Math 291-2: Intensive Linear Algebra & Multivariable Calculus

Northwestern University, Lecture Notes

Written by Santiago Cañez

These are notes which provide a basic summary of each lecture for Math 291-2, the second quarter of "MENU: Intensive Linear Algebra & Multivariable Calculus", taught by the author at Northwestern University. The books used as references are the 5th edition of *Linear Algebra with Applications* by Bretscher and the 4th edition of *Vector Calculus* by Colley. Watch out for typos! Comments and suggestions are welcome.

These notes will focus on material covered in Math 291 which is not normally covered in Math 290, and should thus be used in conjunction with my notes for Math 290, which are available at http://www.math.northwestern.edu/∼scanez/courses/290/notes.php.

Contents

Lecture 1: Dot Products and Transposes

Dot product. The *dot product* of two vectors

$$\mathbf{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \text{ and } \mathbf{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$$

in $\mathbb{R}^n$ is defined to be

$$\mathbf{x} \cdot \mathbf{y} = x_1 y_1 + \cdots + x_n y_n.$$

This simple algebraic expression turns out to encode important geometric properties, as we'll see. For now, note that

$$\mathbf{x} \cdot \mathbf{x} = x_1^2 + \cdots + x_n^2$$

is never negative, so it makes sense to take the square root. We define the *norm* (another word for length) of $\mathbf{x}$ to be

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

Of course, in 2 and 3 dimensions this gives the usual notion of length.

Orthogonality. The most basic geometric fact about the dot product is that is fully determines whether or not two vectors are *orthogonal*, which is just another word for perpendicular. Indeed, first consider the 2-dimensional case. In this case $\mathbf{x} \cdot \mathbf{y} = 0$ when

$$x_1 y_1 + x_2 y_2 = 0, \text{ or } x_1 y_1 = -x_2 y_2.$$

Assuming for now that x_1, y_1, x_2, y_2 are nonzero, this final equality is the same as

$$\frac{x_1}{x_2} = -\frac{y_2}{y_1}.$$

The fraction $\frac{x_2}{x_1}$ is the slope of the line spanned by $\mathbf{x}$ and $\frac{y_2}{y_1}$ is the slope of the line spanned by $\mathbf{y}$, so this says that the slopes of these two lines are negative reciprocals of one another, which means the lines are perpendicular. Hence $\mathbf{x} \cdot \mathbf{y} = 0$ does mean that $\mathbf{x}$ and $\mathbf{y}$ are perpendicular in the two dimensional case.

In three dimensions the same reasoning doesn't work since "slope" is harder to define, but instead we can use the following fact, which is essentially the *Pythagorean Theorem*: $\mathbf{x}$ and $\mathbf{y}$ form the non-hypothenuse sides of a right-triangle if and only if $\|\mathbf{x}\|^2 + \|\mathbf{y}\|^2 = \|\mathbf{x} - \mathbf{y}\|^2$. Indeed, with $\mathbf{x}$ and $\mathbf{y}$ as sides of a triangle, $\mathbf{x} - \mathbf{y}$ describes the third side and the given equality is then the well known Pythagorean characterization of right triangles. Thus, $\mathbf{x}$ and $\mathbf{y}$ are perpendicular if and only if they form the non-hypothenuse sides of a right triangle, which is true if and only if

$$\|\mathbf{x}\|^2 + \|\mathbf{y}\|^2 = \|\mathbf{x} - \mathbf{y}\|^2,$$

which becomes

$$(x_1^2 + x_2^2 + x_3^2) + (y_1^2 + y_2^2 + y_3^2) = (x_1 - y_1)^2 + (x_2 - y_2)^2 + (x_3 - y_3)^2.$$

Expanding the terms on the right and simplifying gives

$$0 = -2(x_1 y_1 + x_2 y_2 + x_3 y_3),$$

so $\mathbf{x}$ and $\mathbf{y}$ are perpendicular if and only if $\mathbf{x} \cdot \mathbf{y} = x_1y_1 + x_2y_2 + x_3y_3 = 0$ as claimed.

In higher dimensions vectors can no longer be visualized, so we simply take $\mathbf{x} \cdot \mathbf{y} = 0$ as our *definition* of what it means for $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ to be orthogonal in general.

Dot product properties. The dot product as the following key properties:

- $(\mathbf{x} + \mathbf{y}) \cdot \mathbf{z} = \mathbf{x} \cdot \mathbf{z} + \mathbf{y} \cdot \mathbf{z}$, which we call "distributivity"
- $(a\mathbf{x} \cdot \mathbf{y}) = a(\mathbf{x} \cdot \mathbf{y})$
- $\mathbf{x} \cdot \mathbf{y} = \mathbf{y} \cdot \mathbf{x}$, which we call "symmetry"
- $\mathbf{x} \cdot \mathbf{x} \geq 0$ for all $\mathbf{x}$, and $\mathbf{x} \cdot \mathbf{x} = 0$ if and only if $\mathbf{x} = \mathbf{0}$.

All of these can be verified by working everything out use components and the definition of the dot product. For instance, $\mathbf{x} \cdot \mathbf{x} = x_1^2 + \cdots + x_n^2$ consists of all nonnegative terms, so this equals 0 if and only if each term x_i^2 is zero, which means that each x_i is zero, which gives the "$\mathbf{x} \cdot \mathbf{x} = 0$ if and only if $\mathbf{x} = \mathbf{0}$" claim. Phrased another way, since $\|\mathbf{x}\| = \sqrt{\mathbf{x} \cdot \mathbf{x}}$, this says that $\mathbf{0}$ is the only vector of length zero.

The first two properties can be phrased more succinctly as saying that for a fixed $\mathbf{x} \in \mathbb{R}^n$, the function $\mathbb{R}^n \to \mathbb{R}$ defined by

$$\mathbf{x} \mapsto \mathbf{x} \cdot \mathbf{z}$$

is a linear transformation. Indeed, the distributivity property says that this function preserves addition and the second property above says that it preserves scalar multiplication. We summarize this all by saying that the dot product is "linear in the first argument", meaning in the first location in which we plug in a vector. Combined with symmetry, this also implies that the dot product is linear in the second argument since

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

and similarly for the scalar multiplication property.

Orthogonal projections. We can now derive the formula for orthogonally projecting one vector onto another from last quarter. Say we want to orthogonally project $\mathbf{x} \in \mathbb{R}^n$ onto $\mathbf{v} \in \mathbb{R}^n$. The resulting vector $\text{proj}_{\mathbf{v}} \mathbf{x}$ is on the line spanned by $\mathbf{v}$, so it is a multiple of $\mathbf{v}$: $\text{proj}_{\mathbf{v}} \mathbf{x} = \lambda \mathbf{v}$ for some $\lambda \in \mathbb{R}$. The goal is to figure out what λ must be.

The orthogonal projection $\text{proj}_{\mathbf{v}} \mathbf{x} = \lambda \mathbf{v}$ is characterized by the property that the vector $\mathbf{x} - \text{proj}_{\mathbf{v}} \mathbf{x}$ should be orthogonal to $\mathbf{v}$, which means we require that

$$(\mathbf{x} - \lambda \mathbf{v}) \cdot \mathbf{v} = 0.$$

Using the linearity properties of the dot product, this gives

$$\mathbf{x} \cdot \mathbf{v} = \lambda(\mathbf{v} \cdot \mathbf{v}), \text{ so } \lambda = \frac{\mathbf{x} \cdot \mathbf{v}}{\mathbf{v} \cdot \mathbf{v}}.$$

Thus the orthogonal projection of $\mathbf{x}$ onto $\mathbf{v}$ is concretely given by

$$\text{proj}_{\mathbf{v}} \mathbf{x} = \left(\frac{\mathbf{x} \cdot \mathbf{v}}{\mathbf{v} \cdot \mathbf{v}} \right) \mathbf{v}.$$

Tranposes. Now with the notion of the dot product at hand we can give the *real* meaning behind the concept of the transpose of a matrix. Suppose that A is an $m \times n$ matrix. The claim is that A^T is the unique matrix satisfying the equality

$$A\mathbf{x} \cdot \mathbf{y} = \mathbf{x} \cdot A^T\mathbf{y} \text{ for all } \mathbf{x} \in \mathbb{R}^n, \mathbf{y} \in \mathbb{R}^m.$$

Thus, in any dot product expression where a vector is being multiplied by a matrix, taking the transpose allows us to move that matrix into the other argument. In particular, if A is symmetric, then

$$A\mathbf{x} \cdot \mathbf{y} = \mathbf{x} \cdot A\mathbf{y} \text{ for all } \mathbf{x}, \mathbf{y},$$

which as we'll see is the key property which explains why symmetric matrices have so many amazing properties.

To justify the property of transposes given by, note that if we express $\mathbf{x} = x_1\mathbf{e}_1 + \cdots + x_n\mathbf{e}_n$ and $\mathbf{y} = y_1\mathbf{e}_1 + \cdots + y_m\mathbf{e}_m$ in terms of the standard basis vectors, linearity of the dot product gives

$$\begin{aligned} A\mathbf{x} \cdot \mathbf{y} &= (x_1A\mathbf{e}_1 + \cdots + x_nA\mathbf{e}_n) \cdot (y_1\mathbf{e}_1 + \cdots + y_m\mathbf{e}_m) \\ &= x_1y_1(A\mathbf{e}_1 \cdot \mathbf{e}_1) + x_1y_2(A\mathbf{e}_1 \cdot \mathbf{e}_2) + (\text{other terms involving constants times } A\mathbf{e}_i \cdot \mathbf{e}_j), \end{aligned}$$

and the same is true of the right side $\mathbf{x} \cdot A^T\mathbf{y}$. This shows that if the given equality is true when $\mathbf{x}$ and $\mathbf{y}$ are standard basis vectors, it will be true for all vectors.

Thus we look at $A\mathbf{e}_i \cdot \mathbf{e}_j$ and $\mathbf{e}_i \cdot A^T\mathbf{e}_j$. Actually, for now forget A^T and suppose that B was *some* matrix which satisfied $A\mathbf{x} \cdot \mathbf{y} = \mathbf{x} \cdot B\mathbf{y}$. We will show that B *must* in fact be A^T, which not only gives the equality in question but also the uniqueness part of the statement. The key observation is that $A\mathbf{e}_i \cdot \mathbf{e}_j$ and $\mathbf{e}_i \cdot B\mathbf{e}_j$ are actually pretty simple values:

$$A\mathbf{e}_i \cdot \mathbf{e}_j = (i\text{-th column of } A) \cdot \mathbf{e}_j = j\text{-th entry in the } i\text{-th column of } A = a_{ji},$$

where we use the notation $m_{k\ell}$ for the entry in the k-th row and ℓ-th column of a matrix, and

$$\mathbf{e}_i \cdot B\mathbf{e}_j = \mathbf{e}_i \cdot (j\text{-th column of } B) = i\text{-th entry in the } j\text{-th column of } B = b_{ij}.$$

Hence in order for $A\mathbf{e}_i \cdot \mathbf{e}_j = \mathbf{e}_i \cdot B\mathbf{e}_j$ to be true, we must have $b_{ij} = a_{ji}$. But this says precisely that the i-th row of B is the i-th column of A and the j-th column of B is the j-th row of A, so B must be A^T as claimed.

Don't underestimate the importance of the equality $A\mathbf{x} \cdot \mathbf{y} = \mathbf{x} \cdot A^T\mathbf{y}$; this is really the only reason why we care about transposes at all.

Lecture 2: Orthogonal Bases

Warm-Up 1. Suppose that A is an $m \times n$ matrix and B is an $n \times k$ matrix, so that AB is defined. We justify the fact that $(AB)^T = B^TA^T$, which we used a few times last quarter. The "messy" way of doing this is to write out expressions for the entries of A and B and try to compute AB and B^TA^T to see that $(AB)^T = B^TA^T$. This is doable, but way too much work. Instead, we use the characterization of transposes given in terms of the dot product. We have for any $\mathbf{x} \in \mathbb{R}^k$ and $\mathbf{y} \in \mathbb{R}^m$:

$$(AB)\mathbf{x} \cdot \mathbf{y} = A(B\mathbf{x}) \cdot \mathbf{y} = B\mathbf{x} \cdot A^T\mathbf{y} = \mathbf{x} \cdot B^TA^T\mathbf{y}$$

where in the second equality we use the defining property of A^T and in the final equality the defining property of B^T. This shows that B^TA^T satisfies the defining property of $(AB)^T$ as the unique matrix satisfying

$$(AB)\mathbf{x} \cdot \mathbf{y} = \mathbf{x} \cdot (AB)^T\mathbf{y},$$

so we must have $(AB)^T = B^T A^T$ as claimed.

Warm-Up 2. Suppose that A, B are $m \times n$ matrices satisfying

$$A\mathbf{x} \cdot \mathbf{y} = B\mathbf{x} \cdot \mathbf{y} \text{ for all } \mathbf{x}, \mathbf{y} \in \mathbb{R}^n.$$

We show that A and B must be the same. One way to show this is to use the defining property of transposes to say that

$$A\mathbf{x} \cdot \mathbf{y} = B\mathbf{x} \cdot \mathbf{y} = \mathbf{x} \cdot B^T\mathbf{y} \text{ for all } \mathbf{x}, \mathbf{y} \in \mathbb{R}^n,$$

which implies that $A^T = B^T$ and hence that $A = B$. However, let's give another argument which doesn't involve transposes.

The given equality gives

$$A\mathbf{x} \cdot \mathbf{y} - B\mathbf{x} \cdot \mathbf{y} = 0, \text{ so } (A\mathbf{x} - B\mathbf{x}) \cdot \mathbf{y} = 0 \text{ for all } \mathbf{x}, \mathbf{y}.$$

This says that for a fixed $\mathbf{x}$, $A\mathbf{x} - B\mathbf{x}$ is a vector which is orthogonal to every other vector, which means that it must be zero since the zero vector is the only vector with this property. To be precise, applying the equality above to the vector $\mathbf{y} = A\mathbf{x} - B\mathbf{x}$ itself gives

$$(A\mathbf{x} - B\mathbf{x}) \cdot (A\mathbf{x} - B\mathbf{x}) = 0, \text{ so } \|A\mathbf{x} - B\mathbf{x}\|^2 = 0$$

and thus $A\mathbf{x} - B\mathbf{x} = \mathbf{0}$ as claimed. Thus $A\mathbf{x} = B\mathbf{x}$ for any $\mathbf{x} \in \mathbb{R}^n$, so $A = B$.

Complex dot product. Before continuing, we note that there is an analog of the dot product in the complex setting, and that many of the same properties we'll see remain true in the complex setting as well. To be precise, for *complex* vectors $\mathbf{z}, \mathbf{w} \in \mathbb{C}^n$ we define the complex (or *Hermitian*) dot product as

$$\mathbf{z} \cdot \mathbf{w} = z_1\overline{w_1} + \cdots + z_n\overline{w_n}$$

where $z_1, \ldots, z_n \in \mathbb{C}$ and $w_1, \ldots, w_n \in \mathbb{C}$ are the components of $\mathbf{z}$ and $\mathbf{w}$ respectively. Note that this is very similar to the expression for the real product in $\mathbb{R}^n$ only that here we take the conjugate of the entries of the second vector. The reason for this is the following. As in the real case, we would like to define the *length* of $\mathbf{z}$ as $\sqrt{\mathbf{z} \cdot \mathbf{z}}$, and we would like for this value to be real so that it makes sense to think of it as a "length". If we had simply defined

$$\mathbf{z} \cdot \mathbf{w} = z_1 w_1 + \cdots + z_n w_n$$

without conjugates, we would get $\mathbf{z} \cdot \mathbf{z} = z_1^2 + \cdots + z_n^2$, an expression which may still be complex. Defining the complex dot product as we did gives

$$\mathbf{z} \cdot \mathbf{z} = z_1\overline{z_1} + \cdots + z_n\overline{z_n},$$

which is a real expression since each individual term $z_i\overline{z_i}$ is real. Thus it makes sense to take the square root of $\mathbf{z} \cdot \mathbf{z}$ and get a nonnegative real number as a result.

The complex dot product satisfies the same kinds of properties as does the real dot product, with a few modifications as will be elaborated on in the homework. As in the real case, we say that $\mathbf{z}, \mathbf{w} \in \mathbb{C}^n$ are *orthogonal* if $\mathbf{z} \cdot \mathbf{w} = 0$. The point is that much of what we talk about for $\mathbb{R}^n$ using the real dot product will be true for $\mathbb{C}^n$ using the complex product, and it will be useful to recognize when this is so.

Exercise. Suppose that $A \in M_{m,n}(\mathbb{C})$. Show that $A\mathbf{z} \cdot \mathbf{w} = \mathbf{z} \cdot \overline{A^T}\mathbf{w}$ for any $\mathbf{z} \in \mathbb{R}^n$ and $\mathbf{w} \in \mathbb{R}^m$, where $\overline{A^T}$ denotes the conjugate transpose of A. (Thus, the conjugate transpose of a complex matrix is the "correct" analog of the ordinary transpose of a real matrix.)

Definition. Suppose that V is a subspace of $\mathbb{R}^n$. A basis $\mathbf{b}_1, \dots, \mathbf{b}_k$ is an *orthogonal* basis for V if each vector in this basis is orthogonal to every other vector in the basis: $\mathbf{b}_i \cdot \mathbf{b}_j = 0$ for $i \neq j$. The given basis is an *orthonormal* basis if it is orthogonal and in addition each vector in the basis has length 1, which is equivalent to $\mathbf{b}_i \cdot \mathbf{b}_i = 1$ for all i. For instance, the standard basis $\mathbf{e}_1, \dots, \mathbf{e}_n$ is an orthonormal basis of $\mathbb{R}^n$.

Why we care about orthogonal bases. Orthogonal bases are important because it is straightforward to write an arbitrary vector as a linear combination of orthogonal basis vectors. Indeed, suppose that $\mathbf{b}_1, \dots, \mathbf{b}_k$ is an orthogonal basis for a subspace V of $\mathbb{R}^n$. The claim is that for any $\mathbf{v} \in V$, we have

$$\mathbf{v} = \text{proj}_{\mathbf{b}_1} \mathbf{v} + \cdots + \text{proj}_{\mathbf{b}_k} \mathbf{v}.$$

The point is that since $\mathbf{b}_1, \dots, \mathbf{b}_k$ is a basis of V, we know that

$$\mathbf{v} = c_1\mathbf{b}_1 + \cdots + c_k\mathbf{b}_k$$

for some $c_1, \dots, c_k \in \mathbb{R}$, but as opposed to having to solve some system to determine $c_1, \dots, c_k$, when our basis is orthogonal we know immediately that coefficients needed are those that describe the orthogonal projection of $\mathbf{v}$ onto the various orthogonal basis vectors.

Indeed, taking the dot product of both sides of $\mathbf{v} = c_1\mathbf{b}_1 + \cdots + c_k\mathbf{b}_k$ with some $\mathbf{b}_\ell$ gives

$$\mathbf{v} \cdot \mathbf{b}_\ell = (c_1\mathbf{b}_1 + \cdots + c_k\mathbf{b}_k) \cdot \mathbf{b}_\ell = c_\ell(\mathbf{b}_\ell \cdot \mathbf{b}_\ell)$$

since when expanding $(c_1\mathbf{b}_1 + \cdots + c_k\mathbf{b}_k) \cdot \mathbf{b}_\ell$ we get that any term of the form $\mathbf{b}_i \cdot \mathbf{b}_\ell$ for $i \neq \ell$ is zero since the basis $\mathbf{b}_1, \dots, \mathbf{b}_\ell$ is orthogonal. Solving for c_ℓ (note that $\mathbf{b}_\ell \cdot \mathbf{b}_\ell \neq 0$ since $\mathbf{b}_\ell$ is not the zero vector) gives

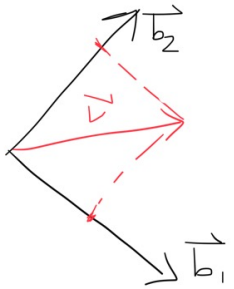
$$c_\ell = \frac{\mathbf{v} \cdot \mathbf{b}_\ell}{\mathbf{b}_\ell \cdot \mathbf{b}_\ell}.$$

Thus these must be the coefficients needed in our linear combination, so

$$v = \left(\frac{\mathbf{v} \cdot \mathbf{b}_1}{\mathbf{b}_1 \cdot \mathbf{b}_1}\right)\mathbf{b}_1 + \cdots + \left(\frac{\mathbf{v} \cdot \mathbf{b}_k}{\mathbf{b}_k \cdot \mathbf{b}_k}\right)\mathbf{b}_k,$$

and the i-th term here is precisely the orthogonal projection of $\mathbf{v}$ onto $\mathbf{b}_i$.

This all makes sense geometrically. For instance, consider two orthogonal vectors $\mathbf{b}_1, \mathbf{b}_2$ in $\mathbb{R}^2$ and some other vector $\mathbf{v}$:



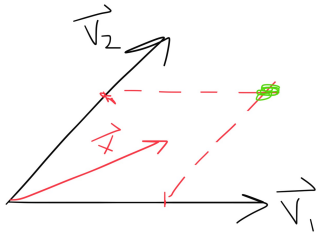
Then indeed adding up the orthogonal projections of $\mathbf{v}$ onto $\mathbf{b}_1$ and $\mathbf{b}_2$ do visually give back $\mathbf{v}$ itself.

Exercise. So, when we have an orthogonal basis, writing a vector in terms of that basis amounts to projecting it onto each basis vector and then adding up all the resulting projections. Conversely, show that only orthogonal bases have this property: that is, if $\mathbf{v}_1, \dots, \mathbf{v}_k$ is a basis of V such that

$$\mathbf{v} = \operatorname{proj}_{\mathbf{v}_1} \mathbf{v} + \cdots + \operatorname{proj}_{\mathbf{v}_k} \mathbf{v} \text{ for all } \mathbf{v} \in V,$$

show that $\mathbf{v}_1, \dots, \mathbf{v}_k$ must be orthogonal.

Indeed, if $\mathbf{v}_1, \mathbf{v}_2 \in \mathbb{R}^2$ are not orthogonal, we have something like:



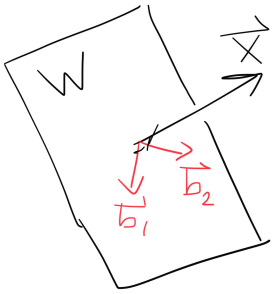
so adding together the orthogonal projections of $\mathbf{x}$ onto $\mathbf{v}_1, \mathbf{v}_2$ does not give back $\mathbf{x}$ itself.

Existence of orthonormal bases. The fact above only applied to orthogonal bases, so in order for it to be useful we have to know whether orthogonal bases exist. The fact is that any subspace of $\mathbb{R}^n$ has an orthogonal bases, and hence an orthonormal basis as well. (To turn an orthogonal basis into an orthonormal basis we just have to divide each basis vector by its length.) We'll give one proof of this here using induction, and will outline another proof next time using the so-called *Gram-Schmidt process*.

To be clear, suppose that V is a subspace of $\mathbb{R}^n$, and proceed by induction on $m = \dim V$. In the base case, $\dim V = 1$, so a basis of V consists of any nonzero vector $\mathbf{v}$ in V, and this vector itself forms an orthogonal basis for V, so we are done with the base case. Suppose for our induction hypothesis that we have any k-dimensional subspace of V has an orthogonal basis, and suppose now that V is $(k+1)$-dimensional. Take any nonzero $\mathbf{x} \in V$, and consider the space W of all things in V which are orthogonal to $\mathbf{x}$:

$$W := \{\mathbf{v} \in V \mid \mathbf{x} \cdot \mathbf{v} = 0\}.$$

(This is a subspace of $\mathbb{R}^n$ since the sum or scalar multiple of vectors orthogonal to $\mathbf{x}$ is itself orthogonal to $\mathbf{x}$.) The idea is to get an orthogonal basis of W from the induction hypothesis and then tack $\mathbf{x}$ onto this basis to get an orthogonal basis for all V. In the case of 3-dimensions we have something like:



where W is orthogonal to $\mathbf{x}$, $\mathbf{b}_1$ and $\mathbf{b}_2$ form an orthogonal basis of W, and then $\mathbf{b}_1, \mathbf{b}_2, \mathbf{x}$ all together give an orthogonal basis of our 3-dimensional space.

To be able to apply the induction hypothesis we need to know that $\dim W = k$, and we'll prove this next time as a Warm-Up. Given this, the induction hypothesis implies that W has an orthogonal basis, say $\mathbf{b}_1, \dots, \mathbf{b}_k$. Then $\mathbf{b}_1, \dots, \mathbf{b}_k, \mathbf{x}$ consists of orthogonal vectors since the $\mathbf{b}$'s are orthogonal to each other since they came from an orthogonal basis of W, and the $\mathbf{b}$'s are orthogonal to $\mathbf{x}$ since the $\mathbf{b}$'s are in W and anything in W is orthogonal to $\mathbf{x}$. The remaining claim is that $\mathbf{b}_1, \dots, \mathbf{b}_k, \mathbf{x}$ actually gives a basis for V, which we'll also prove as a Warm-Up next time. The end result is an orthogonal basis for V, so by induction we conclude that every subspace of $\mathbb{R}^n$ has an orthogonal basis.

Lecture 3: Orthogonal Projections

Warm-Up 1. Referring back to our final proof from last time, the setup is we had a $(k+1)$-dimensional subspace V of $\mathbb{R}^n$ and a nonzero vector $\mathbf{x} \in V$. We defined W to be the space of all things in W orthogonal to $\mathbf{x}$:

$$W := \{\mathbf{v} \in V \mid \mathbf{x} \cdot \mathbf{v} = 0\},$$

and claimed that $\dim W = k$. Intuitively, the point is that $\mathbf{x} \cdot \mathbf{v} = 0$ is a single linear "constraint" on vectors in V, and each such (independent) constraint cuts the dimension down by 1.

To make this precise, consider the linear transformation $T : V \to \mathbb{R}$ defined by

$$T(\mathbf{v}) = \mathbf{x} \cdot \mathbf{v}.$$

The image of this contains $T(\mathbf{x}) = \mathbf{x} \cdot \mathbf{x} \neq 0$ and so is nonzero, meaning that the image must be all of $\mathbb{R}$. Hence by rank-nullity:

$$\dim(\ker T) = \dim V - \dim(\operatorname{im} T) = (k+1) - 1 = k,$$

but since $\ker T = W$ by the definition of W, we have $\dim W = k$ as claimed.

Warm-Up 2. In the conclusion of the final proof from last time we end up with orthogonal vectors $\mathbf{b}_1, \dots, \mathbf{b}_k, \mathbf{x} \in V$, where the $\mathbf{b}$'s came from an orthogonal basis of W. The final claim was that these give a basis for V. Indeed, the point here is that nonzero orthogonal vectors are always linearly independent, and so we have $k+1$ linearly independent vectors in a $(k+1)$-dimensional space, meaning that they must form a basis.

To see that nonzero orthogonal vectors $\mathbf{v}_1, \dots, \mathbf{v}_m$ are always linearly independent, start with an equation

$$\mathbf{0} = c_1\mathbf{v}_1 + \cdots + c_m\mathbf{v}_m$$

expressing $\mathbf{0}$ as a linear combination of said vectors. Since $\mathbf{v}_1, \dots, \mathbf{v}_m$, the coefficients c_i in this expression must be

$$c_i = \frac{\mathbf{0} \cdot \mathbf{v}_i}{\mathbf{v}_i \cdot \mathbf{v}_i} = 0$$

since to write a vector as a linear combination of orthogonal vectors we only need to orthogonally project it onto those orthogonal vectors. Thus $\mathbf{v}_1, \dots, \mathbf{v}_m$ are linearly independent as claimed.

Gram-Schmidt. The proof we gave for the fact that subspaces of $\mathbb{R}^n$ always have orthogonal (and hence orthonormal) bases using induction wasn't constructive, in that it doesn't tell you how to actually construct such a basis. Another proof uses what's called the *Gram-Schmidt* process to actually construct an orthogonal basis. Given a basis $\mathbf{v}_1, \dots, \mathbf{v}_k$ of V, the claim is that vectors defined by:

$$\begin{aligned}
\mathbf{b}_1 &= \mathbf{v}_1 \\
\mathbf{b}_2 &= \mathbf{v}_2 - \operatorname{proj}_{\mathbf{b}_1} \mathbf{v}_2 \\
\mathbf{b}_3 &= \mathbf{v}_3 - \operatorname{proj}_{\mathbf{b}_1} \mathbf{v}_3 - \operatorname{proj}_{\mathbf{b}_2} \mathbf{v}_3 \\
&\vdots = \vdots \\
\mathbf{b}_k &= \mathbf{v}_k - \text{the projections of } \mathbf{v}_k \text{ onto all previously constructed } \mathbf{b} \text{ vectors}
\end{aligned}$$

are orthogonal, and hence form an orthogonal basis for V. To get an orthonormal basis $\mathbf{u}_1, \dots, \mathbf{u}_k$ for V, we finish by dividing each of these vectors by their lengths:

$$\mathbf{u}_1 = \frac{\mathbf{b}_1}{\|\mathbf{b}_1\|}, \dots, \mathbf{u}_k = \frac{\mathbf{b}_k}{\|\mathbf{b}_k\|}.$$

Check the book or my Math 290 lecture notes for various computation examples of applying the Gram-Schmidt process.

The key point for us is proving that the vectors $\mathbf{b}_1, \dots, \mathbf{b}_k$ arising in this process are in fact orthogonal. Intuitively the idea is that each projection we subtract at each step has the effect of making the resulting vector orthogonal to the vector being projected onto, so at each step we get a vector orthogonal to all the ones constructed before it. To be clear, suppose we have shown already that $\mathbf{b}_1, \dots, \mathbf{b}_{i-1}$ are orthogonal to each other. Then we must show that $\mathbf{b}_i$ is orthogonal to each of $\mathbf{b}_1, \dots, \mathbf{b}_{i-1}$. The expression for $\mathbf{b}_i$ is

$$\mathbf{b}_i = \mathbf{v}_i - \operatorname{proj}_{\mathbf{b}_1} \mathbf{v}_i - \cdots - \operatorname{proj}_{\mathbf{b}_{i-1}} \mathbf{v}_i.$$

Take some $\mathbf{b}_\ell$ for $1 \leq \ell \leq i-1$, and compute the dot product of both sides here with $\mathbf{b}_\ell$:

$$\mathbf{b}_i \cdot \mathbf{b}_\ell = (\mathbf{v}_i - \operatorname{proj}_{\mathbf{b}_1} \mathbf{v}_i - \cdots - \operatorname{proj}_{\mathbf{b}_{i-1}} \mathbf{v}_i) \cdot \mathbf{b}_\ell = \mathbf{v}_i \cdot \mathbf{b}_\ell - (\operatorname{proj}_{\mathbf{b}_\ell} \mathbf{v}_i) \cdot \mathbf{b}_\ell,$$

where we use the fact that $(\operatorname{proj}_{\mathbf{b}_m} \mathbf{v}_i) \cdot \mathbf{b}_\ell = 0$ for $m \leq \ell$ since $\operatorname{proj}_{\mathbf{b}_m} \mathbf{v}_i$ is a multiple of $\mathbf{b}_m$ and $\mathbf{b}_\ell$ is orthogonal to $\mathbf{b}_m$. Using the formula for orthogonal projections we have:

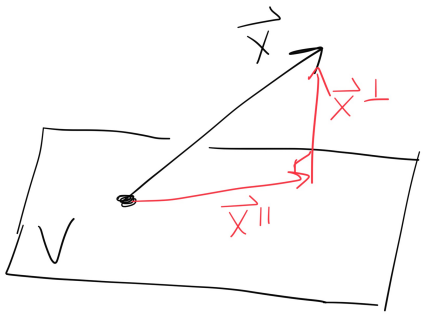
$$\mathbf{b}_i \cdot \mathbf{b}_\ell = \mathbf{v}_i \cdot \mathbf{b}_\ell - \left(\frac{\mathbf{v}_i \cdot \mathbf{b}_\ell}{\mathbf{b}_\ell \cdot \mathbf{b}_\ell}\right) \mathbf{b}_\ell \cdot \mathbf{b}_\ell = \mathbf{v}_i \cdot \mathbf{b}_\ell - \mathbf{v}_i \cdot \mathbf{b}_\ell = 0,$$

so $\mathbf{b}_i$ is orthogonal to $\mathbf{b}_\ell$ for $1 \leq \ell \leq i-1$ as desired.

Orthogonal decompositions. We can now justify a nice geometric fact, which will lead to a more general definition of "orthogonal projection" when we want to project onto an entire subspace of $\mathbb{R}^n$ as opposed to simply projecting onto a vector. Let V be a subspace of $\mathbb{R}^n$. The claim is that given any $\mathbf{x} \in \mathbb{R}^n$, there exist *unique* vectors $\mathbf{x}^{\parallel} \in V$ and $\mathbf{x}^{\perp} \in V^{\perp}$ such that

$$\mathbf{x} = \mathbf{x}^{\parallel} + \mathbf{x}^{\perp}.$$

Here $V^{\perp}$ denotes the *orthogonal complement* to V and consists of all vectors in $\mathbb{R}^n$ which are orthogonal to everything in V: $V^{\perp} := \{\mathbf{y} \in \mathbb{R}^n \mid \mathbf{y} \cdot \mathbf{v} = 0 \text{ for all } \mathbf{v} \in V\}$. So, this result says that any $\mathbf{x}$ can be decomposed into the sum of something in V (we say "parallel" to V) and something orthogonal to V. In the 3-dimensional case we have something like:



where $\mathbf{x}$ is indeed obtained by adding $\mathbf{x}^{\parallel} \in V$ and $\mathbf{x}^{\perp} \in V^{\perp}$.

Here we prove the existence of such a decomposition, and leave the uniqueness for the Warm-Up next time. Let $\mathbf{u}_1, \dots, \mathbf{u}_k$ be an orthonormal basis of V. Let $\mathbf{x} \in \mathbb{R}^n$ and define $\mathbf{x}^{\parallel}$ to be given by

$$\mathbf{x}^{\parallel} = \text{proj}_{\mathbf{u}_1} \mathbf{x} + \cdots + \text{proj}_{\mathbf{u}_k} \mathbf{x} = (\mathbf{x} \cdot \mathbf{u}_1)\mathbf{u}_1 + \cdots + (\mathbf{x} \cdot \mathbf{u}_k)\mathbf{u}_k.$$

(The intuition for this comes from the fact that if $\mathbf{x}^{\parallel}$ is going to be in V, it should expressible as a sum of orthogonal projections onto the given orthonormal basis vectors, so we use this fact to actually define what $\mathbf{x}^{\parallel}$ should be.) With this definition, we then define $\mathbf{x}^{\perp}$ to be given by

$$\mathbf{x}^{\perp} = \mathbf{x} - \mathbf{x}^{\parallel}$$

since this is the only way in which $\mathbf{x} = \mathbf{x}^{\parallel} + \mathbf{x}^{\perp}$ can be true. We thus have our desired decomposition of V as long as we verify that the $\mathbf{x}^{\parallel}$ and $\mathbf{x}^{\perp}$ thus defined indeed have the properties they are meant to have. First, $\text{proj}_{\mathbf{u}_i} \mathbf{x}$ is a multiple of $\mathbf{u}_i$, so the expression defining $\mathbf{x}^{\parallel}$ is a linear combination of the vectors $\mathbf{u}_1, \dots, \mathbf{u}_k$ in V, so $\mathbf{x}^{\parallel} \in V$ as desired.

Finally we need to know that $\mathbf{x}^{\perp}$ as defined is orthogonal to everything in V. Let $\mathbf{v} \in V$ and write it according to the given orthonormal basis as

$$\mathbf{v} = (\mathbf{v} \cdot \mathbf{u}_1)\mathbf{u}_1 + \cdots + (\mathbf{v} \cdot \mathbf{u}_k)\mathbf{u}_k,$$

where each term is individually a projection of $\mathbf{v}$ onto a basis vector. We then compute:

$$\mathbf{x}^{\perp} \cdot \mathbf{v} = (\mathbf{x} - \mathbf{x}^{\parallel}) \cdot \mathbf{v}$$

$$
\begin{aligned}
&= \mathbf{x} \cdot \mathbf{v} - \mathbf{x}^{\parallel} \cdot \mathbf{v} \\
&= \mathbf{x} \cdot [(\mathbf{v} \cdot \mathbf{u}_1)\mathbf{u}_1 + \cdots + (\mathbf{v} \cdot \mathbf{u}_k)\mathbf{u}_k] - [(\mathbf{x} \cdot \mathbf{u}_1)\mathbf{u}_1 + \cdots + (\mathbf{x} \cdot \mathbf{u}_k)\mathbf{u}_k] \cdot \mathbf{v} \\
&= (\mathbf{v} \cdot \mathbf{u}_1)(\mathbf{x} \cdot \mathbf{u}_1) + \cdots ((\mathbf{v} \cdot \mathbf{u}_k)(\mathbf{x} \cdot \mathbf{u}_k) - (\mathbf{x} \cdot \mathbf{u}_1)(\mathbf{u}_1 \cdot \mathbf{v}) - \cdots - (\mathbf{x} \cdot \mathbf{u}_k)(\mathbf{u}_k \cdot \mathbf{v}) \\
&= 0.
\end{aligned}
$$

Thus $\mathbf{x}^{\perp}$ is orthogonal to everything in V as claimed, so $\mathbf{x}^{\perp} \in V^{\perp}$ and $\mathbf{x} = \mathbf{x}^{\parallel} + \mathbf{x}^{\perp}$ is indeed the desired orthogonal decomposition.

Lecture 4: Orthogonal Matrices

Warm-Up. As we saw last time, given a subspace V of $\mathbb{R}^n$ and a vector $\mathbf{x} \in \mathbb{R}^n$, there exists a decomposition

$$\mathbf{x} = \mathbf{x}^{\parallel} + \mathbf{x}^{\perp}$$

where $\mathbf{x}^{\parallel} \in V$ and $\mathbf{x}^{\perp} \in V^{\perp}$. Here we show that such a decomposition is unique, meaning that if

$$\mathbf{x}^{\parallel} + \mathbf{x}^{\perp} = \mathbf{y}^{\parallel} + \mathbf{y}^{\perp}$$

with $\mathbf{x}^{\parallel}, \mathbf{y}^{\parallel} \in V$ and $\mathbf{x}^{\perp}, \mathbf{y}^{\perp} \in V^{\perp}$, then it must be true that $\mathbf{x}^{\parallel} = \mathbf{y}^{\parallel}$ and $\mathbf{x}^{\perp} = \mathbf{y}^{\perp}$.

Indeed, rearranging terms in the given equation yields

$$\mathbf{x}^{\parallel} - \mathbf{y}^{\parallel} = \mathbf{y}^{\perp} - \mathbf{x}^{\perp}.$$

Both terms on the left side are in V, so the left side is in V since V is a subspace of $\mathbb{R}^n$, and both terms on the right side are in the orthogonal complement $V^{\perp}$ so the entire right side is as well since $V^{\perp}$ is also a subspace of $\mathbb{R}^n$. (You should be able to show that if $\mathbf{a}, \mathbf{b}$ are both orthogonal to everything in V, then so are $\mathbf{a} + \mathbf{b}$ and $c\mathbf{a}$ for any scalar $c \in \mathbb{R}$.) Thus the common vector $\mathbf{x}^{\parallel} - \mathbf{y}^{\parallel} = \mathbf{y}^{\perp} - \mathbf{x}^{\perp}$ belongs to both V and $V^{\perp}$, and so in particular must be orthogonal to itself. Since the only thing orthogonal to itself is the zero vector we get $\mathbf{x}^{\parallel} - \mathbf{y}^{\parallel} = \mathbf{0}$ and $\mathbf{y}^{\perp} - \mathbf{x}^{\perp} = \mathbf{0}$, and hence $\mathbf{x}^{\parallel} = \mathbf{y}^{\parallel}$ and $\mathbf{y}^{\perp} = \mathbf{x}^{\perp}$ as claimed.

Orthogonal projections. For a subspace V of $\mathbb{R}^n$, we define the *orthogonal projection* $\operatorname{proj}_V \mathbf{x}$ of $\mathbf{x} \in \mathbb{R}^n$ onto V to be the parallel component $\mathbf{x}^{\parallel} \in V$ in the orthogonal decomposition $\mathbf{x} = \mathbf{x}^{\parallel} + \mathbf{x}^{\perp}$. Thus concretely, if $\mathbf{u}_1, \ldots, \mathbf{u}_k$ is an orthonormal basis of V we have

$$\operatorname{proj}_V \mathbf{x} = \operatorname{proj}_{\mathbf{u}_1} \mathbf{x} + \cdots + \operatorname{proj}_{\mathbf{u}_k} \mathbf{x}.$$

You can check that the function $T : \mathbb{R}^n \to \mathbb{R}^n$ defined by $T(\mathbf{x}) = \operatorname{proj}_V \mathbf{x}$ is a linear transformation whose image is V and kernel is $V^{\perp}$.

Given this, the standard matrix of T is easy to describe. Let $Q = \begin{pmatrix} \mathbf{u}_1 & \cdots & \mathbf{u}_k \end{pmatrix}$ be the $n \times k$ matrix having $\mathbf{u}_1, \ldots, \mathbf{u}_k$ as columns. Note the result of computing the product $QQ^T\mathbf{x}$ for any $\mathbf{x} \in \mathbb{R}^n$:

$$
\begin{aligned}
QQ^T\mathbf{x} &= \begin{pmatrix} | & & | \\ \mathbf{u}_1 & \cdots & \mathbf{u}_k \\ | & & | \end{pmatrix} \begin{pmatrix} — & \mathbf{u}_1 & — \\ & \vdots & \\ — & \mathbf{u}_k & — \end{pmatrix} \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \\
&= \begin{pmatrix} | & & | \\ \mathbf{u}_1 & \cdots & \mathbf{u}_k \\ | & & | \end{pmatrix} \begin{pmatrix} \mathbf{x} \cdot \mathbf{u}_1 \\ \vdots \\ \mathbf{x} \cdot \mathbf{u}_k \end{pmatrix}
\end{aligned}
$$

$$= (\mathbf{x} \cdot \mathbf{u}_1)\mathbf{u}_1 + \cdots + (\mathbf{x} \cdot \mathbf{u}_k)\mathbf{u}_k.$$

This is precisely the orthogonal projection of $\mathbf{x}$ onto V, so we have

$$QQ^T\mathbf{x} = \text{proj}_V \mathbf{x},$$

meaning that QQ^T is the standard matrix of this orthogonal projection. Thus in practice, finding the matrix of an orthogonal projection is fairly straightforward: get an orthonormal basis for the space being projected onto using the Gram-Schmidt process, use those basis vectors as the columns of a matrix Q, and compute QQ^T.

Note that the product Q^TQ is also simple to describe:

$$\begin{pmatrix} — & \mathbf{u}_1 & — \\ & \vdots & \\ — & \mathbf{u}_k & — \end{pmatrix} \begin{pmatrix} | & & | \\ \mathbf{u}_1 & \cdots & \mathbf{u}_k \\ | & & | \end{pmatrix} = \begin{pmatrix} \mathbf{u}_1 \cdot \mathbf{u}_1 & \cdots & \mathbf{u}_1 \cdot \mathbf{u}_k \\ \vdots & \ddots & \vdots \\ \mathbf{u}_k \cdot \mathbf{u}_1 & \cdots & \mathbf{u}_k \cdot \mathbf{u}_k \end{pmatrix} = \begin{pmatrix} 1 & & \\ & \ddots & \\ & & 1 \end{pmatrix}$$

where we use the fact that the vectors $\mathbf{u}_1, \ldots, \mathbf{u}_k$ are orthonormal. Thus, any matrix Q with orthonormal columns satisfies $Q^TQ = I$, and the result of the product QQ^T is the matrix of the orthogonal projection onto the image of Q.

Orthogonal matrices. In the setup above Q was an $n \times k$ matrix, so not necessarily square. In the case were Q is square we give such a matrix a special name: an *orthogonal* matrix is a square matrix with orthonormal columns. As above, for an orthogonal matrix Q we definitely have $Q^TQ = I$, but now since Q is square we know changing the order of Q and Q^T will still give the identity, so $QQ^T = I$.

Note that this makes sense: above we said that QQ^T was the matrix of an orthogonal projection, specifically the orthogonal projection onto the image of Q—when Q is square, the columns of Q form an orthonormal basis for all of $\mathbb{R}^n$, and the orthogonal projection of a vector in $\mathbb{R}^n$ onto $\mathbb{R}^n$ itself is the identity since projecting a vector which is already in the space being projected onto does nothing.

Equivalent characterization. The condition $Q^TQ = I = QQ^T$ for an orthogonal matrix says that Q is invertible with inverse equal to its transpose, so an orthogonal matrix can equivalently be characterized as such a matrix: Q is orthogonal if and only if Q is invertible and $Q^{-1} = Q^T$.

Length preserving. Here is a non-obvious property which orthogonal matrices have: if Q is orthogonal, then $\|Q\mathbf{x}\| = \|\mathbf{x}\|$ for all $\mathbf{x} \in \mathbb{R}^n$. We say that Q *preserves length*. We give two justifications of this fact, using the two characterizations we had above of orthogonal matrices.

First, suppose Q has orthonormal columns $\mathbf{u}_1, \ldots, \mathbf{u}_n$ and let $\mathbf{x} \in \mathbb{R}^n$. Then

$$\begin{aligned} Q\mathbf{x} \cdot Q\mathbf{x} &= \begin{pmatrix} \mathbf{u}_1 & \cdots & \mathbf{u}_n \end{pmatrix} \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \cdot \begin{pmatrix} \mathbf{u}_1 & \cdots & \mathbf{u}_n \end{pmatrix} \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \\ &= (x_1\mathbf{u}_1 + \cdots + x_n\mathbf{u}_n) \cdot (x_1\mathbf{u}_1 + \cdots + x_n\mathbf{u}_n) \\ &= x_1^2(\mathbf{u}_1 \cdot \mathbf{u}_1) + \cdots + x_n^2(\mathbf{u}_n \cdot \mathbf{u}_n) \\ &= x_1^2 + \cdots + x_n^2 \\ &= \mathbf{x} \cdot \mathbf{x} \end{aligned}$$

where in the third line we distribute and use the fact that $\mathbf{u}_i \cdot \mathbf{u}_j = 0$ for $i \neq j$, and in the fourth line the fact that each $\mathbf{u}_i$ has length 1. This gives $\|Q\mathbf{x}\| = \sqrt{Q\mathbf{x} \cdot Q\mathbf{x}} = \sqrt{\mathbf{x} \cdot \mathbf{x}} = \|\mathbf{x}\|$ as claimed.

For a second (quicker) proof, suppose that Q is a square matrix satisfying $QQ^T = I = Q^TQ$. Then

$$Q\mathbf{x} \cdot Q\mathbf{x} = \mathbf{x} \cdot Q^TQ\mathbf{x} = \mathbf{x} \cdot \mathbf{x},$$

where in the second step we use the defining property of transposes and in the second the fact that $Q^TQ = I$. As before, this implies $\|Q\mathbf{x}\| = \|\mathbf{x}\|$ as well.

The book defines an *orthogonal transformation* $T : \mathbb{R}^n \to \mathbb{R}^n$ to be a linear transformation which preserves length in the sense that $\|T(\mathbf{x})\| = \|\mathbf{x}\|$ for all $\mathbf{x} \in \mathbb{R}^n$. Here we have shown that orthogonal matrices always give rise to orthogonal transformations, but in the fact the converse is true: if T is an orthogonal transformation, then the standard matrix of T is an orthogonal matrix. Thus we could also have defined an orthogonal matrix to be one which preserves length. We'll prove that an orthogonal transformation is defined by an orthogonal matrix next time.

Rotations and reflections. At this point we can start to figure out geometrically what orthogonal transformations actually do. Thinking back to last quarter, the only transformations we previously saw which preserve length are rotations and reflections. In fact, in turns out that these are the only possible orthogonal transformation, meaning that if T preserves length then T must be either a rotation or a reflection. We don't yet have enough developed to be able to prove this, but we'll come back to it after we learn about eigenvectors.

As a first step, how do we determine whether a given orthogonal matrix is meant to represent either a rotation or a reflection? For instance, the matrix

$$\begin{pmatrix} 2/3 & -2/3 & 1/3 \\ 1/3 & 2/3 & 2/3 \\ 2/3 & 1/3 & -2/3 \end{pmatrix}$$

has orthonormal columns and hence is orthogonal, so it must represent either a rotation or a reflection. However, just by looking at it it is not at all clear which it should be. Even so, if it is a rotation, what is it a rotation around, or if it is a reflection, what is it a reflection across? We'll see that determinants will give us an easy way of determine whether this is a rotation or a reflection, and then eigenvectors will allow us to describe explicitly what it does.

Lecture 5: More on Orthogonal Matrices

Warm-Up 1. Recall that an orthogonal matrix is a square matrix whose columns are orthonormal. We show that the rows of an orthogonal matrix are orthonormal as well; in other words, the claim is that if Q is orthogonal, then Q^T is also orthogonal.

Indeed, if Q is orthogonal, then $QQ^T = I = Q^TQ$. But $Q = (Q^T)^T$, so this says that

$$(Q^T)^TQ^T = I = Q^T(Q^T)^T,$$

meaning that Q^T itself is a square matrix with the property that its inverse is its transpose. This shows that Q^T is orthogonal, so the columns of Q^T, and hence the rows of Q, are orthonormal.

Warm-Up 2. This next problem is meant to illustrate various properties of orthogonal matrices, although the problem as stated seems to come out of nowhere. The claim is that if $Q_1R_1 = Q_2R_2$ where all matrices are $n \times n$, Q_1 and Q_2 are orthogonal, and R_1 and R_2 are upper triangular with positive diagonal entries, then $Q_1 = Q_2$ and $R_1 = R_2$.

Before proving this, here is the point. It is a true fact that any $n \times k$ matrix A (square or not) with linearly independent columns can be written as

$$A = QR$$

where Q is an $n \times k$ matrix with orthonormal columns and R is a $k \times k$ upper triangular matrix with positive diagonal entries. Such an expression is called a QR *factorization* of A, and this Warm-Up is showing that this factorization is unique in the case where A is a square matrix. It is true that the QR factorization of a matrix is also unique in the non-square case as well, only this requires a different proof. We'll talk more about QR factorizations soon, where we'll see that they play a crucial role in understanding the geometric interpretation of determinants. Again, for now this problem is only meant to illustrate some properties of orthogonal matrices.

Since Q_2 is orthogonal, it is invertible with inverse equal to its transpose, and since R_1 is upper triangular with positive diagonal entries, it too is invertible. Thus multiplying both sides of $Q_1R_1 = Q_2R_2$ on the left by Q_2^T and on the right by R_1^{-1} gives

$$Q_2^TQ_1 = R_2R_1^{-1}.$$

Now, the first Warm-Up implies that Q_2^T is orthogonal, so on the left sides we have a product of orthogonal matrices, and such products are themselves always orthogonal. Indeed, we just check that the transpose of $Q_2^TQ_1$ is its inverse:

$$(Q_2^TQ_1)(Q_2^TQ_1)^T = Q_2^TQ_1Q_1^T(Q_2^T)^T = Q_2^TQ_2 = I$$

and

$$(Q_2^TQ_1)^T(Q_2^TQ_1) = Q_1^TQ_2Q_2^TQ_1 = Q_1^TQ_1 = I.$$

On the other hand, since R_1 is upper triangular with positive diagonal entries R_1^{-1} is upper triangular with positive diagonal entries as well (show this!), so the right side $R_2R_1^{-1}$ of the equation above is a product of upper triangular matrices with positive diagonal entries and so is itself upper triangular with positive diagonal entries.

Thus we get that the common matrix $Q_2^TQ_1 = R_2R_1^{-1}$ is at the same time both orthogonal and upper triangular with positive diagonal entries. There aren't many matrices with these two properties, and indeed we claim that the identity matrix is the only one. If so, this will show that $Q_2^TQ_1 = I$ and $R_2R_1^{-1} = I$, which implies that $Q_1 = Q_2$ and $R_1 = R_2$ as claimed. So to finish this off, suppose that

$$\begin{pmatrix} * & \cdots & * \\ & \ddots & \vdots \\ & & * \end{pmatrix}$$

is upper triangular with positive diagonal entries and at the same time orthogonal. Since the first column should have length 1 and the only nonzero entry is the first entry, the first column must be either $\mathbf{e}_1$ or $-\mathbf{e}_1$. But the first entry, being on the diagonal, must be positive, so the first column must be $\mathbf{e}_1$.

Now, the second column is of the form $a\mathbf{e}_1 + b\mathbf{e}_2$ since only the first two entries can be nonzero. But this column must be orthogonal to the first, so we need

$$0 = (a\mathbf{e}_1 + b\mathbf{e}_2) \cdot \mathbf{e}_1 = a,$$

so the second column is of the form $b\mathbf{e}_2$. Again, for this to have length 1 and for b to be positive the second column must be $\mathbf{e}_2$. And so on, suppose we have shown already that the first k columns must be $\mathbf{e}_1, \ldots, \mathbf{e}_k$. The $(k+1)$-st column is of the form

$$a_1\mathbf{e}_1 + \cdots + a_{k+1}\mathbf{e}_{k+1},$$

and in order for this to be orthogonal to $\mathbf{e}_1, \ldots, \mathbf{e}_k$ requires that $a_1 = \cdots = a_i = 0$, as you can check. Thus the $(k+1)$-st column is of the form $a_{k+1}\mathbf{e}_{k+1}$, and so must be $\mathbf{e}_{k=1}$ itself in order to have length 1 and have a_{k+1} be positive. Thus an orthogonal upper triangular matrix with positive diagonal entries must have columns $\mathbf{e}_1, \ldots, \mathbf{e}_n$, and so must be the identity as claimed.

Exercise. (This only uses material from last quarter.) Show that if A is upper triangular with positive diagonal entries, then A^{-1} is upper triangular with positive diagonal entries as well, and show that if A and B are upper triangular with positive diagonal entries, then AB is upper triangular with positive diagonal entries.

Orthogonal equivalences. We now build on the list of properties which are equivalent to orthogonality for a matrix. Let Q be an $n \times n$ matrix. Then the following properties are equivalent, meaning they all imply each other:

(1) Q is orthogonal in the sense that it has orthonormal columns,

(2) $Q^TQ = I = QQ^T$, so Q is invertible and $Q^{-1} = Q^T$,

(3) Q preserves length in the sense that $\|Q\mathbf{x}\| = \|\mathbf{x}\|$ for all $\mathbf{x} \in \mathbb{R}^n$,

(4) Q preserves dot products in the sense that $Q\mathbf{x} \cdot Q\mathbf{y} = \mathbf{x} \cdot \mathbf{y}$ for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$.

We showed that properties (1) and (2) were equivalent last time, and that properties (1) or (2) implied property (3). We will show that property (3) implies property (1) in a bit. The fact that property (4) implies property (3) comes from taking $\mathbf{x} = \mathbf{y}$ in the statement of property (4), and the fact that any of the other properties imply property (4) is left to the homework.

Length preserving implies right angle preserving. To show that property (3) implies property (1), we first show that property (3) implies Q preserves right angles, in the sense that if $\mathbf{x}$ and $\mathbf{y}$ are orthogonal, then $Q\mathbf{x}$ and $Q\mathbf{y}$ are orthogonal. (Note that this is clear if we instead start by assuming property (4) holds since vectors being orthogonal is the same as saying their dot product is zero.) The key fact is the Pythagorean Theorem.

First we have

$$\|Q\mathbf{x} + Q\mathbf{y}\|^2 = \|Q(\mathbf{x} + \mathbf{y})\|^2 = \|\mathbf{x} + \mathbf{y}\|^2$$

where we use the linearity of Q and the fact that is preserves length. Now, since $\mathbf{x}$ and $\mathbf{y}$ are orthogonal the Pythagorean Theorem gives $\|\mathbf{x} + \mathbf{y}\|^2 = \|\mathbf{x}\|^2 + \|\mathbf{y}\|^2$, so

$$\|Q\mathbf{x} + Q\mathbf{y}\|^2 = \|\mathbf{x} + \mathbf{y}\|^2 = \|\mathbf{x}\|^2 + \|\mathbf{y}\|^2 = \|Q\mathbf{x}\|^2 + \|Q\mathbf{y}\|^2$$

where in the final step we again use the fact that Q preserves length. Thus $Q\mathbf{x}$ and $Q\mathbf{y}$ satisfy the requirements of the Pythagorean Theorem, so $Q\mathbf{x}$ and $Q\mathbf{y}$ are orthogonal as claimed.

Other angles. It is in fact true that an orthogonal transformation preserves not only right angles, but all angles in general. That is, if Q is orthogonal, the angle between $Q\mathbf{x}$ and $Q\mathbf{y}$ is the same as the one between $\mathbf{x}$ and $\mathbf{y}$. However, proving this requires more work, and in particular uses

the fact that arbitrary angles can be described solely in terms of dot products. This is left to the homework.

However, note that angle-preserving does not imply length-preserving. For instance, any scaling transformation preserves angles since scaling vectors doesn't affect the angle between them, but the only scaling transformations which are orthogonal are those where you scale by a factor of 1 (i.e. the identity) or -1.

(3) implies (1). Going back to the claimed equivalent characterizations of orthogonal transformations, we can now show that property (3) implies property (1), which is actually quite short. If Q is orthogonal, then Q preserves lengths and right angles. Thus since the standard basis vectors $\mathbf{e}_1, \dots, \mathbf{e}_n$ are orthonormal, the vectors

$$Q\mathbf{e}_1, \dots, Q\mathbf{e}_n$$

are orthonormal as well. But these vectors are precisely the columns of Q, so Q has orthonormal columns as required.

Unitary matrices. Finally, we briefly talk about the complex analogs of orthogonal matrices. Recall that the complex dot product between complex vectors is

$$\mathbf{z} \cdot \mathbf{w} = z_1\overline{w_1} + \cdots + z_n\overline{w_n}.$$

A square complex matrix $U \in M_n(\mathbb{C})$ is said to be *unitary* if it has orthonormal columns with respect to the complex dot product. Recalling that the complex analog of the transpose is the conjugate transpose, it turns out that U is unitary if and only if $UU^* = I = U^*U$ where U^* denotes the conjugate transpose of U. The proof is essentially the same as the one we gave in the real setting, only using the complex dot product instead. Similarly, U is unitary is also equivalent to U being length-preserving and to U preserving the complex dot product.

The point, as we alluded to when first introducing the complex dot product, is that many properties we've seen and will see for orthogonal matrices also hold for unitary matrices, and it will help to make these connections clear.

Lecture 6: Determinants

Warm-Up. We find all 3×3 unitary matrices of the form

$$\begin{pmatrix} a & 0 & \lambda i/\sqrt{2} \\ b & d & f \\ c & 0 & -\lambda i/\sqrt{2} \end{pmatrix}, \text{ where } a, b, c, d, f \in \mathbb{C} \text{ and } \lambda \in \mathbb{R}.$$

This is kind of random, but the point is simply to get practice working with orthonormal vectors with respect to the complex dot product.

To be unitary the columns must be orthonormal. To start with, the second column should have length 1, so:

$$\begin{pmatrix} 0 \\ d \\ 0 \end{pmatrix} \cdot \begin{pmatrix} 0 \\ d \\ 0 \end{pmatrix} = d\overline{d} = 1.$$

(In class I at first said this implied $d = \pm 1$ or $d = \pm i$, which is nonsense because there are many complex numbers of absolute value 1, as many of you pointed out. So we will just keep this

requirement written as $|d| = 1$ where $|z| = \sqrt{z\overline{z}}$ for $z \in \mathbb{C}$.) Next, in order for the first and third columns to be orthogonal to the second we need:

$$\begin{pmatrix} a \\ b \\ c \end{pmatrix} \cdot \begin{pmatrix} 0 \\ d \\ 0 \end{pmatrix} = a\overline{d} = 0 \quad \text{and} \quad \begin{pmatrix} \lambda i/\sqrt{2} \\ f \\ -\lambda i/\sqrt{2} \end{pmatrix} \cdot \begin{pmatrix} 0 \\ d \\ 0 \end{pmatrix} = f\overline{d} = 0.$$

Since $d \neq 0$, this means $b = 0$ and $f = 0$.

Now, in order for the third column to have norm 1 we need:

$$\begin{pmatrix} \lambda i/\sqrt{2} \\ 0 \\ -\lambda i/\sqrt{2} \end{pmatrix} \cdot \begin{pmatrix} \lambda i/\sqrt{2} \\ 0 \\ -\lambda i/\sqrt{2} \end{pmatrix} = \frac{\lambda i}{\sqrt{2}}\frac{\overline{\lambda i}}{\sqrt{2}} + \left(-\frac{\lambda i}{\sqrt{2}}\right)\overline{\left(-\frac{\lambda i}{\sqrt{2}}\right)} = -\frac{\lambda^2 i^2}{2} - \frac{\lambda^2 i^2}{2} = \lambda^2 = 1,$$

so $\lambda = \pm 1$ since $\lambda \in \mathbb{R}$. For the first column to be orthogonal to the third we need:

$$\begin{pmatrix} a \\ 0 \\ c \end{pmatrix} \cdot \begin{pmatrix} \lambda i/\sqrt{2} \\ 0 \\ -\lambda i/\sqrt{2} \end{pmatrix} = a\frac{\overline{\lambda i}}{\sqrt{2}} + c\overline{\left(-\frac{\lambda i}{\sqrt{2}}\right)} = -a\frac{\lambda i}{\sqrt{2}} + c\frac{\lambda i}{\sqrt{2}} = (c-a)\frac{\lambda i}{\sqrt{2}} = 0.$$

Since $\lambda \neq 0$, this means $a = c$. Finally, in order for the first column to have length 1 we need:

$$\begin{pmatrix} a \\ 0 \\ a \end{pmatrix} \cdot \begin{pmatrix} a \\ 0 \\ a \end{pmatrix} = a\overline{a} + a\overline{a} = 2a\overline{a} = 1,$$

so $a\overline{a} = \frac{1}{2}$, and hence $|a| = \frac{1}{\sqrt{2}}$. Thus we conclude the only unitary matrices of the required form are those which look like:

$$\begin{pmatrix} a & 0 & \pm i/\sqrt{2} \\ 0 & d & 0 \\ a & 0 & \mp i/\sqrt{2} \end{pmatrix} \text{ where } a, d \in \mathbb{C} \text{ satisfy } |a| = \frac{1}{\sqrt{2}} \text{ and } |d| = 1.$$

It is possible to work out more explicitly what any 3×3 unitary matrix must look like, but this is not at all an easy task; this is simpler to figure out in general in the 2×2 case.

Exercise. Determine what a 2×2 unitary matrix must look like.

Determinants. Determinants are numbers we associate to matrices, and depending on the way in which determinants are presented it may or not may not be apparent why we care about them. We'll approach determinants from a higher-level point of view than what most linear algebra courses (such as Math 290) would go into in order to get a sense as to why they are truly important.

But to start with, we should have some sense as to what a determinant actually is numerically before we start talking about more elaborate properties. So, first we'll say that determinants can be computed concretely using so-called *cofactor expansions*. We won't describe this procedure in these notes, but you can check the book or my Math 290-1 (first quarter!) lecture notes to see how this is done. Practicing the procedure a few times will likely be enough to understand how it works. A lingering question here is: why is it that cofactor expansion along any row or along any column always gives the same answer? We'll come back to this after we describe the determinant from another point of view.

Patterns and inversions. Carrying out a cofactor expansion in the 3×3 case gives:

$$\det \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} = a_{11}a_{22}a_{33} - a_{11}a_{23}a_{32} - a_{12}a_{21}a_{33} + a_{12}a_{23}a_{31} + a_{13}a_{21}a_{32} - a_{13}a_{22}a_{31}.$$

The key observation is that each term in this expression is a product consisting of exactly one entry from each row and column of the given matrix. We'll use the book's terminology and call such a collection a *pattern*, so a pattern is a choice of one entry from each row and column. In general, an $n \times n$ matrix will have $n!$ patterns. Note that it makes sense that any type of cofactor expansion should result in an expression made up of patterns: when we cross out the row or a column a given entry is in in a cofactor expansion, we are guaranteeing that that given entry will never be multiplied by another entry from the same row or column, which implies that we only multiply entries coming from different rows and columns.

What about the signs in the above expression? These can also be characterized using patterns in two ways: either using *inversions* or "swaps". An *inversion* of a pattern is any arrangement where an entry in that pattern is above and to the right of another entry. So, for instance for the pattern consisting of the entries

$$\begin{pmatrix} & & a_{13} \\ a_{21} & & \\ & a_{32} & \end{pmatrix},$$

one inversion comes from the a_{13} entry being above and to the right of a_{21}, and another comes from the a_{13} entry being above and to the right of a_{32}; the a_{21} and a_{32} entries themselves don't give an inversion since a_{21} is not to the right of a_{32} and a_{32} is not above a_{21}. The sign of the corresponding term in the cofactor expansion is then

$$(-1)^{\text{number of inversions}}.$$

As another example, the pattern

$$\begin{pmatrix} & & a_{13} \\ & a_{22} & \\ a_{31} & & \end{pmatrix}$$

has three inversions: a_{13} above and to the right of a_{22}, a_{13} above and to the right of a_{31}, and a_{22} above and to the right of a_{31}. Thus the coefficient of the corresponding term in the cofactor expansion is $(-1)^3 = 1$.

It turns out that this coefficient can also be found by determining how many times we have to swap columns or rows to put out pattern into the "standard" diagonal form:

$$\begin{pmatrix} * & & \\ & * & \\ & & * \end{pmatrix},$$

in that the coefficient needed is

$$(-1)^{\text{number of swaps}}.$$

For the first pattern given above, we need two swaps: swap the first and third column, and then the second and third column, while for the second pattern above we need only one swap: swap the first and third column. The first observation is that the number of swaps is NOT necessarily the same as the number of inversions, as we can see in the second example. In fact, the number

of swaps needed is not fixed, since say in the second pattern above we can also use three swaps: swap column one and two, then two and three, then one and two again. The AMAZING (almost amazingly awesome) fact is that although the number of swaps used might be different, we will always have either an even number or an odd number of them for a given pattern. Thus, we always have

$$(-1)^{\#\text{ inversions}} = (-1)^{\#\text{ swaps}}$$

regardless of which swaps we actually use.

Exercise. Show that $(-1)^{\#\text{ inversions}} = (-1)^{\#\text{ swaps}}$ for a given pattern, regardless of which swaps are actually used. This is actually quite challenging to show, so we'll just take it for granted. If you ever take an abstract algebra (or possibly combinatorics) course, you'll see (and likely prove) this fact in the course of giving the definition of what's called the *sign* of a *permutation*.

First definition of the determinant. We now give our first definition of the determinant. The point is that cofactor approach we started off with doesn't really serve as a good definition unless we show that the number we get is the same regardless of which row or column we expand along. This can be done (and you'll do it on a homework problem), but it is much cleaner to define the determinant in another way instead and then derive the cofactor expansion approach from this.

For an $n \times n$ matrix A, we define the *determinant* of A to be the value given by:

$$\det A = \sum_{\text{all patterns}} (-1)^{\#\text{ inversions}}(\text{product of terms in the pattern}),$$

where the sum is taken over all possible $n!$ many patterns. This is a "good" definition in that it doesn't depend on having to pick any specific row or column.

Now, why exactly this expression and not some other one? Whatever motivated someone to think that this sum was a good thing to consider? The answer is that, in fact, this expression doesn't just pop out of nowhere, but rather it can be *derived* from basic properties the determinant has. We'll take this approach next time, where we'll give an alternate definition of the determinant as an operation satisfying some simple properties, and use these to show that the pattern definition above is in fact the only thing which the determinant could possibly be.

Looking ahead. Some of you might have seen determinants in a previous linear algebra course, where I'm betting you saw how to compute them but didn't get a sense as to what they actually mean. The point is that the approach we're taking gets more to the point of their importance and usefulness, and will make it simpler to derive important properties determinants have. In particular, our approach will shed a good amount of light on the actual geometric interpretation of determinants, which absolutely exists.

Lecture 7: More on Determinants

Upper triangular. Before moving on, we note that the determinant of an upper triangular matrix is straightforward to compute. The claim is that

$$\det \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ & \ddots & \vdots \\ & & a_{nn} \end{pmatrix} = a_{11} \cdots a_{nn},$$

so the determinant of an upper triangular matrix is simply the product of its diagonal entries. (In this notation blanks denote entries which are zero.) Indeed, the only pattern which gives rise to a nonzero term in

$$\det A = \sum_{\text{all patterns}} (-1)^{\#\text{ inversions}} (\text{product of terms in the pattern}),$$

is the one consisting of $a_{11}, a_{22}, \dots, a_{nn}$ since any other pattern will contain at least one zero. Moreover, this pattern has no inversions, so the sum above is just

$$(-1)^0 a_{11} \cdots a_{nn} = a_{11} \cdots a_{nn}$$

as claimed. This fact will be the basis behind the method of computing determinants using row operations, which we'll soon talk about.

Warm-Up 1. We show that for any $n \times n$ matrix A, $\det A = \det A^T$, so a square matrix and its transpose always have the same determinant. The first observation is that A and A^T have precisely the same patterns. Indeed, when describing a pattern of A we pick exactly one entry from each row and column, which is the same as picking exactly one entry from each column and row of A^T. To be clear, if

$$a_{i_1 1},\ a_{i_2 2},\ \dots,\ a_{i_n n}$$

is a pattern of A, where $a_{i_k k}$ is the entry we pick from the k-th column and $i_1 \neq i_2 \neq \cdots \neq i_n$ since the rows these entries are in should be different, then this is also a pattern of A^T where $a_{i_k k}$ is now the entry we pick from the k-th row of A^T.

What is left to show is that the coefficient $(-1)^{\#\text{ inversions}}$ corresponding to a pattern is the same for A as it is for A^T. If in a given pattern

$$a_{i_1 1},\ a_{i_2 2},\ \dots,\ a_{i_n n}$$

of A, the entry $a_{i_k k}$ is above and to the right of $a_{i_\ell \ell}$, then in this same pattern of A^T the entry $a_{i_\ell \ell}$ will be above and to the right of $a_{i_k k}$, so that this pair of entries gives one inversion in each of A and A^T. This is true no matter which inversion of A we start with, so the number of inversions corresponding to this pattern in A is the same as the number of inversions corresponding to this pattern in A^T, and we conclude that $(-1)^{\#\text{ inversions}}$ is the same for A as it is for A^T. Thus the pattern/inversion formula for $\det A$ is *exactly* the same as the pattern/inversion formula for $\det A^T$, so these determinants are the same.

Warm-Up 2. Fix $\mathbf{v}_1, \dots, \widehat{\mathbf{v}_j}, \dots, \mathbf{v}_n \in \mathbb{R}^n$ (so there is no $\mathbf{v}_j$ term) and consider the function $T : \mathbb{R}^n \to \mathbb{R}$ defined by

$$T(\mathbf{x}) = \det \begin{pmatrix} \mathbf{v}_1 & \cdots & \mathbf{x} & \cdots & \mathbf{v}_n \end{pmatrix}.$$

We show that this is a linear transformation, which is what it means to say that the determinant is linear in each column. In other words, if in the j-th column we have a sum $\mathbf{x} + \mathbf{y}$ of two vectors, then

$$\det \begin{pmatrix} \mathbf{v}_1 & \cdots & \mathbf{x}+\mathbf{y} & \cdots & \mathbf{v}_n \end{pmatrix} = \det \begin{pmatrix} \mathbf{v}_1 & \cdots & \mathbf{x} & \cdots & \mathbf{v}_n \end{pmatrix} + \det \begin{pmatrix} \mathbf{v}_1 & \cdots & \mathbf{y} & \cdots & \mathbf{v}_n \end{pmatrix},$$

and if the j-th column is multiplied by a scalar c, we can pull the scalar out:

$$\det \begin{pmatrix} \mathbf{v}_1 & \cdots & c\mathbf{x} & \cdots & \mathbf{v}_n \end{pmatrix} = c \det \begin{pmatrix} \mathbf{v}_1 & \cdots & \mathbf{x} & \cdots & \mathbf{v}_n \end{pmatrix}.$$

In class we did this using cofactor expansions, but since we haven't yet show that cofactor expansion gives a valid way of computing the determinant, here we'll use the pattern/inversion definition.

Denote the entries of $\mathbf{x}$ by x_i, those of $\mathbf{y}$ by y_i, and the rest of the entries in the given matrix (the ones coming from the $\mathbf{v}$'s) by a_{ij}. Each pattern of $\begin{pmatrix}\mathbf{v}_1 & \cdots & \mathbf{x}+\mathbf{y} & \cdots & \mathbf{v}_n\end{pmatrix}$ will contain exactly one entry of the form $x_{i_j}+y_{i_j}$ from the j-th column (with corresponding row i_j), so the determinant is given by:

$$\begin{aligned}&\sum_{\text{patterns}}(-1)^{\#\text{ inversions}}(\text{product of terms in pattern})\\&=\sum_{i_1\neq i_2\neq\cdots\neq i_n}(-1)^{\#\text{ inversions}}[a_{i_1 1}a_{i_2 2}\cdots(x_{i_j}+y_{i_j})\cdots a_{i_n n}].\end{aligned}$$

As before, here $a_{i_k k}$ is coming from row i_k and column k, and the rows these come form are all different since $i_1\neq i_2\neq\cdots\neq i_n$. The term in brackets can be split up into

$$a_{i_1 1}\cdots(x_{i_j}+y_{i_j})\cdots a_{i_n n}=a_{i_1 1}\cdots x_{i_j}\cdots a_{i_n n}+a_{i_1 1}\cdots y_{i_j}\cdots a_{i_n n}.$$

With this the sum above can be split into

$$\begin{aligned}&\sum_{i_1\neq i_2\neq\cdots\neq i_n}(-1)^{\#\text{ inversions}}[a_{i_1 1}\cdots(x_{i_j}+y_{i_j})\cdots a_{i_n n}]\\&=\sum_{i_1\neq i_2\neq\cdots\neq i_n}(-1)^{\#\text{ inversions}}a_{i_1 1}\cdots x_{i_j}\cdots a_{i_n n}+\sum_{i_1\neq i_2\neq\cdots\neq i_n}(-1)^{\#\text{ inversions}}a_{i_1 1}\cdots y_{i_j}\cdots a_{i_n n}.\end{aligned}$$

Here the first term is $\det\begin{pmatrix}\mathbf{v}_1 & \cdots & \mathbf{x} & \cdots & \mathbf{v}_n\end{pmatrix}$ while the second is $\det\begin{pmatrix}\mathbf{v}_1 & \cdots & \mathbf{x}+\mathbf{y} & \cdots & \mathbf{v}_n\end{pmatrix}$, so we have $T(\mathbf{x}+\mathbf{y})=T(\mathbf{x})+T(\mathbf{y})$.

In a similar way, in each pattern of $\begin{pmatrix}\mathbf{v}_1 & \cdots & c\mathbf{x} & \cdots & \mathbf{v}_n\end{pmatrix}$ there will be exactly one entry of the form cx_{i_j}, so the determinant of this matrix is:

$$\sum_{i_1\neq\cdots\neq i_n}(-1)^{\#\text{ inversions}}a_{i_1 1}\cdots(cx_{i_j})\cdots a_{i_n n}=c\sum_{i_1\neq\cdots\neq i_n}(-1)^{\#\text{ inversions}}a_{i_1 1}\cdots x_{i_j}\cdots a_{i_n n}.$$

The final expression is $c\det\begin{pmatrix}\mathbf{v}_1 & \cdots & \mathbf{x} & \cdots & \mathbf{v}_n\end{pmatrix}$, so $T(c\mathbf{x})=cT(\mathbf{x})$ and we conclude that T is linear as claimed.

Alternating. We refer to the property that the determinant is linear in each column as saying that the determinant is *multilinear*. The fact that $\det A=\det A^T$ then implies that the determinant is also linear in each row. One other key property of the determinant is the fact that it is *alternating*, which means that swapping two columns changes the overall sign of the determinant. To be clear, if A' is the matrix obtained after swapping two columns of A, we claim that $\det A'=-\det A$.

Indeed, first notice that swapping columns does not affect the possible patterns. Say we swap columns i and j. Then, in a pattern of A we have one entry from column i, one from column j, and one from the other columns as well. But then, in A' we can get a corresponding pattern where the entry picked from column i of A is now picked from column j of A', the entry picked from column j of A is now picked from column j of A', and the other entries picked are the same in A as in A'. This shows that in the

$$\sum_{\text{patterns}}(-1)^{\#\text{ inversions}}(\text{product of terms in pattern})$$

formula, the "product of terms in a pattern" are the same for A as for A'. The point is that the coefficient changes if we think of it as

$$(-1)^{\#\text{ swaps}},$$

where "swaps" is the number of column swaps needed to put the pattern into standard diagonal form. This is due to the fact that if we require k column swaps in A to put the pattern into standard form, we require $k+1$ column swaps in A' since we first have to "undo" the column swap which produced A'. Thus, while the coefficient in the formula for $\det A$ is $(-1)^k$, for $\det A'$ it is

$$(-1)^{k+1} = -(-1)^k,$$

which gives that $\det A'$ is:

$$-\sum_{\text{patterns}} (-1)^{\#\text{ swaps in }A}(\text{product of terms in pattern}) = -\det A$$

as claimed. Moreover, since $\det A = \det A^T$, we immediately conclude that swapping two rows of a matrix also changes the sign of the determinant.

This alternating property implies that if two columns in a matrix are the same, then its determinant is zero. Indeed, if say columns i and j in A are the same, swapping these columns still gives A back, but we now have

$$\det A = -\det A$$

since this column swap must change the sign of the determinant. The only way this can be true is for $\det A = 0$, as claimed.

Second definition of the determinant. We can now give a second way of defining the determinant, where instead of giving a formula we give three properties which completely characterize it. The point is that from these properties alone we will derive the pattern/inversion formula, where the emphasis is one viewing these properties as the "true" reason why determinants are important. In other words, the numerical value the determinant gives often times isn't as important as the fact that it satisfies these properties.

The claim is that the determinant is the unique function $D : M_n(\mathbb{R}) \to \mathbb{R}$ satisfying the following the properties:

- D is multilinear, meaning linear in each column,
- D is alternating, meaning the swapping two columns changes the sign, and
- $D(I) = 1$, where I is the identity matrix.

The third property is required, since without it the function $D(A) = 0$ which sends every matrix to zero satisfies the first and second properties, but clearly isn't equal to the determinant function. We have shown before that the determinant has these properties, so the point now is arguing that any such D must in fact be the same as the determinant function.

Proof. Let $A \in M_n(\mathbb{R})$ and denote its entries by a_{ij}. We write each column of A as a linear combination of standard basis vectors, so that the j-th column is

$$\begin{pmatrix} a_{1j} \\ \vdots \\ a_{nj} \end{pmatrix} = a_{1j}\mathbf{e}_1 + \cdots + a_{nj}\mathbf{e}_n = \sum_{k=1}^{n} a_{kj}\mathbf{e}_k.$$

Thus we can write A column-by-column as

$$A = \begin{pmatrix} \sum_{i_1=1}^{n} a_{i_1 1}\mathbf{e}_{i_1} & \cdots & \sum_{i_n=1}^{n} a_{i_n n}\mathbf{e}_{i_n} \end{pmatrix}.$$

Now, since D is linear in each column, $D(A)$ can be broken up into:

$$D(A) = \sum_{i_1,\ldots,i_n} D\begin{pmatrix} a_{i_1 1}\mathbf{e}_{i_1} & \cdots & a_{i_n n}\mathbf{e}_{i_n} \end{pmatrix}$$

where the sum is taken over all possible values of $i_1, \ldots, i_n$, each ranging from 1 to n. For instance, in the 2×2 case we are saying that the result of applying D to

$$\begin{pmatrix} a_{11}\mathbf{e}_1 + a_{21}\mathbf{e}_2 & a_{12}\mathbf{e}_1 + a_{22}\mathbf{e}_2 \end{pmatrix}$$

can be broken up into a sum of four terms, where in each term we take one standard basis term from each column:

$$D\begin{pmatrix} a_{11}\mathbf{e}_1 & a_{12}\mathbf{e}_1 \end{pmatrix} + D\begin{pmatrix} a_{11}\mathbf{e}_1 & a_{22}\mathbf{e}_2 \end{pmatrix} + D\begin{pmatrix} a_{21}\mathbf{e}_2 & a_{12}\mathbf{e}_1 \end{pmatrix} + D\begin{pmatrix} a_{21}\mathbf{e}_1 & a_{22}\mathbf{e}_2 \end{pmatrix}.$$

Still using multilinearity, we can take each scalar we are multiplying a column by out of the resulting value, so

$$D(A) = \sum_{i_1,\ldots,i_n} a_{i_1 1} \cdots a_{i_n n} D\begin{pmatrix} \mathbf{e}_{i_1} & \cdots & \mathbf{e}_{i_n} \end{pmatrix}.$$

In the 2×2 case this looks like:

$$a_{11}a_{12}D\begin{pmatrix} \mathbf{e}_1 & \mathbf{e}_1 \end{pmatrix} + a_{11}a_{22}D\begin{pmatrix} \mathbf{e}_1 & \mathbf{e}_2 \end{pmatrix} + a_{21}a_{12}D\begin{pmatrix} \mathbf{e}_2 & \mathbf{e}_1 \end{pmatrix} + a_{21}a_{22}D\begin{pmatrix} \mathbf{e}_1 & \mathbf{e}_2 \end{pmatrix}.$$

We saw before that as a consequence of the alternating property, any matrix with repeated columns has determinant zero. Thus in the sum above,

$$D\begin{pmatrix} \mathbf{e}_{i_1} & \cdots & \mathbf{e}_{i_n} \end{pmatrix}$$

is zero whenever two of the columns are the same, so the only nonzero such expressions arise when the columns are all different, or in other words when $i_1 \neq \cdots \neq i_n$. Thus the sum reduces to

$$D(A) = \sum_{i_1 \neq \cdots \neq i_n} a_{i_1 1} \cdots a_{i_n n} D\begin{pmatrix} \mathbf{e}_{i_1} & \cdots & \mathbf{e}_{i_n} \end{pmatrix}.$$

But here the entries $a_{i_1 1}, \ldots, a_{i_n n}$ make up precisely one pattern of A! (That's an exclamation point, not a factorial.) Moreover, the matrix

$$\begin{pmatrix} \mathbf{e}_{i_1} & \cdots & \mathbf{e}_{i_n} \end{pmatrix}$$

to which D is being applied is then an identity matrix with columns swapped around; to be precise, the 1's are in the locations corresponding to the pattern entries. Since swapping columns changes the sign of D, we get

$$D\begin{pmatrix} \mathbf{e}_{i_1} & \cdots & \mathbf{e}_{i_n} \end{pmatrix} = (-1)^{\#\text{ swaps}} D(I),$$

where "swaps" is the number of swaps needed to put the pattern into standard diagonal form. Since $D(I) = 1$, we finally get

$$D(A) = \sum_{i_1 \neq \cdots \neq i_n} (-1)^{\#\text{ swaps}} a_{i_1 1} \cdots a_{i_n n}.$$

The point is that this expression is precisely the pattern/inversion formula for $\det A$, so we conclude that $D(A) = \det A$ as claimed.

To emphasize once more: from the three properties given above (multilinearity, alternating, and sending I to 1), we are in fact able to *derive* the usual formula for the determinant. This approach is closer to the way in which determinants were historically developed. □

Row operations. With this at hand, we can now explain how row operations affect determinants, which we'll see next time gives a more efficient way of computing determinants than does cofactor expansion. We use that the fact determinants are linear in each row and alternating in the rows.

We have:

- swapping two rows changes the sign of the determinant,
- multiplying a row a nonzero scalar scales the determinant by that same scalar, and
- adding a multiple of one row to another does not affect the determinant.

The first property is simply the alternating property. The second property says that if A' is obtained from A by scaling a row by $c \neq 0$, then $\det A' = c \det A$, which comes from the fact that the determinant is linear in each row. Finally, say that A' is obtained from A by adding c times row i to row j:

$$A = \begin{pmatrix} \vdots \\ \mathbf{r}_i \\ \vdots \\ \mathbf{r}_j \\ \vdots \end{pmatrix} \to A' = \begin{pmatrix} \vdots \\ \mathbf{r}_i \\ \vdots \\ c\mathbf{r}_i + \mathbf{r}_j \\ \vdots \end{pmatrix}.$$

Using linearity in the j-th row, we have that

$$\det \begin{pmatrix} \vdots \\ \mathbf{r}_i \\ \vdots \\ c\mathbf{r}_i + \mathbf{r}_j \\ \vdots \end{pmatrix} = c \det \begin{pmatrix} \vdots \\ \mathbf{r}_i \\ \vdots \\ \mathbf{r}_i \\ \vdots \end{pmatrix} + \det \begin{pmatrix} \vdots \\ \mathbf{r}_i \\ \vdots \\ \mathbf{r}_j \\ \vdots \end{pmatrix}.$$

The first term here has a matrix with repeated rows, so the alternating property implies that this first term is zero, so

$$\det \begin{pmatrix} \vdots \\ \mathbf{r}_i \\ \vdots \\ c\mathbf{r}_i + \mathbf{r}_j \\ \vdots \end{pmatrix} = \det \begin{pmatrix} \vdots \\ \mathbf{r}_i \\ \vdots \\ \mathbf{r}_j \\ \vdots \end{pmatrix},$$

which gives $\det A = \det A'$ as claimed.

The idea behind using row operations to compute determinants is: use row operations to transform our matrix into upper triangular form, and use the properties above to keep track of how these row operations affect the determinant in order to relate the determinant of the original matrix to the determinant of the reduced upper triangular form, which has an easy determinant to compute. We'll see an example or two next time.

Lecture 8: Determinants and Products

Warm-Up 1. We look at one example of computing a determinant using row operations. Let

$$A = \begin{pmatrix} 3 & 4 & -1 & 2 \\ 3 & 0 & 1 & 5 \\ 0 & -2 & 1 & 0 \\ -1 & -3 & 2 & 1 \end{pmatrix}.$$

Performing the row operations: swap I and IV, $3I+II$, $3I+IV$, $\frac{1}{5}IV$, swap II and IV, $-2II+III$, $-9II+IV$, and $-2III+IV$, reduces A to

$$U = \begin{pmatrix} -1 & -3 & 2 & 1 \\ 0 & -1 & 1 & 1 \\ 0 & 0 & -1 & -2 \\ 0 & 0 & 0 & 3 \end{pmatrix}.$$

The only row operations used which affect the value of the determinant were the two swaps and the scaling by $\frac{1}{5}$, so we get that

$$\det U = (-1)(-1)\frac{1}{5}\det A.$$

Hence $\det A = -5\det U = -5(-1)^3 = -15$, where we use the fact that U is upper triangular to say that its determinant is the product of its diagonal entries.

In general, computing determinants using row operations is much more efficient than by using patterns or cofactor expansions, *except* for when using determinants to compute eigenvalues, which we'll look at next week.

Warm-Up 2. Suppose A is an invertible matrix. We show that $\det A^{-1} = \frac{1}{\det A}$. Now, we will soon see that for any matrices A and B, we have

$$\det(AB) = (\det A)(\det B).$$

Applying this to $\det(AA^{-1}) = \det I = 1$ gives us the claimed equality, but the point is that here we want to avoid using this fact, and will use row operations instead. Indeed, the proof we'll give here using row operations will actually (with slight modifications) lead to a proof that $\det(AB) = (\det A)(\det B)$, which is why we're looking at this Warm-Up first.

The key is in recalling that A^{-1} can be explicitly computed using row operations by reducing:

$$\begin{pmatrix} A & I \end{pmatrix} \to \begin{pmatrix} I & A^{-1} \end{pmatrix}.$$

Consider the row operations which turn A into I. Of these, only row swaps and scalings of rows by nonzero values affect determinants. Say that in the process of reducing A to I we perform n row swaps and scale some rows by the nonzero values $c_1, \dots, c_k$. Then we get

$$\det I = (-1)^n c_1 \cdots c_k \det A, \text{ so } \det A = \frac{1}{(-1)^n c_1 \cdots c_k}.$$

Now, the same operations transform I into A^{-1}, so we also get

$$\det(A^{-1}) = (-1)^n c_1 \cdots c_k \det I = (-1)^n c_1 \cdots c_k,$$

and hence $\det A^{-1} = \frac{1}{\det A}$ as claimed.

The determinant of a product. Let A and B now be square matrices of the same size. If A is not invertible, then AB is not invertible so $\det(AB) = 0$ and $\det A = 0$, meaning that $\det(AB) = (\det A)(\det B)$ is true since both sides are zero.

Now suppose A is invertible and consider reducing the augmented matrix

$$\begin{pmatrix} A & AB \end{pmatrix} \to \begin{pmatrix} I & ? \end{pmatrix}.$$

Viewing the process of performing row operations as multiplying on the left by elementary matrices, if $E_1, \ldots, E_t$ are the elementary matrices satisfying

$$E_t \cdots E_1 A = I,$$

then

$$E_t \cdots E_1(AB) = (E_t \cdots E_1 A)B = B.$$

Thus the unknown ? above is B, meaning that the row operations transforming A into I have the following effect:

$$\begin{pmatrix} A & AB \end{pmatrix} \to \begin{pmatrix} I & B \end{pmatrix}.$$

Using the same notation as in the second Warm-up, we have

$$\det I = (-1)^n c_1 \cdots c_k \det A, \text{ so } \det A = \frac{1}{(-1)^n c_1 \cdots c_k}.$$

But these same operations transform AB into B, so

$$\det B = (-1)^n c_1 \cdots c_k \det(AB) = \frac{1}{\det A} \det AB,$$

and thus $\det(AB) = (\det A)(\det B)$. (Note that $\frac{1}{\det A}$ makes sense since $\det A \neq 0$ given that A is invertible.)

Alternate proof. The proof that $\det(AB) = (\det A)(\det B)$ given above is the one in the book, but let us give another proof of this fact, this time based on the characterization of the determinant as the unique multilinear, alternating, map $M_n(\mathbb{K}) \to \mathbb{K}$ which sends I to 1. The case where A is not invertible is the same as above, so let us assume that A is invertible.

Define the function $D : M_n(\mathbb{K}) \to \mathbb{K}$ by

$$D(B) = \frac{\det(AB)}{\det A}.$$

First, we have

$$D(I) = \frac{\det(AI)}{\det A} = \frac{\det A}{\det A} = 1.$$

Next we check multilinearity in each column. Suppose that the j-th column of B is written as $\mathbf{x}+\mathbf{y}$ where $\mathbf{x}, \mathbf{y} \in \mathbb{K}^n$. Then AB has columns:

$$A\begin{pmatrix} \mathbf{b}_1 & \cdots & \mathbf{x}+\mathbf{y} & \cdots & \mathbf{b}_n \end{pmatrix} = \begin{pmatrix} A\mathbf{b}_1 & \cdots & A\mathbf{x}+A\mathbf{y} & \cdots & A\mathbf{b}_n \end{pmatrix}.$$

Since the determinant is linear in each column, we have

$$\det(AB) = \det\begin{pmatrix} A\mathbf{b}_1 & \cdots & A\mathbf{x} & \cdots & A\mathbf{b}_n \end{pmatrix} + \det\begin{pmatrix} A\mathbf{b}_1 & \cdots & A\mathbf{y} & \cdots & A\mathbf{b}_n \end{pmatrix}$$

$$= \det(AB_1) + \det(AB_2)$$

where B_1 is the same as B only with $\mathbf{x}$ as the j-th column and B_2 is B with $\mathbf{y}$ as the j-th column. Thus

$$D(B) = \frac{\det(AB)}{\det A} = \frac{\det(AB_1)}{\det A} + \frac{\det(AB_2)}{\det A} = D(B_1) + D(B_2).$$

If B has $r\mathbf{x}$ as its j-th column for some $r \in \mathbb{K}$, then

$$AB = \begin{pmatrix} A\mathbf{b}_1 & \cdots & rA\mathbf{x} & \cdots & A\mathbf{b}_n \end{pmatrix}.$$

Thus

$$\det(AB) = r \det(AB_1)$$

where B_1 has $\mathbf{x}$ alone as the j-th column, so

$$D(B) = \frac{\det(AB)}{\det A} = r\frac{\det(AB_1)}{\det A} = rD(B_1).$$

Hence D is linear in each column.

Finally we show that D is alternating. Switching the i and j-th columns of B also switches the i and j-th columns of

$$AB = \begin{pmatrix} A\mathbf{b}_1 & \cdots & A\mathbf{b}_n \end{pmatrix}.$$

Since the determinant is alternating, we have that this column swap changes the sign of $\det(AB)$, and hence the sign of $D(B)$. Thus D is alternating, and we conclude that D must be the same as the determinant function since the determinant is the only multilinear, alternating function sending I to 1. Thus

$$\det B = D(B) = \frac{\det(AB)}{\det A},$$

which implies $\det(AB) = (\det A)(\det B)$ as claimed.

Determinants of orthogonal matrices. Since an orthogonal matrix Q satisfies $QQ^T = I$, we get

$$1 = \det I = \det(QQ^T) = (\det Q)(\det Q^T) = (\det Q)^2.$$

We conclude that the determinant of an orthogonal matrix is $\det Q = \pm 1$. In fact, those with $\det Q = 1$ are rotations and those with $\det Q = -1$ are reflections, which comes from the geometric interpretation of the sign of the determinant, which we'll look at next time.

Determinants of linear transformations. So far we have only spoken about determinants of matrices, but now with the formula for the determinant of a product we can give meaning to the determinant of an arbitrary linear transformation from a finite-dimensional vector space to itself.

Suppose that V is a finite-dimensional and that $T : V \to \mathbf{v}$ is a linear transformation. Pick a basis $\mathcal{B}$ of V, and let $[T]_{\mathcal{B}}$ denote the matrix of T relative to $\mathcal{B}$. We define the *determinant* of T to be the determinant of this matrix:

$$\det T := \det[T]_{\mathcal{B}}.$$

In order for this to make sense, we have to know that regardless of what basis is chosen, the determinant of $[T]_{\mathcal{B}}$ remains the same. In other words, if $\mathcal{B}'$ is another basis of V, $[T]_{\mathcal{B}'}$ might be different from $[T]_{\mathcal{B}}$, but nonetheless we have

$$\det[T]_{\mathcal{B}} = \det[T]_{\mathcal{B}'}.$$

To see this, recall that these two matrices are related by an invertible matrix S via:

$$[T]_{\mathcal{B}} = S[T]_{\mathcal{B}'}S^{-1}.$$

Concretely, S is the so-called "change of basis" matrix from the basis $\mathcal{B}'$ to the basis $\mathcal{B}$. (Check Chapter 4 in the book or the lecture notes from reading week of last quarter to review these facts.) Using properties of the determinant we've seen, this gives

$$\begin{aligned}\det[T]_{\mathcal{B}} &= (\det S)(\det[T]_{\mathcal{B}'})(\det S^{-1}) \\ &= (\det S)(\det[T]_{\mathcal{B}'})(\det S)^{-1} \\ &= \det[T]_{\mathcal{B}'}\end{aligned}$$

as claimed, so $\det T$ is well-defined and doesn't depend on the specific basis chosen.

Lecture 9: The Geometry of Determinants

Warm-Up. We compute the determinant of the linear transformation $T : M_n(\mathbb{K}) \to M_n(\mathbb{K})$ defined by $T(A) = A^T$. We pick as our basis $\mathcal{B}$ of $M_n(\mathbb{K})$ the one consisting of the matrices E_{ij}, where E_{ij} has a 1 in the ij-th entry and zeroes elsewhere. To be clear, we list our basis elements as

$$E_{11}, \dots, E_{1n}, E_{21}, \dots, E_{2n}, \dots, E_{n1}, \dots, E_{nn}.$$

To compute the matrix of T relative to this basis we need to determine the coordinate vector of each output $T(E_{ij})$, which as we recall encodes the coefficients needed to express each of these outputs as a linear combination of the given basis elements. We have $T(E_{ij}) = E_{ij}^T = E_{ji}$, and thus the coordinate vector of $T(E_{ij})$ has a single 1 in some location and zeroes elsewhere. For instance, in the $n = 2$ case we have

$$T(E_{11}) = E_{11},\ T(E_{12}) = E_{21},\ T(E_{21}) = E_{12},\ T(E_{22}) = E_{22},$$

so the matrix of T is

$$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

In the $n = 3$ case, the matrix of T is:

$$\begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

Note the pattern in each: the column corresponding to $T(E_{ij})$ has a 1 in the entry corresponding to the location at which E_{ji} occurs in our list of basis elements.

The determinant of $[T]_{\mathcal{B}}$ can now be computed by determining how many column swaps are needed to turn this matrix into the $n^2 \times n^2$ identity matrix. The columns corresponding to each $T(E_{ii})$ are already in the correct location they should be in for the identity matrix, and swapping the columns corresponding to $T(E_{ij})$ and $T(E_{ji})$ for $i \neq j$ puts the 1's in these columns in the correct locations they should be in for the identity since $T(E_{ij}) = E_{ji}$ and $T(E_{ji}) = E_{ij}$ for $i \neq j$. Thus all together there are $n^2 - n = n(n-1)$ columns which need to be swapped (i.e. the total number of columns minus those not corresponding to some $T(E_{ii})$), so the total number of swaps needed is

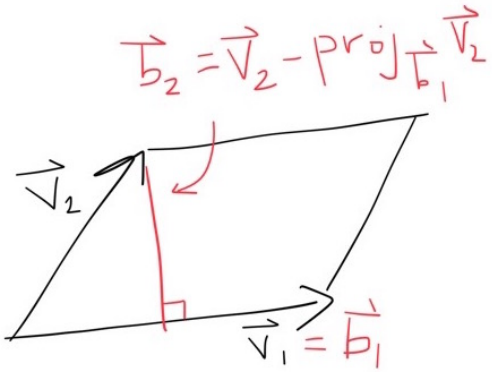
$$\frac{n(n-1)}{2}$$

since two columns are swapped at a time. Thus we conclude that

$$\det T = (-1)^{n(n-1)/2}.$$

You can verify that this is true in the $n = 2$ and $n = 3$ cases above, where when $n = 2$ we need only 1 swap while when $n = 3$ we need 3 swaps.

Area. Consider the parallelogram in $\mathbb{R}^2$ having $\mathbf{v}_1, \mathbf{v}_2$ as edges:



The area of this parallelogram is the length of the base $\mathbf{v}_1$ times the "height", which is the length of the perpendicular vector $\mathbf{b}_2$ drawn above. The main observation is that $\mathbf{b}_2$ is the difference

$$\mathbf{b}_2 = \mathbf{v}_2 - \text{proj}_{\mathbf{b}_1} \mathbf{v}_2$$

where $\mathbf{b}_1 = \mathbf{v}_2$, so the area is $\|\mathbf{b}_1\| \, \|\mathbf{b}_2\|$. Note that these vectors $\mathbf{b}_1, \mathbf{b}_2$ are precisely the vectors resulting from applying the Gram-Schmidt process to $\mathbf{v}_1, \mathbf{v}_2$ (before we divide by lengths to get unit vectors), so the conclusion is that in $\mathbb{R}^2$:

$$\text{area} = \text{ products of lengths of orthogonal vectors arising from Gram-Schmidt.}$$

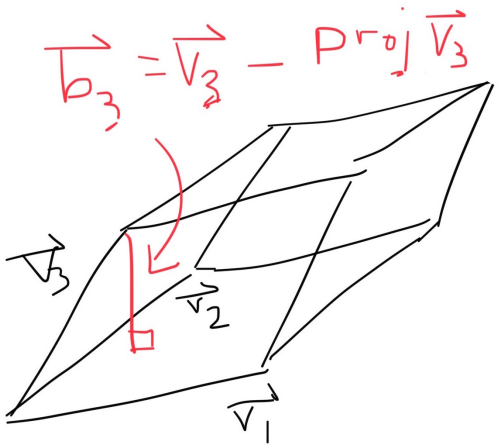
Parallelopipeds. The same is true in higher dimensions, but first we need a definition for the correct generalization of "parallelogram". The key is in recognizing that the parallelogram with edges $\mathbf{v}_1, \mathbf{v}_2$ exactly consists of linear combinations of the form:

$$c_1\mathbf{v}_1 + c_2\mathbf{v}_2 \text{ where } 0 \leq c_1 \leq 1 \text{ and } 0 \leq c_2 \leq 1.$$

We define the *parallelopiped* determined by $\mathbf{v}_1, \dots, \mathbf{v}_n \in \mathbb{R}^n$ to be the object described by

$$\{c_1\mathbf{v}_1 + \cdots + c_n\mathbf{v}_n \mid 0 \leq c_i \leq 1 \text{ for all } i\}.$$

In two dimensions this gives an ordinary parallelogram. In three dimensions it gives a "slanted rectangular box" with base a parallelogram:



Note that in the three-dimensional case the volume of this parallelopiped is given by

$$\text{(area of base parallelogram)(height)}.$$

As in the parallelogram case, the height can be computed by taking the length of

$$\mathbf{b}_3 = \mathbf{v}_3 - \text{proj}_{\text{base}}\,\mathbf{v}_3,$$

which is precisely the third vector appearing in the Gram-Schmidt process applied to $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3$. Thus we get that the desired volume is

$$\|\mathbf{b}_1\|\,\|\mathbf{b}_2\|\,\|\mathbf{b}_3\|$$

since $\|\mathbf{b}_1\|\,\|\mathbf{b}_2\|$ is the area of the base parallelogram by what we did before.

In general, the *volume* of the parallelopiped determined by $\mathbf{v}_1, \ldots, \mathbf{v}_n \in \mathbb{R}^n$ is given by

$$\|\mathbf{b}_1\|\,\|\mathbf{b}_2\| \cdots \|\mathbf{b}_n\|$$

where $\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_n$ are the orthogonal vectors

$$\begin{aligned}
\mathbf{b}_1 &= \mathbf{v}_1 \\
\mathbf{b}_2 &= \mathbf{v}_2 - \text{proj}_{\mathbf{b}_1}\,\mathbf{v}_2 \\
&\vdots \\
\mathbf{b}_n &= \mathbf{v}_n - \text{proj}_{\mathbf{b}_1}\,\mathbf{v}_n - \cdots - \text{proj}_{\mathbf{b}_{n-1}}\,\mathbf{v}_n
\end{aligned}$$

arising in the Gram-Schmidt process before normalization.

|Determinant| = volume. We can now give the first geometric interpretation of the determinant. As we have seen, any square matrix A has a factorization

$$A = QR$$

where Q is orthogonal and R upper triangular with positive diagonal entries. To be precise, if $\mathbf{v}_1, \ldots, \mathbf{v}_n$ are the columns of A, then the diagonal entries of R are the lengths $\|\mathbf{b}_1\|, \ldots, \|\mathbf{b}_n\|$ of the vectors resulting from Gram-Schmidt. Since the determinant of an orthogonal matrix is ± 1, we get

$$|\det A| = |(\det Q)(\det R)| = |\det Q||\det R| = 1|\det R| = \|\mathbf{b}_1\| \cdots \|\mathbf{b}_n\|,$$

so the conclusion is that $|\det A|$ is precisely the volume of the parallelopiped determined by the columns of A!

Expansion factors. There is another geometric interpretation of determinants which follows from the one above, and is in many ways more important. Indeed, this is the interpretation we'll see come up when we discuss multivariable integration next quarter.

Consider the linear transformation $\mathbf{x} \mapsto A\mathbf{x}$ determined by a square matrix A. Let P be a parallelopiped in $\mathbb{R}^n$, determined by vectors $\mathbf{v}_1, \dots, \mathbf{v}_n$. The *image* of P under the transformation A is the set of all points obtained by applying A to points of P, and it turns out that this image is precisely the parallelopiped determined by the vectors

$$A\mathbf{v}_1, \dots, A\mathbf{v}_n.$$

This comes from the linearity of A: if $c_1\mathbf{v}_! + \cdots + c_n\mathbf{v}_n$ with $0 \leq c_i \leq 1$ for each i, then applying A to this point gives

$$c_1(A\mathbf{v}_1) + \cdots + c_n(A\mathbf{v}_n),$$

which is a point in the parallelopiped with edges $A\mathbf{v}_1, \dots, A\mathbf{v}_n$. By the geometric interpretation of determinants give above, we have:

$$\begin{aligned} \operatorname{Vol} A(P) &= |\det \begin{pmatrix} A\mathbf{v}_1 & \cdots & A\mathbf{v}_n \end{pmatrix}| \\ &= |\det(A \begin{pmatrix} \mathbf{v}_1 & \cdots & \mathbf{v}_n \end{pmatrix})| \\ &= |\det A||\det \begin{pmatrix} \mathbf{v}_1 & \cdots & \mathbf{v}_n \end{pmatrix}| \\ &= |\det A| \operatorname{Vol} P. \end{aligned}$$

This says that $|\det A|$ is the factor by which volumes change under the transformation determined by A, and we say that $|\det A|$ is the *expansion factor* of this transformation. (Whether A literally "expands" volumes depends on how large $|\det A|$ is: if $|\det A| < 1$, A contracts volumes; if $|\det A| = 1$, A leaves volumes unchanged; and if $|\det A| > 1$, A indeed expands volumes. Nonetheless, we refer to $|\det A|$ as an "expansion" factor regardless.)

In fact, this is true not only for parallelograms, but for *any* regions in $\mathbb{R}^n$, at least ones for which the notion of "volume" makes sense. That is, if Ω is any "nice" region in $\mathbb{R}^n$, and $A(\Omega)$ denotes its image under A, we have

$$\operatorname{Vol} A(\Omega) = |\det A| \operatorname{Vol} \Omega.$$

We'll actually prove this next quarter, but the idea is simple: approximate Ω using parallelopipeds, apply the expansion property we derived above to each of these parallelopipeds, and then take a "limit" as the parallelopipeds we use "better and better" approximate Ω. Note that surprising fact that the volume of a region, no matter what it looks like, is always expanded/contracted by the same amount, namely $|\det A|$!

Orientations. The two facts above thus give a meaning to the value of $|\det A|$. What remains is to give a meaning to the sign of $\det A$, or in other words, to understand what is different between matrices of positive determinant and those of negative determinant. This was touched upon in the problems from the second discussion section (see Discussion 2 Problems on canvas), and is discussed in Problem 5 of Homework 3. The answer is that when $\det A$ is positive, A is *orientation-preserving*, while when $\det A$ is negative, A is *orientation-reversing*. See the problems mentioned above for a further discussion of what this means. We'll talk more about orientations when we discuss integration next quarter.

Lecture 10: Eigenvalues and Eigenvectors

Warm-Up. We justify the fact that $\det(AB) = (\det A)(\det B)$, using only the geometric interpretations of determinants we derived last time. In other words, if we were define the determinant via these geometric interpretations, we show that we can prove the product formula above.

In particular, we use the fact that determinants give expansion factors. Consider the linear transformations $\mathbb{R}^n \to \mathbb{R}^n$ defined by A and B, and let P be a parallelopiped in $\mathbb{R}^n$. Applying B to P gives a parallelopiped $B(P)$ whose volume is given by

$$\operatorname{Vol} B(P) = |\det B| \operatorname{Vol} P.$$

Now applying A to $B(P)$ gives a parallelopiped $A(B(P))$ whose volume is

$$\operatorname{Vol} A(B(P)) = |\det A| \operatorname{Vol} B(P),$$

which using the first equation above gives

$$\operatorname{Vol} A(B(P)) = |\det A||\det B| \operatorname{Vol} P.$$

On the other hand, $A(B(P))$ can also be thought of as what we get when we apply the composition AB to P, so its volume is

$$\operatorname{Vol} A(B(P)) = |\det AB| \operatorname{Vol} P.$$

Comparing these last two equations gives $|\det AB| = |\det A||\det B|$.

Now, to get rid of the absolute values we consider orientations. If A and B are both orientation-preserving, then AB is also orientation-preserving and hence in this case all absolute values above are positive, giving $\det(AB) = (\det A)(\det B)$. If one of A or B is orientation-preserving and the other orientation-reversing, then AB will also be orientation reversing, so one of $\det A$ or $\det B$ is positive and the other negative and $\det(AB)$ is negative. Hence $\det(AB) = (\det A)(\det B)$ in this case as well. Finally, if A and B both reverse orientation, then AB preserves orientation (since the orientation is reversed twice), so $\det A$ and $\det B$ are negative while $\det(AB)$ is positive, so $\det(AB) = (\det A)(\det B)$ in this case as well. Thus this product formula is true in general.

In fact, note that all three defining properties of the determinant—that it is multilinear, alternating, and sends I to 1—can be justified from geometry alone. The fact that $\det I = 1$ comes from the fact that the parallelopiped determined by the columns of I is a cube of volume 1, and the fact that changing columns changes sign comes from the fact that changing columns in a sense switches orientation. Finally, the fact that

$$\det\begin{pmatrix}\cdots & r\mathbf{x} & \cdots\end{pmatrix} = r\det\begin{pmatrix}\cdots & \mathbf{x} & \cdots\end{pmatrix}$$

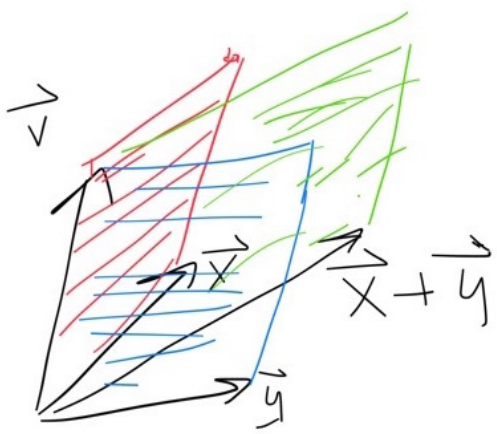
comes from the fact that scaling one edge of a parallelopiped ends up scaling the volume by that same amount, and the fact that

$$\det\begin{pmatrix}\cdots & \mathbf{x}+\mathbf{y} & \cdots\end{pmatrix} = \det\begin{pmatrix}\cdots & \mathbf{x} & \cdots\end{pmatrix} + \det\begin{pmatrix}\cdots & \mathbf{y} & \cdots\end{pmatrix}$$

comes from the fact that the parallelopipeds determined by the two matrices on the right can be split up and rearranged to fill out the parallelopiped determined by the matrix on the left. For instance, in the $n = 2$ case, this looks like

$$\det\begin{pmatrix}\mathbf{x}+\mathbf{y} & \mathbf{v}\end{pmatrix} = \det\begin{pmatrix}\mathbf{x} & \mathbf{v}\end{pmatrix} + \det\begin{pmatrix}\mathbf{y} & \mathbf{v}\end{pmatrix}.$$

The various columns give the parallelograms:



and it is a true fact that the red and blue parallelograms can be cut up and rearranged to give the green parallelogram, so the area of the green one is the sum of the areas of the red and blue ones. (This is a fun thing to try to show.)

Diagonalizable matrices. Recall that a square matrix A is *diagonalizable* if it is similar to a diagonal matrix, which means that there exists an invertible matrix S and a diagonal matrix D such that

$$A = SDS^{-1}.$$

Letting T denote the linear transformation determined by A, recall that this can be interpreted as saying that there exists a basis $\mathcal{B} = \{\mathbf{v}_!, \dots, \mathbf{v}_n\}$ of $\mathbb{R}^n$ such that the matrix of T relative to this basis is diagonal:

$$[T]_{\mathcal{B}} = \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{pmatrix}.$$

Since the columns of this matrix describe the coefficients needed to express each output $T(\mathbf{v}_i)$ as a linear combination of the basis vectors in $\mathcal{B}$, in order to get this form for $[T]_{\mathcal{B}}$ is must be true that each basis vector satisfies

$$T(\mathbf{v}_i) = \lambda_i \mathbf{v}_i.$$

Again, check the end of the lecture notes from last quarter to review all this.

Eigenvalues and eigenvectors. The upshot is that vectors satisfying an equation of the form $T(v) = \lambda v$ have special properties, so we give them and the scalars λ involved special names.

Let $T : V \to V$ be a linear transformation, where V is a vector space over $\mathbb{K}$. We say that a scalar $\lambda \in \mathbb{K}$ is an *eigenvalue* of T if there exists a nonzero vector $v \in V$ such that

$$T(v) = \lambda v.$$

(Note that we require $v \neq 0$ since any scalar λ satisfies $T(0) = \lambda 0$.) We then call v an *eigenvector* corresponding to λ. Hence, eigenvectors are vectors with the property that applying T has the effect of scaling them. The intuition is that eigenvectors describe particularly "nice" directions associated to T, in that they give directions in which the action of T is simple to describe.

Amazingly Awesome. Note what it means for 0 to be an eigenvalue of T. This is true if and only if there is a nonzero vector $v \in V$ such that

$$T(v) = 0v = 0,$$

and such a v is thus in $\ker T$. Hence 0 is an eigenvalue of T if and only if there is a nonzero vector in $\ker T$, which is true if and only if T is not injective. When V is finite-dimensional, this is equivalent to saying that T is not invertible, so the conclusion is the following addition to the Amazingly Awesome Theorem: a square matrix is invertible if and only if 0 is not an eigenvalue.

Example 1. The matrix

$$A = \begin{pmatrix} 13 & -6 \\ -1 & 12 \end{pmatrix}$$

is one we looked at on the final day of last quarter. The point was that relative to the basis

$$\begin{pmatrix} 2 \\ 1 \end{pmatrix}, \begin{pmatrix} -3 \\ 1 \end{pmatrix},$$

the matrix of A becomes

$$\begin{pmatrix} 10 & 0 \\ 0 & 15 \end{pmatrix}.$$

This is because 10 and 15 are eigenvalues of A with eigenvectors $\left(\begin{smallmatrix} 2 \\ 1 \end{smallmatrix}\right)$ and $\left(\begin{smallmatrix} -3 \\ 1 \end{smallmatrix}\right)$ respectively, as we can verify:

$$\begin{pmatrix} 13 & -6 \\ -1 & 12 \end{pmatrix} \begin{pmatrix} 2 \\ 1 \end{pmatrix} = \begin{pmatrix} 20 \\ 10 \end{pmatrix} = 10 \begin{pmatrix} 2 \\ 1 \end{pmatrix}$$
$$\begin{pmatrix} 13 & -6 \\ -1 & 12 \end{pmatrix} \begin{pmatrix} -3 \\ 1 \end{pmatrix} = \begin{pmatrix} -45 \\ 15 \end{pmatrix} = 15 \begin{pmatrix} -3 \\ 1 \end{pmatrix}.$$

Example 2. The matrix

$$\begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$$

has no real eigenvalues. Indeed, this matrix represents rotations by $\pi/2$, and there is no nonzero (real) vector with the property that rotating it by $\pi/2$ results in a multiple of that same vector.

However, this matrix does have *complex* eigenvalues. Indeed, we have:

$$\begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} -1 \\ i \end{pmatrix} = \begin{pmatrix} -i \\ -1 \end{pmatrix} = i \begin{pmatrix} -1 \\ i \end{pmatrix},$$

so i is an eigenvalue, and

$$\begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} -1 \\ -i \end{pmatrix} = \begin{pmatrix} i \\ -1 \end{pmatrix} = -i \begin{pmatrix} -1 \\ -i \end{pmatrix},$$

so $-i$ is also an eigenvalue. The point is that the scalars being used matters when discussing the possible eigenvalues of a transformation.

Example 3. Let V denote the space of infinitely-differentiable functions from $\mathbb{R}$ to $\mathbb{R}$, and let $D : V \to V$ be the linear transformation which sends a function to its derivative:

$$D(f) = f'.$$

Any scalar $\lambda \in \mathbb{R}$ is an eigenvalue of D, since

$$D(e^{\lambda x}) = \lambda e^{\lambda x},$$

so functions of the form $e^{\lambda x}$ are eigenvectors of D.

Example 4. Let $R : \mathbb{K}^\infty \to \mathbb{K}^\infty$ be the right-shift map:

$$R(x_1, x_2, x_3, \ldots) = (0, x_1, x_2, \ldots).$$

An eigenvector of R would satisfy

$$R(x_1, x_2, x_3, \ldots) = (0, x_1, x_2, \ldots) = \lambda(x_1, x_2, x_3, \ldots) = (\lambda x_1, \lambda x_2, \lambda x_3, \ldots)$$

for some $\lambda \in \mathbb{K}$. Comparing entries shows that $\lambda x_1 = 0$, so either $\lambda = 0$ or $x_1 = 0$. Either way, comparing the rest of the entries implies that

$$x_1 = x_2 = x_3 = \cdots = 0,$$

so the only element of $\mathbb{K}^\infty$ which is sent to a multiple of itself is the zero vector, so R has no eigenvectors and hence no eigenvalues. (In particular, the fact that 0 is not an eigenvalue reflects the fact that R is injective.)

Example 5. Consider now the left-shift map $L : \mathbb{K}^\infty \to \mathbb{K}^\infty$ defined by

$$L(x_1, x_2, x_3, \ldots) = (x_2, x_3, x_4, \ldots).$$

An eigenvector of L must satisfy

$$L(x_1, x_2, x_3, \ldots) = (x_2, x_3, x_4, \ldots) = \lambda(x_1, x_2, x_3, \ldots)$$

for some $\lambda \in \mathbb{K}$. For instance,

$$L(1, 1, 1, \ldots) = (1, 1, 1, \ldots) = 1(1, 1, 1, \ldots),$$

so 1 is an eigenvalue with eigenvector $(1, 1, 1, \ldots)$. More generally, any element of $\mathbb{K}^\infty$ where all entries are the same (and nonzero) is an eigenvector with eigenvalue 1.

But L has other eigenvalues as well. For instance the vector $(2^0, 2^1, 2^2, \ldots)$ whose entries are powers of 2 satisfies:

$$L(1, 2, 4, \ldots) = (2, 4, 8, \ldots) = 2(1, 2, 4, \ldots),$$

so 2 is an eigenvalue with eigenvector $(1, 2, 4, \ldots)$. In fact, any scalar $\lambda \in \mathbb{K}$ is an eigenvalue of L, as you will show on a homework problem.

Characteristic polynomials. The fact which makes eigenvalues and eigenvectors actually possible to compute in general (at least in the finite-dimensional setting) is the fact the eigenvalues can be determined independently of the eigenvectors. Then, once we have the eigenvalues, the corresponding eigenvectors are simple to characterize.

The key to this is in recognizing that the eigenvalue/eigenvector equation $T(v) = \lambda v$ can be expressed in another way. Indeed, $\lambda \in \mathbb{K}$ is an eigenvalue of T if and only if there exists $v \neq 0$ such that

$$Tv = \lambda v, \text{ or equivalently } (T - \lambda I)v = 0,$$

which we get after subtracting λv from both sides and factoring. But this says that $v \in \ker(T - \lambda I)$, so λ is an eigenvalue of T if and only if $T - \lambda I$ is not injective (since it has something nonzero in

its kernel. When V is finite-dimensional, this is equivalent to saying that $T - \lambda I$ is not invertible, which we can further characterize by saying $\det(T - \lambda I) = 0$. Thus the conclusion is that

$$\lambda \text{ is an eigenvalue of } T \iff \det(T - \lambda I) = 0.$$

Hence, the eigenvalues can be found independently of the eigenvectors by solving the equation $\det(T - \lambda I) = 0$ for λ. The expression $\det(T - \lambda I)$ turns out to be a polynomial in the variable λ which is called the *characteristic polynomial* of T. The eigenvalues of T are then the roots of its characteristic polynomial.

Example. Let

$$A = \begin{pmatrix} 4 & 2 & 2 \\ 2 & 4 & 2 \\ 2 & 2 & 4 \end{pmatrix}.$$

The matrix $A - \lambda I$ is then

$$A - \lambda I = \begin{pmatrix} 4-\lambda & 2 & 2 \\ 2 & 4-\lambda & 2 \\ 2 & 2 & 4-\lambda \end{pmatrix}.$$

The determinant of $A - \lambda I$, which is the characteristic polynomial of A, can be computed using a cofactor expansion. (The process of finding eigenvalues is probably the only time when using cofactor expansions to compute determinants is more efficient than using row operations.) You can check my Math 290-1 lecture notes to see this computation in detail, but the end result is

$$\det(A - \lambda I) = -(\lambda - 2)^2(\lambda - 8).$$

Again, you get a polynomial of degree 3 in the variable λ, and here we have factored it. Hence the eigenvalues of A are 2 and 8.

Indeed, just to verify, note that

$$\begin{pmatrix} 4 & 2 & 2 \\ 2 & 4 & 2 \\ 2 & 2 & 4 \end{pmatrix} \begin{pmatrix} -1 \\ 1 \\ 0 \end{pmatrix} = \begin{pmatrix} -2 \\ 2 \\ 0 \end{pmatrix} = 2 \begin{pmatrix} -1 \\ 1 \\ 0 \end{pmatrix},$$

which shows that 2 is an eigenvalue, and

$$\begin{pmatrix} 4 & 2 & 2 \\ 2 & 4 & 2 \\ 2 & 2 & 4 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 8 \\ 8 \\ 8 \end{pmatrix} = 8 \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix},$$

which shows that 8 is an eigenvalue.

Lecture 11: More Eigenstuff

Number of eigenvalues. Before moving on, we give an answer to the question as to how many eigenvalues a linear transformation $T : V \to V$ can actually have, when V is finite-dimensional. Recall that the eigenvalues of T are the roots of the characteristic polynomial $\det(T - \lambda I)$, which is a polynomial of degree $\dim V$ in the variable λ. This comes from the following fact. If A is the matrix of T relative to some basis, then $A - \lambda I$ looks like

$$\begin{pmatrix} a_{11} - \lambda & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nn} - \lambda \end{pmatrix},$$

so the highest degree term comes from the pattern consisting of the diagonal entries

$$a_{11} - \lambda, \ldots, a_{nn} - \lambda.$$

Since there are n such terms, each contributing one power of λ to the characteristic polynomial, this polynomial has degree n as claimed.

Thus, the conclusion is that a linear transformation $T : V \to V$ can have at most $\dim V$ eigenvalues. Depending on what kind of scalars we allow, T can have fewer than $\dim V$ eigenvalues, but certainly not more.

Warm-Up 1. We determine the eigenvalues and eigenvectors of $T : P_n(\mathbb{K}) \to P_n(\mathbb{K})$ defined by

$$T(p(x)) = xp'(x),$$

meaning that T sends a polynomial to its derivative times x. Note that for any $k = 0, 1, \ldots, n$ we have

$$T(x^k) = x(kx^{k-1}) = kx^k,$$

which shows that any such k is an eigenvalue of T with eigenvector x^k. (Moreover, any nonzero multiple of x^k is also an eigenvector with eigenvalue k.) Since gives $k + 1$ eigenvalues so far, but since $\dim P_n(\mathbb{K}) = k + 1$, this must be the only eigenvalues of T.

Note that with respect to the basis $\mathcal{B} = \{1, x, \ldots, x^n\}$, the matrix of T is

$$[T]_{\mathcal{B}} = \begin{pmatrix} 0 & & & \\ & 1 & & \\ & & \ddots & \\ & & & n \end{pmatrix}.$$

The eigenvalues of any diagonal matrix (or more generally any upper-triangular matrix) are just its diagonal entries, which gives another way of seeing that the eigenvalues of T are $0, 1, \ldots, n$.

Warm-Up 2. We show that similar matrices have the same eigenvalues. Write $A = SBS^{-1}$ for some invertible S. Note that then

$$A - \lambda I = SBS^{-1} - \lambda I = SBS^{-1} - \lambda SIS^{-1} = S(B - \lambda I)S^{-1}.$$

Taking determinants gives

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

where we use the fact that $\det S^{-1} = \frac{1}{\det S}$. This shows that similar matrices have the same characteristic polynomial, and hence the same eigenvalues since eigenvalues are roots of the characteristic polynomial.

Here is another way to see that any eigenvalue of B is an eigenvalue of A. Suppose that λ is an eigenvalue of B, so there exists $\mathbf{v} \neq \mathbf{0}$ such that $B\mathbf{v} = \lambda\mathbf{v}$. Since $A = SBS^{-1}$, we have $AS = SB$. Thus

$$A(S\mathbf{v}) = SB\mathbf{v} = S(\lambda\mathbf{v}) = \lambda(S\mathbf{v}).$$

Since S is invertible and $\mathbf{v} \neq \mathbf{0}$, $S\mathbf{v} \neq \mathbf{0}$ as well, so this equation above shows that λ is an eigenvalue of A with eigenvector $S\mathbf{v}$. Since $B = S^{-1}AS$, the same reasoning implies that any eigenvalue of A is an eigenvalue of B, so A and B indeed have the same eigenvalues.

Algebraic multiplicity. If c is an eigenvalue of T, the characteristic polynomial of T factors into the form:

$$\det(T - \lambda I) = (\lambda - c)^k (\text{polynomial of smaller degree which doesn't have } c \text{ as a root})$$

for some $k \geq 1$. The exponent k here is called the *algebraic multiplicity* of c. For instance, for the matrix

$$\begin{pmatrix} 4 & 2 & 2 \\ 2 & 4 & 2 \\ 2 & 2 & 4 \end{pmatrix}$$

which has characteristic polynomial $-(\lambda - 2)^2(\lambda - 8)$, the eigenvalue 2 has algebraic multiplicity 2 and the eigenvalue 8 has algebraic multiplicity 8.

We'll see that the algebraic multiplicity of an eigenvalue places a restriction on the number of linearly independent eigenvectors an eigenvalue can have, which will be useful when considering diagonalizability.

Determinant = product of eigenvalues. Suppose that $T : V \to V$ is a linear transformation where V is a finite-dimensional *complex* vector space. Then the characteristic polynomial of T can be factored into linear terms as:

$$\det(T - \lambda I) = (-1)^n (\lambda - \lambda_1)^{k_1} \cdots (\lambda - \lambda_t)^{k_t}$$

where $\lambda_1, \dots, \lambda_t$ are the distinct eigenvalues of T, $k_1, \dots, k_t$ are their corresponding multiplicities, and $n = \dim V$. (The fact that we are working over the complex numbers guarantees that any polynomial can be factored in this way; in particular, note that any such transformation always has at least one complex eigenvalue.) The $(-1)^n$ term is the coefficient of λ^n in the characteristic polynomial, and comes from the

$$(a_{11} - \lambda) \cdots (a_{nn} - \lambda)$$

product in the pattern/inversion formula for $\det(T - \lambda I)$.

Now, note what happens if we set $\lambda = 0$:

$$\det T = (-1)^n (-\lambda_1)^{k_1} \cdots (-\lambda_t)^{k_t} = (-1)^{n + k_1 + \cdots + k_t} \lambda_1^{k_1} \cdots \lambda_t^{k_t}.$$

The sum $k_1 + \cdots + k_t$ of the algebraic multiplicities is just n since this gives the degree of characteristic polynomial, so

$$(-1)^{n + k_1 + \cdots + k_t} = (-1)^{2n} = 1.$$

Thus we get

$$\det T = \lambda_1^{k_1} \cdots \lambda_t^{k_t}.$$

The conclusion is that the determinant of T is simply the product of its eigenvalues, where we count each value according to its multiplicity, meaning that each eigenvalue appears in this product as many times as its multiplicity. This makes sense geometrically, at least in the case where we can find a basis of V consisting of eigenvectors of T: if $v_1, \dots, v_n$ is such a basis, each v_i is scaled by an appropriate eigenvalue when applying T, so volumes in general will be scaled by the product of the eigenvalues, which says that this product is $\det T$ when viewing this as an expansion factor.

Eigenspaces. Now that we know how to find eigenvalues, we turn to finding eigenvectors. However, this is something we already figured out in the course of deriving the characteristic polynomial:

$v \neq 0$ is an eigenvector of T with eigenvalue λ if and only if $(T - \lambda I)v = 0$. Thus, the eigenvectors of T with eigenvalues λ are precisely the nonzero vectors in $\ker(T - \lambda I)$.

We define the *eigenspace* E_λ corresponding to the eigenvalue λ to be this kernel:

$$E_\lambda := \ker(T - \lambda I).$$

Thus, the eigenspace consists of all eigenvectors with that given eigenvalue together with the zero vector. Note that this immediately implies any eigenspace is a subspace of V, which we could have seen when we first gave the definition of an eigenvector without making reference to kernels: if u, v are eigenvectors with eigenvalue λ, then

$$T(u + v) = Tu + Tv = \lambda u + \lambda v = \lambda(u + v)$$

so $u + v$ is also an eigenvector with eigenvalue λ, and if $a \in \mathbb{K}$ then

$$T(au) = aTu = a(\lambda u) = \lambda(au),$$

so au is also an eigenvector with eigenvalue λ.

Hence to find eigenvectors for a given eigenvalue we determine $\ker(T - \lambda I)$. Usually we will be interested in bases for the various eigenspaces; for matrices, such bases are found by row-reducing $A - \lambda I$ to find a basis for its kernel.

Example. For the matrix

$$A = \begin{pmatrix} 4 & 2 & 2 \\ 2 & 4 & 2 \\ 2 & 2 & 4 \end{pmatrix}$$

with eigenvalues 2 and 8, row-reducing $A - 2I$ shows that

$$\begin{pmatrix} -1 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} -1 \\ 0 \\ 1 \end{pmatrix} \text{ is a basis for } E_2$$

and row-reducing $A - 8I$ shows that

$$\begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \text{ is a basis for } E_8.$$

Note that in this case, the dimension of each eigenspace in fact equals the algebraic multiplicity of the corresponding eigenvalue.

For the matrix

$$B = \begin{pmatrix} 2 & 1 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & -3 \end{pmatrix},$$

which has eigenvalues 2 (of algebraic multiplicity 2) and -3 (of algebraic multiplicity 1), row-reducing $A - 2I$ gives

$$\begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} \text{ as a basis for } E_2$$

and row-reducing $A + 3I$ gives

$$\begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \text{ as a basis for } E_{-3}.$$

In this case the dimension of the eigenspace corresponding to 2 is strictly less than the corresponding algebraic multiplicity.

Geometric multiplicity. Given an eigenvalue λ of T, we define the *geometric multiplicity* of λ to be the dimension of the eigenspace corresponding to λ:

$$\text{geometric multiplicity of } \lambda := \dim E_\lambda = \dim \ker(T - \lambda I).$$

As the examples above show, geometric multiplicities do not necessarily equal algebraic multiplicities, but we'll see that in fact we always have

$$\text{geometric multiplicity } \leq \text{ algebraic multiplicity.}$$

Whether or not these two multiplicities are always equal will give one characterization of what it means for T to be diagonalizable.

Lecture 12: Diagonalizability

Warm-Up 1. We find bases for the eigenspaces of

$$B = \begin{pmatrix} 2 & -5 & 5 \\ 0 & 3 & -1 \\ 0 & -1 & 3 \end{pmatrix}.$$

This example comes from my Math 290-1 lecture notes, so I am just copying that solution here. Using a cofactor expansion along the first column, the characteristic polynomial of B is

$$\begin{aligned} \det(B - \lambda I) &= \begin{vmatrix} 2-\lambda & -5 & 5 \\ 0 & 3-\lambda & -1 \\ 0 & -1 & 3-\lambda \end{vmatrix} \\ &= (2-\lambda)\begin{vmatrix} 3-\lambda & -1 \\ -1 & 3-\lambda \end{vmatrix} \\ &= (2-\lambda)(\lambda^2 - 6\lambda + 8) \\ &= -(\lambda-2)^2(\lambda-4). \end{aligned}$$

Thus the eigenvalues of B are 2 with algebraic multiplicity 2 and 4 with multiplicity 1. We have:

$$B - 2I = \begin{pmatrix} 0 & -5 & 5 \\ 0 & 1 & -1 \\ 0 & -1 & 1 \end{pmatrix} \to \begin{pmatrix} 0 & 1 & -1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \text{ so a basis for } E_2 \text{ is } \left\{ \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix} \right\}$$

$$B - 4I = \begin{pmatrix} -2 & -5 & 5 \\ 0 & -1 & -1 \\ 0 & -1 & -1 \end{pmatrix} \to \begin{pmatrix} -2 & -5 & 5 \\ 0 & 1 & 1 \\ 0 & 0 & 0 \end{pmatrix}, \text{ so a basis for } E_4 \text{ is } \begin{pmatrix} 5 \\ -1 \\ 1 \end{pmatrix}.$$

Warm-Up 2. Suppose that $A = SDS^{-1}$ with D diagonal. We show that the columns of S are eigenvectors of A. Say that the diagonal entries of D are $\lambda_1, \dots, \lambda_n$. Then since $AS = SD$, we have:

$$A(S\mathbf{e}_i) = S(D\mathbf{e}_i) = S(\lambda_i \mathbf{e}_i) = \lambda_i(S\mathbf{e}_1),$$

so $S\mathbf{e}_i$, which is the i-th column of S, is an eigenvector of A with eigenvalue λ. This is essentially the same reasoning we gave in one of the approaches to the second Warm-Up from Lecture 11, only in that case D wasn't assumed to be diagonal.

The upshot is that whenever we try to write A as SDS^{-1} with D diagonal, the matrix S is found by computing some eigenvectors of A; in particular, we need n (if A is $n \times n$) linearly independent eigenvectors forming the columns of S if we want S to be invertible.

Diagonalizability. Recall that an $n \times n$ matrix was said to be *diagonalizable* if we can indeed write it as $A = SDS^{-1}$ with D diagonal. Based on the second Warm-Up above, we can now see that this is equivalent to the existence of n linearly independent eigenvectors of A. These n linearly independent eigenvectors will then form a basis of $\mathbb{R}^n$.

More generally, we can take this latter condition as the definition of what it means for an arbitrary linear transformation to be diagonalizable: a linear transformation $T : V \to V$ is diagonalizable if there exists a basis for V consisting of eigenvectors of T. We call such a basis an *eigenbasis* corresponding to T. Of course, based on the original motivation we gave for eigenvectors a few lectures ago, the existence of an eigenbasis is equivalent to the existence of a basis $\mathcal{B}$ of V relative to which $[T]_{\mathcal{B}}$ is diagonal, which is where the "diagonal" in "diagonalizable" comes from.

Example 1. We considered the matrix

$$\begin{pmatrix} 2 & 1 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & -3 \end{pmatrix}$$

in an example last time. The eigenvalues were 2 and -3, and we found that

$$\begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} \text{ and } \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$$

were bases for E_2 and E_{-3} respectively. Thus so far we have two linearly independent eigenvectors. If $\mathbf{v}$ was a third eigenvector which was linearly independent from these two, it would have to come from either E_2 or E_{-3} since the only eigenvalues are 2 and -3. But $\mathbf{v} \in E_2$ would mean that it was a multiple of the first vector above, while $\mathbf{v} \in E_{-3}$ would mean it was a multiple of the second, so it is not possible to find an eigenvector $\mathbf{v}$ which is linearly independent from the two above. Hence there is no basis of $\mathbb{R}^3$ consisting of eigenvectors of this matrix, so this matrix is not diagonalizable.

Example 2. Consider the linear transformation $D : P_n(\mathbb{C}) \to P_n(\mathbb{C})$ which sends a polynomial to its derivative:

$$T(p(x)) = p'(x).$$

Since taking a derivative decreases the degree of a polynomial, no non-constant polynomial can be sent to a multiple of itself. Thus the only eigenvectors are constant polynomials:

$$T(c) = 0 = 0(c),$$

with eigenvalue 0, and at most there is one linearly independent such constant polynomial. Hence D is not diagonalizable, except for the special case where $n = 0$, in which case $P_0(\mathbb{C})$ is just $\mathbb{C}$.

Example 3. The transformation $T : P_n(\mathbb{K}) \to P_n(\mathbb{K})$ from the first Warm-Up last time, which was defined by

$$p(x) \mapsto xp'(x),$$

is diagonalizable. Indeed, we saw in that Warm-Up that $1, x, \ldots, x^n$ were each eigenvectors of T, so that they form an eigenbasis for $P_n(\mathbb{K})$. This is also reflected in the fact, as we saw, that the matrix of T relative to this basis is diagonal.

Example 4. Consider the matrix B from the first Warm-Up. The three basis eigenvectors we found were

$$\begin{pmatrix}1\\0\\0\end{pmatrix}, \begin{pmatrix}0\\1\\1\end{pmatrix} \text{ for eigenvalue 2, and } \begin{pmatrix}5\\-1\\1\end{pmatrix} \text{ for eigenvalue 4.}$$

is a basis of $\mathbb{R}^3$ consisting of eigenvectors of B. Note that here the geometric multiplicity of each eigenvalue is the same as its algebraic multiplicity. The first two vectors are linearly independent because they constitute a basis for the same eigenspace, but do we know that putting all three vectors together still gives linearly independent vectors?

Here is one way to see this, based on Problem 2 from the Discussion 3 Problems. If, say, the second vector above was a linear combination of the other two, then we would be in a scenario where adding an eigenvector with eigenvalue 2 to one with eigenvalue 4 still gives an eigenvector—from the result of Problem 2 of the Discussion 3 Problems, this is not possible unless the two summands corresponded to the same eigenvalue. Hence it is not possible for the second vector above to be a linear combination of the other two, and similar it is not possible for the first to be a linear combination of the other two. The third is definitely not a linear combination of the first two since the first two correspond to the same eigenvalue, meaning that any linear combination of them would have to belong to the same eigenspace.

We conclude that the three eigenvectors found are linearly independent, so that they form an eigenbasis of $\mathbb{R}^3$. As a result, we have:

$$\begin{pmatrix}2&-5&5\\0&3&-1\\0&-1&3\end{pmatrix} = \begin{pmatrix}1&0&5\\0&1&-1\\0&1&1\end{pmatrix}\begin{pmatrix}2&&\\&2&\\&&4\end{pmatrix}\begin{pmatrix}1&0&5\\0&1&-1\\0&1&1\end{pmatrix}^{-1}.$$

Writing a matrix in this form (i.e. SDS^{-1}) is what it means to *diagonalize* that matrix.

Distinct eigenvalues implies independence. The check that combining basis eigenvectors from different eigenspaces still gives linearly independent vectors above was a little tedious, and in fact unnecessary: it is *always* true that eigenvectors corresponding to distinct eigenvalues are linearly independent. Thus, when determining whether or not T is diagonalizable, the only question is whether finding bases for each eigenspace gives enough vectors overall, meaning $\dim V$ many vectors. The total number of vectors found in this way is given by the sum of the geometric multiplicities of the eigenvectors (since each geometric multiplicity tells us how many basis eigenvectors we find for that one eigenvalue), so we get that T is diagonalizable if and only if

$$\text{the sum of the geometric multiplicities of all eigenvalues } = \dim V.$$

To verify that eigenvectors corresponding to distinct eigenvalues are linearly independent, first we work it out in the case of three distinct eigenvalues to see how the argument works. We'll give a formal proof of the general case using induction afterwards. So, suppose that $\lambda_1, \lambda_2, \lambda_3$ are distinct eigenvalues of T with corresponding eigenvectors v_1, v_2, v_3. Suppose that

$$c_1v_1 + c_2v_2 + c_3v_3 = 0$$

for some $c_1, c_2, c_3 \in \mathbb{K}$. We want to show $c_1 = c_2 = c_3 = 0$. Applying T to both sides gives

$$c_1Tv_1 + c_2Tv_2 + c_3Tv_3 = 0,$$

which becomes

$$c_1\lambda_1v_1 + c_2\lambda_2v_2 + c_3\lambda_3v_3 = 0$$

since v_i is an eigenvector of T with eigenvalue λ_i. Multiplying the original equation through by λ_1 gives

$$c_1\lambda_1v_1 + c_2\lambda_1v_2 + c_3\lambda_1v_3 = 0,$$

and subtracting these last two equations gives

$$c_2(\lambda_2 - \lambda_1)v_2 + c_3(\lambda_3 - \lambda_1)v_3 = 0.$$

Now, applying T again gives

$$c_2(\lambda_2 - \lambda_1)Tv_2 + c_3(\lambda_3 - \lambda_1)Tv_3 = 0,$$

which becomes

$$c_2(\lambda_2 - \lambda_1)\lambda_2v_2 + c_3(\lambda_3 - \lambda_1)\lambda_3v_3 = 0.$$

Multiplying the previous equation by λ_2 gives

$$c_2(\lambda_2 - \lambda_1)\lambda_2v_2 + c_3(\lambda_3 - \lambda_1)\lambda_2v_3 = 0,$$

and subtracting these last two equations gives

$$c_3(\lambda_3 - \lambda_1)(\lambda_3 - \lambda_2)v_3 = 0.$$

Since the eigenvalues are distinct, $\lambda_3 - \lambda_1 \neq 0$ and $\lambda_3 - \lambda_2 \neq 0$, and since eigenvectors are nonzero, this equation implies $c_3 = 0$. Then

$$c_2(\lambda_2 - \lambda_1)v_2 + c_3(\lambda_3 - \lambda_1)v_3 = 0$$

becomes

$$c_2(\lambda_2 - \lambda_1)v_2 = 0,$$

which implies $c_2 = 0$ since the other terms are nonzero. Then the original equation we started with becomes

$$c_1v_1 = 0,$$

so $c_1 = 0$ since $v_1 \neq 0$. Thus $c_1v_1 + c_2v_2 + c_3v_3 = 0$ implies $c_1 = c_2 = c_3 = 0$, so v_1, v_2, v_3 are linearly independent.

The same idea works no matter how many distinct eigenvectors we start with, but to clean up the writing we phrase this as an induction.

Proof. Suppose that $\lambda_1, \ldots, \lambda_t$ are distinct eigenvalues of T and that $v_1, \ldots, v_t$ are corresponding eigenvectors. Suppose that

$$c_1 v_1 + \cdots + c_t v_t = 0$$

for some $c_1, \ldots, c_t \in \mathbb{K}$. Applying T to both sides gives

$$c_1 \lambda_1 v_1 + c_2 \lambda_2 v_2 + \cdots + c_t \lambda_t v_t = 0,$$

and multiplying the original equation through by λ_1 gives

$$c_1 \lambda_1 v_1 + c_2 \lambda_1 v_2 \cdots + c_t \lambda_1 v_t = 0.$$

Subtracting these two equations givse

$$c_2(\lambda_2 - \lambda_1)v_2 + \cdots + c_t(\lambda_t - \lambda_1)v_t = 0.$$

We may assume by induction that $v_2, \ldots, v_t$ are linearly independent. (Again, this is just another way of phrasing an induction proof. The base case of one eigenvalue was skipped since there is nothing to check in that case, and the induction hypothesis is that if we have $t-1$ distinct eigenvalues, corresponding eigenvectors are linearly independent, which is why we can assume $v_2, \ldots, v_t$ are independent.) Then this equation implies

$$c_2(\lambda_2 - \lambda_1) = 0, \ldots, c_t(\lambda_t - \lambda_1) = 0,$$

which since the eigenvalues are distinct requires that

$$c_2 = \cdots = c_t = 0.$$

Then the original equation becomes

$$c_1 v_1 = 0,$$

so $c_1 = 0$ as well. Hence $v_1, \ldots, v_t$ are linearly independent as claimed. □

Lecture 13: More on Diagonalization

Warm-Up 1. Given $\theta \in \mathbb{R}$, we diagonalize the rotation matrix

$$A_\theta = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}.$$

Note that if θ is not an integer multiple of π, then this matrix has no real eigenvalues since no nonzero vector can be sent to a multiple of itself under such a rotation.

The characteristic polynomial of A_θ is

$$\det(A_\theta - \lambda I) = \lambda^2 + 2(\cos\theta)\lambda + 1.$$

Using the quadratic formula, the roots, and hence the eigenvalues of A_θ, are:

$$\cos\theta \pm i\sin\theta.$$

(As expected, for $\theta \neq n\pi$ these eigenvalues are complex since the $\sin\theta$ term is nonzero.) Note that this reflects a property given in Problem 9 of Homework 3: if λ is a complex eigenvalue of a *real* matrix, then $\overline{\lambda}$ is an eigenvalue as well.

Now, for $\lambda = \cos\theta + i\sin\theta$, we have:

$$A_\theta - \lambda I = \begin{pmatrix} -i\sin\theta & -\sin\theta \\ \sin\theta & -i\sin\theta \end{pmatrix},$$

so a possible for basis for the eigenspace corresponding to λ is

$$\begin{pmatrix} -1 \\ i \end{pmatrix}.$$

By Problem 9 of Homework 3 again, a basis for the eigenspace corresponding $\cos\theta - i\sin\theta$ is

$$\begin{pmatrix} -1 \\ -i \end{pmatrix}.$$

Thus we can diagonalize A_θ over $\mathbb{C}$ as

$$\begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} = \begin{pmatrix} -1 & -1 \\ i & -i \end{pmatrix} \begin{pmatrix} \cos\theta + i\sin\theta & 0 \\ 0 & \cos\theta - i\sin\theta \end{pmatrix} \begin{pmatrix} -1 & -1 \\ i & -i \end{pmatrix}^{-1}.$$

This is all that was asked for, but note one possible application of this. Taking powers of both sides gives

$$A_\theta^n = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}^n = \begin{pmatrix} -1 & -1 \\ i & -i \end{pmatrix} \begin{pmatrix} (\cos\theta + i\sin\theta)^n & 0 \\ 0 & (\cos\theta - i\sin\theta)^n \end{pmatrix} \begin{pmatrix} -1 & -1 \\ i & -i \end{pmatrix}^{-1}$$

for any $n \geq 1$. On the other hand, A_θ^n should be the matrix of rotation by the angel $n\theta$, so A_θ^n should also equal:

$$\begin{pmatrix} \cos n\theta & -\sin n\theta \\ \sin n\theta & \cos n\theta \end{pmatrix} = \begin{pmatrix} -1 & -1 \\ i & -i \end{pmatrix} \begin{pmatrix} \cos n\theta + i\sin n\theta & 0 \\ 0 & \cos n\theta - i\sin n\theta \end{pmatrix} \begin{pmatrix} -1 & -1 \\ i & -i \end{pmatrix}^{-1}.$$

Comparing entries of both resulting expressions for A_θ^n shows that

$$(\cos\theta + i\sin\theta)^n = \cos n\theta + i\sin n\theta$$

for any $n \geq 1$, an equality which is usually referred to as *Euler's formula*. The point is that here we've given a derivation of Euler's formula using linear algebra.

Warm-Up 2. Suppose that $T : V \to V$ is a diagonalizable linear transformation from a finite-dimensional vector space V to itself, and suppose that T only has one eigenvalue λ. We show that T must be a scalar multiple of the identity.

Since T is diagonalizable, there exists a basis $v_1, \ldots, v_n$ of V consisting of eigenvectors of T. Since each of these are eigenvectors, we have $Tv_i = \lambda v_i$ for all i. Let $x \in V$. Write x in terms of the given basis as:

$$x = c_1 v_1 + \cdots + c_n v_n \text{ for some } c_1, \ldots, c_n.$$

Hence

$$Tx = c_1 Tv_1 + \cdots + c_n Tv_n = c_1\lambda v_1 + \cdots + c_n \lambda v_n = \lambda(c_1 v_1 + \cdots + c_n v_n) = \lambda x,$$

so T scales *any* vector by λ and thus $T = \lambda I$ as claimed.

More on multiplicities. We showed last time that eigenvectors corresponding to distinct eigenvalues are always linearly independent, and as a consequence if we find a basis for each eigenspace and then put all basis vectors obtained together in one big list, the resulting list also consists of linearly independent vectors. Thus the sum of the geometric multiplicities of the various eigenvalues gives the maximum number of linearly independent eigenvectors it is possible to find. Hence, as mentioned last time, a linear transformation $T : V \to V$ is diagonalizable if and only if the sum of the geometric multiplicities equals the dimension of V.

In addition, we can characterize diagonalizability in terms of the relation between geometric and algebraic multiplicities. It is always true that for any eigenvalue:

$$\text{geometric multiplicity} \ \leq \ \text{algebraic multiplicity}.$$

You can find a proof of this in the book. As a result, algebraic multiplicities restrict the number of linearly independent eigenvectors we can find for a given eigenvalue, so to be diagonalizable we must be able to find precisely "algebraic multiplicity"-many linearly independent eigenvectors for that eigenvalue. That, a linear transformation is diaganalizable if and only if for each eigenvalue we have

$$\text{geometric multiplicity} \ = \ \text{algebraic multiplicity}.$$

This gives an efficient way of testing whether or not a given transformation is diagonalizable: we simply determine the dimension of each eigenspace and see if this dimension matches the algebraic multiplicity.

Exponentials. We finished with an application of diagonalizability in computing the *exponential* of a square matrix A, which is defined via the infinite sum

$$e^A := I + \sum_{n=1}^{\infty} \frac{1}{n!} A^n.$$

You can check my Math 290-1 lecture notes to see how this works. We may come back to this a bit when discussing *vector fields* next quarter, but otherwise it won't play a role going forward.

Lecture 14: Symmetric Matrices

Warm-Up. Suppose A is a 5×5 of rank 3, and that 1 and -1 are eigenvalues of A with geometric multiplicities 2 and 1 respectively. We show that $A^3 = A$. Since A has rank 3, A is not invertible so 0 is also an eigenvalue of A. The geometric multiplicity of 0 is:

$$\dim \ker A = 5 - \operatorname{rank} A = 2.$$

Thus A is diagonalizable since the geometric multiplicities add up to 5.

Diagonalize A as:

$$A = S \begin{pmatrix} 0 & & & & \\ & 0 & & & \\ & & 1 & & \\ & & & 1 & \\ & & & & -1 \end{pmatrix} S^{-1}$$

for some invertible S. Then

$$A^3 = S \begin{pmatrix} 0^3 & & & & \\ & 0^3 & & & \\ & & 1^3 & & \\ & & & 1^3 & \\ & & & & (-1)^3 \end{pmatrix} S^{-1} = S \begin{pmatrix} 0 & & & & \\ & 0 & & & \\ & & 1 & & \\ & & & 1 & \\ & & & & -1 \end{pmatrix} S^{-1} = A$$

as claimed.

Fibonacci numbers. We briefly outlined an application of diagonalization to the *Fibonacci numbers*. Check Problem 1 of Homework 5 for more information about this.

Orthogonal diagonalization. Having an eigenbasis $\mathbf{v}_1, \ldots, \mathbf{v}_n$ of $\mathbb{R}^n$ corresponding to a linear transformation T is good since, once we know to write an arbitrary $\mathbf{x} \in \mathbb{R}^n$ in terms of this basis:

$$\mathbf{x} = c_1\mathbf{v}_1 + \cdots + c_n\mathbf{v}_n,$$

it is easy to describe the action of T on $\mathbf{x}$:

$$T\mathbf{x} = c_1\lambda_1\mathbf{v}_1 + \cdots + c_n\lambda_n\mathbf{v}_n$$

where λ_i is the eigenvalue corresponding to $\mathbf{v}_i$. However, in general writing $\mathbf{x}$ in terms of this basis is not so straightforward. If in addition, our basis is also *orthonormal*, then writing $\mathbf{x}$ in the above manner is simple since the coefficients needed are simply

$$c_i = \mathbf{x} \cdot \mathbf{v}_i.$$

Thus, having an *orthonromal eigenbasis* is the best of both worlds: the orthonormal condition makes it easy to describe any vector in terms of this basis, and the eigenbasis condition makes it easy to describe the action of T on the resulting expression.

We say that a linear transformation $T : \mathbb{R}^n \to \mathbb{R}^n$ is *orthogonally diagonalizable* if there exists a basis for $\mathbb{R}^n$ consisting of orthonormal eigenvectors of T. Equivalently, T is orthogonally diagonalizable when its standard matrix A can be written as

$$A = QDQ^T$$

with D diagonal and Q orthogonal. (The columns of Q are the vectors making up the orthonormal eigenbasis.) To *orthogonally diagonalize* A means to write it is in form.

What types of matrices are orthogonally diagonalizable? The first observation is that if $A = QDQ^T$ with D diagonal and Q orthogonal, then A must be symmetric since:

$$A^T = (QDQ^T) = (Q^T)^T D^T Q^T = QDQ^T = A.$$

Thus, only symmetric matrices have the hope of being orthogonally diagonalized. We will soon show that on top of this, *any* symmetric matrix can be orthogonally diagonalized, so that "orthogonally diagonalizable" means the same thing as "symmetric".

Before proving this we need to build up some special properties symmetric matrices have, which come from the characterization of symmetric matrices as being those matrices for which

$$A\mathbf{x} \cdot \mathbf{y} = \mathbf{x} \cdot A\mathbf{y} \text{ for all } \mathbf{x}, \mathbf{y} \in \mathbb{R}^n.$$

In order to remain as general as possible, we will phrase these properties in terms of *symmetric linear transformations*: given a subspace V of $\mathbb{R}^n$, a linear transformation $T : V \to V$ is *symmetric* if $T\mathbf{u} \cdot \mathbf{v} = \mathbf{u} \cdot T\mathbf{v}$ for all $\mathbf{u}, \mathbf{v} \in V$. Of course, when $V = \mathbb{R}^n$, a symmetric linear transformation is one whose standard matrix is symmetric, but the point is that now we've extended this notion to spaces which aren't necessarily all of $\mathbb{R}^n$. It is true, however, that with respect to an orthonormal basis of V, the matrix of a symmetric linear transformation is indeed symmetric.

Real eigenvalues. We show that a symmetric linear transformation $T : V \to V$ can only have real eigenvalues. Let λ be a complex eigenvalue of T and let $\mathbf{u}$ be an associated complex eigenvector. Then

$$T\mathbf{u} \cdot \mathbf{u} = (\lambda\mathbf{u}) \cdot \mathbf{u} = \lambda(\mathbf{u} \cdot \mathbf{u}).$$

To be clear, the dot product being used is the complex dot product. On the other hand, since T is symmetric this is the same as

$$\mathbf{u} \cdot T\mathbf{u} = \mathbf{u} \cdot (\lambda\mathbf{u}) = \overline{\lambda}(\mathbf{u} \cdot \mathbf{u}),$$

where we get $\overline{\lambda}$ since we are pulling this scalar out of the *second* component of the dot product. Thus

$$\lambda(\mathbf{u} \cdot \mathbf{u}) = \overline{\lambda}(\mathbf{u} \cdot \mathbf{u}),$$

and since $\mathbf{u} \cdot \mathbf{u} \neq 0$ (since $\mathbf{u}$ is nonzero), we get $\lambda = \overline{\lambda}$. Thus λ must actually be real, so T only has real eigenvalues.

Orthogonal eigenvectors. Next we show that eigenvectors of a symmetric transformation $T : V \to V$ corresponding to distinct eigenvalues are orthogonal. Suppose that $\lambda \neq \mu$ are distinct eigenvalues of T and that $\mathbf{u}, \mathbf{v}$ are corresponding eigenvectors. Then

$$T\mathbf{u} \cdot \mathbf{v} = (\lambda\mathbf{u}) \cdot \mathbf{v} = \lambda(\mathbf{u} \cdot \mathbf{v}).$$

Since T is symmetric, this is the same as

$$\mathbf{u} \cdot T\mathbf{v} = \mathbf{u} \cdot (\mu\mathbf{v}) = \mu(\mathbf{u} \cdot \mathbf{v}),$$

where we use the fact that μ is real when factoring it out of the second component at the end. Thus

$$\lambda(\mathbf{u} \cdot \mathbf{v}) = \mu(\mathbf{u} \cdot \mathbf{v}), \text{ so } (\lambda - \mu)(\mathbf{u} \cdot \mathbf{v}).$$

Since $\lambda \neq \mu$, this implies that $\mathbf{u} \cdot \mathbf{v} = 0$ as claimed. This property guarantees that vectors in a basis for one eigenspace will always be orthogonal to vectors in a basis for a different eigenspace.

Spectral Theorem. Finally we come to the main result, which is truly one of the most important facts in all of linear algebra: a square matrix is orthogonally diagonalizable if and only if it is symmetric. This is known as the *Spectral Theorem*. We will only see a glimpse as to the importance of this in this course, but this theorem (and its infinite-dimensional analogues) form the foundation of many techniques in modern mathematics. We saw previously that being orthogonally diagonalizable implies being symmetric, so we need only prove the converse direction. The key point is the result of Problem 4 of Homework 1, which says that if A is symmetric and U is A-invariant, the orthogonal complement $U^\perp$ is also A-invariant.

The idea of the proof is to "split off" one eigenvector at a time, using properties of symmetric matrices to guarantee this can always be done. Here is how the proof works in the 3×3 case. If A is a 3×3 symmetric matrix, there exists a complex eigenvalue (since any characteristic polynomial

if nothing else has complex roots), which by a fact derived above must actually be real. Let $\mathbf{v}_1$ be a corresponding eigenvector. Look at the orthogonal complement of $\text{span}(\mathbf{v}_1)$, which is a plane. Since A is symmetric, Problem 4 of Homework 1 guarantees that A sends this orthogonal complement to itself, so A can be viewed as giving a symmetric linear transformation from this plane to itself. Applying the same reasoning as above, this transformation then also has a complex and hence real eigenvalue, so let $\mathbf{v}_1$ be a corresponding eigenvector. Then $\text{span}(\mathbf{v}_1, \mathbf{v}_2)$ is A-invariant, so A sends the orthogonal complement of this (which is a line) to itself. Viewing A as giving a symmetric linear transformation from this line to itself, the same reasoning shows that A has an eigenvector $\mathbf{v}_3$ on this line. Then all together $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3$ as constructed are orthogonal eigenvectors of A, and normalizing them gives an orthonormal eigenbasis of $\mathbb{R}^3$ as required. The proof for general n is similar, only we'll phrase it in terms of induction to make it cleaner.

Proof of Spectral Theorem. Suppose A is an $n \times n$ symmetric matrix. Then A has a complex eigenvalue, which must actually be real since A is symmetric. Let $\mathbf{v}_1$ be a corresponding eigenvector. Then any nonzero vector in $\text{span}(\mathbf{v}_1)$ is an eigenvector of A, so this span is A-invariant. Hence the orthogonal complement $V := \text{span}(\mathbf{v}_1)^\perp$ is A-invariant as well since A is symmetric.

Thus we can view A as defining a symmetric linear transformation $A : V \to V$. Since $\dim V = n - 1$, we may assume by induction that V has a basis consisting of orthonormal eigenvectors of A. (In other words, our induction hypothesis is that any symmetric linear transformation from an $(n-1)$-dimensional space to itself has a basis of orthonormal eigenvectors of A.) Call this basis $\mathbf{u}_2, \dots, \mathbf{u}_n$. Since each of these are in the orthogonal complement of $\text{span}(\mathbf{v}_1)$, each of these are orthogonal to $\mathbf{v}_1$, so $\mathbf{v}_1, \mathbf{u}_2, \dots, \mathbf{u}_n$ is an orthogonal basis of $\mathbb{R}^n$. Setting $\mathbf{u}_1 = \frac{\mathbf{v}_1}{\|\mathbf{v}_1\|}$, we have that

$$\mathbf{u}_1, \dots, \mathbf{u}_n$$

is then an orthonormal basis of $\mathbb{R}^n$ consisting of eigenvectors of A, so A is orthogonally diagonalizable as claimed. □

Example. Let

$$A = \begin{pmatrix} 3 & 1 & 1 \\ 1 & 3 & 1 \\ 1 & 1 & 3 \end{pmatrix}.$$

This has eigenvalues 2 and 5, with bases for the eigenspaces given by

$$\begin{pmatrix} -1 \\ 0 \\ 1 \end{pmatrix}, \begin{pmatrix} -1 \\ 1 \\ 0 \end{pmatrix} \text{ for } E_2 \text{ and } \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \text{ for } E_5.$$

Note that each basis eigenvector for E_2 is indeed orthogonal to the basis eigenvector for E_5, as expected since A is orthogonal.

To get an orthonormal eigenbasis, we simply apply the Gram-Schmidt process to each eigenspace separately. We get the following orthonormal basis for the two eigenspaces:

$$\begin{pmatrix} -1/\sqrt{2} \\ 0 \\ 1/\sqrt{2} \end{pmatrix}, \begin{pmatrix} -1/\sqrt{6} \\ 2/\sqrt{6} \\ -1/\sqrt{6} \end{pmatrix} \text{ for } E_2 \text{ and } \begin{pmatrix} 1/\sqrt{3} \\ 1/\sqrt{3} \\ 1/\sqrt{3} \end{pmatrix} \text{ for } E_5.$$

Each vector obtained is still an eigenvector, and we are guaranteed that all three together will be orthonormal since, again, the vectors from different eigenspaces will necessarily be orthogonal since

A is symmetric. Thus

$$\begin{pmatrix} -1/\sqrt{2} \\ 0 \\ 1/\sqrt{2} \end{pmatrix}, \begin{pmatrix} -1/\sqrt{6} \\ 2/\sqrt{6} \\ -1/\sqrt{6} \end{pmatrix}, \begin{pmatrix} 1/\sqrt{3} \\ 1/\sqrt{3} \\ 1/\sqrt{3} \end{pmatrix}$$

is an orthonormal eigenbasis of $\mathbb{R}^3$ corresponding to A.

We can thus orthogonally diagonalize A as:

$$\begin{pmatrix} 3 & 1 & 1 \\ 1 & 3 & 1 \\ 1 & 1 & 3 \end{pmatrix} = \begin{pmatrix} -1/\sqrt{2} & -1/\sqrt{6} & 1/\sqrt{3} \\ 0 & 2/\sqrt{6} & 1/\sqrt{3} \\ 1/\sqrt{2} & -1/\sqrt{6} & 1/\sqrt{3} \end{pmatrix} \begin{pmatrix} 2 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 5 \end{pmatrix} \begin{pmatrix} -1/\sqrt{2} & -1/\sqrt{6} & 1/\sqrt{3} \\ 0 & 2/\sqrt{6} & 1/\sqrt{3} \\ 1/\sqrt{2} & -1/\sqrt{6} & 1/\sqrt{3} \end{pmatrix}^T.$$

This same procedure (find a basis for each eigenspace and then apply Gram-Schmidt to each eigenspace separately) will work to orthogonally diagonalize any symmetric matrix.

Unitary diagonalization. Finally, we note that everything we did works pretty similarly for Hermitian matrices, which are complex matrices which equal their own conjugate transpose. Indeed, the same proofs we gave in the symmetric case show that the eigenvalues of a Hermitian matrix are always real, eigenvectors corresponding to distinct eigenvalues of a Hermitian matrix are always orthogonal, and any Hermitian matrix A is *unitarily diagonalizable*, which means there there exists a unitary matrix U and a diagonal matrix D such that

$$A = UDU^*$$

where U^* denotes the conjugate transpose of U. Equivalently, there is an orthonormal basis of $\mathbb{C}^n$ consisting of eigenvectors of A.

Note, however, that in the complex case, being unitarily diagonalizable does NOT imply being Hermitian, meaning there are non-Hermitian matrices which are unitarily diagonalizable. The correct version of the Spectral Theorem in the complex case is: a complex matrix A is unitarily diagonalizable if and only if $AA^* = A^*A$, which is what it means for A to be what's called *normal*. We won't study normal matrices in this class; you would learn more about them in Math 334.

Lectures 15 through 17

This portion of the notes still needs to be updated, but honestly my old lecture notes for Math 290-2 includes *all* of the material from Lectures 15, 16, and 17. So, I might not get around to actually updating these notes since I'd just be repeating myself, but I probably will anyway.

Lecture 18: Topology of $\mathbb{R}^n$

Warm-Up. We use level curves to describe the graph of the function $f : \mathbb{R}^2 \to \mathbb{R}$ defined by $f(x, y) = \cos(x + y)$. (Looking at sections at vertical planes $x = k$ or $y = k$ would also be an option, but I think the graph is a bit simpler to describe via level curves.) The level curve at $z = k$ consists of all points (x, y) satisfying

$$k = \cos(x + y).$$

For instance, for $k = 0$ this requires that $x + y$ be one of the values

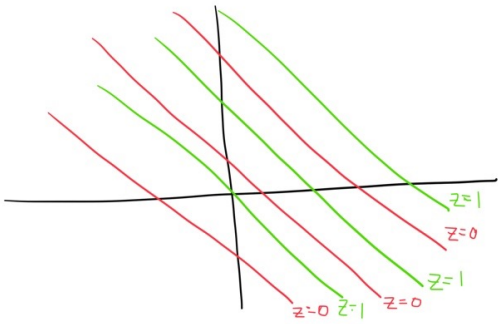
$$x + y = \pm\frac{\pi}{2}, \pm\frac{3\pi}{2}, \pm\frac{\text{odd}}{2}.$$

Each of these equations describes a line, which all together form the level curve at $z = 0$. This means that these lines form the intersection of the graph of f with the plane $z = 0$.

At $k = 1$ we get points satisfying $1 = \cos(x + y)$, so

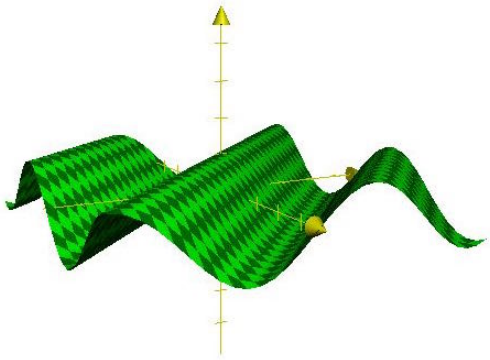
$$x + y = 0, \pm 2\pi, (\text{even})\pi.$$

This again gives a collection of lines, which are interspaced between the lines forming the level curve at $z = 0$ as follows:



For a general $z = k$, $k = \cos(x + y)$ again requires that $x + y$ be one of a discrete possible set of values, so these equations also give a collection of lines.

Now, to describe the graph of f itself in $\mathbb{R}^3$, imagine that we take one of the level curves at $z = 0$. As we move to other level curves, this line gets translated (remaining parallel to the original line) and at the same time moves up or down depending on which $z = k$ we are moving towards. So, a line at $z = 0$ is translated up to a line $z = 1$, then back down to another line at $z = 0$, then down to a line at $z = -1$, then back up, and so on. As this occurs, the lines are tracing out a surface, which is precisely the graph of f. This graph then looks like a "ripple of waves", or some kind of 3-dimensional analog of a cosine curve:



Topological notions. We now move towards studying aspects of $\mathbb{R}^n$ which go beyond the vector space structure we know so well. That is, we will now consider so-called *topological* properties subsets of $\mathbb{R}^n$ can have, and will in the coming weeks talk about their analytic aspects as well. For our purposes, by "topological" notions we mean those notions which can be expressed in terms of so called *open* and *closed* sets. *Topology* in general is a huge area of mathematics which pushes these ideas further, and we will only see a glimpse of it here. (Take Math 344 to learn more about topology; a real analysis course would also deal with many of these ideas.)

To give one brief motivation as to why such notions are important, consider the idea of taking a limit in single-variable calculus. If we have a function $f : [a, b] \to \mathbb{R}$ defined on some closed interval, the idea of taking limit as we approach some point $c \in (a, b)$ within the interval is slightly different than the idea of taking the limit as we approach either of the endpoints a or b: when approaching $c \in (a, b)$, we can approach from *either* direction (left or right) and still be within the domain of the function, while when approaching a or b we can only do so from *one* direction since anything outside $[a, b]$ is not in the domain of f. (The difference is that between a "limit" versus a "one-sided limit".) Differentiability, in particular, is one notion where this difference can lead to different phenomena, which is why differentiable functions in single-variable calculus are usually assumed to have domain an open interval (or unions of open intervals) or all of $\mathbb{R}$.

In higher dimensions the situation is even more delicate, since when approaching $\mathbf{a} \in \mathbb{R}^n$ (for $n > 1$) there are *infinitely* many directions in which this can be done. In order to have a good notion of "limit" (or "derivative") we should consider functions defined only on sets where it is in fact possible to approach a point in it from *any* possible direction. Intuitively this says that such a point cannot be on the "boundary" of the given set, and the notions we discuss below help to make this idea precise.

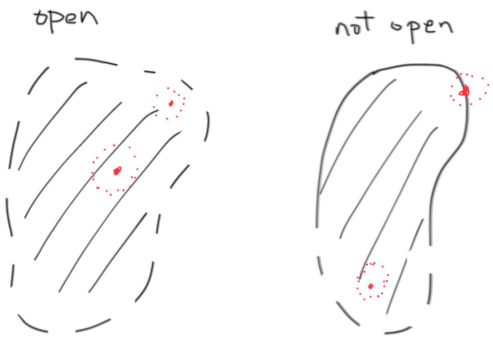
Open. For $\mathbf{p} \in \mathbb{R}^n$ and $r > 0$, the *open ball* of radius r centered at $\mathbf{p}$ is the set $B_r(p)$ of all points in $\mathbb{R}^n$ whose distance to $\mathbf{p}$ is smaller than r:

$$B_r(\mathbf{p}) := \{\mathbf{q} \in \mathbb{R}^n \mid \|\mathbf{q} - \mathbf{p}\| < r\}.$$

(Note that the norm $\|\mathbf{q} - \mathbf{p}\|$ of the difference of the vectors $\mathbf{q}$ and $\mathbf{p}$ indeed gives the distance between them.) When $n = 1$, the open ball of radius r around $p \in \mathbb{R}$ is the open interval $(p-r, p+r)$; when $n = 2$ the open ball $B_r(\mathbf{p})$ is the open disk (not including the boundary circle) of radius r centered at $\mathbf{p}$; and when $n = 3$ an open ball looks like the region enclosed by a sphere (not including the sphere itself), or in other words a "ball", which is where the name for general n comes from.

We say that a subset U of $\mathbb{R}^n$ is *open* if for any $\mathbf{p} \in U$, there exists an open ball $B_r(\mathbf{p})$ around $\mathbf{p}$ which is fully contained in U. Relating this to the motivation we outlined above, having such a ball around $\mathbf{p} \in U$ guarantees that it is indeed possible to approach $\mathbf{p}$ from any given direction and still remain within U itself. Thus, open sets are the "natural" types of domains on which to consider limits and differentiable functions, as we'll see.

Visually, open subsets of $\mathbb{R}^n$ are easy to detect. Consider the following subsets of $\mathbb{R}^2$:



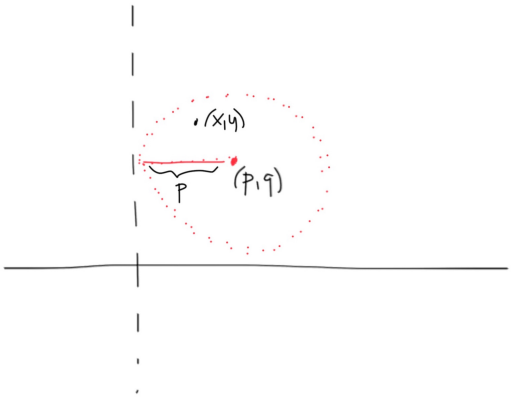
Here and in other pictures we draw, a dashed-line indicates that those points are not included in the given region, while a solid line indicates that they are. In the region on the left we can see that given any point within it, we can always draw a ball around that point which remains fully within the region. As our point gets closer to the "boundary", we might have to take smaller balls,

but nonetheless such open balls can always be found. As a contrast, in the set on the right it is not always possible to find such open balls; for points it is, but for a point say on the "boundary" curve itself, any ball around that point will always contains points outside the region, so no such ball can be fully contained within the region in question.

Example. We show that $\mathbb{R}^2$ with the y-axis removed is open in $\mathbb{R}^2$. Concretely, this is the set

$$U = \{(x, y) \in \mathbb{R}^2 \mid x \neq 0\}.$$

Let $(p, q) \in U$. To show that U is open we must describe a radius $r > 0$ such that the ball $B_r(p, q)$ remains within U, so a radius r such any point in the corresponding ball has nonzero x-coordinate. For simplicity, let us assume that (p, q) is actually in the first quadrant. Visually, it is clear that such a ball exists:



Based on this picture, it seems that the largest radius which should work is the value p itself. (If (p, q) was to the left of the y-axis, we would use $|p|$ instead.) We thus show that $B_p(p, q) \subseteq U$. Let $(x, y) \in B_p(p, q)$, so the distance from (x, y) to (p, q) is less than p:

$$\sqrt{(x - p)^2 + (y - q)^2} < p.$$

We need to show that $x \neq 0$ (in fact, $x > 0$ in the case where (p, q) is in the first quadrant) in order to conclude that $(x, y) \in U$. The distance $|x - p|$ between the x-coordinates of (x, y) and (p, q) thus satisfies:

$$|x - p| \leq \sqrt{(x - p)^2 + (y - q)^2} < p.$$

Since $|x - p| < p$, we have

$$-p < x - p < p, \text{ so } 0 < x < 2p$$

after adding p throughout. Hence $x > 0$, so $(x, y) \in U$ as desired. Thus $B_p(p, q) \subseteq U$, so U is open in $\mathbb{R}^2$. (Another way to show that $x \neq 0$ is to note that if x was zero, then

$$\sqrt{(x - p)^2 + (y - q)^2} < p$$

gives after squaring:

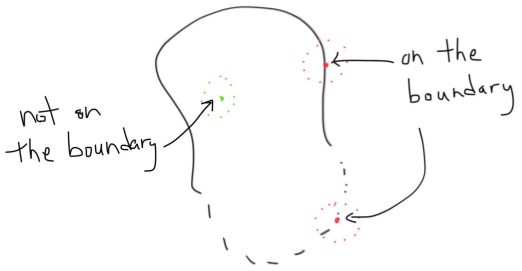
$$p^2 + (y - q)^2 < p^2,$$

so $(y - q)^2 < 0$, which is not possible.)

Boundary. Intuitively, the boundary of a region describes where the region "stops" at. The precise definition is as follows. Let S be a subset of $\mathbb{R}^n$. A *boundary point* of S is a point $\mathbf{p} \in \mathbb{R}^n$ with the

property that *every* open ball $B_r(\mathbf{p})$ around it contains something in S and something not in S. The collection of all boundary points of S is called the *boundary* of S and is denoted by ∂S.

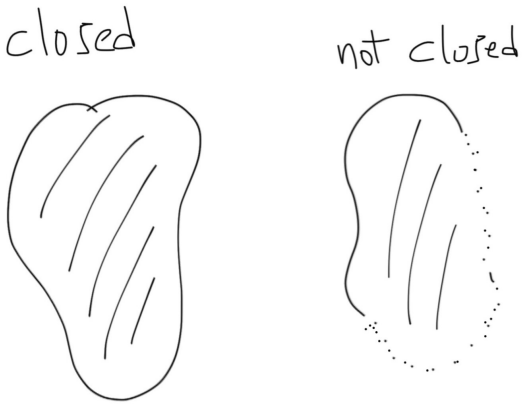
As in the case of open sets, boundaries are also easy to describe visually:



As the name suggests, the boundary points are those which occur at the "edge" of the region in question. Such points might be in the region itself, say the ones on the solid curve portion above, but other points might not be, say the ones on the dashed-curve portion. For the green point, we have drawn a ball around it which contains no point outside the given region, so this point is not a boundary point.

Note also that we can characterize open sets in terms of their boundaries: a set is open if and only if it contains none of its boundary. This is visually clear in the various pictures we've drawn above, and you will give a proof of this fact on the homework.

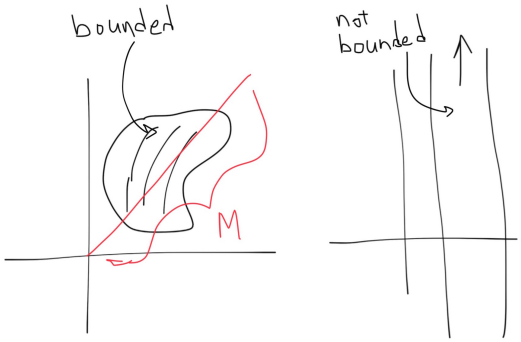
Closed. A subset A of $\mathbb{R}^n$ is said to be *closed* in $\mathbb{R}^n$ if it contains *all* of its boundary. Visually we have:



The boundary of the given region in each case is the "outside" curve enclosing the region, and in the region in the left this boundary is contained in the region itself, whereas it is not fully contained in the region in the picture on the right.

Bounded. Intuitively, a region is bounded if it does not go "off to infinity" in any direction. More precisely, a subset K of $\mathbb{R}^n$ is *bounded* if there exists $M > 0$ such that $\|\mathbf{x}\| \leq M$ for all $\mathbf{x} \in K$. Thus, there is a restriction as to how far away points of K can be from the origin. Note that for any $r > M$ we then have $\|\mathbf{x} - \mathbf{0}\| < r$ for all $\mathbf{x} \in K$, which says that all points of K belong to the open ball $B_r(\mathbf{0})$. Hence we can rephrase the definition of bounded as saying that K is contained in *some* open ball of a finite radius centered at the origin.

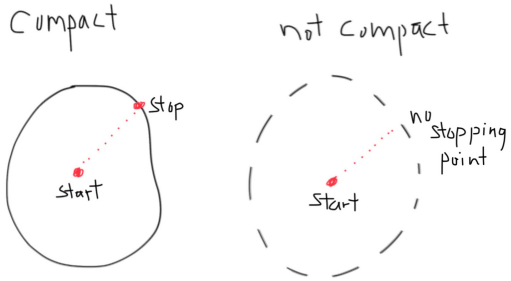
Visually we have:



The region on the left is bounded, and we have indicated a possible $M > 0$ which is longer than the length of any vector in that region. The vertical strip on the right is not bounded since it extends to infinite either vertically up or down, and points in these directions get further and further away from the origin.

Compact. A subset K of $\mathbb{R}^n$ is *compact* if it is closed and bounded. This notion is trickier to get some good intuition for, but in some sense compact sets are ones which only "extend finitely". The boundedness condition makes it clear that compact sets don't extend to infinity.

The closed condition says the following. Imagine you start at point of a bounded set and walk towards the boundary:



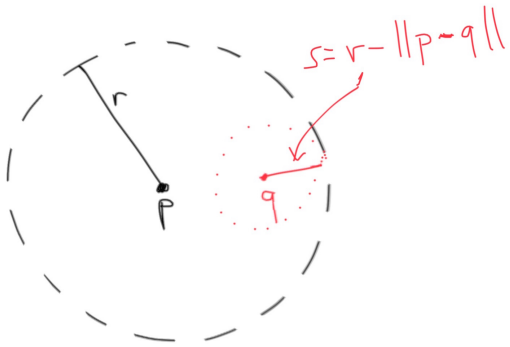
In the compact (closed and bounded) case, you can reach the boundary all while remaining inside the given set. However, in the open and bounded (so not compact) case, if we require that you remain within the set the entire time, then you will never actually "reach" the boundary since the boundary points are not in the given set. Thus, an open and bounded set "extends indefinitely" since there is no "stopping point", while a compact set can only extend so much before you stop. The real reason why we care compactness is because of the special properties which continuous functions defined on them have, as we'll soon see.

Lecture 19: Multivariable Limits

Warm-Up 1. Let $\mathbf{p} \in \mathbb{R}^n$ and $r > 0$. We show that the open ball $B_r(\mathbf{p})$ of radius r centered at $\mathbf{p}$ is open in $\mathbb{R}^n$. Recall that this open ball is the set of all points in $\mathbb{R}^n$ whose distance away from $\mathbf{p}$ is less than r:

$$B_r(\mathbf{p}) = \{\mathbf{q} \in \mathbb{R}^n \mid \|\mathbf{p} - \mathbf{q}\| < r\}.$$

Let $\mathbf{q} \in B_r(\mathbf{p})$. Then we must find a ball around $\mathbf{q}$ which is fully contained in $B_r(\mathbf{p})$. We claim that the ball of radius $s := r - \|\mathbf{p} - \mathbf{q}\|$ centered at $\mathbf{q}$ works:



As the picture suggests, this value of s appears to be the largest possible radius we can draw around $\mathbf{q}$ to give a ball fully contained in $B_r(\mathbf{p})$. Note that $s > 0$ since $\|\mathbf{p} - \mathbf{q}\| < r$, so s is a valid candidate for a radius.

To show that $B_s(\mathbf{q}) \subseteq B_r(\mathbf{p})$, let $\mathbf{x} \in B_s(\mathbf{q})$. We must show that $\|\mathbf{x} - \mathbf{p}\| < r$ in order to conclude that $\mathbf{x} \in B_r(\mathbf{p})$. Since $\mathbf{x} \in B_s(\mathbf{q})$, $\|\mathbf{x} - \mathbf{q}\| < s$. We need some way of relating the various distances being considered, and this is given by the so-called *triangle inequality*:

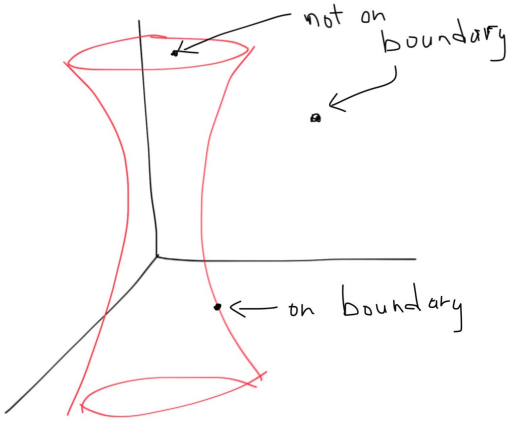
$$\|\mathbf{x} - \mathbf{p}\| \leq \|\mathbf{x} - \mathbf{q}\| + \|\mathbf{q} - \mathbf{p}\|.$$

(The three terms in this inequality are the lengths of the sides of a triangle, and this inequality says that that the length of any one side is always less than or equal to the sum of the lengths of the other two sides; this is where the name "triangle inequality" comes from. The triangle inequality is one of the most important inequalities you'll come across in a Real Analysis course.) Since $\|\mathbf{x} - \mathbf{q}\| < s$, we then get

$$\begin{aligned} \|\mathbf{x} - \mathbf{p}\| &\leq \|\mathbf{x} - \mathbf{q}\| + \|\mathbf{q} - \mathbf{p}\| \\ &< s + \|\mathbf{q} - \mathbf{p}\| \\ &= r \end{aligned}$$

since $s = r - \|\mathbf{q} - \mathbf{p}\|$. Thus $\|\mathbf{x} - \mathbf{p}\| < r$, so $\mathbf{x} \in B_r(\mathbf{p})$ as desired. Hence $B_s(\mathbf{q}) \subseteq B_r(\mathbf{p})$, so $B_r(\mathbf{p})$ is open in $\mathbb{R}^n$ as claimed.

Warm-Up 2. We show that the hyperboloid (of one sheet) defined by $x^2 + y^2 - z^2 = 1$ is closed and unbounded in $\mathbb{R}^3$. (So in particular, it is not compact. As a general rule, subsets of $\mathbb{R}^n$ defined by *equalities* are likely to be closed, whereas sets defined by *strict inequalities* are likely to be open.) First, the boundary of this hyperboloid is the hyperboloid itself. Indeed, around any point not on the hyperboloid we can find a small enough ball which does not intersect the hyperboloid at all, so no such point can be a boundary point of the hyperboloid:



Moreover, given any point on the hyperboloid, any ball around it will contains points both on the hyperboloid and not on the hyperboloid, so any point on the hyperboloid is a boundary point. Thus the hyperboloid contains its own boundary, so it is closed.

To say that the hyperboloid is unbounded means that we can find points on it which are arbitrarily far away from the origin. To be precise, this means that given any $M > 0$ we can find a point on the hyperboloid whose distance to the origin is larger than M. For $M > 0$, there is a point on the hyperboloid with z-coordinate M, say for instance

$$(\sqrt{1+M^2}, 0, M).$$

The distance from this point to the origin is

$$\sqrt{\sqrt{1+M^2}^2 + 0^2 + M^2} = \sqrt{1+2M^2} > \sqrt{M^2} = M,$$

showing that the hyperboloid is unbounded. Alternatively, note that the intersection of the hyperboloid with the $y = 1$ plane is the curve with equation

$$x^2 + 1 - z^2 = 1, \text{ or } x^2 = -z^2.$$

The distance from a point on this intersection to the origin is

$$\sqrt{z^2 + 1 + z^2} = \sqrt{1+2z^2}.$$

As z increases, $\sqrt{1+2z^2}$ grows without bound, so we can always find points on the hyperboloid whose distance to the origin is larger than any prescribed positive number.

Limits. Given a function $f : U \to \mathbb{R}^m$, where U is an open subset of $\mathbb{R}^n$, and a point $\mathbf{a} \in U$, we want to make sense of the *limit* of f as $\mathbf{x}$ approaches $\mathbf{a}$. Intuitively, this should be a point $\mathbf{L} \in \mathbb{R}^m$ such that as $\mathbf{x}$ gets closer and closer to $\mathbf{a}$, $f(\mathbf{x})$ gets closer and closer to $\mathbf{L}$. Here is the formal definition:

> We say that the *limit* of f as $\mathbf{x}$ approaches $\mathbf{a}$ is $\mathbf{L}$ if for every $\epsilon > 0$, there exists $\delta > 0$ such that
>
> $$\text{if } 0 < \|\mathbf{x} - \mathbf{a}\| < \delta, \text{ then } \|f(\mathbf{x}) - \mathbf{L}\| < \epsilon.$$

For notation, we write $\lim_{\mathbf{x}\to\mathbf{a}} f(\mathbf{x}) = \mathbf{L}$ when this condition holds.

What exactly does this definition mean, and how does it capture the intuitive idea we mentioned above? We'll get to that in a bit, but to fully understand this type of definition and its consequences would require a course in what's called *real analysis*, such as Math 320 or 321. Our course is a multivariable calculus course and not a course in analysis, meaning that we won't explore this concept in too much depth. We will outline the intuition below and see how to use it in some examples, but for our purposes this will be enough. Nonetheless, it is good to know that everyone we will do can be completely derived from this formal definition, which incidentally took literally thousands of years to fully develop and is a true testament to the power of human thought.

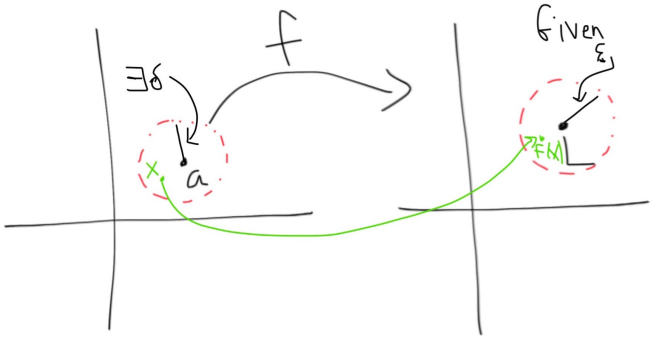
Intuition. The key to the intuition behind the formal definition of a limit comes from interpreting the inequalities used in the definition in terms of open balls: to say that $0 < \|\mathbf{x} - \mathbf{a}\| < \delta$ means that

$$\mathbf{x} \in B_\delta(\mathbf{a}) \text{ and } \mathbf{x} \neq \mathbf{a},$$

and to say that $\|f(\mathbf{x}) - \mathbf{L}\| < \epsilon$ means

$$f(\mathbf{x}) \in B_\epsilon(\mathbf{L}).$$

Thus, the definition can be rephrased as saying: for any open ball $B_\epsilon(\mathbf{L})$ around $\mathbf{L}$, there exists an open ball $B_\delta(\mathbf{a})$ around $\mathbf{a}$ such that that any point in this open ball apart from $\mathbf{a}$ itself is sent into the open ball $B_\epsilon(\mathbf{L})$. Visually, this looks like:



Interpreting ϵ as a measure for how close we want to end up to $\mathbf{L}$, δ is then a measure for how close we have to be to $\mathbf{a}$ in order to guarantee that we end up within ϵ away from $\mathbf{L}$. As ϵ gets smaller (i.e. as the open ball around $\mathbf{L}$ we are considering shrinks), δ gets smaller as well, meaning that we have to get closer to $\mathbf{a}$ to guarantee we end up however close we wanted to be to $\mathbf{L}$, but nonetheless there is a δ which will work. Thus, we we get "closer and closer" to $\mathbf{a}$ (characterized by shrinking δ's), we end up "closer and closer" to $\mathbf{L}$ (characterized by shrinking ϵ's), precisely as the intuitive notion of "limit" suggests should happen.

Example. We prove rigorously that

$$\lim_{(x,y)\to(1,2)} (x + 2y + 1) = 6.$$

Denote by f the function $\mathbb{R}^2 \to \mathbb{R}$ given by $f(x, y) = x + 2y + 1$. The candidate value of 6 for the limit comes from using what you know about single-variable limits and the intuition that $x+2y+1$ should approach $1 + 2(2) + 1$.

Let $\epsilon > 0$. We want to find $\delta > 0$ such that any point (x, y) satisfying $0 < \|(x, y) - (1, 2)\| < \delta$ in sent under f to a point satisfying $|f(x, y) - 6| < \epsilon$. We have:

$$|f(x, y) - 6| = |x + 2y + 1 - 6| = |(x - 1) + 2(y - 2)|.$$

Note that we are writing the resulting expression as $(x - 1) + 2(y - 2)$ in order to emphasize that the x coordinate being considered is approaching 1 and the y coordinate is approaching 2; as we'll see, the point is that saying x approaches 1 and y approaches 2 will let us make some estimates as to how large $x - 1$ and $y - 2$ can be. Now we use a basic fact about absolute values, which is the version of the *triangle inequality* alluded to earlier in $\mathbb{R}^1$: $|a + b| \leq |a| + |b|$ for any $a, b \in \mathbb{R}$. In our setting, this gives

$$|f(x, y) - 6| = |(x - 1) + 2(y - 2)| \leq |x - 1| + |2(y - 2)| = |x - 1| + 2|y - 2|.$$

Now, our goal is to make $|f(x, y)-6|$ smaller than ϵ for some choice of δ, which characterizes how close (x, y) is to $(1, 2)$. The idea is that if we can make the final expression above $|x - 1| + 2|y - 2|$ smaller than epsilon, then will in turn force $|f(x, y) - 6|$ to be smaller than ϵ as well. The point (x, y) being considered is meant to satisfy

$$0 < \|(x, y) - (1, 2)\| < \delta,$$

for a still-unknown δ, which can be written as:

$$0 < \sqrt{(x - 1)^2 + (y - 2)^2} < \delta.$$

Since each term under the square root is nonnegative, dropping either one gives a smaller expression. Thus

$$|x - 1| = \sqrt{(x - 1)^2} \leq \sqrt{(x - 1)^2 + (y - 2)^2} < \delta$$

and

$$|y - 2| = \sqrt{(y - 2)^2} \leq \sqrt{(x - 1)^2 + (y - 2)^2} < \delta.$$

Thus saying that (x, y) is within a distance of δ away from $(1, 2)$ guarantees that the x-coordinate is within δ away from 1 and the y-coordinate is within δ away from 2. Using these bounds, our previous expression is bounded by:

$$|f(x, y) - 6| \leq |x - 1| + 2|y - 2| < \delta + 2\delta = 3\delta.$$

Recall that our goal was to find a $\delta > 0$ which guarantees $|f(x, y) - 6| < \epsilon$. Now we're in business: if we choose $\delta > 0$ which satisfies $3\delta \leq \epsilon$, then we will indeed have

$$|f(x, y) - 6| < 3\delta \leq \epsilon$$

as required. In particular, picking $\delta = \frac{\epsilon}{3}$ works. The point is that for any $\epsilon > 0$, points (x, y) to be within a distance of $\frac{\epsilon}{3}$ away from $(1, 2)$ are guaranteed to be sent to points $f(x, y) = x + 2y + 1$ within a distance of ϵ away from 6, which is what is needed in order to conclude that

$$\lim_{(x,y)\to(1,2)} (x + 2y + 1) = 6.$$

Non-existence of limits. One of the main reasons we are looking at the formal definition of a limit in this course is to make precise the following idea: if approaching **a** along different directions

gives different candidate values for the limit, then the limit does not exist. This is going to be, for us, the most practical way of showing a limit does not exist.

Here is the precise statement. Suppose there exists a curve passing through $\mathbf{a}$ along which $f(\mathbf{x})$ approaches $\mathbf{L}_1$, and a curve passing through $\mathbf{a}$ along which $f(\mathbf{x})$ approaches L_2. If $L_1 \neq L_2$, then $\lim_{\mathbf{x}\to\mathbf{a}} f(\mathbf{x})$ does not exist. We'll see in the following example precisely what we mean by "the limit as we approach $\mathbf{a}$ along a given curve", which is actually just a type of single-variable limit. We'll look at why this statement follows from the formal definition of a limit next time.

Example. We show that

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

does not exist. Along the line $y = x$, our points (x, y) are of the form (x, x). Thus a two-variable function $f(x, y)$ "restricts" to give a single-variable function $f(x, x)$ along this curve, and we can ask whether or not this single-variable function has a limit. In our case, approaching $(0, 0)$ along $y = x$ gives

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

This says that if the limit in question did exist, it should equal 0.

On the other hand, when approaching along the x-axis, our points are of the form $(x, 0)$, and we have:

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

so if the limit in question existed this suggests it would have to equal one. Since we have found two curves such that approaching $(0, 0)$ along each gives different candidate values for the limit, we conclude that the limit in question does not exist. Note that approaching the origin along the y-axis results in:

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

which would be yet another candidate value for the limit.

Lecture 20: More on Limits

Warm-Up 1. We show that

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

using the formal definition of limits. Let $\epsilon > 0$. We want $\delta > 0$ such that

$$0 < \|(x, y) - (0, 0)\| < \delta \text{ implies } |(x^2 - y^2) - 0| < \epsilon.$$

The expression we want to make smaller than ϵ can first be bounded by:

$$|x^2 - y^2| \leq |x^2| + |y^2| = |x|^2 + |y|^2$$

using the triangle inequality for absolute values. Now, the point (x, y) we are considering will satisfy

$$\sqrt{x^2 + y^2} < \delta$$

for the still-unknown δ. But in particular, this implies that $|x|$ and $|y|$ themselves are bounded by δ:

$$|x| = \sqrt{x^2} \leq \sqrt{x^2 + y^2} < \delta \text{ and } |y| = \sqrt{y^2} \leq \sqrt{x^2 + y^2} < \delta.$$

(Visually, (x, y) lies in the disk of radius δ centered at the origin, and the point is that the x and y-coordinates of any such point must also be within $-\delta$ and δ.) Thus, for the points (x, y) being considered we have:

$$|x^2 - y^2| \leq |x|^2 + |y|^2 < \delta^2 + \delta^2 = 2\delta^2.$$

Hence picking $\delta = \sqrt{\frac{\epsilon}{2}}$, for instance, guarantees that $|x^2 - y^2| < \epsilon$.

To be clear, given $\epsilon > 0$, let $\delta = \sqrt{\frac{\epsilon}{2}}$. Since ϵ is positive this δ is positive as well. Let (x, y) be a point satisfying $0 < \|(x, y) - (0, 0)\| < \delta$. Then in particular

$$|x| \leq \sqrt{x^2 + y^2} < \delta \text{ and } |y| \leq \sqrt{x^2 + y^2} < \delta.$$

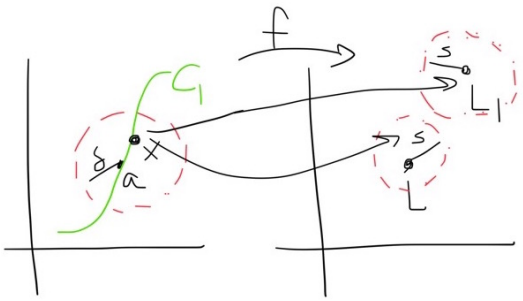
Thus

$$|(x^2 - y^2) - 0| \leq |x|^2 + |y|^2 < \delta^2 + \delta^2 = \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon,$$

showing that $\lim_{(x,y)\to(0,0)}(x^2 - y^2) = 0$ as claimed.

Warm-Up 2. We now justify the method we introduced last time for showing that limits do not exist. The claim was that if the limit of $f(\mathbf{x})$ as we approach $\mathbf{a}$ along some curve C_1 gives the value $\mathbf{L}_1$, and approaching $\mathbf{a}$ along another curve C_2 gives a different value $\mathbf{L}_2$, then $\lim_{\mathbf{x}\to\mathbf{a}} f(\mathbf{x})$ does not exist. This will follow from the fact that if $\lim_{\mathbf{x}\to\mathbf{a}} f(\mathbf{x})$ exists and equals $\mathbf{L}$, then the limit as we approach $\mathbf{a}$ along any curve will also be $\mathbf{L}$. (Hence if $\lim_{\mathbf{x}\to\mathbf{a}} f(\mathbf{x})$ did exist, then the $\mathbf{L}_1$ and $\mathbf{L}_2$ described above would have to be the same since they would both equal the value of $\lim_{\mathbf{x}\to\mathbf{a}} f(\mathbf{x})$; so if $\mathbf{L}_1 \neq \mathbf{L}_2$, then $\lim_{\mathbf{x}\to\mathbf{a}} f(\mathbf{x})$ cannot exist.)

Suppose that $\lim_{\mathbf{x}\to\mathbf{a}} f(\mathbf{x}) = \mathbf{L}$ exists, and that approaching $\mathbf{a}$ along a curve C_1 gives a limit of $\mathbf{L}_1$. For a contradiction, suppose that $\mathbf{L} \neq \mathbf{L}_1$. Then we can find small enough balls $B_s(\mathbf{L})$ and $B_s(\mathbf{L}_1)$ around $\mathbf{L}$ and $\mathbf{L}_1$ respectively which do not intersect; for instance, taking the radius to be $s = \frac{1}{2}\|\mathbf{L} - \mathbf{L}_1\| > 0$ works:



Since $\lim_{\mathbf{x}\to\mathbf{a}} f(\mathbf{x}) = \mathbf{L}$, there exists a ball $B_\delta(\mathbf{a})$ around $\mathbf{a}$ such that any $\mathbf{x} \in B_\delta(\mathbf{a})$—apart from $\mathbf{a}$ itself—is sent to $f(\mathbf{x}) \in B_s(\mathbf{L})$. But if the limit of f as we approach $\mathbf{a}$ along C_1 is $\mathbf{L}_1$, then for $\mathbf{x} \in C_1$ close enough to $\mathbf{a}$ (and not equal to $\mathbf{a}$) we have $\mathbf{x} \in B_s(\mathbf{L}_1)$. In particular, there are points on C_1 which also lie in the ball $B_\delta(\mathbf{a})$, and such points $\mathbf{x}$ must thus satisfy both $\mathbf{x} \in B_s(\mathbf{L})$ and $\mathbf{x} \in B_s(\mathbf{L}_1)$, which is a contradiction since these balls do not intersect. Hence we conclude that the limit of f as we approach $\mathbf{a}$ along C_1 must be the same as $\mathbf{L}$ as claimed.

Example. We show that

$$\lim_{(x,y)\to(0,0)} \frac{x^4y^4}{(x^2 + y^4)^3}$$

does not exist. The interesting observation here is that, as we'll see, the limit as we approach $(0,0)$ along *any* line will in fact be 0, and yet the limit in question does not exist since approaching some curve which is not a line can give a nonzero value. Thus, knowing that the limit along any line always gives the same value is not enough to say that the overall limit exists.

Take a line $y = mx$ for some $m \in \mathbb{R}$. Approaching the origin along this line gives:

$$\lim_{(x,mx)\to(0,0)} \frac{x^4(mx)^4}{(x^2+(mx)^4)^3} = \lim_{x\to 0} \frac{mx^8}{(x^2+m^4x^4)^3} = \lim_{x\to 0} \frac{mx^8}{x^6(1+m^4x^2)^3} = \lim_{x\to 0} \frac{mx^2}{(1+m^4x^2)^3} = 0$$

since the numerator goes to 0 and the denominator to 1. (We will take for granted such well-known facts about single-variable limits.) The lines in question include all lines through the origin except for the y-axis, but along the y-axis we have

$$\lim_{(0,y)\to(0,0)} \frac{0}{y^{12}} = 0$$

as well. Thus approaching the origin along any line gives the same candidate limit value of zero.

However, approaching the origin along the parabola $x = y^2$ gives:

$$\lim_{(y^2,y)} \frac{(y^2)^4y^4}{((y^2)^2+y^4)^3} = \lim_{y\to 0} \frac{y^{12}}{8y^{12}} = \frac{1}{8}.$$

Thus

$$\lim_{(x,y)\to(0,0)} \frac{x^4y^4}{(x^2+y^4)^3}$$

does not exist as claimed.

Other coordinates. In class we looked at how we can determine limits by converting to other coordinate systems, such as polar coordinates in $\mathbb{R}^2$ or spherical coordinates in $\mathbb{R}^3$. This material can be found in my Math 290-2 lecture notes, so I'll avoid reproducing it here. Note that to make some of these limit computations precise requires the use of the *squeeze theorem* from Problem 6 of Homework 6, which isn't made completely clear in my 290-2 notes. You should be able to pick out on your own when the exactly the squeeze theorem is required. See the solutions to Problem 9 of Homework 6 to see examples which correctly use the squeeze theorem.

Other properties. Check the book (or the Discussion 5 Problems) for other properties of limits we will take for granted, such as the fact that limits are unique (when they exist), the limit of a sum is the sum of the limits (when both exist), multiplying a function by a scalar multiplies the limit by that same scalar, etc. The proofs of these are in the book (some are in the Discussion 5 Solutions), but going through these in class would move us away from our goal and really belong in a real analysis course instead. Feel free to use such properties whenever needed without justification unless stated otherwise.

What we will use in practice. Moving forward, for us multivariable limits will show up in the definition of the *derivative* of a multivariable function and in determining differentiability. For the types of limits which show up in this setting, all we really care about is whether a limit is zero or not. Thus, for most purposes, the idea of approaching along different curves or converting to other coordinate systems will be good enough; that is, beyond some problems on the homework or possibly one on the midterm, we won't really use the formal definition anymore. This formal definition, however, would play a much bigger role in an analysis course.

Lecture 21: Differentiability

Continuity. Before moving on, we give the definition of what it means for a function to be continuous, which is easy to state in terms of limits. A function $f : U \subseteq \mathbb{R}^n \to \mathbb{R}^m$, where U is open in $\mathbb{R}^n$, is *continuous* at $\mathbf{p} \in U$ if

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

Thus, to be continuous at a point just means that the limit as you approach that point should just be the value of the function at that point. We say f is *continuous* if it is continuous at each point of its domain.

One thing to note is that since the limit $\lim_{\mathbf{x}\to\mathbf{p}} f(\mathbf{x})$ does not depend on what happens *at* $\mathbf{p}$ (which comes from only considering $\mathbf{x}$ satisfying $0 < \|\mathbf{x} - \mathbf{p}\| < \delta$ in the definition), continuity says that the behavior of f at $\mathbf{p}$ is fully determined by the behavior of f nearby $\mathbf{p}$. Intuitively, the idea is that moving away from $\mathbf{p}$ by a small amount does not greatly alter the value of f.

We will take for granted the continuity of the types of continuous functions you saw in a single-variable calculus course (polynomials, trig functions, exponentials, etc), as well as basic facts like those saying that the sum, product, composition, etc. of continuous functions is continuous.

Warm-Up 1. We determine if the function $f : \mathbb{R}^2 \to \mathbb{R}$ defined by

$$f(x,y) = \begin{cases} \frac{x \ln y}{y-x-1} & \text{if } y \neq x+1 \\ 0 & \text{if } y = x+1 \end{cases}$$

is continuous at $(0,1)$. Concretely, this is asking whether or not

$$\lim_{(x,y)\to(0,1)} f(x,y) = f(0,1).$$

Approaching $(0,1)$ along the line $y = 1$ gives:

$$\lim_{(x,1)\to(0,1)} \frac{x \ln 1}{1-x-1} = \lim_{(x,1)\to(0,1)} \frac{0}{-x} = 0.$$

However, approaching along the curve $y = e^x$ gives:

$$\lim_{(x,e^x)\to(0,1)} \frac{x \ln(e^x)}{e^x - x - 1} = \lim_{x\to 0} \frac{x^2}{e^x - x - 1} = \lim_{x\to 0} \frac{2x}{e^x - 1} = \lim_{x\to 0} \frac{2}{e^x} = 2$$

as a result of two applications of L'Hopital's rule. Thus $\lim_{(x,y)\to(0,1)} f(x,y)$ does not exist, so it certainly does not equal $f(0,1) = 0$. Hence f is not continuous at $(0,1)$.

Warm-Up 2. We determine the value of $c \in \mathbb{R}$ for which $f : \mathbb{R}^3 \to \mathbb{R}$ defined by

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

is continuous at the origin. Thus, this is the value of c for which

$$\lim_{(x,y,z)\to(0,0,0)} f(x,y,z) = f(0,0,0) = c.$$

After converting to spherical coordinates, we have:

$$\frac{x^4 - y^4}{x^2 + y^2 + z^2} = \frac{\rho^4(\sin^4\phi\cos^4\theta - \sin^4\phi\sin^4\theta)}{\rho^2} = \rho^2(\sin^4\phi\cos^4\theta - \sin^4\phi\sin^4\theta),$$

and the squeeze theorem implies that this expression has limit zero as we approach the origin. Hence

$$\lim_{(x,y,z)\to(0,0,0)} f(x,y,z) = 0,$$

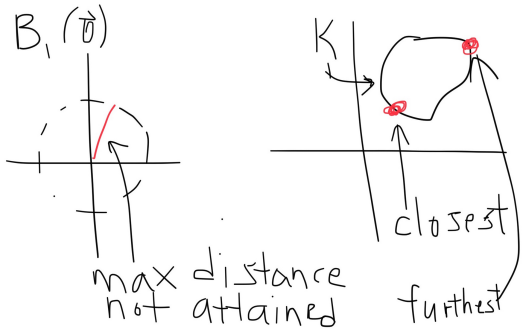
saying that we must take $c = 0$ in order for f to be continuous at the origin.

Extreme Value Theorem. We now give one important property of continuous functions, which is also the main reason for us as to why compactness will be an important concept. Suppose that $f : \mathbb{R}^n \to \mathbb{R}^m$ is continuous and that $K \subseteq \mathbb{R}^n$ is compact. The *Extreme Value Theorem* then says that f attains a maximum and a minimum when restricted to K; that is, there exist $\mathbf{p}, \mathbf{q} \in K$ such that

$$f(\mathbf{q}) \leq f(\mathbf{x}) \leq f(\mathbf{p})$$

for all $\mathbf{x} \in K$. (In this scenario, the maximum value is attained at $\mathbf{p}$ and the minimum value is attained at $\mathbf{q}$.) A proof of this fact belongs to the realm of real analysis. This theorem will have useful consequences for us when we talk about optimization and later integration.

Example. Here is one application of the Extreme Value Theorem. Given a compact subset K of $\mathbb{R}^2$, we claim that there is a point in K which is furthest away from the origin and a point which is closest to the origin. (There may be more than one such point in each case, so all we are saying is that there is *at least* one such point in each case.) This is not true without the assumption that K is compact: for instance, there is no point in the open ball $B_1(0,0)$ which is further away from the origin than any other point, essentially because the boundary of the ball is not included within it:



Let $f : \mathbb{R}^n \to \mathbb{R}$ be the function which sends a point to its norm: $f(\mathbf{x}) = \|\mathbf{x}\|$. This is continuous since

$$\|\mathbf{x}\| = \sqrt{x_1^2 + \cdots + x_n^2}$$

is made up of continuous expressions (i.e. polynomials and square roots), so the Extreme Value Theorem implies that it has a maximum and a minimum when restricted to K; the maximum gives the point further away from the origin, and the minimum the point closest to the origin.

Partial derivatives. We now move to understanding what it means for a function $f : \mathbb{R}^n \to \mathbb{R}^m$ to be differentiable, and what we mean by the *derivative* of f.

Suppose for instance we consider the function $f : \mathbb{R}^3 \to \mathbb{R}$ defined by

$$f(x, y, z) = xye^{xz} + y \sin z + x.$$

The first thing to say is that, if we simply try to compute derivatives as we always have, there are actually *three* derivatives we can compute: the derivative with respect to x, with respect to y, and with respect to z. These give the so-called *partial derivatives* of f, which are the ordinary single-variable derivatives obtained by differentiating with respect to one variable while holding the others constant. In our case, we get:

$$\begin{aligned}
\frac{\partial f}{\partial x} &= ye^{xz} + xyze^{xz} + 1 \\
\frac{\partial f}{\partial y} &= xe^{xz} + \sin x \\
\frac{\partial f}{\partial z} &= x^2 ye^{xz} + y \cos z
\end{aligned}$$

as the partial derivatives of this particular f, which, as mentioned above, are obtained by thinking of two variables as constant and differentiating with respect to the remaining variable. The symbol "∂" is pronounced "del", and distinguishes partial derivatives from derivatives of single-variable functions. (Note that this is the same symbol used to denote the boundary of a subset of $\mathbb{R}^n$—there are reasons for this, which we'll touch upon next quarter.) To be precise, the partial derivative of $f : \mathbb{R}^n \to \mathbb{R}$ with respect to x_i at $\mathbf{a} = (a_1, \dots, a_n)$ is given by the limit:

$$\frac{\partial f}{\partial x_i}(\mathbf{a}) = \lim_{h \to 0} \frac{f(a_1, \dots, a_i + h, \dots, a_n) - f(a_1, \dots, a_i, \dots, a_n)}{h},$$

which is the ordinary derivative at a_i of the single-variable function

$$g(x) = f(a_1, \dots, \underbrace{x}_{i\text{-th location}}, \dots, a_n)$$

obtained by holding every variable except for x_i constant.

So, partial derivatives give us some notion of "derivative", but the question to remains as to what we should think of as *the* derivative of f as a single object. There are three differential partial derivatives in this case, and there is no reason why one should be preferred over another as *the* derivative of f, so the answer has to involve more than partial derivatives alone. The derivative of f should be an object which, in particular, encodes *all* partial derivatives at once.

Single-variable derivatives revisited. To motivate the correct notion of derivative, let us return to single-variable derivatives first. A function $f : \mathbb{R} \to \mathbb{R}$ is differentiable at $\mathbf{x}$ if

$$\lim_{h \to 0} \frac{f(x + h) - f(x)}{h}$$

exists, in which we denote the value of this limit by $f'(x)$ and call it the derivative of f at x. So, our first guess might be to try to use the same limit in defining the derivative of a function $f : \mathbb{R}^n \to \mathbb{R}^m$. However, this is nonsense: for a function $f : \mathbb{R}^n \to \mathbb{R}^m$, the numerator

$$f(\mathbf{x} + \mathbf{h}) - f(\mathbf{x})$$

of the expression analogous to the one above is a vector in $\mathbb{R}^m$, while the denominator $\mathbf{h}$ is a vector in $\mathbb{R}^n$, and it makes no sense to divide vectors by one another, let alone ones which belong to spaces of different dimensions. Thus, we cannot define the derivative of f in as simple a way.

Instead, note that we can rewrite the limit expression

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

as

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

which comes from moving $f'(x)$ in the first expression to the left and then combining everything into the same limit and denominator. So, saying that $f:\mathbb{R}\to\mathbb{R}$ is differentiable at $x\in\mathbb{R}$ is the same as saying that there exists $c\in\mathbb{R}$ such that

$$\lim_{h\to 0}\frac{f(x+h)-f(x)-ch}{h}=0,$$

and this scalar c, if it exists, is then the derivative of f at x.

Trying to make sense of this expression when $f:\mathbb{R}^n\to\mathbb{R}^m$ seems to run into the same issues as before, in particular because the terms in the numerator again involve vectors of different dimensions: $f(\mathbf{x}+\mathbf{h})-f(\mathbf{x})$ in $\mathbb{R}^m$ and $c\mathbf{h}$ in $\mathbb{R}^n$. However, the fix is to replace c not just by a number, but with something which transforms vectors in $\mathbb{R}^n$ into vectors in $\mathbb{R}^m$; namely, a matrix! Even in the single-variable case, we can think of c not just as a number but instead as a 1×1 matrix, and then ch is the result of applying the corresponding linear transformation to h. Then, after using the fact that we can replace the limit above with the limit of the same expression only taking absolute values of the numerator and denominator (since an expression in $\mathbb{R}$ goes to 0 if and only if its absolute value goes to 0), we can get an expression which we can make sense of even when f maps $\mathbb{R}^n$ into $\mathbb{R}^m$.

Differentiability. We say that a function $f:U\subseteq\mathbb{R}^n\to\mathbb{R}^m$, defined on an open subset of $\mathbb{R}^n$, is *differentiable* at $\mathbf{x}\in U$ if there exists an $m\times n$ matrix A such that

$$\lim_{\mathbf{h}\to\mathbf{0}}\frac{f(\mathbf{x}+\mathbf{h})-f(\mathbf{x})-A\mathbf{h}}{\|\mathbf{h}\|}=0.$$

Note that since A is $m\times n$ and $\mathbf{h}\in\mathbb{R}^n$, $A\mathbf{h}\in\mathbb{R}^m$ so the numerator is the norm of a vector in $\mathbb{R}^m$, and it makes perfect sense to divide this by the length of a vector in $\mathbb{R}^n$ (the denominator) since this just involves dividing numbers. Also note that in the $A\mathbf{h}$ term, $\mathbf{h}$ is written as a column vector so that it makes sense to multiply it by A. When f is differentiable and this limit is 0, we say that the matrix A is *the* derivative of f at $\mathbf{x}\in U$. We say f is differentiable on U when it is differentiable at each point of U.

The upshot is that the derivative of a multivariable function is not just a number, but rather an entire matrix! We'll learn more about this matrix next time, but note again that this also makes sense in single-variable case: when $f:\mathbb{R}\to\mathbb{R}$, the derivative at a point will be a 1×1 matrix, which consists of just a single number, which is the ordinary derivative you're already used to.

Example. We show that the function $f:\mathbb{R}^2\to\mathbb{R}$ defined by

$$f(x,y)=x^2+y^2$$

is differentiable at $(0,1)$. We claim that the 1×2 matrix $A=\begin{pmatrix}0 & 2\end{pmatrix}$ satisfies the requirement in the definition. (Again, we'll see why this is the correct matrix to use next time.) Setting $\mathbf{h}=(h,k)$ and $\mathbf{x}=(0,1)$, the quotient used in the expression defining differentiability in this case is:

$$\frac{f(\mathbf{x}+\mathbf{h})-f(\mathbf{x})-A\mathbf{h}}{\|\mathbf{h}\|}=\frac{f(0+h,k+1)-f(0,1)-\begin{pmatrix}0 & 2\end{pmatrix}\begin{pmatrix}h\\k\end{pmatrix}}{\|(h,k)\|}$$

$$= \frac{h^2 + (k+1)^2 - 1^2 - 2k}{\sqrt{h^2+k^2}}$$
$$= \frac{h^2+k^2}{\sqrt{h^2+k^2}}$$
$$= \sqrt{h^2+k^2}.$$

Thus

$$\lim_{\mathbf{h}\to\mathbf{0}} \frac{f(\mathbf{x}+\mathbf{h}) - f(\mathbf{x}) - A\mathbf{h}}{\|\mathbf{h}\|} = \lim_{(h,k)\to(0,0)} \sqrt{h^2+k^2} = 0,$$

so f is differentiable at $(0,1)$ as claimed. The derivative of f at $(0,1)$ is thus the 1×2 matrix $\begin{pmatrix} 0 & 2 \end{pmatrix}$. (We'll see next time that in fact there can only be one matrix satisfying the property required in the definition of differentiable, which is why is makes sense to talk about the "the" derivative of f at a point as a unique thing.)

Linear (affine) approximations. We'll continue with more examples and further development of differentiability next time, but we finish with one more motivation for the definition. In the single-variable case, one often thinks of the derivative geometrically as the thing which tells you the slope of the tangent line at a point: the tangent line to the graph of f at a point x is

$$y = f(x) + f'(x)h,$$

where h is the variable. However, this can also be interpreted in a more "analytic" way as saying that the function

$$g(h) = f(x) + f'(x)h$$

provides the best "linear approximation" to f at x. (In other words, the tangent line approximation is the best linear approximation.)

In fact, this is precisely what the multivariable definition of differentiability provides as well. For a function $f : \mathbb{R}^n \to \mathbb{R}^m$, a "linear approximation" to f at $\mathbf{x} \in \mathbb{R}^n$ should be a function of the form

$$g(\mathbf{h}) = f(\mathbf{x}) + A\mathbf{h}$$

where A is an $m \times n$ matrix ($f(\mathbf{x})$ here is a constant vector) since such an expression is the higher dimensional analogues of $g(h) = f(x) + ah$. (In fact, we know from last quarter that such a function is not really "linear" in the linear-algebraic sense, but rather "affine", so that really we should be talking about the best *affine* approximation to f at $\mathbf{x}$. However, we'll stick with standard terminology and use the phrase "linear approximation" instead, but it is good to realize that a linear approximation is actually given by an affine transformation.) The derivative (as a matrix) A of f at $\mathbf{x}$ indeed characterizes the best linear approximation to f at $\mathbf{x}$, meaning that for "small" vectors $\mathbf{h}$, the value of $f(\mathbf{x}+\mathbf{h})$ is pretty to close to the value $f(\mathbf{x}) + A\mathbf{h}$.

Indeed, denote by $R(\mathbf{h})$ the "error" or "remainder" arising when approximating f near $\mathbf{x}$ by the linear approximation:

$$R(\mathbf{h}) = f(\mathbf{x}+\mathbf{h}) - [f(\mathbf{x}) + A\mathbf{h}].$$

Then to say that f is differentiable at $\mathbf{x}$ means precisely that this error term satisfies

$$\lim_{\mathbf{h}\to\mathbf{0}} \frac{R(\mathbf{h})}{\|\mathbf{h}\|} = 0.$$

Thus, to say that f at $\mathbf{x}$ is differentiable means that we can express f near $\mathbf{x}$ as:

$$f(\mathbf{x}+\mathbf{h}) = f(\mathbf{x}) + A\mathbf{h} + R(\mathbf{h}),$$

where the error $R(\mathbf{h})$ gets smaller and smaller as $\mathbf{x}+\mathbf{h}$ gets closer and closer to $\mathbf{h}$. We'll come back to this point of view when discussing multivariable *Taylor polynomials*. For now, the upshot is that, intuitively, differentiable functions are ones which can be "locally approximated" by matrices.

Lecture 22: Jacobian Matrices

Warm-Up 1. We show that the function $f : \mathbb{R}^2 \to \mathbb{R}^2$ defined by

$$f(x, y) = (x^2 + y^2, xy - y)$$

is differentiable at $\mathbf{a} = (0, 1)$, using the matrix $A = \left(\begin{smallmatrix} 0 & 2 \\ 1 & -1 \end{smallmatrix}\right)$ as the derivative of f at $\mathbf{a}$. Setting $\mathbf{h} = (h, k)$, the quotient of which we must take the limit when determining differentiability is:

$$\begin{aligned}
\frac{f(\mathbf{a}+\mathbf{h}) - f(\mathbf{a}) - A\mathbf{h}}{\|\mathbf{h}\|} &= \frac{f(0+h, 1+k) - f(0,1) - \left(\begin{smallmatrix} 0 & 2 \\ 1 & -1 \end{smallmatrix}\right)\left(\begin{smallmatrix} h \\ k \end{smallmatrix}\right)}{\|(h,k)\|} \\
&= \frac{(h^2 + (1+k)^2, h(1+k) - (1+k)) - (1,-1) - (2k, h-k)}{\sqrt{h^2+k^2}} \\
&= \frac{(h^2 + 1 + 2k + k^2 - 1 - 2k, h + hk - 1 - k + 1 - h + k)}{\sqrt{h^2+k^2}} \\
&= \frac{(h^2+k^2, hk)}{\sqrt{h^2+k^2}} \\
&= \left(\frac{h^2+k^2}{\sqrt{h^2+k^2}}, \frac{hk}{\sqrt{h^2+k^2}}\right).
\end{aligned}$$

Converting to polar coordinates $h = r\cos\theta, k = r\sin\theta$ shows that the limit of both components is 0 as $(h, k) \to (0, 0)$, so

$$\lim_{\mathbf{h}\to\mathbf{0}} \frac{f(\mathbf{a}+\mathbf{h}) - f(\mathbf{a}) - A\mathbf{h}}{\|\mathbf{h}\|} = \mathbf{0}$$

and hence f is differentiable at $\mathbf{a} = (0, 1)$.

Warm-Up 2. Suppose that $T : \mathbb{R}^n \to \mathbb{R}^m$ is a linear transformation. We show that T is differentiable at any $\mathbf{a} \in \mathbb{R}^n$, and the derivative of T at any $\mathbf{a}$ is the standard matrix A of T. This is the higher-dimensional analog of the fact that any function $f : \mathbb{R} \to \mathbb{R}$ of the form $f(x) = ax$ is differentiable with derivative at any point equal to a; here we are saying that $T(\mathbf{x}) = A\mathbf{x}$ is always differentiable with derivative A at any point.

Indeed, say that $T(\mathbf{x}) = A\mathbf{x}$. Since T is linear, the quotient whose limit defines differentiability at $\mathbf{a} \in \mathbb{R}^n$ is:

$$\frac{T(\mathbf{a}+\mathbf{h}) - T(\mathbf{a}) - A\mathbf{h}}{\|\mathbf{h}\|} = \frac{T\mathbf{a} + T\mathbf{h} - T\mathbf{a} - T\mathbf{h}}{\|\mathbf{h}\|} = \frac{\mathbf{0}}{\|\mathbf{h}\|} = \mathbf{0},$$

and the limit of $\mathbf{0}$ as $\mathbf{h} \to 0$ is of course $\mathbf{0}$. Hence T is differentiable at any $\mathbf{a}$ and the derivative at any point is A as claimed.

Observation. Note that a linear transformation in terms of components is explicitly of the form:

$$T(\mathbf{x}) = \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mn} \end{pmatrix} \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} a_{11}x_1 + \cdots + a_{1n}x_n \\ \vdots \\ a_{m1}x_1 + \cdots + a_{mn}x_n \end{pmatrix}.$$

Each component is thus a polynomial expression, which is always differentiable (as in the single-variable case, so it makes sense that T should be differentiable as well. Note that in this case the matrix A can be extracted from the explicit formula above by taking *partial derivatives*:

$$a_{ij} = \text{ the partial derivative of the } j\text{-th component with respect to } x_i.$$

Hence, the derivative of T, as a matrix, is formed by taking as entries all possible partial derivatives of T itself. This is true in general, and characterizes the matrix showing up in the definition of differentiability.

Jacobian matrices. Let $f : U \subseteq \mathbb{R}^n \to \mathbb{R}^m$ (where U is open in $\mathbb{R}^n$) be a function. This can be written as

$$f(\mathbf{x}) = (f_1(\mathbf{x}), \dots, f_m(\mathbf{x}))$$

where each f_i is a function $f_i : \mathbb{R}^n \to \mathbb{R}$, which all together describe the components of the result of $f(\mathbf{x})$. Suppose that all partial derivatives of $f_1, \dots, f_m$ exist at a point $\mathbf{a} \in U$. The *Jacobian matrix* of f at $\mathbf{a}$ is the matrix whose entries are the partial derivatives of the f_i evaluated at $\mathbf{a}$:

$$\begin{pmatrix} \frac{\partial f_1}{\partial x_1}(\mathbf{a}) & \cdots & \frac{\partial f_1}{\partial x_n}(\mathbf{a}) \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1}(\mathbf{a}) & \cdots & \frac{\partial f_m}{\partial x_n}(\mathbf{a}) \end{pmatrix}.$$

Note that each row focuses on specific component function of f, taking partial derivatives with respect to all variables as we move along that row. We denote the Jacobian matrix of f at $\mathbf{a}$ by $Df(\mathbf{a})$, which emphasizes the idea that $Df(\mathbf{a})$ should be thought of as *the* derivative of f at $\mathbf{a}$.

Proposition. The claim is that if $f : U \subseteq \mathbb{R}^n \to \mathbb{R}^m$ is differentiable at $\mathbf{a} \in U$, then

$$\lim_{\mathbf{h} \to \mathbf{0}} \frac{f(\mathbf{a}+\mathbf{h}) - f(\mathbf{a}) - Df(\mathbf{a})\mathbf{h}}{\|\mathbf{h}\|} = \mathbf{0},$$

so the matrix satisfying the required property in the definition of differentiable must be the Jacobian matrix. As a consequence, there can only be one matrix satisfying this property (since it must be the Jacobian matrix), and differentiability of f implies the existence of all the partial derivatives of its components. We'll prove this result next time, which just comes from analyzing the limit above along well-chosen curves approaching $\mathbf{0}$.

Existence of partials does not imply differentiability. As a warning, we point out that even though the existence of partial derivatives is needed in order to define the Jacobian matrix and hence check the definition of differentiable, the existence of these partials alone is not enough to guarantee differentiability.

An example of this is given by the function $f : \mathbb{R}^2 \to \mathbb{R}$ defined by

$$f(x, y) = \begin{cases} x & \text{if } |y| < |x| \\ -x & \text{otherwise.} \end{cases}$$

My Math 290-2 lecture notes (specifically, the notes from February 17, 2014) show that for this function, both partial derivatives at the origin exist:

$$\frac{\partial f}{\partial x}(0,0) = 1 \quad \text{and} \quad \frac{\partial f}{\partial y}(0,0) = 0,$$

so the Jacobian matrix at the origin is $Df(0,0) = \begin{pmatrix} 1 & 0 \end{pmatrix}$, and yet this function is not differentiable at the origin. Check those lecture notes to see this worked out: One minor point: those notes make reference to the so-called *tangent plane*, which we haven't spoken about, but it should not be hard to translate what we did there to what we're doing now.

As explained in those notes, visually the reason why f fails to be differentiable at the origin is that its graph has a "corner" point at the origin, as opposed to being "smooth" at the origin.

Lecture 23: More on Derivatives

Warm-Up. We show that the function $f : \mathbb{R}^2 \to \mathbb{R}$ defined by

$$f(x,y) = \begin{cases} \frac{x^3+y^3}{\sqrt{x^2+y^2}} & (x,y) \neq (0,0) \\ 0 & (x,y) = (0,0) \end{cases}$$

is differentiable on all of $\mathbb{R}^2$. First, at a non-origin point $(x,y) \neq (0,0)$, f is given by a quotient of differentiable functions with nonzero denominator, so it is differentiable at any such point. (To be sure, the numerator $x^3 + y^3$ is differentiable since it is a polynomial, and the denominator is the cube root of a differentiable expression, so it is differentiable itself away from the origin. Here we are taking for granted the fact that such square roots are differentiable; we will soon be able to derive this from the fact that the function f is what's called "C^1" away from the origin.)

Now we check differentiability at the origin. The partial derivatives of f at the origin are:

$$\frac{\partial f}{\partial x}(0,0) = \lim_{h \to 0} \frac{f(h,0) - f(0,0)}{h} = \lim_{h \to 0} \frac{h^3/\sqrt{h^2}}{h} = \lim_{h \to 0} \frac{h^2}{|h|} = 0$$

and

$$\frac{\partial f}{\partial y}(0,0) = \lim_{h \to 0} \frac{f(0,h) - f(0,0)}{h} = \lim_{h \to 0} \frac{h^3/\sqrt{h^2}}{h} = \lim_{h \to 0} \frac{h^2}{|h|} = 0.$$

The Jacobian matrix of f at the origin is thus $Df(0,0) = \begin{pmatrix} 0 & 0 \end{pmatrix}$. Then:

$$\frac{f(\mathbf{0}+\mathbf{h}) - f(\mathbf{0}) - Df(\mathbf{0})\mathbf{h}}{\|\mathbf{h}\|} = \frac{f(h,k) - f(0,0) - \begin{pmatrix} 0 & 0 \end{pmatrix}\begin{pmatrix} h \\ k \end{pmatrix}}{\sqrt{h^2+k^2}} = \frac{h^3+y^3}{h^2+k^2}.$$

After converting to polar coordinates the squeeze theorem shows that this expression has limit 0 as $\mathbf{h} \to \mathbf{0}$, so we conclude that f is differentiable at the origin as claimed.

Entries of Jacobian. We now prove the fact mentioned last time that if there is a matrix satisfying the property required in the definition of differentiable, it must be the Jacobian matrix. That is, suppose $f : U \subseteq \mathbb{R}^n \to \mathbb{R}^m$ is differentiable at $\mathbf{a} \in U$, so there exists an $m \times n$ matrix B such that

$$\lim_{\mathbf{h} \to \mathbf{0}} \frac{f(\mathbf{a}+\mathbf{h}) - f(\mathbf{a}) - B\mathbf{h}}{\|\mathbf{h}\|} = \mathbf{0}.$$

We claim that the entries of A are the partial derivatives of (the components of) f evaluated at $\mathbf{a}$.

Indeed, if the limit above exists, then approaching $\mathbf{0}$ along any specific direction must still give a limit of $\mathbf{0}$, and in particular approaching along the x_i-axis gives a limit of $\mathbf{0}$. Approaching along the x_i-axis means we approach using points of the form $\mathbf{h} = h\mathbf{e}_i$ as the scalar h goes to 0. For such $\mathbf{h}$, we have:

$$\lim_{h \to 0} \frac{f(\mathbf{a}+h\mathbf{e}_i) - f(\mathbf{a}) - B(h\mathbf{e}_i)}{\|h\mathbf{e}_i\|} = \mathbf{0}.$$

The numerator is:

$$f(\mathbf{a}+h\mathbf{e}_i)-f(\mathbf{a})-hA\mathbf{e}_i = f(a_1,\dots,a_i+h,\dots,a_n)-f(a_1,\dots,a_i,\dots,a_n)-h\begin{pmatrix} b_{1i} \\ \vdots \\ b_{mi}\end{pmatrix},$$

where in the first term we only add h to the i-th variable and where the vector on the right is the i-th column of B, while the denominator is $|h|$. Thus we have:

$$\lim_{h\to 0}\frac{f(a_1,\dots,a_i+h,\dots,a_n)-f(a_1,\dots,a_i,\dots,a_n)-h\begin{pmatrix} b_{1i} \\ \vdots \\ b_{mi}\end{pmatrix}}{|h|} = \mathbf{0}.$$

This is the limit of an expression in $\mathbb{R}^m$ (since $f=(f_1,\dots,f_m)$ has m components), and picking out only the j-th component of this expression gives:

$$\lim_{h\to \mathbf{0}}\frac{f_j(a_1,\dots,a_i+h,\dots,a_n)-f_j(a_1,\dots,a_i,\dots,a_n)-b_{ji}h}{|h|} = 0,$$

where $b_{ji}h$ is the j-th component of

$$h\begin{pmatrix} b_{1i} \\ \vdots \\ b_{mi}\end{pmatrix}.$$

Since this is a limit which equals zero, we get the same limit if we omit absolute values, so

$$\lim_{h\to \mathbf{0}}\frac{f_j(a_1,\dots,a_i+h,\dots,a_n)-f_j(a_1,\dots,a_i,\dots,a_n)-b_{ji}h}{h} = 0.$$

Breaking this expression up into two fractions and rearranging terms gives:

$$\lim_{h\to \mathbf{0}}\frac{f_j(a_1,\dots,a_i+h,\dots,a_n)-f_j(a_1,\dots,a_i,\dots,a_n)}{h} = b_{ji}.$$

The left side is precisely the definition of the partial derivative of f_j with respect to x_i at $\mathbf{a}$, so this partial derivative exists and:

$$\frac{\partial f_j}{\partial x_i}(\mathbf{a}) = b_{ji}.$$

This shows that the entries of B are the partial derivatives of the components of f at $\mathbf{a}$, so $B=Df(\mathbf{a})$ as claimed.

Geometric meaning of partials. We'll talk about the "geometric meaning" behind the Jacobian matrix itself later, but here we note partial derivatives themselves have a simple geometric interpretation, which indeed is the ordinary geometric interpretation of single-variable derivatives:

$$\frac{\partial f_j}{\partial x_i}(\mathbf{a}) = \text{ the slope of the graph of } f_j \text{ in the } x_i\text{-direction at the point } \mathbf{a}.$$

Check my Math 290-2 lecture notes (February 14, 2014) for a further explanation of this. Equivalently, partial derivatives give the rate of change of a function in a direction parallel to one of the axes. We'll later talk about the notion of a *directional derivative*, which gives the rate of change (or slope) in an arbitrary direction.

Definition of C^1. Clearly, having to check the formal definition of differentiability every time we wanted to determine whether or not a function was differentiable would be tedious. However, in many cases this is not required, since a simple requirement on the partial derivatives always implies differentiability, namely when they are continuous.

We say that a function $f : U \subseteq \mathbb{R}^n \to \mathbb{R}^m$ is C^1 if all partial derivatives of all components of f exist and are continuous throughout U. The basic fact is a C^1 function is always differentiable, so even though existence of partial derivatives alone is not enough to guarantee differentiability (which we saw in an example last time), having the partials be continuous on top of this *does* guarantee differentiability. Note however that C^1 is not equivalent to differentiability: a differentiable function need not have continuous partial derivatives.

Before giving a sense as to why this is true, let us point out that we can rephrase the definition of C^1 in the following way. If all partial derivatives of $f : \mathbb{R}^n \to \mathbb{R}^m$ exist, then the Jacobian matrix of f exists at every point. We can then consider the function

$$Df : \mathbb{R}^n \to M_{mn}(\mathbb{R}), \quad \mathbf{x} \mapsto Df(\mathbf{x})$$

which sends a point of $\mathbb{R}^n$ to the Jacobian matrix of f at that point. Thinking of the space of $m \times n$ matrices as being the same as $\mathbb{R}^{mn}$ after we identity a matrix with the vector containing all of its entries (more precisely, the space of $m \times n$ matrices is isomorphic to $\mathbb{R}^{mn}$), we can interpret Df s a function

$$Df : \mathbb{R}^n \to \mathbb{R}^{mn}.$$

To say that f is C^1 means that all components of this function Df are continuous since these components are precisely the partial derivatives of f, and hence saying that f is C^1 means that the function Df is itself continuous. Once we know that f is differentiable, this function Df should be thought of as being *the* derivative of f (analogously to how we interpret the derivative of a single-variable function $f : \mathbb{R} \to \mathbb{R}$ itself as a function $f' : \mathbb{R} \to \mathbb{R}$), so C^1 means that the derivative of f is itself continuous. Because of this, it is also common to say that f is *continuously differentiable* when it is C^1.

C^1 implies differentiable. We now give an idea as to why you should believe that continuity of partial derivatives implies differentiability. A full proof of this can be found in the book, but here we will only get to the point which shows where continuity of the partials comes into play. To simplify matters, we will only do this in the case of a function $f : \mathbb{R}^2 \to \mathbb{R}$. The one fact we need from single-variable calculus is the *Mean Value Theorem*: if $g : \mathbb{R} \to \mathbb{R}$ is differentiable and $a, b \in \mathbb{R}$, there exists c between a and b such that $g(a) - g(b) = g'(c)(a - b)$.

Suppose that $f : \mathbb{R}^2 \to \mathbb{R}$ is C^1, meaning the $\frac{\partial f}{\partial x}$ and $\frac{\partial f}{\partial y}$ exist and are continuous everywhere, and let $\mathbf{a} = (a, b) \in \mathbb{R}^2$. Setting $\mathbf{h} = (h, k)$, the numerator of the fraction whose limit defines differentiability is:

$$f(\mathbf{a} + \mathbf{h}) - f(\mathbf{a}) - Df(\mathbf{a})\mathbf{h} = f(a + h, b + k) - f(a, b) - \begin{pmatrix} \frac{\partial f}{\partial x}(a, b) & \frac{\partial f}{\partial y}(a, b) \end{pmatrix} \begin{pmatrix} h \\ k \end{pmatrix}.$$

Now, we can rewrite the difference of the first two terms as

$$f(a + h, b + k) - f(a, b) = f(a + h, b + k) - f(a + h, b) + f(a + h, b) - f(a, b),$$

where we use the age-old trick of adding and subtracting the same term, so that overall we've simply added zero. Note that in the first two terms, the first variable $a + h$ is the same and only the

second variable varies. Thus, applying the single-variable Mean Value Theorem to this expression (viewed as a single-variable function of the second variable alone) gives:

$$f(a+h,b+k)-f(a+h,b)=\frac{\partial f}{\partial y}(a+h,c)(b+k-b)=\frac{\partial f}{\partial y}(a+h,c)k$$

for some c between b and $b+h$. Similarly, applying the Mean Value Theorem to $f(a+h,b)-f(a,b)$ viewed as an expression of only the first variable alone gives:

$$f(a+h,b)-f(a,b)=\frac{\partial f}{\partial x}(d,b)h$$

for some d between a and $a+h$. Thus

$$f(a+h,b+h)-f(a+h,b)+f(a+h,b)-f(a,b)=\frac{\partial f}{\partial y}(a+h,c)k+\frac{\partial f}{\partial x}(d,b)h,$$

which can be written as the matrix product

$$\begin{pmatrix}\frac{\partial f}{\partial x}(d,b) & \frac{\partial f}{\partial y}(a+h,c)\end{pmatrix}\begin{pmatrix}h\\k\end{pmatrix}.$$

Hence the numerator in the limit defining differentiability is

$$f(\mathbf{a}+\mathbf{h})-f(\mathbf{a})-Df(\mathbf{a})\mathbf{h}=\begin{pmatrix}\frac{\partial f}{\partial x}(d,b) & \frac{\partial f}{\partial y}(a+h,c)\end{pmatrix}\begin{pmatrix}h\\k\end{pmatrix}-\begin{pmatrix}\frac{\partial f}{\partial x}(a,b) & \frac{\partial f}{\partial y}(a,b)\end{pmatrix}\begin{pmatrix}h\\k\end{pmatrix}.$$

Now the idea is that since d is between a and $a+h$, $d\to a$ as $h\to 0$, and since c is between b and $b+k$, $c\to b$ as $k\to 0$. Thus since $\frac{\partial f}{\partial x}$ and $\frac{\partial f}{\partial y}$ are continuous:

$$\lim_{(h,k)\to(0,0)}\frac{\partial f}{\partial x}(d,b)=\frac{\partial f}{\partial x}(a,b)\text{ and }\lim_{(h,k)\to(0,0)}\frac{\partial f}{\partial y}(a+h,c)=\frac{\partial f}{\partial y}(a,b),$$

which says that as $\mathbf{h}\to\mathbf{0}$

$$\begin{pmatrix}\frac{\partial f}{\partial x}(d,b) & \frac{\partial f}{\partial y}(a+h,c)\end{pmatrix}\to\begin{pmatrix}\frac{\partial f}{\partial x}(a,b) & \frac{\partial f}{\partial y}(a,b)\end{pmatrix},$$

which will then imply (after some work) that the limit defining differentiability is zero. Again, this is far from an actual proof, but is only meant to show where continuity of the partial derivatives comes in: it is used to to show that $f(\mathbf{a}+\mathbf{h})-f(\mathbf{a})$ will approach $Df(\mathbf{a})\mathbf{h}$ as $\mathbf{h}\to\mathbf{0}$ in a way which will force the overall limit to be zero. As stated earlier, check the book (or ask me) for more details behind the full proof if interested.

Lecture 24: Second Derivatives

Warm-Up. This warm-up is just for fun, and is meant to illustrate how multivariable derivatives show up in other settings. This specific type of problem is not something you'd be responsible for.

Consider the function $f:M_n(\mathbb{R})\to M_n(\mathbb{R})$ defined by

$$f(X)=X^2\text{ for }X\in M_n(\mathbb{R}).$$

In other words, f is the function from matrices to matrices which sends a matrix to its square. We claim that f is differentiable, and that its derivative has an explicit description. The main

thing to understand is the following: what does it mean to say that a function mapping matrices to matrices is differentiable? If we think of an $n \times n$ matrix as being a vector in $\mathbb{R}^{n^2}$ (i.e. form a very long vector whose entries are the entries of the matrix) then we can interpret f as a function $f : \mathbb{R}^{n^2} \to \mathbb{R}^{n^2}$, and we are saying that *this* function is differentiable. For instance, in the $n = 2$ case we have:

$$\begin{pmatrix} x & y \\ z & w \end{pmatrix}^2 = \begin{pmatrix} x & y \\ z & w \end{pmatrix} \begin{pmatrix} x & y \\ z & w \end{pmatrix} = \begin{pmatrix} x^2 + yz & xy + yw \\ xz + wz & zy + w^2 \end{pmatrix},$$

so the function $f : \mathbb{R}^4 \to \mathbb{R}^4$ corresponding to this is

$$f(x, y, z, w) = (x^2 + yz, xy + yw, xz + wz, zy + w^2).$$

This function is differentiable everywhere since, for instance, all of its partial derivatives exist and are continuous everywhere since they are all polynomials.

However, the point is that we can interpret differentiability of $f : M_n(\mathbb{R}) \to M_n(\mathbb{R})$ without making using of the isomorphism $M_n(\mathbb{R}) \cong \mathbb{R}^{n^2}$. The claim is that f is differentiable and its derivative at $X \in M_n(\mathbb{R})$ is the linear transformation $Df(X) : M_n(\mathbb{R}) \to M_n(\mathbb{R})$ defined by

$$H \mapsto XH + HX.$$

The idea is that, just as the derivative of a function $\mathbb{R}^n \to \mathbb{R}^m$ at a point is an $m \times n$ matrix, which gives a linear transformation $\mathbb{R}^n \to \mathbb{R}^m$, the derivative of a function $M_n(\mathbb{R}) \to M_n(\mathbb{R})$ should be a linear transformation $M_n(\mathbb{R}) \to M_n(\mathbb{R})$. To say that f is differentiable at X with this derivative then means that:

$$\lim_{H \to 0} \frac{f(X + H) - f(X) - Df(X)H}{\|H\|} = 0,$$

which mimics the definition of differentiable for a function $\mathbb{R}^n \to \mathbb{R}^m$. Now, we won't be able to make this fully rigorous, since for instance we haven't defined what it means to take the limit of a *matrix* expression as a "matrix" goes to 0, nor have we defined what the norm of a matrix is. This can all be made completely precise, but that is better left to another course. (Topological notions, in particular open and closed sets, can also be precise in the setting of spaces of matrices.)

Given the proposed derivative $Df(X)$ above, we have:

$$f(X + H) - f(X) - Df(X)H = (X + H)^2 - X^2 - (XH + HX) = H^2.$$

Thus the limit above becomes

$$\lim_{H \to 0} \frac{H^2}{\|H\|} = \lim_{H \to 0} \left(\frac{H}{\|H\|} \right) H.$$

The idea is that the term in parentheses is "bounded" so a version of the squeeze theorem will imply that this limit is indeed 0. Again, making this all precise would require knowing more about limits in other contexts, but it is conceivable that this type of argument should work out. Thus f is differentiable at any X and the derivative at X is the linear transformation $H \mapsto XH + HX$.

Note what this says in the case of 1×1 matrices. In this case, f is a function $f : \mathbb{R} \to \mathbb{R}$ of the type covered in a single-variable calculus course. The derivative at $x \in \mathbb{R}$ derived above is the linear transformation $Df(x) : \mathbb{R} \to \mathbb{R}$ defined by

$$h \mapsto xh + hx.$$

However, the nice thing is that 1×1 matrices commute, so this becomes

$$h \mapsto 2xh,$$

which has standard matrix given by the 1×1 matrix $\begin{pmatrix} 2x \end{pmatrix}$. Thus we get that the derivative of f at x is the matrix $\begin{pmatrix} 2x \end{pmatrix}$, which agrees with the ordinary single-variable derivative of $f(x) = x^2$ we all know and love. Note that, in some sense, the derivative of $f(X) = X^2$ is also "$X + X''$, only that one copy of H is multiplied on the left by H and the other copy on the right.

We'll mention one more example of this type. Consider the function $g : U \to M_n(\mathbb{R})$, where $U \subseteq M_n(\mathbb{R})$ is the space of $n \times n$ *invertible* matrices, defined by

$$g(A) = A^{-1}.$$

(The space of invertible matrices turns out to be *open* in the space of all matrices, so it is a valid domain for a differentiable function.) Then g is differentiable and its derivative at $A \in U$ is the linear transformation $M_n(\mathbb{R}) \to M_n(\mathbb{R})$ given by

$$H \mapsto -A^{-1}HA^{-1}.$$

In the $n = 1$ case, this becomes

$$h \mapsto -a^{-1}ha^{-1} = -\frac{1}{a^2}h,$$

so the derivative is given by the 1×1 matrix $\begin{pmatrix} -\frac{1}{a^2} \end{pmatrix}$, which agrees with the usual single-variable derivative of the function $g(x) = \frac{1}{x}$ evaluated at $a \neq 0$.

Second derivatives. As mentioned last time, the derivative of a differentiable function $f : \mathbb{R}^n \to \mathbb{R}^m$ can be viewed as the function

$$Df : \mathbb{R}^n \to M_{mn}(\mathbb{R}),\ \mathbf{x} \mapsto Df(\mathbf{x})$$

which sends a point to the value of the Jacobian matrix of f at that point. Using the isomorphism $M_{mn}(\mathbb{R}) \to \mathbb{R}^{mn}$, we can now ask whether Df is itself differentiable. We say that f is *twice-differentiable* when Df is differentiable, and we refer to the derivative of Df as the *second derivative* of f. This second derivative is the map which sends a point to the "Jacobian matrix of the Jacobian matrix" of f at that point, which gets difficult to think about in general.

Instead, we will restrict ourselves to considering functions $f : \mathbb{R}^n \to \mathbb{R}$. In this case, $Df(\mathbf{x})$ is a $1 \times n$ matrix at each $\mathbf{x} \in \mathbb{R}^n$, so the derivative of f is a map

$$Df : \mathbb{R}^n \to \mathbb{R}^n.$$

The derivative of Df at $\mathbf{x}$ is the $n \times n$ matrix $D^2f(\mathbf{x})$ given by the Jacobian matrix of the Jacobian matrix of f at $\mathbf{x}$:

$$D^2f(\mathbf{x}) = D(Df)(\mathbf{x}).$$

(We'll see what this looks like in an explicit example in a bit, which will make this notation clearer.) The matrix $D^2f(\mathbf{x})$ is called the *Hessian* matrix of f at $\mathbf{x}$, and should be viewed as the "second derivative" of f at $\mathbf{x}$.

Example. Consider the function $f : \mathbb{R}^2 \to \mathbb{R}$ defined by

$$f(x, y) = x^2y + xe^{xy}.$$

The Jacobian matrix is given by

$$Df(x, y) = \begin{pmatrix} 2xy + e^{xy} + xye^{xy} & x^2 + x^2e^{xy} \end{pmatrix}.$$

Viewing this as giving a function $Df : \mathbb{R}^2 \to \mathbb{R}^2$, the Jacobian matrix of Df, or in other words the Hessian matrix of f, is:

$$D^2 f(x,y) = \begin{pmatrix} 2y + ye^{xy} + ye^{xy} + xy^2 e^{xy} & 2x + xe^{xy} + xe^{xy} + x^2 ye^{xy} \\ 2x + 2xe^{xy} + x^2 ye^{xy} & x^3 e^{xy} \end{pmatrix}.$$

Note that this matrix is symmetric—this is no accident, and is a consequence of what's called *Clairaut's Theorem*, which we'll state in a bit.

Second and higher-order partials. Going back to a function $f : \mathbb{R}^n \to \mathbb{R}$, the Jacobian is:

$$Df = \begin{pmatrix} \frac{\partial f}{\partial x_1} & \cdots & \frac{\partial f}{\partial x_n} \end{pmatrix}.$$

The entries of the Hessian of f are obtained by differentiating each of these components with respect to some x_j, which gives the so-called *second order partial derivatives* of f:

$$\frac{\partial^2 f}{\partial x_j \partial x_i} := \frac{\partial}{\partial x_j}\left(\frac{\partial f}{\partial x_i}\right).$$

In other words, the second order partial derivative $\frac{\partial^2 f}{\partial x_j \partial x_i}$ is obtained by first differentiating f with respect to x_i and then differentiating the result with respect to x_j. Another common notation for this is

$$f_{x_i x_j}$$

where the order of the subscripts indicates the order in which we differentiate. Note that the order of the variables in the notation $f_{x_i x_j}$ is opposite the order in $\frac{\partial^2 f}{\partial x_j \partial x_i}$; nowadays most people use the notation

$$\frac{\partial^2 f}{\partial x_j \partial x_i},$$

which emphasizes the idea that this meant to be what is obtained when applying the differentiation operator $\frac{\partial}{\partial x_j}$ to the function $\frac{\partial f}{\partial x_i}$. In the special case where $x_i = x_j$, so when we differentiate with respect to the same variable twice, the notation is simplified to:

$$\frac{\partial^2 f}{\partial x_i^2}.$$

The Hessian of f thus looks like:

$$D^2 f = \begin{pmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_2 \partial x_1} & \cdots & \frac{\partial^2 f}{\partial x_n \partial x_1} \\ \frac{\partial^2 f}{\partial x_1 \partial x_2} & \frac{\partial^2 f}{\partial x_2^2} & \cdots & \frac{\partial^2 f}{\partial x_n \partial x_2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_1 \partial x_n} & \frac{\partial^2 f}{\partial x_2 \partial x_n} & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{pmatrix},$$

and so is the matrix encoding all possible second order partial derivatives. Note that first row derivatives contains all partial derivatives of f_{x_1}, the second row all partial derivatives of f_{x_2}, and so on. We can keep going and talk about *third order* partial derivatives, and higher order partial derivatives, but these can no longer be easily encoded by single matrices.

Definition of C^2 and C^k. We say that a function $f : U \subseteq \mathbb{R}^n \to \mathbb{R}^m$ is C^2 if all second order partial derivatives of all components of f exist and are continuous throughout U. More generally, f is C^k is all k-th order partial derivatives exist and are continuous throughout U.

Clairaut's Theorem. Now we can explain the observation noticed before that Hessians are often symmetric. The fact is that if $f : U \subseteq \mathbb{R}^n \to \mathbb{R}$ is C^2, then

$$\frac{\partial^2 f}{\partial x_j \partial x_i} = \frac{\partial^2 f}{\partial x_i \partial x_j} \text{ for all } i, j.$$

Thus, if all second order partial derivatives are in fact continuous, then the so-called *mixed* partial derivatives (the ones where we differentiate with respect to the same two variables only in different orders) are the same. This is a highly non-obvious fact, and the proof (although not very difficult) is not worth giving in this course. You can check the book or ask in office hours if you're interested.

The equality of mixed partial derivatives then says that the Hessian is symmetric. Indeed, $\frac{\partial^2 f}{\partial x_j \partial x_i}$ is the entry in the j-th column and i-th row of $D^2 f$, and $\frac{\partial^2 f}{\partial x_i \partial x_j}$ is the entry in the i-th column and j-th row. We'll see next quarter what consequences we can derive from the fact that Hessians are symmetric.

Geometric meaning of second derivatives. Finally, we give the geometric interpretation of second order partial derivatives. Recall that first order partial derivatives of $f : \mathbb{R}^n \to \mathbb{R}$ give slopes in directions parallel to one of the coordinate axes. The second order partial derivative

$$\frac{\partial^2 f}{\partial x_j \partial x_i}$$

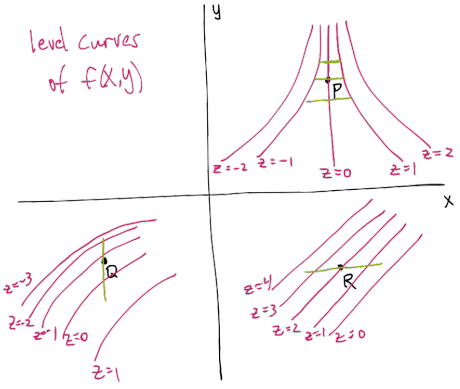
then gives the rate of change of the quantity $\frac{\partial f}{\partial x_i}$ with respect to x_j, or in other words is measures the rate at which the slopes in the x_i-direction change as we move in the x_j-direction. Thus, when differentiating with respect to the same variable twice,

$$\frac{\partial^2 f}{\partial x_i^2}$$

measures the *concavity* of the graph of f in the x_i-direction, mimicking the geometric interpretation of second derivatives in single-variable calculus. The mixed second order partials are a bit tougher to interpret in this way, but we'll look at an example next time that illustrates what it means to look at how the "slope in one direction change as you move in another direction."

Lecture 25: The Chain Rule

Warm-Up. Suppose we are given level curves of a C^2 function f as follows:



We want to determine the signs of some second-order partial derivatives. First we consider $f_{xx}(R)$, which is

$$\frac{\partial}{\partial x}\left(\frac{\partial f}{\partial x}\right)(R).$$

This gives the rate of change in the x-direction of $\frac{\partial f}{\partial x}$, so in other words the rate of change in the x-direction of the slope in the x-direction. Imagine moving horizontally through the point R. The slope in the x-direction at R is negative since z decreases moving horizontally through R, and the same is true a bit to the left of R and a bit to the right. Now, the equal spacing between the level curves tells us that the negative slope at the point R is the same as the negative slope a bit to the left and the same as the negative slope a bit to the right, so the slope in the x-direction $\frac{\partial f}{\partial x}$ stays constant as we move through R in the x-direction. Thus

$$f_{xx}(R) = \frac{\partial^2 f}{\partial x^2}(R) = \frac{\partial}{\partial x}\left(\frac{\partial f}{\partial x}\right)(R) = 0$$

since $\frac{\partial f}{\partial x}$ does not change with respect to x. Geometrically, the graph of f in the x-direction at R looks like a straight line, so it has zero concavity.

Next we look at $f_{yy}(Q)$, which is the rate of change in the y-direction of the slope in the y-direction. At Q the slope in the y-direction is negative since z is decreasing vertically through Q, and the slope in the y-direction is also negative a bit below Q as well as a bit above Q. However, the level curves here are not equally spaces: below Q is takes a longer distance to decrease by a height of 1 than it does at Q, so the slope in the y-direction below Q is a little less negative than it is at Q itself. Similarly, above Q the slope in the y-direction is even more negative than it is at Q since it takes a shorter distance to decrease by a height of 1. Thus moving vertically through Q, the slope in the y-direction gets more and more negative, so $\frac{\partial f}{\partial y}$ is decreasing with respect to y at Q, meaning that

$$f_{yy}(Q) = \frac{\partial^2 f}{\partial y^2}(Q) = \frac{\partial}{\partial y}\left(\frac{\partial f}{\partial y}\right)(Q) < 0.$$

Geometrically, the graph of f at Q in the y-direction is concave down since the downward slope gets steeper and steeper.

Finally we look at $f_{xy}(P)$, which is the rate of change in the y-direction of the slope in the x-direction. At P the slope in the x-direction is positive since z increases when moving horizontally through P. Now, a bit below P the slope in the x-direction is also positive but not has positive as it is at P since it takes a longer distance to increase the height than it does at P. A bit above P it takes an even shorter distance to increase the height in the x-direction, so $\frac{\partial f}{\partial x}$ is larger above P than it is at P. Hence the slope $\frac{\partial f}{\partial x}$ in the x-direction is increasing (getting more and more positive) as you move through P in the y-direction, so

$$f_{xy}(P) = \frac{\partial^2 f}{\partial y \partial x}(P) = \frac{\partial}{\partial y}\left(\frac{\partial f}{\partial x}\right)(P) > 0.$$

Then by Clairaut's Theorem, $f_{yx}(P)$ is also positive, which we can also figure out be looking at how the slopes of the graph in the y-direction change as we move in the x-direction at P.

Rates of change. Before moving on, we give an interpretation of a Jacobian matrix as an "infinitesimal rate of change", analogous to the similar interpretation of single-variable derivatives. Indeed, suppose that $f : \mathbb{R}^n \to \mathbb{R}^m$ is differentiable at $\mathbf{a} \in \mathbb{R}^n$. The idea that the Jacobian matrix $Df(\mathbf{a})$ provides the best linear approximation to f near $\mathbf{a}$ says that for $\mathbf{h}$ "close" to $\mathbf{0}$, we have

$$f(\mathbf{a}+\mathbf{h}) \approx f(\mathbf{a}) + Df(\mathbf{a})\mathbf{h}.$$

(Note that if $\mathbf{h}$ is "small", then $\mathbf{a}+\mathbf{h}$ is "close" to $\mathbf{a}$.) Rewrite this as

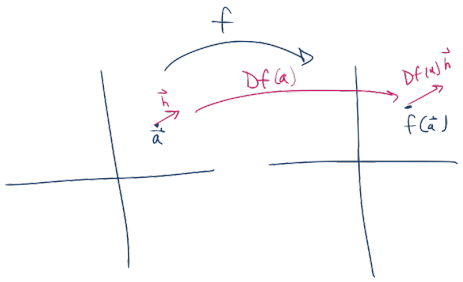
$$f(\mathbf{a}+\mathbf{h}) - f(\mathbf{a}) \approx Df(\mathbf{a})\mathbf{h},$$

and note that on the right side, $\mathbf{h}$ is the difference in the inputs $\mathbf{a}+\mathbf{h}$ and $\mathbf{a}$ showing up on the left. Thus, this says that the change in outputs $f(\mathbf{a}+\mathbf{h}) - f(\mathbf{a})$ can be approximated via the corresponding change in inputs $\mathbf{h} = (\mathbf{a}+\mathbf{h}) - \mathbf{a}$ via the "derivative" of f at $\mathbf{a}$:

$$\text{(small change in outputs)} \approx Df(\mathbf{a})\text{(small change in inputs)}.$$

From this point of view, as $\mathbf{h} \to \mathbf{0}$, so that $\mathbf{a}+\mathbf{h} \to \mathbf{a}$, $Df(\mathbf{a})$ indeed provides the "infinitesimal" rate of change of f at $\mathbf{a}$. The point is that while we do not have a straightforward geometric interpretation of $Df(\mathbf{a})$ as a "slope", we certainly have an interpretation of $Df(\mathbf{a})$ as a rate of change.

Infinitesimal transformations. Even though we cannot interpret $Df(\mathbf{a})$ as a slope, we can still give it a geometric interpretation based on the description above. Indeed, think of $\mathbf{h} = (\mathbf{a}+\mathbf{h}) - \mathbf{a}$ (for very small $\mathbf{h}$) as describing an "infinitesimal vector" at the point $\mathbf{a}$. (We will not make the notion of "infinitesimal" precise and only use this for the sake of intuition. There is, however, a way to make this fully precise, as you would likely see in a course on differential geometry.) Then $f(\mathbf{a}+\mathbf{h}) - f(\mathbf{a})$ is in some sense an "infinitesimal vector" at $f(\mathbf{a})$, and the point is that $Df(\mathbf{a})$ is the transformation which describes how infinitesimal vectors are transformed under f:



In this setting, we say that $Df(\mathbf{a}) : \mathbb{R}^n \to \mathbb{R}^m$, viewed as a linear transformation, is the *infinitesimal transformation* induced by f at $\mathbf{a}$. This is as close to an interpretation of $Df(\mathbf{a})$ as a "slope" as we're going to get in general.

Chain Rule. With the setup above, the multivariable chain rule now makes total sense. The statement is that if $g : \mathbb{R}^n \to \mathbb{R}^m$ is differentiable at $\mathbf{a}$ and $f : \mathbb{R}^m \to \mathbb{R}^k$ are differentiable at $g(\mathbf{a})$, then the composition $f \circ g : \mathbb{R}^n \to \mathbb{R}^k$ is differentiable at $\mathbf{a}$ and

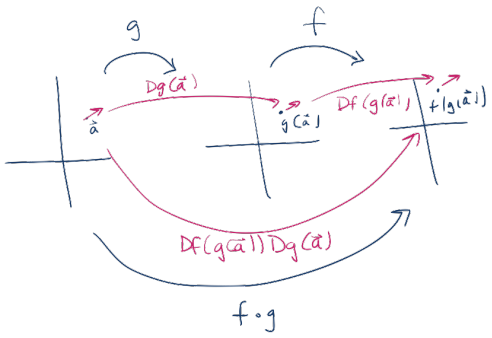
$$D(f \circ g)(\mathbf{a}) = Df(g(\mathbf{a}))Dg(\mathbf{a}).$$

To be clear, $D(f \circ g)(\mathbf{a})$ is a $k \times n$ matrix, and the claim is that this Jacobian matrix is the product of the $k \times m$ matrix $Df(g(\mathbf{a}))$ and the $m \times n$ matrix $Dg(\mathbf{a})$. Note that when $n = m = 1$, so g and f are both single-variable, each of these Jacobian matrices is just a scalar (a 1×1 matrix) and this version of the chain rule gives:

$$(f \circ g)(a) = f'(g(a))g'(a),$$

which is the ordinary single-variable chain rule.

We will not prove the chain rule in this course as the proof is quite involved and requires more experience with analysis, but the book has a proof if you're interested in seeing one. I do claim, however, that the chain rule is completely intuitive from the point of view of a Jacobian matrix as an infinitesimal rate of change or as an infinitesimal transformation. Indeed, take an infinitesimal vector $\mathbf{h}$ at $\mathbf{a}$. Then $Dg(\mathbf{a})$ transforms this into an infinitesimal vector $Dg(\mathbf{a})\mathbf{h}$ at $g(\mathbf{a})$, which in turn is transformed by $Df(g(\mathbf{a}))$ into an infinitesimal vector $Df(g(\mathbf{a}))Dg(\mathbf{a})\mathbf{h}$ at $f(g(\mathbf{a}))$:



But of course, this final infinitesimal vector should be the same as the one obtained by applying the infinitesimal transformation corresponding to $f \circ g$ to the original $\mathbf{h}$:

$$D(f \circ g)(\mathbf{a})\mathbf{h} = Df(g(\mathbf{a}))Dg(\mathbf{a})\mathbf{h}.$$

Since this is true for all infinitesimal vectors $\mathbf{h}$, the matrix $D(f \circ g)(\mathbf{a})$ must be the same as the matrix $Df(g(\mathbf{a}))Dg(\mathbf{a})$, which is the statement of the chain rule.

Example. Suppose $g : \mathbb{R}^2 \to \mathbb{R}^2$ is defined by

$$g(x, y) = (x^2 y, y + \cos x)$$

and $f : \mathbb{R}^2 \to \mathbb{R}^3$ by

$$f(u, v) = (u + v, uv, v^2).$$

The composition $f \circ g$ is

$$(f \circ g)(x, y) = f(g(x, y)) = f(x^2 y, y + \cos x) = (x^2 y + y + \cos x, x^2 y(y + \cos x), (y + \cos x)^2).$$

From this we can determine the Jacobian matrix $D(f \circ g)(x, y)$, or we can instead use the chain rule. Since

$$Df(u, v) = \begin{pmatrix} 1 & 1 \\ v & u \\ 0 & 2v \end{pmatrix} \quad \text{and} \quad Dg(x, y) = \begin{pmatrix} 2xy & x^2 \\ -\sin x & 1 \end{pmatrix},$$

we get

$$\begin{aligned} D(f \circ g)(x, y) &= Df(g(x, y))Dg(x, y) \\ &= \begin{pmatrix} 1 & 1 \\ y + \cos x & x^2 y \\ 0 & 2y + 2\cos x \end{pmatrix} \begin{pmatrix} 2xy & x^2 \\ -\sin x & 1 \end{pmatrix} \\ &= \begin{pmatrix} 2xy - \sin x & x^2 + 1 \\ 2xy(y + \cos x) - x^2 y \sin x & x^2 y + x^2 \cos x + x^2 y \\ -\sin x(2y + 2\cos x) & 2y + 2\cos x \end{pmatrix}. \end{aligned}$$

Chain rule via partial derivatives. The chain rule as stated is a compact, succinct way to express information about the Jacobian of a composition, but often times what matters more in practice is knowing how to express partial derivatives taken with respect to one set of variables in terms of another set of variables.

Suppose $f : \mathbb{R}^2 \to \mathbb{R}$ is a function of (x, y) and $g : \mathbb{R}^2 \to \mathbb{R}^2$ is a function of variables (u, v):

$$g(u, v) = (x(u, v), y(u, v)).$$

Then we can think of $f \circ g$ as expressing f in terms of u and v instead:

$$f(g(u, v)) = f(x(u, v), y(u, v)).$$

The chain rule gives:

$$D(f \circ g) = (Df)(Dg), \text{ or } \begin{pmatrix} \frac{\partial f}{\partial u} & \frac{\partial f}{\partial v} \end{pmatrix} = \begin{pmatrix} \frac{\partial f}{\partial x} & \frac{\partial f}{\partial y} \end{pmatrix} \begin{pmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{pmatrix} = \begin{pmatrix} \frac{\partial f}{\partial x}\frac{\partial x}{\partial u} + \frac{\partial f}{\partial y}\frac{\partial y}{\partial u} & \frac{\partial f}{\partial x}\frac{\partial x}{\partial v} + \frac{\partial f}{\partial y}\frac{\partial y}{\partial v} \end{pmatrix}.$$

Comparing entries gives

$$\frac{\partial f}{\partial u} = \frac{\partial f}{\partial x}\frac{\partial x}{\partial u} + \frac{\partial f}{\partial y}\frac{\partial y}{\partial u} \quad \text{and} \quad \frac{\partial f}{\partial v} = \frac{\partial f}{\partial x}\frac{\partial x}{\partial v} + \frac{\partial f}{\partial y}\frac{\partial y}{\partial v}.$$

Thus when differentiating f with respect to one of the "new" variables u or v, we get one term coming each "intermediate" variable x and y, obtained by multiplying the partials of f with respect to an intermediate variables times the partial of that intermediate variable with respect to the new variable.

The same pattern holds no matter how many variables (new or intermediate) are involved: if f depends on $x_1, \ldots, x_n$ and each x_i depends on some variables $u_1, \ldots, u_m$, then

$$\frac{\partial f}{\partial u_i} = \sum_{j=1}^{n} \frac{\partial f}{\partial x_j}\frac{\partial x_j}{\partial u_i},$$

which comes from comparing entires in various Jacobian matrices. This type of formula, or rather the dependence of f on the various variables, is often encoded in a "tree diagram", which you can find more information about in the book or my Math 290-2 notes.

Lecture 26: More on the Chain Rule

Warm-Up 1. Suppose $f : \mathbb{R} \to \mathbb{R}$ is differentiable and define $u : \mathbb{R}^2 \to \mathbb{R}$ by

$$u(x, y) = f(xy).$$

Then u is differentiable by the chain rule since it is the composition of f with the function $(x, y) \mapsto xy$. We claim that

$$x\frac{\partial u}{\partial x} - y\frac{\partial u}{\partial y} = 0.$$

Letting t be the variable of f, we have $t = xy$, so the chain rule gives:

$$\frac{\partial u}{\partial x} = \frac{\partial u}{\partial t}\frac{\partial t}{\partial x} = \frac{\partial f}{\partial t}y \quad \text{and} \quad \frac{\partial u}{\partial y} = \frac{\partial u}{\partial t}\frac{\partial t}{\partial y} = \frac{\partial f}{\partial t}x.$$

Thus

$$x\frac{\partial u}{\partial x} - y\frac{\partial u}{\partial y} = xy\frac{\partial f}{\partial t} - yx\frac{\partial f}{\partial t} = 0$$

as claimed. Note that since f is only a function of one variable, it is more common to denote its derivative using the standard $\frac{df}{dt}$ notation instead of $\frac{\partial f}{\partial t}$.

Warm-Up 2. Suppose $f, g : \mathbb{R}^n \to \mathbb{R}$ are both differentiable at $\mathbf{a}$. Let $fg : \mathbb{R}^n \to \mathbb{R}$ denote the function $(fg)(\mathbf{x}) = f(\mathbf{x})g(\mathbf{x})$. We derive the product rule:

$$D(fg)(\mathbf{a}) = Df(\mathbf{a})g(\mathbf{a}) + f(\mathbf{a})Dg(\mathbf{a})$$

from the chain rule. The key is in interpreting the product fg as a composition of functions: $fg = m \circ h$ where $h : \mathbb{R}^n \to \mathbb{R}^2$ and $m : \mathbb{R}^2 \to \mathbb{R}$ are the functions

$$h(\mathbf{x}) = (f(\mathbf{x}), g(\mathbf{x})) \quad \text{and} \quad m(x, y) = xy.$$

We have $Dm(x, y) = \begin{pmatrix} y & x \end{pmatrix}$, so by the chain rule gives:

$$D(fg)(\mathbf{x}) = Dm(h(\mathbf{x}))Dh(\mathbf{x}) = \begin{pmatrix} g(\mathbf{a}) & f(\mathbf{a}) \end{pmatrix} \begin{pmatrix} Df(\mathbf{a}) \\ Dg(\mathbf{a}) \end{pmatrix} = g(\mathbf{a})Df(\mathbf{a}) + f(\mathbf{a})Dg(\mathbf{a})$$

as desired. (To be clear, $Df(\mathbf{a})$ and $Dg(\mathbf{a})$ are $1 \times n$ matrices, and so make up the rows of the $2 \times n$ matrix $\left(\begin{smallmatrix} Df(\mathbf{a}) \\ Dg(\mathbf{a}) \end{smallmatrix}\right)$.)

Applications. Here is a typical type of problem which requires the use of the chain rule. Suppose that the temperature at a point (x, y) of a lake is given by

$$u(x, y) = x^2 e^y - xy^3$$

and that a duck moves in the lake according to the equation

$$r(t) = (\cos t, \sin t)$$

where t denotes time. (This says that the duck moves in a circle.) We want to know the rate at which the temperature changes as the ducks moves, which is given by the derivative $\frac{du}{dt}$. The chain rule gives:

$$\begin{aligned} \frac{du}{dt} &= \frac{\partial u}{\partial x}\frac{dx}{dt} + \frac{\partial u}{\partial y}\frac{dy}{dt} \\ &= (2xe^y - y^3)(-\sin t) + (x^2 e^y - 3xy^2)\cos t \\ &= (2e^{\sin t}\cos t - \sin^3 t)(-\sin t) + (e^{\sin t}\cos^2 t - 3\cos t \sin^2 t)\cos t, \end{aligned}$$

where in the last step we set $x = \cos t$ and $y = \sin t$ as given by the position of the duck.

Partials in polar coordinates. A common thing the chain rule is used for is expressing partial derivatives in terms of another set of coordinates, such as polar coordinates. Suppose $f : \mathbb{R}^2 \to \mathbb{R}$ is differentiable and that we use standard rectangular coordinates (x, y) for $\mathbb{R}^2$. In polar coordinates, $x = r\cos\theta$ and $y = r\sin\theta$, so

$$\frac{\partial f}{\partial r} = \frac{\partial f}{\partial x}\frac{\partial x}{\partial r} + \frac{\partial f}{\partial y}\frac{\partial y}{\partial r} = \cos\theta\frac{\partial f}{\partial x} + \sin\theta\frac{\partial f}{\partial y}$$

and

$$\frac{\partial f}{\partial \theta} = \frac{\partial f}{\partial x}\frac{\partial x}{\partial \theta} + \frac{\partial f}{\partial y}\frac{\partial y}{\partial \theta} = -r\sin\theta\frac{\partial f}{\partial x} + r\cos\theta\frac{\partial f}{\partial y}.$$

This is often summarized by writing

$$\begin{aligned}\frac{\partial}{\partial r} &= \cos\theta\frac{\partial}{\partial x} + \sin\theta\frac{\partial}{\partial y}\\ \frac{\partial}{\partial \theta} &= -r\sin\theta\frac{\partial}{\partial x} + r\cos\theta\frac{\partial}{\partial y},\end{aligned}$$

which is interpreted as an equality of *differential operators*: the process of applying the differential operator $\frac{\partial}{\partial r}$ to a function (i.e. the linear transformation from a space of functions to a space of functions given by differentiating with respect to r) is the same as the process of applying the differential operator $\cos\theta\frac{\partial}{\partial x} + \sin\theta\frac{\partial}{\partial y}$ to a function, and similarly for the second equality.

Now, note that the two equations above can be written in matrix form as:

$$\begin{pmatrix}\frac{\partial}{\partial r}\\ \frac{\partial}{\partial \theta}\end{pmatrix} = \begin{pmatrix}\cos\theta & \sin\theta\\ -r\sin\theta & r\cos\theta\end{pmatrix}\begin{pmatrix}\frac{\partial}{\partial x}\\ \frac{\partial}{\partial y}\end{pmatrix},$$

where we use "vectors" whose entries are operators instead of scalars. Think of this as a change of basis type of formula, telling us how to express operators with respect to the one set of coordinates (r,θ) in terms of operators with respect to another set (x,y). Using the inverse of the given matrix we get:

$$\begin{pmatrix}\frac{\partial}{\partial x}\\ \frac{\partial}{\partial y}\end{pmatrix} = \begin{pmatrix}\cos\theta & \sin\theta\\ -r\sin\theta & r\cos\theta\end{pmatrix}^{-1}\begin{pmatrix}\frac{\partial}{\partial r}\\ \frac{\partial}{\partial \theta}\end{pmatrix} = \frac{1}{r}\begin{pmatrix}r\cos\theta & -\sin\theta\\ r\sin\theta & \cos\theta\end{pmatrix}\begin{pmatrix}\frac{\partial}{\partial r}\\ \frac{\partial}{\partial \theta}\end{pmatrix},$$

which gives

$$\begin{aligned}\frac{\partial}{\partial x} &= \cos\theta\frac{\partial}{\partial r} - \frac{\sin\theta}{r}\frac{\partial}{\partial \theta}\\ \frac{\partial}{\partial y} &= \sin\theta\frac{\partial}{\partial r} + \frac{\cos\theta}{r}\frac{\partial}{\partial \theta},\end{aligned}$$

expressing (x,y)-derivatives in terms of (r,θ)-derivatives. Again, the point of these equalities is to say that the process of applying the thing on the left is the same as the process of applying the thing on the right.

Second-order chain rule. The chain rule can be extended to second-order (and higher) partial derivatives, which we we illustrate using the polar coordinate example above. Suppose $f:\mathbb{R}^2\to\mathbb{R}$ is C^2 and written with respect to x,y. Say we want to compute

$$\frac{\partial^2 f}{\partial\theta\partial r} = \frac{\partial}{\partial\theta}\left(\frac{\partial f}{\partial r}\right).$$

We saw above that

$$\frac{\partial f}{\partial r} = \cos\theta\frac{\partial f}{\partial x} + \sin\theta\frac{\partial f}{\partial y},$$

and we now want to differentiate this expression with respect to θ.

We start with:

$$\frac{\partial^2 f}{\partial\theta\partial r} = \frac{\partial}{\partial\theta}\left(\cos\theta\frac{\partial f}{\partial x}\right) + \frac{\partial}{\partial\theta}\left(\sin\theta\frac{\partial f}{\partial y}\right).$$

Now, each of $\cos\theta$ and $\frac{\partial f}{\partial x}$ depend on θ ($\frac{\partial f}{\partial x}$ depends on x, y, and so depends on r, θ as well), so the derivative of the first piece must be computed using the product rule, and similarly for the second piece:

$$\frac{\partial^2 f}{\partial\theta\partial r} = \left(\frac{\partial}{\partial\theta}\cos\theta\right)\frac{\partial f}{\partial x} + \cos\theta\frac{\partial}{\partial\theta}\left(\frac{\partial f}{\partial x}\right) + \left(\frac{\partial}{\partial\theta}\sin\theta\right)\frac{\partial f}{\partial y} + \sin\theta\frac{\partial}{\partial\theta}\left(\frac{\partial f}{\partial y}\right).$$

Now, since $\frac{\partial f}{\partial x}$ and $\frac{\partial f}{\partial y}$ both depend on x and y (and so on r and θ), differentiating these with respect to θ requires another chain rule:

$$\frac{\partial}{\partial\theta}\left(\frac{\partial f}{\partial x}\right) = \frac{\partial}{\partial x}\left(\frac{\partial f}{\partial x}\right)\frac{\partial x}{\partial\theta} + \frac{\partial}{\partial y}\left(\frac{\partial f}{\partial x}\right)\frac{\partial y}{\partial\theta} = -r\sin\theta\frac{\partial^2 f}{\partial x^2} + r\cos\theta\frac{\partial^2 f}{\partial y\partial x}$$

and similarly for $\frac{\partial}{\partial\theta}\left(\frac{\partial f}{\partial y}\right)$. (This can also be obtained using the expression for $\frac{\partial}{\partial\theta}$ in terms of $\frac{\partial}{\partial x}$ and $\frac{\partial}{\partial y}$ derived previously.) Putting it all together gives:

$$\begin{aligned}
\frac{\partial^2 f}{\partial\theta\partial r} &= \left(\frac{\partial}{\partial\theta}\cos\theta\right)\frac{\partial f}{\partial x} + \cos\theta\frac{\partial}{\partial\theta}\left(\frac{\partial f}{\partial x}\right) + \left(\frac{\partial}{\partial\theta}\sin\theta\right)\frac{\partial f}{\partial y} + \sin\theta\frac{\partial}{\partial\theta}\left(\frac{\partial f}{\partial y}\right) \\
&= -\sin\theta\frac{\partial f}{\partial x} + \cos\theta\left(-r\sin\theta\frac{\partial^2 f}{\partial x^2} + r\cos\theta\frac{\partial^2 f}{\partial y\partial x}\right) \\
&\quad + \cos\theta\frac{\partial f}{\partial y} + \sin\theta\left(-r\sin\theta\frac{\partial^2 f}{\partial x\partial y} + r\cos\theta\frac{\partial^2 f}{\partial y^2}\right) \\
&= -\sin\theta\frac{\partial f}{\partial x} - r\cos\theta\sin\theta\frac{\partial^2 f}{\partial x^2} + r(\cos^2\theta - \sin^2\theta)\frac{\partial^2 f}{\partial y\partial x} + \cos\theta\frac{\partial f}{\partial y} + r\sin\theta\cos\theta\frac{\partial^2 f}{\partial y^2},
\end{aligned}$$

where we use the fact that f is C^2 when combining the $\frac{\partial^2 f}{\partial y\partial x}$ and $\frac{\partial^2 f}{\partial x\partial y}$ terms. More concisely, we can express this as an equality of differential operators:

$$\frac{\partial^2}{\partial\theta\partial r} = -\sin\theta\frac{\partial}{\partial x} - r\cos\theta\sin\theta\frac{\partial^2}{\partial x^2} + r(\cos^2\theta - \sin^2\theta)\frac{\partial^2}{\partial y\partial x} + \cos\theta\frac{\partial}{\partial y} + r\sin\theta\cos\theta\frac{\partial^2}{\partial y^2}$$

Laplacians. A few weeks ago when we outlined the problem of "hearing the shape of a drum" (just for fun) we mentioned a certain operator called the Laplacian. We can now say precisely what this is, and mention a bit about why it is important.

The *Laplacian* operator on $\mathbb{R}^n$ is defined by

$$\frac{\partial^2}{\partial x_1^2} + \cdots + \frac{\partial^2}{\partial x_n^2}.$$

To be clear, this is the linear transformation from, say, the space of infinitely differentiable functions on $\mathbb{R}^n$ to itself defined by

$$f \mapsto \frac{\partial^2 f}{\partial x_1^2} + \cdots + \frac{\partial^2 f}{\partial x_n^2}.$$

(This is the transformation whose eigenvalues and eigenvectors were relevant in the "hearing the shape of a drum" problem.) An important realization is that the Laplacian on $\mathbb{R}^3$ can be expressed in terms of cylindrical or spherical coordinates in a fairly nice way. You'll work this out on the homework, which will involve expressing second-order partial derivatives in one set of coordinates in terms of another set using the chain rule.

Outside of the drum problem, the Laplacian has important applications in physics. In particular, the expression for the Laplacian in terms of spherical coordinates derived on the Homework gives

a straightforward description of the Laplacian operator on a sphere, and the knowing how to describe the kernel of this operator (which is possible given the spherical description) is crucial in understanding the quantum mechanics behind the behavior of a hydrogen atom. Look up "spherical harmonics" for more. We won't delve into this more and I only mention the Laplacian here as a key example of why being able to express second derivatives in terms of various coordinates is important. (Nonetheless, we might say a few things about the Laplacian next quarter in relation to integration.)

Lecture 27: Directional Derivatives

Warm-Up. Suppose A is an $n \times n$ matrix and $f : \mathbb{R}^n \to \mathbb{R}$ is C^2. Define the function $g : \mathbb{R}^n \to \mathbb{R}$ by $g(\mathbf{x}) = f(A\mathbf{x})$. We show that the Hessian matrix of g at a point is given by

$$D^2g(\mathbf{x}) = A^T D^2 f(A\mathbf{x})A.$$

This is meant to be a chain rule application: we can think of g as the composition of f with the linear transformation $\mathbb{R}^n \to \mathbb{R}^n$ determined by A, so we can think of g as what f becomes after making a *linear* change of variables. You saw on the homework that when making a non-linear change of variables (say when expressing the Laplacian in cylindrical or spherical coordinates), the effect on second-order partial derivatives is not-so-nice to describe, but the equality we are proving here gives a relatively simply description of this effect when the change of variables is linear. In the $n = 1$ case this says that the second derivative of $f(ax)$ with respect to x is $a^2 f''(ax)$, so this problem is meant to be a higher-dimensional version of this fact.

Expressing g as the composition $g = f \circ A$ shows that g is differentiable by the chain rule and that

$$Dg(\mathbf{x}) = Df(A\mathbf{x})DA(\mathbf{x}) = Df(A\mathbf{x})A,$$

where we use the fact that $DA(\mathbf{x}) = A$ since A is a linear transformation. (This is the higher-dimensional analogue of the fact that the derivative of $f(ax)$ with respect to x is $af'(ax)$.) To compute the derivative of Dg—and thereby the Hessian of g—we first figure out a nice way to express the product $Df(A\mathbf{x})A$. Denote the entries of A by a_{ij}. Then the product $Df(A\mathbf{x})A$ is a $1 \times n$ matrix whose i-th column is

$$a_{1i}\frac{\partial f}{\partial x_1}(A\mathbf{x}) + \cdots + a_{ni}\frac{\partial f}{\partial x_n}(A\mathbf{x}).$$

Differentiating this with respect to all possible variables gives the i-th row of $D^2g(\mathbf{x})$, which we can view as the $1 \times n$ Jacobian matrix of the function $\mathbb{R}^n \to \mathbb{R}$ defined by this expression. Thus the i-th row of $D^2g(\mathbf{x})$ is

$$D(a_{1i}\frac{\partial f}{\partial x_1}(A\mathbf{x}) + \cdots + a_{ni}\frac{\partial f}{\partial x_n}(A\mathbf{x})) = a_{1i}D\left(\frac{\partial f}{\partial x_1}(A\mathbf{x})\right) + \cdots + a_{ni}D\left(\frac{\partial f}{\partial x_n}(A\mathbf{x})\right).$$

Applying the expression for Dg we above only replacing f by $\frac{\partial f}{\partial x_i}$ gives that

$$D\left(\frac{\partial f}{\partial x_i}(A\mathbf{x})\right) = D\left(\frac{\partial f}{\partial x_i}\right)(A\mathbf{x})A.$$

(In other words, view $\frac{\partial f}{\partial x_i}(A\mathbf{x})$ as the composition $\frac{\partial f}{\partial x_i} \circ A$ and apply the chain rule.) Thus the i-th row of $D^2g(\mathbf{x})$ is

$$a_{1i}D\left(\frac{\partial f}{\partial x_1}\right)(A\mathbf{x})A + \cdots + a_{ni}D\left(\frac{\partial f}{\partial x_n}\right)(A\mathbf{x})A.$$

Factoring out an A from the right, this i-th row becomes

$$(a_{1i}D\left(\frac{\partial f}{\partial x_1}\right)(A\mathbf{x}) + \cdots + a_{ni}D\left(\frac{\partial f}{\partial x_n}\right)(A\mathbf{x}))A,$$

which can then be written as

$$\begin{pmatrix} a_{1i} & \cdots & a_{ni} \end{pmatrix} \begin{pmatrix} D\left(\frac{\partial f}{\partial x_1}\right)(A\mathbf{x}) \\ \vdots \\ D\left(\frac{\partial f}{\partial x_n}\right)(A\mathbf{x}) \end{pmatrix} A$$

where the first piece is a $1 \times n$ matrix and the second the $n \times n$ matrix whose rows are the given $1 \times n$ Jacobians. Thus all together, $D^2g(\mathbf{x})$ is

$$\begin{pmatrix} a_{11} & \cdots & a_{n1} \\ \vdots & \ddots & \vdots \\ a_{1n} & \cdots & a_{nn} \end{pmatrix} \begin{pmatrix} D\left(\frac{\partial f}{\partial x_1}\right)(A\mathbf{x}) \\ \vdots \\ D\left(\frac{\partial f}{\partial x_n}\right)(A\mathbf{x}) \end{pmatrix} A.$$

The first term is A^T and the second consists of the second-order partial derivatives of f evaluated at $A\mathbf{x}$, so this expression is indeed

$$D^2g(\mathbf{x}) = A^T D^2 f(A\mathbf{x})A$$

as desired. (Overall, the point of this problem is to illustrate how writing derivatives concisely in terms of Jacobians makes them simpler to work with. Try to compute $D^2g(\mathbf{x})$ using only first- and second-order partial derivatives to see why this approach is actually simpler.)

Directional derivatives. We have seen that partial derivatives give the rate of change of a function in certain directions (the x- and y-directions), or equivalently in the $\mathbb{R}^n \to \mathbb{R}$ they give the slope of the graph in certain directions. However, there are any number of other directions in which we can talk about the rate of change (or slope) of f.

Given $f : \mathbb{R}^n \to \mathbb{R}$, a point $\mathbf{a} \in \mathbb{R}^n$, and a unit vector $\mathbf{u} \in \mathbb{R}^n$, we define the *directional derivative* of f at $\mathbf{a}$ in the direction of $\mathbf{u}$, denoted by $D_{\mathbf{u}}f(\mathbf{a})$, by the limit:

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

The point is that as h varies, $\mathbf{a} + h\mathbf{u}$ describes the line through $\mathbf{a}$ in a direction parallel to $\mathbf{u}$, and $f(a + h\mathbf{u})$ looks at the behavior of f only along points on this line so that the given limit indeed gives the rate of change of f in this direction. Alternately, this directional derivative is the ordinary derivative of the single-variable function $h \mapsto f(\mathbf{a} + h\mathbf{u})$ at $h = 0$. (Note that $h = 0$ describes the point $\mathbf{a}$ we are looking at.)

One thing to note is that we only give this definition when $\mathbf{u}$ is a unit vector. This is to guarantee that the definition only depends on the direction of $\mathbf{u}$ and not on the length of a vector used to specify that direction: there are many vectors giving the same direction, but we want to get the same value no matter which such vector we choose.

Theorem. Suppose $f : \mathbb{R}^n \to \mathbb{R}$ is differentiable. We claim that in this case the direction derivative $D_{\mathbf{u}}f(\mathbf{a})$ is actually equal to

$$D_{\mathbf{u}}f(\mathbf{a}) = Df(\mathbf{a})\mathbf{u}.$$

Indeed, the single-variable function $h \mapsto f(\mathbf{a} + h\mathbf{u})$ can be written as the composition of $g : h \mapsto \mathbf{a} + h\mathbf{u}$ with f, so the chain rule (we need f to be differentiable so that this applies) gives:

$$Dg(0) = Df(g(0))Dg(0) = Df(\mathbf{a})\mathbf{u}$$

since differentiating $\mathbf{a} + h\mathbf{u}$ with respect to h gives $\mathbf{u}$.

Thus, when f is differentiable we get a very simple expression for directional derivatives. Note that this gives yet one more interpretation of the Jacobian $Df(\mathbf{a})$: it is the standard matrix of the linear transformation which sends a vector to the directional derivative of f at $\mathbf{a}$ in that specific direction.

Gradients. In the expression $Df(\mathbf{a})\mathbf{u}$, $Df(\mathbf{a})$ is thought of as a $1 \times n$ matrix, so that $Df(\mathbf{a})\mathbf{u}$ is a row vector times a column vector. Alternatively, we can think of $Df(\mathbf{a})$ a a vector and think of $Df(\mathbf{a})\mathbf{u}$ as a dot product instead. In this case we refer to $Df(\mathbf{a})$ as being the *gradient* of f at $\mathbf{a}$:

$$\nabla f(\mathbf{a}) = \begin{pmatrix} \frac{\partial f}{\partial x_1}(\mathbf{a}) & \cdots & \frac{\partial f}{\partial x_n}(\mathbf{a}) \end{pmatrix}.$$

Using this vector, the directional derivative is given by

$$D_{\mathbf{u}} f(\mathbf{a}) = \nabla f(\mathbf{a}) \cdot \mathbf{u}.$$

Thus, $\nabla f(\mathbf{a})$ and $Df(\mathbf{a})$ are essentially the same thing, only we use $Df(\mathbf{a})$ when thinking of this object as a matrix and we use $\nabla f(\mathbf{a})$ when thinking of it as a vector.

The gradient has some important geometric properties which are derived from the expression

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

for directional derivatives, where θ is the angle between $\nabla f(\mathbf{a})$ and $\mathbf{u}$. In particular, this directional derivative is at a maximum when $\mathbf{u}$ and $\nabla f(\mathbf{a})$ point in the same direction, in which case this maximum value is $\|\nabla f(\mathbf{a})\|$. Check my Math 290-2 notes for more about this and examples which illustrate how to apply this fact.

Lecture 28: Gradient Vectors

Remaining lecture. For now, I'll leave the rest of this lecture and following lecture to my Math 290-2 notes, which contains basically the same material. The main point of the final lecture is that gradients are perpendicular to level sets, which is made clear in my 290-2 notes.