single-variable derivatives for scalar-valued functions of several variables. In particular, partial derivatives are what we obtain when we take the single-variable derivative with respect to one of the variables while holding all of the other variables constant. Here is the precise definition.

Definition 42. Let $\Omega \subseteq \mathbb{R}^n$ be open, let $\vec{a} \in \Omega$, and let $f : \Omega \to \mathbb{R}$. For $1 \leq j \leq n$, the **partial derivative** of f with respect to x_j at $\vec{a}$ is

$$\frac{\partial f}{\partial x_j}(\vec{a}) \stackrel{def}{=} \lim_{h\to 0} \frac{f(\vec{a}+h\vec{e}_j) - f(\vec{a})}{h} = \lim_{h\to 0} \frac{f(a_1,\ldots,a_j+h,\ldots,a_n) - f(a_1,\ldots,a_j,\ldots,a_n)}{h},$$

provided that the limit exists. Similarly, if $\vec{f} : \Omega \to \mathbb{R}^m$, then we define

$$\frac{\partial \vec{f}}{\partial x_j}(\vec{a}) \stackrel{def}{=} \lim_{h\to 0} \frac{1}{h}(\vec{f}(\vec{a}+h\vec{e}_j) - \vec{f}(\vec{a})),$$

provided that the limit exists.

Remark 61. Note that the Limits and Component Functions Theorem implies that for $\vec{f} : \Omega \to \mathbb{R}^m$, $\vec{f}(\vec{x}) = (f_1(\vec{x}),\ldots,f_m(\vec{x}))$, the partial derivative of $\vec{f}$ with respect to x_j at $\vec{a}$ exists if, and only if, for each $1 \leq i \leq m$ the partial derivative of f_i with respect to x_j at $\vec{a}$ exists, and

$$\frac{\partial \vec{f}}{\partial x_j}(\vec{a}) = \lim_{h\to 0} \frac{1}{h}(\vec{f}(\vec{a}+h\vec{e}_j) - \vec{f}(\vec{a})) = \lim_{h\to 0} \begin{bmatrix} \frac{f_1(\vec{a}+h\vec{e}_j)-f_1(\vec{a})}{h} \\ \vdots \\ \frac{f_m(\vec{a}+h\vec{e}_j)-f_m(\vec{a})}{h} \end{bmatrix} = \begin{bmatrix} \frac{\partial f_1}{\partial x_j}(\vec{a}) \\ \vdots \\ \frac{\partial f_m}{\partial x_j}(\vec{a}) \end{bmatrix}.$$

Notation 2. The notation $\frac{\partial f}{\partial x_j}(\vec{a})$ is the multivariable analog of the Leibniz notation $\frac{df}{dx}(a)$ for scalar-valued functions of a single variable. We avoid confusion, we will sometimes use the traditional Leibniz notation when dealing with a scalar-valued function of a single variable.
It will sometimes be helpful to have a multivariable analogue of Newton's notation f' for the derivative. The common convention is to replace the $'$ with a subscript indicating with respect to which variable we have differentiated. For example, we might write $f_y(a,b)$ for $\frac{\partial f}{\partial y}(a,b)$, or $f_{x_2}(\vec{a})$ for $\frac{\partial f}{\partial x_2}(\vec{a})$.

Example 67. Because partial derivatives are just single-variable derivatives in one variable (while holding all of the other variables constant), in many cases you can compute partial derivatives using all of the properties of derivatives that you know and love from single-variable calculus. For example, we have that

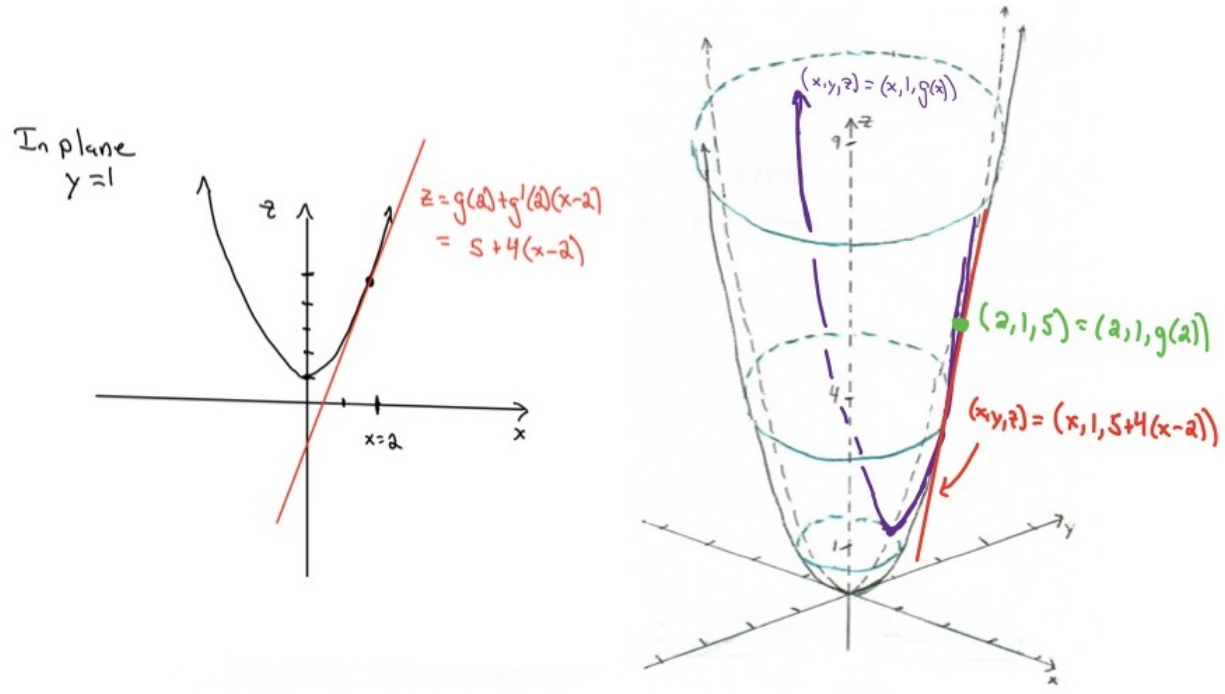
$$\frac{\partial}{\partial y}\left[y^2 e^{xyz}\right] = 2ye^{xyz} + y^2 xze^{xyz}$$

and

$$\frac{\partial}{\partial x}\left[\cos\left(zy - \sin\left(\frac{z}{y^2+1}\right)\right)\right] = 0.$$

Remark 62. It is also possible to import your understanding of single-variable derivatives to give an interpretation of partial derivatives.

For example, the partial derivative of $f(x,y) = x^2+y^2$ at $(2,1)$ with respect to x can be interpreted as $g'(2)$, where $g(x) = f(x,1) = x^2+1$. Therefore, in the plane $y = 1$, the line given by equation $z = g(2) + g'(2)(x-2) = 5 + 4(x-2)$ is tangent to the y-section (at $y = 1$) of the graph of $z = f(x,y)$ at $(2,1,f(2,1)) = (2,1,5)$.



The name "partial derivative" suggests that partial derivatives are not the entire story of differentiation. To generalize the notation of differentiability to functions of several variables, note that the definition of differentiability for single-variable functions can be stated as follows: f is differentiable at a if there exists $c \in \mathbb{R}$ with

$$0 = \lim_{x \to a} \left(\frac{f(x) - f(a)}{x - a} - c \right) = \lim_{x \to a} \frac{f(x) - (f(a) + c(x - a))}{x - a}.$$

Note that $L(x) = f(a) + c(x - a)$ is an affine function, and therefore the statement that f is differentiable at a is really a statement that there is an affine function $L(x) = f(a) + c(x - a)$ that agrees with f at a and that is a "good approximation" for f near a, in the sense that as x approaches a, $|f(x) - L(x)|$ becomes very small relative to $|x - a|$. Of course, the number c is what we call the derivative of f at a.

In the multivariable setting, we will say that $\vec{f}$ is differentiable at $\vec{a}$ if there is an affine function $\vec{f}(\vec{a}) + D(\vec{x} - \vec{a})$ that is a "good approximation" for $\vec{f}(\vec{x})$ when $\vec{x}$ is near $\vec{a}$. More precisely, we have the following.

Definition 43. Let $\Omega \subseteq \mathbb{R}^n$ be open, let $\vec{a} \in \Omega$, and let $\vec{f} : \Omega \to \mathbb{R}^m$. Then $\vec{f}$ is **differentiable** at $\vec{a}$ if there exists $D \in M_{m \times n}(\mathbb{R})$ such that

$$\lim_{\vec{x} \to \vec{a}} \frac{\vec{f}(\vec{x}) - (\vec{f}(\vec{a}) + D(\vec{x} - \vec{a}))}{\|\vec{x} - \vec{a}\|} = \vec{0}.$$

The matrix D will be "the derivative" of $\vec{f}$ at $\vec{a}$, but it is not obvious from the definition that there couldn't be two different matrices D that would make the limit $\vec{0}$. We will prove this soon. To help us see that this definition differentiability produces reasonable results, let's consider the case of affine functions.

Example 68. Let $A \in M_{m \times n}(\mathbb{R})$, $\vec{b} \in \mathbb{R}^n$, and $\vec{c} \in \mathbb{R}^m$. Consider the affine function

$$\vec{f} : \mathbb{R}^n \to \mathbb{R}^m, \quad \vec{f}(\vec{x}) \stackrel{def}{=} A(\vec{x} - \vec{b}) + \vec{c}.$$

Then $\vec{f}$ is differentiable at each $\vec{a} \in \mathbb{R}^n$, for if $\vec{a} \in \mathbb{R}^n$ and $D = A$ then

$$\lim_{\vec{x}\to\vec{a}} \frac{\vec{f}(\vec{x}) - (\vec{f}(\vec{a}) + D(\vec{x} - \vec{a}))}{\|\vec{x} - \vec{a}\|} = \lim_{\vec{x}\to\vec{a}} \frac{A(\vec{x} - \vec{b}) + \vec{c} - (A(\vec{a} - \vec{b}) + \vec{c} + A(\vec{x} - \vec{a}))}{\|\vec{x} - \vec{a}\|} = \lim_{\vec{x}\to\vec{a}} \vec{0} = \vec{0}.$$

In other words, "the derivative" of $\vec{f}(\vec{x}) = A(\vec{x} - \vec{b}) + \vec{c}$ is A! (Again, we'll remove the quotes once we show that this is the matrix D in Definition 43 is unique.)

Lecture 23: More Differentiability

Learning Objectives:

- Describe the relationship between differentiability and continuity.
- Determine the relationship between the matrix of partial derivatives of a vector-valued function and differentiability.
- Show that a vector-valued function is differentiable at a point by checking that its component functions are C^1 in a ball centered at that point.
- Describe the meaning of the derivative as an affine approximation.

This definition also implies that differentiable functions are continuous (just as in single-variable calculus), with virtually the same proof as in single-variable calculus.

Proposition 13. Let $\Omega \subseteq \mathbb{R}^n$ be open, let $\vec{a} \in \Omega$, and suppose $\vec{f} : \Omega \to \mathbb{R}^m$ is differentiable at $\vec{a}$. Then $\vec{f}$ is continuous at $\vec{a}$.

Proof. Let $D \in M_{m\times n}(\mathbb{R})$ be as in the definition of differentiability, and note that for $\vec{x} \in \Omega$ we have

$$\vec{f}(\vec{x}) = \|\vec{x} - \vec{a}\|\Big(\frac{\vec{f}(\vec{x}) - \vec{f}(\vec{a}) - D(\vec{x} - \vec{a})}{\|\vec{x} - \vec{a}\|}\Big) + \vec{f}(\vec{a}) + D(\vec{x} - \vec{a}).$$

Because $\lim\limits_{\vec{x}\to\vec{a}} \|\vec{x} - \vec{a}\| = 0$ and $\lim\limits_{\vec{x}\to\vec{a}} \frac{\vec{f}(\vec{x}) - \vec{f}(\vec{a}) - D(\vec{x}-\vec{a})}{\|\vec{x}-\vec{a}\|} = \vec{0}$,

$$\lim_{\vec{x}\to\vec{a}} \|\vec{x} - \vec{a}\|\Big(\frac{\vec{f}(\vec{x}) - \vec{f}(\vec{a}) - D(\vec{x} - \vec{a})}{\|\vec{x} - \vec{a}\|}\Big) = \vec{0}.$$

It follows (using continuity of affine functions) that

$$\lim_{\vec{x}\to\vec{a}} \vec{f}(\vec{x}) = \lim_{\vec{x}\to\vec{a}} \left(\|\vec{x} - \vec{a}\|\Big(\frac{\vec{f}(\vec{x}) - \vec{f}(\vec{a}) - D(\vec{x} - \vec{a})}{\|\vec{x} - \vec{a}\|}\Big) + \vec{f}(\vec{a}) + D(\vec{x} - \vec{a}) \right) = \vec{0} + \vec{f}(\vec{a}) + \vec{0} = \vec{f}(\vec{a}).$$

□

As promised, the matrix D in the definition of differentiability is unique, so that (as one expects) there is only one 'derivative' of a function at a point.

Theorem 31. Let $\Omega \subseteq \mathbb{R}^n$, let $\vec{a} \in \Omega$, and suppose $\vec{f} : \Omega \to \mathbb{R}^m$ is differentiable at $\vec{a}$. Then the partial derivatives of $\vec{f}$ exist at $\vec{a}$, and $D\vec{f}(\vec{a}) = \begin{bmatrix} \vec{f}_{x_1}(\vec{a}) & \cdots & \vec{f}_{x_n}(\vec{a}) \end{bmatrix} \in M_{m\times n}(\mathbb{R})$ is the unique matrix satisfying

$$\lim_{\vec{x}\to\vec{a}} \frac{\vec{f}(\vec{x}) - (\vec{f}(\vec{a}) + D\vec{f}(\vec{a})(\vec{x} - \vec{a}))}{\|\vec{x} - \vec{a}\|} = \vec{0}.$$

The matrix $D\vec{f}(\vec{a})$ is called the **derivative** of $\vec{f}$ at $\vec{a}$.

Proof. Let $D \in M_{m\times n}(\mathbb{R})$ be a matrix satisfying

$$\lim_{\vec{x}\to\vec{a}} \frac{\vec{f}(\vec{x}) - (\vec{f}(\vec{a}) + D(\vec{x}-\vec{a}))}{\|\vec{x}-\vec{a}\|} = \vec{0}.$$

Let $1 \le j \le n$, and note that taking the limit along the path $\vec{x} = \vec{a} + h\vec{e}_j$ gives

$$\vec{0} = \lim_{h\to 0} \frac{\vec{f}(\vec{a}+h\vec{e}_j) - (\vec{f}(\vec{a}) + D(h\vec{e}_j))}{\|h\vec{e}_j\|} = \lim_{h\to 0} \frac{\vec{f}(\vec{a}+h\vec{e}_j) - (\vec{f}(\vec{a}) + h\vec{d}_j)}{|h|},$$

where $\vec{d}_j$ is the j-th columns of D. By the Coordinate Characterization of limits, for $1 \le i \le m$ we have

$$0 = \lim_{h\to 0} \frac{f_i(\vec{a}+h\vec{e}_j) - (f_i(\vec{a}) + hd_{ij})}{|h|},$$

where d_{ij} is the entry of D in the i-th row and j-th column. Then

$$0 = \lim_{h\to 0} \frac{|f_i(\vec{a}+h\vec{e}_j) - (f_i(\vec{a}) + hd_{ij})|}{|h|} = \lim_{h\to 0} \left| \frac{f_i(\vec{a}+h\vec{e}_j) - (f_i(\vec{a}) + hd_{ij})}{h} \right|,$$

and therefore

$$0 = \lim_{h\to 0} \frac{f_i(\vec{a}+h\vec{e}_j) - (f_i(\vec{a}) + hd_{ij})}{h} = \lim_{h\to 0} \Big(\frac{f_i(\vec{a}+h\vec{e}_j) - f_i(\vec{a})}{h} - d_{ij} \Big).$$

Because $\lim_{h\to 0} d_{ij} = d_{ij}$,

$$d_{ij} = \lim_{h\to 0} \frac{f_i(\vec{a}+h\vec{e}_j) - f_i(\vec{a})}{h},$$

so that $\frac{\partial f_i}{\partial x_j}(\vec{a})$ exists and is equal to d_{ij}. Because the entries of D are uniquely determined as the partial derivatives of the component functions of $\vec{f}$ at $\vec{a}$, D is unique and the theorem is proved. □

Remark 63. Note that, in the context of the previous theorem, $(D\vec{f}(\vec{a}))_{i,j} = \frac{\partial f_i}{\partial x_j}(\vec{a})$.

Definition 44. If we merely know that the partial derivatives of $\vec{f}$ at $\vec{a}$ exist (but not necessarily that $\vec{f}$ is actually differentiable at $\vec{a}$), then the matrix $D\vec{f}(\vec{a}) \stackrel{def}{=} \begin{bmatrix} \vec{f}_{x_1}(\vec{a}) & \cdots & \vec{f}_{x_n}(\vec{a}) \end{bmatrix}$ is called the **Jacobian matrix** of $\vec{f}$ at $\vec{a}$.

Remark 64. As you might expect, we can also characterize the differentiability of a vector-valued function in terms of differentiability of its component functions. To this end, you will prove the following result on your homework.

Proposition 14 (Differentiability and Component Functions)**.** Let $\Omega \subseteq \mathbb{R}^n$ be open, let $\vec{a} \in \Omega$, let $\vec{f} : \Omega \to \mathbb{R}^m$, and suppose that $\vec{f}(\vec{x}) = (f_1(\vec{x}), \ldots, f_m(\vec{x}))$. Then $\vec{f}$ is differentiable at $\vec{a}$ if, and only if, $f_1, \ldots, f_m$ are differentiable at $\vec{a}$. Moreover, for $1 \le k \le m$ the k-th row of $D\vec{f}(\vec{x})$ is $Df_k(\vec{x})$.

Remark 65. Note that if we are trying to prove that a function $\vec{f}:\Omega\to\mathbb{R}^m$ is differentiable at a point $\vec{a}\in\Omega$ using the definition, then the previous theorem implies that $\vec{f}$ has partial derivatives with respect to each variable at $\vec{a}$, and that the Jacobian matrix $Df(\vec{a})$ must satisfy

$$\lim_{\vec{x}\to\vec{a}}\frac{\vec{f}(\vec{x})-(\vec{f}(a)+D(\vec{x}-\vec{a}))}{\|\vec{x}-\vec{a}\|}=\vec{0}.$$

Perhaps surprisingly, even if all of these partial derivatives do exist at $\vec{a}$ and we can write down the matrix of partial derivatives D above, it is not guaranteed that $\vec{f}$ is differentiable at $\vec{a}$. Indeed, the following example illustrates that the existence of partial derivatives is not sufficient to guarantee differentiability.

Example 69. Define $f:\mathbb{R}^2\to\mathbb{R}$, $f(x,y)\stackrel{def}{=}||x|-|y||-|x|-|y|$.

Then $\frac{\partial f}{\partial x}(0,0)$ and $\frac{\partial f}{\partial y}(0,0)$ both exist (and are 0) since

$$\lim_{h\to0}\frac{f(0+h,0)-f(0,0)}{h}=\lim_{h\to0}\frac{||h|-|0||-|h|-|0|-(||0|-|0||-|0|-|0|)}{h}=\lim_{h\to0}0=0$$

and

$$\lim_{h\to0}\frac{f(0,0+h)-f(0,0)}{h}=\lim_{h\to0}\frac{||0|-|h||-|0|-|h|-(||0|-|0||-|0|-|0|)}{h}=\lim_{h\to0}0=0.$$

Therefore, if f were differentiable at $(0,0)$, then the Jacobian matrix $Df(0,0)=\left[\frac{\partial f}{\partial x}(0,0)\quad\frac{\partial f}{\partial y}(0,0)\right]=\begin{bmatrix}0&0\end{bmatrix}$ of f should satisfy the definition of differentiability of f at $(0,0)$. But f is *not* differentiable at $(0,0)$ because

$$\lim_{(x,y)\to(0,0)}=\frac{f(x,y)-(f(0,0)-\begin{bmatrix}0&0\end{bmatrix}\begin{bmatrix}x-0\\y-0\end{bmatrix})}{\|(x,y)-(0,0)\|}=\lim_{(x,y)\to(0,0)}\frac{||x|-|y||-|x|-|y|}{\sqrt{x^2+y^2}}$$

does not exist. To see why, note that approaching along the x-axis gives

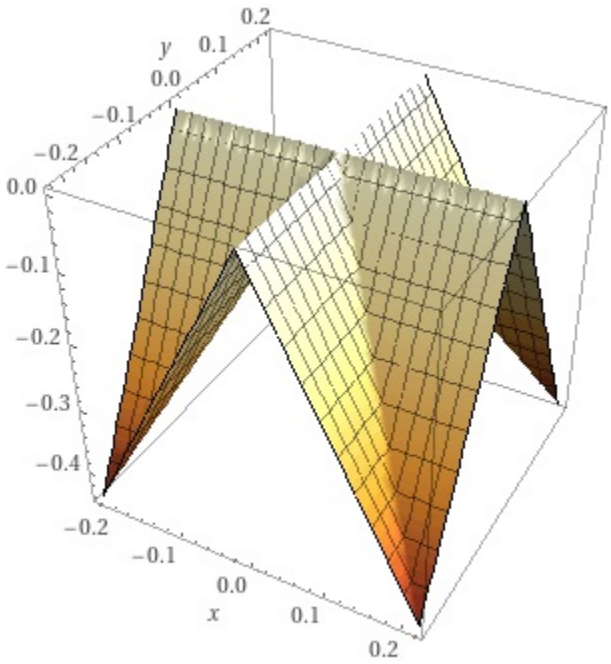
$$\lim_{\substack{(x,y)\to(0,0)\\(x,y)=(t,0)}}\frac{||x|-|y||-|x|-|y|}{\sqrt{x^2+y^2}}=\lim_{t\to0}\frac{|t|-|t|}{|t|}=\lim_{t\to0}0=0,$$

while approaching along the line $y=x$ gives

$$\lim_{\substack{(x,y)\to(0,0)\\(x,y)=(t,t)}}\frac{||x|-|y||-|x|-|y|}{\sqrt{x^2+y^2}}=\lim_{t\to0}\frac{-2|t|}{\sqrt{2}|t|}=\lim_{t\to0}0=-\sqrt{2}.$$

Because approaching $(0,0)$ along two different paths produces a different result, the Nonexistence Theorem implies that this limit does not exist (and therefore that f) is not differentiable.

For intuition about how f can fail to be differentiable at $(0,0)$ even though its partial derivatives exist at $(0,0)$, consider that the definition of differentiability would only be satisfied if the affine function $L(x,y)=f(0,0)+\begin{bmatrix}0&0\end{bmatrix}\begin{bmatrix}x-0\\y-0\end{bmatrix}$ is a good approximation for $f(x,y)$ near $(0,0)$. But the existence of the partial derivatives merely tells us that L that approximates f well for (x,y) near $(0,0)$ that is *on the coordinate axes*—it doesn't say whether we can guarantee that $L(x,y)$ is a good approximation for $f(x,y)$ when (x,y) does not lie on the coordinate axes. Indeed, the graph of $L(x,y)$ will be a plane, but the graph of f is "crumpled" near $(0,0,0)$ and therefore cannot be well-approximated by a plane:



Graph of f(x,y)=| |x|-|y| |-|x|-|y|, produced by Wolfram|Alpha.

The previous example (and one of your homework problems) illustrates that even if the partial derivatives of $\vec{f} : \mathbb{R}^n \to \mathbb{R}^m$ exist at a point $\vec{a}$, it need not be the case that $\vec{f}$ is differentiable at $\vec{a}$. The intuition behind this is that the existence of the partial derivative only shows that $\vec{f}(\vec{x})$ is well-approximated by an affine function when $\vec{x} = \vec{a} + h\vec{e}_j$ for small h and some $1 \leq j \leq n$, but not necessarily when $\vec{x}$ is near $\vec{a}$ but doesn't have this particular form. Luckily, there is a slightly stronger condition on the partial derivatives of the component functions of $\vec{f}$ that guarantees that $\vec{f}$ will be differentiable. We start by making a definition.

Definition 45. Let $\Omega \subseteq \mathbb{R}^n$, let $\vec{a} \in \Omega$, and let $U \subseteq \Omega$ be an open set with $\vec{a} \in U$. We say $\vec{f} : \Omega \to \mathbb{R}$ is C^1 on U if $\vec{f}_{x_1}, \ldots, \vec{f}_{x_n}$ exist and are continuous throughout U.

Example 70. Let $U = \{(r, \theta) \in \mathbb{R}^2 \ : \ r > 0\}$ be the open right-half plane, and consider the function $\vec{P} : U \to \mathbb{R}^2 - \{(0, 0)\}$ defined by

$$\vec{P}(r, \theta) = (r\cos(\theta), r\sin(\theta)).$$

Note that $\vec{P}$ is the map that sends a polar coordinate pair (r, θ) to the point (x, y) represented by those polar coordinates. Because

$$\vec{P}_r(r, \theta) = \begin{bmatrix} \cos(\theta) \\ -r\sin(\theta) \end{bmatrix} \quad \text{and} \quad \vec{P}_\theta(r, \theta) = \begin{bmatrix} \sin(\theta) \\ r\cos(\theta) \end{bmatrix}$$

exist and are continuous throughout U, $\vec{P}$ is C^1 on U.

Remark 66. Note that if $\vec{f} : \Omega \to \mathbb{R}^m$, $\vec{f}(\vec{x}) = (f_1(\vec{x}), \ldots, f_m(\vec{x}))$, then $\vec{f}$ is C^1 on an open set U if, and only if, each component function f_j is C^1 on U. This follows immediately from the observation that $(\vec{f}_{x_i}(\vec{x}))_j = (f_j)_{x_i}(\vec{x})$ (from the remark after the definition of partial derivative), and by the Continuity and Component Functions Theorem.

Theorem 32. Let $\Omega \subseteq \mathbb{R}^n$, let $\vec{a} \in \Omega$, and let $\vec{f} : \Omega \to \mathbb{R}^m$. If $\vec{f}$ is C^1 on an open set $U \subseteq \Omega$ with $\vec{a} \in U$, then $\vec{f}$ is differentiable at $\vec{a}$.

Proof. Let $D\vec{f}(\vec{a})$ denote the Jacobian matrix of $\vec{f}$ at $\vec{a}$. Note that since the k-th row of $D\vec{f}(\vec{a})$ (if it exists) is exactly $Df_k(\vec{a}) = \left[\frac{\partial f_k}{\partial x_1}(\vec{a}) \quad \cdots \quad \frac{\partial f_k}{\partial x_n}(\vec{a})\right]$, the Coordinate Characterization of Limits implies that

$$\lim_{\vec{x}\to\vec{a}} \frac{\vec{f}(\vec{x}) - (\vec{f}(\vec{a}) + D\vec{f}(\vec{a})(\vec{x}-\vec{a}))}{\|\vec{x}-\vec{a}\|} = \vec{0}$$

if and only if

$$\lim_{\vec{x}\to\vec{a}} \frac{f_k(\vec{x}) - (f_k(\vec{a}) + Df_k(\vec{a})(\vec{x}-\vec{a}))}{\|\vec{x}-\vec{a}\|} = 0$$

for each $1 \leq k \leq m$. Therefore it suffices to prove the result for scalar-valued functions (i.e. when $m = 1$).

For a further simplification, we will prove the theorem for $f : \Omega \to \mathbb{R}$ when $\Omega \subseteq \mathbb{R}^2$ (the case where $n > 2$ is almost identical, just with worse notation). Our assumption is that f is C^1 on an open subset U of Ω containing $\vec{a}$, and therefore on an open ball $B_r(\vec{a})$ centered at $\vec{a} = (a, b)$. We wish to show that

$$\begin{aligned} 0 &= \lim_{(x,y)\to(a,b)} \frac{f(x,y) - \left(f(a,b) + \begin{bmatrix} f_x(a,b) & f_y(a,b)\end{bmatrix}\begin{bmatrix} x-a \\ y-b\end{bmatrix}\right)}{\|(x,y)-(a,b)\|} \\ &= \lim_{(x,y)\to(a,b)} \frac{f(x,y) - f(a,b) - f_x(a,b)(x-a) - f_y(a,b)(y-b)}{\|(x,y)-(a,b)\|}. \end{aligned}$$

The key step in the proof is to estimate $|f(x,y) - f(a,b) - f_x(a,b)(x-a) - f_y(a,b)(y-b)|$. To do this, we first write

$$\begin{aligned} &f(x,y) - f(a,b) - f_x(a,b)(x-a) - f_y(a,b)(y-b) \\ &\qquad = (f(x,y) - f(a,y)) - f_x(a,b)(x-a) + (f(a,y) - f(a,b)) - f_y(a,b)(y-b). \end{aligned}$$

Applying the Mean Value Theorem to the function $y \mapsto f(a,y) - f(a,b)$ implies that there is d (dependent on y) between b and y such that $f(a,y) - f(a,b) = f_y(a,d)(y-b)$.

Similarly, applying the Mean Value Theorem to the function $x \mapsto f(x,y) - f(a,y)$ implies that there is c (dependent on x and y) between x and a such that $f(x,y) - f(a,y) = f_x(c,y)(x-a)$. Therefore

$$f(x,y) - f(a,b) - f_x(a,b)(x-a) - f_y(a,b)(y-b) = (f_x(c,y) - f_x(a,b))(x-a) + (f_y(a,d) - f_y(a,b))(y-b),$$

so that

$$\begin{aligned} f(x,y) - f(a,b) - f_x(a,b)(x-a) - &f_y(a,b)(y-b)| \\ &= \left|\begin{bmatrix} f_x(c,y) - f_x(a,b) \\ f_y(a,d) - f_y(a,b)\end{bmatrix} \cdot \begin{bmatrix} x-a \\ y-b\end{bmatrix}\right| \\ &\leq \|(f_x(c,y) - f_x(a,b), f_y(a,d) - f_y(a,b))\|\|\vec{x}-\vec{a}\|. \end{aligned}$$

For $(x,y) \in B_r(\vec{a})$, define $g(x,y) = \|(f_x(c,y) - f_x(a,b), f_y(a,d) - f_y(a,b))\|$, where c and d are as above. We first claim that $\lim_{\vec{x}\to\vec{a}} g(x,y) = 0$. To see this, let $\epsilon > 0$ and choose $\delta > 0$ such that

$|f_x(w,z) - f_x(a,b)| < \frac{\epsilon}{2}$ and $|f_y(w,z) - f_y(a,b)| < \frac{\epsilon}{2}$ whenever $0 < \|(w,z) - (a,b)\| < \delta$. Then whenever $0 < \|(x,y) - (a,b)\| < \delta$, note that since c is between a and x, $\|(c,y) - (a,b)\| \leq \|(x,y) - (a,b)\| < \delta$ and, because d is between b and y, $\|(a,d) - (a,b)\| \leq \|(a,y) - (a,b)\| \leq \|(x,y) - (a,b)\| < \delta$. Therefore

$$\begin{aligned} |g(x,y) - 0| &= \|(f_x(c,y) - f_x(a,b), f_y(a,d) - f_y(a,b))\| \\ &\leq |f_x(c,y) - f_x(a,b)| + |f_y(a,d) - f_y(a,b)| \\ &< \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon. \end{aligned}$$

Now we note that for each (x,y) in a punctured ball centered at $\vec{a}$,

$$\begin{aligned} &\left| \frac{f(x,y) - f(a,b) - f_x(a,b)(x-a) - f_y(a,b)(y-b)}{\|(x,y) - (a,b)\|} \right| \\ &= \frac{|f(x,y) - f(a,b) - f_x(a,b)(x-a) - f_y(a,b)(y-b)|}{\|(x,y) - (a,b)\|} \\ &\leq \frac{g(x,y)\|(x,y) - (a,b)\|}{\|(x,y) - (a,b)\|} = g(x,y). \end{aligned}$$

Because $\lim_{(x,y)\to(a,b)} g(x,y) = 0$, the Squeeze Theorem implies that

$$\lim_{(x,y)\to(a,b)} \frac{f(x,y) - f(a,b) - f_x(a,b)(x-a) - f_y(a,b)(y-b)}{\|(x,y) - (a,b)\|} = 0,$$

and the theorem is proved. □

Example 71. Let $U = \{(r,\theta) \in \mathbb{R}^2 \ : \ r > 0\}$ be the open right-half plane, and consider the function $\vec{P} : U \to \mathbb{R}^2 - \{(0,0)\}$ defined by

$$\vec{P}(r,\theta) = (r\cos(\theta), r\sin(\theta)).$$

Note that $\vec{P}$ is the map that sends a polar coordinate pair (r,θ) to the point (x,y) represented by those polar coordinates.

Show that $\vec{P}$ is differentiable at each point in U, and compute $D\vec{P}(r,\theta)$.

We have already noted that $\vec{P}$ is C^1 on U. Therefore $\vec{P}$ is differentiable at each point in U with

$$D\vec{P}(r,\theta) = \begin{bmatrix} \frac{\partial P_1}{\partial r}(r,\theta) & \frac{\partial P_1}{\partial \theta}(r,\theta) \\ \frac{\partial P_2}{\partial r}(r,\theta) & \frac{\partial P_2}{\partial \theta}(r,\theta) \end{bmatrix} = \begin{bmatrix} \cos(\theta) & -r\sin(\theta) \\ \sin(\theta) & r\cos(\theta) \end{bmatrix}.$$

Remark 67. Differentiability illustrates a strong connection between linear algebra and calculus. In particular, if $\vec{f}$ is differentiable at $\vec{a}$ then

$$\vec{f}(x) = \vec{f}(\vec{a}) + D\vec{f}(\vec{a})(\vec{x} - \vec{a}) + \underbrace{(\vec{f}(\vec{x}) - (\vec{f}(\vec{a}) + D\vec{f}(\vec{a})(\vec{x} - \vec{a})))}_{\vec{R}(\vec{x})},$$

where $\lim_{\vec{x}\to\vec{a}} \frac{\vec{R}(\vec{x})}{\|\vec{x}-\vec{a}\|} = \vec{0}$. In particular, for $\vec{x}$ close to $\vec{a}$,

$$\vec{f}(\vec{x}) \approx \vec{f}(\vec{a}) + D\vec{f}(\vec{a})(\vec{x} - \vec{a}).$$

Intuitively, this says that the affine function $L(\vec{x}) = \vec{f}(\vec{a}) + D\vec{f}(\vec{a})(\vec{x} - \vec{a})$ is a good approximation for $\vec{f}(\vec{x})$ for $\vec{x}$ near $\vec{a}$. To truly understand the meaning of the derivative $D\vec{f}(\vec{a})$, it can be illuminating to rewrite the approximation above as

$$\vec{f}(\vec{x}) - \vec{f}(\vec{a}) \approx D\vec{f}(\vec{a})(\vec{x} - \vec{a}).$$

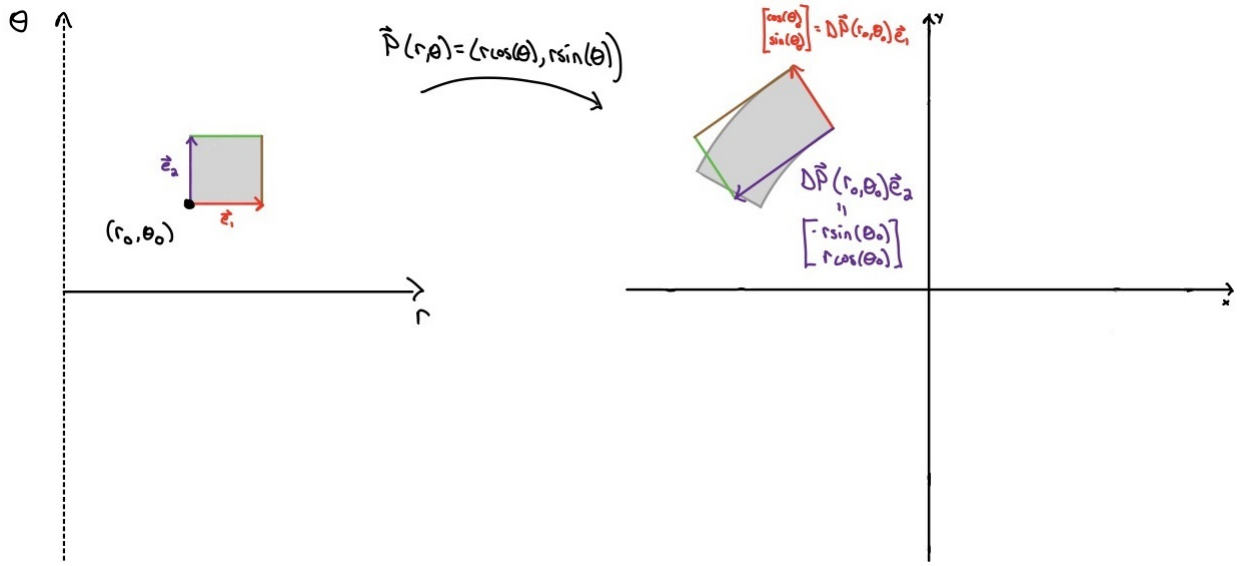
In particular, the left-hand side $\vec{f}(\vec{x}) - \vec{f}(\vec{a})$ is the vector in $\mathbb{R}^m$ that describes the displacement of $\vec{f}(\vec{x})$ from $\vec{f}(\vec{a})$. By the approximation above, this (actual) displacement can be approximated by $D\vec{f}(\vec{a})(\vec{x}-\vec{a})$, which is the image of $\vec{x}-\vec{a}$ (the displacement of $\vec{x}$ from $\vec{a}$) under the linear transformation described by $D\vec{f}(\vec{a})$. In this sense, the matrix $D\vec{f}(\vec{a})$ (interpreted as representing a linear transformation that sends displacement vectors based at $\vec{a}$ to displacement vectors based at $\vec{f}(\vec{a})$) captures how $\vec{f}$ deforms space near $\vec{a}$.

Example 72. For an example of how to think about this, consider the polar coordinate example $\vec{P}(r, \theta) = (r\cos(\theta), r\sin(\theta))$. Then the derivative of $\vec{P}$ at (r, θ) has the form

$$D\vec{P}(r, \theta) = \begin{bmatrix} \cos(\theta) & -r\sin(\theta) \\ \sin(\theta) & r\cos(\theta) \end{bmatrix}.$$

For fixed (r_0, θ_0), the matrix $D\vec{P}(r_0, \theta_0)$ rotates $\vec{e}_1$ counterclockwise by θ, and also rotates $\vec{e}_2$ counterclockwise by θ and then scales the result by r. The result is that $D\vec{P}(r_0, \theta_0)$ sends a rectangle of the form $E(a\vec{e}_1, b\vec{e}_2)$ to another rectangle of the form $E(aR_\theta(\vec{e}_1), brR_\theta(\vec{e}_2))$, where $R_\theta : \mathbb{R}^2 \to \mathbb{R}^2$ is counterclockwise rotation by θ.

The notion that $D\vec{P}(r_0, \theta_0)$ captures (at least approximately) how $\vec{P}$ "deforms space" near (r_0, θ_0) is captured in the following picture.



Lecture 24: Second Derivatives

Learning Objectives:

- Determine when a function is C^k on a set U.
- Fluently employ the notation for higher-order partial derivatives.
- Establish sufficient conditions on a function that ensure that the order in which one computes iterated partial derivatives is irrelevant.

Second Derivatives

Just as we can consider the derivative of a differentiable function $f : I \subseteq \mathbb{R} \to \mathbb{R}$ as a function in its own right, when a scalar-valued function $f : X \subseteq \mathbb{R}^n \to \mathbb{R}$ is differentiable with respect to one of its variables (say x_j) in a neighborhood of a point (say $\vec{a}$) in its domain, then we can consider $\frac{\partial f}{\partial x_j}(\vec{x})$ as a function and ask whether or not $\frac{\partial f}{\partial x_j}(\vec{x})$ is differentiable with respect to one of its variables. In this way we can talk about higher-order partial derivatives. To make this precise, we give the following definition.

Definition 46. Let $\Omega \subseteq \mathbb{R}^n$ be open, let $\vec{a} \in \Omega$, let $\vec{f} : \Omega \to \mathbb{R}^m$, and let $1 \leq j, k \leq n$. If the partial derivative $\frac{\partial \vec{f}}{\partial x_j}(\vec{x})$ exists throughout an open set containing $\vec{a}$, and if $\frac{\partial \vec{f}}{\partial x_j}$ is differentiable with respect to x_k at $\vec{a}$, then we say that

$$\frac{\partial^2 \vec{f}}{\partial x_k \partial x_j}(\vec{a}) \stackrel{def}{=} \frac{\partial}{\partial x_k}\Big[\frac{\partial \vec{f}}{\partial x_j}\Big](\vec{a})$$

is the **second-order partial derivative of $\vec{f}$ with respect to x_j and x_k**. If $j \neq k$ then the second-order partial derivative is called **mixed**, while if $j = k$ the second-order partial derivative is called **pure**.

Remark 68. Higher order partial derivatives (i.e. third-order, fourth-order, etc) are defined analogously. As an example,

$$\frac{\partial^5 \vec{f}}{\partial x_3 \partial x_4 \partial x_2 \partial x_2 \partial x_4}(\vec{a}) = \frac{\partial}{\partial x_3}\Big[\frac{\partial}{\partial x_4}\Big[\frac{\partial}{\partial x_2}\Big[\frac{\partial}{\partial x_2}\Big[\frac{\partial \vec{f}}{\partial x_4}\Big]\Big]\Big]\Big](\vec{a})$$

would denote a mixed fifth-order partial derivative, while $\frac{\partial^3 \vec{f}}{\partial x_2 \partial x_2 \partial x_2}(\vec{a})$ denotes a pure third-order derivative.

Notation 3. In the setting of the above definition, we also use the notation $\vec{f}_{x_j x_k}(\vec{a}) = (\vec{f}_{x_j})_{x_k}(\vec{a})$. Note that the order for the 'subscript' notation is the reverse of that of the 'operator' notation; this is merely an artifact of the notation, but you should be sure to adhere to the standard conventions to avoid confusion.

To keep track of the 'smoothness' of a function $\vec{f}$, we extend our definition of C^1 to describe situations where the higher-order derivatives of $\vec{f}$ exist and are continuous.

Definition 47. Let $\Omega \subseteq \mathbb{R}^n$, let $\vec{a} \in \Omega$, and let $U \subseteq \Omega$ be an open set with $\vec{a} \in U$. We say $\vec{f} : \Omega \to \mathbb{R}$ is C^m on U if $\vec{f}$ is continuous and if all partial derivative of f of order m or less exist and are continuous throughout U.

The order in which one takes partial derivatives is very important. After all $\frac{\partial^2 f}{\partial x \partial y}(\vec{a})$ is the derivative of $\frac{\partial f}{\partial y}$ with respect to x at $\vec{a}$, while $\frac{\partial^2 f}{\partial y \partial x}(\vec{a})$ is the derivative of $\frac{\partial f}{\partial x}$ with respect to y at $\vec{a}$. There is no reason to expect that these functions are related. Indeed, you will explore the 'standard' counterexample to this on your homework.

Example 73. The following function is a standard example of one for which the mixed partial derivatives are not equal (at $(0,0)$):

$$f : \mathbb{R}^2 \to \mathbb{R}, \quad f(x,y) = \begin{cases} xy\left(\frac{x^2-y^2}{x^2+y^2}\right) & \text{if } (x,y) \neq (0,0), \\ 0 & \text{if } (x,y) = (0,0). \end{cases}$$

On your homework you will give the details.

Perhaps unexpectedly, if your function is 'smooth enough' then these mixed partial derivatives are equal. This result, due to Clairaut, shows that if a function is C^2 then its mixed second-order partial derivatives do not depend on the order in which you compute them.

Theorem 33 (Clairaut)**.** Let $\Omega \subseteq \mathbb{R}^n$, let $\vec{a} \in \Omega$, and let $\vec{f} : \Omega \to \mathbb{R}^m$. If $\vec{f}$ is C^2 on an open set containing $\vec{a}$, then for each $1 \leq j, k \leq n$ we have

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

The proof of this result is similar in flavor (but slightly trickier than) the proof that C^1 functions are differentiable. The proof in your book (on pp. 139-140) is the standard proof and is well-written, so we will not reproduce it here.

Example 74. Consider the function $f(x,y) = x^3 \sin(xy)$. The first-order partial derivatives of f are

$$f_x(x,y) = 3x^2 \sin(xy) + x^3 y \cos(xy) \quad \text{and} \quad f_y(x,y) = x^4 \cos(xy),$$

which are continuous throughout $\mathbb{R}^2$. Therefore f is C^1 on $\mathbb{R}^2$. But each of these partial derivatives is C^1 on $\mathbb{R}^2$ as well, since

$$f_{xx}(x,y) = (f_x)_x(x,y) = 6x \sin(xy) + 6x^2 y \cos(xy) - x^3 y^2 \sin(xy),$$

$$f_{xy}(x,y) = (f_x)_y(x,y) = 4x^3 \cos(xy) - x^4 y \sin(xy)$$

and

$$f_{yx}(x,y) = (f_y)_x(x,y) = 4x^3 \cos(xy) - x^4 y \sin(xy), \quad f_{yy}(x,y) = (f_y)_y(x,y) = -x^5 \sin(xy)$$

all exist and are continuous throughout $\mathbb{R}^2$. Therefore f is C^2 on $\mathbb{R}^2$.

We also have that $f_{xy}(x,y) = 4x^3 \cos(xy) - x^4 y \sin(xy) = f_{yx}(x,y)$. Because f is C^2, Clairaut's Theorem implies that these mixed partial derivatives must necessarily be equal; our computation merely confirmed this.

Remark 69. Second-order partial derivatives and Clairaut's Theorem will play a big role later on in the course in the following way. If $f : \Omega \to \mathbb{R}$ is C^2 on an open set $U \subseteq \Omega \subseteq \mathbb{R}^n$, then f is C^1 on U and is therefore differentiable at each point $\vec{x} \in U$, with

$$Df(\vec{x}) = \begin{bmatrix} \frac{\partial f}{\partial x_1}(\vec{x}) & \cdots & \frac{\partial f}{\partial x_n}(\vec{x}) \end{bmatrix}.$$

But because f is C^2 on U, each of $\frac{\partial f}{\partial x_1}, \ldots, \frac{\partial f}{\partial x_n}$ is C^1 on U and is therefore differentiable. By the Differentiability and Component Functions proposition, this means that Df (thought of as having outputs in $\mathbb{R}^n$) is differentiable and

$$D(Df)(\vec{a}) = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1 \partial x_1}(\vec{a}) & \frac{\partial^2 f}{\partial x_2 \partial x_1}(\vec{a}) & \cdots & \frac{\partial^2 f}{\partial x_n \partial x_1}(\vec{a}) \\ \frac{\partial^2 f}{\partial x_1 \partial x_2}(\vec{a}) & \frac{\partial^2 f}{\partial x_2 \partial x_2}(\vec{a}) & \cdots & \frac{\partial^2 f}{\partial x_n \partial x_2}(\vec{a}) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_1 \partial x_n}(\vec{a}) & \frac{\partial^2 f}{\partial x_2 \partial x_n}(\vec{a}) & \cdots & \frac{\partial^2 f}{\partial x_n \partial x_n}(\vec{a}) \end{bmatrix} = \begin{bmatrix} f_{x_1 x_1}(\vec{a}) & f_{x_1 x_2}(\vec{a}) & \cdots & f_{x_1 x_n}(\vec{a}) \\ f_{x_2 x_1}(\vec{a}) & f_{x_2 x_2}(\vec{a}) & \cdots & f_{x_2 x_n}(\vec{a}) \\ \vdots & \vdots & \ddots & \vdots \\ f_{x_n x_1}(\vec{a}) & f_{x_n x_2}(\vec{a}) & \cdots & f_{x_n x_n}(\vec{a}) \end{bmatrix}.$$

This matrix (sometimes notated as $D^2 f(\vec{a})$ or $Hf(\vec{a})$) is called the **Hessian** matrix of f at $\vec{a}$. It is an appropriate notion of "second derivative" of f. Note that if f is C^2, the Clairaut's Theorem implies that $D^2 f(\vec{a})$ is symmetric, and therefore the Spectral Theorem applies. Indeed, we will see next quarter that for $\vec{x}$ near $\vec{a}$,

$$f(\vec{x}) \approx f(\vec{a}) + Df(\vec{a})(\vec{x} - \vec{a}) + \frac{1}{2}(\vec{x} - \vec{a}) \cdot \Big(D^2 f(\vec{a})(\vec{x} - \vec{a}) \Big).$$

In other words, a C^2 scalar-valued function can be "well-approximated" locally by a polynomial of degree 2 or less. As you might predict, exploring the consequences of this will involve a full application of all of our work on quadratic forms.

Lecture 25: The Chain Rule

Learning Objectives:

- Compute derivatives and partial derivatives of compositions of differentiable functions.

The single-variable chain rule $(f \circ g)'(x) = f'(g(x))g'(x)$ generalizes to vector-valued functions of several variables in an extremely satisfying way.

Remark 70. The intuition behind the chain rule is best understood from our understanding of the derivative as a measure of how much a function "deforms" space. To see why, suppose that $\vec{g} : \mathbb{R}^p \to \mathbb{R}^m$ is differentiable at $\vec{a}$, and that $\vec{f} : \mathbb{R}^n \to \mathbb{R}^m$ is differentibale at $\vec{g}(\vec{a})$. One conclusion of the Chain Rule is that $\vec{f} \circ \vec{g}$ will be differentiable at $\vec{a}$, but it may not be obvious what we should expect $D(\vec{f} \circ \vec{g})(\vec{a})$ to be.

To predict this, note that since $\vec{g}$ is differentiable at $\vec{a}$ we have

$$\vec{g}(\vec{x}) - \vec{g}(\vec{a}) \approx D\vec{g}(\vec{a})(\vec{x} - \vec{a}) \quad \text{when } \vec{x} \text{ is near } \vec{a},$$

and if $\vec{b} = \vec{g}(\vec{a})$, then since $\vec{f}$ is differentiable at $\vec{b}$ we have

$$\vec{f}(\vec{t}) - \vec{f}(\vec{b}) \approx D\vec{f}(\vec{b})(\vec{t} - \vec{b}) \quad \text{when } \vec{t} \text{ is near } \vec{b}.$$

But differentiable functions are continuous, and therefore $\vec{g}(\vec{x})$ is near $\vec{b} = \vec{g}(\vec{a})$ whenever $\vec{x}$ is near $\vec{a}$, so that (in terms of approximations)

$$\vec{f} \circ \vec{g}(\vec{x}) - \vec{f} \circ \vec{g}(\vec{a}) = \vec{f}(\vec{g}(\vec{x})) - \vec{f}(\vec{g}(\vec{a})) \approx D\vec{f}(\vec{g}(\vec{a}))(\vec{g}(\vec{x}) - \vec{g}(\vec{a})) \approx D\vec{f}(\vec{g}(\vec{a}))D\vec{g}(\vec{a})(\vec{x} - \vec{a}).$$

Therefore, as long as $\vec{f} \circ \vec{g}$ is differentiable at $\vec{a}$, we expect that $D(\vec{f} \circ \vec{g})(\vec{a}) = D\vec{f}(\vec{g}(\vec{a}))D\vec{g}(\vec{a})$. This is indeed the case.

Theorem 34 (Chain Rule)**.** Let $\Omega \subseteq \mathbb{R}^n$ and $Y \subseteq \mathbb{R}^p$ be open, let $\vec{f} : \Omega \to \mathbb{R}^m$ and $\vec{g} : Y \to \mathbb{R}^n$, and suppose that $\vec{g}$ is differentiable at $\vec{a} \in Y$, that[13] $\vec{g}(Y) \subseteq \Omega$, and that $\vec{f}$ is differentiable at $\vec{g}(\vec{a})$. Then $\vec{f} \circ \vec{g} : Y \to \mathbb{R}^m$ is differentiable at $\vec{a}$ and

$$D(\vec{f} \circ \vec{g})(\vec{a}) = D\vec{f}(\vec{g}(\vec{a}))D\vec{g}(\vec{a}).$$

Proof. Write $\vec{b} = \vec{g}(\vec{a})$. For $\vec{t} \in \Omega$, define

$$\vec{H}(\vec{t}) = \begin{cases} \dfrac{1}{\|\vec{t} - \vec{b}\|}(\vec{f}(\vec{t}) - \vec{f}(\vec{b}) - D\vec{f}(\vec{b})(\vec{t} - \vec{b})) & \text{if } \vec{t} \neq \vec{b}, \\ \vec{0} & \text{if } \vec{t} = \vec{b}. \end{cases}$$

[13] Recall that if $g : A \to B$ is a function and $C \subseteq A$, then $g(C) = \{g(c) \ : \ c \in C\} \subseteq B$ is called the **image of C under g**.

Since $\vec{f}$ is differentiable at $\vec{b}$, $\lim_{\vec{t}\to\vec{b}} \vec{H}(\vec{t}) = \vec{0} = \vec{H}(\vec{b})$, so that $\vec{H}$ is continuous at $\vec{b}$. Moreover, we have

$$\|\vec{t}-\vec{b}\|\vec{H}(\vec{t}) = \vec{f}(\vec{t}) - \vec{f}(\vec{b}) - D\vec{f}(\vec{b})(\vec{t}-\vec{b}), \quad \text{for all } \vec{t}\in\Omega.$$

In particular, since $\vec{g}(Y)\subseteq\Omega$ we have

$$\|\vec{g}(\vec{x})-\vec{b}\|\vec{H}(\vec{g}(\vec{x})) = \vec{f}(\vec{g}(\vec{x})) - \vec{f}(\vec{b}) - D\vec{f}(\vec{b})(\vec{g}(\vec{x})-\vec{b}), \quad \text{for all } \vec{x}\in Y.$$

Since $\vec{g}$ is continuous at $\vec{a}$, the composition $\vec{H}\circ\vec{g}$ is continuous at $\vec{a}$, and therefore the continuity of the norm function (and the fact that the composition of continuous functions is continuous) gives that

$$\lim_{\vec{x}\to\vec{a}} \|\vec{H}(\vec{g}(\vec{x}))\| = \|\vec{H}(\vec{b})\| = 0. \tag{9}$$

For $\vec{x}\in Y\backslash\{\vec{a}\}$, define

$$\begin{aligned}
\vec{G}(\vec{x}) &\stackrel{def}{=} \frac{1}{\|\vec{x}-\vec{a}\|}(\vec{f}(\vec{g}(\vec{x})) - \vec{f}(\vec{b}) - D\vec{f}(\vec{b})D\vec{g}(\vec{a})(\vec{x}-\vec{a})).\\
&= \frac{1}{\|\vec{x}-\vec{a}\|}(\vec{f}(\vec{g}(\vec{x})) - \vec{f}(\vec{b}) - D\vec{f}(\vec{b})(\vec{g}(\vec{x})-\vec{b}))\\
&\qquad + \frac{1}{\|\vec{x}-\vec{a}\|}D\vec{f}(\vec{b})[\vec{g}(\vec{x}) - \vec{g}(\vec{a}) - D\vec{g}(\vec{a})(\vec{x}-\vec{a})]\\
&= \frac{\|\vec{g}(\vec{x})-\vec{b}\|}{\|\vec{x}-\vec{a}\|}\vec{H}(\vec{g}(\vec{x})) + \frac{1}{\|\vec{x}-\vec{a}\|}D\vec{f}(\vec{b})(\vec{g}(\vec{x}) - \vec{g}(\vec{a}) - D\vec{g}(\vec{a})(\vec{x}-\vec{a})).
\end{aligned}$$

By Exercise 5 on Homework 7, there are positive constants M and N such that $\|D\vec{f}(\vec{b})\vec{t}\| \le M\|\vec{t}\|$ for all $\vec{t}\in\Omega$, and $\|D\vec{g}(\vec{a})\vec{x}\| \le N\|\vec{x}\|$ for all $\vec{x}\in Y$. Since

$$\begin{aligned}
\|\vec{g}(\vec{x})-\vec{b}\| &\le \|\vec{g}(\vec{x}) - \vec{b} - D\vec{g}(\vec{a})(\vec{x}-\vec{a})\| + \|D\vec{g}(\vec{a})(\vec{x}-\vec{a})\|\\
&\le \|\vec{g}(\vec{x}) - \vec{b} - D\vec{g}(\vec{a})(\vec{x}-\vec{a})\| + N\|\vec{x}-\vec{a}\|,
\end{aligned}$$

the triangle inequality and some light algebra imply that for all $\vec{x}\in Y\backslash\{\vec{a}\}$,

$$\begin{aligned}
0 \le \|\vec{G}(\vec{x})\| &\le \frac{\|\vec{g}(\vec{x}) - \vec{b} - D\vec{g}(\vec{a})(\vec{x}-\vec{a})\|}{\|\vec{x}-\vec{a}\|}\|\vec{H}(\vec{g}(\vec{x}))\| + N\|\vec{H}(\vec{g}(\vec{x}))\|\\
&\qquad + M\frac{\|\vec{g}(\vec{x}) - \vec{b} - D\vec{g}(\vec{a})(\vec{x}-\vec{a})\|}{\|\vec{x}-\vec{a}\|}.
\end{aligned} \tag{10}$$

Since $\vec{g}$ is differentiable we have $\lim_{\vec{x}\to\vec{a}} \frac{\|\vec{g}(\vec{x})-\vec{b}-D\vec{g}(\vec{a})(\vec{x}-\vec{a})\|}{\|\vec{x}-\vec{a}\|} = 0$. By (9), $\lim_{\vec{x}\to\vec{a}}$(Right-Hand-Side of (10)) $= 0$. The Squeeze Theorem then implies that

$$0 = \lim_{\vec{x}\to\vec{a}} \|\vec{G}(\vec{x})\| = \lim_{\vec{x}\to\vec{a}} \frac{\|\vec{f}(\vec{g}(\vec{x})) - \vec{f}(\vec{g}(\vec{a})) - D\vec{f}(\vec{g}(\vec{a}))D\vec{g}(\vec{a})(\vec{x}-\vec{a})\|}{\|\vec{x}-\vec{a}\|},$$

so that

$$\lim_{\vec{x}\to\vec{a}} \frac{\vec{f}(\vec{g}(\vec{x})) - \vec{f}(\vec{g}(\vec{a})) - D\vec{f}(\vec{g}(\vec{a}))D\vec{g}(\vec{a})(\vec{x}-\vec{a})}{\|\vec{x}-\vec{a}\|} = \vec{0}.$$

We conclude that $\vec{f}\circ\vec{g}$ is differentiable at $\vec{a}$ and

$$D[\vec{f}\circ\vec{g}](\vec{a}) = D\vec{f}(\vec{g}(\vec{a}))D\vec{g}(\vec{a}).$$

□

Example 75. Note that if $\vec{f}:\mathbb{R}^n\to\mathbb{R}^m$ is an affine function of the form $\vec{f}(\vec{t})=A\vec{t}+\vec{a}$, and if $\vec{g}:\mathbb{R}^p\to\mathbb{R}^n$ is also affine and has the form $\vec{g}(\vec{x})=B\vec{x}+\vec{b}$, then because affine functions are differentiable at every point the Chain Rule implies that $\vec{f}\circ\vec{g}:\mathbb{R}^p\to\mathbb{R}^m$ is differentiable throughout $\mathbb{R}^p$ and $D(\vec{f}\circ\vec{g})(\vec{x})=AB$ for every $\vec{x}\in\mathbb{R}^p$.

Of course, we could also have seen this by noting that $\vec{f}\circ\vec{g}$ is affine with matrix AB, since

$$\vec{f}\circ\vec{g}(\vec{x})=A\vec{g}(\vec{x})+\vec{a}=A(B\vec{x}+\vec{b})+\vec{a}=(AB)\vec{x}+(A\vec{b}+\vec{a}),$$

and therefore $\vec{f}\circ\vec{g}$ is differentiable at every $\vec{x}\in\mathbb{R}^p$ with $D(\vec{f}\circ\vec{g})(\vec{x})=AB$.

Example 76. Define

$$f:\mathbb{R}^3\to\mathbb{R},\quad f(x,y,z)\stackrel{def}{=}x+y+z+\ln(x^2+z^2+1)$$

and

$$\vec{g}:\mathbb{R}^2\to\mathbb{R}^3,\quad \vec{g}(s,t)\stackrel{def}{=}(st,s+t,3s^2t).$$

Show that $f\circ\vec{g}$ is differentiable at $(1,-2)$, and compute $D(f\circ\vec{g})(1,-2)$.

Note that f is C^1 throughout $\mathbb{R}^3$, and therefore is differentiable on $\mathbb{R}^3$ with

$$Df(x,y,z)=\left[1+\tfrac{2x}{x^2+z^2+1}\quad 1\quad 1+\tfrac{2z}{x^2+z^2+1}\right].$$

Because the component functions st, $s+t$, and $3s^2t$ of $\vec{g}$ are C^1 throughout $\mathbb{R}^2$, $\vec{g}$ is differentiable throughout $\mathbb{R}^2$ (and therefore at $(1,-2)$) and

$$Dg(s,t)=\begin{bmatrix}t & s\\ 1 & 1\\ 6st & 3s^2\end{bmatrix}.$$

Therefore $f\circ\vec{g}$ is differentiable at $(1,-2)$ with

$$\begin{aligned}D(f\circ\vec{g})(1,-2)&=Df(\vec{g}(1,-2))D\vec{g}(1,-2)\\&=Df(-2,-1,-6)D\vec{g}(1,-2)\\&=\left[\tfrac{37}{41}\quad 1\quad \tfrac{29}{41}\right]\begin{bmatrix}-2 & 1\\ 1 & 1\\ -12 & 3\end{bmatrix}\\&=\left[-\tfrac{381}{41}\quad \tfrac{165}{41}\right].\end{aligned}$$

Of course, we could also have first simplified $f\circ\vec{g}$ to get

$$f\circ\vec{g}(s,t)=f(st,s+t,3s^2t)=st+s+t+3s^2t+\ln(s^2t^2+9s^4t^2+1),$$

and then computed

$$(f\circ\vec{g})_s(s,t)=t+1+6st+\frac{2st^2+36s^3t^2}{s^2t^2+9s^4t^2+1}$$

and

$$(f\circ\vec{g})_t(s,t)=s+1+3s^2+\frac{2s^2t+18s^4t}{s^2t^2+9s^4t^2+1},$$

whence we see that $f\circ\vec{g}$ is C^1 throughout $\mathbb{R}^2$, and therefore differentiable at $(1,-2)$ with

$$D(f\circ\vec{g})(1,-2)=\left[(f\circ\vec{g})_s(1,-2)\quad (f\circ\vec{g})_t(1,-2)\right]=\left[-\tfrac{381}{41}\quad \tfrac{165}{41}\right],$$

which agrees with what we obtained above!

Lecture 26: More Chain Rule

Learning Objectives:

- Compute derivatives and partial derivatives of compositions of differentiable functions.

Example 77. The chain rule can provide an easy way to establish various differentiation rules. For example, suppose that $f : \mathbb{R}^n \to \mathbb{R}$ and $g : \mathbb{R}^n \to \mathbb{R}$ are both differentiable at $\vec{a} \in \mathbb{R}^n$ and that $g(\vec{a}) \neq 0$. Then we can view $D(\frac{f}{g})(\vec{a})$ as the derivative of a composition of functions $\frac{f(\vec{x})}{g(\vec{x})} = Q(\vec{P}(\vec{x}))$, where $\vec{P} : \mathbb{R}^n \to \mathbb{R}^2$ is $\vec{P}(\vec{x}) = (f(\vec{x}), g(\vec{x}))$ and $Q : \{(s,t) \ : t \neq 0\} \to \mathbb{R}$ is $Q(s,t) = \frac{s}{t}$. Then $\vec{P}$ is differentiable at $\vec{a}$ (because its component functions f and g are), and Q is differentiable at $\vec{P}(\vec{a})$ (because $g(\vec{a}) \neq 0$), and therefore the chain rule gives (writing $(s,t) = \vec{P}(\vec{a}) = (f(\vec{a}), g(\vec{a}))$)

$$\begin{aligned}
D\Big(\frac{f}{g}\Big)(\vec{a}) &= DQ(\vec{P}(\vec{a}))D\vec{P}(\vec{a}) \\
&= \begin{bmatrix} \frac{1}{t} & -\frac{s}{t^2} \end{bmatrix} \begin{bmatrix} f_{x_1}(\vec{a}) & \cdots & f_{x_n}(\vec{a}) \\ g_{x_1}(\vec{a}) & \cdots & g_{x_n}(\vec{a}) \end{bmatrix} \\
&= \begin{bmatrix} \frac{1}{g(\vec{a})} & -\frac{f(\vec{a})}{g(\vec{a})^2} \end{bmatrix} \begin{bmatrix} f_{x_1}(\vec{a}) & \cdots & f_{x_n}(\vec{a}) \\ g_{x_1}(\vec{a}) & \cdots & g_{x_n}(\vec{a}) \end{bmatrix} \\
&= \frac{1}{g(\vec{a})^2} \begin{bmatrix} g(\vec{a}) & -f(\vec{a}) \end{bmatrix} \begin{bmatrix} f_{x_1}(\vec{a}) & \cdots & f_{x_n}(\vec{a}) \\ g_{x_1}(\vec{a}) & \cdots & g_{x_n}(\vec{a}) \end{bmatrix} \\
&= \frac{1}{g(\vec{a})^2} \Big(g(\vec{a}) \begin{bmatrix} f_{x_1}(\vec{a}) & \cdots & f_{x_n}(\vec{a}) \end{bmatrix} - f(\vec{a}) \begin{bmatrix} g_{x_1}(\vec{a}) & \cdots & g_{x_n}(\vec{a}) \end{bmatrix} \Big) \\
&= \frac{g(\vec{a})Df(\vec{a}) - f(\vec{a})Dg(\vec{a})}{g(\vec{x})^2},
\end{aligned}$$

which is the quotient rule for scalar-valued functions.

In practice, we are usually more interested in computing partial derivatives of compositions of functions than computing the full derivative all at once. To facilitate these computations, let's make a few observations.

Remark 71. To motivate the general case, suppose that $f : \mathbb{R}^3 \to \mathbb{R}$ is a scalar-valued differentiable function, and that $\vec{g} : \mathbb{R}^2 \to \mathbb{R}^3$, $\vec{g}(s,t) = (x(s,t), y(s,t), z(s,t))$ is differentiable. Then the composition $f \circ \vec{g} : \mathbb{R}^2 \to \mathbb{R}$ can be written as $f \circ \vec{g}(s,t) = f(x(s,t), y(s,t), z(s,t))$. Note that the partial derivatives $\frac{\partial(f \circ \vec{g})}{\partial s}$ and $\frac{\partial(f \circ \vec{g})}{\partial t}$ are exactly the entries in the matrix

$$D(f \circ \vec{g})(s,t) = \begin{bmatrix} \frac{\partial(f \circ \vec{g})}{\partial s}(s,t) & \frac{\partial(f \circ \vec{g})}{\partial t}(s,t) \end{bmatrix}.$$

On the other hand, we know that (simplifying notation by omitting the $\vec{g}(s,t)$ and (s,t) used to indicate where functions are evaluated) $D(f \circ \vec{g})(s,t)$ is given by

$$Df(\vec{g}(s,t))D\vec{g}(s,t) = \begin{bmatrix} \frac{\partial f}{\partial x} & \frac{\partial f}{\partial y} & \frac{\partial f}{\partial z} \end{bmatrix} \begin{bmatrix} \frac{\partial x}{\partial s} & \frac{\partial x}{\partial t} \\ \frac{\partial y}{\partial s} & \frac{\partial y}{\partial t} \\ \frac{\partial z}{\partial s} & \frac{\partial z}{\partial t} \end{bmatrix} = \begin{bmatrix} \frac{\partial f}{\partial x}\frac{\partial x}{\partial s} + \frac{\partial f}{\partial y}\frac{\partial y}{\partial s} + \frac{\partial f}{\partial z}\frac{\partial z}{\partial s} & \frac{\partial f}{\partial x}\frac{\partial x}{\partial t} + \frac{\partial f}{\partial y}\frac{\partial y}{\partial t} + \frac{\partial f}{\partial z}\frac{\partial z}{\partial t} \end{bmatrix}$$

Therefore, we have

$$\frac{\partial(f\circ\vec{g})}{\partial s}=\frac{\partial f}{\partial x}\frac{\partial x}{\partial s}+\frac{\partial f}{\partial y}\frac{\partial y}{\partial s}+\frac{\partial f}{\partial z}\frac{\partial z}{\partial s}\quad\text{and}\quad\frac{\partial(f\circ\vec{g})}{\partial t}=\frac{\partial f}{\partial x}\frac{\partial x}{\partial t}+\frac{\partial f}{\partial y}\frac{\partial y}{\partial t}+\frac{\partial f}{\partial z}\frac{\partial z}{\partial t},$$

where each partial derivative of f is evaluated at $\vec{g}(s,t)$, and where each partial derivative of x,y,z are evaluated at (s,t).

In general, if $f:\mathbb{R}^n\to\mathbb{R}$ is a differentiable scalar-valued function of $x_1,\dots,x_n$, and if $\vec{g}:\mathbb{R}^p\to\mathbb{R}^n$, $\vec{g}(\vec{u})=(x_1(\vec{u}),\dots,x_n(\vec{u}))$ is differentiable, then

$$\frac{\partial(f\circ\vec{g})}{\partial u_i}(\vec{u})=\sum_{j=1}^{n}\frac{\partial f}{\partial x_j}(\vec{g}(\vec{u}))\frac{\partial x_j}{\partial u_i}(\vec{u}).$$

Note that (taking $p=1$) this is a generalization of the formula $\frac{df}{dt}=\frac{df}{dx}\frac{dx}{dt}$ from single-variable calculus, except now that $\vec{x}\in\mathbb{R}^n$ we have that $\frac{\partial f}{\partial t}$ is a sum of terms of the form $\frac{\partial f}{\partial x_j}\frac{\partial x_j}{\partial t}$, one for each coordinate of $\vec{x}$.

Example 78. As I run up a spiral staircase my position after t seconds is given by

$$\vec{r}(t)=(3\cos(t^2/\pi),3\sin(t^2/\pi),\sqrt{t}).$$

The staircase is full of gnats, with the density (in kilograms per cubic meter) is given by

$$G(x,y,z)=(1+\cos(xy))\arctan(z).$$

At what rate does the density of gnats (in kilograms per cubic meter) change from my perspective as I pass through the point $(-3,0,\sqrt{\pi})$?

The density of gnats I experience as I run up the stairs is given as a function of t by $G\circ\vec{r}(t)$ Since $\vec{r}(\pi)=(-3,0,\sqrt{\pi})$, I need to compute $\frac{d(G\circ\vec{r})}{dt}(\pi)$. Note that $G\circ\vec{r}(t)$ is a scalar-valued function of a single variable, and therefore I could (in principle) simplify the expression for $G\circ\vec{r}(t)$, and then compute the derivative using single-variable calculus techniques.

Instead, we apply the previous remark[14] to write (using $(x,y,z)=\vec{r}(t)$)

$$\begin{aligned}\frac{d(G\circ\vec{r})}{dt}&=\frac{\partial G}{\partial x}\frac{dx}{dt}+\frac{\partial G}{\partial y}\frac{dy}{dt}+\frac{\partial G}{\partial z}\frac{dz}{dt}\\&=-y\sin(xy)\arctan(z)\Big(-\frac{6t}{\pi}\sin\Big(\frac{t^2}{\pi}\Big)\Big)-x\sin(xy)\arctan(z)\Big(\frac{6t}{\pi}\cos\Big(\frac{t^2}{\pi}\Big)\Big)\\&\quad+\frac{1+\cos(xy)}{1+z^2}\frac{1}{2\sqrt{t}}.\end{aligned}$$

Therefore, evaluating at $t=\pi$ (so that $(x,y,z)=\vec{r}(\pi)=(-3,0,\sqrt{\pi})$) we have

$$\frac{d(G\circ\vec{r})}{dt}(\pi)=0\cdot0+0\cdot(-6)+\frac{2}{1+\pi^2}\cdot\frac{1}{2\sqrt{\pi}}=\frac{1}{\sqrt{\pi}(1+\pi^2)}.$$

Therefore, from my perspective, the density of gnats is changing at a rate of $\frac{1}{\sqrt{\pi}(1+\pi)}$ kilograms per cubic meter per second when I pass through $(-3,0,\sqrt{\pi})$. Gross.

[14]Here are are writing $\frac{dx}{dt}$ instead of $\frac{\partial x}{\partial t}$ because x is actually a scalar-valued function of a single variable, and therefore $\frac{dx}{dt}$ is actually the derivative of x, not a partial derivative.

Lecture 27: Directional Derivatives

Learning Objectives:

- Define and interpret the directional derivative of a scalar-valued function.
- Use the gradient vector to compute directional derivatives.

So far we have generalized the definition of the derivative from single-variable calculus to capture both partial derivatives of scalar-valued functions and differentiability for vector-valued functions. In particular, we motivated the idea of differentiability from our desire for a "suitable affine approximation" for a function. This was one interpretation of the single-variable derivative that generalized to functions of several variables, and with a bit of work we can generalize other interpretations of the single-variable derivative as well. Today we generalize our interpretation of the single-variable derivative as capturing "rate of change" of the function as its input increases. Because "as its input increases" is really a statement about the direction in which we are measuring the rate of change, we will build a notion of "directional derivative" with this in mind.

Definition 48. Let $\Omega \subseteq \mathbb{R}^n$ be open, let $\vec{a} \in \Omega$, and let $f : \Omega \to \mathbb{R}$. For a unit vector $\vec{u} \in \mathbb{R}^n$, we define the **directional derivative of f in the direction $\vec{u}$ at $\vec{a}$** to be

$$D_{\vec{u}} f(\vec{a}) \stackrel{def}{=} \lim_{h \to 0} \frac{f(\vec{a} + h\vec{u}) - f(\vec{a})}{h} = \lim_{h \to 0} \frac{f(a_1 + hu_1, \ldots, a_n + hu_n) - f(a_1, \ldots, a_n)}{h},$$

provided that the limit exists.

Remark 72. $D_{\vec{u}} f(\vec{a})$ (when it exists) can be interpreted as the rate of change of f at $\vec{a}$ in the direction of $\vec{u}$. We might expect, therefore, that $D_{-\vec{u}} f(\vec{a}) = -D_{\vec{u}} f(\vec{a})$, since if f has "slope M" at $\vec{a}$ in the direction of $\vec{u}$, then f should have "slope $-M$" at $\vec{a}$ in the direction of $-\vec{u}$. Or, even more intuitively, if you are standing on a hillside and walking forward will cause you to more uphill at a certain grade, the walking backward will cause you to move downhill at that same grade.

Remark 73. There is no harm or difficulty in defining $D_{\vec{u}} \vec{f}(\vec{a})$ when $\vec{f} : \Omega \to \mathbb{R}^m$ is vector valued, other than losing the interpretation of the directional derivative as the "rate of change" of the function in the direction of $\vec{u}$.

Remark 74. In the previous definition, we specify the direction of the directional derivative using a unit vector $\vec{u}$ to avoid having the definition of the directional derivative depend on the length of the vector used. To see why we need to make this restriction, suppose that $\vec{u}$ is a unit vector and $D_{\vec{u}} f(\vec{a}) = M$. Then for $\lambda \neq 0$ we have

$$\lim_{h \to 0} \frac{f(\vec{a} + h(\lambda\vec{u})) - f(\vec{a})}{h} = \lambda \lim_{h \to 0} \frac{f(\vec{a} + (h\lambda)\vec{u}) - f(\vec{a})}{h\lambda} = \lambda \lim_{t \to 0} \frac{f(\vec{a} + t\vec{u}) - f(\vec{a})}{t} = \lambda M.$$

Therefore scaling the vector $\vec{u}$ scales the resulting limit by the same amount.

Remark 75. Note that $\frac{\partial f}{\partial x_i}(\vec{a}) = D_{\vec{e}_i} f(\vec{a})$. Therefore the partial derivative of f and $\vec{a}$ with respect to $\vec{x}_i$ is exactly the directional derivative of f in the direction of $\vec{e}_i$ at $\vec{a}$.

Remark 76. You might be tempted to believe that if $D_{\vec{u}}f(\vec{a})$ exists for *every* unit vector $\vec{u}$, then f must be differentiable at $\vec{a}$. For a counterexample, consider the function $f : \mathbb{R}^2 \to \mathbb{R}$ given by

$$f(x,y) = \begin{cases} \dfrac{x^3}{x^2+y^2} & \text{if } (x,y) \neq (0,0), \\ 0 & \text{if } (x,y) = (0,0). \end{cases}$$

As an exercise, you can verify that for every unit vector $\vec{u} = \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} \in \mathbb{R}^2$, $D_{\vec{u}}f(0,0)$ exists and $D_{\vec{u}}f(0,0) = u_1^3$. You can also check that f is not differentiable at $(0,0)$.

To avoid proliferation of ideas, it would be prudent to try to understand directional derivatives $D_{\vec{u}}f(\vec{a})$ in terms of the ordinary derivative $Df(\vec{a})$ of f at $\vec{a}$. The following definition helps us accomplish this while opening up some surprising geometric interpretations of directional derivatives.

Definition 49. Let $\Omega \subseteq \mathbb{R}^n$ be open, let $\vec{a} \in \Omega$, and let $f : \Omega \to \mathbb{R}$. If f is differentiable at $\vec{a}$, then we define the **gradient vector of f at $\vec{a}$**, $\nabla f(\vec{a})$, to be

$$\nabla f(\vec{a}) \stackrel{def}{=} (Df(\vec{a}))^T = \begin{bmatrix} f_{x_1}(\vec{a}) \\ \vdots \\ f_{x_n}(\vec{a}) \end{bmatrix}.$$

Remark 77. The vector $\nabla f(\vec{a})$, while it encodes the same information that is encoded by the derivative $Df(\vec{a})$, has the additional geometric interpretation as a vector in $\mathbb{R}^n$. This geometric interpretation will be crucial for developing a full understanding of directional derivatives.

Theorem 35 (Directional Derivatives)**.** Let $\Omega \subseteq \mathbb{R}^n$ be open, let $\vec{a} \in \Omega$, and let $f : \Omega \to \mathbb{R}$ be differentiable at $\vec{a}$. For each unit vector $\vec{u} \in \mathbb{R}^n$, $D_{\vec{u}}f(\vec{a})$ exists and

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

Proof. Let $\vec{u} \in \mathbb{R}^n$ be a unit vector, and define[15] $\vec{r} : \mathbb{R} \to \mathbb{R}^n$ by $\vec{r}(t) = \vec{a} + t\vec{u}$. Then $\vec{r}(t) = \vec{a} + t\vec{u} = \vec{u}[t] + \vec{a}$ is affine, so it is differentiable with $D\vec{r}(t) = \vec{u}$ for each $t \in \mathbb{R}$. Because f is differentiable at $\vec{a} = \vec{r}(0)$, the Chain Rule implies that $f \circ \vec{r}$ is differentiable at $t = 0$ and

$$(f \circ \vec{r})'(0) = Df(\vec{a})\vec{u} = (Df(\vec{a}))^T \cdot \vec{u} = \nabla f(\vec{a}) \cdot \vec{u}.$$

Therefore the definition of the single-variable derivative gives

$$\nabla f(\vec{a}) \cdot \vec{u} = (f \circ \vec{r})'(0) = \lim_{h \to 0} \frac{f(\vec{r}(h)) - f(\vec{r}(0))}{h} = \lim_{h \to 0} \frac{f(\vec{a} + h\vec{u}) - f(\vec{a})}{h} = D_{\vec{u}}f(\vec{a}).$$

□

[15] Technically we may only be able to define $\vec{r}(t)$ for t in a small interval centered at 0 to ensure that $\vec{r}(t) \in X$, but let's not worry too much about this technical detail.

Example 79. Compute the directional derivative of $f(x,y) = \sqrt{x+y^2-3}$ in the direction of $\vec{v} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$ at $(3,-2)$.

Note first that $\vec{v}$ is not a unit vector, so we first need to find a unit vector $\vec{u}$ in the same direction as $\vec{v}$. Therefore we can take $\vec{u} = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 1 \end{bmatrix}$.

To compute $D_{\vec{u}} f(3,-2)$, we first need to compute the gradient

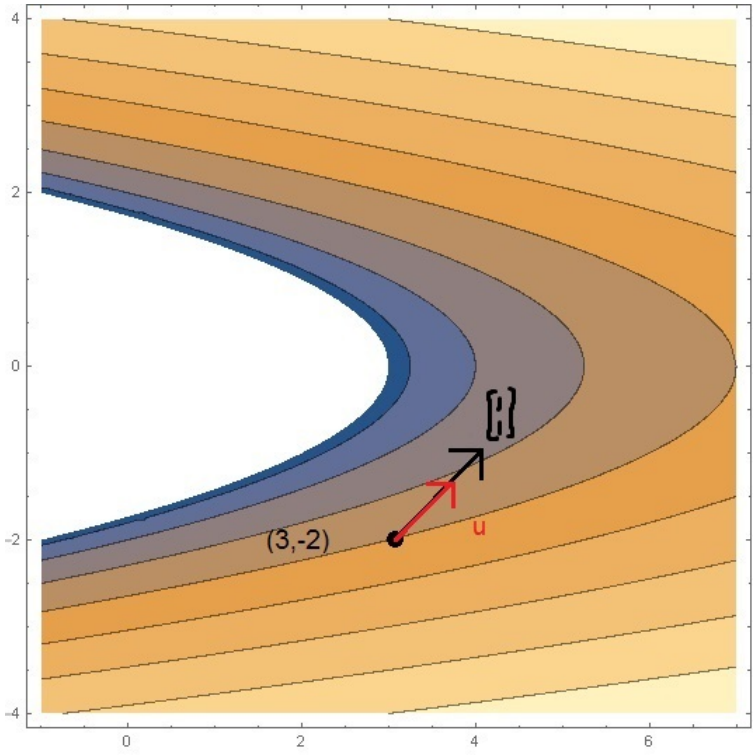
$$\nabla f(x,y) = \begin{bmatrix} f_x(x,y) \\ f_y(x,y) \end{bmatrix} = \begin{bmatrix} \frac{1}{2\sqrt{x+y^2-3}} \\ \frac{y}{\sqrt{x+y^2-3}} \end{bmatrix},$$

so that $\nabla f(3,-2) = \begin{bmatrix} f_x(3,-2) \\ f_y(3,-2) \end{bmatrix} = \begin{bmatrix} \frac{1}{4} \\ -1 \end{bmatrix}$. Therefore

$$D_{\vec{u}} f(3,-2) = \nabla f(3,-2) \cdot \vec{u} = \begin{bmatrix} \frac{1}{4} \\ -1 \end{bmatrix} \cdot \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix} = \frac{1}{4\sqrt{2}} - \frac{1}{\sqrt{2}} = -\frac{3}{4\sqrt{2}}.$$

In other words, the rate of change of $f(x,y) = \sqrt{x+y^2-3}$ in the direction of $\begin{bmatrix} 1 \\ 1 \end{bmatrix}$ at $(3,-2)$ is $-\frac{3}{4\sqrt{2}}$.

To visualize this, consider the following plot of level curves for $f(x,y) = \sqrt{x+y^2-3}$. If we start at $(3,-2)$ and move in the direction of $\begin{bmatrix} 1 \\ 1 \end{bmatrix}$, then we expect to go 'downhill' (so that $D_{\begin{bmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{bmatrix}} f(3,-2) < 0$). This was borne out by our computation above.



Example 80. The Directional Derivatives Theorem also allows us to easily prove that if f is differentiable at $\vec{a}$ and if $\vec{u}$ is a unit vector, then (as expected)

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

Lecture 28: Gradient Vectors

Learning Objectives:

- Explore the geometric significance of the gradient vector.
- Exploit the geometric properties of the gradient vector to solve geometric problems.

Combining the Directional Derivatives Theorem with our geometric understanding of the dot product immediately gives a strong geometric interpretation of the gradient vector.

Theorem 36. Let $\Omega \subseteq \mathbb{R}^n$ be open, let $\vec{a} \in \Omega$, let $f : \Omega \to \mathbb{R}$, and suppose that f is differentiable at $\vec{a}$. If $\nabla f(\vec{a}) \neq \vec{0}$, then

(i) $\nabla f(\vec{a})$ points in the direction $\vec{u}_{max}$ that maximizes $D_{\vec{u}} f(\vec{a})$ (i.e. the direction of steepest increase of f at $\vec{a}$), and $D_{\vec{u}_{max}} f(\vec{a}) = \|\nabla f(\vec{a})\|$.

(ii) $-\nabla f(\vec{a})$ points in the direction $\vec{u}_{min}$ that minimizes $D_{\vec{u}} f(\vec{a})$ (i.e. the direction of steepest decrease of f at $\vec{a}$), and $D_{\vec{u}_{min}} f(\vec{a}) = -\|\nabla f(\vec{a})\|$.

Proof. For any unit vector $\vec{u}$, let $\theta \in [0, \pi]$ be the angle between $\nabla f(\vec{a})$ and $\vec{u}$. Then we have

$$D_{\vec{u}} f(\vec{a}) = \nabla f(\vec{a}) \cdot \vec{u} = \|\nabla f(\vec{a})\| \underbrace{\|\vec{u}\|}_{=1} \cos(\theta) = \|\nabla f(\vec{a})\| \cos(\theta).$$

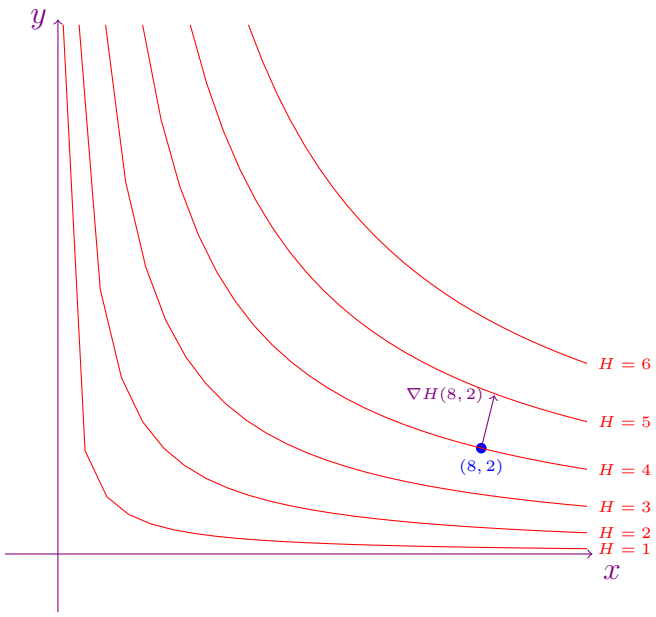
$D_{\vec{u}} f(\vec{a})$ is largest when $\cos(\theta) = 1$ (or rather when $\theta = 0$), which occurs exactly when $\vec{u}$ points in the same direction as $\nabla f(\vec{a})$. Therefore $\nabla f(\vec{a})$ points in the direction $\vec{u}_{max}$, and $D_{\vec{u}_{max}} f(\vec{a}) = \|\nabla f(\vec{a})\| \cos(0) = \|\nabla f(\vec{a})\|$. The statement about the direction of $-\nabla f(\vec{a})$ follows from setting $\theta = \pi$. □

Example 81. A hiker is on a very oddly-shaped mountain, whose height at x°N,y°E is given by $H(x, y) = \sqrt{xy}$. When the hiker is at longitude-latitude $(8, 2)$, then in which direction should the hiker walk to climb the fastest?

This function is C^1 near $(8, 2)$ since the partial derivatives $f_x(x, y) = \frac{\sqrt{y}}{2\sqrt{x}}$ and $f_y(x, y) = \frac{\sqrt{x}}{2\sqrt{y}}$ exist and are continuous near $(8, 2)$. Therefore, f is differentiable and we can compute

$$\nabla f(8, 2) = \begin{bmatrix} f_x(8, 2) \\ f_y(8, 2) \end{bmatrix} = \begin{bmatrix} \frac{1}{4} \\ 1 \end{bmatrix}.$$

Because the gradient points in the direction of steepest ascent, the hiker should walk in the direction of $\begin{bmatrix} 1/4 \\ 1 \end{bmatrix}$ to climb as quickly as possible.

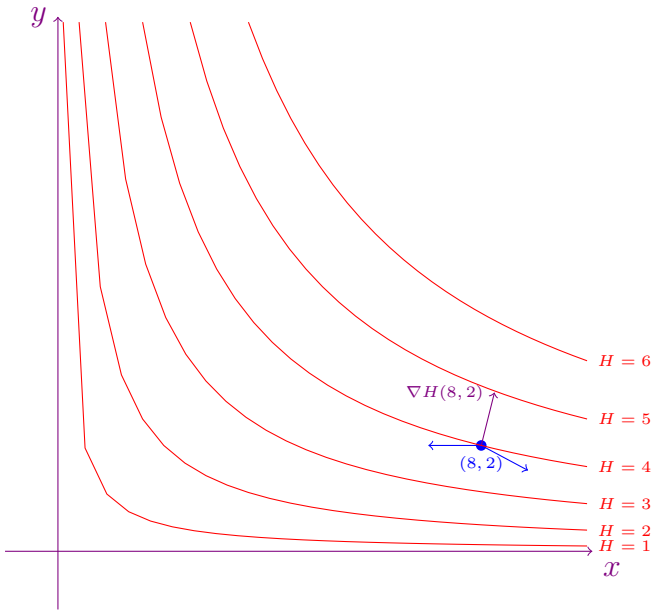


Example 82. The hiker from the last problem is starting to become short of breath, and wants to descend the hill. To avoid undue stress, the hiker should walk downhill on a grade no steeper than 25%. In which direction(s) should the hiker walk?

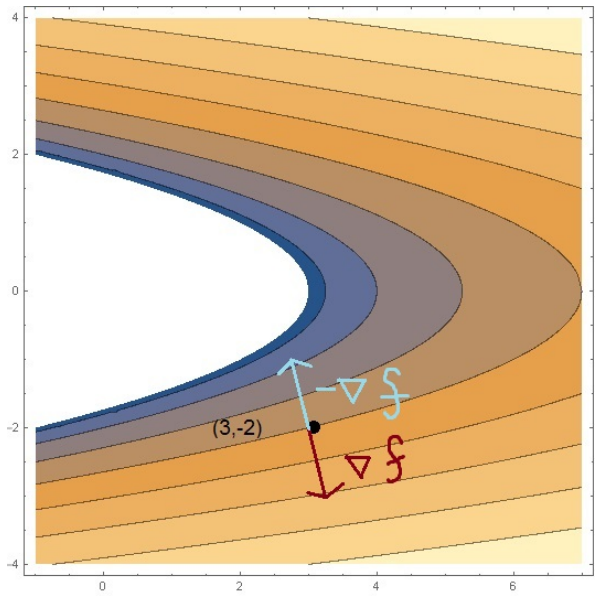
Here we need to find a unit vector $\vec{u}$ with $D_{\vec{u}}f(8,2) = -\frac{1}{4}$. That is, we are looking for a vector of the form $\left[\begin{smallmatrix} a \\ b \end{smallmatrix}\right]$ with

$$a^2 + b^2 = 1, \quad \text{and} \quad -\frac{1}{4} = \nabla f(8,2) \cdot \begin{bmatrix} a \\ b \end{bmatrix} = \frac{a}{4} + b.$$

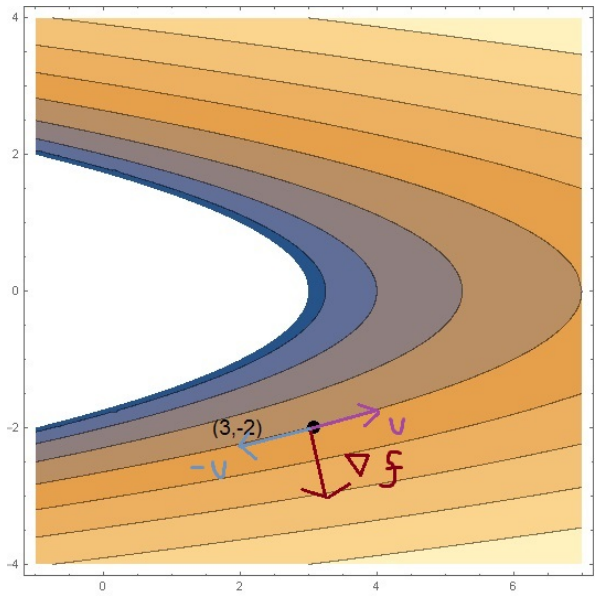
This is a nonlinear set of equations, and we can solve for a and b by writing $b = -\frac{1}{4} - \frac{a}{4}$, so that $a^2 + \frac{(a+1)^2}{16} = 1$, or rather $0 = 17a^2 + 2a - 15 = (a+1)(17a - 15)$. This tells us that if the hiker walks downhill in the direction of either $\left[\begin{smallmatrix} -1 \\ 0 \end{smallmatrix}\right]$ or $\begin{bmatrix} 15/17 \\ -8/17 \end{bmatrix}$, then the hiker will (initially) descend at a grade no steeper than 25%.



Example 83. To see this result intuitively, let's go back to the example of $f(x,y) = \sqrt{x+y^2-3}$. We computed that $\nabla f(3,-2) = \begin{bmatrix} \frac{1}{4} \\ -1 \end{bmatrix}$. Drawing $\nabla f(3,-2)$ and $-\nabla f(3,-2)$ in the contour plot illustrates the directions (at $(3,-2)$) in which f increases and decreases the fastest (respectively).



Note also that the vectors $\begin{bmatrix} 1 \\ \frac{1}{4} \end{bmatrix}$ and $-\begin{bmatrix} 1 \\ \frac{1}{4} \end{bmatrix}$ are orthogonal to $\nabla f(3,-2)$, and therefore the directional derivative of f in the directions $\pm\vec{u}$ of these two vectors is $D_{\pm\vec{u}} f(3,-2) = \nabla f(3,-2) \cdot (\pm\vec{u}) = 0$.



Something in the previous example requires further comment. The vectors $\vec{u} = \begin{bmatrix} 1 \\ \frac{1}{4} \end{bmatrix}$ and $-\vec{u} = \begin{bmatrix} -1 \\ -\frac{1}{4} \end{bmatrix}$ that were orthogonal to $\nabla f(3,-2)$ appear also to be *tangent* to the level curve of $f(x,y)$ passing through the point $(3,-2)$, so that $\nabla f(3,-2)$ is **normal** to (i.e. perpendicular to) the level curve of $f(x,y)$ passing through the point $(3,-2)$.

This is no coincidence. To make this precise, we need to introduce a notion of what is meant by saying that a vector is "normal to a level set". To this end, here is a definition.

Definition 50. Let $S \subseteq \mathbb{R}^n$, and let $\vec{a} \in S$. Then we say that a vector $\vec{u} \in \mathbb{R}^n$ is **tangent** to S at $\vec{a}$ if there is a differentiable path $\vec{r} : \mathbb{R} \to \mathbb{R}^n$ with $\vec{r}(t) \in S$ for all t, and where $\vec{r}(0) = \vec{a}$ and $D\vec{r}(0) = \vec{u}$. We say that a vector $\vec{v}$ is **normal** to S at $\vec{a}$ if $\vec{v} \cdot \vec{u} = 0$ for every vector $\vec{u}$ that is tangent to S at $\vec{a}$.

The above definition is intended to capture the idea that a vector $\vec{u}$ is tangent to S at $\vec{a}$ only if $\vec{u}$ "lies flat against" S at $\vec{a}$. With this definition in hand, the following theorem is a breeze.

Theorem 37. Let $\Omega \subseteq \mathbb{R}^n$ be open, let $\vec{a} \in \Omega$, let $f : \Omega \to \mathbb{R}$, and suppose that f is differentiable at $\vec{a}$. Then $\nabla f(\vec{a})$ is normal to the level set $S = \{\vec{x} \in \Omega \ : \ f(\vec{x}) = f(\vec{a})\}$ of f passing through $\vec{a}$.

Proof. Suppose that $\vec{u} \in \mathbb{R}^n$ is tangent to S at $\vec{a}$. Let $\vec{r} : \mathbb{R} \to \mathbb{R}^n$ be a differentiable path with $\vec{r}(t) \in S$ for all t, and with $\vec{r}(0) = \vec{a}$ and $D\vec{r}(0) = \vec{u}$. Then the (single-variable, scalar-valued) function $(f \circ \vec{r})(t) \equiv f(\vec{a})$ is constant, so that $(f \circ \vec{r})'(0) = 0$. By the Chain Rule, we therefore have

$$0 = (f \circ \vec{r})'(0) = Df(\vec{r}(0))D\vec{r}(0) = Df(\vec{a})\vec{u} = \nabla f(\vec{a}) \cdot \vec{u},$$

which completes the proof. □

Remark 78. One remarkable fact—the proof of which relies on the inverse function theorem and *definitely* belongs in a course in real analysis—is that if $f : \mathbb{R}^n \to \mathbb{R}$ is C^1 on an open set containing $\vec{a}$ with $\nabla f(\vec{a}) \neq \vec{0}$, then $\vec{u} \in \mathbb{R}^n$ is tangent to the level set S of f containing $\vec{a}$ at $\vec{a}$ *if, and only if,* $\nabla f(\vec{a}) \cdot \vec{u} = 0$. In particular, if we treat $\vec{a}$ as the origin of the space V of vectors tangent to S at $\vec{a}$, then $V = (\text{span}(\nabla f(\vec{a})))^{\perp}$, so that $\dim(V) = n - 1$. When $n = 2$ we conclude that V forms the line tangent to S at $\vec{a}$, and when $n = 3$ we conclude that V forms the plane tangent to S at $\vec{a}$.

Example 84. Find all points on the hyperbolic paraboloid S described by $z = 2xy$ where the normal line (the line perpendicular to the surface) is parallel to the vector $\begin{bmatrix} 1 \\ -1 \\ 2 \end{bmatrix}$.

By rewriting the equation for S as $0 = 2xy - z$, we can treat it as the level surface (at level 0) of the function $F(x, y, z) = 2xy - z$. The normal line to S at a point (x, y, z) is parallel to $\nabla F(x, y, z) = \begin{bmatrix} 2y \\ 2x \\ -1 \end{bmatrix}$.

Therefore the normal line to S at (x, y, z) is parallel to $\begin{bmatrix} 1 \\ -1 \\ 2 \end{bmatrix}$ exactly when there is a nonzero $c \in \mathbb{R}$ with

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

We therefore must have $-1 = 2c$ (so that $c = -\frac{1}{2}$), and therefore $2y = c = -\frac{1}{2}$ and $2x = -c = \frac{1}{2}$. Solving yields $x = \frac{1}{4}$ and $y = -\frac{1}{4}$, and therefore $z = 2xy = -\frac{2}{16} = -\frac{1}{8}$.

Therefore $(\frac{1}{4}, -\frac{1}{4}, -\frac{1}{4})$ is the only point on the hyperbolic paraboloid S at which the normal line is parallel to the vector $\begin{bmatrix} 1 \\ -1 \\ 2 \end{bmatrix}$.